جامعة سيدي محمد بن عبد الله بفاس
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵙⵉⴷⵉ ⵎⵓⵃⴰⵎⵎⴷ ⵓ ⵄⴱⴷⵍⵍⴰⵀ ⴱⴼⴰⵙ
UNIVERSITÉ SIDI MOHAMED BEN ABDELLAH DE FES

كلية العلوم ظهر المهراز فاس
ⵜⴰⵖⵢⵍⵓⵜ ⵓ ⵜⵣⵔⴼⵜ ⵎⴷⵓⵖ ⵎⵣⵔⴰⵔ ⴼⴰⵙ
FACULTE DES SCIENCES DHAR EL MAHRAZ FES

N° d'ordre : 40/2024

Centre des Etudes : ***Sciences et Techniques et Sciences Médicales***

Formation Doctorale : ***Sciences et Technologies de l'Information et de la Communication***

Spécialité :***Informatique***

Laboratoire : ***Laboratoire de Physique Appliquée, Informatique et Statistique***

# THÈSE DE DOCTORAT

Titre de la thèse :

## Predicting Students Dropout In MOOCS

Présentée par :

**ALJ Zakaria**

Soutenue le 22 Juillet 2024 devant le jury :

| Nom et Prénom | Grade | Etablissement | Qualité |
|---|---|---|---|
| Pr EL BEKKALI Moulhim | PES | Ecole Nationale des Sciences Appliquées, Fès | Président |
| Pr CHAOUI Nour El Houda | PES | Ecole Supérieure de Technologie, Kénitra | Rapporteur& examinateur |
| Pr EL AZAMI Ikram | MCH | Faculté des Sciences, Kénitra | Rapporteur& examinateur |
| Pr ZINEDINE Ahmed | PES | Faculté des Sciences Dhar El Mahraz, Fès | Rapporteur& examinateur |
| Pr BOUHOUTE Afaf | MCH | Faculté des Sciences Dhar El Mahraz, Fès | Examinateur |
| Pr BENNANI Mohamed Taj | MCH | Faculté des Sciences Dhar El Mahraz, Fès | Examinateur |
| Pr FARDOUSSE Khalid | MCH | Faculté Chariaa, Fès | Examinateur |
| Pr CHERKAOUI MALKI Mohammed Ouçamah | PES | Université Sidi Mohamed Ben Abdellah, Fès | Invité |
| Pr BOUAYAD Anas | MCH | Faculté des Sciences Dhar El Mahraz, Fès | Directeur de thèse |

Année Universitaire : 2023-2024

# Remerciements

La réalisation de cette thèse représente l'aboutissement de plusieurs années de travail, de recherche et de sacrifices. Elle n'aurait pas été possible sans le soutien, l'encadrement et les encouragements de nombreuses personnes. À tous, je souhaite exprimer ma plus profonde gratitude.

En premier lieu, je tiens à remercier chaleureusement M. Anas Bouayad, mon directeur de thèse principal. Votre expertise, votre patience et vos conseils avisés ont été essentiels tout au long de ce parcours. Vos orientations m'ont permis de surmonter de nombreux obstacles et d'apporter de la rigueur à mon travail. Votre soutien indéfectible et votre disponibilité ont été d'un grand réconfort dans les moments difficiles. Vous avez apporté une perspective unique et complémentaire, enrichissant ainsi ce projet de recherche de manière significative.

Je tiens également à exprimer ma plus profonde gratitude à M. Cherkaoui Malki Mohammad Oucamah, qui a commencé à m'encadrer dès le début et m'a soutenu jusqu'à sa retraite. Votre soutien initial et vos conseils avisés ont jeté les bases solides de ce travail. Merci pour votre dévouement et votre engagement.

Je souhaite également exprimer ma profonde gratitude aux rapporteurs de cette thèse et aux membres du jury pour avoir accepté d'évaluer ce travail et de participer à la soutenance. Vos remarques pertinentes et vos critiques constructives ont enrichi ce projet et m'ont permis de l'améliorer de manière significative. Votre présence et vos évaluations sont extrêmement précieuses et témoignent de votre soutien à ce travail de recherche. Merci pour votre temps et votre engagement.

Je souhaite également remercier la faculté des sciences Dhar Mahraz pour l'environnement de recherche stimulant et les ressources mises à ma disposition. Un grand merci à l'ensemble du personnel administratif et technique pour leur aide précieuse et leur professionnalisme.

Un remerciement particulier à mes collègues doctorants. Vos discussions enrichissantes, vos conseils et votre amitié ont grandement contribué à rendre ce voyage plus agréable et moins solitaire. Merci pour les moments de partage, les échanges intellectuels et les encouragements mutuels.

Je tiens aussi à exprimer ma gratitude envers mes enseignants et professeurs tout au long de mon parcours académique. Vos enseignements ont posé les fondations nécessaires à la réalisation de ce travail de recherche.

Un immense merci à ma famille pour votre amour, votre soutien inconditionnel et votre patience. Votre encouragement m'a permis de rester motivé et de persévérer malgré les défis rencontrés.

Enfin, je remercie mes amis pour leur compréhension et leur soutien tout au long de ces années. Vos encouragements et votre présence ont été une source inestimable de motivation et de réconfort.

À tous, je vous exprime ma sincère reconnaissance.

# Dédicace

À la mémoire de mon père, ton souvenir et ton amour ont été une source inépuisable de courage et d'inspiration. Ta présence m'a manqué à chaque étape, mais ton esprit et tes valeurs m'ont guidé jusqu'au bout.

À ma mère, sans qui je ne serais pas là. Ton soutien indéfectible et ton amour inconditionnel m'ont permis de surmonter tous les obstacles. Tu es et resteras toujours mon pilier.

À mon épouse, pour son amour, sa patience et son soutien constant tout au long de ce parcours. Merci d'avoir été mon roc et ma source de réconfort dans les moments de doute.

À mon petit-fils, source de joie et de motivation. Ta présence a illuminé mes journées et apporté une lumière nouvelle à ce travail.

À mes frères et sœurs, pour leur soutien moral et leurs encouragements. Vous avez toujours été là pour moi et m'avez aidé à traverser les moments difficiles. Votre solidarité et votre affection m'ont porté jusqu'ici.

À mes amis, pour leur soutien, leur écoute et leur amitié. Merci d'avoir été présents à chaque étape de ce voyage. Votre présence a été une source précieuse de motivation et de réconfort.

À tous ceux qui, de près ou de loin, ont contribué à ce travail, je vous dédie cette thèse avec une profonde gratitude.

# Résumé

Cette thèse aborde le défi significatif des taux élevés d'abandon dans les cours en ligne ouverts et massifs (MOOC), en proposant des solutions innovantes utilisant l'intelligence artificielle et l'apprentissage automatique. Les MOOCs offrent un accès gratuit à une éducation de qualité à des millions de personnes dans le monde. Cependant, ils sont confrontés à un problème important de taux d'abandon élevés, avec une moyenne de seulement 12,6% des étudiants complétant un cours.

L'étude commence par examiner les facteurs contribuant aux taux d'abandon, en présentant les méthodes d'analyse de l'apprentissage utilisées pour prédire l'abandon. Elle identifie des limitations dans les méthodes existantes, telles que la détection tardive des étudiants à risque, le manque de priorisation pour réduire le nombre d'étudiants qu'un instructeur doit gérer à un moment donné.

Pour surmonter ces limitations, la thèse introduit une approche qui utilise des techniques de réduction de la dimensionnalité et de sélection de caractéristiques pour traiter les données, suivies de la construction et de la prédiction de modèles. Le processus culmine dans la priorisation et la personnalisation des interventions, visant à fournir un soutien opportun et efficace aux étudiants à risque d'abandon. En outre, la thèse examine les déterminants multidimensionnels de la performance académique des étudiants dans les universités marocaines à travers une enquête statistique. De plus, la thèse explore l'impact de la gamification sur l'engagement et la motivation des étudiants dans les MOOCs. L'étude explore également les effets de la motivation sur l'apprentissage et la réussite des étudiants dans les MOOCs. Elle identifie plusieurs facteurs influençant la motivation des apprenants, tels que les besoins académiques, les intérêts personnels, les facteurs sociaux et technologiques. Elle discute également des stratégies pour améliorer la motivation et la rétention des apprenants dans les MOOCs.

Cette thèse contribue à une meilleure compréhension de l'abandon dans les MOOCs et propose des stratégies efficaces pour lutter contre ce problème. Cependant, il reste beaucoup à faire pour améliorer la rétention des apprenants dans les MOOCs. Les résultats et les méthodologies présentés dans ce travail sont censés servir de base pour de futures recherches dans ce domaine. Les recherches futures pourraient explorer l'utilisation de méthodes basées sur les graphes pour la prédiction de l'abandon, car ces méthodes peuvent modéliser des relations et des interactions complexes au sein du système éducatif, améliorant potentiellement la précision des prédictions. De plus, l'intégration des technologies de réalité virtuelle et augmentée (VR/AR) dans les MOOCs pourrait améliorer l'engagement et la rétention des étudiants en offrant des environnements d'apprentissage immersifs et interactifs.

**Mots-clés : MOOCs, Abandon, Réduction de dimensionnalité, Motivation, Gamification**

# Abstract

This thesis delves into the significant challenge of high dropout rates in Massive Open Online Courses (MOOCs), proposing innovative solutions using artificial intelligence and machine learning. MOOCs have gained popularity due to their ability to provide free access to quality education for millions of people worldwide. However, they face a substantial issue of high dropout rates, with an average of only 12.6% of students completing a course.

The study begins by examining the factors contributing to dropout rates, presenting a comprehensive literature review on learning analytics methods used to predict dropout. It identifies limitations in existing methods, such as late detection of at-risk students, lack of prioritization to significantly reduce the number of students an instructor or monitor must handle at a time, and the absence of clues for the monitor to propose personalized intervention.

To overcome these limitations, the thesis introduces an approach that utilizes dimensionality reduction and variable selection techniques to process data, followed by model construction and prediction. The process culminates in the prioritization and personalization of interventions, aiming to provide timely and effective support to students at risk of dropping out. Furthermore, the thesis examines the multidimensional determinants of academic performance among undergraduate students in Moroccan universities through a statistical survey. Additionally, the thesis investigates the impact of gamification on student engagement and motivation in MOOCs. Gamification, the application of game mechanisms in non-game contexts, is found to enhance learner motivation and improve their learning experience. The study also explores the effects of motivation on student learning and success in MOOCs. It identifies several factors influencing learner motivation, such as academic needs, personal interests, social and technological factors. It also discusses strategies to improve motivation and learner retention in MOOCs.

This thesis contributes to a better understanding of dropout in MOOCs and proposes effective strategies to combat this issue. However, much remains to be done to improve learner retention in MOOCs. The findings and methodologies presented in this work are expected to serve as a foundation for future research in this area. Future research could explore the use of graph-based methods for dropout prediction, as these methods can model complex relationships and interactions within the educational system, potentially improving prediction accuracy. Additionally, integrating virtual reality and augmented reality (VR/AR) technologies in MOOCs could enhance student engagement and retention by providing immersive and interactive learning environments.

**Keywords : MOOCs, Dropout, Dimensionality Reduction, Motivation, Gamification**

# ملخص

تحظى الدورات التعليمية المفتوحة عبر الإنترنت بشعبية كبيرة نظرًا لقدرتها على توفير تعليم عالي الجودة مجانًا لملايين الأشخاص حول العالم. ومع ذلك، يواجه العديد من الطلاب صعوبة في إكمال هذه الدورات، حيث أن معدل الإكمال لا يتجاوز 12.6% فقط . تناقش هذه الأطروحة التحدي الكبير المتمثل في ارتفاع معدلات الانسحاب في هذه الدورات، وتقدم حلولًا مبتكرة باستخدام الذكاء الاصطناعي وتعلم الآلة.

تفحص الدراسة العوامل التي تسهم في معدلات الانسحاب، وتستعرض بشكل شامل الأدبيات المتعلقة بطرق تحليل التعلم المستخدمة للتنبؤ بالانسحاب. وتحدد الدراسة قيودًا في الطرق الحالية، مثل الكشف المتأخر عن الطلاب المعرضين لخطرالانسحاب، وعدم إعطاء الأولوية لتقليل عدد الطلاب الذين يجب على المعلم أو المراقب التعامل معهم في وقت واحد، وغياب الآليات التي تساعد المراقب في اقتراح تدخلات مخصصة.

تقدم الأطروحة نهجًا يتغلب على هذه القيود باستخدام تقنيات تخفيض الأبعاد واختيار المتغيرات لمعالجة البيانات، يتبعها بناء النموذج والتنبؤ. تهدف هذه العملية إلى تحديد الأولويات وتخصيص التدخلات لتقديم دعم فعال وفي الوقت المناسب للطلاب المعرضين لخطر الانسحاب. بالإضافة إلى ذلك، تفحص الأطروحة المحددات متعددة الأبعاد للأداء الأكاديمي بين الطلاب الجامعيين في الجامعات المغربية من خلال دراسة إحصائية. كما تبحث في تأثير التحفيز من خلال الألعاب على مشاركة الطلاب وتحفيزهم في هذه الدورات. وتستنتج الدراسة أن التحفيز من خلال الألعاب، وهو تطبيق آليات الألعاب في سياقات غير الألعاب، يعزز تحفيز المتعلمين ويحسن تجربتهم التعليمية.

تستكشف الدراسة أيضًا تأثيرات التحفيز على تعلم الطلاب ونجاحهم في الدورات التعليمية المفتوحة عبر الإنترنت. وتحدد عدة عوامل تؤثر على تحفيز المتعلم، مثل الاحتياجات الأكاديمية، الاهتمامات الشخصية، والعوامل الاجتماعية والتكنولوجية. كما تناقش استراتيجيات لتحسين التحفيز والاحتفاظ بالمتعلمين في هذه الدورات.

تساهم هذه الأطروحة في فهم أفضل لمشكلة الانسحاب في الدورات التعليمية المفتوحة عبر الإنترنت، وتقترح استراتيجيات فعالة لمكافحة هذه المشكلة. ومع ذلك، يبقى الكثير لتحقيق تحسين أكبر في الاحتفاظ بالمتعلمين في هذه الدورات. من المتوقع أن تكون النتائج والمنهجيات المقدمة في هذا العمل أساسًا للبحث المستقبلي في هذا المجال. يمكن أن يستكشف البحث المستقبلي استخدام الطرق القائمة على الرسوم البيانية للتنبؤ بالانسحاب، حيث يمكن لهذه الطرق أن تنمذج العلاقات والتفاعلات المعقدة داخل النظام التعليمي، مما يحسن دقة التنبؤ. بالإضافة إلى ذلك، يمكن أن يعزز دمج تكنولوجيا الواقع الافتراضي والواقع المعزز في هذه الدورات من مشاركة الطلاب واحتفاظهم من خلال توفير بيئات تعليمية غامرة وتفاعلية.

**الكلمات المفتاحية : الدورات التعليمية المفتوحة عبر الإنترنت، الانسحاب، تقليل الأبعاد، التحفيز، التلعيب**

# Table des matières


**I Partie I : État de l'art** **18**

**1 Décrochage en MOOC** **19**

1.1 Introduction . . . 19

1.2 Facteurs de décrochage . . . 20

1.3 Jeu de données . . . 22

1.4 Apprentissage automatique dans la détection du décrochage . . . 24

1.5 Conclusion . . . 25

**2 Réduction de la dimensionnalité** **26**

2.1 Introduction . . . 26

2.2 Réduction basée sur la sélection des caractéristiques . . . 27

2.2.1 Processus de sélection . . . 27

2.2.2 Approche générale pour la sélection des fonctionnalités . . . 29

2.2.2.1 Filter . . . 29

2.2.2.2 Wrapper . . . 31

2.2.2.3 Embedded . . . 33

2.2.3 Revue de quelques méthodes de sélection . . . 35

2.2.3.1 SFS et SBS . . . 36

2.2.3.2 Branch and Bound . . . 38

2.2.3.3 FOCUS . . . 39

2.2.3.4 LVW et LVF . . . 40

2.2.3.5 Les Algorithmes Génétiques . . . 42

2.3 Réduction basée sur la transformation des caractéristiques . . . 46

2.3.1 Les méthodes linéaires . . . 46

2.3.1.1 Analyse en Composantes Principales . . . 46

2.3.1.2 Positionnement Multi-Dimensionnel . . . 47

2.3.2 Les méthodes non-linéaires . . . 49

2.3.2.1 Isomap . . . 49

2.3.2.2 T-SNE . . . 49

2.4 Conclusion . . . 50

**3 Apprentissage Automatique** **52**

3.1 Introduction . . . 52

3.2 L'apprentissage supervisé . . . 53

3.2.1 Classification . . . 53

3.2.1.1 Régression Logistique . . . 54

3.2.1.2 K-plus proches voisins (k-NN) . . . 54

3.2.1.3 Machines à vecteurs de support (SVM) . . . 55

3.2.1.4 Arbres de décision . . . 56

3.2.1.5 Forêts aléatoires . . . 57

3.2.1.6 Gradient Boosting . . . 58

3.2.1.7 Apprentissage profond . . . 59

3.2.2 Régression . . . 60

3.2.2.1 Régression linéaire . . . 60

3.2.2.2 Régression polynomiale . . . 60

3.2.2.3 Régression par processus gaussien . . . 61

3.2.2.4 Réseaux de neurones pour la régression . . . 61

3.2.2.5 Support Vector Regression (SVR) . . . 62

3.3 Apprentissage non supervisé . . . 63

3.3.1 Clustering . . . 63

3.3.1.1 K-means . . . 63

3.3.1.2 DBSCAN . . . 63

3.3.1.3 Clustering hiérarchique . . . 64

3.3.2 Détection d'anomalies . . . 64

3.3.2.1 Algorithme LOF (Local Outlier Factor) . . . 64

3.3.2.2 One-Class SVM . . . 65

3.4 Évaluation approfondie des modèles . . . 65

3.4.1 Métriques pour les problèmes de classification . . . 65

3.4.2 Métriques pour les problèmes de régression . . . 67

3.5 Validation croisée et sélection des modèles . . . 67

3.5.1 K-Fold Cross-Validation . . . 68

3.5.2 Leave-One-Out Cross-Validation . . . 68

3.5.3 Sélection des paramètres . . . 69

3.5.3.1 Recherche exhaustive (Grid Search) . . . 69

3.5.3.2 Recherche aléatoire (Random Search) . . . 69

3.5.3.3 Optimisation bayésienne . . . 70

3.6 Surrapprentissage, Sous-apprentissage et Techniques de Régularisation . . . . . 70
3.6.1 Surapprentissage et Sous-apprentissage . . . . . 70
3.6.2 Techniques de régularisation . . . . . 71
3.6.2.1 Régularisation $L_1$ (Lasso) . . . . . 71
3.6.2.2 Régularisation $L_2$ (Ridge) . . . . . 71
3.7 Classes non balancées . . . . . 71
3.8 Conclusion . . . . . 72

**II Partie II : Contributions 73**

**4 Prédiction du Décrochage des étudiants dans les MOOCs 74**
4.1 Introduction . . . . . 74
4.2 Travaux Connexes . . . . . 75
4.3 Présentation des données . . . . . 77
4.4 Ingénierie des caractéristiques . . . . . 80
4.5 Réduction de la dimensionalité . . . . . 83
4.6 Résultats . . . . . 84
4.7 Personnalisation de l'intervention . . . . . 92

**5 Déterminants Multidimensionnels de la Performance Académique 98**
5.1 Introduction . . . . . 98
5.2 Méthodologie . . . . . 99
5.2.1 Collecte des Données . . . . . 99
5.2.2 Analyse des Données . . . . . 100
5.2.3 Engagement Éthique . . . . . 100
5.2.4 Limites de l'Étude . . . . . 100
5.3 Questionnaire . . . . . 101
5.4 Résultats . . . . . 102
5.5 Discussion . . . . . 108
5.6 Conclusion . . . . . 111

**6 Effets de la gamification sur la motivation et l'engagement des étudiants 113**
6.1 Introduction . . . . . 113
6.2 Méthodologie . . . . . 114
6.2.1 Conception de l'étude . . . . . 114
6.2.2 Collecte de données . . . . . 114
6.2.3 Analyse des données . . . . . 114

6.2.4 Évaluation de la qualité . . . 114
6.3 Résultats . . . 115
6.3.1 Motivation . . . 116
6.3.2 Engagement . . . 117
6.4 Implications pratiques . . . 118
6.5 Conclusion . . . 118

**7 Théories de la motivation et les stratégies pour la rétention 119**
7.1 Introduction . . . 119
7.2 Méthodologie . . . 120
7.2.1 Conception de l'étude . . . 120
7.2.2 Stratégie de recherche . . . 120
7.2.3 Critères d'inclusion et d'exclusion . . . 121
7.2.3.1 Critères d'inclusion . . . 121
7.2.3.2 Critères d'exclusion . . . 121
7.2.4 Évaluation de la qualité . . . 121
7.2.5 Analyse des données . . . 121
7.3 Résultats . . . 122
7.3.1 Facteurs de motivation . . . 123
7.3.2 Stratégies motivationnelles et cadres théoriques . . . 125
7.3.3 Limitations pour améliorer la motivation et la rétention dans les MOOCs . . . 131
7.4 Discussion . . . 132
7.5 Conclusion . . . 134

# Liste des Figures


2.1 Procédure de sélection des caractéristiques . . . . . 27
2.2 Processus de l'approche Filter . . . . . 30
2.3 Processus de l'approche Wrapper . . . . . 31
2.4 Processus de l'approche Embedded . . . . . 33
2.5 Arborescence de B&B avec n=4 x=2 . . . . . 38
2.6 Les étapes de l'algorithme génétique . . . . . 42
2.7 Opérateur de croisement à un point . . . . . 44
2.8 Opérateur de croisement à deux points . . . . . 44
2.9 Exemple d'une opération de mutation . . . . . 45
2.10 Isomap sur des données non-linéaires . . . . . 49

3.1 Illustration du kernel trick : transformation des données séparées à l'aide d'un hyperplan 56
3.2 Matrice de confusion . . . . . 66
3.3 5-Fold Cross-Validation . . . . . 68

4.1 Nombre d'inscrits par cours . . . . . 79
4.2 Taux de décrochage par cours . . . . . 79
4.3 Analyse ROC Curve . . . . . 88
4.4 Évolution de la Précision de l'AG sur les Générations . . . . . 90
4.5 Évolution de la Perte à travers les Époques . . . . . 91
4.6 Précision par nombre de composantes . . . . . 91
4.7 Visualisation des points en 2D . . . . . 92
4.8 Erreur de prédiction sur Nombre de Logs avec ARIMA . . . . . 93
4.9 Interface de la page web . . . . . 94
4.10 Pourcentage de décrochage par semaine . . . . . 94

7.1 Statistique des articles examinés . . . . . 122

# Liste des Tables


1.1 Facteurs de décrochage . . . . . 20
1.2 Résumé des jeux de données des MOOCs . . . . . 23

2.1 Catégorisation des algorithmes de sélection des caractéristiques . . . . . 35
2.2 Matrice des distances entre les logs de 6 étudiants . . . . . 48

4.1 Informations sur les cours . . . . . 77
4.2 Informations sur les modules . . . . . 78
4.3 Informations sur les inscriptions . . . . . 78
4.4 Informations sur les logs . . . . . 78
4.5 Résultats des Étudiants . . . . . 79
4.6 Caractéristiques extraites . . . . . 82
4.7 Classement des caractéristiques . . . . . 85
4.8 Scores des modèles sur les données d'apprentissage . . . . . 86
4.9 Scores des modèles . . . . . 87
4.10 Effet de la réduction sur les scores . . . . . 89
4.11 Scores des modèles AG et DL . . . . . 90

5.1 Facteurs démographiques et académiques associés aux résultats des étudiants . . . . . 103
5.2 Facteurs psychologiques, sociaux et de mode de vie associés aux résultats académiques des étudiants . . . . . 105

6.1 Approches et méthodes utilisées dans les études extraites . . . . . 115
6.2 Éléments de gamification et leur impact sur les caractéristiques des étudiants . . . . . 116
6.3 Caractéristiques des étudiants affectés par la gamification . . . . . 116

7.1 Aperçu des catégories et types de motivations influençant la réussite des MOOCs . . . . 123
7.2 Catégorisation des Cadres Théoriques dans la Recherche sur les MOOCs . . . . . 126
7.3 Catégories de stratégies pour améliorer la motivation et la rétention dans les MOOCs . 127

# Liste des Algorithmes


1 Algorithme Filter . . . . . 30
2 Algorithme Wrapper . . . . . 32
3 Algorithme Embedded . . . . . 34
4 Algorithme SFS . . . . . 37
5 Algorithme SBS . . . . . 37
6 Algorithme FOCUS . . . . . 39
7 Algorithme LVW . . . . . 40
8 Algorithme LVF . . . . . 41
9 Minutes Connectées . . . . . 81

# Liste des Acronymes

**MOOC** Massive Open Online Course

**KDD** Knowledge Discovery in Databases

**SMOOC** Social Massive Open Online Course

**LA** Learning Analytics

**EDM** Educational Data Mining

**VLE** Virtual Learning Environment

**ML** Machine Learning

**LR** Logistic Regression

**GA** Genetic Algorithm

**CFIN** Context Feature Interaction Network

**SVM** Support Vector Machine

**RF** Random Forest

**XGBoost** Extreme Gradient Boosting

**ROC** Receiver Operating Characteristic

**AUC** Area Under the Curve

**DNN** Deep Neural Network

**NN** Neural Network

**GB** Gradient Boosting

**DT** Decision Tree

**SFS** Sequential Forward Selection

**SBS** Sequential Backward Selection

**LVW** Las Vegas Wrapper

**LVF** Las Vegas Filter

**APC** Principal Component Analysis

**MDS** Multidimensional Scaling

**TSNE** t-Distributed Stochastic Neighbor Embedding

**KNN** K-Nearest Neighbors

**DL** Deep Learning

**GPT** Generative Pre-trained Transformer

**RNN** Recurrent Neural Network

**RMSE** Root Mean Square Error

**MAE** Mean Absolute Error

**SVR** Support Vector Regression

**SMOTE** Synthetic Minority Over-sampling Technique

**ADASYN** Adaptive Synthetic Sampling

**RFE** Recursive Feature Elimination

**ARIMA** AutoRegressive Integrated Moving Average

**GPA** Grade Point Average

**PRISMA** Preferred Reporting Items for Systematic Reviews and Meta-Analyses

**MMAT** Mixed Methods Appraisal Tool

**PICOC** Population, Intervention, Comparator, Outcome, Context

**SDT** Self-Determination Theory

**SRLT** Self-Regulated Learning Theory

**CVTAE** Cross-Validation and Test Averaged Estimation

**EVT** Expectancy-Value Theory

**AGT** Achievement Goal Theory

**ECT** Extrinsic Cognitive Theory

**CLT** Cognitive Load Theory

**ALT** Activity Learning Theory

**ELT** Experiential Learning Theory

**UDLTAM** Unified Design and Learning Theory of Adaptive Multimedia

**FT** Flow Theory

**TTAT** Technology Threat Avoidance Theory

**MRT** Media Richness Theory

**TNE** Theory of Narrative Ethics

**CHAT** Cultural-Historical Activity Theory

**SCT** Social Cognitive Theory

# Introduction Générale

Suite à l'émergence des nouvelles technologies numériques visant la modernisation du système éducatif traditionnel, les cours en ligne ouverts et massifs ont gagné en popularité ces dernières années [1]. Depuis leur apparition en 2008, les MOOCs ont gagné en notoriété et sont devenus un mode d'apprentissage populaire, notamment pendant l'épidémie de COVID-19. En 2020, 16300 cours ont été proposés par 950 universités, le nombre d'inscriptions ayant atteint plus de 180 millions d'apprenants dans le monde [2]. Ces MOOCs ont été proposés par de nombreux fournisseurs, dont Udacity, Coursera, FutureLearn et edX [3].

Les MOOCs offrent un accès gratuit à l'éducation et au développement professionnel pour toutes les personnes ayant accès à internet, en fournissant de nombreuses ressources pédagogiques précieuses aux apprenants. Ils sont considérés comme un pilier de la démocratisation de l'éducation. En effet, les articles récents sur les MOOCs dans la littérature scientifique et la presse populaire soulignent le fait que des centaines de milliers de personnes dans le monde ont désormais accès à des cours offerts par des universités d'élite. La plupart des MOOCs sont gratuits et ne suggèrent aucun frais pour suivre les cours, n'exigent aucun pré-requis autre que l'intérêt de l'étudiant pour le sujet et une bonne connexion internet [4]. Avec les récents changements des besoins du marché de l'emploi et les exigences académiques, une évolution vers les MOOCs comme une alternative de ressource d'apprentissage a été constatée dans la communauté étudiante. Les MOOCs sont devenus une source idéale d'auto-développement qui permet de combler le fossé entre les exigences de l'industrie et les compétences acquises dans le cursus universitaire [5]. Malgré ces avantages qui apportent une amélioration substantielle à l'expérience d'apprentissage des étudiants, le défi le plus important auquel les MOOCs sont confrontés est la faible rétention des apprenants. La moyenne des taux de complétion d'un MOOC est de 12,6% [6]. Par exemple, dans un MOOC proposé par Duke University en 2012, parmi les 12175 étudiants inscrits, 7761 ont déjà regardé une vidéo. Seulement 3658 ont répondu au moins à un quiz, 346 ont tenté l'examen final et 313 ont reçu un certificat, indiquant un taux d'achèvement d'environ 3% [7]. Les MOOCs offerts par des universités d'élite telles que MIT, Stanford et Berkley ont rapporté un taux d'abandon pouvant atteindre 95% [8].

Les approches pédagogiques traditionnelles peinent à identifier à l'avance les étudiants en risque d'abandon, en raison du grand nombre d'inscrits. Il en résulte une difficulté à concevoir une intervention efficace pour prévenir le décrochage. Ce taux élevé d'abandon dans les MOOCs a incité les chercheurs à utiliser les méthodes d'analyse de l'apprentissage (Learning Analytics) [9] pour prédire de manière précoce les apprenants en danger de décrocher. L'analyse de l'apprentissage permet d'analyser les données de trace ou de journal collectées pendant que les étudiants interagissent avec les cours en

ligne. À partir de ces données, des modèles de prédiction peuvent être construits grâce à l'apprentissage automatique supervisé. Cela permet de former un modèle mathématique capable de distinguer automatiquement les décrocheurs des persévérants en fonction de leur comportement et de leurs interactions précédentes pendant le cours. En se basant sur les résultats obtenus, une intervention appropriée peut être conçue et prodiguée aux étudiants à risque. L'analyse de l'apprentissage aide à relever le défi de l'identification à grande échelle des étudiants à risque de décrochage tout en permettant également de répondre à l'exigence d'une intervention précoce [10].

La construction de modèles de prédiction du décrochage en utilisant l'analyse de l'apprentissage a révélé un énorme potentiel pour améliorer la conception des interventions pour réduire les taux d'abandon dans les MOOCs. Cependant, ces modèles de prédiction sont limités, en raison du grand nombre de décrocheurs, soit par 1) la détection tardive des étudiants à risque d'abandon, soit par 2) l'absence de priorité pour réduire considérablement le nombre d'étudiants que l'instructeur ou le moniteur doit traiter à chaque fois. Finalement, 3) les modèles de prédiction ne fournissent aucun indice pour que le moniteur puisse proposer une intervention personnalisée.

Le but de ce mémoire est d'utiliser l'intelligence artificielle et l'apprentissage supervisé pour combler ces trois limitations. Ce mémoire décrira le processus suivi pour résoudre les limitations de l'analyse de l'apprentissage, comprenant la présentation et l'analyse de données, le traitement de données (réduction de la dimensionnalité et sélection de caractéristiques), la construction de modèles, la prédiction, et enfin la priorité et la personnalisation des interventions. La réduction de la dimensionnalité a montré des résultats positifs en réduisant le bruit et la complexité du problème, ce qui a permis d'améliorer la précision des modèles de prédiction. En outre, cette thèse examine également les déterminants multidimensionnels de la performance académique des étudiants dans les universités marocaines, à travers une étude statistique par questionnaire, afin de mieux comprendre les facteurs qui influencent leur réussite et leur échec.Les résultats de cette analyse ont révélé plusieurs facteurs clés, notamment les facteurs démographiques tels que l'âge, le sexe et le milieu socio-économique. De plus, les facteurs académiques comme les habitudes d'étude et les antécédents éducatifs jouent un rôle important. Enfin, les facteurs psychologiques et sociaux, tels que la motivation intrinsèque et extrinsèque, le soutien social des pairs et de la famille, ainsi que le sentiment d'appartenance à l'institution universitaire, sont également déterminants pour la performance académique des étudiants.

Cette thèse examine également l'impact de la gamification sur l'engagement et la motivation des apprenants dans les MOOCs. La gamification, qui est l'application de mécanismes de jeu dans des contextes non ludiques, est de plus en plus utilisée dans l'éducation en ligne pour augmenter la motivation des apprenants et améliorer leur expérience d'apprentissage. Les résultats de cette étude ont montré que la gamification a un impact positif significatif sur la motivation des apprenants, en augmentant leur engagement et leur persévérance dans les cours en ligne. Par exemple, les données collectées indiquent

que l'utilisation d'éléments de gamification, tels que les badges, les classements et les récompenses, a permis d'améliorer les taux de participation et de rétention des étudiants, avec une augmentation de 20% du taux de complétion des cours pour ceux incluant des éléments de gamification par rapport aux cours sans ces éléments..

Enfin, nous explorons les effets de la motivation sur l'apprentissage et la réussite des étudiants dans les MOOCs. Nous examinons les différents facteurs qui influencent la motivation des apprenants, tels que les besoins académiques, les intérêts personnels, les facteurs sociaux et technologiques, et discutons des stratégies pour améliorer la motivation et la rétention des apprenants dans les MOOCs.

Cette thèse est divisée en deux parties principales : la première partie est consacrée à l'état de l'art, et la deuxième partie est axée sur l'application pratique de l'analyse de l'apprentissage pour améliorer la rétention des apprenants dans les MOOCs.

Le document est structuré comme suit :

- **État de l'art** - Cette partie comprend trois chapitres :
  - Le premier chapitre traite du décrochage dans les MOOCs, en examinant les raisons pour lesquelles les apprenants abandonnent ces cours en ligne.
  - Le deuxième chapitre explore les méthodes de réduction de la dimensionnalité, une technique essentielle pour analyser les grandes quantités de données générées par les MOOCs.
  - Le troisième chapitre se concentre sur les techniques d'apprentissage automatique utilisées pour prédire et analyser le comportement des apprenants dans les MOOCs.
- **Contributions** - Cette partie comprend cinq chapitres :
  - Le premier chapitre présente une application pratique des techniques d'apprentissage sur les données de la KDD Cup 2015, une compétition de data mining.
  - Le deuxième chapitre examine les déterminants multidimensionnels de la performance académique des étudiants dans les universités marocaines, tels que les facteurs démographiques, académiques, psychologiques, sociaux et liés au mode de vie, à travers une étude statistique par questionnaire.
  - Le troisième chapitre examine les effets de la gamification sur la motivation et l'engagement des étudiants dans les MOOCs.
  - Le quatrième chapitre discute des théories de la motivation et propose des stratégies pour améliorer la rétention des apprenants dans les MOOCs.

Enfin, une conclusion générale résume les principaux résultats et contributions de cette thèse, et propose des pistes pour des recherches futures, notamment l'utilisation de méthodes basées sur les graphes pour prédire le décrochage et l'intégration des technologies de réalité virtuelle et augmentée (VR/AR) dans les MOOCs pour améliorer l'engagement et la rétention des apprenants.

# Partie I : État de l'art

# Chapitre 1

# Décrochage en MOOC

## 1.1 Introduction

Les cours en ligne ouverts et massifs (MOOC) sont apparus dans le domaine des technologies éducatives en 2008. L'idée du MOOC a continué à se développer de manière rapide et en 2012, le New York Times a déclaré cette année comme "l'année des MOOC" en raison de l'apparition de nombreuses grandes plates-formes de MOOC telles que Coursera, EdX et Udacity [11]. Le nombre d'inscriptions aux MOOC a connu une croissance explosive. Par exemple, à l'Université de Californie, plus de 250000 étudiants étaient inscrits à six cours qui ont commencé en janvier 2013. Au cours de sa première année, Coursera avait signé 62 partenariats universitaires et collégiaux, offrant plus de 250 cours gratuits à 3,1 millions d'utilisateurs.

L'étude de l'évolution des MOOCs a fait l'objet d'analyses depuis le siècle dernier. Afin de répondre aux évolutions rapides du marché de l'emploi et des exigences académiques, de nombreux étudiants dans le monde considèrent de plus en plus les MOOCs comme des contenus d'apprentissage qui les aident à combler le fossé entre les compétences requises par l'industrie et celles acquises dans les universités. Les MOOCs offrent un choix idéal pour de nombreuses raisons, en fournissant un accès ouvert, en ligne et à faible coût à une éducation de haute qualité et aux cours des universités d'élite et des experts de l'industrie. Les MOOCs sont également un outil idéal pour le développement professionnel autodirigé, grâce aux ressources de cours disponibles en ligne pour les étudiants à tout moment. Au cours des 15 dernières années, l'intérêt pour les MOOCs a augmenté en raison de l'augmentation de l'offre éducative et des besoins en matière d'apprentissage [12]. Dans le contexte de l'éducation virtuelle, il existe différents fondements didactiques et pédagogiques selon les différents types de MOOCs. Plusieurs auteurs [13, 14, 15, 16] ont classé les MOOCs en : (1) cMOOC (Connectivist MOOC), où le travail collaboratif et la coopération sont favorisés ; (2) xMOOC (eXtended MOOC), où l'interaction est fortement liée à la relation élève-enseignant et l'évaluation se concentre sur les questions fermées ; (3) MadeMOOC, qui favorise l'utilisation de vidéos, de l'interactivité et de la co-évaluation ; (4) SynchMOOC, qui fixe des délais temporels pour les cours et les évaluations ; (5) adaptiveMOOC, qui permet le développement d'évaluations adaptatives et dynamiques ; (6) gMOOC, qui inclut une plus large mesure la gamification ; (7) SMOOC (Social Massive Open Online Course), qui favorise une interaction accrue dans l'apprentissage, une accessibilité constante et l'ubiquité des ressources pédagogiques ; (8)

iMOOC (intelligent MOOC), qui favorise la personnalisation de la formation.

## 1.2 Facteurs de décrochage

Malgré la diversification et les avantages apportant une amélioration significative à l'expérience d'apprentissage des étudiants, les MOOC sont confrontés à divers défis, tels que le taux élevé d'abandon. La moyenne des taux de complétion pour un MOOC est de 12,6% selon une étude [6]. Ce taux élevé d'abandon peut être attribué à de nombreux facteurs, qui peuvent généralement être divisés en deux catégories : 1) les facteurs liés aux étudiants, 2) les facteurs liés au MOOC. Cela est illustré dans la table 1.1.

Table 1.1: Facteurs de décrochage

| Les facteurs de décrochage | |
|---|---|
| facteurs liés aux étudiants | Manque de motivation<br>Manque de temps<br>Manque de pré-requis et de bases |
| facteurs liés aux MOOCs | Conception des cours<br>Isolement et manque d' interactivité<br>Coûts cachés |

1. **Les facteurs liés aux étudiants**
   - Manque de motivation : Le manque de motivation est considéré comme l'un des facteurs les plus influents sur la décision des étudiants d'abandonner un MOOC. Selon [17], la motivation peut être stimulée par de nombreux facteurs tels que l'avenir économique, le développement personnel et professionnel, le défi, la réussite, le plaisir et l'amusement. Il est donc important d'examiner les motivations qui incitent les étudiants à s'inscrire à un MOOC. Les auteurs de [18] ont interrogé des étudiants pour étudier leurs motivations. Ils ont constaté que 95% des étudiants considèrent le divertissement et le plaisir comme la principale raison de s'inscrire à un MOOC, 87% ont choisi l'intérêt général pour un sujet et 15% ont choisi de s'inscrire pour aider à choisir entre les spécialités dans l'enseignement supérieur. Seulement 10% des étudiants ont décidé de s'inscrire à un MOOC par manque de moyens pour suivre une éducation formelle.
   - Manque de temps : Le manque de temps est un autre facteur qui peut empêcher les étudiants de suivre les exigences d'un MOOC. Selon une enquête menée par Belanger [19], regarder les vidéos, faire les devoirs et les quiz dans un MOOC peut prendre beaucoup de temps, ce qui peut inciter les étudiants à abandonner.

— Manque de pré-requis et de connaissances de base : Le manque de connaissances de base et de compétences requises pour suivre un cours peut également être une raison pour laquelle les étudiants abandonnent un MOOC. En particulier, les difficultés en mathématiques peuvent être perçues comme un problème qui empêche les étudiants de terminer les cours [19]. Comme une grande partie de l'interaction dans les MOOCs repose sur le texte, il est nécessaire pour les étudiants d'avoir des compétences en lecture, écriture et dactylographie solides, en plus des compétences techniques. Le manque de ces compétences est souvent perçu comme une cause de l'abandon dans les MOOCs [5].

2. **Les facteurs liés aux MOOCs**

— Conception des cours : La conception du cours est un des facteurs clés qui peut inciter les étudiants à abandonner les MOOCs. Elle repose sur trois composants : le contenu, la structure et la technologie de diffusion. Parmi ces trois éléments, le contenu est le plus déterminant pour la rétention. En effet, les étudiants ayant participé à des MOOCs ont indiqué qu'ils sont plus intéressés par les cours qui fournissent un contenu concret accompagné d'exemples et d'applications pratiques [20]. De plus, les étudiants apprécient les cours qui donnent des conseils et des astuces ainsi que les cours sur les soft skills. Les problèmes liés au contenu, tels que la difficulté technique du cours, la complexité du langage utilisé, le nombre élevé de modules, sont certaines des réclamations rapportées par les étudiants qui n'ont pas terminé le MOOC [21].

— Isolation et manque d'interactivité : L'isolation est un autre facteur qui a un impact direct sur les étudiants abandonnant les MOOCs. Une enquête menée par Hone [22] a montré que les étudiants ne participent pas aux discussions et aux débats, ce qui réduit l'interaction et la rétroaction entre le conférencier et les étudiants. Les étudiants ont également signalé dans l'enquête que l'absence de travail d'équipe et la manque de communication entre les étudiants ont créé un sentiment d'isolement.

— Coûts cachés : Les coûts cachés dans un MOOC sont une autre raison qui peut causer un taux élevé d'abandon. Ils représentent des frais qui doivent parfois être payés par les étudiants pour obtenir leur certificat ou pour acheter des manuels coûteux recommandés par les enseignants des cours [23].

L'identification et l'exploration des facteurs ayant un impact direct sur le décrochage des étudiants en MOOC permettent aux chercheurs, enseignants et technologues en éducation d'étudier et de proposer de nouvelles stratégies et techniques de rétention pour aider les étudiants à persévérer plus longtemps et à réussir dans les MOOC.

## 1.3 Jeu de données

La prédiction du succès ou de l'échec de l'apprentissage en ligne (E-learning) chez un élève est un des sujets les plus fréquemment abordés dans l'exploration de données éducatives (EDM) et de l'analyse de l'apprentissage (LA) (Learning Analytics) [24]. EDM et LA se basent sur l'analyse des données liées aux logs et aux informations démographiques des étudiants pour améliorer l'efficacité de l'apprentissage en révélant des informations utiles pour la modification de la structure des cours ou la prédiction de la performance et du comportement des étudiants. Les approches de prédiction actuelles reposent généralement sur ces méthodes. Les chercheurs dans le domaine de l'éducation sont préoccupés par la prédiction de la probabilité de décrochage d'un étudiant dans un cours, surtout dès les premières semaines [25]. Le facteur temporel est important dans le processus de prédiction de décrochage, car certaines analyses ont montré que 75% des abandons surviennent dans les premières semaines [26]. La prédiction du décrochage grâce à l'analyse de l'apprentissage a montré un grand potentiel pour réduire les taux d'abandon dans les MOOCs. Les logs obtenus lors de l'interaction des étudiants avec les cours MOOC sont collectés et organisés sous forme de vecteurs, ou exportés vers des fichiers Excel ou csv pour former un jeu de données. Dans la littérature, plusieurs jeux de données d'analyse de l'apprentissage sont disponibles pour les chercheurs. La sous-section suivante donne un aperçu des jeux de données les plus couramment utilisés dans les travaux de recherche antérieurs sur l'analyse de l'apprentissage.

Une prédiction efficace dépend fortement de l'utilisation de données fiables et significatives [27]. Selon Ifenthaler et Yau [28], augmenter le nombre de caractéristiques (variables) n'assure pas nécessairement une prédiction optimale. Ils ont conclu que la fiabilité et la validité des données et leur analyse rigoureuse sont essentielles pour obtenir une meilleure prédiction. L'augmentation de données peu significatives peut en effet nuire à la complexité et au temps de calcul.

De nombreuses études ont examiné l'impact des caractéristiques sur le taux de rétention en utilisant différentes méthodes statistiques. Guo et Reinecke ont analysé l'activité des étudiants dans quatre MOOCs sur edX [29] en utilisant la régression linéaire multiple sur des données de 140546 étudiants. Ils ont constaté une corrélation positive entre l'âge et le niveau d'études avec le volume de matériel d'apprentissage étudié par un apprenant. Dans une étude similaire, Cisel [30] a analysé les données de 3029 étudiants inscrits dans un xMOOC français en utilisant le logiciel R pour identifier les indicateurs clés du taux d'achèvement. Les résultats ont montré que le taux d'achèvement dépend principalement du statut professionnel et des contraintes de temps, et que la participation active des étudiants aux forums peut améliorer leur performance globale. Morris [31] a également examiné les relations entre l'achèvement et les caractéristiques démographiques des étudiants inscrits dans cinq MOOCs en utilisant des méthodes non paramétriques. Il a constaté une forte corrélation entre l'achèvement et l'expérience antérieure avec les MOOCs, le niveau d'études, le statut professionnel et l'âge des participants.

La table 1.2 donne un aperçu des ensembles de données disponibles pour la recherche sur les MOOC provenant de différentes plateformes.

TABLE 1.2: Résumé des jeux de données des MOOCs

| Dataset | Année | Description |
|---|---|---|
| Student Academics Performance Data Set [32] | 2018 | 300 instances, 22 caractéristiques nominaux, 43 cours, utilisé pour prédire le rendement scolaire des élèves |
| Embeddings and topic vectors for MOOC lectures dataset [33] | 2019 | L'ensemble de 12 032 vidéos et de 200 cours de Coursera. |
| NAEP 2019 : Educational Data Mining Competition [34] | 2019 | 438 392 étudiants, 5 programmes de cours. |
| Data for : Effectiveness of flip teaching on engineering students [35] | 2019 | 1 233 étudiants, les matières de physique et d'électricité sur 4 ans, cours inversés et méthodes pédagogiques traditionnelles. |
| OULAD [36] | 2017 | 32 592 étudiants, plus de 10 millions de lignes de données sur le VLE, 22 cours, données disponibles sous forme de 7 fichiers csv. |
| Khan Academy [37] | 2018 | 1,4 million d'exercices complétés par 47 495 étudiants, 69 différents types d'exercices. |
| XuetangX (OpenEdx) [38] | 2016 | Contenu des cours, inscription des étudiants, ensembles de données du journal d'accès au MOOC, 120 542 échantillons de 39 cours. |
| KDD Cup 2015 [39] | 2015 | 39 cours, 7 fichiers csv. |
| XuetangX [40] | 2017 | 11 cours, 56 800 000 enregistrements horodatés. |
| Cours de Harvard et MIT [41] | 2014 | 597 692 participants, 15 cours, 800 000 fichiers journaux. |
| edX [41] | 2014 | 8 cours en edX et 14 en FutureLearn, un total de 22 cours, 8 935 inscrits. |
| Coursera [42] | 2015 | 3 millions d'étudiants, 5 000 messages sur le forum. |
| Open University China [43] | 2015 | 62 375 étudiants inscrits en trois semestres. |

Harvard et MIT qui hébergent de nombreux cours en ligne sur la plateforme edX, fournissent des

ensembles de données pour les études analytiques sur l'apprentissage. La Open University UK, la plus grande université du Royaume-Uni avec environ 170000 étudiants inscrits dans divers programmes, propose des supports de cours via un environnement virtuel d'apprentissage (VLE) . L'ensemble de données OULAD comprend les données de 32592 étudiants et 22 modules, ainsi que des informations démographiques sur les étudiants, les dates d'évaluation, les scores et les données d'interaction avec le VLE. D'autres ensembles de données tels que KDD Cup et XuetangX ne comprennent que des données de logs et n'ont aucune information démographique sur les étudiants.

## 1.4 Apprentissage automatique dans la détection du décrochage

Plusieurs approches d'apprentissage automatique (ML) ont été adoptées pour détecter les élèves susceptibles de décrocher dans un MOOC. Kizilcec [44] a appliqué une méthode de regroupement (clustering) pour catégoriser les étudiants en fonction de leurs schémas d'engagement. Le cluster le plus notable comprenait les apprenants qui sont restés engagés tout au long du cours sans passer d'évaluations. De la même manière, une enquête systématique sur l'abandon des MOOC a été menée à l'aide d'une Régression Logistique (LR) sur les données de 100000 étudiants inscrits dans 21 filières [45]. L'étude a révélé que la probabilité d'abandon augmente considérablement si les étudiants se désengagent pendant 14 jours ou plus. De plus, la probabilité de se réengager augmente avec le nombre de vidéos que les élèves ont visionnées avant l'absence. La régression logistique a également été utilisée pour prédire l'abandon des étudiants dans les MOOCs avec une sélection de variables [46] et aussi avec des algorithmes génétiques (GA) pour le réglage des paramètres d'un classificateur [39]. Feng [47] a proposé un réseau d'interaction de caractéristiques basé sur le contexte (CFIN) pour prédire les potentiels décrocheurs dans deux MOOC. Des expériences ont été effectuées pour évaluer l'efficacité du CFIN par rapport à des méthodes de classification courantes telles que LR, les machines à vecteurs de support (SVM) et la forêt aléatoire (RF). En outre, une méthode d'ensemble a été conçue en combinant CFIN avec XGBoost . CFIN a mieux performé que les méthodes de base en termes d'aire sous la courbe ROC (AUC) et de score F1. Dans le même ordre d'idées, la méthode d'ensemble RF a été utilisée pour trouver les caractéristiques les plus importantes qui influencent les performances des élèves [41].

Des algorithmes de machine learning courants ont été utilisés pour prédire les performances des étudiants dans les MOOC. Les résultats ont montré que RF était supérieur en termes de précision, de sensibilité et de coefficient kappa de Cohen. Les réseaux de neurones profonds (DNN) ont également été utilisés pour prédire la réussite et l'abandon des élèves dans un MOOC [48]. Plusieurs modèles prédictifs ont été construits après une analyse pour sélectionner les caractéristiques les plus importantes. De plus, des règles d'association ont été extraites pour découvrir des similitudes dans les modèles de comportement des élèves. Un réseau de neurones à propagation (NN feedforward) a également été utilisé pour résoudre le problème de décrochage dans les MOOC [49]. Le modèle a produit une grande

précision et un faible taux de faux négatifs par rapport aux méthodes de classification courantes.

