N° d'ordre :

UNIVERSITE SULTAN MOULAY SLIMANE

Faculté des Sciences et Techniques

Béni-Mellal

Centre d'Études Doctorales : Sciences et Techniques

Formation Doctorale : Mathématiques et Physique Appliquées

THÈSE

Présentée par

OUATIK FAROUK

Pour l'obtention du grade de

DOCTEUR

Spécialité : Informatique

Contribution à la prédiction de la réussite et l'orientation des étudiants à base de E-learning et du Big Data

---

Soutenue le : 02/07/2022

Pr. Mohamed BAHAJ	Professeur, Université Hassan Premier, F.S.T. Settat, Maroc.	Président/Rapporteur
Pr. Mohamed BASLAM	Professeur, Université Sultan Moulay Slimane, F.S.T. Béni-Mellal, Maroc.	Rapporteur
Pr. Rachid EL AYACHI	Professeur, Université Sultan Moulay Slimane, F.S.T. Béni-Mellal, Maroc	Rapporteur
Pr. Hicham ZOUGAGH	Professeur, Université Sultan Moulay Slimane, F.S.T. Béni-Mellal, Maroc	Rapporteur
Pr. Mohammed ERRITALI	Professeur, Université Sultan Moulay Slimane, F.S.T. Béni-Mellal, Maroc.	Co-Directeur de Thèse
Pr.Mostafa Jourhmane	Professeur, Université Sultan Moulay Slimane, F.S.T. Béni-Mellal, Maroc	Directeur de Thèse

Dédicaces

Nous offrons ce travail :

- À nos chers parents,
  Aucune dédicace ne pourait refléter notre amour, notre immense gratitude, car nous ne pourrions jamais oublier la tendresse et l'amour dévoué par lesquels ils nous ont toujours entourés depuis notre enfance. Ils ont toujours présents pour les bons conseils. Leur affection et leur soutien ont un grand secours au long de ma vie professionnelle et personnelle. Qu'ils trouvent dans ce modeste travail notre reconnaissance pour tous leurs efforts.

- Nous dédions aussi ce travail : À mes frères et toute la famille, à tous nos amis, et à toutes les personnes qui nous ont prodigué des encouragements et se sont donné la peine de nous soutenir durant ces ans de formation.

- À nos chers Encadrant Mr. Mostafa Jourhmane, Mr. Mohammed Erritali et nos Professeurs qui doivent voir dans ce travail la fierté d'un savoir bien acquis. À tous les membres de la faculté des Sciences et Techniques de Beni Mellal. Et à tous les chers lecteurs.

Remerciements

Les remerciements sont des marques de politesse, mais insuffisante pour montrer nos gratitudes envers ceux qui nous ont soutenus. Au terme de ce travail, nous tenons à remercier tous ceux qui nous ont aidés et guidés dans la réalisation de ce travail.

Nos remerciements s'adressent plus précisemment à Mr. Mostafa Jourhmane notre encadrant et Mr. Mohammed Erritali mon co-encadrant, qui m'a fait confiance et nous ont honorés avec ce sujet de thèse, leurs encouragements, leurs soutiens, leurs remarques précieuses nous ont été d'une grande utilité.

Nous tenons également à remercier les membres du jury pour l'intérêt qu'ils ont porté à notre travail et aussi pour les remarques cruciaux qu'ils ont faites sur notre sujet. Enfin, merci à tous ceux qui nous ont encouragés de loin et de près pendant la réalisation de notre travail.

Table des matières

Sommaire

Liste des figures

Liste des tableaux

Liste des acronymes

MCDM : Multiple-criteria decision-making

LMS : learning management system

CMS : Content Management System

LCMS : earning Content Management System

SCORM : Sharable Content Object Reference Model

SMOTE : Synthetic Minority Oversampling Technique

TOPSIS : Technique for Order of Preference by Similarity to Ideal Solution

AHP : Analytic Hierarchy Process

KNN : k-nearest neighbors

SVM : support vector machine

SMO : Sequential minimal optimization

mRMR : Maximum Relevance - Minimum Redundancy

Liste des publications

Articles de revues indexs Thomson Reuters & Scopus

1-Farouk ouatik, mohammed erritali, fahd ouatik, mostafa jourhmane "comparative study of mapreduce classification algorithms for students orientation" elseiver procedia computer science 170 (2020) 1192–1197.

2-Farouk ouatik, mohammed erritali, fahd ouatik, mostafa jourhmane " student orientation using machine learning under mapreduce with hadoop'' journal of ubiquitous systems & pervasive networks volume 13, no. 1 (2020) pp. 21-26.

3-Farouk ouatik, mohammed erritali, fahd ouatik, mostafa jourhmane " students' orientation using machine learning and big data '' international journal of online and biomedical engineering (ijoe) – eissn: 2626-8493.

4-Farouk ouatik, Fahd ouatik "Learning Management System Comparison: New Approach Using Multi-Criteria Decision Making". In: Fakir M., Baslam M., El Ayachi R. (eds) Business Intelligence. CBI 2021. Lecture Notes in Business Information Processing, vol 416. Springer, Cham. https://doi.org/10.1007/978-3-030-76508-8_17.

5-Farouk Ouatik, Mohammed Erritali, Fahd Ouatik, Mostafa Jourhmane " Predicting student success using big data and machine learning algorithms " International Journal of Emerging Technologies in Learning (iJET) , Volume 12 (2022).

Participations à des conférences et congrès scientifique international

1. Farouk ouatik, mohammed erritali, fahd ouatik, mostafa jourhmane, "Decision Making System for students' orientation using big data and Data Mining Algorithm", The 5th International Conference on Business Intelligence CBI'19, (2019).

2. Farouk ouatik, mohammed erritali, fahd ouatik, mostafa jourhmane, "comparative study of mapreduce classification algorithms for students' orientation", The 11th International Conference on Ambient Systems, Networks and Technologies (ANT-2020).

3. Farouk ouatik, mohammed erritali, fahd ouatik, mostafa jourhmane, "Learning Management System Comparison: New Approach Using Multi-Criteria Decision Making", The 6th Edition of the International Conference on Business Intelligence CBI'21(2021).

Résumé

Les technologies d'information et de communication, sont récemment introduites dans plusieurs domaines, notamment l'enseignement, pour l'intérêt de renforcer et faciliter le flux de l'apprentissage. E-learning est un type de formation en ligne et à distance. Il utilise l'internet et les nouvelles technologies numériques. Son point fort, il fournit aux apprenants un accès à distance aux différents contenus et des supports d'apprentissage. Et pour promouvoir et consolider la formation des apprenants afin d'avoir de bonnes qualités et de bonnes connaissances dans leurs domaines. Dans le cadre de cette thèse, notre but, est de profiter les avantages et les données issues des plateformes E-Learning, afin de construire un modèle et un concept plus efficace, pour arriver à l'objectif principal de l'enseignement qui est l'apprentissage et la réussite de l'étudiant.

La réussite de l'étudiant dépend de plusieurs paramètres, notamment l'orientation scolaire. C'est une opération primordiale et aussi une étape décisive pour l'étudiant. Elle dirige le parcours scolaire d'une manière positive ou négative. Il existe plusieurs facteurs qui influencent le choix de la spécialité, à savoir les notes de l'étudiant, le nombre d'absence, selon les matières, le désir, les penchants de l'étudiant, etc. Dans le cadre de cette thèse, pour mettre en évidence et analyser l'effet de ces facteurs, nous avons récupéré les activités et les comportements des étudiants à partir des plateformes du E-Learning : Moodle, Sakai, Claroline, Easy LMS, OpenEdx, TalentLMS. De plus, au cours de cette thèse, nous avons montré que ces plateformes ne fournissent pas les mêmes propriétés, ainsi que le choix de la plateforme est lié à l'objectif de l'utilisation de cette dernière. Pour trouver la plateforme adéquate à notre objectif, on a comparé les plateformes précédentes en utilisant la méthode MCDM avec un ensemble des critères.

En outre, vu le nombre progressif des étudiants, le nombre des spécialités, ainsi que la diversité des sources de données, les méthodes de stockage et de traitement utilisé par les spécialistes de l'enseignement et de l'apprentissage, sont limitées. Elles ne peuvent pas rendre les services de l'orientation scolaire et la prédiction de réussite ou l'échec de l'étudiant en temps réel. La solution

proposée dans cette thèse, est l'utilisation de la technologie Big Data. Cette technologie permet le stockage distribué des données ainsi que le traitement des données qui se fait d'une manière parallèle à l'aide des algorithmes de Data Mining sous Map Reduce. Les algorithmes utilisés sont : Les réseaux de neurones, les réseaux bayésiens, le K-plus proche voisins, les forêts aléatoires et les SVMs.

Les mots clés : MCDM, Big Data, E-Learning, les réseaux de neurones, les réseaux bayésiens, le K-plus proche voisins, les forêts aléatoires, les SVMs, E-learning, M-learning, D-learning, l'orientation scolaire, la réussite scolaire.

Abstract

Information and communication technologies have recently been introduced in a number of fields, notably teaching, in order to strengthen and facilitate the flow of learning. E-learning is a type of online and distance learning. It uses the internet and new digital technologies. Its strength is that it provides learners with remote access to various content and learning materials. And to promote and consolidate the training of learners in order to have good qualities and good knowledge in their fields. Within the framework of this thesis, our goal is to take advantage of the advantages and the data from the E-Learning platforms, in order to build a more efficient model and concept, to achieve the primary goal of teaching which is the student's learning and success.

Student success depends on several parameters, including academic orientation. It is a crucial operation and also a decisive step for the student. It directs the academic path in a positive or negative way. There are several factors that influence the choice of specialty, namely the student's grades, the number of absences, according to the subjects, the desire, the inclinations of the student, etc. In the context of this thesis, to highlight and analyze the effect of these factors, we have recovered the activities and behaviors of the students from the platforms of E-Learning: Moodle, Sakai, Claroline, Easy LMS, OpenEdx, TalentLMS. In addition, during this thesis, we showed that these platforms do not provide the same properties, as the choice of the platform is related to the purpose of the use of the latter. To find the right platform for our goal, we compared the previous platforms using the MCDM method with a set of criteria.

In addition, given the increasing number of students, the number of specialties, as well as the diversity of data sources, the storage and processing methods used by teaching and learning specialists are limited. They cannot render the services of academic guidance and the prediction of success or failure of the student in real time. The solution proposed in this thesis, is the use of Big Data technology. This technology allows the distributed storage of data as well as the processing of data

that is done in a parallel way using the algorithms of Data Mining under Map Reduce. The algorithms used are: Neural networks, Bayesian networks, neighboring K-nearest, random forests and SVMs.

Keywords: MCDM, Big Data, E-Learning, neural networks, Bayesian Networks, K-nearest neighbors, Random Forest, SVMs, E-learning, M-learning, D-learning, academic orientation, academic success.

Introduction générale

Le succès ou la réussite de l'étudiant dans les évaluations, est l'objectif crucial de l'éducation et de l'enseignement. Il reflète le niveau de compréhension, et la maitraise des compétenses par l'étudiant, d'une part, et d'autre part, le développement du niveau d'enseignement dans les établissements. Pour améliorer les méthodes de l'enseignement et de l'apprentissage et aussi les compétences des étudiants. Il faut que ces méthodes prennent en considération, le développement des technologies d'information et de communication.

Parmi les méthodes de l'enseignement et de l'apprentissage intégrant les technologies d'information, on trouve « E-Learning ». Ce dernier met en évidence son importance sur la continuité de l'enseignement après l'apparition du virus Corona. Parce qu'il n'est pas lié à l'existence physique des étudiants et de l'enseignant dans le même lieu d'enseignement et en même temps. Il représente la clé d'ouverture de l'enseignement sur la technologie.

Le développement de l'enseignement n'est pas limité à la recherche des nouvelles méthodes de l'enseignement seulement, il existe d'autres aspects de recherche pour mettre en valeur l'enseignement. Ce qui nous a poussé dans cette thèse à chercher des méthodes pour la prédiction du succès scolaire de l'étudiant et les facteurs influençant, et aussi d'améliorer la qualité d'enseignement et d'apprentissage.

Afin d'augmenter le taux de la réussite des étudiants, cette thèse aussi propose des modèle pour l'amélioration de la qualité de l'orientation scolaire des étudiants. Elle détermine les facteurs qui influencent le choix de la spécialité, ainsi elle propose un système de recommandation d'orientation scolaire pour la spécialité adéquate à l'étudiant. Ce système propose aussi la réorientation au cas où il est apparu que l'étudiant rencontre des problèmes. La recommandation de la spécialité se fait par des algorithmes du Data Mining, et vu le grand nombre de paramétres et les propriétés utilisées par le système, la solution trouvée est de rendre le stockage des données distribuées et le traitement parallèle de ces données à l'aide de la technologie Big Data.

Le plan de cette thèse est comme suit:

- Le premier chapitre est consacré à l'état de l'art sur le E-Learning. Ce chapitre présent les définitions du E-Learning ainsi que ses objectifs et son historique, il présente aussi la différence entre E-Learning, D-Learning et M-Learning. Puis, les facteurs du succès et l'application de E-learning au Maroc.
- Dans le deuxième chapitre, nous allons présenter les systèmes de gestion d'apprentissage, ainsi que les plateformes de E-Learning. Puis, nous allons comparer les plateformes de E-Learning à base de la méthode MCDM, pour sélectionner la plateforme adéquate à l'orientation scolaire et la prédiction de la réussite des étudiants.
- Au cours du troisième chapitre, nous allons présenter le système d'orientation scolaire et les facteurs qui l'influencent. Le système d'orientation scolaire réalisé est basé sur la méthode SMOTE pour l'équilibrage des données qui sont utilisées par la méthode TOPSIS et la méthode AHP. Pour appliquer la méthode TOPSIS et la méthode AHP, nous avons besoin des poids de chaque critère utilisé. Nous avons calculé ces poids par le gain d'information des attributs (une hybridation des deux méthodes précédentes avec le gain d'information).
- Le quatrième chapitre présente la technologie Big Data et ses caractéristiques ainsi que ses bénéfices. Il présente aussi les outils du Big Data, à savoir Hadoop et son architecture. A la fin du chapitre, les aspects du Big Data en E-Learning sont présentés.
- Le cinquième chapitre présente notre système d'orientation scolaire, mais cette fois-ci, notre système est basé sur la technologie Big Data, pour résoudre le problème de la diversité des sources des données, et pour minimiser le temps d'exécution. Parce que, le nombre des étudiants est en progrès et aussi le nombre des spécialités d'une part, et de l'autre part, les utilisateurs du système (les étudiants et les administrateurs) ont besoin de la décision en temps réel.
- Enfin, le sixième chapitre décrit notre système de la prédiction de la réussite scolaire des étudiants à base de la technologie Big Data et les algorithmes de Data Mining. Il présente les attributs utilisés pour la prédiction de la réussite scolaire des étudiants à l'aide des méthodes

de sélection des attributs et le modèle utilisé ainsi que les algorithmes de classification adoptés pour la prise de décision.

- Ce rapport se termine par une conclusion générale et des perspectives tracées.

1. Chapitre I : Etat de l'art sur E-learning

1.1 Introduction

Le terme « e-learning » existe depuis 1999, il est utilisé la première fois dans un séminaire CBT Systems à la ville américaine, Los Angeles. Il est associé au terme, l'apprentissage en ligne et le terme, l'apprentissage virtuel. Il désigne la manière d'apprentissage, basée sur l'utilisation de nouvelles technologies, qui permet aux utilisateurs d'accéder à des formations en ligne, soit interactive ou personnalisée à travers l'internet ou à travers les autres médias électroniques (extranet, intranet, télévision interactive, CD-ROM, etc.), pour développer et améliorer leurs compétences. Le processus d'apprentissage n'est pas lié au temps ou au lieu [1]. Il existe des définitions qui considèrent que E-Learning n'utilise pas l'internet seulement comme un support technique pour l'apprentissage, mais aussi il utilise l'intranet, la télévision interactive, la diffusion de conférences par satellite, les disques audio et les disques vidéo, les appareils mobiles et sans fil [2]. Et aussi il y a des définitions qui traduisent E-Learning par l'apprentissage sur Internet, du fait qu'il utilise principalement les technologies internet pour la réalisation, l'utilisation, l'adoption, le transfert et la facilitation d'apprentissage [3].

Dans ce chapitre, nous avons présenté les objectifs et l'historique de E-learning. Puis nous avons comparé D-learning, E-learning et M-learning. Et après, nous avons cite les avantages et les inconvénients de E-learning. Ensuite nous avons présenté la méthode d'intégration du Big data en E-learning. Enfin, nous avons montré les facteurs du succès du E-learning et E-learning au Maroc.

1.2 Objectifs du E-Learning

E-Learning a plusieurs objectifs à savoir :

- L'individualisation et l'adaptation du contenu d'apprentissage aux étudiants ;
- Il rend l'apprentissage et l'enseignement en temps réel et augmente leur qualité ;

- E-Learning peut être synchrone ou asynchrone, comme il peut être en classe ou en dehors de la classe, comme il peut être mixte ;
- Réponds aux besoins des étudiants ;
- Adaptable au style d'apprentissage ;
- Améliorer l'efficacité et l'efficience d'apprentissage et d'enseignement ;

1.3 Historique de E-Learning

Le saut qualitatif de la technologie a créé de grands défis dans le domaine de l'éducation. Comme cela a conduit à l'invention de nouveaux modes d'éducation. Les méthodes d'enseignement à distance ne sont pas nouvelles aujourd'hui, mais existaient plutôt avant l'apparition d'internet. Là où il y avait des cours à distance pour les étudiants, par exemple. Isaac Bateman a enseigné la sténographie à ses étudiants dans les années 1840 par correspondance. Comme les élèves de Bateman utilisaient le courrier pour envoyer leurs devoirs, il leur envoyait également d'autres travaux à faire. Aussi, en 1924, une machine a été inventée pour tester les connaissances des étudiants.

1.4 Comparaison de Distance Learning, Electronique Learning et Mobile Learning

D'abord. Il faut différencier entre Distance Learning (D-Learning), Électronique Learning (E-Learning) et Mobile Learning (M-Learning).

L'enseignement et l'apprentissage s'effectuaient selon le mode traditionnel en face-à-face (F2F), il était considéré comme le pilier du système éducatif. Avec les nouvelles technologies, le modèle d'enseignement et d'apprentissage est passé de l'enseignement traditionnel à l'enseignement à distance. Le système d'éducation est dans une période de la recherche du nouveau système d'apprentissage pour passer du modèle traditionnel aux modèles D-Learning, E-Learning et M-Learning.

1.4.1 Définition de Distance Learning (D-Learning)

Le concept d'enseignement à distance n'est pas nouveau, puisqu'il a remplacé les cours par correspondance [4], [5]. La méthode de correspondance est transformée à d'autres méthodes avec l'apparition de la télévision et la radio éducatives.

L'apprentissage à distance est un processus éducatif dans lequel les apprenants et les enseignants sont séparés par le lieu et le temps, ou les deux. Mais Valentin [5] a lié l'enseignement à distance à la séparation entre les étudiants et l'enseignant par lieu, mais pas forcément par le temps.

L'apparition de la technologie de l'information et aussi la communication via internet ont permis de passer à E-Learning (le mode électronique).

1.4.2 Défintion de Electronic Learning (E-Learning)

La révolution de l'information, ou la révolution électronique (e-révolution), a eu un impact presque sur tous les domaines social, gouvernemental, commercial et économique. L'introduction d'un environnement électronique, tel que E-Learning, le courrier électronique, les services sociaux, le commerce électronique et les services bancaires en ligne, a apporté de nouvelles opportunités avec plus de services et la facilité des taches [6]. Ce qui augmente la demande de e-Learning dans les universités, et aussi augmente la rapidité de propagation de E-Learning, il est devenu très populaire [7]. Il y a aujourd'hui pas mal d'universités qui sont uniquement virtuelles et qui proposent des cours à l'aide du e-Learning.

1.4.3 Défintion de Mobile Learning (M-Learning)

Le M-Learning est une extension de e-Learning qui utilise des appareils d'apprentissage portables et mobiles [8].

Le m-Learning est une nouvelle période du e-Learning qui permet un apprentissage à tout moment et de partout par des appareils mobiles [9]. C'est un type d'apprentissage très flexible.

L'appareil mobile peut vous emmener virtuellement dans la salle de classe .Tels que les Smartphones, les appareils GPS .Ces appareils ont annulé la salle de classe en simplifiant la communication entre les

instructeurs et les étudiants. Maintenant, les élèves ont plus de contrôles sur leurs outils d'apprentissage et aussi leurs besoins éducatifs.

1.4.4 Difference entre le D-Learning, E-Learning et M-Learning

Le E-Learning, l'apprentissage à distance, le 'web-based Learning' et l'apprentissage en ligne sont utilisés de manière interchangeable [10]. Mais pour D-Learning, E-Learning et M-Learning, il y a une relation entre eux, le terme e-Learning est utilisé pour indiquer tous type d'apprentissage qui utilise les technologies. Le d-Learning est un type d'apprentissage dans lequel les apprenants et les enseignants sont séparés par le lieu ou le temps, ou les deux.

La figure suivante présente la relation entre e-Learning, m-Learning et d-Learning.

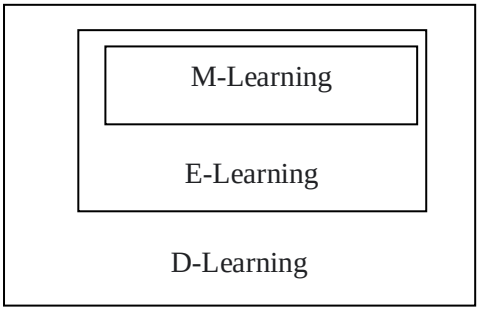

Figure 1.1 La relation entre M-Learning, E-Learning et D-Learning avant les inventions technologiques.

D'après la figure 1.1, M-Learning appartient à E-Learning et lui aussi est inclus à D-Learning [9]. Avec les inventions technologiques rapides le D-Learning se transforme progressivement en E-Learning, ce qui donne la relation exprimée par la figure suivante.

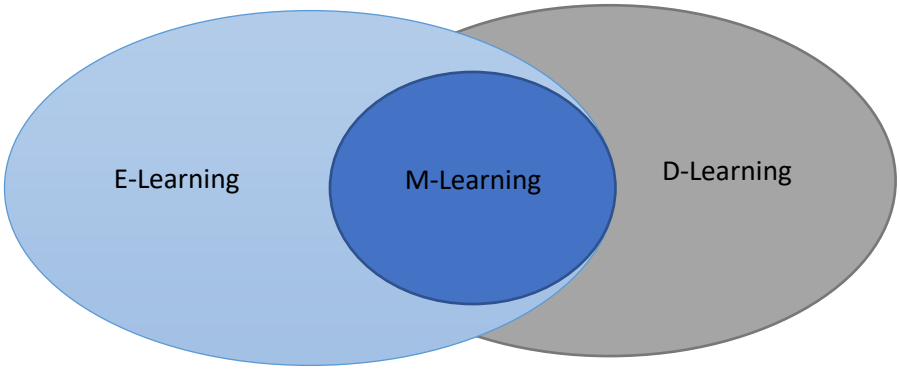

Figure 1.2 La relation entre M-Learning, E-Learning et D-Learning après les inventions technologiques.

1.4.5 Perspective de D-Learning, E-Learning et M-Learning

D-Learning, E-Learning et M-Learning ont tous des perspectives fondamentaux, la perspective contextuelle, la perspective émotionnelle, la perspective cognitive et la perspective comportementale.

Pour M-Learning, il a trois perspectives fondamentales : la mobilité de l'apprenant, la mobilité de l'apprentissage et la mobilité de la technologie. Pour D-Learning, il a aussi trois perspectives, telles que, l'enseignement, le contenu numérique et la technologie. Enfin, les perspectives de E-Learning, qui a quatre perspectives fondamentales qui sont interdépendantes et aussi nécessaires pour transformer les appareils électroniques à des instruments pour servir les établissements d'enseignement.

1.4.5.1 La description des perspectives fondamentales du E-Learning :

Perspective cognitive :

D'après [11], la perspective cognitive s'intéresse aux processus cognitifs qui participent à l'apprentissage et comment le cerveau fonctionne. Pour appliquer les modèles pédagogiques cognitifs dans l'environnement de E-Learning, on peut utiliser les systèmes d'apprentissage intelligent (smart learning system) et la technologie d'apprentissages adaptatifs, afin d'optimiser les progrès des étudiants. Le monde virtuel où les simulateurs et les autres environnements d'apprentissage peuvent aider les étudiants dans le contenu. Le système de support peut guider rapidement les étudiants et les aide à apprendre et de communiquer. Les outils collaboratifs et sociaux peuvent utiliser pour promouvoir l'interaction, le dialogue et l'apprentissage par procuration chez les étudiants [12].

Perspective émotionnelle :

Pour la perspective émotionnelle, elle se base sur l'engagement, la motivation, et aussi sur d'autres parties émotionnelles de l'apprentissage. Alors que [13], déclare diverses émotions, telles que, la peur, la fierté, l'anxiété, le soulagement, la frustration, la résistance, l'attente, la confiance, le désespoir, et l'envie. Il les lie avec l'intégration de la cognition et de l'action et aussi de la motivation.

Perspective comportementale :

D'après [11], la perspective comportementale s'intéresse aux compétences et aux résultats comportementaux de l'apprentissage. Et selon [14], la perspective comportementale s'intéresse aussi au jeu de rôles et la mise en place du milieu de travail [14].

Perspective contextuelle :

La perspective contextuelle se base sur les parties environnementales et les parties sociales qui stimulent l'apprentissage [14], et aussi elle se base sur la réaction des gens, le soutien, l'apprentissage par les pairs, ainsi que la collaboration.

La figure suivante présente les perspectives fondamentales du E-Learning.