Liang a utilisé le Gradient Boosting (GB) pour prédire le décrochage des étudiants dans les MOOCs [50]. Les données sur les activités d'apprentissage des étudiants de 39 cours ont été collectées et utilisées pour créer des modèles d'arbre de décision (DT) avec une précision de 89%. Récemment, un modèle de régression vectorielle amélioré a été utilisé pour la même tâche [27]. Le modèle était basé sur les comportements d'apprentissage des étudiants collectés sur une base hebdomadaire. La prédiction des performances obtenues surpassaient les autres méthodes de classification.

Quel que soit l'algorithme choisi, une grande partie de l'efficacité d'un modèle de classification dépend de la sélection adéquate des vecteurs de caractéristiques. Les caractéristiques sélectionnées doivent être pertinentes et informatives pour le problème à résoudre. Dans le contexte des MOOCs, cela peut inclure des caractéristiques liées aux comportements d'apprentissage des étudiants, à leur engagement dans le cours, à leurs interactions avec les autres participants, et bien plus encore. Cependant, la sélection de caractéristiques peut être un défi en soi, en particulier lorsque l'on travaille avec des ensembles de données de grande dimension. C'est pourquoi le chapitre suivant se concentrera sur la réduction de la dimensionnalité, une technique essentielle pour simplifier les ensembles de données, réduire le bruit et le sur-apprentissage, et améliorer l'efficacité des modèles de classification.

## 1.5 Conclusion

Ce chapitre a exploré l'évolution des MOOCs, leur importance croissante dans le paysage éducatif actuel, ainsi que les défis auxquels ils sont confrontés, notamment le taux élevé d'abandon. Nous avons examiné les facteurs qui contribuent à ce taux d'abandon, qui peuvent être liés soit aux étudiants, soit aux MOOCs eux-mêmes. Nous avons également discuté de l'importance des données dans l'analyse des MOOCs, en mettant en lumière plusieurs ensembles de données disponibles pour la recherche. Enfin, nous avons examiné comment l'apprentissage automatique peut être utilisé pour prédire le décrochage des étudiants dans les MOOCs, en soulignant l'importance de la sélection des caractéristiques pour la performance des modèles de classification. Dans le chapitre suivant, nous approfondirons la question de la réduction de la dimensionnalité, une technique essentielle pour gérer les ensembles de données de grande dimension et améliorer l'efficacité des modèles de classification.

# Chapitre 2

# Réduction de la dimensionnalité

## 2.1 Introduction

La réduction de dimensionnalité est un domaine de recherche en constante évolution, qui s'applique principalement aux problèmes qui impliquent un grand nombre de variables, tels que la visualisation des données, l'apprentissage automatique, la génomique, le traitement d'images, la reconnaissance de formes et la médecine. Dans de nombreux domaines, la résolution du problème repose sur un ensemble de caractéristiques qui modélisent ce problème. Cependant, l'augmentation non significative du nombre de variables (caractéristiques) peut entraîner des difficultés telles que la complexité du problème, le temps de calcul et la dégradation du système de résolution en présence de données bruyantes [51]. Par conséquent, la réduction de dimensionnalité facilite la classification, la visualisation et la compression des données de grande dimension. La réduction de dimensionnalité consiste à trouver une représentation optimale des données initiales dans un espace plus petit. Les méthodes de réduction de dimensionnalité peuvent être classées en deux catégories.

- **Sélection des caractéristiques** : une réduction basée sur la sélection des caractéristiques consiste à sélectionner un sous-ensemble de caractéristiques les plus pertinentes à partir de l'ensemble de données initiales.
- **Transformation des données** : une réduction basée sur la transformation des données consiste à transformer les données de l'ensemble de départ en utilisant une combinaison des caractéristiques pour obtenir une dimension réduite.

Dans la sélection de caractéristiques, la représentation originale des caractéristiques n'est pas modifiée. La sélection de caractéristiques est généralement préférée par rapport à la transformation lorsque l'on souhaite conserver la signification et la sémantique originales des caractéristiques, et déterminer lesquelles d'entre elles sont les plus importantes.

Le reste de ce chapitre se concentrera sur les méthodes de sélection de caractéristiques en décrivant d'abord les mécanismes de sélection et les avantages et inconvénients des différentes techniques de sélection. Ensuite, nous examinerons les techniques de réduction basées sur la transformation des données en introduisant les différentes approches, telles que les approches linéaires et non linéaires

## 2.2 Réduction basée sur la sélection des caractéristiques

### 2.2.1 Processus de sélection

La sélection de caractéristiques est définie comme un processus de recherche pour trouver un sous-ensemble optimal de caractéristiques parmi l'ensemble de départ [52]. Le processus de sélection de caractéristiques comprend quatre étapes, illustrées sur la figure 2.1 : génération du sous-ensemble, évaluation du sous-ensemble, critère d'arrêt et validation des résultats. La sélection de caractéristiques peut être modélisée comme suit :

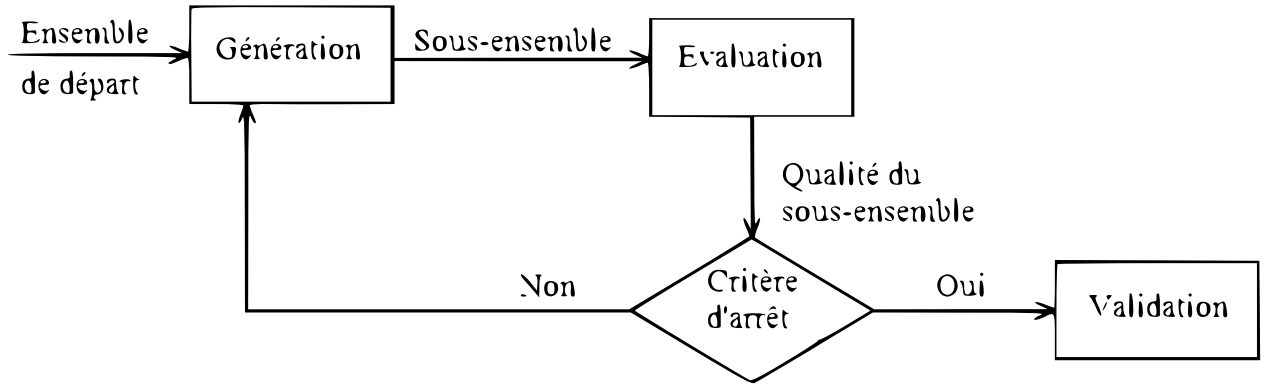



FIGURE 2.1: Procédure de sélection des caractéristiques

1. **Génération** : La génération de sous-ensembles est un processus de recherche heuristique, qui peut commencer par un ensemble vide et ajouter des caractéristiques (Forward selection), ou commencer par l'ensemble entier et supprimer des caractéristiques (Backward elimination). Des stratégies comme la recherche exhaustive, la recherche séquentielle ou heuristique et la recherche aléatoire ont été proposées [53].
   - **Recherche exhaustive** : Trouver le sous-ensemble minimal de caractéristiques est un problème NP-complet, nécessitant une recherche exhaustive qui devient coûteuse en temps de calcul avec l'augmentation du nombre de caractéristiques. Des techniques comme Branch and bound et Best first peuvent réduire le nombre de sous-ensembles examinés sans compromettre les chances de trouver le sous-ensemble optimal [54, 55].
   - **Recherche heuristique** : La recherche heuristique, incluant la recherche vers l'avant (SFS), la recherche vers l'arrière (SBS), et la recherche bidirectionnelle (BDS), est une approche courante pour la sélection des caractéristiques. Elle est simple à mettre en œuvre et rapide, mais peut ne pas explorer complètement l'espace de recherche.
   - **Recherche aléatoire** : Cette méthode génère aléatoirement un nombre limité de sous-ensembles de caractéristiques pour sélectionner le plus optimal. Les algorithmes de recherche aléatoire convergent rapidement vers une solution approchée optimale en évitant les optimums locaux dans l'espace de recherche.

2. **Évaluation** :

Dans le cadre du processus de recherche aléatoire, chaque sous-ensemble généré doit être évalué à l'aide d'un critère d'évaluation. Les critères d'évaluation peuvent être regroupés en deux catégories : les critères dépendants, influencés par les algorithmes de classification, et les critères indépendants, insensibles aux algorithmes de classification. Il est important de choisir un critère d'évaluation adéquat pour assurer la qualité et la fiabilité des résultats.

— **Critères indépendants** : Utilisés dans les algorithmes d'évaluation de type filtre, ils se basent sur les caractéristiques intrinsèques des données sans impliquer d'algorithmes d'apprentissage. Parmi les critères indépendants les plus populaires figurent les mesures de distance, les mesures d'information, les mesures de dépendance et les mesures de cohérence.

— **Mesures de distance** : Jouent un rôle crucial dans l'évaluation de la qualité des caractéristiques dans la classification des données. Elles permettent de mesurer la différence entre les probabilités conditionnelles associées aux différentes classes. Les types populaires incluent la divergence de Kullback-Leibler, la distance Euclidienne et la distance de Mahalanobis [52].

— *Distance Euclidienne* : $d(x, y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$.

— *Distance de Mahalanobis* : Prend en compte la covariance entre les caractéristiques.

— *Distance de Jaccard* : Utilisée pour les ensembles non numériques, mesure la similarité entre deux ensembles.

— *Distance de Cosinus* : Utilisée pour les vecteurs et les ensembles numériques, mesure l'angle entre deux vecteurs.

— **Mesures d'information** : Quantifient l'importance d'une caractéristique en termes d'apport d'informations. Elles se classent en mesures basées sur l'entropie et mesures basées sur la réduction d'incertitude. L'entropie de Shannon est un exemple de mesure basée sur l'entropie, tandis que le gain d'information est une mesure basée sur la réduction d'incertitude [56].

— **Mesures de dépendance** : Évaluent la force de la relation entre deux caractéristiques aléatoires. Elles sont utilisées pour déterminer la capacité à prédire la valeur d'une variable à partir de la valeur d'une autre. Des exemples incluent la corrélation de Pearson, la corrélation de Spearman, l'indice de Kendall et le coefficient de détermination.

— **Mesures de cohérence** : Diffèrent des mesures précédentes par leur forte dépendance aux informations de classe. Elles tentent de trouver un nombre minimum de caractéristiques pouvant séparer les classes avec la même justesse que l'ensemble complet. Une incohérence est définie comme deux instances ayant les mêmes valeurs des caractéristiques mais des étiquettes de classe différentes.

— **Critères dépendants** : Utilisés principalement dans le modèle d'évaluation de type Wrapper, ces critères sélectionnent les caractéristiques les mieux adaptées à l'algorithme de minage (Mining algorithm) prédéterminé, offrant souvent des performances supérieures. Cependant, les caractéristiques sélectionnées sont très dépendantes de l'algorithme utilisé et peuvent ne pas convenir pour d'autres [57]. Par exemple, dans une tâche de classification, la précision (Accuracy) est une mesure principale et peut être utilisée comme un critère dépendant pour la sélection des fonctionnalités, bien que cela puisse être coûteux en termes de calcul [58]. Pour le clustering, le modèle wrapper évalue la qualité d'un sous-ensemble de caractéristiques par la qualité des clusters résultants, en utilisant des critères heuristiques tels que la compacité des clusters, la séparabilité, la dispersion et le maximum de vraisemblance.

3. **Critère d'arrêt** : Un critère d'arrêt détermine le moment où le processus de sélection des fonctionnalités doit s'arrêter. Les critères d'arrêt fréquemment utilisés sont :

   a) La recherche est terminée.

   b) Un nombre spécifié fixé par l'utilisateur est atteint (nombre minimum de fonctionnalités ou nombre maximum d'itérations).

   c) L'ajout ou la suppression de toute caractéristique ne produit pas un meilleur sous-ensemble.

   d) Un sous-ensemble suffisamment pertinent est sélectionné (par exemple, un sous-ensemble peut être suffisamment bon si son erreur de classification est inférieure au taux d'erreur autorisé).

4. **Validation** : Le moyen le plus simple pour valider les résultats consiste à utiliser les connaissances préalables sur les données. Comme dans le cas des données synthétiques, si nous connaissons à l'avance les caractéristiques pertinentes, on peut comparer cet ensemble connu de caractéristiques avec les caractéristiques sélectionnées. Ainsi de même on s'attend que les caractéristiques redondantes ou non informatives ne soient pas sélectionnés. Cependant, en cas d'absence de telles connaissances préalables. Il faut surveiller le changement des performances de l'algorithme de minage avec le changement des fonctionnalités.

Tenant compte des quatre étapes du processus de sélection des caractéristiques, les approches utilisées fonctionnent d'une manière très diverse par rapport aux méthodes de recherche et aux critères d'évaluation. Dans la section suivante, nous distinguerons les trois approches utilisées dans la littérature : Filter , Wrapper et Embedded.

### 2.2.2 Approche générale pour la sélection des fonctionnalités

#### 2.2.2.1 Filter

L'approche Filter est une méthode de sélection de caractéristiques utilisée en phase de pré-traitement des données d'un modèle, évaluant les caractéristiques selon les propriétés des données d'apprentissage

et étant indépendante du processus d'apprentissage.

Comme illustré dans la figure 2.2, ce processus comprend trois étapes principales :

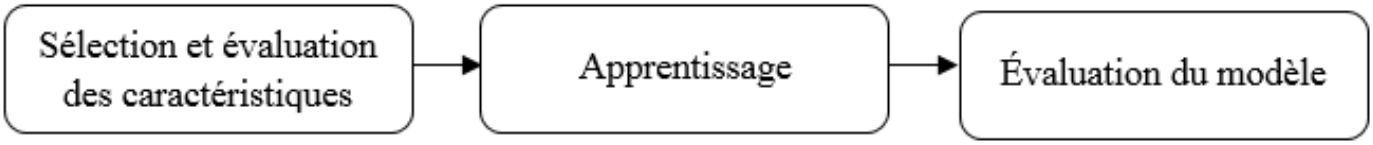



FIGURE 2.2: Processus de l'approche Filter

1. **Sélection et évaluation des caractéristiques :** Sélectionne et évalue les caractéristiques en fonction des propriétés des données d'apprentissage.
2. **Apprentissage :** Le modèle utilise les caractéristiques sélectionnées pour s'entraîner.
3. **Évaluation du modèle :** Évalue la performance du modèle pour déterminer l'efficacité des caractéristiques sélectionnées et leur impact sur la performance globale du modèle.

Les avantages de l'approche Filter incluent son indépendance vis-à-vis de l'algorithme utilisé, son faible coût en temps de calcul et sa robustesse face au sur-apprentissage. Cependant, son inconvénient majeur est qu'elle ignore l'effet du sous-ensemble sélectionné sur la performance de l'algorithme d'apprentissage [59, 60].

**Algorithm 1** Algorithme Filter

**Entrées :**
  $D(f_1, f_2, ..f_n)$ ▷ Ensemble d'apprentissage avec n caractéristiques
  $S_0$ ▷ Ensemble de départ
  $S_c$ ▷ Critère d'arrêt
  $S_{optimal}$ ▷ Sous-ensemble optimal
**Sortie :** $S_{optimal}$
**Début :**
  **Initialisation :** $S_{optimal} \leftarrow S_0$
  $\gamma_{optimal} \leftarrow \text{eval}(S_0, D, M)$ ▷ Évaluation de $S_0$ dans D selon le critère M
**repeat**
    $S \leftarrow \text{generate}(D)$ ▷ Génération du sous-ensemble pour l'évaluation
    $\gamma \leftarrow \text{eval}(S, D, M)$ ▷ Évaluation de S
  **if** $\gamma$ est meilleur que $\gamma_{optimal}$ **then**
      $S_{optimal} \leftarrow S$
      $\gamma_{optimal} \leftarrow \gamma$
**until** $S_c$ est atteint
**Retour** $S_{optimal}$
**Fin**

Le processus de l'approche Filter commence avec un sous-ensemble initial $S_0$ et utilise une stratégie de recherche (exhaustive, heuristique, aléatoire) pour trouver l'ensemble optimal. Chaque sous-ensemble généré est évalué et comparé au précédent selon un critère d'évaluation $M$ (mesure de distance, d'information, de dépendance, etc.). La recherche continue jusqu'à atteindre un critère d'arrêt $S_C$, et le dernier meilleur sous-ensemble $S_{optimal}$ est retenu comme résultat final.

#### 2.2.2.2 Wrapper

Le principal inconvénient de l'approche Filter est que le sous-ensemble des caractéristiques sélectionnées n'a pas d'effet sur les performances de l'algorithme d'apprentissage. Pour résoudre ce problème, Kohavi et John ont introduit le concept du Wrapper pour la sélection des caractéristiques [61]. L'approche Wrapper utilise la précision de l'algorithme d'apprentissage pour évaluer la pertinence d'une variable.

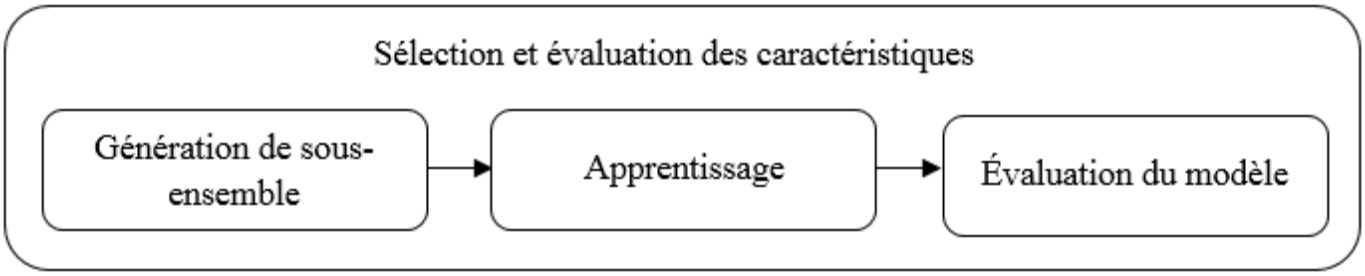



FIGURE 2.3: Processus de l'approche Wrapper

Le processus de l'approche Wrapper est illustré dans la Figure 2.3 et se déroule en trois étapes principales :

1. **Génération de sous-ensemble :** Dans cette étape, plusieurs sous-ensembles de caractéristiques sont générés. Ces sous-ensembles peuvent être créés de manière aléatoire, systématique, ou en utilisant des méthodes heuristiques. L'objectif est de considérer différentes combinaisons de caractéristiques pour trouver celles qui sont les plus pertinentes.
2. **Apprentissage :** Chaque sous-ensemble de caractéristiques généré est utilisé pour entraîner l'algorithme d'apprentissage. L'algorithme est appliqué aux données en utilisant uniquement les caractéristiques du sous-ensemble considéré. Cette étape permet de construire un modèle prédictif basé sur les caractéristiques sélectionnées.
3. **Évaluation du modèle :** Les performances du modèle sont évaluées en fonction de la précision de l'algorithme d'apprentissage. Les sous-ensembles de caractéristiques qui produisent les meilleurs résultats en termes de précision sont considérés comme les plus pertinents. Cette évaluation peut être réalisée à l'aide de techniques de validation croisée pour assurer la robustesse des résultats.

Les algorithmes qui utilisent l'approche Wrapper sont souvent considérés comme meilleures que celles de l'approche Filter, en termes de performance de prédiction[62]. Cependant, les caractéristiques

sélectionnées par cette méthode sont spécifiquement adaptées à l'algorithme de classification utilisé et peuvent ne pas être appropriées si un autre classificateur est utilisé. De plus, la complexité de l'algorithme d'apprentissage rend l'utilisation de la méthode Wrapper très coûteuse en temps de calcul contrairement à l'approche Filter. L'approche Wrapper utilise un algorithme de classification 'A' au lieu d'une mesure 'M' pour évaluer les sous-ensembles. Le fonctionnement de l'approche Wrapper est présenté dans l'algorithme ci-dessous.

**Algorithm 2** Algorithme Wrapper

**Entrées :**
  $D(f_1, f_2, ..f_n)$ ▷ Ensemble d'apprentissage avec n caractéristiques
  $S_0$ ▷ Ensemble de départ
  $S_c$ ▷ Critère d'arrêt
  $S_{optimal}$ ▷ Sous-ensemble optimal
**Sortie :** $S_{optimal}$
**Début :**
  **Initialisation :** $S_{optimal} \leftarrow S_0$
  $\gamma_{optimal} \leftarrow \text{eval}(S_0, D, A)$ ▷ Évaluation de $S_0$ par l'algorithme A
**repeat**
    $S \leftarrow \text{generate}(D)$ ▷ Génération du sous-ensemble pour l'évaluation
    $\gamma \leftarrow \text{eval}(S, D, A)$ ▷ Évaluation de S
  **if** $\gamma$ est meilleur que $\gamma_{optimal}$ **then**
      $S_{optimal} \leftarrow S$
      $\gamma_{optimal} \leftarrow \gamma$
**until** $S_c$ est atteint
**Retour** $S_{optimal}$
**Fin**

L'algorithme d'approche Wrapper illustré dans l'algorithme 2 généralisé est très similaire à l'algorithme de filtre, à ceci près qu'il utilise un algorithme de minage prédéfini A plutôt qu'une mesure indépendante M pour évaluer le sous-ensemble de caractéristiques. En conséquence, différents algorithmes de minage produiront des résultats de sélection de caractéristiques différents. Plusieurs algorithmes appartenant à la famille Wrapper peuvent être obtenus en variant les stratégies de recherche, les techniques de génération de l'ensemble de départ et les algorithmes de minage (A).

L'approche Wrapper permet ainsi de sélectionner les caractéristiques les plus pertinentes en tenant compte de leur impact direct sur les performances de l'algorithme d'apprentissage, offrant ainsi une amélioration par rapport à l'approche Filter.

#### 2.2.2.3 Embedded

L'approche Embedded permet d'effectuer la sélection des caractéristiques de manière plus efficace et intégrée, en exploitant les capacités intrinsèques des algorithmes d'apprentissage pour identifier et utiliser les caractéristiques les plus pertinentes, ce qui peut conduire à des modèles plus performants et plus simples.

Contrairement aux approches Filter et Wrapper, l'approche intégrée (Embedded) combine la sélection des caractéristiques et l'apprentissage du modèle en un seul processus. Dans cette approche, la sélection des caractéristiques est effectuée de manière intrinsèque par l'algorithme d'apprentissage au cours de l'entraînement. L'approche Embedded tire parti des mécanismes internes de certains algorithmes pour identifier les caractéristiques les plus pertinentes.

Le processus de l'approche Embedded est illustré dans la Figure 2.4 et se déroule en deux étapes principales :

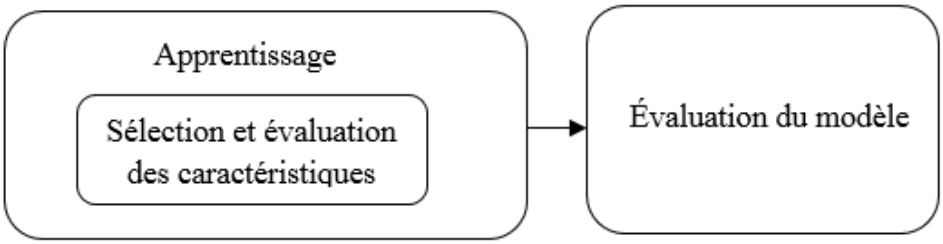



FIGURE 2.4: Processus de l'approche Embedded

1. **Sélection et évaluation des caractéristiques :** Cette étape se produit simultanément avec l'apprentissage du modèle. L'algorithme d'apprentissage intègre des mécanismes pour évaluer la pertinence des caractéristiques pendant l'entraînement. Par exemple, les arbres de décision peuvent utiliser la réduction de l'impureté pour sélectionner les caractéristiques les plus importantes. Les méthodes de régularisation telles que Ridge ou Lasso, intègrent également la sélection des caractéristiques en imposant des pénalités sur les coefficients des caractéristiques moins pertinentes.
2. **Évaluation du modèle :** Après la sélection des caractéristiques intégrée dans l'apprentissage, le modèle est évalué en fonction de sa précision et de ses performances globales. Cette évaluation peut également utiliser des techniques de validation croisée pour garantir la robustesse et la généralisation du modèle.

Cette approche permet de profiter des avantages des deux modèles, Filter et Wrapper. En effet, elle interagit avec l'algorithme d'apprentissage à un coût de calcul inférieur à celui de l'approche Wrapper. Tout comme l'approche Filter, elle capture les dépendances entre les fonctionnalités, et considère non seulement les relations entre une caractéristique et la variable de sortie, mais recherche également

les caractéristiques qui permettent une meilleure discrimination locale. Le modèle Wrapper est particulièrement adapté pour gérer de grands ensembles de données [63].

**Algorithm 3** Algorithme Embedded

**Entrées :**
  $D(f_1, f_2, ..f_n)$ ▷ Ensemble d'apprentissage avec n caractéristiques
  $S_0$ ▷ Ensemble de départ
  $S_c$ ▷ Critère d'arrêt
  $S_{optimal}$ ▷ Sous-ensemble optimal
**Sortie :** $S_{optimal}$
**Début :**
  **Initialisation :** $S_{optimal} \leftarrow S_0$
  $\gamma_{optimal} \leftarrow \text{eval}(S_0, D, M)$ ▷ Évaluation de $S_0$ par la mesure M
  $\beta_{optimal} \leftarrow \text{eval}(S_0, D, A)$ ▷ Évaluation de $S_0$ par l'algorithme A
  $C_0 \leftarrow C(S_0)$ ▷ Cardinalité de $S_0$
**for** $k = C_0$ **to** $n$ **do**
  **for** $i = 0$ **to** $n - k$ **do**
    $S \leftarrow \text{generate}(D)$ ▷ Génération du sous-ensemble de cardinalité c pour l'évaluation
    $\gamma \leftarrow \text{eval}(S, D, M)$ ▷ Évaluer le sous-ensemble courant S par la mesure M
    **if** $\gamma$ est meilleur que $\gamma_{optimal}$ **then**
      $S'_{optimal} \leftarrow S$
      $\gamma_{optimal} \leftarrow \gamma$
    $\beta \leftarrow \text{eval}(S'_{optimal}, D, A)$ ▷ Évaluation de $S'_{optimal}$ par l'algorithme A
    **if** $\beta$ est meilleur que $\beta_{optimal}$ **then**
      $S_{optimal} \leftarrow S'_{optimal}$
      $\beta_{optimal} \leftarrow \beta$
    **else**
      **Break** et retourne $S_{optimal}$
**Retour** $S_{optimal}$
**Fin**

L'approche Embedded utilise à la fois une mesure indépendante M et un algorithme de minage A pour évaluer des sous-ensembles de caractéristiques. Elle utilise d'abord la mesure indépendante pour déterminer les meilleurs sous-ensembles pour une cardinalité donnée. Ensuite, l'algorithme de minage est utilisé pour sélectionner le meilleur sous-ensemble final parmi les meilleurs sous-ensembles trouvés à travers différentes cardinalités. En gros, l'approche Wrapper lance la recherche à partir d'un sous-ensemble donné S0 (généralement un ensemble vide dans la sélection séquentielle) et effectue une itération pour trouver les meilleurs sous-ensembles à chaque cardinalité croissante. À chaque tour,

pour un sous-ensemble avec la cardinalité $c_0$, l'algorithme recherche tous les sous-ensembles possibles de cardinalité $c_0 + 1$ en ajoutant une fonctionnalité parmi les fonctionnalités restantes

### 2.2.3 Revue de quelques méthodes de sélection

Dans cette section, nous présentons quelques algorithmes de sélection de caractéristiques les plus couramment utilisés dans la littérature. Nous avons choisi de présenter des algorithmes appartenant aux différentes approches que nous avons introduit dans la section précédente, basées sur les diverses techniques de recherche et d'évaluation.

Il existe un vaste ensemble d'algorithmes de sélection de fonctionnalités disponibles. Afin de mieux comprendre le fonctionnement de chaque algorithme ainsi que les points communs et les différences entre eux, nous avons développé un cadre de catégorisation en trois dimensions, présenté dans le tableau 2.1, qui repose sur les discussions précédentes.

TABLE 2.1: Catégorisation des algorithmes de sélection des caractéristiques

| | | Stratégies de recherche | | | |
|---|---|---|---|---|---|
| | | Complète | Heuréstique | | Aléatoire |
| Filter | Distance | B&B [64], BFF[65], BOBRO [66], OBLIVIIN [67] | Relief [68], Relief[69], ReliffS [70], SFS [71], SBS [72] | | |
| | Information | MDLM[73] , CARDIE [74], | DTM [74], Koller[75], FG [76] , FCBF [77], | Dash's [78], SBUD [79] , | |
| | Dépendance | Bobrowski's [73] | CFS [80] , RRESET[81], POE+ACC [82] , DVMM[83] | Mitra's [84] | |
| | Consistence | FOCUS [85] , ABB [86] , MIFESI [87] , Schlimmer's [88] , | Set Cover [89] , WINNOW[90] | | LIV [76], QBB [76], LVF [91] , |
| Wrapper | Précision du modèle ou qualité du groupement | BS [92], AMB&B[93] , FSLC[94] , FSBC[95] , CARDIE[74] , OBLIVIIN [90], | SBS-SLASH [96], WSFG [97], WSBG [97] , BDS [92], PQSS [92] , RC [98], SS [99] , BSE [96], K2-AS [100] , RACE [99] , SBS-W [101] | AICC[102], , FSSEM [103], , ELSA [104] , | SA[92] ,RGSS [92] , LVW [105] , RMHC-PF [106], GA [107] , RVE [108] |
| Embedded | Filter + Wrapper | | BBHFS [109], Xing's [110] | Dash-Liu's [78] | |
| | | Classification | Classification | Clustering | Classification |
| | | Objectif du modèle | | | |

Les stratégies de recherche et les critères d'évaluation sont les deux facteurs dominants dans la conception d'un algorithme de sélection de caractéristiques. Dans le tableau 2.1, les algorithmes sont classés selon les stratégies de recherche : complète, séquentielle et aléatoire.

En ce qui concerne les approches d'évaluation, les algorithmes sont classés en Filter, Wrapper et Embedded. Nous considérons les tâches d'exploration de données comme une troisième dimension car la disponibilité d'informations sur l'étiquette de la classe dans les tâches de classification ou de regroupement affecte les critères d'évaluation utilisés dans les algorithmes de sélection de fonctionnalités.

De plus, les algorithmes de l'approche Filter se distinguent par des critères d'évaluation spécifiques, tels que la distance, l'information, la dépendance et la cohérence. Dans la catégorie Wrapper, la précision prédictive est utilisée pour la classification, tandis que la qualité du cluster est utilisée pour la tâche de regroupement.

La catégorisation des algorithmes de sélection de caractéristiques remplit trois rôles importants. Tout d'abord, elle permet de révéler les relations entre les différents algorithmes. En effet, les algorithmes appartenant à la même catégorie sont les plus similaires les uns aux autres, car ils sont conçus avec des stratégies de recherche et des critères d'évaluation similaires, et pour le même type de tâches d'exploration de données. Ensuite, cette catégorisation nous permet de concentrer notre choix d'algorithmes pour une tâche donnée sur un nombre relativement restreint. Par exemple, pour une tâche de classification, la précision d'un classificateur convient pour être le critère d'évaluation, et la recherche complète n'est pas appropriée en raison du temps de calcul nécessaire. Nous pouvons donc limiter nos choix à deux groupes d'algorithmes dans le tableau 2.1 : l'un est défini par la classification, le Wrapper et l'Heuristique, et l'autre est la classification, l'Embedded et l'Aléatoire. Enfin, cette catégorisation nous permet de mettre en évidence les lacunes dans la collection actuelle d'algorithmes de sélection de fonctionnalités. En effet, comme on peut le voir, il y a de nombreux blocs vides dans le tableau, ce qui indique qu'aucun algorithme de sélection de caractéristiques n'existe pour ces combinaisons, ce qui pourrait être exploré dans des travaux futurs. En particulier, les algorithmes actuels de sélection de caractéristiques pour le regroupement se limitent à la recherche séquentielle. En conclusion, nous présentons dans cette section quelques-unes des méthodes de sélection de caractéristiques les plus utilisées dans la littérature.

#### 2.2.3.1 SFS et SBS

La méthode de sélection des caractéristiques SFS (Sequential Forward Selection) ou sélection séquentielle croissante a été proposée par Marill et Green [111]. Elle repose sur une approche heuristique de recherche. L'algorithme 4 commence avec un ensemble vide de caractéristiques. À chaque itération et jusqu'à un critère d'arrêt prédéfini, la meilleure caractéristique est retirée de l'ensemble initial et ajoutée au sous-ensemble sélectionné.

La méthode SBS (Sequential Backward Selection) ou sélection séquentielle arrière est une méthode similaire proposée par Whitney [112]. Contrairement à la méthode SFS, cette méthode commence avec l'ensemble complet de toutes les caractéristiques et élimine à chaque itération la caractéristique la moins pertinente.

Étant donné que le SBS commence par l'ensemble complet des caractéristiques, ses premières évaluations sont relativement coûteuses en termes de temps de calcul [113]. Par conséquent, le SBS est généralement appliqué avec une faible dimension. Le principal inconvénient du SBS est qu'une fois

qu'une caractéristique est supprimée, elle ne sera jamais réévaluée [114]. Ahan et Bankert [115] ont montré que la méthode SBS est plus performante que le SFS, car elle prend en compte l'interaction d'une caractéristique avec un ensemble de caractéristiques plus large, contrairement au SFS qui ne prend en compte que l'interaction de cette caractéristique avec le sous-ensemble déjà sélectionné. Cependant, l'évaluation des sous-ensembles de grande taille avec la méthode SBS pose un problème en termes de temps de calcul [51].

**Algorithm 4** Algorithme SFS

**Entrées :**
  $F = (f_1, f_2, ..f_N)$ ▷ Ensemble de caractéristiques
  $M$ ▷ Taille de l'ensemble final
**Sortie :** $E = (f_{s1}, f_{s2}, ..f_{sN})$
**Initialisation :** $E = \emptyset$
**Début :**
**for** $i = 1$ **to** $M$ **do**
  **for** $j = 1$ **to** $|F|$ **do**
    Évaluer $f_j \cup E$
  $f_{max} =$ meilleur $f_j$
  $E = E \cup \{f_{max}\}, F = F \setminus \{f_{max}\}$
**Retourner** $E$
**Fin**

En 1978, Kittler à proposé GSFS et GSBS [116] qui sont des généralisation SBS et SFS permettant d'inclure ou exclure un ensemble de caractéristiques par itération. Malgré que ces algorithmes surpassent en performance les méthodes de base, ils gardent toujours les mêmes inconvénients.

**Algorithm 5** Algorithme SBS

**Entrées :**
  $F = (f_1, f_2, ..f_N)$ ▷ Ensemble de caractéristiques
  $M$ ▷ Taille de l'ensemble final
**Sortie :** $E = (f_{s1}, f_{s2}, ..f_{sN})$
**Initialisation :** $E = F$
**Début :**
**for** $i = 1$ **to** $N - M$ **do**
  **for** $j = 1$ **to** $|E|$ **do**
    Évaluer $E \setminus \{f_j\}$
  $f_{min} =$ moins pertinente $f_j$
  $E = E \setminus \{f_{min}\}$
**Retourner** $E$
**Fin**

En 1994, Pudil a introduit deux algorithmes de sélection de caractéristiques, SFFS (Sequential Floating Forward Selection) et SFBS (Sequential Floating Backward Selection), qui appartiennent à la même famille que SBS et SFS [117]. Ces méthodes séquentielles alternent entre des sélections croissantes et décroissantes de caractéristiques, permettant une exploration plus vaste de l'espace des caractéristiques. Leur principal avantage réside dans leur capacité à s'adapter dynamiquement à la dimension du sous-ensemble de caractéristiques, ce qui peut conduire à une solution plus proche de l'optimum global. Cependant, le choix des paramètres x et y et le coût de calcul élevé représentent des défis majeurs. Malgré cela, des études telles que celles menées par Yang et al. [113] et Harish [114] ont démontré l'efficacité de SFFS et SFBS dans diverses applications, notamment la classification et la reconnaissance d'images, en surpassant les méthodes traditionnelles telles que SFS et SBS.

#### 2.2.3.2 Branch and Bound

Branch and Bound est un algorithme couramment utilisé en optimisation combinatoire. Il suit une structure arborescente (voir Figure 2.5) pour sélectionner le meilleur sous-ensemble de caractéristiques. Il appartient à l'approche Filter et utilise une recherche exhaustive et garantie pour trouver le sous-ensemble optimal de caractéristiques. L'algorithme B&B tente de satisfaire deux critères : minimiser la taille du sous-ensemble de caractéristiques optimal et respecter le seuil placé sur la valeur calculée par la fonction d'évaluation.

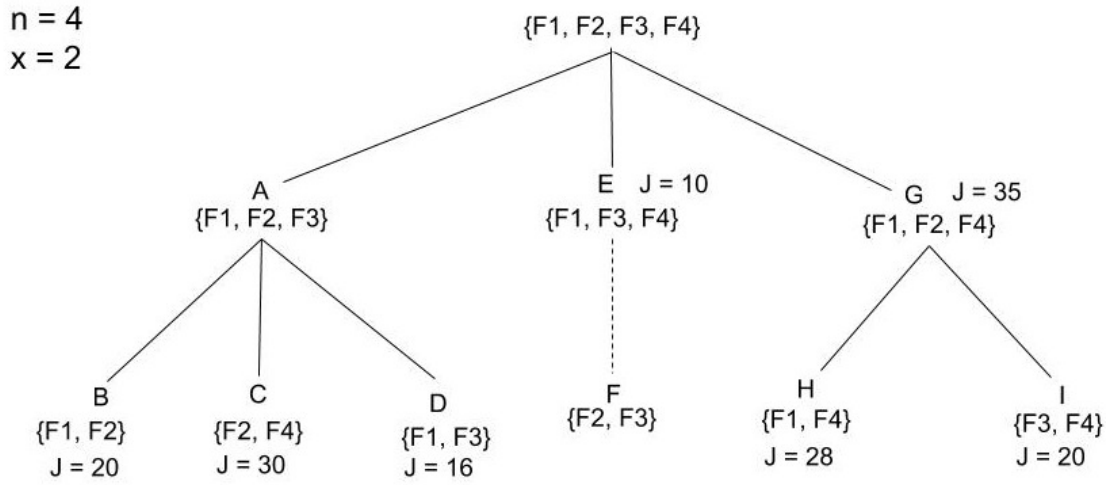



FIGURE 2.5: Arborescence de B&B avec n=4 x=2

L'algorithme B&B (Branch and Bound) commence sa recherche avec l'ensemble complet des caractéristiques n, et procède en supprimant successivement la caractéristique la moins pertinente à chaque étape, en suivant une structure d'arbre de la racine aux feuilles. Le critère de sélection J évalue les sous-ensembles de caractéristiques à chaque nœud, et sa valeur décide de l'expansion des branches. Les branches dont la valeur de J dépasse le seuil sont étendues, tandis que les autres sont élaguées. Le nœud feuille avec la valeur de J la plus élevée est finalement sélectionné. Cependant, le temps de calcul de cette méthode augmente rapidement avec le nombre de caractéristiques, devenant impraticable au-delà de 30 caractéristiques. Pour surmonter cet inconvénient, des améliorations comme le B&B rapide proposé par Somol en 2004 ont été développées pour accélérer le processus de sélection [118].

Sur la figure, n représente le nombre total de caractéristiques disponibles (4 dans cet exemple), et x représente le nombre de caractéristiques à sélectionner (2 dans cet exemple). Les nœuds F, E, et G illustrent différents sous-ensembles de caractéristiques et leur valeur J. Par exemple, F correspond au sous-ensemble F2, F3 avec une valeur J de 16, E correspond au sous-ensemble F1, F3, F4 avec une valeur J de 10.

J représente la valeur calculée par la fonction d'évaluation pour chaque sous-ensemble. Cette valeur est utilisée par l'algorithme Branch and Bound pour évaluer et sélectionner le sous-ensemble de caractéristiques optimal en minimisant J tout en respectant les contraintes de sélection.

#### 2.2.3.3 FOCUS

L'algorithme FOCUS [85] réussit à identifier l'ensemble optimal d'attributs en utilisant le moins de caractéristiques possible grâce à une exploration complète de l'espace des sous-ensembles d'attributs, et en un temps quasi-polynomial. En conséquence, FOCUS est un algorithme idéal à utiliser lorsqu'un ensemble minimal d'attributs est requis et lorsqu'on dispose des échantillons sans bruit.

**Algorithm 6** Algorithme FOCUS

**Entrées :**
  $F = (f_1, f_2, \ldots, f_N)$ ▷ Ensemble d'apprentissage de N caractéristiques
  $T$ ▷ Taille maximale de l'ensemble final
  $\epsilon$ ▷ Seuil
**Sortie :** $S$ ▷ Ensemble final de caractéristiques
**Initialisation :** $E = \emptyset$
**Début :**
**for** $i = 1$ **à** $T$ **do**
  **for** chaque sous-ensemble $S_1$ de taille $i$ **do**
    $Cons =$ Inconsistance$(A, S_1)$
    **if** $Cons \leq \epsilon$ **then**
      $S = S_1$
      **Retourner** $S$
**Fin**

En outre, dans un travail de dash [89], des résultats empiriques très intéressants de l'algorithme FOCUS sont présentés. L'algorithme FOCUS affiche de bonnes performances même sur certains jeux de données très bruités. Cependant, FOCUS et FOCUS-2 [85] (la version optimisée de FOCUS) sont limités aux attributs booléens, tandis que de nombreux problèmes impliquent des caractéristiques discrètes et continues. Pour éviter cet inconvénient, C-FOCUS [119] a été développé en tant qu'extension de FOCUS-2 pour traiter les attributs nominaux, discrèts et continus.

L'algorithme FOCUS présenté par l'algorithme 6 commence par la génération et l'évaluation de tous les sous-ensembles de taille T (initialement égale à 1), puis continue à évaluer les paires d'attributs, les triplets, les quadruplets, et ainsi de suite jusqu'à ce que le critère d'arrêt soit atteint.

#### 2.2.3.4 LVW et LVF

LVW (Las Vegas Wrapper) est un algorithme de sélection de caractéristiques de l'approche Wrapper, proposé par Liu et Setiono [120]. Le fonctionnement de cet algorithme consiste à générer aléatoirement à chaque itération un sous-ensemble de caractéristiques, puis à évaluer sa pertinence à l'aide d'un classificateur. Les algorithmes de Las Vegas utilisent le hasard pour guider leur recherche de manière à garantir une solution correcte. Ce processus est répété jusqu'à ce que T essais consécutifs d'amélioration des performances du modèle échouent, comme présenté dans l'algorithme 7. L'inconvénient majeur de cet algorithme est son temps de calcul très élevé. Malgré cet inconvénient, l'algorithme LVW reste un outil puissant pour la sélection de caractéristiques dans les problèmes de classification et peut être utilisé avec succès dans de nombreux cas pratiques.

**Algorithm 7** Algorithme LVW

**Entrées :**
  $A$ ▷ Base d'apprentissage
  $S$ ▷ Base de caractéristiques
  $T$ ▷ Nombre d'itérations
**Sortie :** $S$ ▷ Ensemble final de caractéristiques
$Err = \text{Classificateur}(A, S)$
$k = 0, n = |S|$
**Début :**
**repeat**
  $S_1 = \text{générer}(A), N_1 = |S_1|$
  $Err_1 = \text{Classificateur}(A, S_1)$
  **if** $Err_1 \leq Err$ ou ($Err = Err_1$ et $N_1 \leq N$) **then**
    $k = 0, N = N_1, S = S_1, Err = Err_1$
  $k = k + 1$
**until** $k = T$
**Retourner** $S$
**Fin**

LVF (Las Vegas Filter) est un algorithme similaire appartenant à l'approche Filter, proposé par les mêmes auteurs. À l'inverse de LVW, l'algorithme LVF utilise la mesure du taux d'incohérence ou d'inconsistance. La pertinence d'un sous-ensemble de caractéristiques est évaluée en se basant sur le taux d'incohérence mesuré dans le jeu de données. L'incohérence est définie par le rapport du nombre d'observations incohérentes sur l'ensemble total des observations du jeu de données. Une observation est dite inconsistante si une autre observation ayant la même présentation sur l'espace des caractéristiques étudiés appartient à une autre classe. L'algorithme LVF présente l'avantage d'être plus rapide que LVW, car il ne nécessite pas l'évaluation de chaque sous-ensemble de caractéristiques par un classificateur. Cependant, il est plus susceptible d'éliminer des caractéristiques pertinentes du modèle, car il ne tient pas compte de leur impact sur la précision du classificateur.

L'algorithme LVF peut également être combiné avec d'autres techniques de sélection de caractéristiques pour obtenir de meilleures performances. Par exemple, il peut être utilisé en combinaison avec des méthodes basées sur la corrélation pour sélectionner des caractéristiques hautement corrélées, puis appliquer l'algorithme LVF pour sélectionner les caractéristiques les plus pertinentes parmi ces caractéristiques hautement corrélées.

**Algorithm 8** Algorithme LVF

**Entrées :**
    $A$ ▷ Base d'apprentissage
    $S$ ▷ Base de caractéristiques
    $T$ ▷ Nombre d'itérations
    $\epsilon$ ▷ Seuil
**Sortie :** $S$ ▷ Ensemble final de caractéristiques
$N = |S|$
**Début :**
**for** $i = 1$ **à** $T$ **do**
    $S_1 = \text{générer}(A)$
    $N_1 = |S_1|$
    **if** $\text{Inconsistance}(A, S_1) \leq \epsilon$ **et** $N_1 \leq N$ **then**
        $N = N_1$
        $S = S_1$
**Retourner** $S$
**Fin**

En conclusion, bien que LVF présente des avantages en termes de temps de calcul, il est important de choisir l'algorithme de sélection de caractéristiques qui convient le mieux aux exigences de chaque problème de classification.

#### 2.2.3.5 Les Algorithmes Génétiques

Les algorithmes génétiques sont des algorithmes de recherche d'optimisation adaptative basés sur l'analogie avec la théorie darwinienne, dans lesquels un groupe de solutions évolue par sélection naturelle. Depuis leur développement par John Holland en 1975, les algorithmes génétiques ont été largement utilisés dans de nombreux domaines tels que l'optimisation, l'apprentissage automatique, la reconnaissance de formes et la sélection des caractéristiques. Les algorithmes génétiques opèrent sur une population d'individus qui représentent des solutions candidates pour un problème donné. Chaque individu ou chromosome rivalise avec les autres pour se reproduire, en se basant du principe darwinien de la "survie du plus apte", à chaque génération d'évolution. Un individu est évalué par une fonction de fitness qui exprime dans quelle mesure ce chromosome est approprié comme solution. Les meilleurs individus parents, qui ont survécu à la lutte dans une population, se croisent pour produire une progéniture améliorée. Pour éviter qu'un algorithme génétique ne soit pris dans une solution bonne mais non optimale, une mutation est effectuée. Après un certain nombre de générations, les chromosomes qui ont survécu dans la population sont les solutions optimales. En général, un algorithme génétique comprend les étapes suivantes, représentées sur la Figure 2.6.

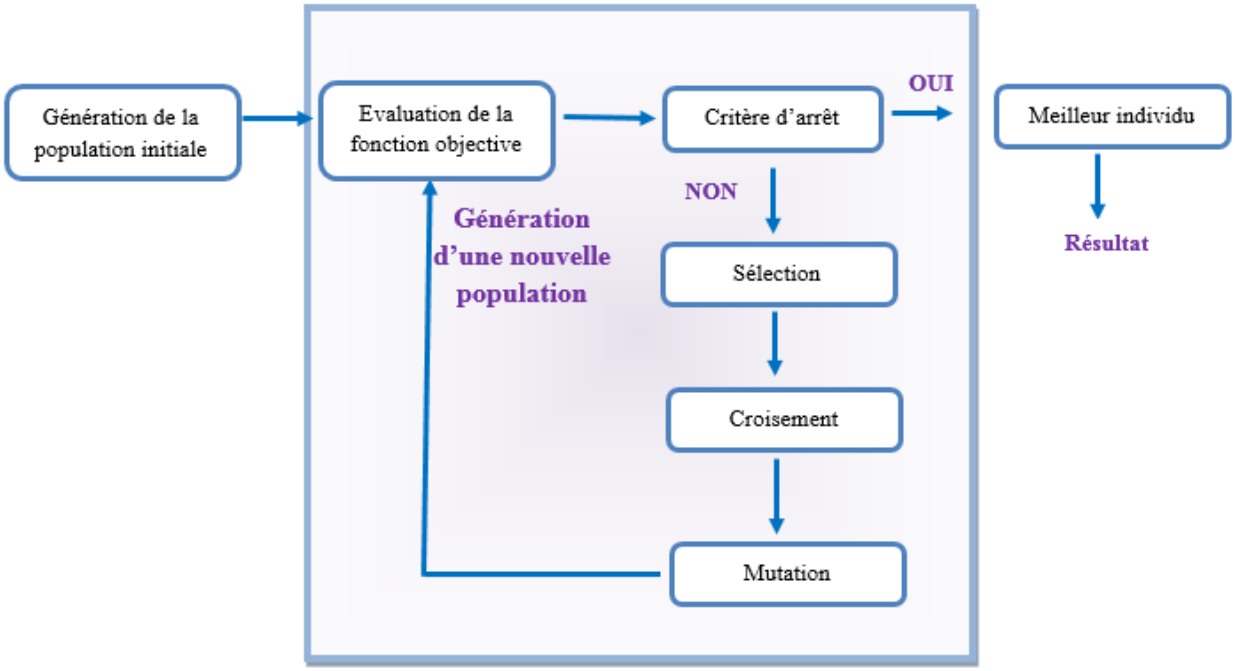



FIGURE 2.6: Les étapes de l'algorithme génétique

1. Commencer avec une population de n chromosomes générée aléatoirement (solutions candidates).
2. Évaluer chaque chromosome de la population en calculant la fonction de fitness.
3. Créer de nouveaux chromosomes en croisant les chromosomes actuels, et appliquer la mutation et la reproduction lors de l'accouplement des chromosomes parents.
4. Supprimer certains chromosomes de la population actuelle pour faire de la place aux nouveaux chromosomes.
5. Évaluer les nouveaux chromosomes et les insérer dans la population.
6. Si la condition d'arrêt est satisfaite, arrêter et retourner le meilleur chromosome ; sinon, passer à l'étape 3.

1. L'initialisation : L'initialisation est la première phase d'un algorithme génétique, où une population initiale de solutions candidates est générée pour offrir de la diversité et éviter le problème de l'optimum local. Cette population peut être créée aléatoirement ou en utilisant des solutions d'autres techniques de recherche, bien que la dernière méthode puisse réduire la diversité et mener à une convergence prématurée vers des solutions sous-optimales. L'objectif est de s'assurer que l'algorithme génétique soit au moins aussi efficace que l'algorithme initial.
2. Représentation : Dans les algorithmes génétiques, différentes représentations des génomes individuels peuvent être utilisées. La représentation la plus courante est celle d'un chromosome codé en binaire, sous forme de vecteur de zéros et de uns, où chaque bit représente un gène. Les deux principales représentations sont le codage binaire et le codage des nombres réels, qui diffèrent dans la mise en œuvre des opérateurs de croisement et de mutation. D'autres représentations

comme les tableaux, les arbres ou les listes peuvent également être utilisées en fonction du problème et du domaine d'application. Pour chaque représentation, les opérateurs génétiques doivent être adaptés et définis en conséquence.