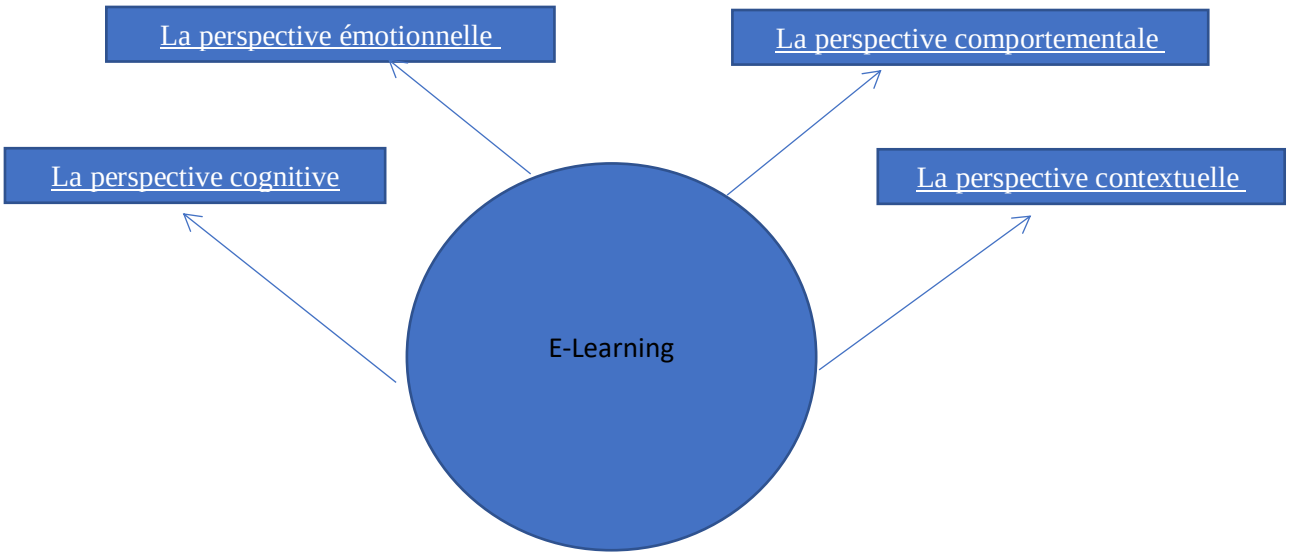

Figure 1.3 Les perspectives fondamentales du E-Learning.

1.4.5.2 La description des perspectives fondamentales du M-Learning :

La mobilité de la technologie :

Selon [15], la technologie mobile se base sur les téléphones portables qui permettent à leurs utilisateurs (les apprenants) d'accéder aux contenus éducatifs. Ces téléphones portables donnent plusieurs services, tels que, e-mail, SMS, MMS, WAP, GPRS, et les autres outils de communication qui sont très nombreux.

La mobilité des apprenants :

A l'apparition du M-Learning, plusieurs possibilités d'apprentissage sont apparues, qui rendent l'apprentissage illimité à un endroit ou un temps spécifique. À l'aide de M-Learning, les apprenants peuvent interagir avec leurs collègues et aussi avec leurs éducateurs qui sont à n'importe quels endroits. Le M-Learning n'est pas destiné à l'éducation formelle seulement, mais aussi à l'éducation semi-formelle et informelle.

La mobilité de l'apprentissage :

Grâce à La mobilité de l'apprentissage les étudiants peuvent développer ses compétences sociales, professionnelles et aussi interculturelles. D'après [16], les appareils mobiles rendent l'apprentissage très spécial, car l'étudiant peut envoyer, traiter et recevoir des données dans leur contexte où il se trouve. C'est-à-dire que le contexte est individuel.

La figure suivante présente les perspectives fondamentales du M-Learning.

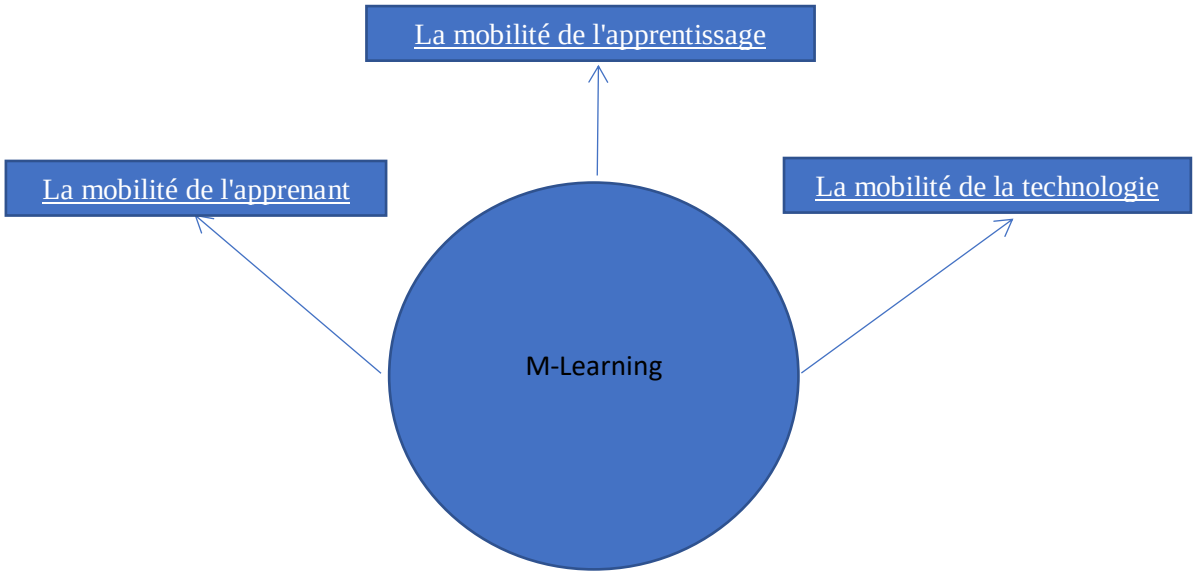

Figure 1.4 Les perspectives fondamentales du M-Learning.

1.4.5.3 La description des perspectives fondamentales du D-Learning :

La technologie :

Les outils et les mécanismes technologiques sont utilisés pour présenter le contenu cognitif de la part des éducateurs et aussi pour recevoir le contenu de la part des étudiants. La technologie contient aussi l'accès à Internet ainsi que les ordinateurs, IPad et les smartphones.

Le contenu numérique :

Les cours en PDF et en Powerpoint, ne sont pas les seuls contenus numériques, mais c'est tous les matériaux académiques qui sont fournis grâce à la technologie.

L'instruction :

L’éducation numérique ne peut pas remplacer les éducateurs. Ces derniers sont nécessaires pour le fonctionnement de l'éducation numérique. Dans l’éducation numérique le rôle de l’éducateur est important, c’est lui qui dirige les étudiants et donne les conseils à ses élèves et qui font le suivi de leurs travaux. Et le rôle le plus important que l’éducation numérique ne peut pas le faire, c’est le soutien personnel pour que les étudiants restent sur la bonne route et continuent leurs apprentissages.

La figure suivante présente les perspectives fondamentales du D-Learning.

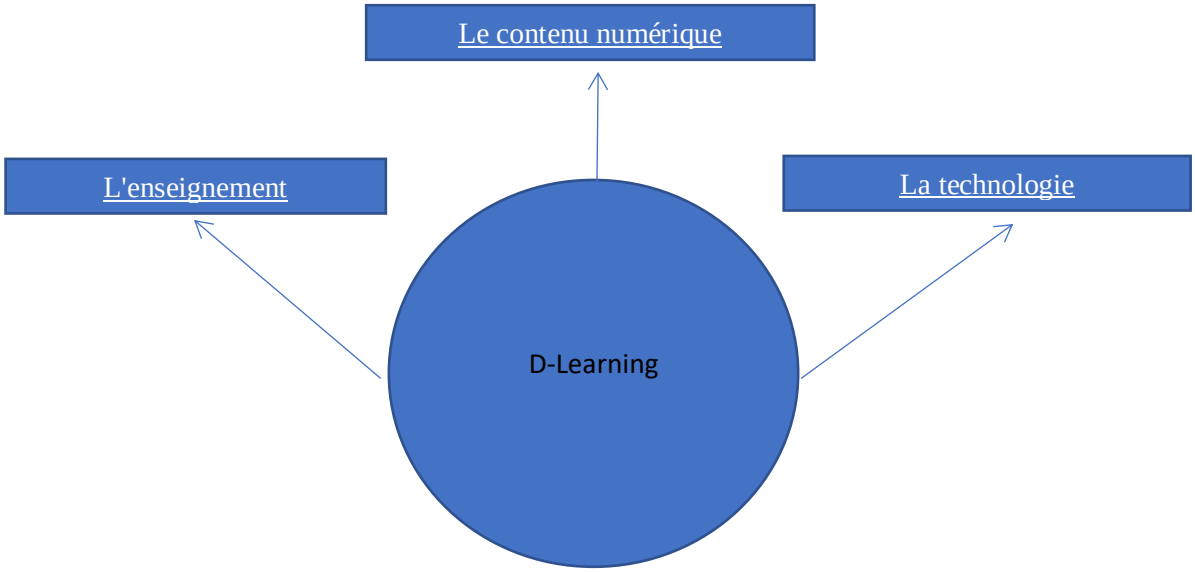

Figure 1.5 Les perspectives fondamentales du D-Learning.

1.5 Avantages et inconvénients de E-Learning

L'apprentissage en ligne a des avantages et des inconvénients. Les avantages de l'apprentissage en ligne comprennent :

- L'apprentissage en ligne ne dépend pas à l'emplacement physique ;
- Il est très rapide et aussi très souple, parce qu’il permet aux étudiants de sauter (passer) le contenu qui est déjà connu ;

- E-Learning permet aux étudiants de personnaliser et de gérer leur matériel en fonction de leurs besoins. Ce qui offre aux étudiants un contrôle sur leurs apprentissages, et une bonne compréhension, et enfin un apprentissage plus efficace ;
- Il est adapté au rythme de l'étudiant ;
- Il est moins cher à livrer ;
- La mise à jour du matériel en ligne se fait facilement et rapidement, parce qu'elle se fait côté serveur ;
- L'apprentissage des étudiants peut être à tout moment et à tout endroit ;
- La possibilité de la gestion de plusieurs classes ;
- E-Learning permet l'interaction entre les enseignants et les étudiants, grâce au forum de discussions et aussi les messageries instantanées ;

Le E-Learning a aussi des inconvénients tels que :

-L'inconvénient majeur de E-Learning, est l'absence d'interaction sociale ou la partie informelle de la communication F2F ;

-La réalisation d'un nouveau système est très couteuse au départ ;

-Dans le E-Learning, l'étudiant peut devenir frustré ou confus ;

-La nécessité du temps et de l'argent afin de Développer de nouveaux cours ou formations ;

1.6 Facteurs du succès du E-Learning

E-Learning a plusieurs défis et plusieurs problèmes, car c'est une nouvelle méthode d'apprentissage. Pour réussir E-Learning de nombreux facteurs intervient lors de la planification de E-Learning, qu'il faut être simplifiés pour réussir et voilà quelques-uns :

- **La motivation de l'apprenant :**

Parmi les importants facteurs qui influencent l'apprentissage, on trouve la motivation, qui est la clé du succès de tous les types d'apprentissage, que ce soit en face à face qui nécessite un effort élevé de la part de l'étudiant ou en E-Learning, qui est aussi, demande de la part des étudiants de développer,

d'améliorer et de maintenir leurs niveaux de motivation par eux-mêmes, et d'être prêt aux moments difficiles au cours de la recherche de l'information, sans l'existence du professeur devant eux. Selon [17]] et [18] E-Learning est très attrayant et motive les étudiants à apprendre et à étudier plus que l'apprentissage en face à face. Cela peut être lié au type de matériels utilisés et la façon de communication utilisée par les professeurs. En outre, il y a l'exhortation des parents de leurs enfants à travailler dur et aussi le suivi de leur enfant.

- **Les attitudes de l'apprenant**

La technologie peut être utilisée pour s'amuser. Aussi les étudiants utilisent des appareils pour communiquer et pour utiliser les différentes applications. Mais, est ce que ces étudiants utilisent ces appareils de façon satisfaisante.

L'attitude des étudiants peut être un facteur pour réussir l'expérience de E-Learning .La familiarisation de nouvelle génération des étudiants avec ces technologies poussent les éducateurs et aussi les concepteurs des programmes scolaires d'en tirer profit et de découvrir leurs méthodologies requises.

- **La Technologie**

Pour arriver aux objectifs du E-Learning, la disponibilité du matériel est requise du côté étudiant et aussi du côté éducateurs. Les étudiants et les éducateurs doivent avoir à la fois d'un ordinateur ou d'un smartphone et les logiciels requis pour l'apprentissage et aussi un accès à internet avec un bon débit, pour apprendre/enseigner sans collision et pour créer une interaction positive entre les éducateurs et les étudiants. L'objectif important que les apprenants doivent atteindre est la capacité de concentrer et de suivre et s'en tenir à leur apprentissage.

- L'état de préparation du matériel

Dans le contexte du E-Learning, le professeur rencontre un autre défi, c'est la capacité et la possibilité de développer à temps le matériel. Le design et le type du matériel sont importants et aussi la formation nécessaire.

- **Le contexte d'apprentissage**

Le niveau de préparation des étudiants, des professeurs et de l'environnement contextuel (les parents et la société) sont très importants. Le facteur social a un grand impact sur les résultats de E-Learning et aussi le soutien des étudiants fait une grande différence. À la différence entre l'enseignement en face à face, E-Learning exigent les éducateurs de prendre en considération ce nouveau style d'apprentissage et d'enseignement et le développe. D'après [19], les éducateurs doivent être conscients de tous les changements, ainsi il faut qu'ils travaillent sur leurs compétences et leurs aptitudes pour arriver grâce à E-Learning à un niveau très élevé d'enseignement. Du côté des étudiants, ils ont obligé de modifier leur style d'apprentissage au sein de E-Learning. Les compétences autodidactes sont très précieuses. E-Learning exige aussi une grande force de la part des étudiants eux-mêmes et aussi les parents, les éducateurs et les prestataires de l'éducation.

1.7 E-Learning au Maroc

Le virus Corona a eu un grand impact sur l'accélération de l'adoption de l'enseignement à distance dans tous les pays, ce qui a incité ceux qui s'intéressent à ce domaine à lui accorder une grande importance et priorité .Car de nombreuses applications, plates-formes et outils ont été créés pour augmenter la valeur de ce type de l'éducation et parmi les plateformes, on trouve:

- **Telmidtice :**

En 2020, le ministère de l'Education nationale a lancé l'application Telmidtice, qui comprend les cours approuvés selon les niveaux et dans le même ordre. Les parents peuvent suivre le niveau de leurs enfants. Cette application a été approuvée afin d'assurer la continuité de l'enseignement et la sécurité des apprenants et des cadres pédagogiques.

- **Site emadrassa :**

C'est un site non gouvernemental, mais il a été soutenu par le ministère de l'éducation nationale. Ce qui caractérise ce site, c'est l'inclusion des universités et les écoles marocaines en plus des possibilités d'étudier à l'étranger, ce qui permet aux bacheliers de choisir plus facilement l'université et l'institut qui leur conviennent.

Il existe également d'autres plateformes, à savoir : Dorouss.ma et academy.tamkin.org.

Les inconvénients de l'enseignement à distance au Maroc :

L'éducation a de nombreux avantages et une grande importance, mais elle souffre d'un ensemble de contraintes et d'obstacles. Ces contraintes peuvent être classées en 7 catégories : contraintes humaines, contraintes administratives, contraintes méthodologiques et didactiques, contraintes techniques, contraintes technologiques, contraintes d'infrastructure et contraintes financières.

- **Contraintes humaines :**

La difficulté à traiter et à utiliser les moyens technologiques, les techniques de communication, ainsi que les logiciels, en raison du manque d'étudiants et de professeurs ayant la formation qui leur permettant d'utiliser ces moyens de manière simple et sans effort.

- Les enseignants sont habitués à la manière traditionnelle de dispenser les cours. Le professeur étant la seule source de connaissance et il est le centre du processus d'enseignement-apprentissage, il se base sur la méthode d'endoctrinement et de la nécessité de l'assiduité ainsi que de la mémorisation. Et c'est ce qui rend difficile au professeur le passage à E-Learning, qui considère l'étudiant est au centre du processus d'enseignement-apprentissage, et c'est lui qui construit son propre apprentissage. Ce qui lui donne les mécanismes d'analyse et de discussion, ainsi que la créativité, tandis que le professeur est un facilitateur et un guide pour l'étudiant.

- E-Learning peut devenir un obstacle en soi au processus d'apprentissage, car certains apprenants peuvent devenir introvertis et ils ont besoin d'une véritable confrontation que les apprenants gagnent dans l'enseignement en face à face lorsqu'ils sont face à face avec leurs professeurs et collègues.

- La nécessité d'avoir un nombre important de spécialistes du E-Learning, afin d'améliorer et de maintenir le système, ainsi que de suivre les différents processus du début jusqu'à la fin, et de communiquer entre eux et aussi avec les professeurs afin d'augmenter la valeur et les performances du E-Learning.

- La présence de certains enseignants et responsables qui n'acceptent pas le changement ou n'ont pas le courage de changer leur travail ou leur méthode d'enseignement. Il y a aussi un type d'enseignant qui n'aime pas se soucier du changement ou se désintéressent. Les parents ont peur de la déviation de leurs enfants de la bonne voie et la possibilité d'utiliser le réseau d'informatique à des fins non éducatives, ce

qui les empêche de laisser les moyens technologiques entre les mains de leurs enfants. Les parents n'ont pas suffisamment de temps pour rester avec leurs enfants pendant qu'ils utilisent les moyens de communication.

- Les appareils électroniques peuvent distraire les élèves et disperser leurs pensées.

- Certains enseignants craignent de la possibilité de publier sur l'internet les erreurs qu'ils ont commises durant la présentation de leurs cours par les élèves ou leurs parents.

- Le manque de communication et d'interaction avec les enseignants.

- Le manque d'aide parentale pour leurs enfants, que ce soit par manque de temps du fait de leurs conditions de travail ou d'analphabétisme, notamment en milieu rural, ou encore par l'indifférence des parents vis-à-vis de l'éducation de leurs enfants, car ils ont besoin de leurs enfants pour faire paître leurs animaux ou les aider dans le champ.

- **Contraintes administratives :**

Le manque de connaissances et de compétences de base de certains directeurs d'établissements d'enseignement pour utiliser les moyens technologiques et les programmes informatiques, ce qui rend le directeur incapable d'assister les professeurs, alors qu'ils font face à un problème avec les applications et les programmes.

- Manque de sensibilisation des chefs de département à l'importance et à la signification de ce type d'apprentissage.

- **Contraintes systématiques et didactiques :**

- La non-accréditation des cours et des programmes d'apprentissage en ligne et le manque de suivi des développements technologiques.

- Il n'existe pas de méthode efficace d'évaluation en E-Learning, ce qui rend le processus d'évaluation difficile lors de l'adoption de ce type d'enseignement.

- L'absence d'un plan et d'une vision clairs en matière de E-Learning par le Ministère de l'Éducation Nationale et de l'Enseignement Supérieur, de la recherche scientifique et de la formation des cadres. Aussi, le manque de production de livres électroniques, des médias interactifs et des logiciels par les experts.

- Les théories d'apprentissage n'ont pas renouvelés ou développés pour suivre le rythme de ce style éducatif.

- **Contraintes artistiques :**

Le manque d'entreprises et le manque des techniciens capables de réparer les appareils électroniques ou les applications utilisées dans ce type d'enseignement. Et aussi, le coût financier de ces entreprises est également important.

- **Contraintes technologiques :**

Le manque d'appareils utilisés dans ce type d'enseignement pour de nombreux élèves, en plus de la connexion internet, rend difficile le suivi des cours. Car le pourcentage d'indisponibilité de ces dispositifs est très élevé pour les zones rurales, puisque ce pourcentage atteint 55% pour l'enseignement primaire, 54% pour l'enseignement préparatoire, 41% pour l'enseignement secondaire et 29% pour l'enseignement supérieur.

- Le débit d'internet est faible, ce qui entraîne parfois des interruptions. De plus, certaines zones n'ont pas de couverture mobile et d'internet.

- Les sites Web officiels cessent souvent de fonctionner en raison de la grande pression exercée sur eux.

- Malgré la mise en disposition de la plateforme d'enseignement Tilmidetice par le ministère de l'éducation, et la contribution des chaînes de télévision au processus d'enseignement à distance, les réseaux sociaux sont les plus utilisés par les apprenants, puis les chaînes de télévision.

- Les statistiques indiquent que le téléphone portable est le plus utilisé par les apprenants, mais le téléphone portable pose de nombreux problèmes. Parmi eux :

- Il y a beaucoup d'étudiants qui ne peuvent pas acheter un smartphone, en raison de son prix et aussi l'écran du téléphone est petit, qui ne permet pas d'afficher suffisamment d'informations, ce qui rend la lecture difficile sur les téléphones portables.

- La capacité de stockage du téléphone portable est faible.

- La nécessité de recharger le téléphone en permanence.

- Le coût de l'internet mobile est élevé pour les parents d'élèves, en particulier dans les zones rurales.

- Le clavier du téléphone mobile est petit, ce qui rend difficile de saisir les informations, c'est-à-dire que le téléphone portable est inconvenable à l'apprentissage.

- **Contraintes liées aux infrastructures :**

La faiblesse des infrastructures qui facilitent l'utilisation de l'E-Learning. Elles sont quasi inexistantes dans les zones rurales, comme l'internet, l'électricité, les téléphones portables et les ordinateurs.

- **Contraintes financières :**

- Le coût de la technologie est très élevé.

- L'incapacité à suivre le rythme du développement rapide de la technologie.

- La formation des cadres spécialisés dans ce type d'apprentissage nécessite un budget important.

1.8 L'intégration du Big Data en E-Learning

1.8.1 Définition du Big Data

Le terme « Big Data » ou les données massives, décrit l'ensemble des tâches suivantes : la collecte des données de tailles très grandes, le traitement et l'analyse de ces données puis la visualisation. Mais le terme du Big Data a plusieurs définitions [20, 21, 22]. Il regroupe une gamme de données, d'applications et de technologies. En E-Learning, Big Data désigne l'ensemble des données qui sont générés par les apprenants lorsqu'ils prennent des cours en E-Learning et lors de leurs interactions avec le système d'apprentissage.

Pour bien définir le terme Big Data, il faut savoir l'évolution du Big Data et ses 17 caractéristiques.

L'informatisation des services augmente le volume des données utilisables de plus en plus et rend l'organisation, le traitement et la gestion des données par les logiciels limités à certains volumes de données. Ces volumes des données sont inclus en Big Data. Ces volumes des données sont en augmentation à chaque seconde à travers le monde en parallèle avec l'augmentation de la capacité de stockage des ordinateurs qui est arrivé aujourd'hui à 35 ZB.

La figure 1.6 montre la chronologie des méthodes de gestion des données.

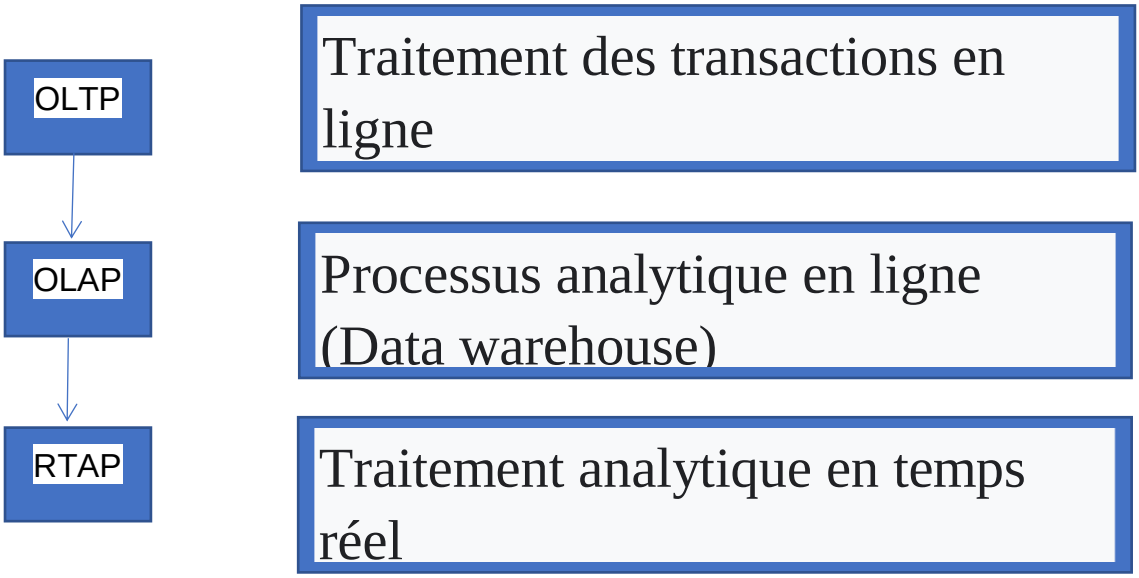

Figure 1.6 La chronologie des méthodes de gestion des données.

Le terme Big Data est très vague. Il existe des définitions formelles du Big Data [23][24][25], et aussi, il existe des définitions informelles du Big data [26][27], mais le problème est la grande différence qui existe entre la définition formelle et la définition informelle du Big Data. Donc, il faut préciser la définition de ce terme pour que ces définitions soient cohérentes et donnent une bonne compréhension du terme Big Data et ses éléments et ses qualités des données, qui sont indispensables à l'analyse et le traitement du Big Data.