La reproduction occupe une place centrale dans la transmission de caractéristiques d'une génération à une autre. Elle constitue un cycle essentiel de l'algorithme génétique, qui implique l'utilisation de processus génétiques tels que la sélection, le croisement et la mutation.

3. Sélection : La sélection est une étape cruciale des algorithmes génétiques, visant à choisir les individus pour la génération suivante. Les méthodes de sélection courantes incluent :

   (a) **Sélection par roulette** (Goldberg et David, 1989 [121]) : Attribue des probabilités de sélection aux individus en fonction de leur qualité. Avantages : simplicité d'implémentation. Inconvénients : peut conduire à une convergence prématurée.

   (b) **Sélection par tournoi** ([122]) : Sélectionne aléatoirement un certain nombre d'individus et choisit le meilleur comme parent. Avantages : favorise la diversité génétique. Inconvénients : peut entraîner une perte de qualité si le nombre de participants est trop faible.

   (c) **Sélection par rang** ([123]) : Classe les individus en fonction de leur qualité et sélectionne les meilleurs en fonction de leur rang. Avantages : favorise la diversité génétique tout en évitant la convergence prématurée.

4. Croisement : l'opérateur de croisement est largement étudié dans la littérature. Il permet de générer de nouvelles solutions en combinant les caractéristiques de deux solutions parentes. Lors du croisement, les chromosomes parentaux échangent des séquences de gènes, ce qui conduit à la création d'un nouvel individu ou d'une nouvelle solution. Cette opération est effectuée sur chaque paire de chromosomes sélectionnée à l'aide d'une des méthodes de sélection décrites précédemment, avec une certaine probabilité de croisement, notée Pcrois. Les paires de chromosomes non croisées sont recopiées tels quels dans la génération suivante avec une probabilité de 1 - Pcrois. Plus la valeur de Pcrois est élevée, plus il' y a des chances que de nouveaux individus apparaissent dans la population. Parmi les méthodes de croisement les plus utilisées, on trouve le croisement à un point et le croisement multi-points. Ces méthodes ont été largement étudiées dans la littérature, notamment par des auteurs tels que Holland (1975) [124], Davis (1985) [125] et Whitley (1989) [126].

   (a) **Croisement à un point** : Consiste à sélectionner aléatoirement un point de croisement pour chaque paire de chromosomes et à permuter les ensembles de séquences situées de part et d'autre de ce point entre les deux parents comme illustré dans la figure 2.7. Bien que cette méthode ait été largement étudiée dans la littérature (Whitley, 1989) [126], elle peut manquer de complexité dans l'échange de segments génétiques entre les parents. Ainsi, le

croisement multi-points a été proposé pour introduire une plus grande diversité génétique dans la descendance (Davis, 1991) [123].

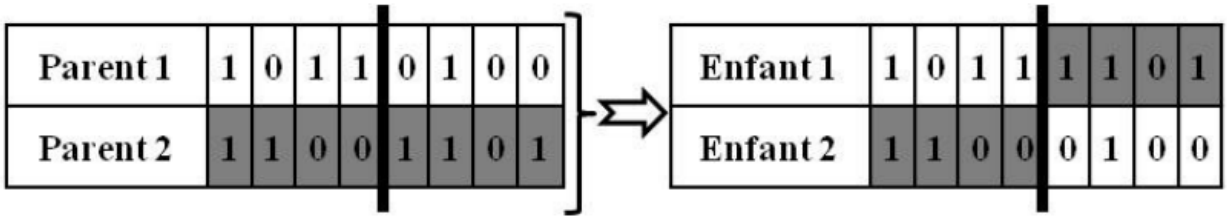



FIGURE 2.7: Opérateur de croisement à un point

(b) **Croisement multi-points** : Cette méthode implique la sélection de plusieurs points de croisement pour chaque paire de chromosomes, entre lesquels les séquences génétiques sont échangées entre les parents comme illustré dans la figure 2.8. Les points de croisement peuvent être sélectionnés de manière aléatoire ou selon des stratégies spécifiques pour optimiser la performance de l'algorithme génétique (Goldberg, 1989) [121]. Le croisement multipoints a également été largement étudié dans la littérature scientifique et est souvent comparé à d'autres méthodes de croisement pour évaluer son efficacité (De Jong, 1975) [127].

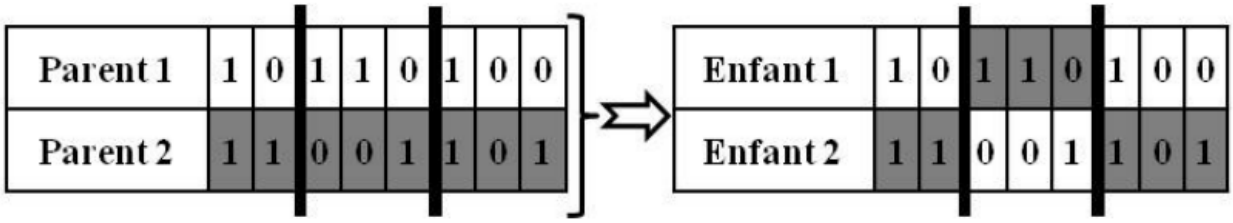



FIGURE 2.8: Opérateur de croisement à deux points

5. Mutation : est un processus aléatoire qui peut modifier la valeur d'un gène dans un chromosome. Cette modification peut aider à éviter la convergence vers des optima locaux en introduisant une certaine variabilité dans la population de chromosomes comme illustré dans la figure 2.9. En d'autres termes, la mutation joue le rôle de bruit dans le système d'évolution, ce qui empêche l'ensemble de se figer dans des optima locaux.




FIGURE 2.9: Exemple d'une opération de mutation

En appliquant une mutation, on peut étendre l'exploration de l'espace des solutions possibles et garantir que l'optimum global puisse être atteint. Ainsi, la mutation est un opérateur essentiel dans les algorithmes d'optimisation basés sur l'évolution, car elle permet d'éviter une convergence prématurée vers des solutions sous-optimales.

6. Critère d'arrêt : L'algorithme génétique est efficace pour résoudre des problèmes d'optimisation complexes, mais nécessite un critère d'arrêt adéquat pour garantir la qualité de la solution finale. Les critères d'arrêt courants sont :
   (a) **Stabilité de la population :** L'algorithme s'arrête lorsque la population n'évolue plus suffisamment, indiquant que la meilleure solution possible a été atteinte.
   (b) **Meilleur compromis pour les problèmes multi-critères :** Dans le cas d'un problème multi-critères, l'algorithme s'arrête lorsque la solution trouvée offre le meilleur équilibre possible entre les différents critères.
   (c) **Nombre limite de générations** : L'algorithme s'arrête après un nombre prédéfini de générations, sans garantie que la meilleure solution ait été trouvée.

   En conclusion, le choix du critère d'arrêt dépend du problème spécifique que l'algorithme génétique est utilisé pour résoudre. Cependant, quel que soit le critère d'arrêt choisi, il est important de s'assurer qu'il garantit une solution de haute qualité.
7. Fonction d'évaluation : La fonction d'évaluation dans les algorithmes génétiques est essentielle pour mesurer la qualité de chaque individu dans une population, en calculant une valeur de "fitness" qui représente la qualité de l'individu par rapport au problème à résoudre. Cette fonction dépend du problème spécifique et est utilisée pour sélectionner, croiser et muter les individus dans les générations suivantes. Pour la sélection de caractéristiques, la fonction de fitness évalue la qualité de chaque sous-ensemble de caractéristiques candidates. Les fonctions de fitness couramment utilisées pour la sélection de caractéristiques incluent :
   (a) **Précision de prédiction** : Mesure la qualité de la prédiction d'un modèle de machine learning, en comparant les prédictions du modèle avec les valeurs réelles [128].
   (b) **Réduction d'erreur de prédiction** : Évalue la qualité de la prédiction en comparant l'erreur de prédiction du modèle avec et sans l'utilisation des caractéristiques sélectionnées ([61]).
   (c) **Critère d'information** : Mesure la qualité de l'ajustement d'un modèle statistique, en utilisant des critères tels que l'AIC ou le BIC ([129].

En conclusion, les algorithmes génétiques sont une approche prometteuse pour la sélection de caractéristiques dans les modèles prédictifs. Ces algorithmes peuvent être utilisés pour rechercher rapidement une combinaison optimale de caractéristiques à partir d'un grand ensemble de caractéristiques potentielles, ce qui peut améliorer la précision et l'interprétabilité des modèles.

Après avoir présenté les méthodes de réduction de dimensionnalité par sélection de caractéristiques, il est temps de passer aux méthodes de transformation de données. Contrairement à la sélection de

caractéristiques, les méthodes de transformation de données modifient les caractéristiques existantes pour créer de nouvelles caractéristiques qui capturent l'essence des données d'origine. Ces méthodes sont plus complexes que la sélection de caractéristiques, mais elles sont également plus puissantes car elles peuvent capturer des relations plus complexes entre les caractéristiques.

## 2.3 Réduction basée sur la transformation des caractéristiques

La réduction de la dimensionnalité par une transformation de données, également appelée extraction de caractéristiques, ne consiste pas simplement à sélectionner certaines caractéristiques, mais plutôt à créer de nouvelles caractéristiques en combinant les caractéristiques initiales. Toutefois, il convient de souligner que cette transformation peut altérer la signification de l'ensemble initial de caractéristiques, ce qui limite l'applicabilité de cette famille de méthodes aux cas où la signification n'a plus d'importance dans les étapes suivant la réduction.

Dans les sections suivantes, nous allons brièvement présenter plusieurs techniques de réduction bien connues, généralement classées en deux catégories : les méthodes linéaires et les méthodes non linéaires.

### 2.3.1 Les méthodes linéaires

#### 2.3.1.1 Analyse en Composantes Principales

L'Analyse en Composantes principales (ACP) est une technique de réduction de dimension qui fait partie des méthodes descriptives multidimensionnelles. Elle permet de transformer un ensemble de variables quantitatives initiales en un ensemble de variables quantitatives non corrélées appelées composantes principales. Les composantes principales sont des combinaisons linéaires des variables initiales qui expliquent le maximum de variance dans les données. L'objectif de l'ACP est de trouver des dimensions plus petites dans lesquelles il est possible d'observer les individus de manière optimale. Cette méthode est souvent utilisée pour explorer la structure des données et réduire la complexité de l'analyse en ne gardant que les informations les plus pertinentes. La première composante principale est déterminée en recherchant la direction dans laquelle le nuage de points des observations s'étire au maximum. La seconde composante principale est calculée de manière à être la plus indépendante possible de la première. Ces deux composantes principales forment le premier plan principal. L'opération est réitérée pour trouver toutes les composantes principales expliquant le maximum de variance. Enfin, une visualisation des données dans le plan principal permet de mieux comprendre la structure des données et les relations entre les variables.

Supposons maintenant que nous avons un ensemble de données $Y = (y_1, y_2, \ldots, y_M)$ composé de $M$ observations où chaque observation $y_i = (y_{i1}, y_{i2}, \ldots, y_{iN})$ est composée de $N$ variables. $Y$ est associé à une matrice de données $\mathbf{B}$ de taille $N \times M$ où chaque colonne représente une variable. En pratique,

le calcul de l'analyse en composantes principales (ACP) pour la matrice $\mathbf{Y}$ revient à effectuer les opérations suivantes afin de trouver les composantes principales :

— Calculer le vecteur $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_N)^T$ qui représente le vecteur moyen où $\mu_i$ est la moyenne de la $i$ème variable des données.

— Calculer la matrice $\boldsymbol{\chi}$ en soustrayant le vecteur moyen à toutes les colonnes de $\mathbf{B}$ dans le but d'obtenir des données centrées, soit $\boldsymbol{\chi} = \mathbf{B} - \boldsymbol{\mu}\mathbf{1}_M^T$, où $\mathbf{1}_M$ est un vecteur colonne de taille $M$ avec toutes les entrées égales à 1.

— Calculer la matrice $\mathbf{S}$ (de taille $N \times N$) de covariance de $\boldsymbol{\chi}$ avec $\mathbf{S} = \boldsymbol{\chi}\boldsymbol{\chi}^T/(M-1)$.

— Calculer la matrice $\mathbf{U}$ (de taille $N \times N$) qui est composée des coordonnées des vecteurs propres $\boldsymbol{u}_j$ de $\mathbf{S}$ triés par ordre décroissant des modules des valeurs propres $\lambda_j$, soit $\mathbf{S}\boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j$ (la première colonne de $\mathbf{U}$ est le vecteur propre qui correspond à la plus grande valeur propre).

— Garder les $R$ premières colonnes de $\mathbf{U}$ pour former la matrice $\mathbf{U}_{\sim R} : N \times R$ qui représente les $R$ premières composantes principales.

#### 2.3.1.2 Positionnement Multi-Dimensionnel

Dans cette partie, nous abordons la méthode de positionnement multidimensionnel (MDS) proposée par [130]. Cette méthode vise à construire une représentation en faible dimension des points d'un espace, en se basant sur une matrice des distances ou des mesures de similarité calculées entre chaque paire de points. L'objectif est d'obtenir une représentation euclidienne des individus dans un espace de dimension réduite qui préserve, autant que possible, ces distances. Imaginons que nous disposons d'un ensemble de données $X = x_1, x_2, \ldots, x_M$ constitué de $M$ observations, où chaque observation $x_i = (x_{i1}, x_{i2}, \ldots, x_{iN})$ comprend $N$ caractéristiques. Soit $d$ une matrice symétrique de dimensions $M \times M$ dont chaque élément $d_{ij}$ représente la distance entre $x_i$ et $x_j$ et $d_{ii} = 0$.L'idée de MDS est de déterminer une configuration de points $y_i$, avec $i = 1, \ldots, N$, dans un espace de dimension plus faible qui conserve les distances entre les points initiaux $x_i$. Autrement dit, la méthode cherche les points $y_i$ dans un espace de dimension $q < N$ tels que $d(y_i, y_j) \approx d_{ij} = d(x_i, x_j)$. Ceci revient à minimiser le critère suivant :

$$J = \sum_{i=1}^{N} \sum_{j=1}^{N} (d_{ij} - d(y_i, y_j))^2$$

où $J$ représente la somme des carrés des différences entre les distances originales $d_{ij}$ et les distances projetées $d(y_i, y_j)$ dans l'espace de dimension réduite.

Prenons par exemple, un ensemble de données contenant des informations sur les durées de connexion (log) d'étudiants à une plateforme d'apprentissage en ligne. Les distances entre les étudiants pourraient être basées sur les différences de temps de connexion, les activités réalisées ou les ressources consultées comme illustré dans la Table 2.2.

TABLE 2.2: Matrice des distances entre les logs de 6 étudiants

| | Étudiant A | Étudiant B | Étudiant C | Étudiant D | Étudiant E | Étudiant F |
|---|---|---|---|---|---|---|
| Étudiant A | 0 | 240 | 710 | 610 | 350 | 360 |
| Étudiant B | 240 | 0 | 790 | 690 | 570 | 540 |
| Étudiant C | 710 | 790 | 0 | 110 | 560 | 510 |
| Étudiant D | 610 | 690 | 110 | 0 | 460 | 400 |
| Étudiant E | 350 | 570 | 560 | 460 | 0 | 59 |
| Étudiant F | 360 | 540 | 510 | 400 | 59 | 0 |

À l'aide de la méthode MDS, nous pourrions représenter ces étudiants dans un espace bidimensionnel (ou tridimensionnel) de manière à ce que les distances entre les points représentant les étudiants reflètent les différences entre leurs logs.

En résumé, la méthode MDS est utile pour visualiser et analyser des ensembles de données complexes en réduisant leur dimensionnalité tout en préservant les relations entre les points. Dans notre exemple, cela permettrait d'identifier des groupes d'étudiants ayant des comportements similaires sur la plateforme d'apprentissage en ligne et ainsi d'adapter les ressources et les activités en fonction de leurs besoins.

### 2.3.2 Les méthodes non-linéaires

#### 2.3.2.1 Isomap

Isomap [131] est une technique de réduction de dimensions qui, tout comme la méthode MDS, part de la connaissance d'une matrice de dissimilarités entre les paires d'individus. Le but est, cette fois, de trouver une variété (non linéaire) contenant les données.

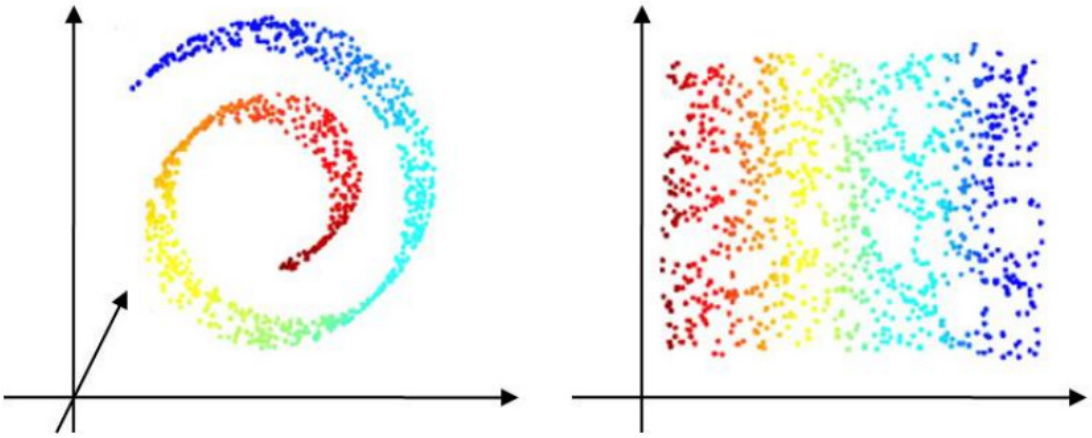


FIGURE 2.10: Isomap sur des données non-linéaires

La Figure 2.10 illustre, à gauche, un exemple de données non linéaires, et à droite, le résultat de la représentation de ces points dans l'espace guidé par les deux premières dimensions obtenues en

appliquant l'algorithme Isomap sur ces données.

Supposons que nous ayons un ensemble de données $\mathbf{X} = \{x_1, x_2, \ldots, x_M\}$ composé de $M$ observations où chaque observation $x_i = (x_{i1}, x_{i2}, \ldots, x_{iN})$ est composée de $N$ caractéristiques. Soit $d$ une matrice symétrique de taille $M \times M$ où chaque élément $d_{ij}$ représente la distance entre $x_i$ et $x_j$ et $d_{ii} = 0$. L'algorithme Isomap, qui a $\mathbf{X}$ et $d$ en entrée, passe en général par trois étapes :

1. Déterminer un voisinage pour chaque point $x_i$ par la méthode des $k$-plus proches voisins calculées à partir des distances euclidiennes $d$, en considérant que les distances euclidiennes approchent les distances géodésiques lorsque les vecteurs se trouvent à petite distance.
2. Estimer la matrice des distances géodésiques ($\mathbf{D}$) entre tous les points. Isomap construit un graphe dont les sommets sont les points et les arêtes sont labellisées par les distances entre eux.
3. Appliquer la méthode MDS à la matrice de distances $\mathbf{D}$ pour obtenir un nouveau système de coordonnées euclidiennes qui préserve la géométrie intrinsèque de la variété.

#### 2.3.2.2 T-SNE

T-SNE (T-Distributed Stochastic Neighbor Embedding) est une autre méthode de réduction de dimensions non-linéaire largement utilisée dans le domaine de l'apprentissage automatique [132]. T-SNE est particulièrement efficace pour visualiser des données de haute dimension en deux ou trois dimensions, tout en préservant la topologie et la structure de voisinage des données. Contrairement à d'autres méthodes de réduction de dimensionnalité telles que PCA et MDS, T-SNE cherche à préserver les structures non-linéaires dans les données. Il repose sur la construction d'un graphe de voisinage local pour les données, en utilisant une mesure de similarité entre les points. Ensuite, il utilise une distribution de probabilité t-student pour modéliser les similarités dans l'espace de faible dimension. La minimisation de la divergence de Kullback-Leibler entre les distributions de probabilité des espaces de haute et de faible dimensions permet d'obtenir la configuration des points dans l'espace de faible dimension. T-SNE est particulièrement utile pour la visualisation des données en cluster, car il permet de distinguer des clusters très proches ou des groupes de points similaires dans l'espace de haute dimension [132]. La fonction de coût utilisée pour minimiser la divergence de Kullback-Leibler est la suivante :

$$J_{T-SNE} = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

où $p_{ij}$ est la probabilité que le point $i$ choisisse le point $j$ comme voisin, calculée à partir de la mesure de similarité entre les points dans l'espace de haute dimension, et $q_{ij}$ est la probabilité que les points $i$ et $j$ soient choisis comme voisins dans l'espace de faible dimension, calculée à partir de la distribution t-student. La minimisation de cette fonction de coût est effectuée par descente de gradient

stochastique, en utilisant des techniques de régularisation pour éviter la formation de clusters artificiels [133].

## 2.4 Conclusion

Ce chapitre s'est concentré sur le domaine de la réduction de dimensionnalité. Nous avons d'abord examiné la sélection de caractéristiques en présentant les différents composants nécessaires à un algorithme de sélection de caractéristiques, ainsi que les alternatives possibles pour leur mise en œuvre. Nous avons ensuite décrit un certain nombre d'algorithmes en fonction de la combinaison de composants utilisée, en présentant les avantages et les inconvénients des techniques de sélection en général, ainsi que ceux des algorithmes décrits. Cependant, les approches existantes présentent des limitations en termes de complexité très élevée des approches Wrapper, ainsi qu'en raison de la dépendance des caractéristiques pertinentes sélectionnées par rapport au classificateur utilisé. Par ailleurs, les approches "Filter" présentent plusieurs limitations concernant la redondance ainsi que les interactions entre les caractéristiques.

Dans la deuxième partie de ce chapitre, nous avons détaillé les techniques de réduction de dimensionnalité par une transformation de données, en présentant différentes approches linéaires et non linéaires. Nous avons particulièrement abordé l'algorithme Isomap qui permet de réduire la dimensionnalité des données non linéaires tout en préservant leur structure géométrique. Nous avons également présenté l'algorithme T-SNE qui utilise une mesure de similarité non linéaire pour projeter les données dans un espace de dimension réduite. Ces méthodes permettent de capturer des informations plus variées et pourraient être combinées avec les méthodes de sélection de caractéristiques pour améliorer leurs performances.

# Chapitre 3

# Apprentissage Automatique

## 3.1 Introduction

L'apprentissage automatique, ou Machine Learning en anglais, est un sous-domaine de l'intelligence artificielle qui se concentre sur le développement d'algorithmes et de modèles permettant aux machines d'apprendre et de s'adapter à partir des données [134]. L'objectif principal de l'apprentissage automatique est d'extraire des connaissances à partir de grandes quantités de données, d'améliorer les performances des algorithmes au fil du temps et de généraliser ces connaissances pour résoudre de nouveaux problèmes [135].L'apprentissage automatique a un impact significatif sur l'éducation, notamment dans le développement de MOOCs. L'apprentissage automatique peut être utilisé pour personnaliser l'expérience d'apprentissage des étudiants, améliorer l'évaluation des performances et détecter les problèmes qui pourraient conduire à l'échec ou au décrochage scolaire [136].

Il existe principalement trois types d'apprentissage automatique :

1. **Apprentissage supervisé :** Il s'agit d'un type d'apprentissage où les algorithmes sont entraînés à partir d'exemples étiquetés. Les données d'entrée sont associées à des étiquettes ou des valeurs de sortie, et l'algorithme apprend à prédire ces valeurs en fonction des entrées.
2. **Apprentissage non supervisé :** Dans ce type d'apprentissage, les algorithmes travaillent avec des données non étiquetées et cherchent à découvrir des structures ou des modèles sous-jacents. Les techniques d'apprentissage non supervisé incluent le regroupement (clustering), la réduction de dimensionnalité et la détection d'anomalies.
3. **Apprentissage par renforcement :** L'apprentissage par renforcement se concentre sur l'entraînement d'agents qui apprennent à prendre des décisions en interagissant avec leur environnement. L'agent reçoit des récompenses ou des pénalités pour ses actions et cherche à optimiser sa stratégie pour maximiser les récompenses cumulées. Ce type d'apprentissage ne sera pas abordé dans ce travail, car il ne fait pas partie du périmètre de notre étude.

L'apprentissage automatique a trouvé des applications dans divers domaines, tels que la reconnaissance d'image et de parole, la recommandation de produits, la détection de fraude, la médecine personnalisée et bien d'autres. Un domaine d'application particulièrement important est la prévention du décrochage scolaire, où l'apprentissage automatique peut aider à identifier les étudiants à risque et à mettre en

place des mesures d'intervention adaptées. Les avancées récentes en matière d'apprentissage profond (Deep Learning) ont considérablement élargi les capacités des algorithmes d'apprentissage automatique, rendant possible des tâches autrefois considérées comme inaccessibles aux machines.

Dans ce chapitre, nous explorerons en détail les différentes facettes de l'apprentissage automatique. Nous débuterons avec l'apprentissage supervisé, en couvrant les techniques de classification et de régression, et en présentant des algorithmes tels que la régression logistique, les k-plus proches voisins, les machines à vecteurs de support, les arbres de décision, les forêts aléatoires, le gradient boosting et l'apprentissage profond. Ensuite, nous nous pencherons sur l'apprentissage non supervisé, en abordant le clustering et la détection d'anomalies à travers des méthodes telles que le K-means, DBSCAN, le clustering hiérarchique, l'algorithme LOF et One-Class SVM.

Dans la troisième partie, nous évaluerons en profondeur les modèles en étudiant les différentes métriques pour les problèmes de classification et de régression. Nous poursuivrons avec la validation croisée et la sélection des modèles, en examinant des techniques telles que la K-Fold Cross-Validation, la Leave-One-Out Cross-Validation et diverses méthodes de sélection des paramètres, notamment la recherche exhaustive, la recherche aléatoire et l'optimisation bayésienne.

Enfin, nous discuterons des problèmes de surapprentissage, de sous-apprentissage, des techniques de régularisation, et des stratégies pour gérer les classes non balancées. Nous aborderons les concepts de régularisation L1 (Lasso) et L2 (Ridge) et examinerons les défis associés aux classes non balancées dans les problèmes d'apprentissage automatique.

## 3.2 L'apprentissage supervisé

L'apprentissage supervisé est une méthode d'apprentissage automatique où les algorithmes sont entraînés à partir d'exemples étiquetés. Les données d'entrée sont associées à des étiquettes ou des valeurs de sortie, et l'algorithme apprend à prédire ces valeurs en fonction des entrées. L'apprentissage supervisé est largement utilisé pour résoudre des problèmes de classification et de régression [137].

### 3.2.1 Classification

La classification est une tâche d'apprentissage supervisé qui consiste à attribuer une étiquette de classe à une entrée donnée. Les algorithmes de classification sont généralement utilisés pour résoudre des problèmes où les étiquettes de sortie sont discrètes et finies. Les exemples incluent la détection de spam, la reconnaissance d'écriture manuscrite et le diagnostic médical. Mathématiquement, un problème de classification peut être formulé comme suit :

Soit $\mathbf{x} \in \mathbb{R}^d$ une entrée et $y \in 1, \ldots, K$ l'étiquette correspondante, où $K$ est le nombre de classes. L'objectif est de trouver une fonction $f : \mathbb{R}^d \to 1, \ldots, K$ qui minimise l'erreur de classification sur un

ensemble d'apprentissage $(\mathbf{x}i, y_i)i = 1^N$, où $N$ est le nombre d'exemples d'apprentissage.

Dans la suite de cette sous-section nous présenterons les algorithmes de classification les plus couramment utilisés dans l'apprentissage automatique pour résoudre des problèmes de classification, tels que la régression logistique, les k-plus proches voisins (K-NN), les machines à vecteurs de support (SVM), les arbres de décision et les réseaux de neurones, suivra cette section.

#### 3.2.1.1 Régression Logistique

La régression logistique est un algorithme de classification linéaire qui utilise une fonction sigmoïde pour transformer la sortie d'une régression linéaire en une probabilité dans l'intervalle [0,1] [72]. Elle est largement utilisée pour résoudre des problèmes de classification binaire, où il y a seulement deux classes de sortie possibles.

La régression logistique modélise la probabilité $P(Y = 1|\mathbf{x})$, c'est-à-dire la probabilité que la classe de sortie soit 1 étant donné l'entrée $\mathbf{x}$. Elle est définie par :

$$P(Y = 1|\mathbf{x}; \boldsymbol{\theta}) = \sigma(\boldsymbol{\theta}^T \mathbf{x})$$

où $\sigma(z) = \frac{1}{1+e^{-z}}$ est la fonction sigmoïde, $\mathbf{x}$ est l'entrée, $\boldsymbol{\theta}$ est le vecteur de poids du modèle.

Le but de l'apprentissage de la régression logistique est de trouver le vecteur de poids $\boldsymbol{\theta}$ qui maximise la vraisemblance de l'ensemble d'apprentissage. La vraisemblance est définie par :

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} P(Y = y_i|\mathbf{x}_i; \boldsymbol{\theta})$$

où $(\mathbf{x}_i, y_i)$ sont les exemples d'apprentissage. L'objectif est de maximiser la log-vraisemblance :

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} y_i \log P(Y = 1|\mathbf{x}_i; \boldsymbol{\theta}) + (1 - y_i) \log P(Y = 0|\mathbf{x}_i; \boldsymbol{\theta})$$

Il est généralement difficile de maximiser la log-vraisemblance analytiquement, donc on utilise des algorithmes d'optimisation numérique pour trouver le vecteur de poids qui maximise la log-vraisemblance. Le plus couramment utilisé est l'algorithme de descente de gradient stochastique (SGD) [138].

La sortie de l'algorithme de régression logistique est une probabilité de la classe positive, c'est-à-dire $P(Y = 1|\mathbf{x})$. Si cette probabilité est supérieure à un seuil prédéfini, alors la classe de sortie est prédite comme positive, sinon elle est prédite comme négative.

#### 3.2.1.2 K-plus proches voisins (k-NN)

L'algorithme des k-plus proches voisins (K-NN) est une méthode d'apprentissage automatique non paramétrique et basée sur la mémoire utilisée pour la classification et la régression [139]. Dans le

contexte de la classification, l'algorithme fonctionne en trouvant les k exemples d'apprentissage les plus proches d'une nouvelle entrée et en attribuant à cette entrée la classe majoritaire parmi ces voisins.

Soit $\mathbf{x}$ une entrée et $y$ son étiquette de classe correspondante. L'algorithme K-NN est défini comme suit :

1. Pour une nouvelle entrée $\mathbf{x}^{'}$, calculez la distance entre $\mathbf{x}^{'}$ et tous les autres points d'entraînement.
2. Sélectionnez les k points d'entraînement les plus proches de $\mathbf{x}^{'}$, c'est-à-dire les k voisins.
3. Attribuez à $\mathbf{x}^{'}$ la classe majoritaire parmi ses k voisins.

Il existe plusieurs façons de mesurer la distance entre deux points, telles que la distance euclidienne, la distance de Manhattan ou la distance de Minkowski. Le choix de la métrique de distance dépend du problème à résoudre et des caractéristiques des données. De plus, le choix de k peut avoir un impact significatif sur les performances de l'algorithme. Une valeur trop petite de k peut entraîner un surapprentissage, tandis qu'une valeur trop grande peut entraîner un sous-apprentissage. Le choix optimal de k est généralement déterminé par la validation croisée.

L'algorithme K-NN est simple à comprendre et à mettre en œuvre, et il peut être efficace pour les problèmes avec un petit nombre d'exemples d'apprentissage. Cependant, il présente quelques inconvénients, notamment sa sensibilité aux caractéristiques non pertinentes et à la malédiction de la dimensionnalité, ainsi que sa complexité computationnelle élevée lorsqu'il est appliqué à de grands ensembles de données.

#### 3.2.1.3 Machines à vecteurs de support (SVM)

Les machines à vecteurs de support (SVM) sont un ensemble d'algorithmes d'apprentissage supervisé pour la classification et la régression, développés principalement par Vapnik et al. [140]. Les SVM cherchent à trouver le meilleur hyperplan de séparation entre les différentes classes, en maximisant la marge entre les classes les plus proches de l'hyperplan. L'hyperplan est déterminé par un sous-ensemble de points d'entraînement appelés vecteurs de support.

Dans le cas d'une classification binaire linéaire, l'objectif est de trouver l'hyperplan qui maximise la marge entre les deux classes. L'hyperplan est défini par :

$$\mathbf{w}^T\mathbf{x} + b = 0$$

où $\mathbf{w}$ est le vecteur normal à l'hyperplan, $\mathbf{x}$ est un point sur l'hyperplan et $b$ est le biais. La marge est définie comme la distance minimale entre l'hyperplan et les points d'entraînement les plus proches de chaque classe. Les SVM cherchent à maximiser la marge sous contrainte de classification correcte :

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \quad \text{pour } i = 1, \ldots, N$$

où $(\mathbf{x}_i, y_i)$ sont les exemples d'apprentissage et $N$ est le nombre d'exemples. Le problème d'optimisation peut être formulé comme un problème d'optimisation quadratique avec des contraintes linéaires, qui peut être résolu efficacement à l'aide de techniques d'optimisation telles que la méthode des multiplicateurs de Lagrange et la décomposition séquentielle minimale (SMO) [141].

Dans le cas où les données ne sont pas linéairement séparables, les SVM utilisent une technique appelée le truc du noyau (kernel trick) pour projeter les données dans un espace de caractéristiques de plus grande dimension, où elles deviennent linéairement séparables 3.1. Les fonctions noyau couramment utilisées sont le noyau linéaire, le noyau polynomial et le noyau radial de base gaussienne (RBF) [142]. Le choix du noyau et des paramètres du noyau a un impact significatif sur les performances de l'algorithme et est généralement déterminé par la validation croisée.

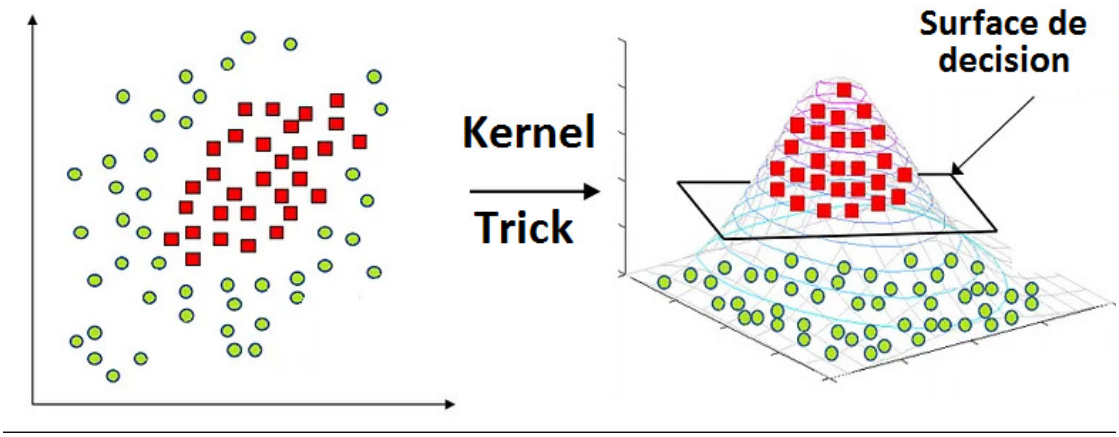



FIGURE 3.1: Illustration du kernel trick : transformation des données séparées à l'aide d'un hyperplan

Les SVM sont particulièrement efficaces pour les problèmes de classification à grande échelle et avec un grand nombre de caractéristiques. Ils sont moins sensibles au surapprentissage que d'autres méthodes, grâce à leur contrôle explicite de la complexité du modèle par la marge. Cependant, les SVM sont sensibles au choix des paramètres et des fonctions noyau, et leur temps d'entraînement peut être élevé pour de très grands ensembles de données.

#### 3.2.1.4 Arbres de décision

Les arbres de décision sont des modèles d'apprentissage automatique utilisés pour la classification et la régression, caractérisés par leur facilité d'interprétation. Ils sont composés de nœuds de décision et de nœuds feuilles, et classent les nouvelles entrées en suivant les branches de l'arbre jusqu'à atteindre un nœud feuille. L'apprentissage d'un arbre de décision implique de minimiser une mesure d'impureté, comme l'indice de Gini ou l'entropie, à l'aide d'algorithmes tels que CART [143] ou ID3 [144]. Cependant, les arbres de décision peuvent être sujets au surapprentissage, ce qui peut être atténué par des techniques de post-élagage ou l'utilisation de méthodes d'ensemble telles que les forêts aléatoires [145] et les machines à gradient boosting [146].

Parmi les développements récents, l'algorithme XGBoost [147] est devenu populaire pour sa performance exceptionnelle et sa robustesse. D'autres travaux se concentrent sur l'amélioration des forêts aléatoires, comme les forêts extrêmement aléatoires [148] et les forêts aléatoires adaptatives en ligne [149]. L'interprétabilité des modèles basés sur des arbres est également un domaine de recherche actif, avec des techniques comme les valeurs SHAP [150] pour expliquer les prédictions de modèles complexes.

#### 3.2.1.5 Forêts aléatoires

Les forêts aléatoires sont une méthode d'ensemble d'apprentissage automatique basée sur les arbres de décision [145]. L'idée principale derrière les forêts aléatoires est de construire plusieurs arbres de décision indépendants $h_i(x)$, où $i = 1, \ldots, T$, et de combiner leurs prédictions pour obtenir une prédiction plus précise et robuste. La prédiction finale est obtenue en effectuant un vote majoritaire pour les problèmes de classification ou en calculant la moyenne pour les problèmes de régression. Le mécanisme de vote est représenté par la fonction $H(x)$ :

$$H(x) = \begin{cases} \text{majorité}\{h_1(x), \ldots, h_T(x)\} & \text{pour classification} \\ \frac{1}{T}\sum_{i=1}^{T} h_i(x) & \text{pour régression} \end{cases}$$

Les forêts aléatoires utilisent le bagging (Bootstrap Aggregating) pour créer des ensembles de données d'apprentissage pour chaque arbre. Chaque ensemble est créé en échantillonnant aléatoirement avec remplacement à partir de l'ensemble de données d'origine $(X, Y)$, où $X$ représente les caractéristiques et $Y$ les étiquettes. De plus, lors de la construction de chaque arbre, un sous-ensemble aléatoire des caractéristiques est sélectionné à chaque nœud pour déterminer la meilleure division.

Cette combinaison de bagging et de sélection aléatoire des caractéristiques permet aux forêts aléatoires de réduire la variance et le biais des modèles individuels d'arbres de décision, ce qui entraîne une amélioration des performances et une réduction du risque de surapprentissage.

Les forêts aléatoires ont fait l'objet de plusieurs améliorations et variantes au fil des années pour optimiser leurs performances et leur adaptabilité à différents problèmes. Voici quelques-uns des développements récents liés aux forêts aléatoires :

— **Forêts extrêmement aléatoires (Extra Trees)** : Les forêts extrêmement aléatoires, proposées par Geurts et al. [148], ajoutent une couche supplémentaire de randomisation aux forêts aléatoires. Contrairement aux forêts aléatoires classiques, où le seuil de division optimal est choisi pour chaque caractéristique, les forêts extrêmement aléatoires sélectionnent les seuils de division de manière aléatoire. Cela conduit à des arbres plus diversifiés et peut réduire la variance du modèle.

— **Forêts aléatoires adaptatives en ligne (Online Adaptive Random Forests)** : Les forêts aléatoires adaptatives en ligne, développées par Gomes et al. [149], sont une variante des forêts

aléatoires qui permet l'apprentissage en ligne et l'adaptation aux changements de distribution des données au fil du temps. Cette méthode est particulièrement utile pour les scénarios où les données sont reçues de manière incrémentielle ou les distributions de données évoluent avec le temps.

— **Optimisation de l'efficacité des forêts aléatoires** : Plusieurs travaux de recherche se sont concentrés sur l'amélioration de l'efficacité des forêts aléatoires, notamment en termes de temps d'apprentissage et de mémoire. Par exemple, les forêts aléatoires compactes [151] réduisent l'utilisation de la mémoire en stockant les arbres de manière compacte et en élaguant les branches inutiles.

— **Interprétabilité des forêts aléatoires** : L'interprétabilité des modèles basés sur des arbres, y compris les forêts aléatoires, est un domaine de recherche actif. Des techniques telles que les valeurs SHAP (SHapley Additive exPlanations) [150] ont été développées pour fournir des explications locales et globales des prédictions de modèles basés sur des arbres, en quantifiant l'importance des caractéristiques pour chaque prédiction individuelle.

Ces développements récents ont contribué à étendre l'applicabilité des forêts aléatoires à divers problèmes et à améliorer leur performance et leur interprétabilité.

#### 3.2.1.6 Gradient Boosting

Le gradient boosting est une technique d'apprentissage automatique ensembliste qui vise à construire un modèle prédictif solide en combinant plusieurs modèles faibles, généralement des arbres de décision [152]. L'objectif principal du gradient boosting est de réduire l'erreur de prédiction en mettant l'accent sur les erreurs commises par les modèles précédents [153]. La méthode fonctionne en ajoutant séquentiellement des modèles faibles qui sont formés pour expliquer les résidus (erreurs) laissés par les modèles précédents [154].

L'algorithme de gradient boosting suit les étapes suivantes :

1. Initialiser le modèle avec une constante :

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$$

2. Pour chaque itération $m = 1, 2, \ldots, M$ :

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right] F(x) = Fm - 1(x)$$

3. Ajuster $h_m(x)$ aux résidus $r_{im}$.

4. Mettre à jour le modèle global :

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x)$$

5. Répéter les étapes 2 à 4.

6. Utiliser le modèle final pour les prédictions :

$$\hat{y}(x) = F_M(x) = F_0(x) + \nu \sum_{m=1}^{M} h_m(x)$$

Le gradient boosting est très efficace pour résoudre des problèmes complexes et non linéaires et est souvent utilisé dans diverses applications, telles que la prédiction du prix des logements [155] et la classification d'images [156].

Récemment, il y a eu plusieurs développements dans la littérature sur le gradient boosting, tels que l'introduction de nouvelles variantes d'optimisation et d'amélioration des performances. Par exemple, XGBoost [154] est une implémentation très populaire et efficace de l'algorithme de gradient boosting qui utilise une optimisation régularisée pour éviter le surapprentissage. LightGBM [157] est un autre développement récent qui vise à accélérer l'entraînement et à améliorer l'efficacité du gradient boosting en utilisant un apprentissage basé sur le gradient dirigé par le gain d'information. CatBoost [158] est également une implémentation avancée qui se concentre sur la prise en charge des données catégorielles et l'amélioration de l'exactitude des prédictions.

#### 3.2.1.7 Apprentissage profond

L'apprentissage profond est une branche de l'apprentissage automatique axée sur l'emploi de réseaux de neurones artificiels multicouches pour résoudre des problèmes complexes de classification et de régression [135]. Il a été largement utilisé dans divers domaines, tels que la reconnaissance d'images [159], le traitement du langage naturel [160], la synthèse vocale [161]et la traduction automatique [162].

Un réseau neuronal profond comprend plusieurs couches cachées situées entre les couches d'entrée et de sortie. Chaque couche se compose d'un ensemble de neurones artificiels, appelés unités, qui transforment les données d'entrée en fonction de leurs poids ($W$) et biais ($b$). Au cours de la phase d'apprentissage, les poids et les biais sont mis à jour pour minimiser une fonction de coût donnée, généralement dénotée par $J(W, b)$.

$$\min_{W,b} J(W, b)$$

L'apprentissage dans les réseaux neuronaux profonds s'effectue généralement à l'aide de l'algorithme de rétropropagation du gradient [163]. Cela consiste à propager l'erreur de la sortie vers l'entrée en ajustant les poids et les biais selon le gradient de la fonction de coût :

$$W_{t+1} = W_t - \eta \nabla_W J(W_t, b_t), \quad b_{t+1} = b_t - \eta \nabla_b J(W_t, b_t)$$

où $\eta$ est le taux d'apprentissage et $t$ représente l'itération actuelle.

Récemment, des avancées significatives ont été réalisées dans le domaine de l'apprentissage profond. Parmi elles, on peut citer l'introduction des Transformers [160], qui ont révolutionné le traitement du

langage naturel en utilisant l'attention auto-régressive pour capter les dépendances à long terme dans les séquences de texte. De plus, les architectures de réseaux neuronaux profonds sont devenues de plus en plus complexes et efficaces, telles que les modèles EfficientNet [164] pour la classification d'images et les modèles GPT [165, 166] pour la génération de texte et la compréhension du langage naturel.

Maintenant que nous avons exploré les principaux algorithmes de classification en apprentissage supervisé, intéressons-nous aux algorithmes de régression. Alors que les algorithmes de classification sont utilisés pour prédire des variables catégorielles, les algorithmes de régression sont utilisés pour prédire des variables continues. Les algorithmes de régression sont largement utilisés dans des domaines tels que la finance, l'économie, les sciences sociales et les sciences naturelles pour prédire des variables telles que les prix des actions, les revenus, les scores de tests et les températures. Dans la section suivante, nous allons explorer les algorithmes de régression les plus couramment utilisés en apprentissage supervisé.

### 3.2.2 Régression

#### 3.2.2.1 Régression linéaire

La régression linéaire est l'une des techniques de régression les plus simples et les plus couramment utilisées en apprentissage supervisé [167]. L'objectif de la régression linéaire est de modéliser la relation entre une variable dépendante $y$ et une ou plusieurs variables indépendantes $x_1, x_2, \ldots, x_p$. Le modèle linéaire est défini comme suit :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

où $\beta_0, \beta_1, \ldots, \beta_p$ sont les coefficients du modèle et $\epsilon$ est l'erreur résiduelle. La régression linéaire vise à estimer les coefficients $\beta$ qui minimisent la somme des erreurs résiduelles quadratiques (SSE) :

$$\min\beta \sum i = 1^n (y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}))^2$$

La solution aux moindres carrés ordinaires (OLS) est une méthode couramment utilisée pour estimer les coefficients de régression linéaire [155]. Les méthodes de régularisation, telles que la régression Ridge [168] et la régression Lasso [169], sont souvent utilisées pour améliorer la robustesse et la stabilité du modèle en ajoutant une pénalité aux coefficients.

#### 3.2.2.2 Régression polynomiale

La régression polynomiale est une extension de la régression linéaire qui modélise la relation entre la variable dépendante $y$ et les variables indépendantes $x_1, x_2, \ldots, x_p$ à l'aide d'un polynôme d'un degré donné $d$ [170]. Le modèle polynômial est défini comme suit :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \gamma_{11} x_1^2 + \gamma_{22} x_2^2 + \cdots + \gamma_{pp} x_p^2 + \cdots + \epsilon$$

où $\beta$ et $\gamma$ sont les coefficients du modèle et $\epsilon$ est l'erreur résiduelle. La régression polynomiale permet de capturer les relations non linéaires entre les variables en incluant des termes de degré supérieur dans le modèle. Toutefois, il est important de sélectionner le degré approprié pour éviter le surapprentissage et l'ajustement excessif du modèle.

#### 3.2.2.3 Régression par processus gaussien

La régression par processus gaussien (GPR) est une méthode non paramétrique basée sur la théorie des processus gaussiens pour estimer la relation entre la variable dépendante $y$ et les variables indépendantes $x_1, x_2, \ldots, x_p$ [171]. Contrairement aux méthodes paramétriques telles que la régression linéaire, la GPR n'utilise pas de fonction prédéfinie pour modéliser la relation entre les variables. Au lieu de cela, elle utilise des fonctions de covariance (ou noyaux) pour définir la similarité entre les points de données et prédire la variable dépendante en fonction de cette similarité.

La GPR suppose que les données suivent un processus gaussien multivarié, qui est défini par une moyenne et une fonction de covariance :

$$y(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

où $m(x)$ est la fonction de moyenne et $k(x, x')$ est la fonction de covariance. La prédiction pour un nouveau point $x_*$ est donnée par la distribution conditionnelle suivante :

$$y(x_)|y(X) \sim \mathcal{N}(\mu(x_), \sigma^2(x_)$$

où $\mu(x_)$ et $\sigma^2(x_)$ sont la moyenne et la variance de la distribution conditionnelle, respectivement. La GPR est particulièrement utile pour modéliser des données bruitées et des relations non linéaires complexes.

La sélection du noyau approprié est cruciale pour obtenir de bonnes performances avec la GPR. Les noyaux couramment utilisés incluent le noyau exponentiel quadratique (RBF), le noyau linéaire et le noyau de Matérn [172]. Les paramètres du noyau sont généralement appris en optimisant la vraisemblance marginale des données.

#### 3.2.2.4 Réseaux de neurones pour la régression

Les réseaux de neurones, comme mentionné précédemment dans la section sur l'apprentissage profond, peuvent également être utilisés pour résoudre des problèmes de régression en adaptant la couche de sortie pour prédire des valeurs continues plutôt que des catégories. Les réseaux de neurones pour

la régression, également appelés réseaux de neurones de régression (NNR), sont capables de modéliser des relations complexes et non linéaires entre les variables indépendantes et dépendantes [173].

Pour un NNR, la fonction de coût est généralement définie comme l'erreur quadratique moyenne (MSE) ou l'erreur absolue moyenne (MAE) entre les prédictions du réseau et les valeurs réelles. L'entraînement du réseau est réalisé en ajustant les poids et les biais pour minimiser la fonction de coût à l'aide de l'algorithme de rétropropagation du gradient et d'autres techniques d'optimisation, comme décrit précédemment dans la section sur l'apprentissage profond.

Les réseaux de neurones de régression sont largement utilisés dans diverses applications, telles que la prédiction de la demande énergétique [174], la modélisation de la qualité de l'eau [175] et la prédiction des prix des actions [176]. Les avantages des NNR incluent leur capacité à apprendre des représentations hiérarchiques et à modéliser des relations non linéaires complexes. Cependant, ils peuvent également être sujets au surapprentissage et nécessitent souvent une sélection minutieuse des hyperparamètres et une validation croisée pour obtenir de bonnes performances.

#### 3.2.2.5 Support Vector Regression (SVR)

La Support Vector Regression (SVR), également connue sous le nom d'Epsilon-Support Vector Regression ($\epsilon$-SVR), est une méthode de régression basée sur les Support Vector Machines (SVM) [177]. Alors que les SVM sont principalement utilisées pour la classification, la SVR étend le concept aux problèmes de régression en introduisant une marge de tolérance autour de la fonction de régression.

L'objectif de la SVR est de trouver une fonction de régression $f(x)$ qui minimise l'erreur de prédiction tout en maintenant la marge de tolérance $\epsilon$ pour les erreurs. La SVR est formulée comme un problème d'optimisation quadratique sous contraintes, qui peut être résolu à l'aide des techniques de programmation quadratique. La fonction de régression est définie comme suit :

$$f(x) = \langle w, \phi(x) \rangle + b$$

où $w$ est le vecteur de poids, $\phi(x)$ est une transformation non linéaire des données d'entrée et $b$ est un biais. La SVR peut également être formulée en termes de noyau, ce qui permet d'utiliser des noyaux tels que le noyau RBF et le noyau polynomial pour modéliser des relations non linéaires.

La SVR est robuste face aux bruits et aux données atypiques et peut modéliser des relations non linéaires complexes. Cependant, elle est plus sensible aux choix d'hyperparamètres (tels que le paramètre de régularisation $C$ et le paramètre de tolérance $\epsilon$) que d'autres méthodes de régression, et peut être plus lente à entraîner et à prédire que d'autres méthodes pour de grands ensembles de données.

Après avoir abordé les algorithmes d'apprentissage supervisé et leurs différentes approches, notamment la classification et la régression, nous allons maintenant nous intéresser à l'apprentissage non supervisé et ses techniques associées.

## 3.3 Apprentissage non supervisé

L'apprentissage non supervisé est un type d'apprentissage automatique où le modèle est entraîné sans utiliser de données étiquetées. Les algorithmes d'apprentissage non supervisé cherchent à découvrir des structures et des motifs sous-jacents dans les données, tels que des groupes ou des représentations à faible dimensionnalité. Les principales techniques d'apprentissage non supervisé comprennent le clustering, la réduction de dimensionnalité et la détection d'anomalies. Dans cette section, nous nous concentrerons sur le clustering et la détection d'anomalies, étant donné que la réduction de dimensionnalité a déjà été présentée dans le chapitre 2 avec des méthodes telles que l'ACP et t-SNE.

### 3.3.1 Clustering

Le clustering, ou partitionnement de données, est une technique d'apprentissage non supervisé visant à diviser un ensemble de données en groupes homogènes, de sorte que les points de données similaires se retrouvent dans le même groupe et les points de données différents se retrouvent dans des groupes séparés. Les algorithmes de clustering les plus couramment utilisés comprennent :

#### 3.3.1.1 K-means

L'algorithme K-means [178] cherche à partitionner les données en K groupes en minimisant la somme des distances au carré entre chaque point de données et le centroïde de son groupe. Les centroïdes sont recalculés à chaque itération jusqu'à convergence. La formule pour le K-means est la suivante :

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2$$

où $S$ représente les groupes, $k$ est le nombre de groupes, $x$ est un point de données, et $\mu_i$ est le centroïde du groupe $i$.

#### 3.3.1.2 DBSCAN

En anglais : Density-Based Spatial Clustering of Applications with Nois, DBSCAN [179] est un algorithme de clustering basé sur la densité qui regroupe les points en fonction de leur proximité et de la densité des points voisins. L'algorithme DBSCAN est défini comme suit :

$$\text{DBSCAN}(D, \epsilon, \text{MinPts}) = C_1, C_2, \ldots, C_k$$

où $D$ est l'ensemble des points de données, $\epsilon$ est le rayon, et MinPts est le nombre minimum de points requis pour former un cluster dense. Les clusters $C_i$ sont déterminés en fonction des critères de densité et de proximité.

#### 3.3.1.3 Clustering hiérarchique

Le clustering hiérarchique [180] construit une hiérarchie de clusters en combinant ou en divisant des clusters à chaque niveau. La distance entre les clusters est généralement mesurée à l'aide d'une fonction de distance, comme la distance euclidienne, la distance de Manhattan ou la distance de Mahalanobis :

$$d(x, y) = ||x - y||_p$$

où $x$ et $y$ sont deux points de données et $p$ est un paramètre de distance (par exemple, $p = 2$ pour la distance euclidienne).

### 3.3.2 Détection d'anomalies

Dans cette section, nous aborderons deux types de détection d'anomalies : la détection d'anomalies basée sur la distance (ou la densité) et la détection d'anomalies basée sur des modèles. Pour chaque type de détection, nous présenterons un algorithme populaire.

1. **Détection d'anomalies basée sur la densité** la détection d'anomalies basée sur la densité cherche à identifier les points de données qui sont situés dans des régions de faible densité de l'espace des données. L'une des méthodes populaires est l'algorithme LOF (Local Outlier Factor) [181].

#### 3.3.2.1 Algorithme LOF (Local Outlier Factor)

L'algorithme LOF [181] est une méthode de détection d'anomalies basée sur la densité. Il compare la densité locale d'un point de données à la densité locale de ses voisins. Un score LOF élevé indique que le point de données a une densité significativement plus faible que celle de ses voisins et est donc considéré comme une anomalie.

La formule pour calculer le score LOF d'un point de données $x$ est la suivante :

$$\text{LOF}(x) = \frac{\sum_{y \in N_k(x)} \frac{\text{reachability distance}(x,y)}{\text{local reachability density}(y)}}{k}$$

où $N_k(x)$ est l'ensemble des $k$ voisins les plus proches de $x$, la distance de reachability est la distance maximale entre $x$ et $y$ et la densité de reachability locale de $y$, et $k$ est le nombre de voisins considérés.

2. **Détection d'anomalies basée sur des modèles**

La détection d'anomalies basée sur des modèles consiste à ajuster un modèle aux données et à identifier les points de données qui s'écartent significativement du modèle. Dans cette sous-section, nous

présentons un exemple d'algorithme basé sur des modèles appelé One-Class SVM (Support Vector Machines).

#### 3.3.2.2 One-Class SVM

L'algorithme One-Class SVM [182] cherche à séparer les points de données normaux de l'origine dans un espace de caractéristiques de haute dimension, en utilisant un hyperplan qui maximise la marge entre les points normaux et l'origine.

La fonction de décision pour One-Class SVM est la suivante :

$$f(x) = \langle w, \phi(x) \rangle - \rho$$

où $w$ est le vecteur normal de l'hyperplan, $\phi(x)$ est la transformation de l'espace des caractéristiques, et $\rho$ est un seuil qui détermine la sensibilité de la détection d'anomalies. Les points de données pour lesquels $f(x) < 0$ sont considérés comme des anomalies.

## 3.4 Évaluation approfondie des modèles

L'évaluation minutieuse et approfondie de la performance d'un modèle d'apprentissage automatique est une étape cruciale pour s'assurer qu'il répond aux exigences de précision et de généralisation. Les métriques d'évaluation varient en fonction du type de problème (classification, régression, etc.). Dans cette section, nous présenterons les métriques d'évaluation courantes pour les problèmes de classification et de régression, ainsi que des détails supplémentaires pour mieux comprendre leur utilité et leur interprétation.

### 3.4.1 Métriques pour les problèmes de classification

Pour les problèmes de classification, plusieurs métriques d'évaluation sont couramment utilisées pour évaluer la performance des modèles. Voici une liste plus détaillée de ces métriques et de leur signification :

— **Accuracy (Exactitude)** : L'accuracy est la proportion de prédictions correctes sur l'ensemble des prédictions. Elle est définie comme suit :

$$\text{Accuracy} = \frac{\text{Nombre de prédictions correctes}}{\text{Nombre total de prédictions}}$$

Bien que l'exactitude soit une métrique simple et facile à comprendre, elle peut être trompeuse dans des situations où les classes sont déséquilibrées. Dans ces cas, d'autres métriques, telles que la précision, le rappel et le F1-score, sont souvent plus informatives.

— **Matrice de confusion** : La matrice de confusion est un tableau qui montre les prédictions correctes et incorrectes pour chaque classe, comme indiqué sur la Figure 3.2. Les lignes représentent les vraies classes, tandis que les colonnes représentent les classes prédites .

| | | **Classe Réelle** | |
|---|---|---|---|
| | | - | + |
| **Classe Prédite** | - | Vrais négatifs **(TN)** | Faux négatifs **(FN)** |
| | + | Faux positifs **(FP)** | Vrais positifs **(TP)** |

FIGURE 3.2: Matrice de confusion

La matrice de confusion est particulièrement utile pour visualiser les erreurs de classification et pour identifier les classes qui sont souvent confondues entre elles.

— **Precision (précision)** : La précision est la proportion de prédictions correctes parmi les prédictions positives. Elle est définie comme suit :

$$\text{Precision} = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs} + \text{Faux Positifs}}$$

La précision est une métrique importante pour les situations où les faux positifs sont coûteux ou indésirables, comme la détection de spams ou la détection de fraudes.

— **Recall (rappel)** : Le rappel est la proportion de vrais positifs parmi l'ensemble des éléments qui devraient être identifiés comme positifs. Il est défini comme suit :

$$\text{Recall} = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs} + \text{Faux Négatifs}}$$

Le rappel est une métrique importante pour les situations où les faux négatifs sont coûteux ou indésirables, comme la détection de maladies ou la détection de pannes de machines.

— **F1-score** : Le F1-score est la moyenne harmonique de la précision et du rappel. Il est défini comme suit :

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Le F1-score est une métrique qui tient compte à la fois de la précision et du rappel et qui est particulièrement utile lorsque les coûts des faux positifs et des faux négatifs sont similaires ou lorsque les classes sont déséquilibrées.

— **AUC-ROC (Area Under the Receiver Operating Characteristic Curve)** : L'AUC-ROC est une métrique qui mesure la performance d'un classificateur pour différents seuils de classification. La courbe ROC est un graphique qui représente le taux de vrais positifs en fonction

du taux de faux positifs. L'AUC (Area Under the Curve) est la surface sous la courbe ROC, et une valeur d'AUC plus élevée indique une meilleure performance du modèle. Une AUC de 0,5 indique un modèle aléatoire, tandis qu'une AUC de 1 indique un modèle parfait.

### 3.4.2 Métriques pour les problèmes de régression

Pour les problèmes de régression, d'autres métriques d'évaluation sont utilisées pour mesurer la performance des modèles. Voici une liste plus détaillée de ces métriques et de leur signification :

— **Erreur quadratique moyenne (Mean Squared Error, MSE)** : La MSE est la moyenne des carrés des erreurs entre les prédictions et les valeurs réelles. Elle est définie comme suit :

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

où $y_i$ est la valeur réelle, $\hat{y}_i$ est la valeur prédite et $n$ est le nombre de prédictions. La MSE est sensible aux valeurs aberrantes et donne plus de poids aux erreurs importantes.

— **Erreur absolue moyenne (Mean Absolute Error, MAE)** : La MAE est la moyenne des erreurs absolues entre les prédictions et les valeurs réelles. Elle est définie comme suit :

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

La MAE est moins sensible aux valeurs aberrantes que la MSE et donne un poids égal à toutes les erreurs.

— **Coefficient de détermination ($\mathbf{R^2}$)** : Le coefficient de détermination, également connu sous le nom de $R^2$, mesure la proportion de la variance dans les valeurs réelles qui est expliquée par le modèle. Il est défini comme suit :

$$R^2 = 1 - \frac{\text{MSE}}{\text{Variance}}$$

Un $R^2$ de 1 indique que le modèle explique parfaitement la variance des données, tandis qu'un $R^2$ de 0 indique que le modèle n'explique pas du tout la variance des données. Un $R^2$ négatif indique que le modèle est pire qu'une simple régression constante (une régression qui prédit toujours la moyenne des valeurs réelles).

## 3.5 Validation croisée et sélection des modèles

La validation croisée est une technique utilisée pour évaluer la performance d'un modèle d'apprentissage automatique de manière plus robuste et fiable. Elle permet également de sélectionner le meilleur modèle parmi un ensemble de modèles candidats. La validation croisée est particulièrement utile lorsque

le nombre de données est limité ou pour éviter le surapprentissage. Cette section présente les concepts clés de la validation croisée et leurs formulations mathématiques.

### 3.5.1 K-Fold Cross-Validation

La K-Fold Cross-Validation est une technique couramment utilisée pour évaluer les modèles d'apprentissage automatique. Elle consiste à diviser l'ensemble des données en K sous-ensembles (ou "folds") de taille égale. Le modèle est entraîné sur K-1 sous-ensembles et testé sur le sous-ensemble restant. Ce processus est répété K fois, chaque fois en utilisant un sous-ensemble différent comme ensemble de test. Les métriques d'évaluation sont ensuite moyennées sur les K itérations pour obtenir une estimation plus robuste de la performance du modèle. La Figure 3.3 illustre le processus de K-Fold Cross-Validation.

Soit $D$ l'ensemble de données de taille $n$, et $D_k$ le $k$-ième sous-ensemble de $D$. Pour chaque itération $i$ avec $1 \leq i \leq K$, nous avons l'ensemble d'apprentissage $D^i_{train}$ et l'ensemble de test $D^i_{test}$. Les ensembles d'apprentissage et de test sont définis comme suit :

$$D^i_{train} = D \setminus D_i, \quad D^i_{test} = D_i$$

Le modèle est entraîné sur $D^i_{train}$ et testé sur $D^i_{test}$ pour chaque itération. La métrique d'évaluation $M$ (par exemple, accuracy, MSE, etc.) est calculée pour chaque itération. La performance globale du modèle est obtenue en moyennant les métriques d'évaluation sur les K itérations :

$$M_{CV} = \frac{1}{K} \sum_{i=1}^{K} M_i$$

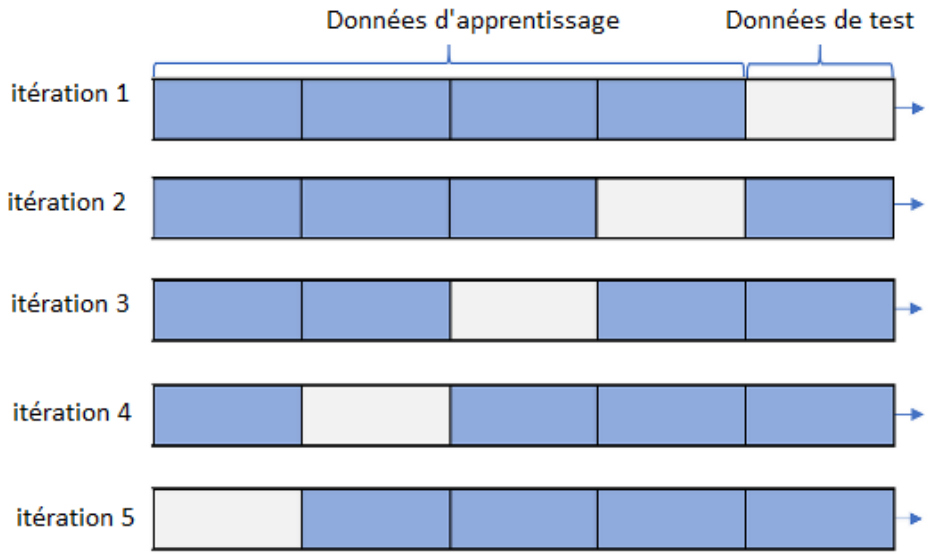



FIGURE 3.3: 5-Fold Cross-Validation

### 3.5.2 Leave-One-Out Cross-Validation

La Leave-One-Out Cross-Validation (LOOCV) est un cas particulier de K-Fold Cross-Validation où K est égal au nombre d'échantillons de données (c'est-à-dire $K = n$) [183]. Dans la LOOCV, le modèle est entraîné sur toutes les données sauf un échantillon et testé sur cet échantillon unique. Ce processus est répété pour chaque échantillon de l'ensemble de données. La métrique d'évaluation $M$ est calculée pour chaque itération, et la performance globale du modèle est obtenue en moyennant les métriques d'évaluation sur les n itérations :

$$M_{LOOCV} = \frac{1}{n} \sum_{i=1}^{n} M_i$$

La LOOCV est souvent utilisée lorsque le nombre de données est faible, car elle maximise l'utilisation des données pour l'apprentissage. Cependant, elle peut être coûteuse en termes de temps de calcul, car le modèle doit être entraîné n fois.

### 3.5.3 Sélection des paramètres

La validation croisée est également utilisée pour déterminer les hyperparamètres optimaux d'un modèle. Le processus de sélection des paramètres implique l'évaluation de plusieurs modèles candidats avec différentes combinaisons d'hyperparamètres. La validation croisée est utilisée pour estimer la performance de chaque modèle candidat, et le modèle avec la meilleure performance en validation croisée est sélectionné.

La recherche des hyperparamètres optimaux peut être effectuée à l'aide de différentes méthodes, telles que la recherche exhaustive (Grid Search), la recherche aléatoire (Random Search) ou des méthodes plus avancées telles que l'optimisation bayésienne [184]. Dans tous les cas, la validation croisée fournit une estimation fiable de la performance du modèle pour chaque combinaison d'hyperparamètres.

#### 3.5.3.1 Recherche exhaustive (Grid Search)

La recherche exhaustive, également connue sous le nom de Grid Search, est une méthode de recherche d'hyperparamètres qui évalue toutes les combinaisons possibles d'hyperparamètres spécifiées. Cette approche est simple et facile à mettre en œuvre, mais elle peut être coûteuse en termes de temps de calcul, en particulier lorsque le nombre d'hyperparamètres et de leurs valeurs possibles est important.

Pour effectuer une recherche exhaustive, un ensemble de valeurs est défini pour chaque hyperparamètre. Les modèles sont ensuite évalués pour chaque combinaison possible d'hyperparamètres à l'aide de la validation croisée. Le modèle qui donne la meilleure performance en validation croisée est sélectionné comme modèle optimal.

#### 3.5.3.2 Recherche aléatoire (Random Search)

La recherche aléatoire est une méthode de recherche d'hyperparamètres qui évalue un nombre fixe de combinaisons d'hyperparamètres tirées au hasard [185]. Contrairement à la recherche exhaustive, la recherche aléatoire ne teste pas toutes les combinaisons possibles, ce qui la rend plus rapide et moins coûteuse en termes de temps de calcul. La recherche aléatoire est particulièrement utile lorsque l'espace des hyperparamètres est large et que certaines combinaisons ont un impact limité sur les performances du modèle.

Pour effectuer une recherche aléatoire, un nombre d'itérations est défini, et à chaque itération, une combinaison d'hyperparamètres est tirée au hasard. Les modèles sont évalués pour chaque combinaison d'hyperparamètres à l'aide de la validation croisée. Le modèle qui donne la meilleure performance en validation croisée est sélectionné comme modèle optimal.

#### 3.5.3.3 Optimisation bayésienne

L'optimisation bayésienne est une méthode avancée de recherche d'hyperparamètres qui utilise des modèles probabilistes pour estimer la fonction de performance en fonction des hyperparamètres [184]. L'optimisation bayésienne sélectionne intelligemment les combinaisons d'hyperparamètres à évaluer en fonction des informations précédemment recueillies, ce qui permet de réduire le nombre d'évaluations de modèles nécessaires pour trouver les hyperparamètres optimaux.

Dans l'optimisation bayésienne, un modèle probabiliste est construit pour estimer la fonction de performance en fonction des hyperparamètres. À chaque itération, le modèle probabiliste est mis à jour en fonction des évaluations précédentes et une nouvelle combinaison d'hyperparamètres est sélectionnée en fonction d'une stratégie d'exploration-exploitation. Les modèles sont évalués pour chaque combinaison d'hyperparamètres à l'aide de la validation croisée. Le modèle qui donne la meilleure performance en validation croisée est sélectionné comme modèle optimal.

En conclusion, la sélection des modèles est une étape essentielle pour obtenir un modèle d'apprentissage automatique performant. La validation croisée est utilisée pour estimer la performance des modèles candidats, et différentes méthodes de recherche d'hyperparamètres, telles que la recherche exhaustive, la recherche aléatoire et l'optimisation bayésienne, peuvent être employées pour trouver les hyperparamètres optimaux.

## 3.6 Surrapprentissage, Sous-apprentissage et Techniques de Régularisation

### 3.6.1 Surapprentissage et Sous-apprentissage

Le **sous-apprentissage** se produit lorsque le modèle ne parvient pas à capturer suffisamment les caractéristiques sous-jacentes du jeu de données. Mathématiquement, cela signifie que l'erreur sur les données d'entraînement est élevée :

$$E_{\text{entraînement}} \approx E_{\text{réel}}$$

Le **surapprentissage** se produit lorsque le modèle apprend trop bien les données d'entraînement, y compris le bruit, au détriment de sa performance sur de nouvelles données. Mathématiquement, cela signifie que l'erreur sur les données d'entraînement est faible, mais l'erreur sur les données de validation est élevée :

$$E_{\text{entraînement}} \ll E_{\text{validation}}$$

### 3.6.2 Techniques de régularisation

La régularisation est une technique pour réduire le surapprentissage en ajoutant un terme de pénalité à la fonction de coût. Il existe plusieurs méthodes de régularisation, parmi lesquelles la régularisation $L_1$ et $L_2$.

#### 3.6.2.1 Régularisation $L_1$ (Lasso)

La régularisation $L_1$ ajoute une pénalité proportionnelle à la somme des valeurs absolues des poids du modèle :

$$J_{\text{Lasso}}(\theta) = J(\theta) + \alpha \sum_{i=1}^{n} |\theta_i|$$

où $J(\theta)$ est la fonction de coût sans régularisation, $\alpha$ est un hyperparamètre de régularisation, et $\theta_i$ sont les poids du modèle.

#### 3.6.2.2 Régularisation $L_2$ (Ridge)

La régularisation $L_2$ ajoute une pénalité proportionnelle à la somme des carrés des poids du modèle :

$$J_{\text{Ridge}}(\theta) = J(\theta) + \alpha \sum_{i=1}^{n} \theta_i^2$$

## 3.7 Classes non balancées

Dans de nombreux problèmes de classification, il est fréquent de rencontrer des situations où les classes ne sont pas également représentées dans les données d'apprentissage. Ce déséquilibre de classes peut entraîner un biais dans la performance du modèle, en particulier lorsque la classe minoritaire est d'un intérêt particulier. Pour remédier à ce problème, plusieurs stratégies peuvent être mises en œuvre afin d'assurer un apprentissage plus équilibré et une meilleure performance de généralisation.

L'une des approches les plus courantes pour traiter le déséquilibre des classes consiste à utiliser des techniques de **rééchantillonnage**. Le rééchantillonnage peut impliquer de **suréchantillonner** la classe minoritaire, c'est-à-dire de créer des copies supplémentaires d'échantillons appartenant à la classe sous-représentée, ou de **sous-échantillonner** la classe majoritaire, en éliminant certains échantillons appartenant à la classe sur-représentée. Les deux approches visent à créer un ensemble de données d'apprentissage plus équilibré, ce qui peut améliorer la performance du modèle lorsqu'il est confronté à des données inédites [186, 187].

En outre, il est crucial d'utiliser des métriques de performance adaptées pour évaluer la performance d'un modèle dans des scénarios où les classes ne sont pas équilibrées. Les métriques telles que la précision, le rappel et le score F1 peuvent être plus informatives que la simple exactitude, car elles tiennent compte du nombre de vrais positifs, de faux positifs et de faux négatifs pour chaque classe.

## 3.8 Conclusion

Dans ce chapitre, nous avons examiné les techniques d'apprentissage automatique, en mettant l'accent sur l'apprentissage supervisé et non supervisé. Nous avons discuté des algorithmes de classification et de régression, ainsi que des méthodes de clustering et de détection d'anomalies. Nous avons également abordé l'évaluation des modèles, les défis tels que le sur-apprentissage et le sous-apprentissage, et les stratégies pour gérer les classes déséquilibrées. Ce chapitre fournit une vue d'ensemble complète des techniques d'apprentissage automatique pertinentes pour les praticiens et les chercheurs.

# Partie II : Contributions

# Chapitre 4

# Prédiction du Décrochage des étudiants dans les MOOCs

## 4.1 Introduction

Depuis leur émergence en 2008, les MOOCs ont radicalement révolutionné le panorama éducatif. Ces plateformes se sont imposées comme un outil indispensable à la démocratisation de l'accès au savoir[1]. Au fil des années, leur popularité n'a cessé de croître. En 2020, plus de 16 300 cours étaient offerts par près de 950 institutions universitaires, bénéficiant à une audience mondiale de 180 millions d'apprenants[2]. Plus qu'un simple outil pédagogique, les MOOCs sont devenus une ressource essentielle pour le développement personnel et la formation continue, comblant l'écart entre les compétences théoriques acquises au cours des études et les besoins concrets du monde professionnel.

Malgré les avantages incontestables des MOOCs et leur potentiel pour améliorer l'expérience d'apprentissage des étudiants, plusieurs défis demeurent. Le plus préoccupant est sans doute le taux d'abandon très élevé. Moins de 10% des étudiants inscrits parviennent à terminer leur parcours [188], ce qui pose un problème majeur. Plusieurs facteurs peuvent expliquer cette situation, notamment le manque de soutien individualisé des tuteurs face à un nombre considérable de participants, les niveaux d'interaction généralement faibles entre les apprenants, et l'absence de sanctions pour l'abandon du cours. Il est important de souligner que les parcours d'engagement des étudiants dans les MOOCs sont variés et comprennent des finalistes, des auditeurs, des désengagés et des échantillonneurs, chacun ayant des besoins et des comportements spécifiques [189].

Pour pallier ce taux d'abandon élevé, des recherches ont été menées pour mettre en place des outils d'analyse d'apprentissage et des algorithmes de machine learning permettant de prédire les étudiants susceptibles d'abandonner [190]. Si ces approches ont montré un certain potentiel, elles restent limitées par le nombre élevé d'inscriptions, le retard dans l'identification des étudiants à risque et l'absence de systèmes de priorisation. Il est donc nécessaire de concevoir un système capable d'identifier rapidement les étudiants à risque et de fournir une intervention personnalisée.

L'objectif de ce travail est de construire un tel système, utilisant les techniques et méthodes exposées dans la partie théorique de cette thèse, telles que les algorithmes d'apprentissage supervisé et les méthodes de réduction de dimensionnalité dans le domaine de l'analytique de l'apprentissage.

Dans la partie pratique de cette thèse, nous appliquerons les techniques et méthodes discutées dans la partie théorique sur l'ensemble de données du KDD Cup 2015 [191]. Cette base de données, riche en enregistrements des activités des étudiants sur la plateforme de MOOC chinoise Xuetang, nécessite toutefois un traitement conséquent avant d'être utilisable pour notre étude. Nous entamerons donc par une phase d'ingénierie des caractéristiques, destinée à extraire des attributs pertinents à partir des données brutes. Ensuite, nous mettrons en œuvre des techniques de sélection d'attributs et de réduction de dimensionnalité pour diminuer la complexité des modèles que nous construirons tout en préservant une représentativité optimale de l'information. En parallèle à l'utilisation de l'apprentissage profond et des algorithmes génétiques, nous implémenterons trois algorithmes de base d'apprentissage automatique : la régression logistique, les forêts aléatoires (Random Forest) et le Gradient Boosting. Ces méthodes serviront à évaluer l'efficacité de notre approche de réduction de dimensionnalité et de sélection d'attributs. Pour pallier le déséquilibre des classes dans notre ensemble de données, nous utiliserons également la technique de suréchantillonnage des minorités synthétiques (SMOTE). Pour évaluer la performance de nos modèles, nous recourrons à la méthode de validation croisée à 10 plis, et nous nous appuierons sur deux métriques d'évaluation : la précision et le score AUC ROC. Enfin, nous établirons un système intelligent pour aborder le problème de la prédiction précoce de l'abandon des étudiants, en priorisant ceux nécessitant une intervention selon un modèle temporel hebdomadaire basé sur leur probabilité d'abandon.

Le système que nous proposons permet d'identifier à grande échelle les étudiants à haut risque d'abandon, tout en étant en mesure de proposer une intervention précoce et adaptée. Ainsi, en fonction des résultats obtenus, une intervention personnalisée pourra être mise en place pour retenir les étudiants à risque, ce qui constitue une avancée majeure dans le domaine des MOOCs. Le potentiel de cette démarche est immense et pourrait significativement réduire le taux d'abandon dans les MOOCs, tout en améliorant la qualité de l'expérience d'apprentissage pour des millions d'étudiants à travers le monde.

## 4.2 Travaux Connexes

De nombreux chercheurs se sont récemment intéressés au phénomène complexe de décrochage des étudiants dans les MOOCs [27] [192] [193] . C'est une problématique qui se situe au carrefour de plusieurs disciplines, de la psychologie de l'éducation à l'analyse des données. La prise en compte du temps s'avère être un aspect particulièrement significatif lorsqu'il s'agit de s'attaquer à ce défi majeur. La détection précoce joue en effet un rôle crucial dans la réduction du taux d'attrition, permettant d'intervenir plus tôt et de façon plus ciblée pour soutenir les étudiants en difficulté.

Des recherches antérieures ont montré que la majorité des abandons, soit 75%, surviennent dès les premières semaines d'un cours [194]. Dans ce contexte, Gitinabard et al [195] ont mis en lumière le potentiel des données massives générées par les plateformes d'apprentissage en ligne. En analysant les

journaux de connexion et les données de forum d'un cours annuel d'un MOOC, ils ont pu identifier des tendances et des motifs de comportement parmi les étudiants. Ils ont appliqué des techniques de régression logistique et des machines à vecteurs de support (SVM) pour prédire le décrochage dès les premières semaines de chaque cours. Les étudiants ont été signalés comme étant à risque d'abandonner avec une précision impressionnante de 70% après la troisième semaine.

Berens et al. [196], en s'appuyant sur une approche complémentaire, ont développé un système innovant de prédiction précoce en exploitant des données démographiques ainsi qu'un algorithme de boosting. Celui-ci combine plusieurs algorithmes de Machine Learning reconnus pour leur efficacité : la Régression Linéaire, les Réseaux Neuronaux, les Arbres de Décision et AdaBoost. Le système a démontré sa pertinence avec une précision de 58,2% pour le premier semestre et de 81,5% pour le quatrième semestre.

Dans un souci d'optimisation encore plus poussée, les auteurs de [197] ont fait appel à l'apprentissage profond, une technique de pointe en matière de Machine Learning. Ils ont ainsi réussi à atteindre une précision de 0,92% dès la première semaine. Cependant, malgré ces avancées, une difficulté demeure. La performance des algorithmes de Machine Learning dépend fortement du jeu de données utilisé. De plus, ces algorithmes ne fournissent pas une méthode directe et efficace pour régler les paramètres, ce qui peut limiter leur utilisation optimale.

C'est ici qu'intervient une autre dimension de notre recherche, la sélection des caractéristiques des paramètres. Nous avons donc proposé une méthode hybride de sélection des caractéristiques, combinant diverses approches pour choisir le sous-ensemble de caractéristiques le plus pertinent. De plus, nous avons recours aux algorithmes génétiques (GA), couramment utilisés pour les problèmes d'optimisation, pour l'ajustement des paramètres. Les GA, bien qu'ils ne soient pas couramment utilisés dans le domaine du décrochage des étudiants, ont montré leur efficacité dans de nombreux domaines tels que la reconnaissance des émotions [198], la médecine [199], et la sélection des caractéristiques [200].

C'est ainsi que notre étude contribue de manière significative à l'état actuel du domaine, en alliant l'exploration de nouvelles approches d'analyse des données à l'application de techniques éprouvées. Elle se distingue par le développement d'une approche globale pour l'étude et la détection précoce du décrochage dans une perspective de données. Mais au-delà de l'aspect technique, elle met également l'accent sur l'importance d'une intervention adaptée aux besoins des étudiants. C'est pourquoi nous avons également conçu une intervention prioritaire et personnalisée pour les étudiants à risque de décrochage, qui prend en compte non seulement leur profil et leur comportement, mais aussi leurs aspirations et leur contexte individuel.

En somme, il est essentiel de continuer à explorer et à développer des méthodes innovantes pour lutter contre le décrochage dans les MOOCs. Les perspectives ouvertes par les avancées technologiques et méthodologiques, comme celles présentées dans cette étude, offrent de nouvelles opportunités pour

améliorer la rétention et la réussite des étudiants dans l'éducation en ligne.

## 4.3 Présentation des données

Cette étude s'appuie sur un ensemble de données issu de la KDD Cup 2015, un événement compétitif de renom organisé par le Groupe d'intérêt spécial de l'ACM (Association for Computing Machinery). Cette compétition avait pour objectif d'encourager les scientifiques des données à prédire le taux de décrochage scolaire en recourant à des techniques avancées d'exploration de données et d'apprentissage automatique.

La KDD Cup 2015 s'est distinguée en proposant un défi centré sur l'éducation, plus précisément sur le phénomène du décrochage scolaire dans le contexte des cours en ligne ouverts et massifs (MOOC). Pour ce faire, elle a mis à disposition un ensemble de données particulièrement riche, capturant les traces d'interaction des utilisateurs avec XuetangX, l'une des plus grandes plateformes d'apprentissage en ligne de Chine. Les données détaillées et exhaustives, comprennent les journaux de suivi des utilisateurs, offrant une vue précise de leurs activités et comportements sur la plateforme. En se plongeant dans ces journaux, les chercheurs et les data scientists ont la possibilité d'extraire des informations précieuses, qui pourraient servir à prédire les comportements futurs des utilisateurs, notamment leur propension à abandonner un cours.

Pour mesurer la précision des modèles prédictifs développés par les participants à la compétition, le score AUC ROC (Area Under the Receiver Operating Characteristic) a été choisi comme métrique d'évaluation principale. Ce choix reflète l'importance de la performance des modèles en termes de précision et de sensibilité, tout en tenant compte du taux de faux positifs.

Voici une description détaillée des cinq parties de cet ensemble de données :

1. La première partie de l'ensemble de données contient des informations sur la date de début et de fin de chaque cours selon la table 4.1.

Table 4.1: Informations sur les cours

| Champs | Type | Description |
|---|---|---|
| Course-ID | Nominal | Identifiant du cours |
| Course-S | Date | Date de début du cours |
| Course-E | Date | Date de fin du cours |

2. Informations sur les modules de chaque cours, les sous-modules et également la catégorie des modules et leur date de début selon la table 4.2
3. Informations sur les inscriptions : une inscription est une entrée (Étudiant, Cours) selon la table 4.3.

TABLE 4.2: Informations sur les modules

| Champs | Type | Description |
|---|---|---|
| Course-ID | Nominal | Identifiant du cours |
| Module-ID | Nominal | Identifiant du module |
| Module-cat | Nominal | Catégorie du module |
| Module-child | Nominal | Sous-module |
| Module-S | Date | Date de début du module |

TABLE 4.3: Informations sur les inscriptions

| Champs | Type | Description |
|---|---|---|
| Enrolment-ID | Nominal | Identifiant de l'inscription |
| Student-ID | Nominal | Identifiant de l'étudiant |
| Course-ID | Nominal | Identifiant du cours |

4. La quatrième partie de l'ensemble de données contient une trace de chaque inscription, des horodatages de journal, et la source et le type de l'événement selon la table 4.4

TABLE 4.4: Informations sur les logs

| Fields | Type | Description |
|---|---|---|
| Enrolment-ID | Nominal | Enrolment Identifier |
| Student-ID | Nominal | Student Identifier |
| Course-ID | Nominal | Course Identifier |
| Event-Time | Timestamps | Time when the event occurs |
| Source | Nominal | Source of Event (Server / Browser) |
| Event | Nominal | Event Type: Problem, Access, Video, Wiki, Discussion, Navigate, Page Close |

5. La dernière partie de l'ensemble de données contient des informations sur la valeur du résultat de l'inscription selon la table 4.5.

Le jeu de données décrit le journal de suivi du MOOC du 27 octobre 2013 au 1er août 2014. Pendant cette période, trente-neuf cours ont été lancés, chaque cours durant 29 jours. Le nombre total

TABLE 4.5: Résultats des Étudiants

| Champs | Type | Description |
|---|---|---|
| Enrolment-ID | Nominal | Identifiant de l'inscription |
| Résultat | Booléen | Résultat de l'étudiant 0 Succès 1 Abandon |

d'étudiants est de 79186, et le nombre d'inscriptions s'élève à 120543. Un étudiant peut s'inscrire à plusieurs cours.

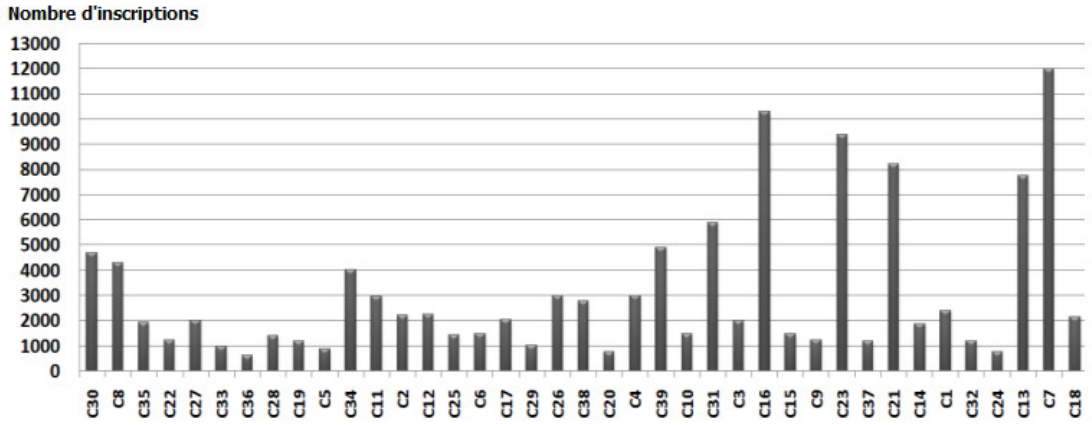



FIGURE 4.1: Nombre d'inscrits par cours

Les Figures 4.1 et 4.2 présentent le nombre d'inscriptions ainsi que le taux d'abandon pour chaque cours. Nous constatons que pour tous les cours, le pourcentage d'abandon dépasse 65%. Par conséquent, dans le tableau des résultats d'inscription, la sortie 1, qui correspond à l'abandon, dépasse 0, qui correspond à la réussite. Nous sommes donc confrontés à un problème de déséquilibre des classes dans notre jeu de données. La classe majoritaire est celle de l'abandon, avec (95581) 79% des inscriptions, comparée à la réussite (24961) 21%.

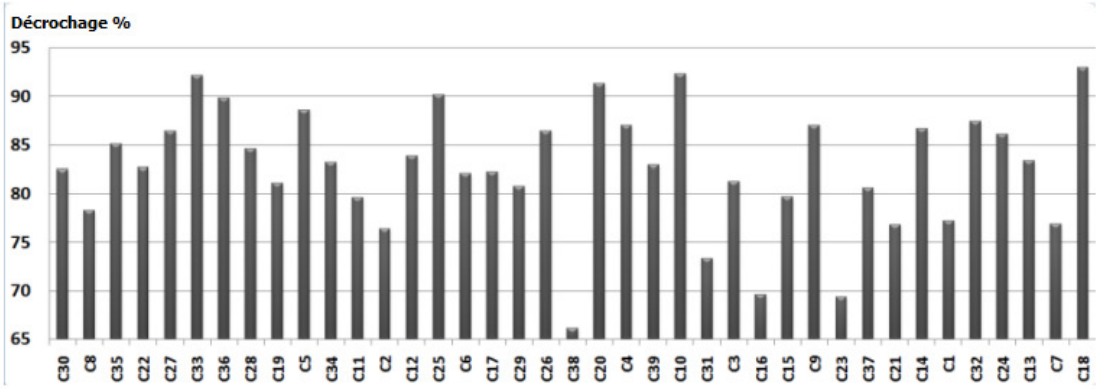



FIGURE 4.2: Taux de décrochage par cours

Dans ce jeu de données, nous nous heurtons à un problème de déséquilibre des classes. Diverses solutions s'offrent à nous pour rééquilibrer le jeu de données. L'une d'entre elles consiste à suréchantillonner

la classe minoritaire, c'est-à-dire à augmenter le nombre de ses instances jusqu'à égaler celui de la classe majoritaire.

Comme évoqué dans la partie théorique, le suréchantillonnage est une technique qui permet de créer des doublons de certaines instances de la classe minoritaire pour augmenter sa cardinalité. On peut générer de nouveaux échantillons de manière aléatoire ou en faisant appel à des méthodes plus sophistiquées telles que SMOTE ou ADASYN.

Le choix de la technique à privilégier dépend étroitement du jeu de données utilisé. Pour notre part, c'est la méthode SMOTE qui a donné les meilleurs résultats.

Pour quantifier l'effet du rééquilibrage, nous avons créé deux sous-ensembles de données : le jeu de données déséquilibré (Unbalanced Dataset - UB) et le jeu de données équilibré (Balanced Dataset - BD). L'étude de ces deux jeux de données va nous permettre d'approfondir notre analyse du phénomène de déséquilibre des classes.

## 4.4 Ingénierie des caractéristiques

Les informations disponibles dans le jeu de données de la KDD Cup 2015 manquent d'informations personnelles (par exemple, âge, sexe, nationalité) et d'informations concernant le cours (par exemple, prérequis, niveau de difficulté). La trace de connexion reste notre source d'information la plus précieuse. Notre méthode d'extraction de caractéristiques se base sur le comptage des logs de chaque inscription ; une inscription est une entrée (étudiant, cours). Les caractéristiques extraites peuvent être divisées en deux parties :

— Caractéristiques de l'historique des inscriptions : Elles contiennent des informations sur l'historique des interactions avec le MOOC, telles que le nombre de cours réussis, le nombre de cours échoués, le nombre cumulatif de jours passés sur le MOOC lors des anciennes inscriptions, et le nombre cumulatif de logs pour chaque événement présent dans le catalogue du tableau des événements.

— Caractéristiques de l'inscription actuelle : Elles contiennent des informations sur le nombre de jours passés sur le MOOC lors de l'inscription actuelle et le nombre de logs pour chaque événement.

Nous extrayons également le nombre de minutes passées pour chaque inscription. En effet, après avoir examiné la trace des logs, nous constatons que toutes les sessions commencent par l'événement "Navigate" et parfois par l'événement "Access" et se terminent par l'événement "Page close". Nous calculons la différence de temps, exprimée en minutes, entre l'un de ces deux événements et la fin de la session. L'accumulation de minutes est enregistrée pour chaque inscription dans la caractéristiques 'ConnectedMin' selon l'algorithme suivant :

**Algorithm 9** Minutes Connectées

```
@ConnectedMin ← 0
@BeginDay ← 00 :00 :00
@EndDay ← 23 :59 :59
@TBegin ← ”
@TEnd ← ”
while pas à la fin des lignes du journal d'inscription do
    lire actuel
    if @Evente = 'Naviguer' ou (@Evente = 'Accès' et @TimeBegin = ”) then
        @TimeBegin ← Temps-Événement
    end if
    if @Evente = 'Fermeture de page' then
        @TimeEnd ← Temps-Événement
    end if
    @Diff ← DiffTemps(MIN, @TBegin, @TEnd)
    if @Diff < 0 then
        @x ← DiffTemps(MIN, @BeginDay, @EndDay)
        @y ← DiffTemps(MIN, @BeginDay, @EndDay)
        @Diff ← @x + @y
    end if
    @ConnectedMin ← @ConnectedMin + @Diff
End while
End
```

Cet algorithme, intitulé “Minutes Connectées”, est utilisé pour calculer le temps total passé par un utilisateur sur une plateforme d'apprentissage en ligne, à travers diverses inscriptions à des cours. L'algorithme s'appuie sur les journaux d'événements de la plateforme pour déterminer le temps de début et de fin des sessions de chaque utilisateur. Voici une brève description de son fonctionnement :

1. Il initialise plusieurs variables, y compris une pour enregistrer le temps total passé connecté (denommée @ConnectedMin), et des variables pour marquer les instants de début (@TBegin) et de fin (@TEnd) d'une session.
2. Ensuite, l'algorithme parcourt chaque ligne du journal d'inscription, en extrayant les informations nécessaires. Si l'événement correspond à une navigation ou un accès, il note le temps de début.
3. Lorsqu'il rencontre un événement de “fermeture de page”, il enregistre ce temps comme le temps

de fin.

4. Il calcule ensuite la différence entre le temps de début et de fin pour obtenir la durée de la session en cours.
5. Si la différence de temps est négative (ce qui pourrait se produire en raison du passage d'un jour à l'autre), l'algorithme ajuste cette valeur en ajoutant la durée d'une journée complète.
6. Finalement, il ajoute la durée de la session actuelle au temps total passé connecté.
7. L'algorithme continue d'itérer à travers chaque ligne du journal jusqu'à ce qu'il ait terminé, fournissant à la fin le temps total passé par l'utilisateur sur la plateforme.

Cet algorithme est utile pour comprendre l'engagement des utilisateurs dans les cours en ligne et pourrait être utilisé pour déduire les comportements d'apprentissage, les taux d'achèvement des cours.

Le tableau 4.6 présente l'ensemble des caractéristiques extraites qui vont constituer le vecteur de données. Ce vecteur sera ensuite soumis à une réduction de dimensionnalité.

Table 4.6: Caractéristiques extraites

| N | Caractéristiques |
|---|---|
| 1 | Identifiant d'inscription |
| 2 | Nombre d'inscriptions précédentes de l'étudiant |
| 3 | Nombre d'inscriptions réussies précédentes de l'étudiant |
| 4 | Nombre d'abandons précédents de l'étudiant |
| 5 | Nombre de logs pour l'inscription en cours |
| 6 | Nombre de logs pour toutes les inscriptions précédentes |
| 7 | Nombre de jours entre le premier et le dernier log |
| 8 | Nombre de jours entre le premier et le dernier log pour toutes les inscriptions précédentes |
| 9 | Nombre de logs pour l'événement : Problème |
| 10 | Nombre de logs pour l'événement : Problème pour toutes les inscriptions précédentes |
| 11 | Nombre de logs pour l'événement : Vidéo |
| 12 | Nombre de logs pour l'événement : Vidéo pour toutes les inscriptions précédentes |
| 13 | Nombre de logs pour l'événement : Navigation |
| 14 | Nombre de logs pour l'événement : Navigation pour toutes les inscriptions précédentes |
| 15 | Nombre de logs pour l'événement : Fermeture de page |
| 16 | Nombre de logs pour l'événement : Fermeture de page pour toutes les inscriptions précédentes |
| 17 | Nombre de logs pour l'événement : Accès |
| 18 | Nombre de logs pour l'événement : Accès pour toutes les inscriptions précédentes |
| 19 | Nombre de logs pour l'événement : Discussion |

| N | Caractéristiques |
|---|---|
| 20 | Nombre de logs pour l'événement : Discussion pour toutes les inscriptions précédentes |
| 21 | Nombre de logs pour l'événement : Wiki |
| 22 | Nombre de logs pour l'événement : Wiki pour toutes les inscriptions précédentes |
| 23 | Nombre de logs dans les 10 premiers jours de cours |
| 24 | Nombre de logs dans les 10 jours suivants du cours |
| 25 | Nombre de logs dans les 10 derniers jours de cours |
| 26 | Nombre de minutes actives |
| 27 | Nombre de jours actifs |
| Résultat de l'inscription : Succès 0 / Abandon 1 | |

## 4.5 Réduction de la dimensionalité

La plupart des problèmes de classification sont basés sur le traitement des données extraites, qui sont structurées sous la forme de vecteurs. La qualité du classificateur est directement liée à la pertinence des éléments sélectionnés pour ces vecteurs. Cependant, dans nombre de cas, la résolution du problème est entravée par la haute dimensionnalité des vecteurs de données. En effet, une augmentation non significative des données peut augmenter la complexité et le temps de calcul. À cet effet, les techniques de sélection de caractéristiques et de réduction de dimensionnalité sont un remède systématique à ce problème, en supprimant les caractéristiques qui ne fournissent pas suffisamment d'informations prédictives au modèle. La diminution du nombre de caractéristiques a un double avantage. D'une part, le modèle devient plus compréhensible et, d'autre part, l'élimination des caractéristiques non informatives améliore la précision de prédiction des modèles utilisés.

Dans le cadre de cette étude, nous avons adopté une approche hybride, combinant différentes méthodes de sélection de caractéristiques et de réduction de la dimensionnalité. Ces méthodes, déjà présentées dans la partie théorique, nous ont permis d'évaluer la pertinence de chaque caractéristique de manière exhaustive.

1. **Sélection de caractéristiques :**
   - **Méthode de Filtrage – Test du Chi-carré :** Cette technique statistique nous a permis de déterminer le degré d'indépendance de chaque caractéristiques par rapport à la variable cible. Ainsi, chaque caractéristique a reçu un score d'indépendance, et celles affichant un score faible, supposées indépendantes de la variable cible, ont été éligibles à l'élimination.
   - **Méthodes de Wrapper – RFE-RF et RFE-ADA :** Ces méthodes ont attribué un classement à chaque caractéristique basé sur son importance dans un modèle spécifique. Le RFE-RF a été utilisé avec un modèle de forêt aléatoire et le RFE-ADA avec un modèle de

boosting adaptatif.

— **Méthodes Intégrées – Lasso et Ridge :** Ces techniques de régularisation ont assigné un coefficient à chaque caractéristique dans un modèle linéaire, reflet de son importance dans la prédiction de la variable cible. Le Lasso a tendance à réduire à zéro les coefficients des caractéristiques moins pertinentes, tandis que la Ridge réduit les coefficients vers zéro, sans les annuler totalement.

2. **Transformation des Données :**
   — **Analyse en Composantes Principales (PCA) :** Nous avons employé cette technique de réduction de dimensionnalité pour examiner la variance inhérente aux données et déterminer le nombre optimal de composantes à conserver.
   — **t-SNE (t-Distributed Stochastic Neighbor Embedding) :** Cette méthode de réduction de dimensionnalité a été privilégiée pour la visualisation des données de haute dimension.

Il convient de souligner que chaque méthode produit un type de sortie différent : coefficients pour le Lasso et le Ridge, classements pour RFE-RF et RFE-ADA, scores d'indépendance pour le test du Chi-carré, nombre de composantes pour la PCA et visualisation pour le t-SNE. Afin de faciliter l'agrégation et l'interprétation de ces diverses sorties, une normalisation des scores s'est avérée nécessaire. Celle-ci nous a permis de ramener ces scores hétérogènes sur une échelle commune allant de 0 (pour le score le plus bas) à 1 (pour le score le plus élevé). De ce fait, les caractéristiques affichant un score moyen élevé ont été retenues, comme illustré dans le tableau 4.7. C'est cette harmonieuse fusion des différentes méthodes qui rend notre approche hybride à la fois robuste et exhaustive, nous assurant une analyse approfondie et nuancée des caractéristiques.

Selon les résultats trouvés dans le tableau 4.7, nous avons choisi 0,64 comme seuil statistique pour la sélection des caractéristiques. Ce seuil a été déterminé afin de maintenir un équilibre optimal : il permet de ne pas écarter un nombre significatif de caractéristiques potentiellement informatives tout en évitant d'inclure un trop grand nombre de caractéristiques, ce qui pourrait entraîner une complexité excessive du modèle. En conséquence, la dimensionnalité de l'ensemble de données sera réduite aux six caractéristiques suivantes : 26 (0,81), 27 (0,80), 25 (0,80), 5 (0,69), 7 (0,67), et 3 (0,64).

## 4.6 Résultats

Poursuivant les discussions théoriques de cette thèse, nous avons identifié notre problème comme relevant d'une classification binaire statistique. Cela est justifié par notre variable de sortie qui admet deux valeurs possibles : 0 pour le succès et 1 pour l'abandon. Les modèles de classification binaire entrent dans le domaine des méthodes d'apprentissage supervisé.