1.8.2 Caractéristiques du Big Data.

1.8.2.1 Les 3V du Big Data

Il existe plusieurs définitions du Big Data. Mais les caractéristiques du big data évoluent avec le temps. Dans un premier temps, on trouve que Big Data est défini par les 3 valeurs, la variété, le volume, la vélocité, qui est la définition la plus utilisée dans les techniques et dans la littérature.

La figure suivante présente les 3V du Big Data.

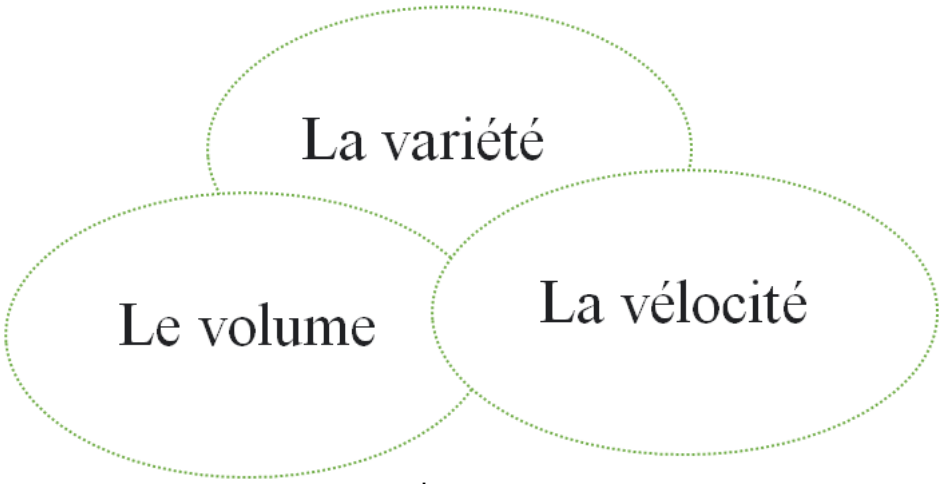

Figure 1.7 Les trois caractéristiques du Big Data.

Ces trois caractéristiques ne sont pas suffisantes pour définir le terme Big Data. Les termes utilisés pour définir les caractéristiques du Big Data ne sont pas bien définis. Par exemple le terme "volume", sa compréhension diffère selon le contexte.

1.8.2.2 Les 4V du Big Data

SAS (Statistical Analysis System) a ajouté la véracité qui contient la complexité et la variabilité. La figure suivante présente les quatre caractéristiques du Big Data.

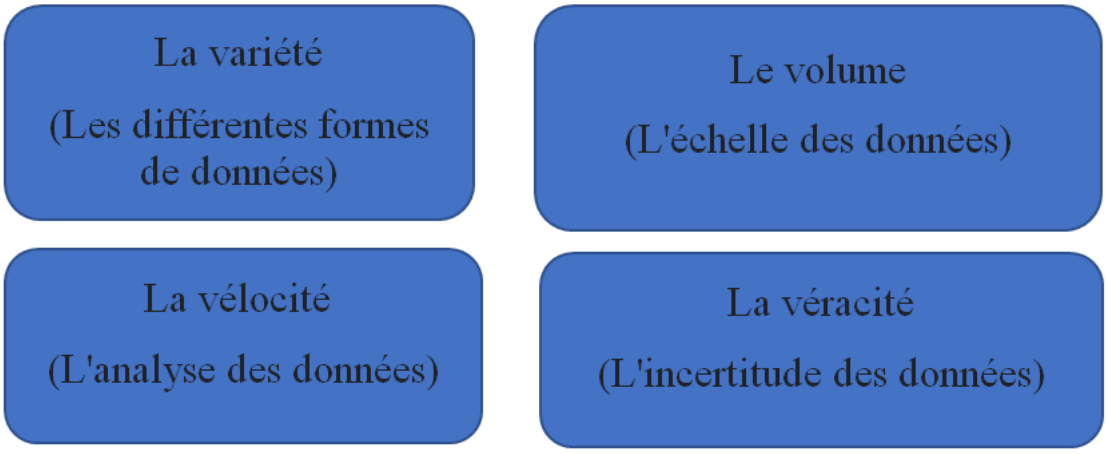

Figure 1 .8 Les quatre caractéristiques du Big Data.

1.8.2.3 Les 5V du Big Data

Après les 4V, on a passé au 5V du Big Data par l'ajout de "la valeur". Ce qui rend le terme Big Data est déterminé par la variété, le volume, la vélocité, la véracité et la valeur.

En 2014, la détermination du Big Data est liée à 10 caractéristiques : La variété, le volume, la vélocité, la véracité, la valeur, la vitesse, la validité, le lieu, la variabilité, l'imprécision, le vocabulaire.

1.8.2.4 Les 14V du Big Data

Et après, le terme Big Data devient caractérisé par 14 caractéristiques. (voir tableau 1.1)

Les caractéristiques du Big Data	**L'élucidation**	**La description des caractéristiques**
La valeur	L'importance des données	Il représente la valeur commerciale à tirer du Big Data
Le volume	La taille des données	Désigne la taille des données collectées et stockées
La variété	Type des données	Les types des données à savoir, l'audio, les vidéos, les images qui sont arrivées au destinataire.
La rapidité	La vitesse des données	Désigne le débit de transfert des données entre le destinateur et le destinataire.
La véracité	La qualité des données	L'analyse de la précision des données : Les données sans valeur, sont des données non précises.
La volatilité	La durée d'utilité	La volatilité des données massives désigne, le temps réservé aux données de l'utilisateur
La validité	L'authenticité des données	L'exactitude et la fiabilité des données utilisées pour produire l'information.
La visualisation	Traitement des données	Traitement de représentation abstraite.
La viscosité	Décalage de l'événement	Le temps entre l'évènement produit et l'évènement décrit.

La viralité	Vitesse d'épandage	La vitesse de propagation des données d'un utilisateur.
La variabilité	La différenciation des données	C'est l'efficacité de différenciation entre les données bruyantes et les données importantes
Le vocabulaire	La terminologie des données	La terminologie des données Qui ce soit les modèles des données ou les structures des données
Le lieu	La diversité des plateformes	Les différents types de données et leurs sources différentes et à partir de différentes plateformes.
La complexité	la corrélation des données	La diversité des sources des données impose l'identification des changements et les comprendre pour que les données arrivent rapidement.
L'imprécision	L'ambiguïté des données	L'imprécision de la réalité des informations des données transmises.

Table 1-1 Les 14 caractéristiques du Big Data.

1.8.2.5 Les 17V du Big Data

Les problèmes de la nature des données massives poussent les scientistes pour trouver des solutions à l'aide du Big Data, ce qui rend la détermination du Big Data se fait par 17 caractéristiques. On a déjà présenté 14 caractéristiques. Les trois caractéristiques restantes sont : La verbosité, le volontariat et la polyvalence. Les significations des trois caractéristiques précédentes sont comme suite :

La verbosité

Big Data est un ensemble des données massives qui sont provenues de différentes sources et qui peuvent être des données structurées ou non structurées, comme elles peuvent être bonnes ou

mauvaises. Les mauvaises données font référence aux informations erronées, obsolètes ou incomplètes. Les conséquences du stockage de ces types d'informations peuvent parfois être dangereuses. Ainsi, il est recommandé de vérifier que les données stockées sont sécurisées, pertinentes, complètes et dignes de confiance. Si une technique appropriée au stade initial est appliquée pour décider si l'information est utile ou non, alors l'espace de stockage, ainsi que le temps de traitement peuvent être économisés. Gardant à l'esprit la nature verbeuse du Big Data, nous avons identifié la « verbosité » comme l'une des caractéristiques du Big Data qui est définie comme « la redondance des informations disponibles à différentes sources ».

Le volontariat

Big Data est un ensemble d'énormes quantités de données qui peuvent être utilisées bénévolement par différentes organisations sans aucune interférence. Big Data aide volontairement de nombreuses entreprises. Elle assiste les détaillants en leur donnant la connaissance des préférences des clients. Et aussi l'urbanisme, en visualisant la modélisation de l'environnement et les modèles de trafic. Ainsi que les fabricants en prédisant les problèmes de produits pour optimiser leur productivité et améliorer les performances des équipements et des clients et aussi les entreprises énergétiques pour répondre aux demandes d'énergie pendant les heures de pointe. Et par conséquent augmenter la production et améliorer l'efficacité en réduisant les pertes, et aussi elle est utilisé par les professionnels de la santé pour prévenir les maladies et améliorer la santé des patients [28], elles aident les organismes de recherche pour obtenir une recherche de qualité et révolutionner les sciences de la vie, les sciences physiques, les sciences médicales et la recherche scientifique[29][30], ainsi qu'elle aident les organisations de services financiers pour identifier et prévenir la fraude, les agences gouvernementales pour améliorer les services dans leurs domaines respectifs. Gardant à l'esprit le comportement volontaire du Big Data, le « volontaire » a été défini comme l'une des caractéristiques du Big Data qui est définie comme « la disponibilité totale des Big Data à utiliser en fonction du contexte ».

La polyvalence

Big Data évolue pour satisfaire les besoins de nombreuses organisations, chercheurs et gouvernements. Il facilite la planification urbaine, la modélisation de l'environnement, la visualisation, l'analyse, la classification de la qualité, la sécurisation de l'environnement, l'analyse informatique, la

compréhension biologique, le processus de conception et de fabrication requis par les organisations et les modèles rentables ainsi que l'exploration élégante du résultat. Gardant à l'esprit les ingénieurs du Big Data, ont identifié la « polyvalence » comme l'une des caractéristiques du Big Data, qui est définie comme « la capacité du big data à être suffisamment flexibles pour être utilisées différemment dans différents contextes ».

1.8.3 Bénéfices du Big Data en E-Learning

Les analyses qui sont faites dans le domaine du E-Learning en utilisant Big Data donne des résultats très favorables. Ce qui encourage les spécialistes en E-Learning pour dépendre sur Big Data. Parmi les avantages qui bénéfice les spécialistes qui utilisent Big Data en E-Learning, on cite :

- Il aide les spécialistes du E-Learning à trouver les points faibles du cours et aussi de les régler pour les rendre compatibles avec les besoins des étudiants en moindre du temps.
- Il permet d'évaluer facilement les cours, et aussi de savoir est ce que les cours sont attirants ou non, à l'aide du nombre de "j'aime" et le nombre de partage.
- Il permet l'analyse du comportement des étudiants pour trouver les difficultés qui les rencontrent avec leurs cours et les remédié et aussi de corriger la méthodologie du cours.
- Il permet une bonne évaluation des performances des étudiants à l'aide de l'analyse de toutes les données qui concernent les étudiants en temps réel, ce qui évite d'utiliser plusieurs évaluations pour savoir les performances des étudiants. Ainsi qu'on peut ajuster les cours selon les besoins des étudiants immédiatement. Donc on gagne du temps lorsqu'on utilise la technologie Big Data.

1.8.4 Outils du big data

1.8.4.1 Apache Hadoop

Hadoop [31] est un logiciel open source pour le traitement du Big Data. Il est caractérisé par : l'évolutivité, la fiabilité, la tolérance aux pannes, la haute disponibilité, le traitement et le stockage local, le calcul distribué et parallèle et la rentabilité. Il utilise un modèle de programmation simple

pour le traitement de grands ensembles de données sur des grappes d'ordinateurs. Il a été développé pour passer d'un seul serveur à des milliers de machines et il permet la colocation du stockage et du calcul. Sa bibliothèque a été conçue de telle manière qu'elle puisse d'auto-identifier les défaillances et de fournir un mécanisme de récupération. Le système de fichiers de Hadoop HDFS (Hadoop File System) [32], est un composant important de Hadoop et possède une architecture maître-esclave. Il se compose de deux composants : NameNode et DataNode. Le système de fichiers distribué offre un débit élevé. Il peut traiter de grands ensembles de données et il a été conçu pour un système à faible coût. L'objectif principal de HDFS est de fournir les fonctionnalités suivantes : débit accru en diminuant la congestion du réseau, la détection et l'isolation des pannes, l'accès à de grands ensembles de données, l'accès aux données en continu, un modèle de cohérence simple et la portabilité entre les différents matériels et logiciels.

1.8.4.2 BlinkDB

Blink DB [33] est utilisé pour le traitement des données à grande échelle et pour les requêtes SQL interactives. Il permet aux clients d'ajuster la précision d'une requête pour un temps de réponse rapide et il permet des requêtes collaboratives de données volumineuses en exécutant les requêtes sur les exemples de données. Il existe deux caractéristiques principales de Blink DB : premièrement, il dispose d'un cadre d'optimisation adaptative, qui construit et conserve des échantillons de données multidimensionnelles à partir des données d'origine au fil du temps. Deuxièmement, il dispose d'une stratégie de sélection d'échantillons appropriée, qui est dynamique, sur la base du temps de réponse de la requête et des besoins de précision. L'objectif principal de Blink DB est de prendre en charge les requêtes interactives, c'est-à-dire les requêtes agrégées sur de gros volumes de données.