Dans le contexte de l'apprentissage supervisé, décrit en profondeur dans la partie théorique de cette

TABLE 4.7: Classement des caractéristiques

| N | Lasso | REF-RF | REF-ADA | Ridge | Chi-2 | Moyenne |
|---|---|---|---|---|---|---|
| 1 | 0.00 | 1.00 | 0.46 | 0.00 | 0.00 | 0.29 |
| 2 | 0.00 | 0.57 | 0.04 | 0.42 | 1.00 | 0.41 |
| 3 | 0.00 | 0.29 | 0.92 | 1.00 | 1.00 | 0.64 |
| 4 | 0.00 | 0.71 | 0.08 | 0.58 | 1.00 | 0.47 |
| 5 | 0.66 | 1.00 | 0.79 | 0.01 | 0.97 | 0.69 |
| 6 | 0.05 | 1.00 | 0.12 | 0.01 | 1.00 | 0.44 |
| 7 | 0.00 | 1.00 | 1.00 | 0.5 | 1.00 | 0.67 |
| 8 | 0.00 | 1.00 | 0.71 | 0.00 | 1.00 | 0.54 |
| 9 | 0.00 | 1.00 | 0.96 | 0.01 | 0.99 | 0.59 |
| 10 | 0.00 | 1.00 | 0.50 | 0.01 | 1.00 | 0.50 |
| 11 | 0.00 | 1.00 | 0.54 | 0.02 | 1.00 | 0.51 |
| 12 | 0.00 | 1.00 | 0.37 | 0.01 | 1.00 | 0.48 |
| 13 | 0.00 | 1.00 | 0.58 | 0.05 | 1.00 | 0.53 |
| 14 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.20 |
| 15 | 0.00 | 1.00 | 0.62 | 0.09 | 1.00 | 0.54 |
| 16 | 0.00 | 1.00 | 0.17 | 0.01 | 1.00 | 0.44 |
| 17 | 0.00 | 1.00 | 0.33 | 0.02 | 0.99 | 0.47 |
| 18 | 0.00 | 1.00 | 0.29 | 0.01 | 1.00 | 0.46 |
| 19 | 0.00 | 0.14 | 0.25 | 0.00 | 1.00 | 0.28 |
| 20 | 0.00 | 0.86 | 0.67 | 0.01 | 1.00 | 0.51 |
| 21 | 0.00 | 1.00 | 0.21 | 0.06 | 1.00 | 0.45 |
| 22 | 0.00 | 0.43 | 0.75 | 0.00 | 1.00 | 0.44 |
| 23 | 0.00 | 1.00 | 0.87 | 0.03 | 0.99 | 0.58 |
| 24 | 0.00 | 1.00 | 0.83 | 0.01 | 0.99 | 0.57 |
| 25 | 1.00 | 1.00 | 1.00 | 0.03 | 0.99 | 0.80 |
| 26 | 0.60 | 1.00 | 0.87 | 0.58 | 1.00 | 0.81 |
| 27 | 0.00 | 1.00 | 1.00 | 0.98 | 1.00 | 0.80 |

thèse, nous avons réparti nos données en deux ensembles : les données d'entraînement et les données de test. Initialement, les données d'entraînement sont introduites dans un algorithme d'apprentissage automatique. Cet algorithme s'efforce de découvrir le modèle le mieux ajusté en optimisant les paramètres pour minimiser le biais. Le modèle ainsi obtenu est ensuite employé pour classer de nouvelles observations, qui sont ultérieurement comparées à la cible dans l'ensemble de test. Par la suite, nous calculons

l'exactitude du modèle, une métrique reflétant sa capacité à effectuer des prédictions correctes.

Afin de détecter des algorithmes potentiellement plus performants pour notre jeu de données, nous avons mis en place une boucle de comparaison d'algorithmes. Cette approche nous a permis d'évaluer la performance des modèles générés en se basant sur la moyenne des scores de précision et de l'AUC ROC obtenus lors d'une validation croisée en 10 plis. Nous avons utilisé à la fois les données avec balancement (BD) et sans balancement (UD) pour cette évaluation dans le tableau 4.8.

TABLE 4.8: Scores des modèles sur les données d'apprentissage

| **Algorithmes d'Apprentissage Automatique** | **Modèles** | **Précision** | **AUC ROC** |
|---|---|---|---|
| Régression Logistique | LR UD | 0.87 | 0.87 |
| | LR BD | 0.80 | 0.80 |
| k-Nearest Neighbours | KNN UD | 0.85 | 0.82 |
| | KNN BD | 0.85 | 0.84 |
| Gradient Boosting | GB UD | 0.87 | 0.87 |
| | GB BD | 0.88 | 0.86 |
| Arbre de Décision | DT UD | 0.80 | 0.68 |
| | DT BD | 0.86 | 0.86 |
| Forêt Aléatoire | RF UD | 0.85 | 0.86 |
| | RF BD | 0.86 | 0.88 |

Il est essentiel de mentionner qu'à ce stade de notre recherche, nous n'avons pas effectué un entraînement complet des modèles. Nous avons plutôt réalisé un "passage simple" sur les algorithmes de base. Cette approche, consistant en une évaluation préliminaire des performances de ces algorithmes, nous a permis d'obtenir une première impression des modèles qui pourraient potentiellement offrir de bons résultats.

La plupart des algorithmes d'apprentissage automatique nécessitent d'ajuster des paramètres pour obtenir les meilleurs résultats possibles. Utiliser une recherche aléatoire pour trouver la bonne combinaison de paramètres peut être très consommateur en termes de temps et de ressources. C'est là que la recherche en grille avec validation croisée intervient. Il s'agit d'une méthode d'optimisation qui permet de tester une série de paramètres, de comparer les performances et de choisir la meilleure configuration. Pour chaque paramètre, on détermine un ensemble de valeurs, puis la recherche en grille évalue chaque combinaison en créant un modèle. Pour chaque modèle, nous utilisons une validation croisée à 10 plis pour valider la combinaison la plus pertinente pour plusieurs paramètres. Les sous-sections suivantes analysent ces paramètres :

1. Régression logistique
   — Paramètre de régularisation C : Le paramètre de régularisation C est un facteur d'échelle

qui inverse la force de régularisation. Une valeur plus élevée de C indique une régularisation plus faible, tandis qu'une valeur plus faible de C entraîne une régularisation plus forte.

2. Forêt aléatoire
   — Nombre d'estimateurs : Le nombre d'arbres utilisés dans la forêt. Un nombre plus élevé d'estimateurs peut améliorer la performance, mais cela doit être équilibré avec le temps de calcul nécessaire pour estimer le modèle.
   — Critère : Le critère utilisé pour évaluer la qualité des divisions dans les arbres, tels que le critère de Gini ou l'entropie.
   — Nombre maximal de caractéristiques (max features) : Le nombre maximal de caractéristiques à prendre en compte lors de la recherche de la meilleure division.
   — Profondeur maximale (max depth) : La profondeur maximale des arbres dans la forêt.
   — Nombre minimum d'échantillons par feuille (min samples leaf) : Le nombre minimum d'échantillons requis dans une feuille de l'arbre.
   — Nombre minimum d'échantillons pour effectuer une division (min samples split) : Le nombre minimum d'échantillons requis dans un nœud interne pour effectuer une division.
3. Gradient Boosting
   — Taux d'apprentissage (learning rate) : Le taux d'apprentissage contrôle l'impact de chaque arbre ajouté à l'ensemble. Des valeurs inférieures ou égales à 0,1 sont souvent utilisées pour obtenir de meilleurs modèles.

Le tableau 4.8 fournit les performances de divers modèles fondés sur plusieurs algorithmes d'apprentissage automatique. Les indicateurs de performance incluent la précision des modèles et l'aire sous la courbe ROC, mesurées sur des ensembles de données équilibrés et non équilibrés. Les données suggèrent que certains algorithmes offrent de meilleurs résultats indépendamment du type de jeu de données. Dans ce contexte, nous utiliserons trois algorithmes répandus : la régression logistique, la forêt aléatoire et le Gradient Boosting. Ces algorithmes de base seront combinés avec des méthodes plus sophistiquées, à savoir l'apprentissage profond et les algorithmes génétiques, dans notre étude.

TABLE 4.9: Scores des modèles

| **Modèle** | **Précision** | **AUC ROC** | **Exactitude** | **Rappel** | **F1** | **TP** | **TN** | **FP** | **FN** |
|---|---|---|---|---|---|---|---|---|---|
| LR UD | 0.87 | 0.88 | 0.87 | 0.87 | 0.86 | 4023 | 27531 | 1223 | 3386 |
| LR BD | 0.82 | 0.87 | 0.81 | 0.80 | 0.80 | 16280 | 16002 | 3958 | 3857 |
| GB UD | 0.87 | 0.89 | 0.87 | 0.88 | 0.88 | 4346 | 27357 | 1318 | 3142 |
| GB BD | 0.90 | 0.90 | 0.88 | 0.88 | 0.88 | 16953 | 18466 | 1673 | 3005 |
| RF UD | 0.86 | 0.88 | 0.87 | 0.87 | 0.87 | 4026 | 27511 | 1243 | 3383 |
| RF BD | 0.86 | 0.90 | 0.83 | 0.83 | 0.83 | 15590 | 17605 | 2534 | 4368 |

Selon le tableau 4.9, nous avons obtenu des résultats avec une précision élevée de 0,90 pour l'ensemble de données équilibré et un score AUC ROC de 0,89 pour l'ensemble déséquilibré. Ces scores ont été obtenus en utilisant la méthode de validation croisée à 10 plis.

— **Effet de l'ajustement des hyperparamètres :**

Comme le montre le tableau 4.9, l'ajustement des hyperparamètres a eu un effet significatif sur les scores des modèles. Dans plusieurs cas, les scores de précision et d'AUC ROC ont augmenté après l'ajustement. Par exemple, l'ajustement des hyperparamètres a augmenté la précision des modèles LR BD et GB BD de 0,2. De plus, il a entraîné une augmentation de 0,1 de l'AUC ROC pour le modèle LR UD et de 0,2 pour les modèles RF UD et GB UD.

— **Impact de l'équilibrage des données :**

L'équilibrage des ensembles de données a eu un effet notable sur l'efficacité des algorithmes d'apprentissage ensembliste. Dans de nombreux cas, les scores de précision et d'AUC ROC ont été nettement améliorés après l'équilibrage. Par exemple, l'équilibrage des données a entraîné une augmentation de 0,1 de l'AUC ROC pour la méthode RF, comme le montre la figure 4.3. Cela démontre l'importance de cette étape préparatoire des données pour améliorer la performance des modèles de classification.

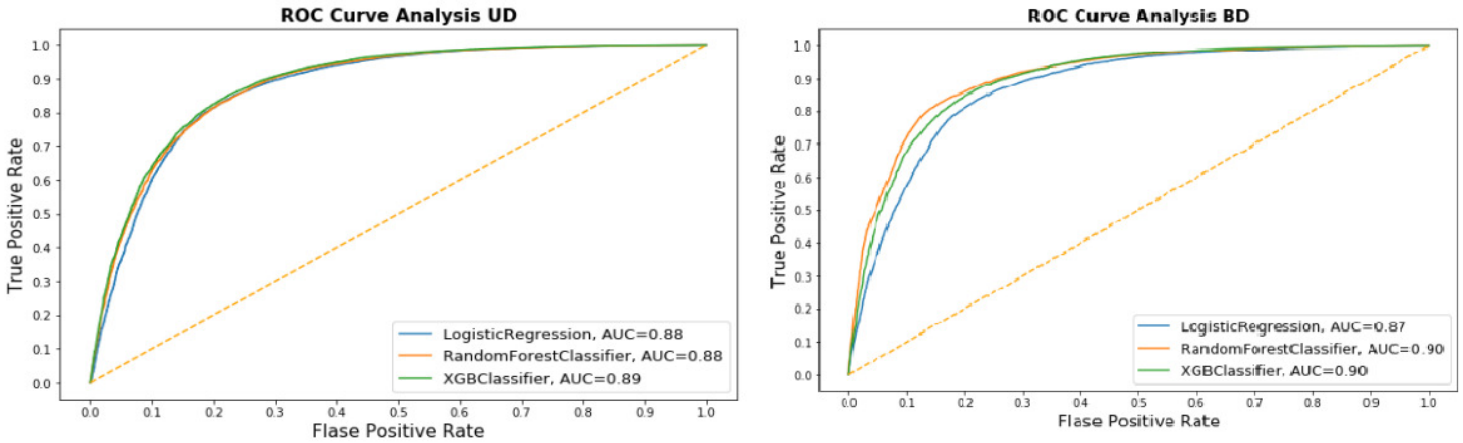



FIGURE 4.3: Analyse ROC Curve

— **TP (True Positive) et FN (False Negative) :**

Les valeurs TP (True Positive) et FN (False Negative) sont essentielles pour évaluer la performance des modèles de classification, surtout dans les contextes où les coûts des erreurs de type 1 et de type 2 peuvent être très différents.

Dans le cas de la détection de décrochage, maximiser les TP signifie identifier correctement le plus grand nombre possible d'élèves susceptibles de décrocher. Cela est essentiel pour pouvoir intervenir rapidement et efficacement pour les aider. En parallèle, minimiser les FN est tout aussi crucial : un FN signifie qu'un élève en danger de décrocher a été classé à tort comme n'étant pas à risque. Ignorer ces étudiants pourrait avoir des conséquences dévastatrices sur leur parcours

éducatif.

Au contraire, les erreurs de type 1 (FP - faux positifs) se produisent lorsqu'un étudiant qui ne risque pas de décrocher est incorrectement identifié comme étant à risque. Bien que ces erreurs puissent entraîner une utilisation inefficace des ressources, elles sont généralement considérées comme moins graves que les erreurs de type 2 (FN - faux négatifs) dans ce contexte.

Les résultats du tableau 4.9 indiquent que le modèle Gradient Boosting (GB BD et GB UD) est le plus performant pour maximiser les TP et minimiser les FN. C'est particulièrement vrai pour le modèle GB BD, qui a correctement identifié 16953 étudiants à risque sur l'ensemble de données équilibré.

Ces résultats soulignent l'efficacité du Gradient Boosting pour traiter ce type de problème de classification. Il semble que cet algorithme soit particulièrement adapté à la détection des élèves à risque de décrochage, ce qui pourrait avoir des implications importantes pour l'élaboration de stratégies d'intervention efficaces dans le domaine de l'éducation.

— **Effets de la réduction de la dimensionalité :Algorithmes supervisés**

Dans l'analyse des ensembles de données obtenus à partir de la méthode hybride de sélection des caractéristiques, et afin d'évaluer la pertinence de notre sélection nous avons examiné les performances de plusieurs modèles . Pour cela, nous avons utilisé deux jeux de données : le premier contenant toutes les caractéristiques extraites (All) et le deuxième ne contenant que les six caractéristiques sélectionnées (Selected). Les scores de précision et AUC ROC de ces deux jeux de données ont été comparés afin d'évaluer la pertinence de cette sélection de caractéristiques.

Table 4.10: Effet de la réduction sur les scores

| Algorithme ML | Caractéristiques | Précision | AUC ROC |
|---|---|---|---|
| Régression Logistique | All | 0.871 | 0.861 |
| | Selected | 0.870 | 0.863 |
| Gradient Boosting | All | 0.875 | 0.879 |
| | Selected | 0.874 | 0.877 |
| Random Forest | All | 0.873 | 0.876 |
| | Selected | 0.873 | 0.876 |

Selon les résultats obtenus dans le Tableau 4.10, nous remarquons que les résultats sont presque les mêmes malgré l'élimination de vingt caractéristiques. Nous remarquons également que pour la Régression Logistique, le score AUC ROC s'est amélioré lors de la suppression des caractéristiques non significatives.

Nous concluons donc que les caractéristiques éliminées ne sont pas importantes pour prédire la variable cible. Nous pouvons uniquement conserver les six variables sélectionnées.

En vue d'aller au-delà des performances offertes par les algorithmes classiques tels que la Régression Logistique, Gradient Boosting et Random Forest, nous avons mis en œuvre des modèles plus sophistiqués, notamment l'Algorithme Génétique (AG) et le Deep Learning (DL). Notre objectif était d'observer si ces méthodes plus avancées pouvaient surpasser les techniques de base en termes de précision et de performance de prédiction.

TABLE 4.11: Scores des modèles AG et DL

| Modèle | Caractéristiques | Précision | AUC ROC |
|---|---|---|---|
| AG | All | 0.926 | 0.898 |
| | Selected | 0.933 | 0.894 |
| DL | All | 0.943 | 0.876 |
| | Selected | 0.938 | 0.887 |

Les résultats obtenus, présentés dans le tableau 4.11, indiquent que les scores restent pratiquement les mêmes, même après l'élimination de plusieurs caractéristiques. Cela s'explique par le fait que certaines des caractéristiques éliminées n'avaient aucun rôle dans la prédiction de la variable cible. Dans certains cas, ces caractéristiques pouvaient même introduire du bruit.
Il est intéressant de noter que le score s'est amélioré pour le modèle AG après l'élimination des caractéristiques inutiles. Cela valide notre processus de sélection des caractéristiques, en mettant en évidence l'importance de n'inclure que les caractéristiques les plus pertinentes pour la prédiction de la variable cible. De plus, nous observons que les deux algorithmes employés dans ce travail, l'AG et le DL, surpassent les algorithmes de base utilisés précédemment. Cela s'explique par la nature évolutive et auto-correctrice de l'AG et les méthodes avancées de DL.

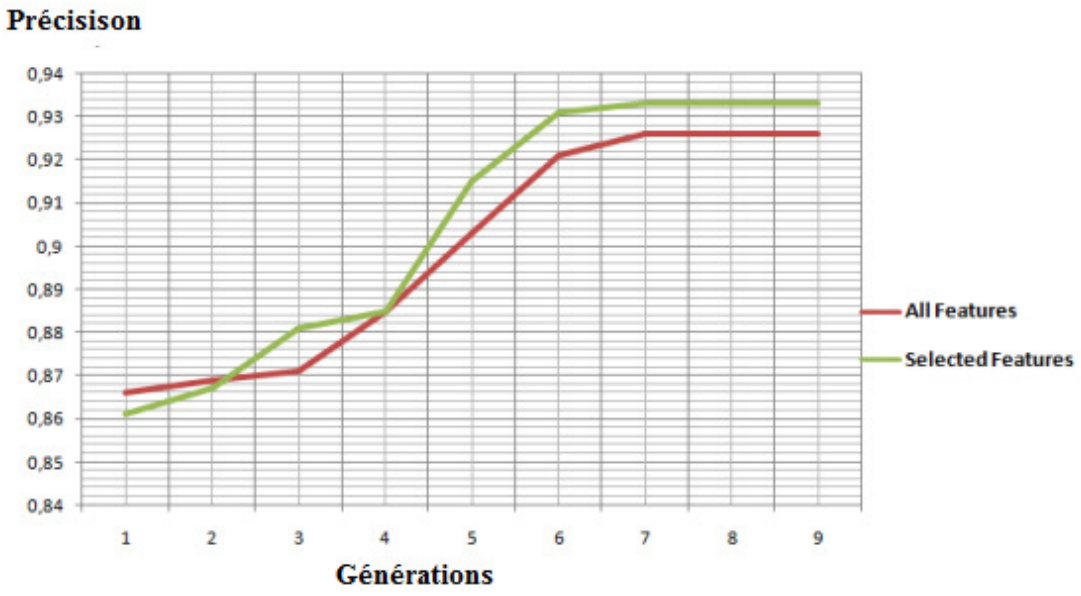



FIGURE 4.4: Évolution de la Précision de l'AG sur les Générations

L'efficacité de l'AG est illustrée par sa convergence rapide vers la solution optimale, comme

le montre la Figure 4.4. Après seulement sept générations, l'algorithme a trouvé une solution optimale pour le problème. Cette évolution rapide de la précision sur les générations dépend fortement du taux de mutation.

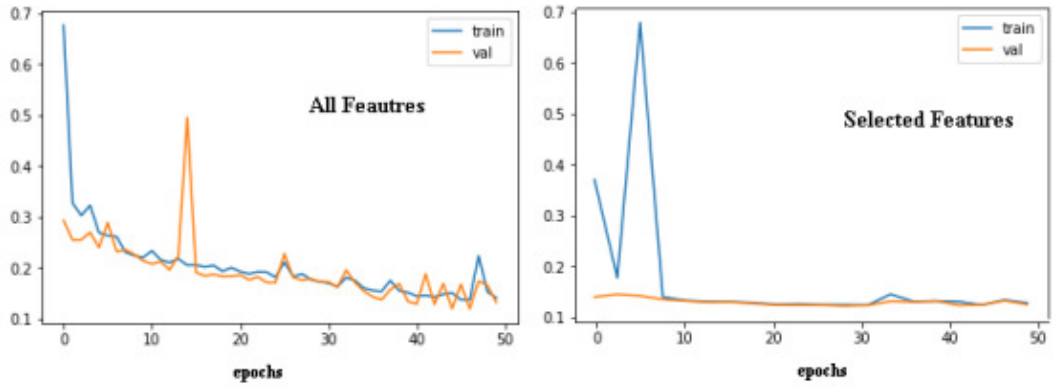



FIGURE 4.5: Évolution de la Perte à travers les Époques

D'un autre côté, le DL a également montré des performances remarquables. Comme illustré dans la Figure 4.5, nous obtenons une précision de 90% après seulement 40 époques pour l'ensemble de données utilisant toutes les caractéristiques, et la même précision est obtenue après seulement 10 époques pour l'ensemble de données utilisant les caractéristiques sélectionnées.

En conclusion, nos résultats démontrent que l'application de techniques de sélection des caractéristiques et l'utilisation de modèles de machine learning avancés peuvent améliorer de manière significative les performances de prédiction. Non seulement l'élimination des caractéristiques non pertinentes a permis d'améliorer les performances des modèles, mais elle a également révélé l'efficacité des modèles AG et DL par rapport aux méthodes classiques.

— **Effets de la réduction de la dimensionalité :Algorithmes non-supervisés**

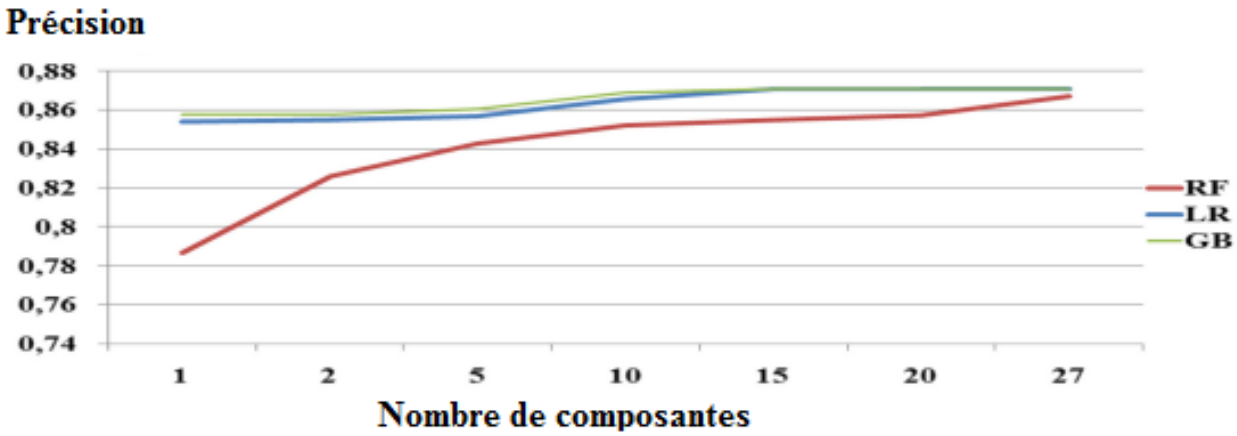



FIGURE 4.6: Précision par nombre de composantes

La Figure 4.6 illustre la variation de la précision en fonction du nombre de composants dans l'analyse PCA (Principal Component Analysis). Nous observons que pour les trois modèles utilisés, la réduction de la dimensionnalité à cinq composants néglige seulement 0,2 de la précision

globale des classificateurs. 90% de la variance se trouve dans les cinq premiers composants, ce qui signifie qu'il est possible de ne conserver que cinq caractéristiques sans perdre une part importante d'information dans les données. L'utilisation de 10 composants améliore légèrement la précision. En effet, pour le Gradient Boosting (GB) et la Régression Logistique (LR), à partir du 10ème composant, le score se stabilise respectivement à 0,871 et 0,87, ce qui est exactement le même score qu'en utilisant toutes les caractéristiques.

La visualisation des données nous permet de reconnaître davantage d'informations sur l'ensemble de données. L'une des raisons d'utiliser T-SNE est de valider la pertinence des caractéristiques extraites. En effet, si les caractéristiques utilisées sont pertinentes et très distinctives entre les classes de données, cette distinction doit être observée lors de la représentation des points avec T-SNE. Comme nous pouvons le voir à la Figure 4.7, T-SNE permet de distinguer deux clusters différents, associés au comportement des étudiants collecté avec les caractéristiques extraites de la trace de log ; les points rouges correspondent aux inscriptions avec un résultat de décrochage.

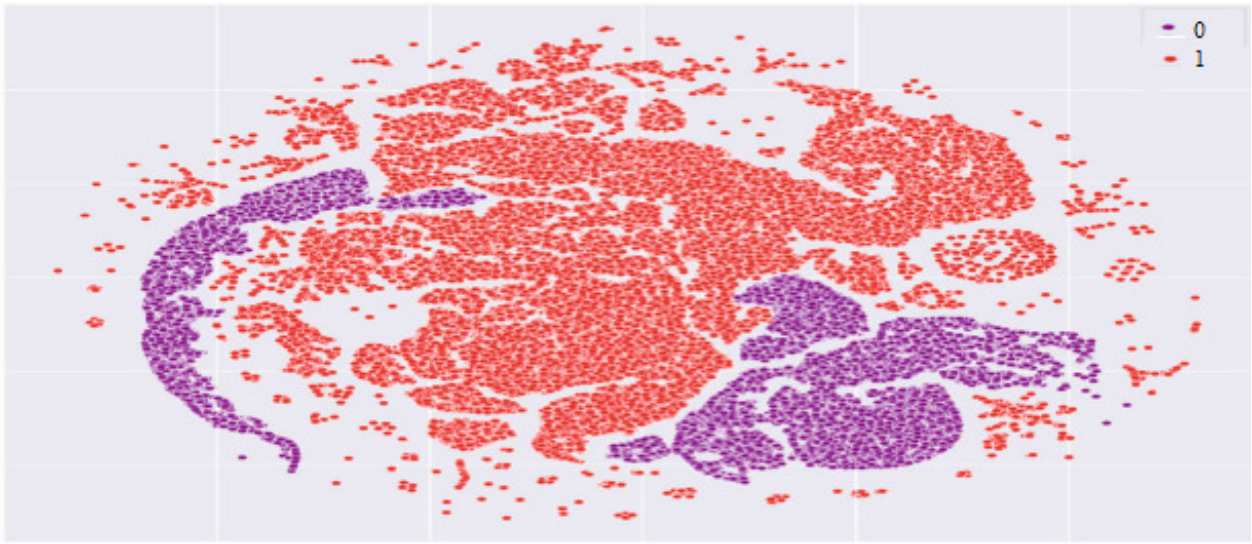



FIGURE 4.7: Visualisation des points en 2D

Les points violets correspondent à des réussites. Les points de décrochage, teints de la même couleur, se regroupent pour former un cluster et se séparent des points violets. Ce qui montre les similitudes et les distinctions dans l'ensemble de données. Cette découverte valide la pertinence des caractéristiques utilisées.

## 4.7 Personnalisation de l'intervention

Après avoir démontré dans la section précédentes la pertinence des résultats, des modèles et de l'ensemble des caractéristiques utilisées , notre attention se porte sur la dimension temporelle présente dans nos données. Pour analyser cet aspect, nous avons recours à la méthode ARIMA (AutoRegressive Integrated Moving Average) , une technique couramment utilisée dans l'analyse des séries temporelles.

ARIMA est un modèle statistique qui utilise des variations et des dépendances temporelles pour prédire des données futures. Il se compose de trois parties principales : le terme autorégressif (AR), le terme intégré (I) et le terme moyenne mobile (MA). Ces composantes permettent au modèle d'appréhender respectivement la relation entre une observation et un certain nombre de retards de cette observation, la différenciation des observations (suppression de la tendance ou de la saisonnalité pour rendre la série stationnaire), et la dépendance entre une observation et une erreur résiduelle à partir d'une moyenne mobile appliquée à des retards de cette observation.

Dans notre cas, nous allons utiliser ARIMA pour prédire les variables indépendantes pour les semaines futures de chaque inscription. Après avoir obtenu ces valeurs, elles seront utilisées comme une observation pour nos modèles afin de prédire la valeur de la variable cible Y. Ensuite, pour estimer la probabilité de chaque prédiction, nous utiliserons la fonction predict.proba() de la bibliothèque python Scikit-learn. Cette fonction permet de fournir une estimation probabiliste de la classe pour chaque instance, qui dans notre cas, sera utilisée pour déterminer le pourcentage hebdomadaire de décrochage. Cette approche nous permet d'obtenir une compréhension plus nuancée de notre prédiction en nous donnant non seulement la classe la plus probable, mais aussi la probabilité associée à cette classe.

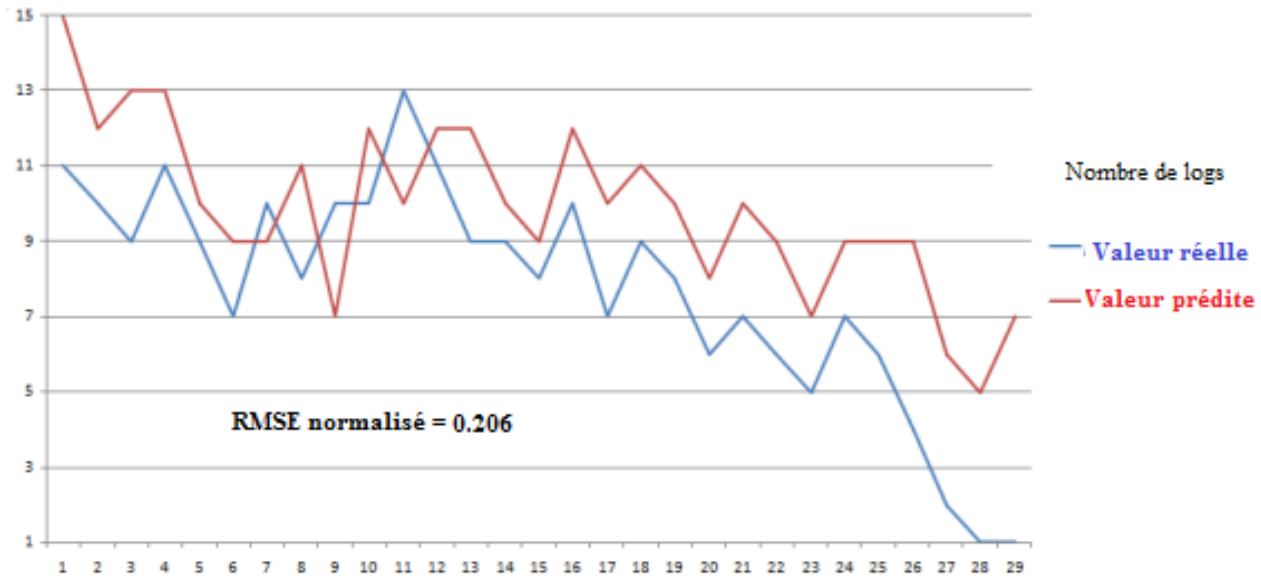



FIGURE 4.8: Erreur de prédiction sur Nombre de Logs avec ARIMA

La précision des prédictions du modèle ARIMA est évaluée à l'aide de la valeur normalisée du RMSE (Root Mean Squared Error). L'erreur entre la valeur réelle du nombre de logs journaliers et la valeur prédite est illustrée à la figure 4.8. La valeur du RMSE normalisé, adaptée à l'ensemble des caractéristiques, indique la précision du modèle ARIMA ainsi que la stationnarité des séries temporelles de ces caractéristiques.

Dans le cadre de notre étude, nous avons choisi d'utiliser le RMSE normalisé, signifiant que le RMSE a été divisé par l'écart type des valeurs observées. Cela fournit une mesure relative de l'erreur de prédiction, indépendante de l'échelle de la variable cible. En conséquence, la faible valeur du RMSE

normalisé trouvée pour la caractéristiques "nombre de logs" (0.206) suggère que notre modèle ARIMA offre une précision significative dans la prédiction des séries temporelles pour les caractéristiques de notre ensemble de données.

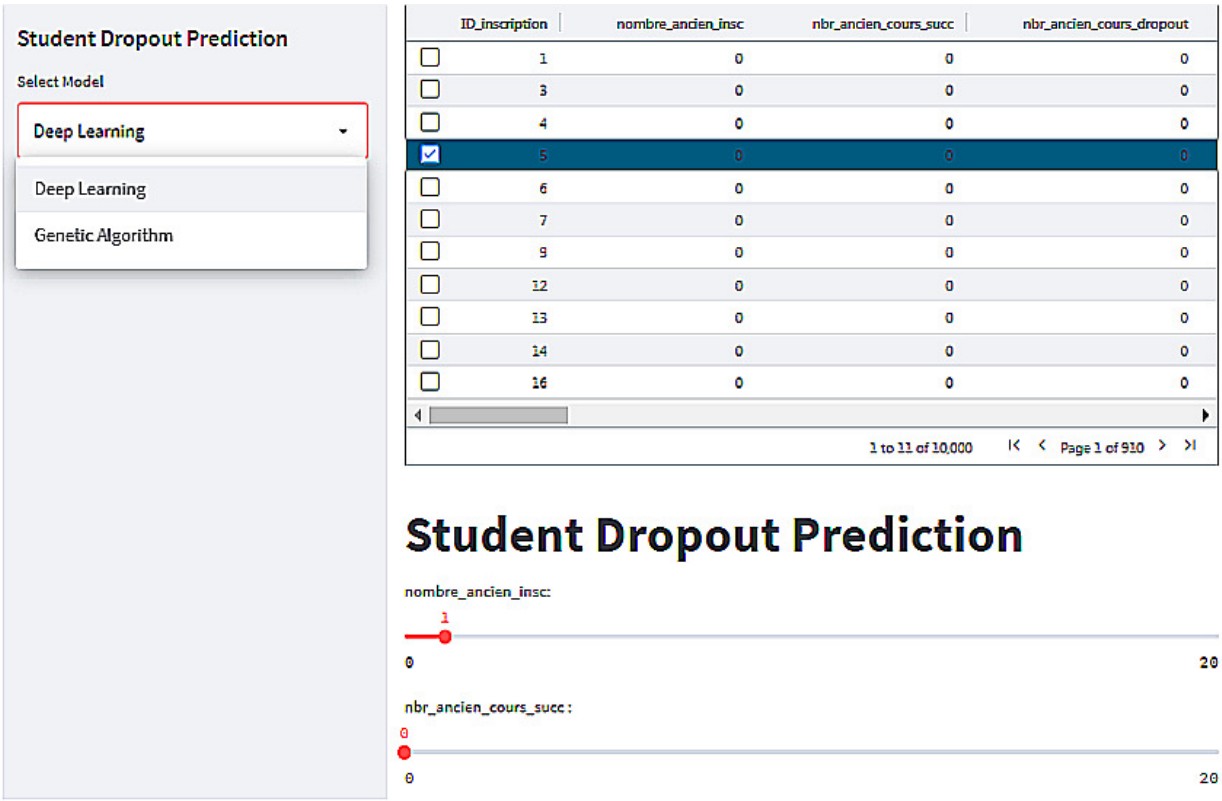



FIGURE 4.9: Interface de la page web

Un framework basé sur une page web unique a été développé en utilisant Python et le framework Streamlit. Ce système traite les inscriptions (étudiants, cours), récupère les logs correspondants, puis les agrège sur une base hebdomadaire. En outre, le modèle ARIMA est mis en œuvre pour anticiper les nouvelles observations pour les semaines à venir en se basant sur les données antérieures. Ce système donne aussi la possibilité de sélectionner le modèle à utiliser pour prédire le taux de décrochage chaque semaine, comme le démontrent les figures 4.9 et 4.10.

Pour identifier les profils d'étudiants et les modèles de décrochage, une analyse de corrélation a été menée entre les caractéristiques extraites et la performance finale de l'étudiant (réussite ou décrochage). Nous avons découvert que lorsque le modèle suivant est respecté : Accès - Vidéo - Devoirs - Discussion, l'exactitude de la rétention augmente notablement. Cela suggère que le comportement attendu de l'étudiant est d'accéder au lien du cours (événement Accès), de visualiser le contenu du cours via des vidéos, puis de se rendre sur la page des Devoirs. Enfin, l'étudiant doit discuter des points ambigus ou mal compris sur la page de discussions. Par ailleurs, le comportement de l'étudiant sur les pages de devoirs, le nombre d'événements de visualisation vidéo, de discussion et de page wiki enregistré chaque semaine, donne une indication de l'engagement de l'étudiant et de sa persévérance. Cela s'est avéré être un excellent prédicteur précoce du taux de décrochage et de la performance.

En utilisant le méthodes d'apprentissage non-supervisée K-means, nous avons identifié trois clus-

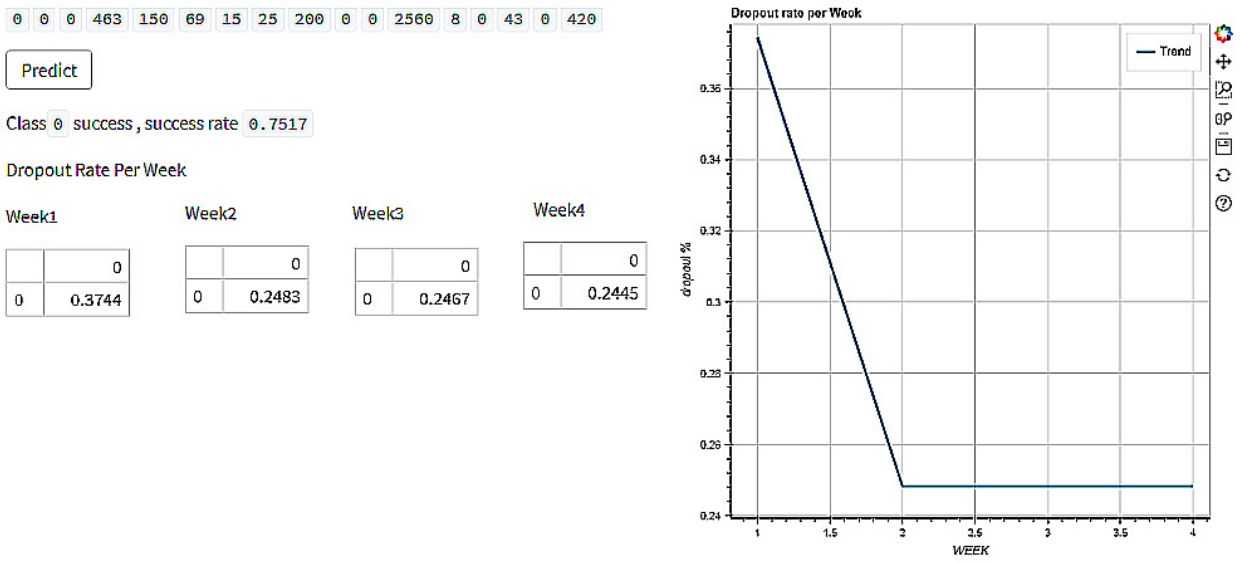



FIGURE 4.10: Pourcentage de décrochage par semaine

ters principaux pour les décrocheurs : Selon l'analyse des données des semaines 1 et 2, trois clusters prédominants ont été observés.

— **Cluster 1** : Ce groupe correspond aux étudiants qui n'ont pas terminé les vidéos du cours. Ce regroupement était un indicateur fort de l'abandon : 95% de ces étudiants n'ont pas terminé le cours.

— **Cluster 2** concerne les étudiants qui terminent le visionnage des vidéos mais qui ont peu de visites et de temps passé sur la page des devoirs. En d'autres termes, ces étudiants ont terminé la plupart du cours mais n'ont pas fait les quiz et les exercices. Ce groupe était également un indicateur fort d'abandon : 80% ont abandonné dans la quatrième semaine.

Les raisons qui expliquent le comportement des étudiants des clusters 1 et 2 peuvent être liées à plusieurs causes extraites des travaux connexes et de la revue de littérature présenté dans la première partie de cette thèse :

— **Causes liées à l'étudiant** : Le manque de motivation et d'engagement des étudiants est considéré comme l'un des facteurs les plus influents empêchant les étudiants de terminer un MOOC. Par exemple, Gutl [201] a interrogé 134 étudiants qui n'avaient pas terminé les cours de MOOC et a constaté que la majorité des étudiants avaient l'intention de terminer leurs études mais n'ont pas pu le faire en raison d'une faible motivation et d'un mauvais feedback. Les étudiants manquent de capacités/compétences et de prérequis pour suivre le cours. Selon des études précédentes [202], [203], [22] démontrant l'effet des compétences académiques des étudiants et de leurs expériences préalables sur le taux d'abandon dans les MOOC.

— **Causes liées au cours** : Elles peuvent aussi être liées à la longueur et à la difficulté du cours. Selon [204], [205], [206] la complexité ou la difficulté du contenu du cours a été trouvée dans de nombreuses études pour être associée au taux d'abandon des étudiants.

— **Causes liées aux instructeurs** : Le manque de supervision de l'instructeur : Le mauvais

feedback fourni par les instructeurs en raison du grand nombre d'étudiants inscrits par cours a été signalé comme un prédicteur significatif de l'abandon des étudiants dans les cours de MOOC [10] et [207].

— **Cluster 3** : concerne les étudiants qui ont passé un temps moyen sur les vidéos et la page des devoirs mais ont peu de visites sur la page de discussion. On peut interpréter que ces étudiants ne sont pas assez sociables pour communiquer avec d'autres étudiants ou poser des questions sur des points ambigus dans les cours ou les exercices. L'isolement et le manque d'interactivité dans les MOOC affectent directement l'abandon des étudiants. Une enquête [22] sur l'abandon des MOOC a montré dans les commentaires des abandons qu'ils se sentaient isolés et démotivés à continuer en raison du faible niveau d'interaction et de communication avec les étudiants et les instructeurs. Ils se sont plaints que l'instructeur ne les félicitait pas ou ne les motivait pas après les quiz. Ils ont également déclaré que les instructeurs n'engageaient pas les apprenants dans la discussion ou ne facilitaient pas le brainstorming.

Après avoir détecté les schémas de l'élève à haut risque de décrochage, la section suivante décrit des exemples d'interventions :

— **Interventions pour les clusters 1 et 2** : Pour les étudiants qui n'ont pas les prérequis nécessaires pour suivre le cours, il est judicieux de les détecter à travers un sondage ou des quiz au début de chaque cours. Ensuite, envoyez-leur des cours et des exercices contenant les prérequis qu'ils leur manquent pour suivre le cours. Selon la littérature, il existe une myriade d'interventions visant à augmenter la motivation et l'engagement des étudiants en créant un intérêt pour les sujets du cours [208]. Certaines interventions ont traité la démotivation par le biais d'un mécanisme d'email [209]. D'autres interventions tentent de ramener les étudiants absents dans le cours et de recueillir leurs raisons de départ [210].

Le manque de supervision de l'instructeur et le mauvais feedback de l'instructeur sont dus au nombre considérable d'inscriptions par cours dans les MOOCs. La solution ici est de se concentrer sur les étudiants signalés à haut risque de décrochage sur une semaine particulière selon le taux de décrochage donné par le système. En ce qui concerne le contenu du cours, le cours doit sembler utile pour les étudiants dans la vie réelle. Il devrait contenir beaucoup d'applications et d'exercices pratiques. Les compétences acquises dans le cours doivent s'appliquer à des problèmes du monde réel, à des objectifs de carrière spécifiques ou à des rôles ultérieurs dans la vie.

— **Interventions pour le cluster 3** : Plusieurs recherches ont trouvé que la connexion sociale à l'école est liée à des taux plus élevés de réussite académique des élèves [211], les enseignants et les pairs peuvent servir de sources pour faciliter cette connexion sociale. L'intervention proposée pour ce cluster est une réunion hebdomadaire de groupe de soutien par les pairs qui se concentre

sur l'amélioration des compétences académiques et interpersonnelles des élèves combinées à des interactions quotidiennes. Cela améliorera les résultats pour les étudiants signalés comme potentiellement décrocheurs. Nous pourrions former un groupe de soutien par les pairs de trois à quatre participants, et le 5ème est l'étudiant signalé comme décrocheur. Cette méthodologie est plus adaptée à l'apprentissage mixte ; elle a été testée à l'université, et les résultats obtenus ont amélioré le comportement en classe, augmenté l'engagement académique, et des interactions positives entre pairs et enseignants.

Le coût d'un abandon scolaire ou d'un abandon dans un MOOC est extrêmement élevé. Il se traduit par une perte considérable de temps, d'efforts et de ressources financières. Lorsqu'un étudiant décide d'abandonner, le lien avec lui est souvent perdu définitivement, et aucune action n'est généralement entreprise pour comprendre les raisons de son départ. Par conséquent, il est impératif pour les institutions, qu'elles soient traditionnelles ou dédiées à l'enseignement en ligne, de mettre en place ce système ou un système similaire pour anticiper et réduire le taux de décrochage.

Des schémas de décrochage supplémentaires peuvent également être identifiés. Des stratégies d'intervention préétablies peuvent être élaborées à partir de l'expérience d'enseignants experts ou de l'analyse de données historiques [20]. Cette étude représente simplement un premier pas vers une approche plus complète et efficace pour gérer le décrochage, que ce soit dans les établissements scolaires traditionnels ou dans le cadre des MOOCs.

Ces interventions, bien qu'elles soient diversifiées, ont toutes un objectif commun : créer un environnement d'apprentissage plus attrayant et propice pour les étudiants, leur permettant de rester engagés et motivés dans leurs parcours éducatifs. En fin de compte, notre but est d'améliorer le taux de rétention des étudiants, d'optimiser leur potentiel et de promouvoir une expérience éducative enrichissante et porteuse de sens, que ce soit dans une salle de classe traditionnelle ou dans l'espace virtuel d'un MOOC.

# Chapitre 5

# Déterminants Multidimensionnels de la Performance Académique

## 5.1 Introduction

Dans le contexte dynamique et compétitif du XXIe siècle, la performance académique est un indicateur crucial du futur statut socio-économique et de la qualité de vie. La performance académique, largement définie, englobe l'accomplissement des objectifs éducatifs dans divers domaines cognitifs et non cognitifs [212]. Le système éducatif marocain, riche en influences historiques, fait face à des défis contemporains uniques, tels que la réforme des programmes, les taux d'abandon scolaire et l'équilibre entre l'expansion des inscriptions et la qualité pédagogique (Fahim et al) [213].

Cette recherche enquête sur les déterminants multifactoriels de la performance académique dans une étude à grande échelle impliquant 5 092 étudiants dans diverses universités marocaines. Nous avons exploré les facteurs démographiques, académiques, psychologiques, sociaux et de mode de vie, fournissant une analyse complète des potentiels influenceurs du succès académique. Ces dimensions ont été identifiées comme des prédicteurs significatifs de la performance académique, mesurée par la moyenne générale des étudiants (GPA). Les résultats révèlent la nature complexe et interconnectée de ces facteurs et leur influence cumulative sur la trajectoire académique d'un étudiant.

Notre étude souligne la nécessité d'une approche holistique intégrant non seulement des stratégies académiques traditionnelles, mais aussi un soutien psychologique, des pratiques de vie équilibrées et un engagement social (Alhajraf et Alasfour, 2014 ; Al-Khani et al.) [214] [215]. Cette recherche explore ces influences multifactoriels dans le contexte de l'enseignement supérieur marocain, une perspective non occidentale où une telle recherche complète n'avait pas encore été entreprise.

La question centrale de notre recherche est la suivante : **Comment les facteurs psychologiques, sociaux et de mode de vie influencent-ils la performance académique des étudiants l'enseignement supérieur marocain ?**

En utilisant une méthodologie d'enquête robuste, cette étude recueille et analyse les données d'un échantillon diversifié d'étudiants dans les universités marocaines. À travers une régression logistique multivariée, nous identifions les principaux facteurs impactant significativement les résultats académiques.

Cette recherche dépasse le simple comblement d'une lacune dans la littérature en offrant une perspective culturellement nuancée sur le succès académique dans un contexte non occidental. Elle remet en question la dominance des récits éducatifs centrés sur l'occident et propose un modèle d'évaluation adapté aux dynamiques culturelles, sociales et éducatives uniques du Maroc. En outre, l'étude contribue au discours mondial sur l'éducation en fournissant des perspectives sur l'intégration des facteurs non cognitifs dans les politiques éducatives pour améliorer le succès des étudiants. Elle vise également à informer les stratégies institutionnelles qui favorisent le bien-être psychologique, l'engagement social et des choix de vie sains parmi les étudiants.

Les sections suivantes de ce chapitre commencent par une description de la méthodologie employée dans notre enquête, présentent les résultats et discutent des implications plus larges de ces résultats pour les politiques et pratiques éducatives au Maroc et dans des contextes similaires. Le chapitre se conclut par des recommandations concrètes pour les parties prenantes et propose des directions pour les recherches futures. En explorant le réseau complexe de facteurs influençant le succès académique au Maroc, cette recherche aspire à catalyser la réforme éducative et à contribuer à la conversation internationale sur l'excellence éducative.

## 5.2 Méthodologie

Notre étude comprend un vaste échantillon de 5 092 étudiants issus de diverses universités marocaines. Pour garantir une représentation adéquate de tous les domaines d'études, années d'études et milieux socio-économiques, nous avons utilisé une technique d'échantillonnage aléatoire stratifié, similaire à celle présentée par (Bhadrawja et al.) [216]. La stratification s'est basée sur les statistiques de chaque université pour refléter fidèlement la diversité de la population étudiante.

### 5.2.1 Collecte des Données

La collecte des données a été réalisée à l'aide d'un questionnaire auto-administré en ligne, semblable à l'enquête sur les comportements et expériences des adolescents menée par (Rico et al.) [217], diffusé via les listes de diffusion des étudiants et les plateformes de médias sociaux. Ce questionnaire, structuré en sections relatives aux informations démographiques, aux pratiques académiques, aux auto-évaluations psychologiques, aux influences sociales et aux habitudes de vie, comprenait des éléments à échelle de Likert, des questions à choix multiple et des questions ouvertes, conçues pour générer à la fois des données quantitatives et qualitatives.

Les données recueillies ont été soigneusement organisées dans un tableur, chaque colonne représentant une variable et chaque ligne reflétant les réponses individuelles des participants. Cette méthode d'organisation a facilité une analyse précise des données et minimisé les erreurs potentielles. Avant la

distribution à grande échelle, nous avons mené une étude pilote avec un sous-ensemble de notre population cible. Cela nous a permis d'identifier et de corriger tout problème potentiel avec le questionnaire, améliorant ainsi sa clarté, sa pertinence et réduisant les biais.

### 5.2.2 Analyse des Données

Une fois les données collectées, nous avons entamé la phase d'analyse des données à l'aide du puissant logiciel *Statistical Package for Social Sciences* (SPSS) version 28.0. Nous avons commencé notre enquête par des statistiques descriptives pour chaque variable, fournissant un aperçu des caractéristiques de l'échantillon, de la distribution des réponses pour chaque déterminant et du GPA.

Nous avons ensuite employé des analyses de régression logistique univariée et multivariée pour examiner les effets individuels et combinés des différents facteurs sur le GPA. L'analyse univariée visait à identifier les associations significatives entre chaque déterminant et le GPA. Cela a été suivi par une analyse multivariée pour contrôler les variables de confusion et explorer l'interaction complexe entre les différents déterminants. En cas de résultats contradictoires, nous avons vérifié les données et effectué des analyses de sensibilité pour valider la robustesse de nos résultats.

### 5.2.3 Engagement Éthique

Notre engagement envers des normes éthiques strictes a été inébranlable tout au long du processus de recherche, similaire à l'approche utilisée dans l'enquête sur les comportements et expériences des adolescents. La participation était volontaire et nous avons garanti l'anonymat des participants pendant la collecte des données. Nous avons clairement communiqué l'objectif de l'étude, promis la confidentialité et informé les participants de leur droit de se retirer de l'étude à tout moment. Outre le biais de non-réponse, nous avons reconnu et abordé les facteurs de confusion potentiels et autres biais grâce à la technique d'échantillonnage aléatoire stratifié, la gestion appropriée des données manquantes et le processus de validation de notre questionnaire.

### 5.2.4 Limites de l'Étude

Bien que notre étude soit robuste, elle présente des limitations. Une limitation significative est la dépendance aux mesures auto-rapportées, ce qui peut introduire un biais, comme le note [218]. De plus, étant donné que notre étude est transversale, elle limite notre capacité à inférer des relations causales à partir des relations observées. Les recherches futures pourraient se concentrer sur des études longitudinales et intégrer des mesures supplémentaires pour capturer une image plus complète des déterminants du succès académique.Voici le questionnaire utilisé pour la collecte des données dans cette étude :

## 5.3 Questionnaire

### Section 1 : Informations Démographiques

1. Âge : Veuillez spécifier votre âge.
2. Genre : Sélectionnez votre genre (Homme / Femme).
3. Année d'étude : Indiquez votre année d'étude actuelle ($1^{ère}$ Année / $2^{ème}$ Année / $3^{ème}$ Année / $4^{ème}$ Année / $5^{ème}$ Année).
4. Discipline : Spécifiez votre domaine d'étude ou discipline académique.

### Section 2 : Pratiques Académiques

5. Assiduité en classe : À quelle fréquence assistez-vous aux cours ? (Régulièrement / Occasionnellement / Rarement)
6. Heures d'étude quotidiennes : Combien d'heures consacrez-vous généralement aux études chaque jour ? (Moins d'une heure / 1-3 heures / Plus de 3 heures)
7. Participation à des groupes d'étude : À quelle fréquence participez-vous à des groupes d'étude ? (Régulièrement / Occasionnellement / Rarement)
8. Utilisation des ressources académiques : À quelle fréquence utilisez-vous des ressources académiques telles que les bibliothèques ou les matériaux en ligne ? (Régulièrement / Occasionnellement / Rarement)

### Section 3 : Auto-évaluations Psychologiques

9. Niveaux de stress : Comment évalueriez-vous votre niveau de stress actuel ? (Faible / Modéré / Élevé)
10. Niveau de motivation : Comment décririez-vous votre niveau de motivation envers vos études ? (Faible / Modéré / Élevé)
11. Anxiété académique : À quelle fréquence ressentez-vous de l'anxiété liée à vos études ? (Rarement / Occasionnellement / Fréquemment)

### Section 4 : Influences Sociales

12. Implication parentale : Dans quelle mesure vos parents ou tuteurs sont-ils impliqués dans votre vie académique ? (Faible / Modérée / Élevée)
13. Influence des pairs : Dans quelle mesure vous sentez-vous influencé par vos pairs dans vos décisions académiques ou habitudes d'étude ? (Faible / Modérée / Élevée)

14. Culture et climat scolaires : Comment décririez-vous la culture et le climat global de votre école ? (Faible / Modéré / Élevé)

### Section 5 : Habitudes de Vie

15. Heures de sommeil : En moyenne, combien d'heures de sommeil obtenez-vous par nuit ? (<6 heures / 6-8 heures / >8 heures)
16. Activité physique régulière : Pratiquez-vous une activité physique régulière ? (Oui / Non)

### Section 6 : Performance Académique

17. Moyenne Générale (GPA) : Veuillez sélectionner la plage qui correspond le mieux à votre GPA actuel.(Inférieur à 2.0 / 2.0 à 3.5 /Supérieur à 3.5 ) [1]

### Questions Ouvertes sur la Performance Académique

18. Pratiques académiques impactantes : Décrivez toute pratique académique spécifique ayant significativement impacté votre GPA.
19. Gestion du stress : Comment gérez-vous le stress lié à vos études ?
20. Influence des autres : Pouvez-vous partager une expérience où l'influence des pairs ou des parents a affecté votre performance académique ?
21. Changements de mode de vie : Quels changements de mode de vie avez-vous apportés qui ont positivement ou négativement affecté votre performance académique ?

## 5.4 Résultats

En avançant pour analyser les résultats de notre étude, nous commencerons par examiner deux ensembles distincts de variables et le degré de leur influence sur la performance académique des étudiants, avec une attention particulière sur leur moyenne générale (GPA). Ces variables sont classées en facteurs démographiques, académiques, psychologiques, sociaux et ceux liés au mode de vie. L'évaluation approfondie de ces ensembles de variables est décrite respectivement dans le tableau 5.1 et le tableau 5.2, et la discussion détaillée de ceux-ci sera présentée dans les sections suivanteSs.

Le tableau 5.1 inaugure notre section des résultats, en clarifiant l'impact des facteurs démographiques et académiques sur les GPAs des étudiants de manière détaillée.

1. Le questionnaire a été administré avec les notes suivantes : moins de 8, entre 8 et 12, et plus de 12, conformément au système de notation marocain. Ces résultats ont ensuite été traduits en GPA lors de l'analyse afin de les rendre l'étude plus universels.

L'analyse du tableau 5.1 offre des révélations intéressantes concernant la corrélation entre les facteurs démographiques et académiques et les résultats académiques des étudiants, mesurés par le GPA. Il est à noter que l'âge présente une corrélation considérable avec le GPA, les étudiants plus âgés affichant généralement des GPAs plus élevés. Cette observation, statistiquement significative (p-value $<$ 0.001), peut être attribuée à divers facteurs (Heretick et Tanguma, 2021) [219]. Au fur et à mesure que les étudiants progressent dans leur parcours académique, ils acquièrent une maturité personnelle et académique. La transition du lycée à l'université est souvent une phase turbulente, les étudiants passant d'un environnement structuré à un cadre d'apprentissage autonome. Au fil du temps, les étudiants développent généralement des habitudes d'étude efficaces et acquièrent une meilleure compréhension de leurs intérêts académiques, ce qui contribue à leur meilleure performance (Komasawa, Terasaki, Takitani, Lee, Kawata et Nakano, 2022) [220]. Les données mettent également en lumière l'influence significative du genre sur le GPA, les hommes semblant plus enclins à se retrouver dans la catégorie des GPAs inférieurs (p-value $<$ 0.001). Le rapport de cotes implique que les hommes ont environ 45% plus de chances de présenter des GPAs inférieurs par rapport aux femmes, probablement en raison d'un mélange de facteurs culturels et personnels, incluant la motivation et l'engagement. Les femmes sont souvent plus engagées dans leurs études, intrinsèquement motivées pour réussir, exceller dans la gestion du temps, et plus enclines à demander de l'aide en cas de besoin, ce qui conduit à une meilleure réussite académique (Power et al., 2020) [221].

TABLE 5.1: Facteurs démographiques et académiques associés aux résultats des étudiants

| Variables | Analyse univariée | | | Analyse multivariée | | |
|---|---|---|---|---|---|---|
| | Résultat | | | | | |
| | Faible (GPA $<$ 2.0) | Modéré (2.0 $\leq$ GPA $<$ 3.5) | Élevé (GPA $\geq$ 3.5) | Odds Ratio | Intervalle de confiance | P-value |
| **Démographiques** | | | | | | |
| Âge (moyenne +/- Déviation standard) | 19.7 +/- 1.5 | 20.8 +/- 1.7 | 22.5 +/- 2.3 | | | <0.001 |
| Sexe | | | | | | |
| Femmes | 1147 (35.5%) | 1107 (34.2%) | 977 (30.3%) | | | <0.001 |
| Hommes | 890 (44.9%) | 516 (26.1%) | 455 (23.0%) | 0.71 | [0.64 - 0.79] | <0.001 |
| **Académiques** | | | | | | |
| Année d'étude | | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 1ère et 2ème années | 1572 (77.1%) | 1005 (61.9%) | 730 (51.0%) | | | <0.001 |
| 3ème au 5ème années (Licence et Master) | 465 (22.8%) | 618 (38.0%) | 702 (49.0%) | 0.5 | [0.45 - 0.55] | |
| Assiduité en classe | | | | | | |
| Régulière | 465 (22.8%) | 776 (47.8%) | 876 (61.2%) | 1.6 | [1.4 - 1.9] | <0.001 |
| Occasionnelle | 857 (42.1%) | 626 (38.6%) | 402 (28.1%) | 0.9 | [0.8 - 1.1] | 0.2 |
| Rare | 715 (35.1%) | 221 (13.6%) | 154 (10.7%) | 0.5 | [0.4 - 0.7] | <0.001 |
| Heures d'étude quotidiennes | | | | | | |
| Moins d'une heure | 1095 (54.0%) | 451 (28.4%) | 88 (15.0%) | 0.4 | [0.3 - 0.5] | <0.001 |
| 1-3 heures | 760 (37.7%) | 852 (53.7%) | 387 (26.5%) | 1 | | <0.001 |
| Plus de 3 heures | 182 (9.0%) | 320 (20.2%) | 957 (65.7%) | 2.5 | [2.2 - 2.9] | <0.001 |
| Participation aux groupes d'étude | | | | | | |
| Régulière | 700 (34.3%) | 610 (37.6%) | 720 (50.3%) | 1.2 | [1.1 - 1.4] | <0.001 |
| Occasionnelle | 685 (33.6%) | 580 (35.7%) | 367 (25.6%) | 1 | | 0.02 |
| Rare | 652 (32.0%) | 433 (26.7%) | 345 (24.1%) | 0.9 | [0.8 - 1.0] | 0.07 |
| Utilisation des ressources académiques ( bibliothèque, matériels en ligne...) | | | | | | |
| Régulière | 480 (23.6%) | 700 (43.1%) | 875 (61.0%) | 1.8 | [1.6 - 2.0] | <0.001 |
| Occasionnelle | 875 (43.0%) | 598 (36.8%) | 395 (27.5%) | 1 | | <0.001 |
| Rare | 682 (33.5%) | 325 (20.0%) | 162 (11.3%) | 0.6 | [0.5 - 0.7] | <0.001 |

Concernant les variables académiques, il est évident que l'année d'étude, l'assiduité en classe, les heures d'étude quotidiennes, la participation aux groupes d'étude et l'utilisation des ressources académiques exercent divers degrés d'influence sur le GPA. Notamment, les étudiants en 3ème à 5ème année sont significativement plus susceptibles d'obtenir des GPAs plus élevés par rapport à ceux de leurs deux premières années, avec un odds ratio de 0.5. Cette observation est en accord avec la théorie de l'évolution académique et de la maturité de Tinto (Tinto, 1997) [222], qui postule qu'à mesure que les étudiants progressent dans leur parcours académique, ils développent des habitudes d'étude plus efficaces, deviennent plus engagés dans leur apprentissage et acquièrent une compréhension plus approfondie de leurs objectifs académiques et professionnels. Ces changements peuvent conduire à une amélioration de la performance académique, renforçant ainsi le pouvoir transformateur de l'enseignement supérieur. L'assiduité régulière en classe et le dévouement à étudier plus de trois heures par jour se manifestent également comme de forts prédicteurs de GPAs élevés, soulignant la valeur de la discipline

et de la persévérance dans le succès académique. En revanche, la participation à des groupes d'étude ne semble pas altérer significativement la performance académique, mettant ainsi en évidence les divers styles d'apprentissage parmi les étudiants. Enfin, l'utilisation fréquente de ressources académiques telles que les bibliothèques et les matériaux en ligne augmente considérablement la probabilité d'obtenir des GPAs plus élevés, soulignant le rôle instrumental de ces ressources dans la promotion de l'apprentissage et le soutien de la réussite académique (Anderson et García, 2020) [223].

En considérant ces observations collectivement, il est évident que les variables démographiques et académiques tissent un réseau complexe influençant les résultats académiques des étudiants. Les résultats accentuent la nécessité d'analyses multivariées plus complexes pour une compréhension approfondie et soulignent le rôle crucial des stratégies individuelles et institutionnelles dans la promotion du succès académique.

Après une revue approfondie des variables démographiques et académiques, nous concentrons notre attention sur un second ensemble de variables : les facteurs psychologiques, sociaux et ceux liés au mode de vie. Les détails statistiques de ces facteurs et leur influence sur les GPAs des étudiants sont présentés dans le tableau 5.2.

TABLE 5.2: Facteurs psychologiques, sociaux et de mode de vie associés aux résultats académiques des étudiants

| Variables | Analyse univariée | | | Analyse multivariée | | |
|---|---|---|---|---|---|---|
| | Résultat | | | | | |
| | Faible (GPA $< 2.0$) | Modéré ($2.0 \leq$ GPA $< 3.5$) | Élevé (GPA $\geq 3.5$) | Odds Ratio | Intervalle de confiance | P-value |
| **Psychologiques** | | | | | | |
| Niveaux de stress | | | | | | |
| Faible | 318 (9.8%) | 293 (18.0%) | 821 (57.2%) | 1 | [0.90, 1.12] | 0.05 |
| Modéré | 965 (29.9%) | 785 (48.4%) | 347 (24.2%) | 0.75 | [0.68, 0.83] | <0.001 |
| Élevé | 754 (54.4%) | 545 (33.6%) | 264 (18.4%) | 0.50 | [0.44, 0.57] | <0.001 |
| Motivation | | | | | | |
| Faible | 537 (26.3%) | 300 (18.5%) | 595 (41.5%) | 1.00 | [0.89, 1.13] | 0.06 |
| Modérée | 830 (40.8%) | 797 (49.1%) | 350 (24.4%) | 1.20 | [1.08, 1.34] | 0.001 |
| Élevée | 670 (32.9%) | 526 (32.4%) | 487 (34.0%) | 1.40 | [1.25, 1.57] | <0.001 |
| Anxiété académique | | | | | | |
| Faible | 485 (23.8%) | 384 (23.7%) | 563 (39.3%) | 1 | [0.89, 1.12] | 0.08 |
| Modérée | 887 (43.5%) | 539 (33.2%) | 400 (27.9%) | 0.95 | [0.85, 1.06] | 0.32 |
| Élevée | 665 (32.7%) | 700 (43.1%) | 469 (32.7%) | 0.85 | [0.71, 0.91] | 0.003 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Sociales** | | | | | | |
| Implication parentale | | | | | | |
| Faible | 550 (27.0%) | 665 (41.0%) | 419 (29.2%) | 1 | [0.90, 1.12] | 0.05 |
| Modérée | 760 (37.3%) | 600 (37.1%) | 401 (28.0%) | 0.75 | [0.68, 0.83] | <0.001 |
| Élevée | 727 (35.7%) | 358 (22.1%) | 612 (42.7%) | 0.50 | [0.44, 0.57] | <0.001 |
| Influence des pairs | | | | | | |
| Faible | 620 (30.4%) | 665 (41.0%) | 470 (32.8%) | 1 | [0.89, 1.13] | 0.06 |
| Modérée | 650 (31.9%) | 600 (37.1%) | 505 (35.2%) | 1.20 | [1.08, 1.34] | 0.001 |
| Élevée | 767 (37.7%) | 358 (22.1%) | 457 (31.9%) | 0.50 | [0.44, 0.57] | <0.001 |
| Culture et climat scolaires | | | | | | |
| Faible | 530 (26.0%) | 455 (28.1%) | 252 (17.5%) | 1.00 | [0.90, 1.12] | 0.05 |
| Modéré | 635 (31.2%) | 510 (31.5%) | 523 (36.5%) | 0.75 | [0.68, 0.83] | <0.001 |
| Élevé | 872 (42.8%) | 658 (40.6%) | 657 (45.8%) | 0.20 | [0.44, 0.57] | <0.001 |
| **Mode de vie** | | | | | | |
| Heures de sommeil | | | | | | |
| <6 heures | 350 (9.8%) | 275 (18.0%) | 520 (36.3%) | 0.4 | [0.3 - 0.5] | <0.001 |
| 6-8 heures | 980 (29.9%) | 725 (48.4%) | 870 (60.7%) | 0.75 | [1.1 - 1.4] | <0.001 |
| >8 heures | 707 (21.9%) | 623 (33.6%) | 237 (16.5%) | 0.50 | [0.68, 0.83] | <0.001 |
| Activité physique régulière | | | | | | |
| Oui | 502 (24.6%) | 578 (35.6%) | 707 (43.4%) | 1.00 | [2.2 - 2.9] | <0.001 |
| Non | 1535 (75.35%) | 1045 (64.3%) | 920 (56.5%) | 1.45 | [1.34 - 1.56] | <0.001 |

En procédant à l'analyse du tableau 5.2, nous nous concentrons sur les facteurs psychologiques, sociaux et ceux liés au mode de vie et leur impact sur le GPA des étudiants. En ce qui concerne les facteurs psychologiques, les données suggèrent que les niveaux de stress, la motivation et l'anxiété académique jouent effectivement un rôle considérable dans l'influence du GPA. Il est évident que les étudiants ayant de faibles niveaux de stress tendent à avoir un GPA plus élevé, avec une p-value significative inférieure à 0.001. Ce résultat est compatible avec la notion largement acceptée selon laquelle le stress chronique peut affecter négativement la fonction cognitive, conduisant à de mauvaises performances académiques (Stegers-Jager, Savas, van der Waal, van Rossum & Woltman, 2020) [224]. Cependant, il ne faut pas sous-estimer les effets bénéfiques potentiels du stress léger et à court terme, qui peuvent motiver les étudiants et améliorer leurs performances (Pascoe, Hetrick & Parker, 2020) [225].

Les niveaux de motivation affectent également de manière significative les résultats académiques.

Une motivation plus élevée correspond à un GPA plus élevé, avec un fort odds ratio de 1.40. Cela s'aligne avec la littérature existante confirmant le rôle crucial de la motivation dans le succès académique (Stark, 2019) [226]. Étonnamment, l'anxiété académique ne semble pas avoir un impact négatif similaire sur le GPA comme le stress. Une anxiété académique modérée et élevée est associée à des GPAs plus élevés. Cela peut être dû au fait qu'un certain degré d'anxiété peut améliorer la concentration et l'effort, améliorant ainsi les performances académiques (Chang, 2021) [227].

Dans le domaine social, l'implication parentale, l'influence des pairs et la culture et le climat scolaires ont un impact notable sur le GPA des étudiants. Les étudiants ayant une forte implication parentale et une culture scolaire élevée ont des GPAs plus élevés. Cela corrobore les études précédentes soulignant l'impact positif des environnements de soutien sur les performances académiques des étudiants (Anthony & Ogg, 2019) [228]. L'influence des pairs, cependant, est à double tranchant, pouvant soit motiver soit distraire les étudiants. Ainsi, le rôle des pairs dans la réussite académique peut nécessiter une approche plus nuancée (Qureshi et Khaskheli 2023) [229]. En ce qui concerne les facteurs liés au mode de vie, la qualité du sommeil et la participation à des activités physiques régulières sont toutes deux significativement corrélées avec le GPA. Un sommeil adéquat (6-8 heures) et une activité physique régulière conduisent à de meilleurs résultats académiques. Ces résultats font écho aux études précédentes qui ont trouvé une corrélation positive entre la qualité du sommeil, l'activité physique et les performances académiques (Wang & Boros, 2021) [230].

En résumé, le tableau 5.2 démontre que les facteurs psychologiques, sociaux et liés au mode de vie influencent de manière significative les performances académiques. Ces résultats suggèrent l'importance des approches holistiques qui prennent en compte ces facteurs diversifiés pour améliorer le succès des étudiants. Par conséquent, des stratégies et des interventions doivent être développées non seulement pour améliorer les compétences académiques des étudiants, mais aussi pour cultiver leur bien-être psychologique, leur engagement social et leurs habitudes de vie saines. Des recherches supplémentaires sont nécessaires pour déterminer comment ces facteurs interagissent et les meilleures façons de les aborder pour maximiser les résultats académiques.

Sur la base de l'analyse de régression logistique multivariée, plusieurs variables ont été associées à un faible GPA (GPA $< 2.0$) chez les étudiants. En termes de facteurs démographiques, être un homme était associé à une probabilité plus faible d'avoir un faible GPA par rapport aux femmes (OR = 0.71, IC à 95% : 0.64-0.79, $P < 0.001$). Cependant, chaque augmentation d'une unité de l'âge était associée à une probabilité plus élevée d'avoir un faible GPA (OR = 1.1, IC à 95% : 1.1-1.2, $P < 0.001$).

Dans le domaine académique, l'année d'étude jouait un rôle significatif. Les étudiants en 3ème à 5ème années (Licence et Master) avaient une probabilité plus faible d'avoir un faible GPA par rapport aux 1ère et 2ème années (Licences) (OR = 0.5, IC à 95% : 0.45-0.55, $P < 0.001$). L'assiduité en classe était un autre facteur important, avec une assiduité régulière associée à une probabilité plus élevée

d'avoir un GPA élevé (OR = 1.6, IC à 95% : 1.4-1.9, P < 0.001).

En ce qui concerne les habitudes d'étude, les étudiants qui étudiaient plus de 3 heures par jour avaient une probabilité plus élevée d'avoir un GPA élevé (OR = 2.5, IC à 95% : 2.2-2.9, P < 0.001). La participation régulière à des groupes d'étude et l'utilisation des ressources académiques, telles que les bibliothèques et les matériaux en ligne, étaient également associées à une probabilité plus élevée d'avoir un GPA élevé (OR = 1.2, IC à 95% : 1.1-1.4, P < 0.001 ; OR = 1.8, IC à 95% : 1.6-2.0, P < 0.001, respectivement).

Les facteurs psychologiques montraient des associations significatives avec le GPA. Des niveaux de stress élevés étaient associés à une probabilité plus faible d'avoir un GPA élevé (OR = 0.50, IC à 95% : 0.44-0.57, P < 0.001). En revanche, les étudiants avec une forte motivation avaient une probabilité plus élevée d'avoir un GPA élevé (OR = 1.40, IC à 95% : 1.25-1.57, P < 0.001). De plus, une forte anxiété académique était liée à une probabilité plus faible d'avoir un GPA élevé (OR = 0.85, IC à 95% : 0.71-0.91, P = 0.003).

Parmi les facteurs sociaux, une forte implication parentale et une influence des pairs élevée étaient associées à une probabilité plus faible d'avoir un GPA élevé (OR = 0.50, IC à 95% : 0.44-0.57, P < 0.001). Cependant, les étudiants dans des écoles avec une culture et un climat élevés avaient une probabilité plus élevée d'avoir un GPA élevé (OR = 0.20, IC à 95% : 0.44-0.57, P < 0.001). En ce qui concerne les facteurs liés au mode de vie, les étudiants qui dormaient moins de 6 heures avaient une probabilité plus faible d'avoir un GPA élevé (OR = 0.4, IC à 95% : 0.3-0.5, P < 0.001). De plus, les étudiants qui ne pratiquaient pas d'activité physique régulière avaient une probabilité plus élevée d'avoir un faible GPA (OR = 1.45, IC à 95% : 1.34-1.56, P < 0.001).

Cette analyse de régression logistique multivariée met en évidence les associations entre divers facteurs et les niveaux de GPA chez les étudiants. Des facteurs tels que le genre, l'âge, l'année d'étude, l'assiduité en classe, les heures d'étude, la participation à des groupes d'étude, l'utilisation des ressources académiques, les niveaux de stress psychologique, la motivation, l'anxiété académique, l'implication parentale, l'influence des pairs, la culture et le climat scolaires, les heures de sommeil et l'activité physique régulière jouent tous un rôle dans la détermination des résultats de GPA des étudiants.

Ayant présenté les principaux résultats de notre analyse, nous nous tournons maintenant vers une discussion de leurs implications pour la compréhension des déterminants multifactoriels de la performance académique et la formulation de stratégies éducatives efficaces. De plus, nous explorerons et expliquerons certaines contradictions apparentes observées dans les données, en mettant en lumière l'interaction complexe entre divers facteurs influençant le succès académique.

## 5.5 Discussion

L'analyse multifactorielle des déterminants influençant la performance académique des étudiants, mesurée par la moyenne générale (GPA), nécessite une approche holistique de la part des éducateurs et des décideurs politiques pour répondre aux besoins diversifiés des étudiants.

L'âge et le sexe ont été identifiés comme des prédicteurs significatifs de la réussite académique. Une augmentation de l'âge est corrélée avec des GPA plus élevés, suggérant que la maturité peut améliorer les compétences académiques. Cette constatation est en harmonie avec les recherches de Poropat (2009), qui a identifié la conscience, un trait souvent associé à la maturité, comme un prédicteur de la performance académique [231]. Pour répondre aux besoins variés liés à l'âge, les institutions pourraient bénéficier de programmes d'orientation spécialisés pour les étudiants plus âgés, afin d'intégrer leurs expériences de vie dans leur parcours académique.

La disparité de performance entre les sexes, où les femmes dépassent généralement les hommes, reflète les résultats de Richardson, Abraham et Bond (2012), qui ont mis en évidence le rôle de l'auto-efficacité dans la réussite académique, souvent plus élevée chez les étudiantes [232]. Reconnaissant cela, des programmes de motivation et d'engagement ciblés pour les étudiants masculins pourraient être essentiels pour réduire l'écart de genre dans les performances académiques.

Les comportements académiques tels que la présence en classe et les habitudes d'étude sont cruciaux pour les résultats en GPA, résonnant avec les insights de Poropat (2009) [231] concernant l'influence des traits de personnalité comme la conscience sur la réussite académique. L'intégration d'outils de collaboration numérique pourrait rendre le processus académique plus engageant et répondre aux routines d'étude disciplinées essentielles à la réussite académique.

Les facteurs psychologiques, y compris le stress et la motivation, ont un impact significatif sur le GPA. Nos observations s'alignent avec celles de Khanna, Singh et Alam (2018) [233], qui ont également souligné le pouvoir prédictif des facteurs psychologiques sur la performance académique. L'étude étend cette compréhension en considérant un spectre plus large de facteurs psychologiques, y compris l'anxiété académique. Les institutions pourraient mettre en œuvre des initiatives de bien-être axées sur la gestion du stress et l'amélioration de la motivation pour soutenir le bien-être mental des étudiants.

Les éléments sociétaux tels que l'implication des parents et l'influence des pairs sont également cruciaux, comme le soutiennent Li, Allen et Casillas (2017) [234], qui ont trouvé que l'implication familiale prédit le GPA. Notre étude ajoute à cela en suggérant que des ateliers pour les parents et la création de sociétés étudiantes pourraient renforcer l'engagement parental et favoriser un environnement d'apprentissage positif.

Les facteurs de mode de vie tels que le sommeil et l'activité physique ont été positivement liés à la performance académique, renforçant l'importance de modes de vie équilibrés pour les étudiants,

comme le suggèrent des études précédentes. Plaider pour des programmes de bien-être qui promeuvent l'exercice régulier et des habitudes de sommeil saines pourrait soutenir le bien-être général et la réussite académique des étudiants.

Dans le monde complexe de la performance académique, des contradictions émergent souvent, révélant la nature nuancée et multifactorielle des facteurs influençant le succès des étudiants. En analysant ces contradictions, il est crucial de considérer non seulement les divergences apparentes mais aussi les limitations inhérentes à nos méthodologies d'étude. Cette approche permet de découvrir des insights plus profonds sur les dynamiques en jeu, illustrant comment des variables telles que l'âge, l'anxiété académique, les heures de sommeil et l'activité physique interagissent de manière inattendue pour façonner les résultats éducatifs.

Une contradiction intrigante concerne la relation entre l'âge et le GPA. Bien que les étudiants plus âgés soient généralement associés à des GPA plus élevés, suggérant que la maturité et les habitudes d'étude développées contribuent positivement à la performance académique, une augmentation de l'âge est également corrélée à des chances plus élevées d'obtenir un GPA plus faible. Ce paradoxe peut être compris en examinant le contexte plus large dans lequel ces étudiants évoluent. Les étudiants plus âgés font souvent face à des défis de vie plus complexes, tels que la conciliation des engagements académiques avec le travail ou les responsabilités familiales. Cette complexité supplémentaire peut, paradoxalement, avoir un impact négatif sur leur performance académique, malgré leur maturité académique. Une telle contradiction souligne la nécessité d'une approche holistique dans l'analyse éducative, mettant en avant l'importance des analyses multivariées pour capturer l'interaction entre l'âge, les circonstances de vie et le succès académique.

De même, l'analyse de l'anxiété académique présente une découverte contre-intuitive. Des niveaux modérés à élevés d'anxiété académique sont associés à des GPA plus élevés, une découverte qui contredit la sagesse conventionnelle selon laquelle l'anxiété entrave la performance. Cependant, à travers le prisme de la loi de Yerkes-Dodson [235], cette contradiction devient moins perplexe. La loi suggère un niveau optimal d'éveil pour une performance de pointe, indiquant qu'un certain degré d'anxiété académique pourrait en fait motiver les étudiants à s'engager plus profondément dans leurs études, améliorant ainsi leurs résultats. Cet insight met en lumière la relation complexe entre le stress et la performance, suggérant que l'impact de l'anxiété sur la réussite académique n'est pas simple mais varie selon son intensité et la capacité de l'individu à la canaliser de manière productive.

Les résultats sur les heures de sommeil illustrent également l'équilibre délicat nécessaire pour la réussite académique. Bien que 6-8 heures de sommeil par nuit soient associées à des GPA plus élevés, dépasser cette plage optimale semble diminuer l'association, défiant la recommandation générale pour un sommeil ample. Ce paradoxe peut être résolu en reconnaissant la loi des rendements décroissants, en particulier dans le contexte de l'impact du sommeil sur la fonction cognitive et la performance

académique. Des durées de sommeil excessivement longues peuvent refléter des problèmes de santé sous-jacents ou entraîner une diminution du temps consacré aux études, suggérant une plage de sommeil optimale qui soutient la restauration cognitive sans compromettre les responsabilités académiques.

Enfin, la corrélation positive entre l'activité physique régulière et les GPA plus élevés renforce les avantages de l'exercice pour la fonction cognitive et la santé mentale. Pourtant, cette association peut sembler contradictoire en considérant le potentiel d'une activité physique excessive pour nuire au temps d'étude. Cette contradiction est résolue en reconnaissant l'équilibre que les étudiants réussis trouvent entre l'activité physique et les obligations académiques, bénéficiant de la concentration accrue et de la réduction du stress que l'exercice procure sans affecter négativement leur temps d'étude.

En abordant ces contradictions et leurs limitations, nous obtenons des insights précieux sur l'équilibre délicat nécessaire pour favoriser la réussite académique. Que ce soit par la gestion de l'anxiété, l'optimisation du sommeil ou l'équilibre entre l'activité physique et les obligations académiques, l'essence de la réussite académique réside dans la recherche d'un équilibre harmonieux qui soutient à la fois le bien-être et les objectifs éducatifs. Ces réflexions nous rappellent la nature complexe et souvent contre-intuitive des facteurs influençant le succès des étudiants, exhortant une analyse complète et nuancée pour comprendre pleinement et améliorer les résultats académiques.

Bien que notre étude fournisse des insights significatifs, elle partage des limitations communes à la recherche, telles que le potentiel de biais de réponse dans les données auto-déclarées et le défi de généraliser les résultats à partir d'un contexte géographique spécifique. Les recherches futures pourraient aborder ces limitations en incorporant des études longitudinales et des contextes éducatifs plus larges pour valider et étendre nos découvertes.

En conclusion, cette étude non seulement corrobore mais également développe la littérature existante, offrant une perspective nuancée sur les dynamiques de la performance académique. Elle souligne l'importance d'une approche interdisciplinaire pour créer des systèmes de soutien complets qui répondent aux besoins académiques, psychologiques et de mode de vie des étudiants, ouvrant la voie à des recherches futures pour améliorer davantage les résultats éducatifs.

## 5.6 Conclusion

Ce chapitre explore les déterminants multidimensionnels de la performance académique parmi un nombre important d'étudiants issus de diverses universités marocaines. En examinant les facteurs démographiques, académiques, psychologiques, sociétaux et liés au mode de vie, nous avons acquis des informations cruciales sur l'interaction complexe de ces variables et leur impact collectif sur la réussite académique des étudiants, mesurée par la moyenne générale (GPA).

Nos résultats soulignent l'importance de divers facteurs influençant la performance académique. L'âge, le sexe, l'année d'études, la présence en classe, les habitudes d'étude, les niveaux de stress, la

motivation, l'anxiété académique, l'implication des parents, l'influence des pairs, la culture scolaire, la durée de sommeil et l'activité physique se sont révélés être des prédicteurs significatifs de la GPA. Ces résultats mettent en évidence la nécessité d'une approche holistique de l'éducation, intégrant le soutien psychologique, l'équilibre de vie et l'engagement sociétal au sein des stratégies académiques.

Les éducateurs et les décideurs politiques peuvent utiliser ces résultats pour développer des interventions ciblées et des systèmes de soutien qui répondent aux déterminants identifiés. Des initiatives adaptées pour différents groupes d'âge et genres peuvent promouvoir la croissance individuelle et la réussite académique. Cultiver une culture de présence régulière en classe, promouvoir des habitudes d'étude efficaces et fournir des ressources académiques stratégiques peuvent optimiser la performance des étudiants. De plus, reconnaître et atténuer les niveaux de stress, encourager la motivation et instiller la résilience peuvent améliorer le bien-être général et les résultats académiques.

Bien que cette étude fournisse des insights précieux, il est essentiel de reconnaître ses limitations. Le design transversal empêche d'établir des relations causales, et les mesures auto-déclarées peuvent introduire des biais de réponse. En outre, l'étude s'est concentrée spécifiquement sur les étudiants au Maroc, limitant la généralisation des résultats. Les recherches futures devraient utiliser des designs longitudinaux ou expérimentaux, des mesures objectives et explorer des facteurs supplémentaires potentiellement influents pour renforcer les connaissances existantes.

En conclusion, cette étude contribue à une compréhension globale des déterminants de la performance académique parmi les étudiants au Maroc. En identifiant l'interaction complexe entre les facteurs démographiques, académiques, psychologiques, sociétaux et liés au mode de vie, nous soulignons l'importance d'adopter une approche holistique en éducation. Les insights tirés de cette recherche peuvent guider les éducateurs, les décideurs politiques et les chercheurs dans l'amélioration des résultats académiques des étudiants et la promotion de la réussite globale des étudiants.

# Chapitre 6

# Effets de la gamification sur la motivation et l'engagement des étudiants

## 6.1 Introduction

Les Cours en Ligne Ouverts à Tous (MOOCs) représentent un formidable outil d'élargissement des opportunités éducatives. Cependant, ils sont confrontés à un défi majeur : un taux d'abandon élevé. Avec un taux de réussite moyen de 7%, selon une étude récente de l'Association Européenne des Universités d'Enseignement à Distance (EADTU) [236]. Beaucoup d'apprenants qui s'inscrivent à un MOOC ne le terminent pas, perdant ainsi les avantages envisagés. Cette problématique concerne non seulement les MOOCs mais aussi d'autres formes d'apprentissage en ligne.

En réponse à ces défis, ce chapitre permet d'explorer le potentiel de la gamification pour motiver les étudiants et réduire le taux d'abandon dans les MOOCs. L'incorporation d'éléments de jeu dans des contextes non ludiques, c'est-à-dire la gamification, a démontré son potentiel pour augmenter l'engagement et les taux d'achèvement dans l'éducation. En utilisant des éléments tels que des points, des badges, des classements et des récompenses, la gamification vise à rendre l'apprentissage plus engageant et motivant. L'idée centrale derrière la gamification est que rendre l'apprentissage plus ludique motivera davantage les apprenants à terminer le cours et à atteindre leurs objectifs. Cependant, malgré son utilisation généralisée, l'efficacité de la gamification doit encore être pleinement démontrée par des recherches empiriques. De plus, on manque de compréhension des effets à court, moyen et long terme de la gamification sur la motivation, l'engagement et les performances des étudiants dans les MOOCs. Le succès des stratégies de gamification dépend du public visé et du contexte spécifique d'application.

De plus, on manque de compréhension des éléments de jeu spécifiques qui fonctionnent le mieux dans différents contextes de MOOC et comment ils impactent le comportement des étudiants. Il est donc nécessaire d'effectuer davantage de recherches pour synthétiser et analyser les résultats existants afin d'approfondir notre compréhension de ce domaine. Ce chapitre vise à combler cette lacune en proposant une analyse complète des aspects du jeu les plus couramment utilisés dans les plateformes d'apprentissage en ligne, en particulier les MOOCs, et leurs effets confirmés sur le comportement humain.

Ce chapitre vise à fournir une compréhension globale de l'état actuel de la recherche sur la gamification dans les MOOCs en abordant les questions de recherche suivantes :

1. **Quels sont les aspects de la gamification les plus couramment utilisés dans les MOOCs ?**
2. **Comment la gamification affecte-t-elle la motivation et l'engagement des étudiants dans les MOOCs ?**
3. **Comment la gamification affecte-t-elle le taux d'abandon des étudiants dans les MOOCs ?**

Pour répondre à ces questions, ce chapitre suit les directives PRISMA 2020 pour les revues systématiques et les méta-analyses. Il est structuré en trois sections : Méthodologie, Résultats, et Discussion.

## 6.2 Méthodologie

### 6.2.1 Conception de l'étude

Ce chapitre examine les effets des techniques de gamification sur la motivation, l'engagement et la rétention des étudiants dans les MOOCs. Une approche méthodologique mixte est utilisée pour évaluer les articles quantitatifs et qualitatifs.

### 6.2.2 Collecte de données

La littérature, incluant la littérature grise, a été examinée via des bases de données comme Google Scholar, Scopus, IEEE-Xplore, APA PsychInfo, ProQuest, et ERIC. Des critères d'inclusion/exclusion prédéfinis ont été utilisés..

### 6.2.3 Analyse des données

Les études ont été classées selon les éléments de gamification, la population cible et l'environnement d'apprentissage. Ensuite, une analyse thématique a identifié les thèmes et tendances communs. Une analyse comparative a été réalisée pour évaluer la cohérence des résultats entre les études similaires. Enfin, les résultats ont été synthétisés pour conclure sur l'impact de la gamification sur la motivation, l'engagement et le taux d'abandon dans les MOOCs et les environnements d'apprentissage mixte.

### 6.2.4 Évaluation de la qualité

Les études ont été évaluées pour leur fiabilité et validité en utilisant des critères clarifiés du MMAT. Les données ont été extraites indépendamment, avec résolution des désaccords par consensus ou un

troisième auteur.

## 6.3 Résultats

ce chapitre a été réalisée pour combler les lacunes prédéfinies liées à l'étude et au rôle de la gamification sur la motivation et l'engagement des étudiants dans les MOOC et l'apprentissage hybride. En se basant sur la stratégie de recherche, ce chapitre a principalement présenté deux catégories d'impact de la gamification sur le comportement et les caractéristiques des étudiants : leur motivation et leur engagement.

Afin de mettre en évidence les différentes approches et méthodes utilisées dans les études examinées, le Tableau 6.1 présente une liste des études, leur approche, l'environnement en ligne et la taille de l'échantillon.

Table 6.1: Approches et méthodes utilisées dans les études extraites

| Étude | Approche | Environnement en ligne | Échantillon |
|---|---|---|---|
| Kyewski & Krämer (2018) [237] | Étude expérimentale | Moodle | 126 étudiants |
| Hakulinen, Auvinen & Korhonen (2015) [238] | Étude expérimentale | Apprentissage hybride | 281 étudiants |
| Gooch et al. (2016) [239] | Tests pré et post | ClassDojo | 22 étudiants |
| Saputra & Risqi (2015) [240] | Qualitative | Apprentissage hybride | 40 étudiants |
| Hanus & Fox (2015) [241] | Méthodes mixtes | Apprentissage hybride | 76 étudiants |
| Utomo & Santoso (2015) [242] | Quantitatif | MOOC | 31 étudiants |
| Huang & Hew (2015) [243] | Quasi-expérimental | MOOC | 40 étudiants |
| Kuo & Chuang (2016) [244] | Quantitatif | MOOC | 73 étudiants |
| Sitra et al. (2017) [245] | Étude de cas | MOOC | 5 étudiants |
| Tenorio et al. (2016) [246] | Expérimental | MeuTutor | 32 étudiants |
| Krause et al. (2015) [247] | Expérimental | MOOC | 172 étudiants |
| De-Marcos et al. (2014) [248] | Expérimental | MOOC | 72 étudiants |
| Mazarakis (2015) [249] | Expérimental | MOOC | 436 étudiants |
| Aldemir, Celik & Kaplan (2018) [250] | Qualitatif | MOOC | 118 étudiants |

En examinant les approches et méthodes utilisées dans les études incluses dans cette étude, le Tableau 6.1 offre un résumé clair. Les études utilisent une gamme d'approches telles que des conceptions expérimentales, quasi-expérimentales, empiriques et des études de cas pour étudier l'impact de la gamification sur la motivation et l'engagement des étudiants dans les MOOC et les environnements d'apprentissage hybride. Les environnements en ligne abordés dans ces études comprennent les MOOC, Moodle, MeuTutor et ClassDojo. Les tailles d'échantillon varient de 5 à 281 participants, offrant aux lecteurs un aperçu complet des différentes conceptions de recherche dans ce domaine.

Le tableau 6.2 offre un aperçu concis des principaux éléments de la gamification qui ont été étudiés

Table 6.2: Éléments de gamification et leur impact sur les caractéristiques des étudiants

| Éléments | Motivation | Engagement | Abandon |
|---|---|---|---|
| Récompenses | Kyewski & Krämer (2018) [237] | Utomo & Santoso (2015) [242] | Krause et al. (2015) [247] |
| Feedback | Hakulinen, Auvinen & Korhonen (2015) [238] | Huang & Hew (2015) [243] | De-Marcos et al. (2014) [248] |
| Points/scores | Gooch et al. (2016) [239] | Kuo & Chuang (2016) [244] | Mazarakis (2015) [249] |
| Défis | Saputra & Risqi (2015) [240] | | Aldemir, Celik & Kaplan (2018) [250] |
| J'aime/Je n'aime pas | Hanus & Fox (2015) [241] | Sitra et al. (2017) [245] | |

et de leur impact sur les caractéristiques des étudiants, résumant les éléments de gamification et leurs effets sur la motivation des étudiants, leur engagement et leur taux d'abandon.

Une étude plus approfondie des impacts de la gamification sur la motivation et l'engagement des étudiants est présentée dans le Tableau 6.3, qui détaille les caractéristiques des étudiants affectés par la gamification.

Table 6.3: Caractéristiques des étudiants affectés par la gamification

| Groupe | Impacts |
|---|---|
| Motivation des étudiants | Motivation intrinsèque et extrinsèque |
| Engagement des étudiants | Rétention, engagement et plaisir |

Le Tableau 6.3 présente les effets de la gamification sur la motivation et l'engagement des étudiants. Les impacts sur la motivation des étudiants couvrent à la fois la motivation intrinsèque et extrinsèque. En ce qui concerne l'engagement des étudiants, les effets comprennent la rétention, l'engagement et la satisfaction. En comprenant ces impacts, les éducateurs et les concepteurs de MOOC peuvent mieux mettre en œuvre les éléments de gamification dans leurs cours pour améliorer la motivation et l'engagement des étudiants.

### 6.3.1 Motivation

Les études discutées fournissent des résultats mitigés sur l'impact des récompenses sur la motivation des étudiants dans les MOOCs et les environnements d'apprentissage gamifiés.

Kyewski et Krämer [237] ont constaté que les badges n'affectaient pas la motivation intrinsèque des étudiants. En revanche, Hakulinen et al. [238] ont rapporté que les récompenses motivaient les étudiants universitaires. Gooch et al. [239] ont démontré les effets positifs des badges sur les étudiants ayant des troubles d'apprentissage. Hanus et Fox [241] ont observé une diminution de la motivation et de l'engagement dans les cours gamifiés sur le long terme.

Utomo et Santoso [242] ont montré que le feedback personnalisé en temps réel augmentait la motivation des étudiants pour un comportement d'apprentissage positif. Huang et Hew [243] ont trouvé que les classements et les récompenses motivaient la majorité des étudiants, mais n'avaient pas d'effet sur le groupe témoin. Kuo et Chuang [244] ont rapporté une augmentation de l'engagement et de la motivation avec leur plateforme gamifiée. Dominguez et al. [98] ont souligné que la gamification des systèmes d'e-learning pouvait accroître l'intérêt des étudiants, mais nécessitait une conception soignée.

En conclusion, les récompenses peuvent être efficaces pour motiver les étudiants dans les MOOCs, mais leur succès dépend de leur conception et de leur mise en œuvre. Les besoins et les motivations des étudiants doivent être pris en compte pour maximiser l'efficacité des systèmes de récompense.

### 6.3.2 Engagement

Kyewski et Krämer [237] ont constaté que les badges n'augmentaient pas l'engagement des étudiants et que ceux sans badges étaient plus engagés. Ces résultats contredisent Sitra et al. [245], qui ont trouvé un effet positif des badges sur l'engagement dans un système LMS-Moodle similaire.

Tenorio et al. [246] ont montré que des éléments de jeu comme les étiquettes, les récits, les scores et les classements augmentaient la motivation et la participation des étudiants dans un environnement d'évaluation par les pairs. Les résultats suggèrent que la gamification stimule l'engagement des étudiants sur la plateforme.

Krause et al. [247] ont découvert que les conditions de jeu social, qui incluaient la compétition entre étudiants, renforçaient l'engagement et la rétention par rapport aux conditions de jeu normales.

DeMarcos et al. [248] ont examiné les effets des réseaux sociaux et de la gamification sur la réussite scolaire et l'engagement des étudiants. Bien que les étudiants aient eu des perceptions positives de la gamification et des réseaux sociaux, aucune différence significative n'a été observée dans les résultats d'apprentissage ou l'engagement.

Mazarakis [249] a étudié l'impact de différents types de rétroaction sur l'engagement des étudiants, trouvant que la rétroaction augmentait l'engagement dans le wiki du cours. Aldemir et al. [250] ont

constaté que les étudiants avaient des attitudes positives envers les éléments de jeu tels que les classements, les défis et les récompenses, tant dans les cours en ligne qu'en présentiel.

En résumé, certaines études, comme celles de Sitra et al. et Tenorio et al., ont trouvé des effets positifs des éléments de jeu sur l'engagement et la motivation des étudiants. D'autres, comme celles de Kyewski et Krämer, et DeMarcos et al., ont trouvé des résultats moins concluants ou contradictoires. L'intégration efficace des éléments de jeu semble essentielle pour améliorer la motivation et l'engagement des étudiants.

## 6.4 Implications pratiques

Après avoir examiné les résultats de cette étude, les implications pratiques pour les éducateurs et les concepteurs de MOOCs sont les suivantes :

1. **Adapter les éléments de gamification au public cible :** Concevoir des cours en tenant compte des besoins et préférences spécifiques des étudiants pour améliorer leur expérience d'apprentissage.
2. **Équilibrer la motivation intrinsèque et extrinsèque :** Trouver un équilibre entre ces deux types de motivation pour maximiser l'engagement et la performance des étudiants.
3. **Suivre et ajuster la gamification :** Évaluer régulièrement l'efficacité des éléments de gamification et apporter des ajustements en fonction des retours des étudiants.
4. **Combiner avec d'autres stratégies pédagogiques :** Intégrer la gamification avec d'autres méthodes d'enseignement pour une expérience d'apprentissage complète.
5. **Fournir des directives et un soutien clairs :** Informer les étudiants sur les éléments de gamification et offrir un soutien pour les aider à les utiliser efficacement.

Ces recommandations permettent d'exploiter le potentiel de la gamification pour créer des expériences d'apprentissage en ligne stimulantes et efficaces, améliorant ainsi les performances des étudiants et réduisant les taux d'abandon.

## 6.5 Conclusion

En conclusion, ce chapitre a exploré l'état actuel des recherches sur la gamification dans les MOOC et son impact sur la motivation, l'engagement et le taux d'abandon des étudiants. Les résultats montrent que des éléments de gamification tels que les badges, les niveaux, les classements, les défis et les récompenses ont un impact positif sur la motivation et l'engagement des étudiants, et peuvent réduire le taux d'abandon. Cependant, l'efficacité de la gamification varie selon les éléments utilisés et le contexte. Des recherches supplémentaires sont nécessaires pour comprendre les effets à long terme

et identifier les meilleures pratiques de gamification dans les MOOC. Les concepteurs de MOOC et les éducateurs devraient envisager d'intégrer des éléments de gamification pour améliorer l'expérience d'apprentissage.

# Chapitre 7

# Théories de la motivation et les stratégies pour la rétention

## 7.1 Introduction

Le problème du taux d'abandon élevé a été un point focal pour les concepteurs de MOOCs depuis 2016 [251]. De nombreuses recherches ont visé à comprendre les causes de ces taux d'abandon, ainsi qu'à identifier des stratégies potentielles pour améliorer la rétention des apprenants [252]. Reconnaissant l'importance de la motivation, les chercheurs explorent les facteurs qui contribuent à ces défis afin d'affiner la conception des MOOCs [253]. Malgré de nombreuses tentatives pour élaborer des stratégies afin de réduire les taux d'abandon, la littérature met en évidence la motivation comme un déterminant clé de la rétention des apprenants [254], [255], [256], [257], [258].

Une compréhension plus profonde de la motivation pourrait aider à aborder à la fois les facteurs centrés sur l'apprenant et ceux liés aux MOOCs, potentiellement en réduisant les taux d'abandon et en enrichissant l'expérience des apprenants. Bien que la motivation joue un rôle significatif dans l'achèvement des MOOCs, les facteurs spécifiques qui poussent les apprenants à persévérer ne sont pas encore entièrement compris [259]. Il existe un besoin pour un examen plus approfondi du rôle de la motivation dans les taux d'abandon des MOOCs, car la plupart des études actuelles ont employé des méthodes quantitatives [260], [261].