1.8.4.3 MongoDB

MongoDB [34] est un outil open source pour le traitement de gros volumes de données stockées dans une base de données. Il possède des fonctionnalités clés telles que la mise à l'échelle automatique, des performances des élevées et la disponibilité des données. Dans la mise à l'échelle automatique, il a une mise à l'échelle horizontale, qui est le principal fragment des fonctionnalités de base. Les données sont réparties entre les clusters à l'aide du partage automatique. L'ensemble de réplication offre une faible latence et un débit élevé. Il a une haute disponibilité car il contient des jeux de réplication qui

assurent automatiquement la récupération des données en cas de panne. Pour des performances élevées, il dispose d'indices, ce qui rend le processus d'interrogation très rapide. De plus, il minimise l'activité d'entrée sortie dans la base de données en intégrant différents modèles.

1.8.4.4 High-Performance Computing Cluster

Le cluster de calcul haute performance (HPCC) [35] est une plate-forme open source pour le traitement d'énormes volumes d'ensemble de données. Il permet à l'utilisateur de traiter le Big Data de manière efficace et efficiente. Il est plus fiable que les autres plateformes. Cette plate-forme offre l'évolutivité, les hautes performances et l'agilité. Il dispose d'un moteur de livraison de données en temps réel pour l'entreposage de données et le traitement des requêtes. Il dispose d'un langage de programmation très puissant pour le traitement de grands ensembles de données. Pour les requêtes Big Data, il dispose d'une plate-forme basée sur des services Web standard. HPCC fournit diverses fonctionnalités qui sont nécessaires pour surmonter les défis du Big Data. Il peut utiliser le matériel de base et il dispose également d'une construction distribuée dans le système de fichiers, la tolérance aux pannes, un environnement de développement intégré (IDE) pour le développement, les différents modules tels que l'apprentissage automatique et les outils d'exploitation. Il comporte trois composants principaux [36] : la raffinerie de données HPCC est un moteur ETL qui permet à l'utilisateur d'intégrer et de manipuler des données. Le deuxième composant est un moteur de livraison de données, qui offre une faible latence et une réponse rapide aux requêtes et un débit élevé. Et le troisième composant est un langage de contrôle d'entreprise (ECL), qui répartit la charge de travail entre les nœuds, avec une bibliothèque pour le développement de l'apprentissage automatique et la synchronisation des algorithmes.

1.8.4.5 L'outil R

R [37] est un outil analytique pour les applications Big Data. Il est utilisé pour l'analyse statistique et la visualisation graphique. R propose une large gamme d'analyses statistiques : la classification, les tests de données classiques, le clustering, l'analyse de séries chronologiques, les méthodes graphiques, la modélisation non linéaire et la modélisation linéaire. Il gère les données de manière efficace et il fournit un stockage de données robuste. Il garantit une procédure fluide pour les matrices et les données vectorielles, afin que les calculs statistiques puissent être optimisés. Il dispose du package R,

qui est un réseau complet d'archives R (CRAN), afin que l'utilisateur puisse obtenir toutes les statistiques nécessaires. Dans le domaine statistique, il peut être promu comme open source. Il est très compatible avec de nombreuses plates-formes telles que Mac OS, UNIX et Windows.

1.8.4.6 Neo4j

Neo4j [38] est une base de données basée sur des graphes, qui est une base de données open source pour le traitement de gros volumes de données. Il dispose de son propre langage de requête et de deux technologies graphiques importantes : le stockage basé sur les graphes et un moteur de traitement basé sur les graphes. Le moteur de traitement basé sur des graphes fournit un traitement de graphe natif. Étant donné que les nœuds sont physiquement connectés au graphe, cela fournit un moyen efficace de traitement du graphe. D'autre part, il fournit un traitement et un stockage natifs pour les bases de données basées sur des graphes. Il offre une meilleure évolutivité que les autres bases de données. Il présente les caractéristiques suivantes : la mise à l'échelle verticale, les performances supérieures et la concurrence. Il prend également en charge diverses plates-formes telles que Java, Python, Ruby, Net et PHP, ce qui le rend plus utile pour les développeurs. Il a également quelques limitations, par exemple, il n'y a pas de support pour le partage. De plus, il existe également des limitations sur les propriétés et les relations des nœuds.

1.8.4.7 Talend

Talend [39] est une plate-forme open source de traitement de big data. L'architecture de Talend couvre tous les besoins des utilisateurs en matière d'intégration et de gouvernance des données. Il présente diverses fonctionnalités telles que l'évolutivité, la facilité d'utilisation et la fiabilité. Ces fonctionnalités le rendent très approprié pour les développeurs. Les outils de Talend se composent de produits pour le développement, la gestion des données, le déploiement et l'intégration de produits. Il existe de nombreux avantages de Talend. Avec son outil ETL, il peut égaliser les charges sur le traitement du serveur sur le cluster. Il utilise également le logiciel Jaspersoft BI. L'interface de l'ETL prend en charge l'importation de Big Data. Sa configuration et la liaison de divers composants permettent aux développeurs de générer plus de productivité que n'importe quel langage de programmation existant. Cependant, il présente également des limitations telles que le besoin de Java Database Connectivity (JDBC) pour l'accès aux ressources. Il n'y a pas non plus de produit pour la

gestion des métadonnées et la qualité des données, et il existe des goulots d'étranglement dans l'automatisation des tâches, le partitionnement et le repartitionnement des données, et l'allocation des ressources à travers la grille.

1.8.4.8 Pentaho

Pentaho [40] est un outil d'intégration de données et d'analyse commerciale. Il s'agit d'une plate-forme open source très populaire pour le mélange, l'analyse et la visualisation du Big Data. Il a de larges capacités pour l'exploration et l'analyse de données. C'est un choix alternatif pour les développeurs et offre une interface utilisateur bien mise à jour. Il fournit un support natif pour Hadoop et tout type de source de données. Les utilisateurs de Pentaho n'ont pas besoin d'écrire de code pour l'intégration. Il fournit également plusieurs autres fonctionnalités aux utilisateurs, par exemple, la veille économique (BI), l'intégration de données, le tableau de bord, les capacités ETL, les services de traitement analytique en ligne (OLAP) et l'exploration de données. Les limitations de Pentaho incluent une visualisation limitée des données, un manque de documentation appropriée et des outils analytiques limités, qui nécessitent davantage d'améliorations.

1.8.4.9 SAS

SAS [41] propose diverses techniques d'analyse du Big Data, en fournissant une infrastructure pour des logiciels d'analyse et des statistiques de hautes performances. Il fournit des fonctionnalités telles que le traitement distribué, la commutation de grille, l'analyse de base de données et le calcul en mémoire. Il peut effectuer un déploiement dans le cloud et sur site. Il fournit également des solutions à des problèmes complexes. C'est un outil analytique avancé utilisé dans les principales industries.

Hadoop est le framework le plus utilisé puisqu'il est open source et aussi très performant. C'est le framework qu'on a utilisé pour le traitement du Big Data.

1.8.5 Architecture de hadoop

Hadoop contient plusieurs modules et technologies à savoir : HDFS, MapReduce, Yarn, Pig, Phoenix, Hbase, Zookeeper, Impala, Hama, Hawq, Spark, Hive, ElasticSearch, Lucene, Sqoop, Mahout, Oozie, Storm et TEZ. L'utilisation de ces technologies dépend du problème de Big Data. Ces

modules sont divisés en deux types, les modules principaux et les modules supplémentaires. Il existe quatre modules principaux.

Le premier module est Hadoop Distributed File System (HDFS). L'apparition des données de grande taille, pose un problème au niveau du stockage centralisé, donc au lieu de faire le stockage des données sur une seule supermachine coûteuse, ces données sont réparties sur plusieurs serveurs qui sont considérées comme une seule machine logique (cluster), grâce à la méthode de stockage distribuée utilisé par Hadoop Distributed File System. Ce système de fichiers est facile à utiliser et très rapide et en plus, il est extensible et ne nécessite pas, ni la modification de l'architecture, ni le déplacement des données existantes. Avec la réplication des données le HDFS limite le risque de perdre les données.

Le deuxième module de Hadoop est "Yet Another Resource Negotiator" (Yarn) [42], qui est l'élément chargé à la planification et à la gestion des clusters Hadoop, il fait l'allocation et la planification des ressources au sein des clusters.

Le troisième module est MapReduce. C'est un modèle de programmation qui permet la parallélisation des traitements sur les nœuds d'un cluster Hadoop. MapReduce s'appuie sur deux opérations : Map () et Reduce(), pour l'analyse efficace des données et en moins du temps.

Le quatrième module est Hadoop Common, c'est un ensemble de bibliothèques qui gèrent les systèmes de fichiers distribués, sur lesquels se basent les différents modules.

Il existe d'autres modules de l'écosystème Hadoop qui le complémentent et aussi améliorent ses performances. Parmi ces modules on trouve, Hive [43], qui est une infrastructure de calcul similaire au Data Warehouse, il est open source et stable. Il permet la simplification de la rédaction, la lecture et la gestion de jeux de données stockées sur un système de fichiers distribué, il réalise des analyses par HiveQL. Le point fort qui caractérise Hive, est la rapidité d'exécution des requêtes, qui nécessite la liaison des tables de grandes tailles. Pour le même but que Hive, la technologie PIG [44] est créée, c'est un environnement d'exécution de flux interactifs de données pour Hadoop. Il dispose d'un environnement d'exécution et le Pig Latin qui est un langage permettant un accès aux requêtes d'une manière simple, il est proche du langage SQL. PIG joue le rôle d'un ETL pour Hadoop. La différence entre Hive et PIG, est que PIG est utile pour la préparation de données, parce qu'il exécute les requêtes

complexes et les jointures facilement, ainsi qu'il fonctionne bien avec les données semi-structurées et aussi les données non structurées. Au contraire HIVE est bien compatible avec les données structurées et les opérations de data warehousing. Un autre composant de l'écosystème Hadoop est le HBase [45], qui est une base de données orientée colonnes pour Hadoop, elle est non relationnelle. Il y a aussi le module Oozie [46], qui est un planificateur des jobs qui sont exécutés sur un cluster Hadoop. Et le module Zookeeper [47], qui est un coordonnateur de traitements distribués, il permet la supervision des échanges entre les différents nœuds d'un cluster. Il y a aussi Sqoop [48], qui est un outil qui permet le transfert des données entre les serveurs de bases de données relationnelles et Hadoop, et FLUME [49] qui est un outil d'ingestion de données de grandes tailles vers Hadoop en se basant sur des événements, ainsi que Kafka [50], qui est une plateforme open source qui permet le traitement de flux des données, elle est évolutive par rapport au flume.

L'utilisation de MapReduce est compatible seulement aux algorithmes scalables, c'est-à-dire que pour exécuter un algorithme en utilisant MapReduce, cet algorithme doit être parallélisable, car, l'exécution des opérations sur MapReduce [51] se fait d'une manière séquentielle et sans retour. Ce qui rend la réutilisation des données par des opérations multiples interdit par MapReduce. Dans ce cas, la solution adéquate est l'utilisation de Spark [52], qui fait les calculs d'une manière distribuée mais en mémoire, il est en mode interactif, en d'autres termes, avant de traiter les données, ces derniers sont montés de la mémoire. Ce mode est plus utilisé dans la majorité des algorithmes et surtout les algorithmes de machine Learning.

La figure suivante présente l'architecture de l'écosystème Hadoop.

<table>
<tr><td rowspan="5">Zookeeper
(La coordination)</td><td>Hive (L'analyse)</td><td>Pig (script)</td><td>Oozie (La planification)</td><td rowspan="5">HBase
(Le stockage en colonne)</td></tr>
<tr><td colspan="3">Spark, MapReduce (Le traitement des données)</td></tr>
<tr><td></td><td>Yarn (la gestion des ressources du cluster)</td><td></td></tr>
<tr><td></td><td>HDFS (Le stockage)</td><td></td></tr>
<tr><td colspan="3">Sqoop, Flume, Kafka (L'ingestion)</td></tr>
</table>

Figure 1.9 Les composantes de l'écosystème Hadoop.

Cette figure présente les différents modules de l'écosystème Hadoop et aussi la fonction de chaque composant. Hive et Pig sont des langages d'abstraction. Hive pour le côté analyse et Pig pour le côté

Scripting. HBase est une base de données qui permet le stockage en colonne. Oozie est un planificateur de job (scheduling). Spark et MapReduce sont utilisés pour le traitement des données. Yarn permet la gestion des ressources du cluster et HDFS permet le stockage distribué des données. Sqoop, Flume et Kafka permettent l'ingestion des données et enfin Zookeeper qui permet la coordination des traitements. Tous ces modules rendent l'écosystème Hadoop plus flexible et efficace et aussi complet.

1.8.6 Classification du Big Data

Les données utilisées en Big Data ont différentes sources, différents formats, multiples méthodes de stockage et de transiter des données et différentes méthodes de traitement. La figure suivante présente une classification du big data.