Cette étude cherche à construire sur cette base de connaissances en fournissant une nouvelle perspective sur le rôle de la motivation dans les MOOCs. Notre objectif principal n'est pas seulement d'identifier les facteurs motivationnels, mais aussi de comprendre comment ces facteurs peuvent être efficacement appliqués dans la conception et la livraison des cours MOOCs pour optimiser l'engagement et le succès des apprenants. Pour atteindre cet objectif, nous avons affiné nos questions de recherche en trois questions centrales basées sur les critères PICOC (Population, Intervention, Comparaison, Résultat, Contexte) :

1. **Q1 : Quels sont les facteurs fondamentaux qui influencent la motivation des apprenants dans le contexte des MOOCs ?**
2. **Q2 : Quels cadres théoriques sous-tendent les stratégies pratiques pour améliorer**

**la motivation et la rétention dans les MOOCs ? Comment ces cadres théoriques peuvent-ils guider une mise en œuvre efficace de ces stratégies ?**

3. **Q3 : Quelles sont les limites ou les défis associés à la mise en œuvre de stratégies pour améliorer la motivation et la rétention dans les MOOCs ?**

Notre étude vise à fournir des informations précieuses pour ceux impliqués dans la conception et l'instruction des MOOCs, ainsi que pour les chercheurs, afin d'améliorer l'efficacité des MOOCs et de réduire les taux d'abandon. Comprendre les facteurs contribuant à la rétention des apprenants est essentiel pour créer des expériences MOOCs plus efficaces. En exploitant ces facteurs, les fournisseurs de MOOCs peuvent mieux aider les apprenants à atteindre leurs objectifs éducatifs et à améliorer leur satisfaction globale vis-à-vis de l'apprentissage en ligne. De plus, l'application des théories motivationnelles dans le contexte des MOOCs peut fournir une base théorique solide pour comprendre le comportement des apprenants et les tendances à l'abandon. Cette compréhension peut orienter le développement d'interventions ciblées et de stratégies pour améliorer les taux de rétention et l'efficacité globale des MOOCs.

Notre étude vise à contribuer à la discussion en cours sur les taux d'abandon des MOOCs et le rôle de la motivation dans la détermination des résultats des apprenants, en fournissant des orientations pratiques pour les acteurs des MOOCs afin d'affiner leurs plateformes et de mieux répondre aux besoins divers des apprenants. Pour répondre à ces questions, ce chapitre respectera les directives du "Preferred Reporting Items for Systematic Reviews and Meta-Analyses" (PRISMA) 2020 [262].

## 7.2 Méthodologie

### 7.2.1 Conception de l'étude

Ce chapitre explore le rôle de la motivation dans la rétention des apprenants dans les MOOCs, en utilisant une approche de méthodes mixtes pour évaluer des études quantitatives et qualitatives. La revue de littérature, incluant la littérature grise, a été réalisée via Google Scholar, Scopus, Web of Science, ERIC, PsycINFO et ProQuest.

### 7.2.2 Stratégie de recherche

La recherche a filtré les résultats par date de publication (1er janvier 2013 - 31 mars 2023) en utilisant des critères PICOC et des mots-clés spécifiques tels que "motivat*", "MOOC", "dropout", "participat*", "retention", "motivation intrinsèque", "motivation extrinsèque" et "théories". La première phase a identifié 750 articles, réduits à 120 après examen des résumés. La deuxième phase a identifié 453 articles, réduits à 70. La troisième phase a identifié 1041 articles, réduits à 372. Finalement, 495

articles ont été analysés en profondeur, laissant 140 articles pour l'étude finale.

### 7.2.3 Critères d'inclusion et d'exclusion

#### 7.2.3.1 Critères d'inclusion

— Études en anglais ou en français, publiées entre le 1er janvier 2013 et le 31 mars 2023.
— Publications évaluées par des pairs, incluant la littérature grise.
— Études axées sur la motivation dans les MOOCs, d'une longueur minimale de huit pages.

#### 7.2.3.2 Critères d'exclusion

— Études non rédigées en anglais ou en français.
— Blogs, articles de magazines, dissertations, newsletters.
— Études non évaluées par des pairs, sauf pour la littérature grise.

### 7.2.4 Évaluation de la qualité

Deux chercheurs ont initialement examiné les articles pour éliminer les duplications et les articles non pertinents. Les critères d'évaluation incluaient la pertinence du titre et du résumé, et les résultats explicites. Le coefficient de kappa de 76% indiquait un accord considérable entre les codeurs. Les désaccords ont été résolus par conciliation avec l'intervention d'un troisième auteur si nécessaire.

### 7.2.5 Analyse des données

— **Catégorisation :** Les études ont été classées par facteurs motivationnels, population cible, environnement MOOC et engagement.
— **Analyse thématique :** Identification des thèmes communs à travers les études.
— **Analyse comparative :** Juxtaposition des résultats pour comprendre la cohérence et identifier les contradictions.
— **Synthèse et interprétation :** Intégration des conclusions pour tirer des leçons sur la motivation et l'engagement dans les MOOCs, et identification des limitations et des domaines de recherche future.

Les résultats montrent 140 recherches académiques, incluant 89 comptes rendus de conférences et 51 articles de revues, soulignant l'importance du sujet pour la communauté académique.

Comme le montre la figure 7.1, le nombre d'articles publiés sur l'abandon des MOOCs a considérablement augmenté en 2016, indiquant un intérêt croissant des chercheurs pour ce sujet. Cependant, la diminution du nombre d'articles publiés dans les années suivantes ne signifie pas nécessairement une baisse d'importance du sujet. Certains articles n'ont peut-être pas été inclus dans l'examen systématique en

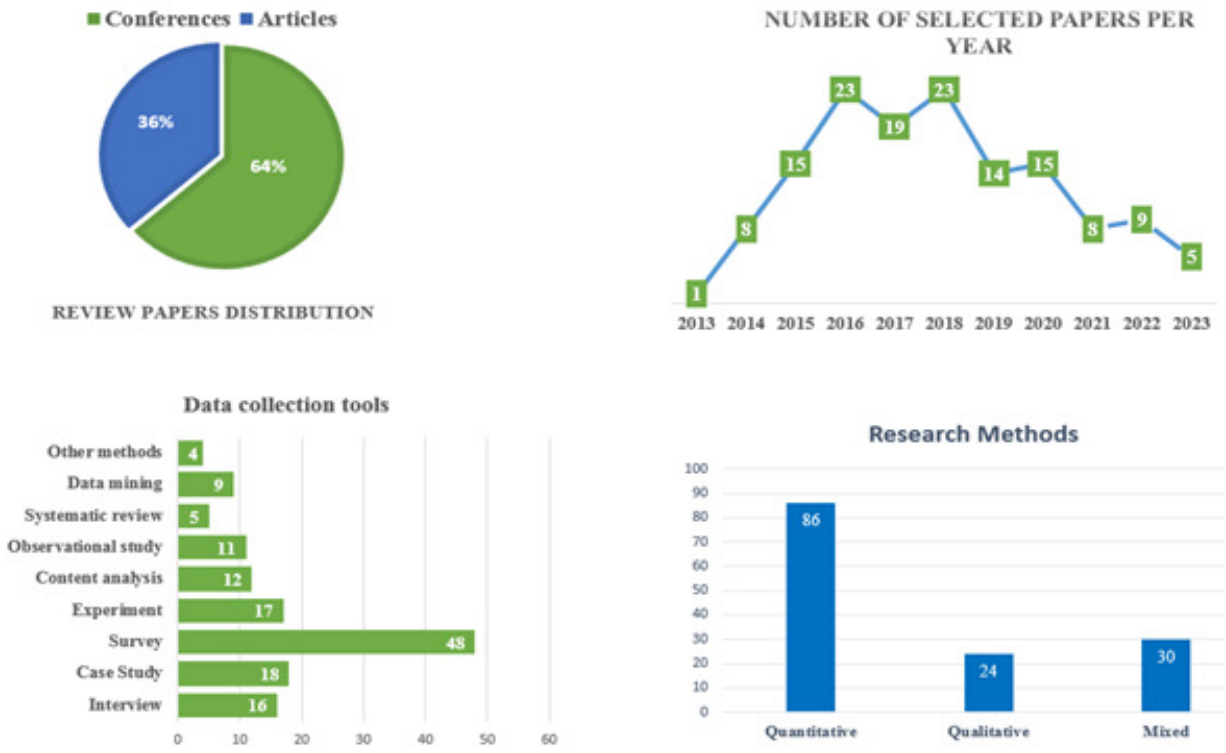



FIGURE 7.1: Statistique des articles examinés

raison de la redondance des idées déjà étudiées, ou cette diminution pourrait refléter une saturation de la recherche plutôt qu'un désintérêt.

L'analyse des données révèle que la majorité des articles utilisent une méthode quantitative, avec une prédilection pour les enquêtes structurées. Les méthodes qualitatives, comprenant des études de cas et des interviews, offrent des enquêtes approfondies. Les méthodes mixtes (30 articles) combinent les avantages des approches quantitatives et qualitatives pour fournir des aperçus complets. Cette diversité de méthodologies indique une variété de conceptions et d'objectifs de recherche au sein des articles examinés.

La section des résultats présente les principales conclusions des articles analysés, synthétisant les méthodologies quantitatives, qualitatives et mixtes pour fournir une compréhension complète. Les résultats sont organisés et discutés selon les objectifs de recherche, offrant une présentation claire des résultats significatifs et de leurs implications. Cette section sert de base pour les discussions et conclusions ultérieures, où les résultats sont interprétés et des recommandations pour des recherches futures sont fournies.

## 7.3 Résultats

**Q1 : Quels sont les facteurs fondamentaux et les cadres théoriques qui influencent la motivation des apprenants dans le contexte des MOOCs ?**

### 7.3.1 Facteurs de motivation

Pour analyser les facteurs influençant la réussite des MOOCs, il faut examiner les motivations des apprenants, divisées en deux types : besoins et intérêts. Les motivations basées sur les besoins incluent des facteurs académiques, spécifiques aux cours, et professionnels. Celles basées sur l'intérêt comprennent des facteurs personnels, sociaux, et technologiques : voir tableau 7.1. La section suivante analysera ces motivations et leur impact sur la réussite des MOOCs.

TABLE 7.1: Aperçu des catégories et types de motivations influençant la réussite des MOOCs

| Catégorie de motivation | Type de Motivation | Références |
|---|---|---|
| Motivations basées sur les besoins | Académique | Jung and Lee (2018) [263]<br>Shao et al (2018) [264]<br>Wang and Baker (2018)[261]<br>Semenova et al (2022) [265]<br>Nleme Ze & Molinari (2022) [266] |
| | Liée aux Cours | James et al (2019)[267]<br>Reparaz et al. (2019)[20]<br>Aldowah et al. (2018) [252]<br>Kuo et al. (2021) [268] |
| | Professionnelle | Doo et al. (2019) [269]<br>Shao & Chen (2021) [270]<br>Cobos & Ruiz-Garcia (2021) [271]<br>Meekers et al. (2022) [272] |
| Motivations basées sur l'intérêt | Personnelle | Buhr et al. (2019) [273]<br>Borrella et al. (2022) [274]<br>Schettino & Capone (2022) [275]<br>Yu et al. (2022) [276] |
| | Sociale | Gregori et al. (2019) [277]<br>Khan et al. (2019) [278]<br>Romero-Frías et al. (2023) [279]<br>Huang et al. (2023) [280] |
| | Liée à la technologie | Joo et al. (2019) [281]<br>Vezne et al. (2023) [282]<br>Cheng et al (2023) [283] |

La motivation basée sur les besoins repose sur l'idée que les apprenants s'engagent dans des activités d'apprentissage pour répondre à leurs besoins éducatifs spécifiques. Selon les théories basées

sur les besoins, cette motivation est influencée par des facteurs externes et surtout internes, tels que les incitations académiques, les objectifs de cours, et les aspirations professionnelles. Ces motivations internes encouragent les apprenants à reconnaître et à satisfaire leurs besoins éducatifs, ce qui stimule leur parcours d'apprentissage.

- **Motivation académique :** La motivation académique, qui caractérise l'enthousiasme des étudiants pour leurs études, joue un rôle crucial dans leur réussite scolaire. Les étudiants ayant une motivation académique plus élevée obtiennent souvent de meilleurs résultats. Cette motivation est influencée par des facteurs tels que l'auto-efficacité académique (Jung & Lee [263], Shao [264], Wang & Baker [23]) et l'utilisation d'instruments pédagogiques innovants (Chang et al. [284], Carrera & Ramírez-Hernández [285], Ortega-Arranz et al [286]).
- **Motivation de cours :** La motivation de cours dépend de la structure, de la conception et du contenu du cours. La complexité et la difficulté du matériel (James [267]), le calendrier du cours (Kizilcec et Halawa [255]) et l'efficacité perçue du contenu (Charo et al. [287], Alario-Hoyos et al. [288]) influencent les taux d'achèvement des MOOCs. La conception du cours est également cruciale (Aldowah et al. [252], El Said [204]).
- **Motivation professionnelle :** La motivation professionnelle découle du désir d'améliorer les perspectives de carrière et les compétences liées à l'emploi. Les MOOCs permettent d'acquérir de nouvelles connaissances et compétences (Milligan & Littlejohn [289]), répondant aux besoins en développement professionnel (Doo et al. [269], Xiong et al. [256]). La pertinence du cours par rapport à l'emploi actuel est également importante (Lu et al. 2017).
- **Motivation personnelle :** Les motivations personnelles incluent la curiosité, la croissance personnelle, le plaisir, la valeur perçue et les expériences antérieures (Bonk & Lee [290], Aldowah et al. [252], Watted & Barak [291]). L'intérêt personnel et les raisons sociales influencent aussi les décisions des apprenants (Petronzi & Hadi [292], Buhr et al. [273], Shao [264]).
- **Motivation sociale :** Les motivations sociales incluent le soutien social, la présence d'amis, les opportunités de réseautage et la possibilité de rencontrer de nouvelles personnes (Aldowah et al. [252], Gregori et al [277], Uchidiuno et al [293]). Les interactions sociales de qualité et la réputation perçue d'un cours contribuent aux taux d'achèvement (Khan et al [278]).
- **Motivation liée à la technologie :** La motivation technologique est influencée par la convivialité de la plateforme, les fonctionnalités innovantes et l'accessibilité (Joo et al [281], Jung & Lee [263]). Des aspects tels que la richesse des médias et l'interactivité ajoutent un attrait dynamique aux MOOCs (Zhao et al [294]). Malgré des défis comme les limites infrastructurelles (Shapiro et al [202]), les avantages des MOOCs, y compris l'accessibilité étendue et la flexibilité de l'apprentissage, sont souvent prédominants (Eriksson et al [205]).

Comprendre les interactions entre les motivations académiques, personnelles, sociales et de cours

est essentiel pour améliorer la complétion des MOOCs. Ces motivations s'influencent mutuellement, renforçant l'engagement des apprenants (Veletsianos et al., 2020). Cette compréhension permet de concevoir des MOOCs plus engageants. Les recherches futures devraient approfondir ces interactions pour identifier les facteurs clés de réussite.

**Q2 : Quels cadres théoriques sous-tendent les stratégies pratiques pour améliorer la motivation et la rétention dans les MOOCs ? Comment ces cadres théoriques peuvent-ils guider une mise en œuvre efficace de ces stratégies ?**

### 7.3.2 Stratégies motivationnelles et cadres théoriques

Cette section examine les stratégies pour renforcer la motivation et la rétention dans les MOOCs, basées sur diverses théories clés. Les recherches identifient cinq dimensions principales pour la motivation à terminer les MOOCs : le contrôle de l'apprenant, les aspects pédagogiques, technologiques, sociaux, et l'engagement comportemental (Barak, 2016 ; Khan, 2018 ; Sun, 2019 ; Maya, 2020 ; Badali, 2022). Ces théories soulignent l'importance de l'implication active de l'apprenant, des méthodes pédagogiques, des technologies, des interactions sociales, et de l'engagement comportemental pour la réussite des MOOCs.

| **Dimension** | **Théories** | **Description** |
|---|---|---|
| Contrôle de l'apprenant | SDT, PRT, SRLT, CVTAE | Théories mettant l'accent sur la capacité de l'apprenant à suivre des cours de MOOC de manière autodirigée, considérant l'apprenant comme un agent actif (Piccoli et al., 2001). |
| Pédagogique | EVT, AGT, ECT, CLT, ALT, ELT, UDL | Théories axées sur les aspects pédagogiques, d'enseignement et d'apprentissage des MOOCs, y compris la création d'environnements d'apprentissage constructifs et les attentes et réalisations dans les MOOCs (Kesim & Altınpulluk, 2015). |
| Technologique | TAM, FT, TTAT, MRT, TNE, CHAT | Théories concernant l'incorporation et l'utilisation de technologies et de médias innovants dans les MOOCs, et leurs avantages potentiels pour l'apprentissage à grande échelle (Khan et al., 2018 ; Wu & Chen, 2017). |

| | | |
|---|---|---|
| Sociale | CT, SCT, ST, TRQ | Théories axées sur les échanges sociaux, les processus d'apprentissage collaboratif et les actions au sein des MOOCs, soulignant l'importance du contexte social et de l'engagement dans l'apprentissage (Sun et al., 2019 ; Barak et al., 2016). |
| Comportementale | TPB, TRA | Théories appliquées pour prévoir le comportement individuel sur la base de leurs attitudes et intentions comportementales existantes. Ces théories élucident la corrélation entre les attitudes et les comportements dans le domaine de l'activité humaine (Maya-Jariego et al., 2020). |
| Autres | MT, ITI, HTFT, RIDT | Ces théories couvrent divers aspects, de la motivation au comportement et au-delà, qui ne s'inscrivent pas nettement dans les cinq catégories précédentes. |

TABLE 7.2: Catégorisation des Cadres Théoriques dans la Recherche sur les MOOCs

Le tableau 7.2 catégorise les théories dominantes utilisées dans l'étude des MOOCs en six groupes : Contrôle de l'apprenant, Pédagogique, Technologique, Sociale, Engagement/Comportementale et Autres. Les théories du Contrôle de l'apprenant, comme SDT, PRT, SRLT et CVTAE, mettent l'accent sur la capacité de l'apprenant à s'engager dans les MOOCs de manière autonome, les considérant comme des agents actifs [295]. Les théories pédagogiques, y compris EVT, AGT, ECT, CLT, ALT, ELT et UDL, se concentrent sur les aspects d'enseignement et d'apprentissage des MOOCs, y compris la création d'environnements d'apprentissage constructifs et la gestion des attentes et des réalisations. Les théories technologiques, comme TAM, FT, TTAT, MRT et TNE, traitent de l'adoption et de l'application de nouvelles technologies et de médias dans les MOOCs, et comment ceux-ci peuvent faciliter l'apprentissage à grande échelle [278]. Les théories sociales, telles que CT, SCT, ST et TRQ, examinent les échanges sociaux, les processus d'apprentissage collaboratif et les activités dans les MOOCs, soulignant l'importance du contexte social et des interactions dans l'apprentissage [254], [296]. Les théories comportementales, telles que TPB et TRA, sont utilisées pour anticiper les actions déterminées par les attitudes et les intentions comportementales existantes, délimitant le lien entre les attitudes et les comportements dans l'activité humaine [297]. Enfin, la catégorie 'Autres' englobe des théories comme MT, ITI, HTFT et RIDT, traitant de divers aspects, allant de la motivation au comportement et plus encore, qui ne s'alignent pas confortablement avec les cinq classifications précédentes.

Dans la littérature examinée, la Théorie de l'Autodétermination (SDT) a émergé comme parti-

culièrement pertinente. C'est un cadre psychologique qui décrit trois besoins intrinsèques fondamentaux qui orientent le comportement humain : autonomie, compétence et appartenance. Dans le contexte des MOOCs, la SDT joue un rôle crucial. Les MOOCs répondent au besoin d'autonomie en permettant aux apprenants de choisir leurs cours préférés, de contrôler leur rythme d'apprentissage et de conserver le choix de leur parcours d'apprentissage, influençant ainsi leurs décisions de poursuivre le cours [278]. L'élément de compétence dans la SDT est évident lorsque les apprenants, en s'engageant dans les MOOCs, améliorent leurs capacités et augmentent leur sentiment d'efficacité, ce qui encourage un engagement plus important [298]. De plus, le besoin d'appartenance est comblé car les MOOCs favorisent un sentiment d'appartenance et facilitent les interactions sociales, ce qui renforce davantage l'engagement [278]. L'interaction entre l'autonomie, la compétence et l'appartenance dans la SDT contribue donc de manière significative à la motivation, à l'engagement et à la persévérance des apprenants dans les MOOCs [281], [296].

En plus de ces perspectives théoriques, nous avons également identifié des stratégies pratiques qui peuvent être utilisées pour améliorer la motivation et la rétention dans les MOOCs. Ces stratégies peuvent être catégorisées sous quatre catégories principales selon le tableau 7.3 : Personnalisation, Engagement, Soutien et Communication. Chacune de ces catégories peut être associée aux dimensions théoriques identifiées précédemment. Par exemple, les stratégies de personnalisation telles que l'apprentissage adaptatif, le contenu pertinent et authentique et la conception centrée sur l'apprenant sont étroitement liées aux théories du contrôle de l'apprenant, y compris la SDT. En adaptant l'expérience d'apprentissage à l'apprenant individuel, ces stratégies peuvent améliorer le sentiment d'autonomie et de compétence, deux composantes clés de la SDT.

TABLE 7.3: Catégories de stratégies pour améliorer la motivation et la rétention dans les MOOCs

| **Catégorie de Stratégie** | **Stratégie** | **Support Théorique** |
|---|---|---|
| Personnalisation | Apprentissage adaptatif | SDT, TAM, SRLT |
| | Contenu pertinent et authentique | SDT, EVT |
| | Conception centrée sur l'apprenant | SDT, CLT, SRLT |
| | Analyse de l'apprentissage | SDT, TAM |
| | Réseaux d'apprentissage personnels | SDT, EVT |
| | Parcours d'apprentissage personnels | SDT, SRLT |
| Engagement | Gamification | SDT, SCT, TRA |
| | Collaboration et soutien entre pairs | SCT, CT, SDT |
| | Présence et interaction de l'instructeur | SCT, CT, EVT |
| | Applications concrètes | SDT, EVT |
| | Expériences immersives (e.g., RV) | SDT, CLT |

| | | |
|---|---|---|
| | Activités d'apprentissage interactives et compétences de pensée de haut niveau (HOTS) | SDT, CLT, ALT |
| Soutien | Feedback régulier et évaluation | SDT, FT, EVT |
| | Apprentissage échafaudé | SCT, CLT, ALT |
| | Conception de cours accessible | TAM,UDL |
| | Amélioration continue | SDT,TAM |
| | Soutien multilingue | CHAT |
| | Encadrement par les pairs et mentorat | SCT, SDT |
| Communication | Communication claire des attentes | SDT, EVT, TRA |
| | Feedback et réflexion | SDT,FT, EVT |
| | Communication multi-modale (e.g., vidéo, audio, texte) | SDT, MRT,TRA |
| | Intégration des médias sociaux | SCT,MRT,TAM |
| | Discussions synchrones | SCT,CT,SDT |

**Personnalisation**

— **Apprentissage adaptatif :** Utilise des algorithmes pour fournir des ressources et des activités adaptées aux besoins uniques de l'apprenant. Basé sur la théorie de l'autodétermination (SDT) et la théorie de l'apprentissage constructiviste (TAC), il améliore l'engagement et la compréhension en ajustant le rythme et la difficulté du matériel selon la progression de l'étudiant. L'application de ces cadres théoriques favorise une expérience d'apprentissage plus personnalisée et efficace, particulièrement dans les environnements d'apprentissage en ligne et numérique [299].

— **Contenu pertinent et authentique :** Informé par la théorie de l'apprentissage expérientiel (TAE) et la théorie de l'apprentissage situé (TAS). En alignant le matériel éducatif avec des situations réelles, les apprenants peuvent appliquer directement leurs connaissances et compétences à des contextes pratiques, stimulant ainsi la motivation et l'engagement [300].

— **Conception centrée sur l'apprenant :** Donne la priorité aux besoins et objectifs individuels des apprenants, encourageant l'apprentissage actif et la coopération dans la résolution de problèmes. Soutenue par la théorie de l'autodétermination (SDT) et la théorie de la charge cognitive (CLT), cette approche cultive l'autonomie et la responsabilité, améliorant les résultats d'apprentissage et augmentant la motivation des étudiants [301].

— **Analyse de l'apprentissage :** Consiste à collecter, évaluer et rapporter des données pour optimiser l'apprentissage. Utilisant des techniques statistiques et d'apprentissage automatique, cette approche prévoit la performance des apprenants et identifie ceux à risque, fournissant un

feedback personnalisé et soutenant la conception des environnements d'apprentissage. Elle est soutenue par le modèle d'acceptation de la technologie (MAT) [299].

— **Réseaux d'apprentissage personnels (RAP) :** Permettent aux apprenants de se connecter et d'apprendre des autres à l'échelle mondiale. Soutenus par la théorie de l'attente-valeur (TAV), les RAP offrent des perspectives diverses et un soutien en temps réel, favorisant l'apprentissage tout au long de la vie et le développement professionnel [302] [303].

— **Parcours d'apprentissage personnels :** Offrent des routes d'apprentissage adaptées aux besoins et intérêts des apprenants. Utilisant des technologies numériques, cette approche favorise l'apprentissage autodirigé et donne aux apprenants de l'agence, améliorant l'engagement et les résultats d'apprentissage [304] [305].

**Engagement**

— **Gamification :** Utilise des éléments de conception de jeu pour stimuler l'engagement et la motivation. Elle est soutenue par la théorie de l'autodétermination (SDT), la théorie sociale cognitive (SCT) et la théorie de l'action raisonnée (TRA). En rendant l'apprentissage plus agréable et compétitif, la gamification améliore les résultats d'apprentissage [306] [283] [307].

— **Collaboration et soutien entre pairs :** Impliquent le travail collectif pour atteindre des objectifs d'apprentissage communs. Enracinée dans la théorie sociale cognitive (SCT) et la théorie du constructivisme (TC), cette stratégie améliore l'engagement, la compréhension et la rétention des étudiants [308].

— **Présence et interaction de l'instructeur :** Concernent l'implication active de l'instructeur dans le processus d'apprentissage. Cela crée un sentiment de communauté et de connexion, essentiel dans les environnements d'apprentissage en ligne, et est soutenu par la théorie du constructivisme social et la théorie de l'attente-valeur [308].

— **Applications concrètes :** Intègrent des exemples et des tâches pratiques dans le contenu éducatif, améliorant l'engagement et la motivation des étudiants. Soutenues par la théorie de l'apprentissage expérientiel et la théorie de l'autodétermination, elles rendent l'apprentissage plus significatif et pertinent [309] [310].

— **Expériences immersives :** Les expériences immersives, telles que la réalité virtuelle (VR) et la réalité augmentée (AR), offrent des environnements d'apprentissage interactifs et réalistes. Soutenues par la théorie de l'apprentissage expérientiel et la théorie de la charge cognitive, elles améliorent les résultats d'apprentissage et l'engagement des apprenants [311].

— **Activités d'apprentissage interactives :** Favorisent la participation active et le développement des compétences de réflexion de haut niveau (CRHN). En alignant les tâches interactives avec des théories telles que la théorie de l'apprentissage constructiviste et la théorie de l'apprentissage

actif, elles améliorent la motivation et les résultats d'apprentissage [312].

**Soutien**

- **Feedback régulier et évaluation :** Offrent une rétroaction constructive et en temps opportun sur les performances des apprenants, améliorant la motivation et l'autorégulation. Ces pratiques sont soutenues par la théorie de l'autodétermination (SDT) et la théorie de la feedback [313].
- **Apprentissage échafaudé :** Fournit un soutien structuré pour maîtriser des tâches complexes, soutenu par la théorie de Vygotsky sur la Zone de Développement Proximal. Cela améliore la compréhension, l'auto-efficacité et l'autonomie des apprenants [314].
- **Conception de cours accessible :** Vise à créer des cours inclusifs pour tous les apprenants, en utilisant les principes de la Conception Universelle pour l'Apprentissage (CUA). Cela améliore l'engagement et les résultats des apprenants [315].
- **Amélioration continue :** Implique un processus d'évaluation et de raffinement des pratiques d'enseignement pour optimiser l'expérience d'apprentissage. Elle est soutenue par le cycle Planifier-Faire-Étudier-Agir (PDSA) et la pratique réflexive [316].
- **Soutien multilingue :** Fournit des ressources et des matériaux répondant aux apprenants de différentes origines linguistiques, améliorant l'accessibilité et les résultats d'apprentissage globaux [317].
- **Encadrement par les pairs et mentorat :** Offrent des conseils et du soutien, améliorant l'engagement et les résultats d'apprentissage dans les MOOCs [318].

**Communication**

- **Communication claire des attentes :** Assure que les apprenants comprennent ce qui est attendu d'eux. Cette stratégie, ancrée dans la théorie de l'apprentissage autorégulé, améliore la gestion du processus d'apprentissage et l'engagement des apprenants [319].
- **Feedback et réflexion :** Favorisent un apprentissage optimal en offrant des informations sur les performances et en encourageant les processus métacognitifs. Soutenue par la théorie de l'autodétermination, la théorie du flow, et la théorie de l'expectancy-value, cette stratégie améliore la motivation et les résultats d'apprentissage [320].
- **Communication multimodale :** Intègre des formats vidéo, audio et texte pour améliorer l'efficacité de la communication. Basée sur la théorie cognitive de l'apprentissage multimédia, cette approche maximise l'engagement des apprenants et les résultats d'apprentissage [321].
- **Intégration des médias sociaux :** Favorise l'interaction sociale et l'apprentissage collaboratif, soutenue par la théorie sociale cognitive, la théorie de la richesse des médias, et le modèle d'acceptation de la technologie. Cette intégration améliore l'engagement et le partage des connaissances

[206].

— **Discussions synchrones :** Facilitent l'interaction en temps réel et le travail d'équipe. En intégrant des outils de communication synchrones, cette stratégie améliore l'engagement, la compréhension et la connexion sociale entre les apprenants et les instructeurs [251].

Les stratégies discutées ont un potentiel considérable pour améliorer les expériences d'apprentissage dans les MOOCs et réduire les taux d'abandon. En favorisant un environnement d'apprentissage engageant, personnalisé et efficace, ces stratégies peuvent améliorer les résultats d'apprentissage et être adaptées pour répondre aux besoins, préférences et objectifs diversifiés des apprenants. Cependant, il est crucial de reconnaître que l'application réussie de ces stratégies n'est pas une solution universelle mais dépend plutôt du contexte. Leur mise en œuvre efficace repose sur une personnalisation et une adaptation soigneuses au contexte unique, au sujet à traiter, au profil de l'apprenant, et aux ressources disponibles. Par conséquent, bien que ces stratégies présentent des avantages potentiels, elles nécessitent une planification méticuleuse, une exécution et un raffinement constants pour réaliser pleinement leur potentiel. Même avec le potentiel évident de ces stratégies, elles ne sont pas dépourvues de défis et de limitations.

Au fur et à mesure que nous progressons dans la sous-section suivante, nous allons nous pencher sur un examen complet des obstacles potentiels et des contraintes associées à la mise en œuvre de ces stratégies dans les MOOCs. Cette exploration couvrira des facteurs tels que l'accessibilité, le coût, les prérequis techniques, le soutien aux apprenants et les considérations éthiques. Reconnaître ces défis est une étape fondamentale vers l'élaboration de solutions efficaces et la garantie que nos efforts pour améliorer l'apprentissage en ligne sont inclusifs, équitables et avantageux pour tous les apprenants.

**Q3 : Quelles sont les limitations ou les défis associés à la mise en œuvre de stratégies pour améliorer la motivation et la rétention dans les MOOCs ?**

### 7.3.3 Limitations pour améliorer la motivation et la rétention dans les MOOCs

Les défis pour améliorer motivation et rétention dans les MOOCs incluent le coût et la scalabilité des technologies, la littératie numérique, les préoccupations de confidentialité et de sécurité, l'évaluation complexe de l'efficacité des stratégies, et le manque d'interaction personnelle. Ces défis touchent la Personnalisation, l'Engagement, le Soutien et la Communication.

Les stratégies de personnalisation dans les MOOCs, comme l'apprentissage adaptatif, le contenu pertinent, la conception centrée sur l'apprenant, l'analyse de l'apprentissage, et les parcours d'apprentissage personnels, peuvent améliorer la motivation et la rétention. Cependant, elles posent des défis en matière de confidentialité des données, de sécurité, d'infrastructure technologique, de création de contenu, et

nécessitent une expertise pédagogique. La mise en œuvre de ces stratégies est gourmande en ressources et requiert des outils appropriés et une gestion active de la communauté [322, 323, 324, 325, 305].

En plus des stratégies de personnalisation, les stratégies d'engagement comme la gamification, la collaboration entre pairs, l'interaction de l'instructeur, les applications du monde réel, les expériences immersives et les activités interactives améliorent la motivation et la rétention dans les MOOCs. Cependant, elles présentent des défis tels que la conception soignée de la gamification [326], les limites de l'apprentissage asynchrone pour la collaboration [307], les exigences pour une interaction authentique de l'instructeur, les contraintes techniques et de ressources pour les expériences immersives [283], et la conception minutieuse des activités interactives pour les HOTS [327].

Les stratégies de soutien dans les MOOCs, telles que les retours réguliers, l'apprentissage structuré, la conception de cours accessibles, l'amélioration continue, le soutien multilingue, et le mentorat par les pairs, sont essentielles pour la motivation et la rétention. Cependant, elles posent des défis comme la gestion chronophage des retours [328], l'adaptation aux divers besoins des apprenants [298], la planification soignée pour l'accessibilité [329], la demande de ressources pour l'amélioration continue, la complexité du soutien multilingue, et la coordination du mentorat [8].

Les stratégies de communication dans les MOOCs, telles que la communication claire des attentes, les retours d'information, la communication multimodale, l'intégration des médias sociaux et les discussions synchrones, sont vitales pour la motivation et la rétention. Cependant, elles présentent des défis : nécessité d'une conception pédagogique efficace [45], gestion de la communication asynchrone [328], besoin d'expertise technique et de ressources pour la communication multimodale [330], préoccupations de confidentialité et sécurité avec les médias sociaux [331], et défis logistiques pour les discussions synchrones [332].

Les stratégies pour améliorer la motivation et la rétention dans les MOOCs présentent des défis techniques, logistiques, pédagogiques et éthiques, liés à la Personnalisation, l'Engagement, le Soutien et la Communication. Les éducateurs et concepteurs doivent en être conscients et prendre des décisions adaptées aux contextes spécifiques et aux profils des apprenants. Une évaluation continue de ces stratégies est essentielle pour garantir l'efficacité des MOOCs en termes d'engagement et de résultats d'apprentissage.

## 7.4 Discussion

Ce chapitre met en lumière le potentiel de diverses stratégies pour renforcer la motivation et la rétention dans les MOOCs. Bien que ces stratégies aient démontré leur potentiel pour stimuler l'engagement des apprenants et améliorer les résultats éducatifs, il est crucial de reconnaître les limites et les difficultés liées à leur application. Nous examinerons les implications de ces connaissances et discuterons de leur importance dans l'environnement des MOOCs, tout en proposant des orientations pour la

recherche future afin de combler les lacunes existantes et de continuer à faire progresser ce domaine.

Ce chapitre souligne la puissance des stratégies de personnalisation pour stimuler la motivation et la rétention dans les MOOCs. Cependant, il est important de noter que ces stratégies nécessitent une approche prudente en ce qui concerne les questions de confidentialité et de sécurité des données, en plus de la nécessité de disposer de solides cadres technologiques. Les recherches futures doivent se pencher sur les facteurs culturels et situationnels qui affectent l'efficacité des stratégies de personnalisation dans divers environnements éducatifs, afin d'assurer leur adaptabilité et leur passage à l'échelle.

Il existe plusieurs orientations et recommandations pour la recherche future, basées sur cette étude, qui pourraient améliorer notre compréhension de la façon d'augmenter la motivation et la rétention dans les MOOCs.

Tout d'abord, la recherche future devrait donner la priorité à l'investigation de l'efficacité de ces stratégies dans des contextes culturels et éducatifs variés, afin d'assurer leur applicabilité et leur universalité. Les aspects culturels peuvent influencer fortement la motivation et l'engagement des apprenants, et comprendre comment ces stratégies fonctionnent dans différents contextes culturels est essentiel pour leur déploiement réussi. Des études comparatives dans différentes zones géographiques et systèmes éducatifs peuvent éclairer les subtilités culturelles qui ont un impact sur l'efficacité de ces stratégies.

Deuxièmement, il est nécessaire de mener des études longitudinales pour explorer les impacts à long terme de ces stratégies sur la motivation et la rétention des apprenants. Bien que ce chapitre offre des informations sur les effets immédiats de ces stratégies, comprendre leur influence prolongée dans le temps est primordial. Les études longitudinales peuvent aider à identifier les facteurs qui conduisent à un engagement et une rétention durables, en fournissant des orientations pour concevoir des interventions qui favorisent la persévérance et le succès des apprenants.

De plus, la recherche future devrait viser à relever les défis et les limites associés au déploiement de ces stratégies. Par exemple, les coûts et la possibilité de mise à l'échelle de certaines stratégies, comme les expériences immersives et l'apprentissage adaptatif, méritent une exploration plus approfondie. Trouver des solutions rentables et des méthodes innovantes pour dépasser les limites de ressources sera essentiel pour rendre ces stratégies plus accessibles et durables dans divers contextes éducatifs.

De plus, la recherche devrait s'intéresser à la possibilité d'intégrer de manière synergique plusieurs stratégies pour stimuler la motivation et la rétention dans les MOOCs. L'interaction entre différentes stratégies et leurs effets combinés sur les résultats des apprenants restent relativement inexploités. En comprenant comment différentes stratégies se complètent et se renforcent mutuellement, les éducateurs et les concepteurs pédagogiques peuvent concevoir des interventions plus holistiques et efficaces pour favoriser l'engagement et le succès des apprenants.

Enfin, les considérations éthiques liées à l'utilisation de la technologie et des données dans les MOOCs doivent être abordées. Comme ces stratégies dépendent souvent de la collecte et de l'analyse

des données des apprenants, il est primordial de garantir la confidentialité, la sécurité et l'utilisation responsable des données. La recherche future devrait se concentrer sur la formulation de directives et de cadres éthiques pour le déploiement de ces stratégies, en tenant compte des droits et du consentement des apprenants, ainsi que des devoirs éthiques des éducateurs et des institutions.

Pour conclure, ce chapitre offre des informations précieuses sur le potentiel de différentes stratégies pour améliorer la motivation et la rétention dans les MOOCs. Les conclusions soulignent l'importance de la personnalisation, de l'engagement, du soutien et de la communication dans la création d'environnements d'apprentissage engageants et efficaces. Cependant, l'application de ces stratégies nécessite une réflexion soigneuse sur les défis et les limites associés. En surmontant ces obstacles et en faisant progresser la recherche dans les directions recommandées, nous pouvons optimiser davantage l'application de ces stratégies dans les MOOCs et contribuer à l'amélioration continue des expériences d'apprentissage en ligne.

## 7.5 Conclusion

En conclusion, ce chapitre offre un examen complet du rôle des stratégies de motivation dans l'amélioration de l'engagement des apprenants et la réduction des taux d'abandon dans les Massive Open Online Courses (MOOCs). Elle met en évidence le potentiel prometteur de techniques telles que l'apprentissage personnalisé, le contenu interactif, la collaboration entre pairs et le feedback sur la progression. Cependant, elle souligne également la nécessité d'une conception et d'une mise en œuvre soigneuses de ces stratégies, en veillant à ce qu'elles soient alignées sur les objectifs uniques et les motivations intrinsèques des apprenants.

Malgré le potentiel prometteur de ces stratégies, la revue reconnaît les limites existantes, notamment les contraintes technologiques, les préoccupations en matière de confidentialité et de sécurité des données, et la nécessité d'une compréhension robuste de la diversité des apprenants. Elle plaide pour l'élaboration de lignes directrices éthiques pour l'utilisation des données et la confidentialité dans les MOOCs, un domaine auquel la recherche future devrait prêter une attention particulière.

De plus, la revue souligne l'importance de la recherche future dans des contextes culturels et éducatifs divers afin d'assurer l'applicabilité et l'efficacité larges de ces stratégies de motivation. Elle souligne également la nécessité d'études longitudinales pour comprendre l'impact soutenu de ces stratégies sur la motivation, l'engagement des apprenants et les taux de réussite des MOOCs.

En fin de compte, ce chapitre apporte des aperçus précieux pour le développement continu et l'optimisation des MOOCs. En fournissant une compréhension robuste des stratégies de motivation et de leur impact, elle ouvre la voie à une conception et une mise en œuvre plus efficaces des MOOCs, améliorant finalement l'expérience d'apprentissage pour les apprenants diversifiés à travers le monde.

# Conclusion générale

Les MOOCs ont révolutionné l'éducation en offrant un accès gratuit à une éducation de qualité à des millions de personnes à travers le monde. Cependant, malgré leur popularité croissante, les MOOCs sont confrontés à un taux d'abandon élevé. Ce mémoire a exploré l'utilisation de l'intelligence artificielle et de l'apprentissage supervisé pour prédire et prévenir le décrochage dans les MOOCs.

Nous avons d'abord examiné les facteurs de décrochage et présenté une revue de la littérature sur les méthodes d'analyse de l'apprentissage utilisées pour prédire le décrochage. Ensuite, nous avons détaillé notre approche pour résoudre les limitations de ces méthodes, en utilisant la réduction de la dimensionnalité, la sélection de variables, la construction de modèles, la prédiction, et enfin la priorisation et la personnalisation des interventions.

Les contributions de cette thèse se déclinent en plusieurs chapitres principaux. Le premier chapitre présente une application pratique des techniques d'apprentissage sur les données de la KDD Cup 2015, une compétition de data mining. Cette étude initiale pose les bases de notre approche en démontrant l'efficacité des techniques d'apprentissage automatique sur un jeu de données complexe.

En second lieu, nous avons inclus un chapitre sur les déterminants multidimensionnels de la performance académique : perspectives des étudiants dans les universités marocaines, à travers une étude statistique par questionnaire. Ce chapitre met en lumière divers facteurs démographiques, académiques, psychologiques, sociaux et liés au mode de vie qui influencent significativement la performance des étudiants. Ces résultats soulignent l'importance d'adopter une approche holistique dans l'éducation, intégrant un soutien psychologique, un équilibre de vie sain et un engagement sociétal au sein des stratégies académiques.

Le troisième chapitre examine les effets de la gamification sur la motivation et l'engagement des étudiants dans les MOOCs. Nous avons constaté que la gamification peut augmenter la motivation des apprenants et améliorer leur expérience d'apprentissage, renforçant ainsi leur engagement et leur persévérance dans les cours.

Enfin, le quatrième chapitre discute des théories de la motivation et propose des stratégies pour améliorer la rétention des apprenants dans les MOOCs. En explorant divers facteurs qui influencent la motivation des apprenants, tels que les besoins académiques, les intérêts personnels, les facteurs sociaux et technologiques, nous avons pu identifier des stratégies efficaces pour améliorer la motivation et la rétention des apprenants.

Pour les perspectives futures, nous envisageons d'explorer d'autres méthodes pour la prédiction du décrochage, notamment en utilisant des graphes. Les graphes offrent l'avantage de pouvoir modéliser les relations et interactions complexes entre les différents éléments du système éducatif, ce qui pourrait

améliorer la précision des prédictions de décrochage en capturant les dynamiques sous-jacentes aux comportements des étudiants. De plus, les graphes permettent d'analyser les réseaux d'apprentissage et de collaboration entre les étudiants, fournissant ainsi une vue plus complète et intégrée des facteurs qui influencent leur engagement et leur réussite académique.

Une autre piste prometteuse pour les recherches futures est l'intégration des technologies de réalité virtuelle et augmentée (VR/AR) dans les MOOCs. Ces technologies immersives peuvent enrichir l'expérience d'apprentissage en créant des environnements interactifs et engageants. La VR/AR peut simuler des situations réelles et pratiques, ce qui peut aider à renforcer la compréhension et la rétention des connaissances. En explorant comment ces technologies peuvent être utilisées pour améliorer l'engagement et la motivation des étudiants, nous pourrions développer de nouvelles approches pédagogiques qui réduisent encore davantage le taux de décrochage.

En conclusion, cette thèse contribue à une meilleure compréhension des facteurs influençant la performance académique et le décrochage dans les MOOCs. Les résultats et les méthodologies présentés dans ce travail offrent une base solide pour de futures recherches visant à améliorer la rétention des apprenants et à promouvoir leur succès académique. Les insights obtenus peuvent guider les éducateurs, les décideurs politiques et les chercheurs dans le développement de stratégies plus efficaces pour soutenir les étudiants dans leur parcours éducatif.

# Bibliographie


[1] D. Lizcano, J. A. Lara, B. White, and S. Aljawarneh, "Blockchain-based approach to create a model of trust in open and ubiquitous higher education," *Journal of Computing in Higher Education*, vol. 32, no. 1, pp. 109–134, 2020.

[2] D. Shah, "The second year of the mooc : A review of mooc stats and trends in 2020," *The Report by class central [Electronic resource]. URL : https ://www. classcentral. com/report/the-second-year-of-the-mooc/(accessed : 09.11. 2021)*, 2020.

[3] A. R. Sari, C. J. Bonk, and M. Zhu, "Mooc instructor designs and challenges : what can be learned from existing moocs in indonesia and malaysia ?," *Asia Pacific Education Review*, vol. 21, no. 1, pp. 143–166, 2020.

[4] A. McAuley, B. Stewart, G. Siemens, and D. Cormier, "The mooc model for digital practice," 2010.

[5] F. Dalipi, A. S. Imran, and Z. Kastrati, "Mooc dropout prediction using machine learning techniques : Review and research challenges," in *2018 IEEE global engineering education conference (EDUCON)*, pp. 1007–1014, IEEE, 2018.

[6] A. Stagg, L. Nguyen, C. Bossu, H. Partridge, J. Funk, and K. Judith, "Open educational practices in australia : A first-phase national audit of higher education," *International Review of Research in Open and Distributed Learning*, vol. 19, no. 3, 2018.

[7] D. Catropa, "Big (mooc) data," *Inside Higher Education*, 2013.

[8] H. Khalil and M. Ebner, "Moocs completion rates and possible methods to improve retention-a literature review," *EdMedia+ innovate learning*, pp. 1305–1313, 2014.

[9] G. Siemens, "Learning analytics : The emergence of a discipline," *American Behavioral Scientist*, vol. 57, no. 10, pp. 1380–1400, 2013.

[10] S. Halawa, D. Greene, and J. Mitchell, "Dropout prediction in moocs using learner activity features," *Proceedings of the second European MOOC stakeholder summit*, vol. 37, no. 1, pp. 58–65, 2014.

[11] S. Henn, "Stanford takes online schooling to the next academic level," *NPR. org*, 2012.

[12] F. J. Palacios Hidalgo, C. A. Huertas Abril, and M. Gómez Parra, "Moocs : Origins, concept and didactic applications : A systematic review of the literature (2012–2019)," *Technology, Knowledge and Learning*, vol. 25, no. 4, pp. 853–879, 2020.

[13] M. Mellati and M. Khademi, "Mooc-based educational program and interaction in distance education : long life mode of teaching," *Interactive Learning Environments*, vol. 28, no. 8, pp. 1022–1035, 2020.

[14] S. Osuna-Acedo, C. Marta-Lazo, and D. Frau-Meigs, "From smooc to tmooc, learning towards professional transference : Eco european project [de smooc a tmooc, el aprendizaje hacia la transferencia profesional : El proyecto europeo eco]," *Comunicar*, no. ART-2018-105258, 2018.

[15] E. Romero-Frías, J. L. Arquero, and S. del Barrio-García, "Exploring how student motivation relates to acceptance and participation in moocs," *Interactive Learning Environments*, pp. 1–17, 2020.

[16] A. M. Teixeira, J. Mota, M. d. C. T. Pinto, and L. Morgado, "Can imoocs close the opportunity gaps ? : the contribution of social inclusive pedagogical design," *Revista Fuentes*, vol. 21, no. 2, pp. 239–252, 2019.

[17] D. Tao, P. Fu, Y. Wang, T. Zhang, and X. Qu, "Key characteristics in designing massive open online courses (moocs) for user acceptance : An application of the extended technology acceptance model," *Interactive Learning Environments*, vol. 30, no. 5, pp. 882–895, 2022.

[18] D. Yang, T. Sinha, D. Adamson, and C. P. Rosé, "Turn on, tune in, drop out : Anticipating student dropouts in massive open online courses," in *Proceedings of the 2013 NIPS Data-driven education workshop*, vol. 11, p. 14, 2013.

[19] Y. Belanger and J. Thornton, "Bioelectricity : A quantitative approach duke university's first mooc," 2013.

[20] C. Reparaz, M. Aznárez-Sanado, and G. Mendoza, "Self-regulation of learning and mooc retention," *Computers in Human Behavior*, vol. 111, p. 106423, 2020.

[21] W. Xing, H. Tang, and B. Pei, "Beyond positive and negative emotions : Looking into the role of achievement emotions in discussion forums of moocs," *The Internet and Higher Education*, vol. 43, p. 100690, 2019.

[22] K. S. Hone and G. R. El Said, "Exploring the factors affecting mooc retention : A survey study," *Computers & Education*, vol. 98, pp. 157–168, 2016.

[23] W. Wang, Y. Zhao, Y. J. Wu, and M. Goh, "Factors of dropout from moocs : a bibliometric review," *Library Hi Tech*, no. ahead-of-print, 2022.

[24] B. Prenkaj, P. Velardi, G. Stilo, D. Distante, and S. Faralli, "A survey of machine learning approaches for student dropout prediction in online courses," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020.

[25] A. Alamri, M. Alshehri, A. Cristea, F. D. Pereira, E. Oliveira, L. Shi, and C. Stewart, "Predicting moocs dropout using only two easily obtainable features from the first week's activities," in *International Conference on Intelligent Tutoring Systems*, pp. 163–173, Springer, 2019.

[26] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, and C. D. Kloos, "Prediction in moocs : A review and future research directions," *IEEE transactions on Learning Technologies*, vol. 12, no. 3, pp. 384–401, 2018.

[27] C. Jin, "Mooc student dropout prediction model based on learning behavior features and parameter optimization," *Interactive Learning Environments*, pp. 1–19, 2020.

[28] D. Ifenthaler and J. Y.-K. Yau, "Utilising learning analytics to support study success in higher education : a systematic review," *Educational Technology Research and Development*, vol. 68, no. 4, pp. 1961–1990, 2020.

[29] P. J. Guo and K. Reinecke, "Demographic differences in how students navigate through moocs," in *Proceedings of the first ACM conference on Learning@ scale conference*, pp. 21–30, 2014.

[30] M. Cisel, "Analyzing completion rates in the first french xmooc," *Proceedings of the European MOOC stakeholder summit*, vol. 26, p. 51, 2014.

[31] N. P. Morris, B. Swinnerton, and S. Hotchkiss, "Can demographic information predict mooc learner outcomes ?," in *Experience track : proceedings of the European MOOC stakeholder*, Leeds, 2015.

[32] G. Ayyappan, "Ensemble classifications for student academics performance data set," *Indian J. Comput. Sci. Eng*, vol. 10, pp. 31–34, 2019.

[33] Z. Kastrati, A. Kurti, and A. S. Imran, "Wet : Word embedding-topic distribution vectors for mooc video lectures dataset," *Data in brief*, vol. 28, p. 105090, 2020.

[34] "NAEP Data Mining Competition 2019 — sites.google.com." `https://sites.google.com/view/dataminingcompetition2019/`. [Accessed 24-Sep-2022].