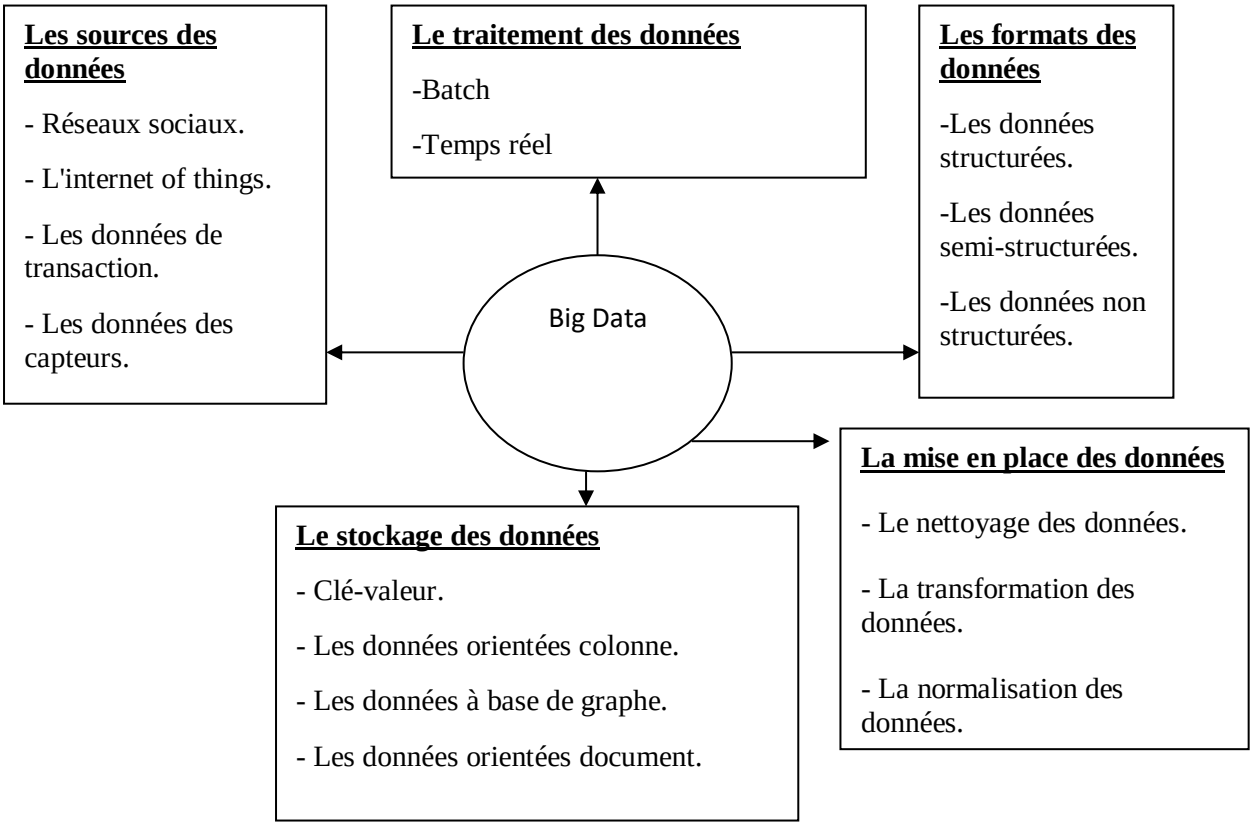

Figure 1.10 La classification du Big Data.

1.8.7 Aspects de l'application du Big Data en éducation

La technologie Big Data n'est pas utilisée dans tous les aspects de l'éducation. Son utilisation est limitée à certains aspects, la figure suivante présente les aspects de l'application du Big Data en éducation.

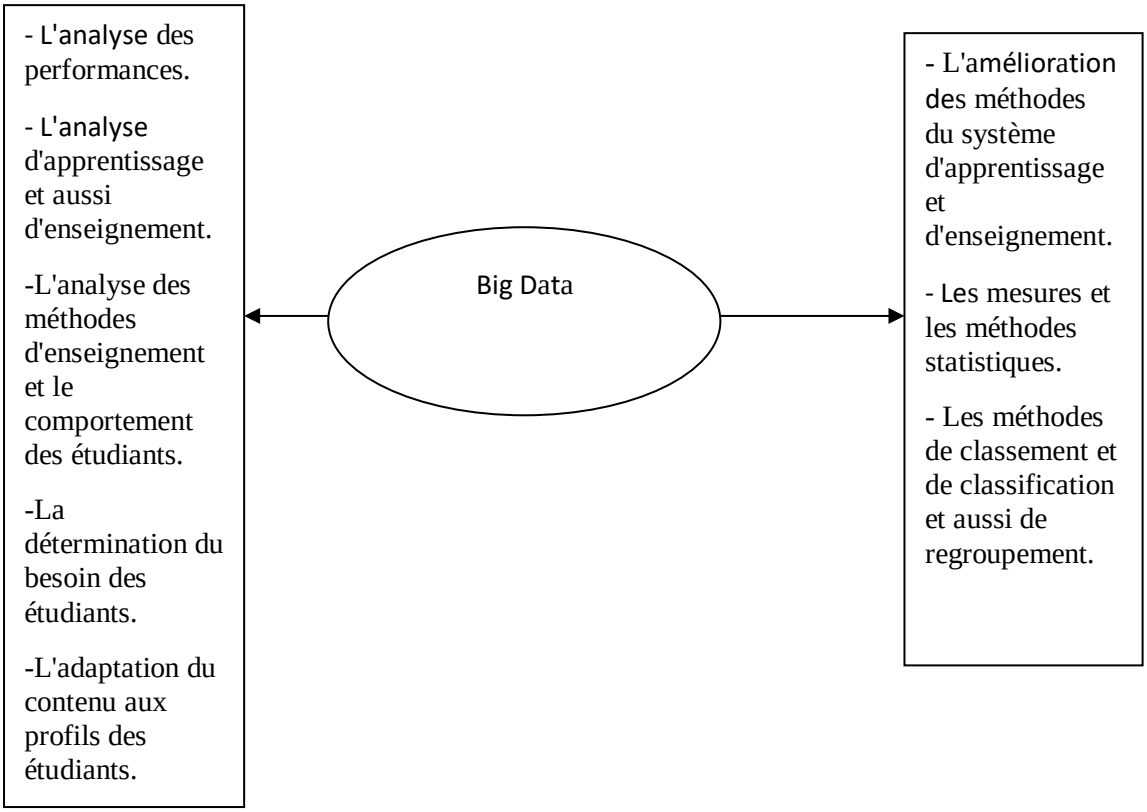

Figure 1.11 Les aspects de l'application du Big Data en éducation.

1.8.8 Big Data en E-Learning

Les systèmes de E-Learning traditionnelles sont limités et aussi ne donnent pas des bons résultats vu à leurs capacités de stockage et de traitement. Ils sont composés de trois couches : la couche client, la couche plateforme du E-Learning et la couche de la base de données.

La figure suivante présente les différentes composantes de chaque couche.

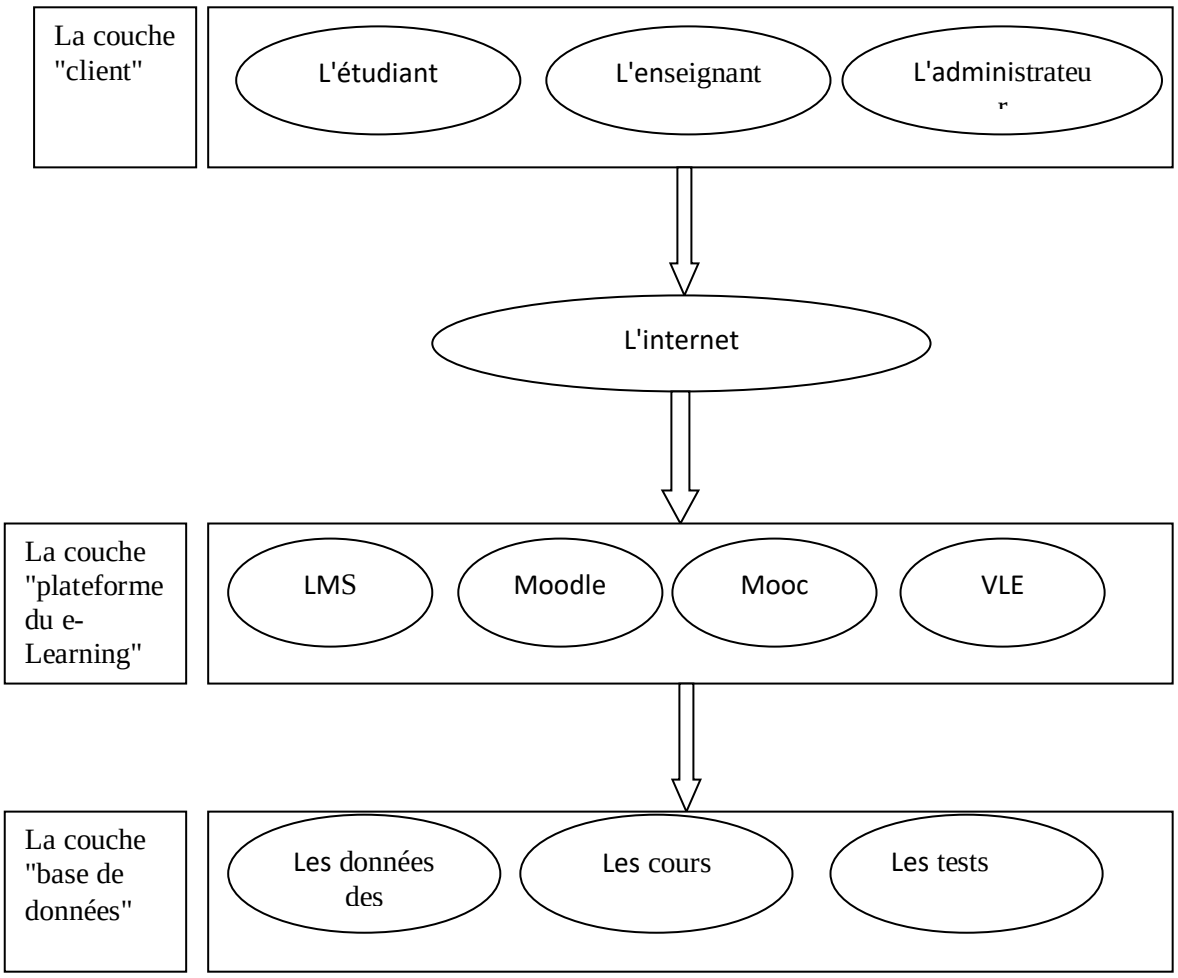

Figure 1.12 L'architecture traditionnelle du e-learning.

L'architecture traditionnelle du E-Learning se compose de la couche "client" qui contient les utilisateurs du système (Les étudiants, les enseignants et les administrateurs), qui peuvent accéder à la plate-forme du E-Learning via l'internet. Cette plate-forme peut être LMS, MOOC, VLE, LCMS. Elle interagit avec la couche de la base de données par lecture ou par écriture. Ces bases de données contiennent les profils des étudiants, les cours et les tests.

Avec l'apparition du Big Data et le cloud computing et aussi pour profiter les avantages du big data qui sont déjà cités (les benéfices du Big Data), cette architecture du système de E-Learning est changée. Il existe plusieurs architectures qui intègrent le Big Data en E-Learning.

La figure suivante présente un exemple d'architecture qui intègre le Big Data en E-Learning.