[35] J. A. Gómez-Tejedor, A. Vidaurre, I. Tort-Ausina, J. Molina-Mateo, M.-A. Serrano, J. M. Meseguer-Dueñas, R. M. M. Sala, S. Quiles, and J. Riera, "Effectiveness of flip teaching on engineering students' performance in the physics lab," *Computers & Education*, vol. 144, p. 103708, 2020.

[36] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," *Scientific data*, vol. 4, no. 1, pp. 1–8, 2017.

[37] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring mathematical problem solving with the math dataset," *arXiv preprint arXiv :2103.03874*, 2021.

[38] B. Hong, Z. Wei, and Y. Yang, "Discovering learning behavior patterns to predict dropout in mooc," in *2017 12th International Conference on Computer Science and Education (ICCSE)*, pp. 700–704, IEEE, 2017.

[39] Z. Alj and M. O. C. Malki, "Predicting students drop-out in mooc from learning behavior using machine learning," *Journal of Uncertain Systems*, p. 2250011, 2022.

[40] J. Qiu, J. Tang, T. X. Liu, J. Gong, C. Zhang, Q. Zhang, and Y. Xue, "Modeling and predicting learning behavior in moocs," in *Proceedings of the ninth ACM international conference on web search and data mining*, pp. 93–102, 2016.

[41] R. Al-Shabandar, A. Hussain, A. Laws, R. Keight, J. Lunn, and N. Radi, "Machine learning approaches to predict learning outcomes in massive open online courses," in *2017 International joint conference on neural networks (IJCNN)*, pp. 713–720, IEEE, 2017.

[42] D. S. Chaplot, E. Rhim, and J. Kim, "Predicting student attrition in moocs using sentiment analysis and neural networks," in *CEUR Workshop Proceedings*, vol. 1432, pp. 7–12, 2015.

[43] M. Tan and P. Shao, "Prediction of student dropout in e-learning program through the use of machine learning method.," *International journal of emerging technologies in learning*, vol. 10, no. 1, 2015.

[44] R. F. Kizilcec, C. Piech, and E. Schneider, "Deconstructing disengagement : analyzing learner subpopulations in massive open online courses," in *Proceedings of the third international conference on learning analytics and knowledge*, pp. 170–179, 2013.

[45] R. F. Kizilcec and S. Halawa, "Attrition and achievement gaps in online learning," in *Proceedings of the second (2015) ACM conference on learning@ scale*, pp. 57–66, 2015.

[46] L. Qiu, Y. Liu, and Y. Liu, "An integrated framework with feature selection for dropout prediction in massive open online courses," *IEEE Access*, vol. 6, pp. 71474–71484, 2018.

[47] W. Feng, J. Tang, and T. X. Liu, "Understanding dropouts in moocs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 517–524, 2019.

[48] Y. Mourdi, M. Sadgal, H. El Kabtane, and W. B. Fathi, "A machine learning-based methodology to predict learners' dropout, success or failure in moocs," *International Journal of Web Information Systems*, 2019.

[49] A. S. Imran, F. Dalipi, and Z. Kastrati, "Predicting student dropout in a mooc : An evaluation of a deep neural network model," in *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence*, pp. 190–195, 2019.

[50] J. Liang, C. Li, and L. Zheng, "Machine learning application in moocs : Dropout prediction," in *2016 11th International Conference on Computer Science & Education (ICCSE)*, pp. 52–57, IEEE, 2016.

[51] H. Chouaib, "Sélection de caractéristiques : méthodes et applications," *Paris Descartes University : Paris, France*, 2011.

[52] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[53] R. Vijayanand and D. Devaraj, "A novel feature selection method using whale optimization algorithm and genetic operators for intrusion detection system in wireless mesh network," *IEEE Access*, vol. 8, pp. 56847–56854, 2020.

[54] J. Hua, W. D. Tembe, and E. R. Dougherty, "Performance of feature-selection methods in the classification of high-dimension data," *Pattern Recognition*, vol. 42, no. 3, pp. 409–424, 2009.

[55] J. C. Schlimmer *et al.*, "Efficiently inducing determinations : A complete and systematic search algorithm that uses optimal pruning," in *Proceedings of the 1993 International Conference on Machine Learning*, pp. 284–290, 1993.

[56] H. Daviet, *Class-Add, une procédure de sélection de variables basée sur une troncature k-additive de l'information mutuelle et sur une classification ascendante hiérarchique en pré-traitement.* PhD thesis, Université de Nantes, 2009.

[57] M. Alweshah, S. Alkhalaileh, D. Albashish, M. Mafarja, Q. Bsoul, and O. Dorgham, "A hybrid mine blast algorithm for feature selection problems," *Soft Computing*, vol. 25, no. 1, pp. 517–534, 2021.

[58] A. Wan, X. Dai, P. Zhang, Z. He, Y. Tian, S. Xie, B. Wu, M. Yu, T. Xu, K. Chen, *et al.*, "Fbnetv2 : Differentiable neural architecture search for spatial and channel dimensions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12965–12974, 2020.

[59] X. Geng, T.-Y. Liu, T. Qin, and H. Li, "Feature selection for ranking," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 407–414, 2007.

[60] N. Settouti and A. Hafa, "Approche filtre pour la sélection des gènes pertinents des données biopuces du cancer du côlon," 2013.

[61] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[62] B. Kumari and T. Swarnkar, "Filter versus wrapper feature subset selection in large dimensionality micro array : A review," 2011.

[63] N. Mahendran and D. R. V. PM, "A deep learning framework with an embedded-based feature selection approach for the early detection of the alzheimer's disease," *Computers in Biology and Medicine*, vol. 141, p. 105056, 2022.

[64] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Transactions on computers*, vol. 26, no. 09, pp. 917–922, 1977.

[65] L. Xu, P. Yan, and T. Chang, "Best first strategy for feature selection," in *9th international conference on pattern recognition*, pp. 706–707, IEEE Computer Society, 1988.

[66] L. Bobrowski, "Feature selection based on some homogeneity coefficient," in *[1988 Proceedings] 9th International Conference on Pattern Recognition*, pp. 544–546, IEEE, 1988.

[67] P. Langley and S. Sage, "Oblivious decision trees and abstract cases," in *Working notes of the AAAI-94 workshop on case-based reasoning*, pp. 113–117, Seattle, WA, 1994.

[68] K. Kira, L. A. Rendell, *et al.*, "The feature selection problem : Traditional methods and a new algorithm," in *Aaai*, vol. 2, pp. 129–134, 1992.

[69] I. Kononenko, "Estimating attributes : Analysis and extensions of relief," in *European conference on machine learning*, pp. 171–182, Springer, 1994.

[70] H. Liu, H. Motoda, and L. Yu, "Feature selection with selective sampling," in *In Proceedings of the Nineteenth International Conference on Machine Learning*, Citeseer, 2002.

[71] P. Pudil and J. Novovičová, "Novel methods for feature subset selection with respect to problem knowledge," in *Feature extraction, construction and selection*, pp. 101–116, Springer, 1998.

[72] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning : data mining, inference, and prediction*, vol. 2. Springer, 2009.

[73] J. Sheinvald, B. Dom, and W. Niblack, "A modeling approach to feature selection," in *[1990] Proceedings. 10th International Conference on Pattern Recognition*, vol. 1, pp. 535–539, IEEE, 1990.

[74] C. Cardie, "Using decision trees to improve case-based learning," in *Proceedings of the tenth international conference on machine learning*, pp. 25–32, 1993.

[75] D. Koller and M. Sahami, "Toward optimal feature selection," tech. rep., Stanford InfoLab, 1996.

[76] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*, vol. 454. Springer Science & Business Media, 2012.

[77] L. Yu and H. Liu, "Feature selection for high-dimensional data : A fast correlation-based filter solution," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, pp. 856–863, 2003.

[78] M. Dash, K. Choi, P. Scheuermann, and H. Liu, "Feature selection for clustering-a filter solution," in *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pp. 115–122, IEEE, 2002.

[79] M. Dash, H. Liu, and J. Yao, "Dimensionality reduction of unsupervised data," in *Proceedings ninth ieee international conference on tools with artificial intelligence*, pp. 532–539, IEEE, 1997.

[80] M. A. Hall, "Correlation-based feature selection of discrete and numeric class machine learning," 2000.

[81] A. Miller, *Subset selection in regression*. chapman and hall/CRC, 2002.

[82] A. N. Mucciardi and E. E. Gose, "A comparison of seven techniques for choosing subsets of pattern recognition properties," *IEEE Transactions on Computers*, vol. 100, no. 9, pp. 1023–1031, 1971.

[83] N. Slonim, G. Bejerano, S. Fine, and N. Tishby, "Discriminative feature selection via multiclass variable memory markov model," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 2, pp. 1–10, 2003.

[84] P. Mitra, C. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 3, pp. 301–312, 2002.

[85] H. Almuallim, T. G. Dietterich, *et al.*, "Learning with many irrelevant features.," in *AAAI*, vol. 91, pp. 547–552, Citeseer, 1991.

[86] H. Liul, H. Motoda, and M. Dash, "A monotonic measure for optimal feature selection," in *European conference on machine learning*, pp. 101–106, Springer, 1998.

[87] A. L. Oliveira and A. Sangiovanni-Vincentelli, "Constructive induction using a non-greedy strategy for feature selection," in *Machine Learning Proceedings 1992*, pp. 355–360, Elsevier, 1992.

[88] Y. Rui, T. S. Huang, and S.-F. Chang, "Image retrieval : Current techniques, promising directions, and open issues," *Journal of visual communication and image representation*, vol. 10, no. 1, pp. 39–62, 1999.

[89] M. Dash, "Feature selection via set cover," in *Proceedings 1997 IEEE knowledge and data engineering exchange workshop*, pp. 165–171, IEEE, 1997.

[90] N. Littlestone, "Learning quickly when irrelevant attributes abound : A new linear-threshold algorithm," *Machine learning*, vol. 2, no. 4, pp. 285–318, 1988.

[91] H. Liu, R. Setiono, *et al.*, "A probabilistic approach to feature selection-a filter solution," in *ICML*, vol. 96, pp. 319–327, Citeseer, 1996.

[92] J. Doak, "An evaluation of feature selection methods and their application to computer security," *Techninal Report CSE-92-18*, 1992.

[93] I. Foroutan and J. Sklansky, "Feature selection for automatic classification of non-gaussian data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 17, no. 2, pp. 187–198, 1987.

[94] M. Ichino and J. Sklansky, "Feature selection for linear classifier," in *Proceedings of the Seventh International Conference on Pattern Recognition*, vol. 1, pp. 124–127, 1984.

[95] M. Ichino and J. Sklansky, "Optimum feature selection by zero-one integer programming," *IEEE Transactions on Systems, Man, and Cybernetics*, no. 5, pp. 737–746, 1984.

[96] R. Caruana and D. Freitag, "Greedy attribute selection," in *Machine Learning Proceedings 1994*, pp. 28–36, Elsevier, 1994.

[97] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition : A review," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[98] P. M. Domingos, "Why does bagging work ? a bayesian account and its implications.," in *KDD*, pp. 155–158, Citeseer, 1997.

[99] A. W. Moore and M. S. Lee, "Efficient algorithms for minimizing cross validation error," in *Machine Learning Proceedings 1994*, pp. 190–198, Elsevier, 1994.

[100] M. Singh and G. M. Provan, "Efficient learning of selective bayesian network classifiers," 1995.

[101] P. A. Devijver and J. Kittler, *Pattern recognition : A statistical approach*. Prentice hall, 1982.

[102] A. Jain and D. Zongker, "Feature selection : Evaluation, application, and small sample performance," *IEEE transactions on pattern analysis and machine intelligence*, vol. 19, no. 2, pp. 153–158, 1997.

[103] J. G. Dy and C. E. Brodley, "Feature subset selection and order identification for unsupervised learning," in *ICML*, pp. 247–254, Citeseer, 2000.

[104] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *Journal of machine learning research*, vol. 5, no. Aug, pp. 845–889, 2004.

[105] H. Liu and R. Setiono, "Feature selection and classification–a probabilistic wrapper approach," in *Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pp. 419–424, CRC Press, 1996.

[106] D. B. Skalak, "Prototype and feature selection by sampling and random mutation hill climbing algorithms," in *Machine Learning Proceedings 1994*, pp. 293–301, Elsevier, 1994.

[107] H. Vafaie, I. F. Imam, *et al.*, "Feature selection methods : genetic algorithms vs. greedy-like search," in *Proceedings of the international conference on fuzzy and intelligent control systems*, vol. 51, p. 28, 1994.

[108] D. J. Stracuzzi and P. E. Utgoff, "Randomized variable elimination," *The Journal of Machine Learning Research*, vol. 5, pp. 1331–1362, 2004.

[109] S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection," in *Icml*, vol. 1, pp. 74–81, Citeseer, 2001.

[110] S. Loscalzo, L. Yu, and C. Ding, "Consensus group stable feature selection," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 567–576, 2009.

[111] T. Marill and D. Green, "On the effectiveness of receptors in recognition systems," *IEEE transactions on Information Theory*, vol. 9, no. 1, pp. 11–17, 1963.

[112] A. W. Whitney, "A direct method of nonparametric measurement selection," *IEEE transactions on computers*, vol. 100, no. 9, pp. 1100–1103, 1971.

[113] X. Yang, J. Tang, X. Yin, L. Chen, S. Zhang, and Q. Gao, "Optimization of an artificial neural network through the analysis of input features," *Applied Sciences*, vol. 11, no. 4, p. 1539, 2021.

[114] S. Harish and S. Sivanandam, "A hybrid feature selection approach using filter and wrapper methods for text classification," *Applied Soft Computing*, vol. 67, pp. 830–838, 2018.

[115] D. W. Aha and R. L. Bankert, "A comparative evaluation of sequential feature selection algorithms," in *Pre-proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, pp. 1–7, PMLR, 1995.

[116] J. Kittler, "Feature set search algorithms," *Pattern recognition and signal processing*, pp. 41–60, 1978.

[117] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern recognition letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[118] P. Somol, P. Pudil, and J. Kittler, "Fast branch & bound algorithms for optimal feature selection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 26, no. 7, pp. 900–912, 2004.

[119] A. Arauzo, J. M. Benitez, and J. L. Castro, "C-focus : A continuous extension of focus," in *Advances in Soft Computing*, pp. 225–232, Springer, 2003.

[120] H. Liu and R. Setiono, "Feature selection via a discretization and wrapper approach," *Computational Intelligence*, vol. 11, no. 3, pp. 311–321, 1995.

[121] D. E. Goldberg and E. David, *Genetic algorithms in search, optimization, and machine learning.* Addison-Wesley, 1989.

[122] B. L. Miller and D. E. Goldberg, "Genetic algorithms, tournament selection, and the effects of noise," *Complex systems*, vol. 9, no. 3, pp. 193–212, 1995.

[123] L. Davis, *Handbook of genetic algorithms*, vol. 1. Van Nostrand Reinhold New York, 1991.

[124] J. H. Holland, *Adaptation in natural and artificial systems.* University of Michigan Press, 1975.

[125] L. Davis, "Job-shop scheduling with genetic algorithms," *Proceedings of the 1st International Conference on Genetic Algorithms*, pp. 136–140, 1985.

[126] D. Whitley, "Genetic algorithms and neural networks : optimization with combinatorial search algorithms," *Journal of Heuristics*, vol. 4, no. 1, pp. 71–88, 1989.

[127] K. A. De Jong, "Analysis of the behavior of a class of genetic adaptive systems," *Doctoral dissertation, University of Michigan*, 1975.

[128] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[129] H. Akaike, "A new look at the statistical model identification," *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.

[130] S. J. Messick and R. P. Abelson, "The additive constant problem in multidimensional scaling.," *Psychometrika*, 1956.

[131] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[132] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[133] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[134] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[135] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[136] R. S. Baker, "Stupid tutoring systems, intelligent humans," *International Journal of Artificial Intelligence in Education*, vol. 26, pp. 600–614, 2016.

[137] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[138] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv :1609.04747*, 2016.

[139] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[140] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[141] J. Platt, "Sequential minimal optimization : A fast algorithm for training support vector machines," *Advances in kernel methods*, vol. 208, pp. 185–208, 1999.

[142] B. Schölkopf and A. J. Smola, *Learning with kernels : Support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[143] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[144] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[145] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[146] J. H. Friedman, "Greedy function approximation : a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[147] T. Chen and C. Guestrin, "Xgboost : A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, ACM, 2016.

[148] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[149] H. M. Gomes, A. Bifet, J. Read, J. P. Barddal, and J. Gama, "Adaptive random forests for evolving data stream classification," in *Data Mining (ICDM), 2017 IEEE International Conference on*, pp. 913–918, IEEE, 2017.

[150] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, pp. 4765–4774, 2017.

[151] J. Mourtada, E. Moutoussamy, R. Lardy, and F. Clérot, "Compact random forests," *arXiv preprint arXiv :1705.07366*, 2017.

[152] J. H. Friedman, "Greedy function approximation : A gradient boosting machine," *Annals of statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[153] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.

[154] T. Chen and C. Guestrin, "Xgboost : A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 785–794, ACM, 2016.

[155] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning : data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[156] G. Biau and E. Scornet, "A review of gradient boosting algorithms," *arXiv preprint arXiv :1603.02754*, 2016.

[157] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm : A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[158] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, "Catboost : unbiased boosting with categorical features," in *Advances in neural information processing systems*, pp. 6638–6648, 2018.

[159] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[160] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[161] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet : A generative model for raw audio," *arXiv preprint arXiv :1609.03499*, 2016.

[162] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[163] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[164] M. Tan and Q. V. Le, "Efficientnet : Rethinking model scaling for convolutional neural networks," *International Conference on Machine Learning*, pp. 6105–6114, 2019.

[165] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, 2019.

[166] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, A. Lee, V. Sankaralingam, J. Bi, and et al., "Language models are few-shot learners," *arXiv preprint arXiv :2005.14165*, 2020.

[167] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer, 2006.

[168] A. E. Hoerl and R. W. Kennard, "Ridge regression : Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[169] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society : Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[170] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[171] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[172] M. L. Stein, "Interpolation of spatial data : Some theory for kriging," 1999.

[173] K. Hornik, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[174] K. W. Hipel and R. Boucher, "Forecasting and management of technology," 1994.

[175] H. R. Maier and G. C. Dandy, "Neural networks for the prediction and forecasting of water resources variables : a review of modelling issues and applications," *Environmental Modelling & Software*, vol. 15, no. 1, pp. 101–124, 2000.

[176] G. P. Zhang, "Neural network forecasting of stock market indices," *International Journal of Forecasting*, vol. 17, no. 3, pp. 357–368, 2001.

[177] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," *Advances in Neural Information Processing Systems*, vol. 9, pp. 155–161, 1997.

[178] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, 1967.

[179] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *KDD*, pp. 226–231, 1996.

[180] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[181] M. M. Breunig, H. Kriegel, R. T. Ng, and J. Sander, "Lof : Identifying density-based local outliers," in *SIGMOD*, pp. 93–104, 2000.

[182] B. Sch"olkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[183] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics surveys*, vol. 4, pp. 40–79, 2010.

[184] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, pp. 2951–2959, 2012.

[185] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 281–305, 2012.

[186] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote : synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[187] H. He and E. Garcia, "Learning from imbalanced data," in *2009 International Conference on Knowledge Discovery and Data Mining*, pp. 1269–1270, IEEE, 2009.

[188] C. Bossu and T. Heck, "Engaging with open science in learning and teaching," *Education for Information*, vol. 36, no. 3, pp. 211–225, 2020.

[189] Z. Xie, "Modelling the dropout patterns of mooc learners," *Tsinghua Science and Technology*, vol. 25, no. 3, pp. 313–324, 2019.

[190] J. Figueroa-Cañas and T. Sancho-Vinuesa, "Changing the recent past to reduce ongoing dropout : An early learning analytics intervention for an online statistics course," *Open Learning : The Journal of Open, Distance and e-Learning*, pp. 1–18, 2021.

[191] B. Kang, J. McAuley, J. McAuley, and C. Elkan, "A power-law approach to cold-start recommendations," in *Proceedings of the 2015 KDD Cup Workshop*, (New York, NY, USA), ACM, 2015. KDD Cup 2015.

[192] V. Muthukumar and N. Bhalaji, "Moocversity-deep learning based dropout prediction in moocs over weeks," *Journal of Soft Computing Paradigm (JSCP)*, vol. 2, no. 03, pp. 140–152, 2020.

[193] S. Dass, K. Gary, and J. Cunningham, "Predicting student dropout in self-paced mooc course using random forest model," *Information*, vol. 12, no. 11, p. 476, 2021.

[194] P. M. Moreno-Marcos, P. J. Muñoz-Merino, J. Maldonado-Mahauad, M. Perez-Sanagustin, C. Alario-Hoyos, and C. D. Kloos, "Temporal analysis for dropout prediction using self-regulated learning strategies in self-paced moocs," *Computers & Education*, vol. 145, p. 103728, 2020.

[195] N. Gitinabard, Y. Xu, S. Heckman, T. Barnes, and C. F. Lynch, "How widely can prediction models be generalized ? performance prediction in blended courses," *IEEE Transactions on Learning Technologies*, vol. 12, no. 2, pp. 184–197, 2019.

[196] J. Berens, K. Schneider, S. Görtz, S. Oster, and J. Burghoff, "Early detection of students at risk–predicting student dropouts using administrative student data and machine learning methods," 2018.

[197] W. Xing and D. Du, "Dropout prediction in moocs : Using deep learning for personalized intervention," *Journal of Educational Computing Research*, vol. 57, no. 3, pp. 547–570, 2019.

[198] R. Munoz, R. Olivares, C. Taramasco, R. Villarroel, R. Soto, T. S. Barcelos, E. Merino, and M. F. Alonso-Sánchez, "Using black hole algorithm to improve eeg-based emotion recognition," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[199] R. Olivares, R. Munoz, R. Soto, B. Crawford, D. Cárdenas, A. Ponce, and C. Taramasco, "An optimized brain-based algorithm for classifying parkinson's disease," *Applied Sciences*, vol. 10, no. 5, p. 1827, 2020.

[200] Y. Xue, H. Zhu, J. Liang, and A. Słowik, "Adaptive crossover operator based multi-objective binary genetic algorithm for feature selection in classification," *Knowledge-Based Systems*, vol. 227, p. 107218, 2021.

[201] C. Gütl, R. H. Rizzardini, V. Chang, and M. Morales, "Attrition in mooc : Lessons learned from drop-out students," in *Learning Technology for Education in Cloud. MOOC and Big Data : Third International Workshop, LTEC 2014, Santiago, Chile, September 2-5, 2014. Proceedings 3*, pp. 37–48, Springer, 2014.

[202] H. B. Shapiro, C. H. Lee, N. E. W. Roth, K. Li, M. Çetinkaya-Rundel, and D. A. Canelas, "Understanding the massive open online course (mooc) student experience : An examination of attitudes, motivations, and barriers," *Computers & Education*, vol. 110, pp. 35–50, 2017.

[203] M. Yamba-Yugsi and S. Luján-Mora, "Cursos mooc : factores que disminuyen el abandono en los participantes," *Enfoque UTE*, vol. 8, pp. 1–15, 2017.

[204] G. R. El Said, "Understanding how learners use massive open online courses and why they drop out : Thematic analysis of an interview study in a developing country," *Journal of Educational Computing Research*, vol. 55, no. 5, pp. 724–752, 2017.

[205] T. Eriksson, T. Adawi, and C. Stöhr, ""time is the bottleneck" : a qualitative study exploring why learners drop out of moocs," *Journal of Computing in Higher Education*, vol. 29, no. 1, pp. 133–146, 2017.

[206] S. Zheng, M. B. Rosson, P. C. Shih, and J. M. Carroll, "Understanding student motivation, behaviors and perceptions in moocs," in *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pp. 1882–1895, 2015.

[207] D. F. Onah, J. Sinclair, and R. Boyatt, "Dropout rates of massive open online courses : behavioural patterns," *EDULEARN14 proceedings*, pp. 5825–5834, 2014.

[208] T. NeCamp, J. Gardner, and C. Brooks, "Beyond a/b testing : Sequential randomization for developing interventions in scaled digital learning environments," in *Proceedings of the 9th International Conference on learning analytics & knowledge*, pp. 539–548, 2019.

[209] I. Borrella, S. Caballero-Caballero, and E. Ponce-Cueto, "Predict and intervene : Addressing the dropout problem in a mooc-based program," in *Proceedings of the sixth (2019) ACM conference on learning@ scale*, pp. 1–9, 2019.

[210] J. Whitehill, J. Williams, G. Lopez, C. Coleman, and J. Reich, "Beyond prediction : First steps toward automatic intervention in mooc student stopout," *Available at SSRN 2611750*, 2015.

[211] L. Bond, H. Butler, L. Thomas, J. Carlin, S. Glover, G. Bowes, and G. Patton, "Social and school connectedness in early secondary school as predictors of late teenage substance use, mental health, and academic outcomes," *Journal of adolescent health*, vol. 40, no. 4, pp. 357–e9, 2007.

[212] J. Bravo-Agapito, S. J. Romero, and S. Pamplona, "Early prediction of undergraduate student's academic performance in completely online learning : A five-year study," *Computers in Human Behavior*, vol. 115, p. 106595, 2021.

[213] A. Fahim, Q. Tan, B. Naz, Q. u. Ain, and S. U. Bazai, "Sustainable higher education reform quality assessment using swot analysis with integration of ahp and entropy models : A case study of morocco," *Sustainability*, vol. 13, no. 8, p. 4312, 2021.

[214] N. M. Alhajraf and A. M. Alasfour, "The impact of demographic and academic characteristics on academic performance," *International Business Research*, vol. 7, no. 4, p. 92, 2014.

[215] A. M. Al-Khani, M. I. Sarhandi, M. S. Zaghloul, M. Ewid, and N. Saquib, "A cross-sectional survey on sleep quality, mental health, and academic performance among medical students in saudi arabia," *BMC research notes*, vol. 12, no. 1, p. 665, 2019.

[216] P. Bhardwaj, "Types of sampling in research," *Journal of Primary Care Specialties*, vol. 5, no. 3, pp. 157–163, 2019.

[217] A. Rico, "Overview and methodology of the adolescent behaviors and experiences survey—united states, january–june 2021," *MMWR supplements*, vol. 71, 2022.

[218] S. M. Jones-Jang, Y.-J. Heo, R. McKeever, J.-H. Kim, L. Moscowitz, and D. Moscowitz, "Good news ! communication findings may be underestimated : Comparing effect sizes with self-reported and logged smartphone use data," *Journal of Computer-Mediated Communication*, vol. 25, no. 5, pp. 346–363, 2020.

[219] D. M. Heretick and J. Tanguma, "Anxiety and attitudes toward statistics and research among younger and older nontraditional adult learners," *The Journal of Continuing Higher Education*, vol. 69, no. 2, pp. 87–99, 2021.

[220] N. Komasawa, F. Terasaki, K. Takitani, S.-W. Lee, R. Kawata, and T. Nakano, "Comparison of younger and older medical student performance outcomes : a retrospective analysis in japan," *Medicine*, vol. 101, no. 45, p. e31392, 2022.

[221] K. Power, "The covid-19 pandemic has increased the care burden of women and families," *Sustainability : Science, practice and policy*, vol. 16, no. 1, pp. 67–73, 2020.

[222] V. Tinto, "Classrooms as communities : Exploring the educational character of student persistence," *The Journal of higher education*, vol. 68, no. 6, pp. 599–623, 1997.

[223] L. L. Anderson and S. A. V. García, "Library usage, instruction, and student success across disciplines : A multilevel model approach," *College & Research Libraries*, vol. 81, no. 3, p. 459, 2020.

[224] K. M. Stegers-Jager, M. Savas, J. van der Waal, E. F. van Rossum, and A. M. Woltman, "Gender-specific effects of raising year-1 standards on medical students' academic performance and stress levels," *Medical education*, vol. 54, no. 6, pp. 538–546, 2020.

[225] M. C. Pascoe, S. E. Hetrick, and A. G. Parker, "The impact of stress on students in secondary school and higher education," *International journal of adolescence and youth*, vol. 25, no. 1, pp. 104–112, 2020.

[226] E. Stark, "Examining the role of motivation and learning strategies in student success in online versus face-to-face courses.," *Online Learning*, vol. 23, no. 3, pp. 234–251, 2019.

[227] Y.-F. Chang, "2-dimensional cognitive test anxieties and their relationships with achievement goals, cognitive resources, motivational engagement, and academic performance," *Learning and Individual Differences*, vol. 92, p. 102084, 2021.

[228] C. J. Anthony and J. Ogg, "Parent involvement, approaches to learning, and student achievement : Examining longitudinal mediation.," *School Psychology*, vol. 34, no. 4, p. 376, 2019.

[229] M. A. Qureshi, A. Khaskheli, J. A. Qureshi, S. A. Raza, and S. Q. Yousufi, "Factors affecting students' learning performance through collaborative learning and engagement," *Interactive Learning Environments*, vol. 31, no. 4, pp. 2371–2391, 2023.

[230] F. Wang and S. Boros, "The effect of physical activity on sleep quality : a systematic review," *European journal of physiotherapy*, vol. 23, no. 1, pp. 11–18, 2021.

[231] A. E. Poropat, "A meta-analysis of the five-factor model of personality and academic performance.," *Psychological bulletin*, vol. 135, no. 2, p. 322, 2009.

[232] M. Richardson, C. Abraham, and R. Bond, "Psychological correlates of university students' academic performance : a systematic review and meta-analysis.," *Psychological bulletin*, vol. 138, no. 2, p. 353, 2012.

[233] L. Khanna, S. N. Singh, and M. Alam, "Multidimensional analysis of psychological factors affecting students academic performance," *arXiv preprint arXiv :1806.03242*, 2018.

[234] Y. Li, J. Allen, and A. Casillas, "Relating psychological and social factors to academic performance : A longitudinal investigation of high-poverty middle school students," *Journal of Adolescence*, vol. 56, pp. 179–189, 2017.

[235] R. M. Yerkes and J. D. Dodson, "The relation of strength of stimulus to rapidity of habit-formation," *Journal of comparative neurology and psychology*, vol. 18, no. 5, pp. 459–482, 1908.

[236] E. A. for Distance Teaching Universities, "Mooc completion rates : An overview of recent studies." Online, 2018.

[237] E. Kyewski and N. C. Krämer, "To gamify or not to gamify ? an experimental field study of the influence of badges on motivation, activity, and performance in an online learning course," *Computers & Education*, vol. 118, pp. 25–37, 2018.

[238] L. Hakulinen, T. Auvinen, and A. Korhonen, "The effect of achievement badges on students' behavior : An empirical study in a university-level computer science course.," *International Journal of Emerging Technologies in Learning*, vol. 10, no. 1, 2015.

[239] D. Gooch, A. Vasalou, L. Benton, and R. Khaled, "Using gamification to motivate students with dyslexia," in *Proceedings of the 2016 CHI Conference on human factors in computing systems*, pp. 969–980, 2016.

[240] M. R. U. Saputra and M. Risqi, "Lexipal : Design, implementation and evaluation of gamification on learning application for dyslexia," *International Journal of Computer Applications*, vol. 131, no. 7, pp. 37–43, 2015.

[241] M. D. Hanus and J. Fox, "Assessing the effects of gamification in the classroom : A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance," *Computers & education*, vol. 80, pp. 152–161, 2015.

[242] A. Y. Utomo and H. B. Santoso, "Development of gamification-enriched pedagogical agent for e-learning system based on community of inquiry," in *Proceedings of the International HCI and UX Conference in Indonesia*, pp. 1–9, 2015.

[243] B. Huang and K. F. Hew, "Do points, badges and leaderboard increase learning and activity : A quasi-experiment on the effects of gamification," in *Proceedings of the 23rd international conference on computers in education*, pp. 275–280, Society for Computer in Education Hangzhou, China, 2015.

[244] M.-S. Kuo and T.-Y. Chuang, "How gamification motivates visits and engagement for online academic dissemination–an empirical study," *Computers in Human Behavior*, vol. 55, pp. 16–27, 2016.

[245] O. Sitra, V. Katsigiannakis, C. Karagiannidis, and S. Mavropoulou, "The effect of badges on the engagement of students with special educational needs : A case study," *Education and Information Technologies*, vol. 22, pp. 3037–3046, 2017.

[246] T. Tenorio, I. I. Bittencourt, S. Isotani, A. Pedro, and P. Ospina, "A gamified peer assessment model for on-line learning environments in a competitive context," *Computers in Human Behavior*, vol. 64, pp. 247–263, 2016.

[247] M. Krause, M. Mogalle, H. Pohl, and J. J. Williams, "A playful game changer : Fostering student retention in online education with social gamification," in *Proceedings of the Second (2015) ACM conference on Learning@ Scale*, pp. 95–102, 2015.

[248] L. De-Marcos, A. Domínguez, J. Saenz-de Navarrete, and C. Pagés, "An empirical study comparing gamification and social networking on e-learning," *Computers & education*, vol. 75, pp. 82–91, 2014.

[249] A. Mazarakis, "Using gamification for technology enhanced learning : The case of feedback mechanisms," *Bulletin of the IEEE Technical Committee on Learning Technology*, vol. 4, no. 17, pp. 6–9, 2015.

[250] T. Aldemir, B. Celik, and G. Kaplan, "A qualitative investigation of student perceptions of game elements in a gamified course," *Computers in Human Behavior*, vol. 78, pp. 235–254, 2018.

[251] M. Zhu, A. Sari, and M. M. Lee, "A systematic review of research methods and topics of the empirical mooc literature (2014–2016)," *The Internet and Higher Education*, vol. 37, pp. 31–39, 2018.

[252] H. Aldowah, H. Al-Samarraie, A. I. Alzahrani, and N. Alalwan, "Factors affecting student dropout in moocs : a cause and effect decision-making model," *Journal of Computing in Higher Education*, vol. 32, pp. 429–454, 2020.

[253] J. Goopio and C. Cheung, "The mooc dropout phenomenon and retention strategies," *Journal of Teaching in Travel & Tourism*, vol. 21, no. 2, pp. 177–197, 2021.

[254] M. Barak, A. Watted, and H. Haick, "Motivation to learn in massive open online courses : Examining aspects of language and social engagement," *Computers & Education*, vol. 94, pp. 49–60, 2016.

[255] R. F. Kizilcec, M. Pérez-Sanagustín, and J. J. Maldonado, "Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses," *Computers & education*, vol. 104, pp. 18–33, 2017.

[256] Y. Xiong, H. Li, M. L. Kornhaber, H. K. Suen, B. Pursel, and D. D. Goins, "Examining the relations among student motivation, engagement, and retention in a mooc : A structural equation modeling approach.," *Global Education Review*, vol. 2, no. 3, pp. 23–33, 2015.

[257] Y. Nong, N. Buavaraporn, and P. Punnakitikashem, "Exploring the factors influencing users' satisfaction and continuance intention of moocs in china," *Kasetsart Journal of Social Sciences*, vol. 43, no. 2, pp. 403–408, 2022.

[258] M. Badali, J. Hatami, S. K. Banihashem, E. Rahimi, O. Noroozi, and Z. Eslami, "The role of motivation in moocs' retention rates : a systematic literature review," *Research and Practice in Technology Enhanced Learning*, vol. 17, no. 1, pp. 1–20, 2022.

[259] N. Hakami, S. White, and S. Chakaveh, "Motivational factors that influence the use of moocs : learners' perspectives-a systematic literature review," in *International Conference on Computer Supported Education*, vol. 2, pp. 323–331, SCITEPRESS, 2017.

[260] M. Zhu, A. R. Sari, and M. M. Lee, "A comprehensive systematic review of mooc research : Research techniques, topics, and trends from 2009 to 2019," *Educational Technology Research and Development*, vol. 68, pp. 1685–1710, 2020.

[261] Y. Wang and R. Baker, "Content or platform : Why do students complete moocs," *MERLOT Journal of Online Learning and Teaching*, vol. 11, no. 1, pp. 17–30, 2015.

[262] M. Subirana, I. Solá, J. M. Garcia, I. Gich, and G. Urrútia, "A nursing qualitative systematic review required medline and cinahl for study identification," *Journal of clinical epidemiology*, vol. 58, no. 1, pp. 20–25, 2005.

[263] Y. Jung and J. Lee, "Learning engagement and persistence in massive open online courses (moocs)," *Computers & Education*, vol. 122, pp. 9–22, 2018.

[264] Z. Shao, "Examining the impact mechanism of social psychological motivations on individuals' continuance intention of moocs : The moderating effect of gender," *Internet Research*, vol. 28, no. 1, pp. 232–250, 2018.

[265] T. Semenova, "The role of learners' motivation in mooc completion," *Open Learning : The Journal of Open, Distance and e-Learning*, vol. 37, no. 3, pp. 273–287, 2022.

[266] Y. S. Nleme Ze and G. Molinari, "Développement et validation psychométrique d'une échelle de mesure de l'engagement des apprenants dans les forums de discussion des mooc," *Distances et médiations des savoirs. Distance and Mediation of Knowledge*, vol. 40, no. 40, 2022.

[267] J. L. James, "Students as stakeholders : understanding expectations can increase student retention," *Journal of College Student Retention : Research, Theory & Practice*, vol. 24, no. 1, pp. 20–42, 2022.

[268] T. M. Kuo, C.-C. Tsai, and J.-C. Wang, "Linking web-based learning self-efficacy and learning engagement in moocs : The role of online academic hardiness," *The Internet and Higher Education*, vol. 51, p. 100819, 2021.

[269] M. Y. Doo, Y. Tang, C. J. Bonk, and M. Zhu, "Mooc instructor motivation and career development," *Distance Education*, vol. 41, no. 1, pp. 26–47, 2020.

[270] Z. Shao and K. Chen, "Understanding individuals' engagement and continuance intention of moocs : the effect of interactivity and the role of gender," *Internet Research*, vol. 31, no. 4, pp. 1262–1289, 2021.

[271] R. Cobos and J. C. Ruiz-Garcia, "Improving learner engagement in moocs using a learning intervention system : A research study in engineering education," *Computer Applications in Engineering Education*, vol. 29, no. 4, pp. 733–749, 2021.

[272] L. Meekers, A. Kumps, K. Boumazguida, G. Temperman, and B. De Lièvre, "Analyse de facteurs qui influencent le choix de parcours des apprenants dans le mooc «l'innovation pédagogique dont vous êtes le héros»," *Distances et médiations des savoirs. Distance and Mediation of Knowledge*, vol. 39, no. 39, 2022.

[273] E. E. Buhr, L. M. Daniels, and L. D. Goegan, "Cognitive appraisals mediate relationships between two basic psychological needs and emotions in a massive open online course," *Computers in Human Behavior*, vol. 96, pp. 85–94, 2019.

[274] I. Borrella, S. Caballero-Caballero, and E. Ponce-Cueto, "Taking action to reduce dropout in moocs : Tested interventions," *Computers & Education*, vol. 179, p. 104412, 2022.

[275] G. Schettino and V. Capone, "Learning design strategies in moocs for physicians' training : A scoping review," *International journal of environmental research and public health*, vol. 19, no. 21, p. 14247, 2022.

[276] Z. Yu, W. Xu, and P. Sukjairungwattana, "A meta-analysis of eight factors influencing mooc-based learning outcomes across the world," *Interactive Learning Environments*, pp. 1–20, 2022.

[277] E. B. Gregori, J. Zhang, C. Galván-Fernández, and F. de Asís Fernández-Navarro, "Learner support in moocs : Identifying variables linked to completion," *Computers & Education*, vol. 122, pp. 153–168, 2018.

[278] I. U. Khan, Z. Hameed, Y. Yu, T. Islam, Z. Sheikh, and S. U. Khan, "Predicting the acceptance of moocs in a developing country : Application of task-technology fit model, social motivation, and self-determination theory," *Telematics and Informatics*, vol. 35, no. 4, pp. 964–978, 2018.

[279] E. Romero-Frías, J. L. Arquero, and S. del Barrio-García, "Exploring how student motivation relates to acceptance and participation in moocs," *Interactive Learning Environments*, vol. 31, no. 1, pp. 480–496, 2023.

[280] A. Y. Huang, O. H. Lu, and S. J. Yang, "Effects of artificial intelligence–enabled personalized recommendations on learners' learning engagement, motivation, and outcomes in a flipped classroom," *Computers & Education*, vol. 194, p. 104684, 2023.

[281] Y. J. Joo, H.-J. So, and N. H. Kim, "Examination of relationships among students' self-determination, technology acceptance, satisfaction, and continuance intention to use k-moocs," *Computers & Education*, vol. 122, pp. 260–272, 2018.

[282] R. Vezne, H. Yildiz Durak, and N. Atman Uslu, "Online learning in higher education : Examining the predictors of students' online engagement," *Education and Information Technologies*, vol. 28, no. 2, pp. 1865–1889, 2023.

[283] Y.-M. Cheng, "How gamification and social interaction stimulate moocs continuance intention via cognitive presence, teaching presence and social presence ?," *Library Hi Tech*, no. ahead-of-print, 2023.

[284] R. I. Chang, Y. H. Hung, and C. F. Lin, "Survey of learning experiences and influence of learning style preferences on user intentions regarding mooc s," *British Journal of Educational Technology*, vol. 46, no. 3, pp. 528–541, 2015.

[285] J. Carrera and D. Ramírez-Hernández, "Innovative education in mooc for sustainability : Learnings and motivations," *Sustainability*, vol. 10, no. 9, p. 2990, 2018.

[286] A. Ortega-Arranz, M. L. Bote-Lorenzo, J. I. Asensio-Pérez, A. Martínez-Monés, E. Gómez-Sánchez, and Y. Dimitriadis, "To reward and beyond : Analyzing the effect of reward-based strategies in a mooc," *Computers & Education*, vol. 142, p. 103639, 2019.

[287] C. Reparaz, M. Aznárez-Sanado, and G. Mendoza, "Autorregulación del aprendizaje y retención de mooc," *Computadoras en el comportamiento humano*, vol. 11, no. 1, 2020.

[288] C. Alario-Hoyos, I. Estévez-Ayres, M. Pérez-Sanagustín, C. D. Kloos, and C. Fernández-Panadero, "Understanding learners' motivation and learning strategies in moocs," *The International Review of Research in Open and Distributed Learning*, vol. 18, no. 3, 2017.

[289] C. Milligan, A. Littlejohn, and A. Margaryan, "Patterns of engagement in connectivist moocs," *Journal of Online Learning and Teaching*, vol. 9, no. 2, pp. 149–159, 2013.

[290] C. J. Bonk and M. M. Lee, "Motivations, achievements, and challenges of self-directed informal learners in open educational environments and moocs," *Journal of Learning for Development*, vol. 4, no. 1, 2017.

[291] A. Watted and M. Barak, "Motivating factors of mooc completers : Comparing between university-affiliated students and general participants," *The Internet and Higher Education*, vol. 37, pp. 11–20, 2018.

[292] D. Petronzi and M. Hadi, "Exploring the factors associated with mooc engagement, retention and the wider benefits for learners," *European Journal of Open, Distance and E-Learning*, vol. 19, no. 2, pp. 112–129, 2016.

[293] J. O. Uchidiuno, A. Ogan, E. Yarzebinski, and J. Hammer, "Going global : Understanding english language learners' student motivation in english-language moocs," *International Journal of Artificial Intelligence in Education*, vol. 28, pp. 528–552, 2018.

[294] H. Zhao, F. Jiang, X. Nie, and Q. Chen, "A review of the application of machine learning in personalized education," *IEEE Access*, vol. 7, pp. 110961–110976, 2019.

[295] G. Piccoli, R. Ahmad, and B. Ives, "Web-based virtual learning environments : a research framework and a preliminary assessment of effectiveness in basic it skills training. mis q 25 : 401–426," *How Does Distance Education Compare with Classroom Instruction ? A Meta-Analysis of the Empirical Literature*, vol. 74, 2001.

[296] Y. Sun, L. Ni, Y. Zhao, X.-L. Shen, and N. Wang, "Understanding students' engagement in moocs : An integration of self-determination theory and theory of relationship quality," *British Journal of Educational Technology*, vol. 50, no. 6, pp. 3156–3174, 2019.

[297] I. Maya-Jariego, D. Holgado, E. González-Tinoco, J. Castaño-Muñoz, and Y. Punie, "Typology of motivation and learning intentions of users in moocs : the moocknowledge study," *Educational technology research and development*, vol. 68, pp. 203–224, 2020.

[298] K. F. Hew, "Promoting engagement in online courses : What strategies can we learn from three highly rated moocs," *British Journal of Educational Technology*, vol. 47, no. 2, pp. 320–341, 2016.

[299] B. Arsovic and N. Stefanovic, "E-learning based on the adaptive learning model : case study in serbia," *Sādhanā*, vol. 45, no. 1, p. 266, 2020.

[300] T. Stanley, *Case studies and case-based learning : Inquiry and authentic learning that encourages 21st-century skills*. Routledge, 2021.

[301] K. Jang, J.-H. Lee, and J.-S. Kim, "Association between interpersonal relationships and 4c (communication, critical thinking, collaborative self-efficacy, and creative problem-solving) core competencies of dental hygiene students," *Journal of Korean society of Dental Hygiene*, vol. 21, no. 2, pp. 109–118, 2021.

[302] M. R. Haas, K. Haley, B. S. Nagappan, F. Ankel, A. Swaminathan, and S. A. Santen, "The connected educator : personal learning networks," *The Clinical Teacher*, vol. 17, no. 4, pp. 373–377, 2020.

[303] K. Oddone, "The nature of teachers' professional learning through a personal learning network : Individual, social and digitally connected," *Teaching and Teacher Education : Leadership and Professional Development*, vol. 1, p. 100001, 2022.

[304] J. Salinas and B. De-Benito, "Construction of personalized learning pathways through mixed methods.," *Comunicar : Media Education Research Journal*, vol. 28, no. 65, pp. 31–41, 2020.

[305] H. Peng, S. Ma, and J. M. Spector, "Personalized adaptive learning : an emerging pedagogical approach enabled by a smart learning environment," *Smart Learning Environments*, vol. 6, no. 1, pp. 1–14, 2019.

[306] E. G. Rincón-Flores, M. S. R. Montoya, and J. Mena, "Engaging mooc through gamification : Systematic mapping review," in *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*, pp. 600–606, 2019.

[307] E. Yıldız, "Investigation of factors affecting the sense of community of distance education learners in online learning environments," *Journal of Qualitative Research in Education*, vol. 8, no. 1, 2020.

[308] H. Çelik, N. B. Baykal, and H. N. K. Memur, "Nitel veri analizi ve temel ilkeleri," *Eğitimde Nitel Araştırmalar Dergisi*, vol. 8, no. 1, pp. 379–406, 2020.

[309] T. Ginige and S. T. Vanderwall, "Effective online learning management system to improve and enhance the online learning and student engagement experience. this document contains how the proposed solution, an extension, can help enhance distance learning.," in *Proceedings of the 6th International Conference on Information System and Data Mining*, pp. 95–100, 2022.

[310] K. Saravanan, "Mooc for student learning and active engagement," in *Optimizing Student Engagement in Online Learning Environments*, pp. 127–146, IGI Global, 2018.

[311] M. Ranieri, D. Luzzi, S. Cuomo, and I. Bruni, "If and how do 360 videos fit into education settings ? results from a scoping review of empirical research," *Journal of Computer Assisted Learning*, vol. 38, no. 5, pp. 1199–1219, 2022.

[312] D. Intan Kurniasih, Baedhowi, and Sudiyanto, "The effectiveness of higher order thinking skills (hots) based e-book to improve student learning outcomes," in *Proceedings of the 5th International Conference on Learning Innovation and Quality Education*, pp. 1–5, 2021.

[313] J. A. McMillan, M. Land, and L. K. Leslie, "Pediatric residency education and the behavioral and mental health crisis : a call to action," *Pediatrics*, vol. 139, no. 1, 2017.

[314] L. S. Vygotsky and M. Cole, *Mind in society : Development of higher psychological processes*. Harvard university press, 1978.

[315] M. Gozza-Cohen, "Session 8 : Accessible world — practicing what we preach : Using universal design for learning (udl) principles to teach udl to pre-service special education teachers," in *WCSNE 2015 Proceedings*, Infonomics Society, 2015.

[316] V. Peters, "Meeting learners where they are : Using microsoft forms to drive improvement in learning outcomes." `https://doi.org/10.51388/20.500.12265/52`, 2018.

[317] L. C. de Oliveira, *A Language-Based Approach to Content Instruction (LACI) for Multilingual Learners*, pp. 1–12. Routledge, 2023.

[318] M. S. A. Mohd Sharif and P. Ramakrisnan, "Log data indicators for identifying learner engagement in moocs," *International Journal of Advanced Research in Education and Society*, vol. 5, no. 1, pp. 35–51, 2023.

[319] U. K. Masrom, N. A. Nik Mohd Alwi, and N. H. Nor Asshidin, "The underlying factors of learner readiness and satisfaction in blended learning environment," in *2018 IEEE 6th International Conference on MOOCs, Innovation and Technology in Education (MITE)*, pp. 1–6, IEEE, 2018.

[320] T. P. Hébert, "Sofia creates a supportive classroom environment," in *Guiding Gifted Students With Engaging Books*, pp. 13–20, Routledge, 2021.

[321] J. Chauhan, S. Taneja, and A. Goel, "Enhancing mooc with augmented reality, adaptive learning and gamification," in *2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE)*, pp. 348–353, IEEE, 2015.

[322] F. Schwenkreis, "Coaching support by collecting and analyzing data (cocoanda)," in *Opportunities and Challenges for European Projects*, SCITEPRESS - Science and Technology Publications, 2018.

[323] S. Prediger, "Implementation research as a task for subject-matter education disciplines : Co-constructive, content-related, and research-based," *Research in Subject-matter Teaching and Learning (RISTAL)*, vol. 5, no. 1, pp. 4–23, 2022.

[324] R. AlOkaily, "Evaluations of learning designs," in *Learner-Centered Instructional Design and Evaluation*, pp. 49–112, Routledge, 2023.

[325] A. Amigud, "Post-traditional learning analytics," in *Emerging Trends in Learning Analytics*, pp. 13–25, BRILL, 2019.

[326] A. Shaban and E. Pearson, "A learning design framework to support children with learning disabilities incorporating gamification techniques," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, ACM, 2019.

[327] A. Manciaracina, "How to design learning using technology and users' needs ?," in *Designing Hybrid Learning Environments and Processes*, pp. 93–128, Springer International Publishing, 2022.

[328] M. Bothe and C. Meinel, "When do learners rewatch videos in moocs ?," in *2020 IEEE Learning With MOOCS (LWMOOCS)*, IEEE, 2020.

[329] K. Bastedo and N. Swenson, "General accessibility guidelines for online course content creation," in *Universal Access Through Inclusive Instructional Design*, pp. 208–217, Routledge, 2019.

[330] A. Battye, "Multi-modal communication," in *Navigating AAC*, pp. 49–52, Routledge, 2022.

[331] C. Changchit and K. Bagchi, "Privacy and security concerns with healthcare data and social media usage," *Journal of Information Privacy and Security*, vol. 13, no. 2, pp. 49–50, 2017.

[332] A. Kong, "Understanding mooc learners from different levels of study : An investigation of hospitality and tourism," in *2022 IEEE Learning with MOOCS (LWMOOCS)*, IEEE, 2022.