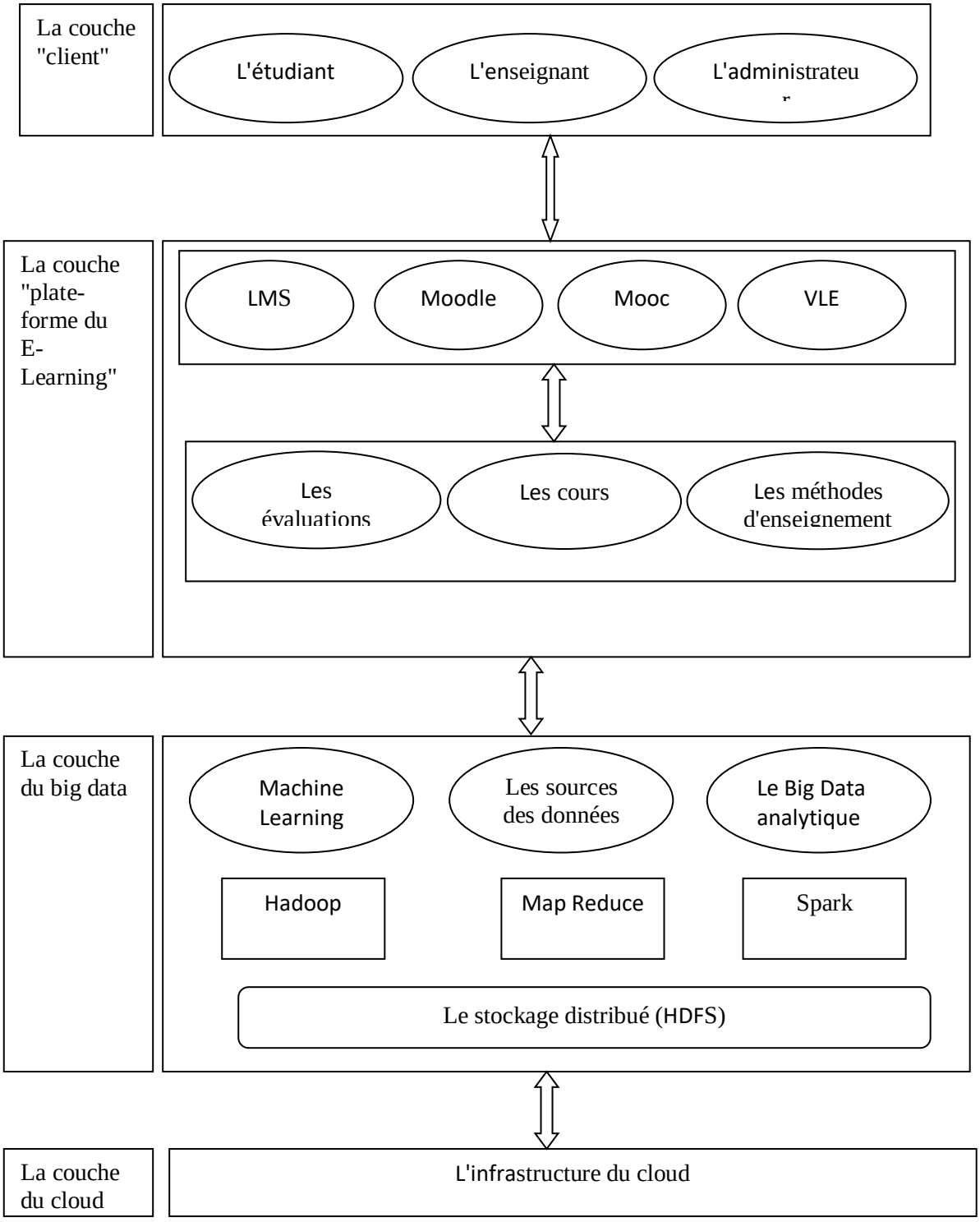

Figure 1.13 Un exemple d'architecture du e-Learning basé sur big data

L'architecture du E-Learning basé sur Big Data est composée de quatre couches. La couche client, la couche plateforme du E-Learning, la couche du Big Data et la couche du cloud.

La première couche est la couche client, qui contient trois types de clients, les étudiants, les administrateurs et les enseignants. Ils sont les acteurs qui utilisent le système du E-Learning.

La deuxième couche est la couche "plate-forme du E-Learning", elle contient les plates-formes d'apprentissage en ligne (LMS, CMS, LCMS, VLE, ...) et les méthodes d'enseignement et d'évaluation. Elle contient aussi les informations des étudiants et des cours. Ces données sont utilisées pour la prise de décision et aussi pour adapter les cours aux profils et aux besoins des étudiants.

La troisième couche est la couche du Big Data. Elle permet le stockage et le traitement distribué et l'analyse des données à l'aide des outils technologiques à savoir, HDFS, Map Reduce, Spark, Cassandra, MongoDB, HBase, Kafka, Zookeeper, Storm, Hive, Pig, Flume, Sqoop et Oozie. Ces outils rendent l'analyse des données très efficace et fiable.

La quatrième couche est la couche du cloud. Cette couche rend l'utilisation du matériel très flexible. Elle contient des ressources de stockage et de calcul et du réseau virtuel.

Les services fournis par Big Data aux utilisateurs des plates-formes du E-Learning sont présentés dans la figure suivante :

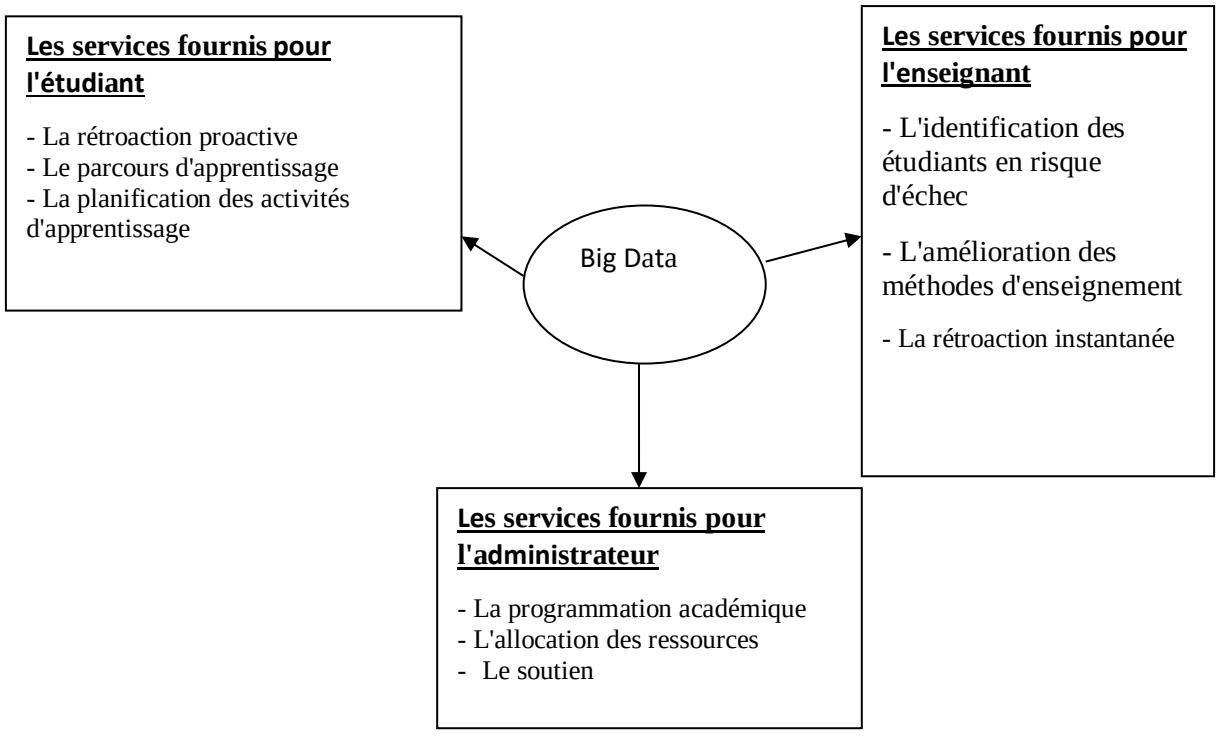

Figure 1.14 Les services fournis par Big Data aux utilisateurs des plates-formes du E-Learning.

D'après la figure 1.14, les services fournis par Big Data sont divers. Ce qui favorise l'utilisation du Big Data dans plusieurs domaines et surtout le domaine de l'éducation et de l'apprentissage.

1.9 Conclusion

Ce chapitre a présenté E-learning, la technologie Big Data et leurs caractéristiques, et aussi leurs outils. Ainsi que les aspects d'intégration du Big Data en éducation. D'après ce chapitre, la technologie Big Data fournit plusieurs services dont le domaine de l'éducation en a désespérément besoin. Mais les recherches dans ce stade sont très limitées. Et vu à l'importance cruciale de l'éducation et surtout le E-Learning, nous sommes très intéressés à utiliser cette technologie pour améliorer le système d'orientation scolaire des étudiants et aussi d'augmenter la qualité et les performances de la prédiction de réussite des étudiants, qui est très utile pour lutter contre l'échec scolaire et le décrochage scolaire. Les chapitres suivants présentent les travaux qu'on a faits dans ce stade.

2. Chapitre II : Systèmes de gestion de l'apprentissage

2.1 Introduction

Parmi les méthodes utilisées en E-Learning [53], on trouve les systèmes de gestion de l'apprentissage, LMS [54], qui contiennent également de nombreuses plateformes d'E-Learning, par exemple Moodle, Sakai, edomodo, ... etc. Ils permettent la gestion des cours, des contenus, des étudiants, etc. Mais le problème qui se pose est de savoir comment choisir la bonne plateforme selon notre objectif, en compte tenu de la diversité de ces outils en termes de leurs fonctions et leurs performances.

Plusieurs études comparatives sont faites sur les LMS. Parmi ces études on trouve [55], dans laquelle l'auteur utilise l'expérience utilisateur (UX) pour comparer deux systèmes d'apprentissage en ligne : iQuality et Moodle. Il trouve que iQuality fonctionne mieux que Moodle. De même [56], évalue l'accessibilité des plates-formes Sakai, Moodle et ABC à l'aide de ces critères d'accessibilité : connexion, configuration, tests de compatibilité, personnalisation, navigation, formulaires, aide et documentation, modules/outils communs aux étudiants, création des outils et création du contenu, fonctionnalités uniques au LMS qui affectent l'accessibilité. Ils ont trouvé l'accessibilité des plates-formes open source est la meilleure que l'accessibilité des plates-formes développées en interne. De plus, dans [57], les auteurs présentent sept critères : la base, interface utilisateur, intégration avec les applications mobiles, outils, fichier de format compatible, prise en charge linguistique et prix de base pour comparer les trois systèmes de gestion de l'apprentissage : Moodle, Edmodo et Jejak Bali. De plus [58], utilisent des trois fonctionnalités: administrateur, le point de vue du tuteur, le point de vue de l'étudiant pour comparer ATutor, Moodle et Chamilo .Ils ont trouvé aussi que Moodle est meilleur que Chamilo et enfin ATutor. Toutes ces études considèrent que les caractéristiques utilisées dans la

comparaison ont la même distribution et la même importance. Pour résoudre ce problème, nous utilisons un algorithme de prise de décision multicritère.

Ce chapitre est structuré comme suit. Le premier sous chapitre présente LMS, l'importance du LMS, un historique des systèmes de gestion de l'apprentissage, les systèmes de gestion de l'apprentissage mobile, la différence entre LMS, CMS et LCMS et les types des outils de LMS. Alors que, le deuxième sous chapitre présente six systèmes de gestion de l'apprentissage : Moodle, Sakai, Claroline, TalentLM, Easy LMS et OpenedX. Le troisième sous chapitre explique les fonctionnalités utilisées pour comparer les plateformes LMS précédentes, alors que le quatrième sous chapitre présente l'algorithme de prise de décision multicritères et le cinqième sous chapitre présente les résultats. Enfin la conclusion dans le dernier sous chapitre.

2.2 Définition des systèmes de gestion d'apprentissage

L'enseignement et l'apprentissage traditionnel se font en face à face, et nécessite l'existence des étudiants et de l'enseignant en même place, qui ce soit dans une classe ou une bibliothèque, et les cours se font en temps réel. Mais maintenant on a passé à un modèle d'apprentissage et d'enseignement à distance basé sur la technologie. On a passé du modèle de l'apprentissage et de l'enseignement traditionnel à un modèle de D-Learning, E-Learning et M-Learning.

Le système de gestion de l'apprentissage (LMS) est un logiciel informatique d'apprentissage électronique .Il dispose des fonctions d'évaluation, de l'enseignement et de l'administration des cours. Il permet le suivi des étudiants, la diffusion du contenu, la création de rapports et l'administration.

Les LMSs sont devenues de plus en plus utilisables dans l'enseignement et surtout l'enseignement supérieur. La progression d'adoption du LMSs ou VLE implique l'importance du e-Learning. Les outils d'apprentissage numérique en ligne sont très utilisés par les étudiants universitaires dans leurs campus et aussi hors de leurs campus, tels que Black board et WebCT [59], [60]. LMS est un outil d'apprentissage très usuel et très rentable, car il gère l'apprentissage face à face et aussi l'apprentissage en ligne ainsi que tous les éléments essentiels de l'apprentissage [61]. C'est un environnement

d'apprentissage virtuel (VLE) [62] et aussi un système de gestion des cours [63]. Il a toutes les fonctions et les outils d'administration des cours, d'évaluation et soutien et aussi d'enseignement [64]. C'est un outil qui permet le suivi et l'évaluation des étudiants pour détecter les capacités des étudiants et leurs lacunes, dans le but de les remédier [65]. Tous ces outils et ces fonctions sont disponibles aux étudiants et aux enseignants [66][67][68]. Parmi les fonctions du LMS, on trouve la possibilité d'interaction entre les apprenants et aussi entre les apprenants et les enseignants à l'aide des messageries instantanées, les courriers électroniques, le chat, les forums de discussions, le suivi des activités des étudiants [67]. On peut définir LMS comme une collection d'outils d'apprentissage basé sur une interface qui permet une gestion partagée en ligne [69], à savoir les outils de prestation des cours (le chargement de cours, les testes, les évaluations, les devoirs et les programmes), les outils d'interaction (les forums de discussions, les e-mails, le chat, les messageries). Ces outils permettent une bonne conception des cours destinés aux étudiants. Ces cours peuvent être réalisés en face à face avec des instructions en lignes [70]. La majorité des universités ont utilisé les LMSs pour mettre en place l'apprentissage en ligne [71]. L'utilisation du LMS n'est pas limitée en éducation, mais elle est utilisée presque dans tous les domaines et tous les secteurs que ce soit gouvernementals ou commerciaux, dans le but de former les employées [72]. Les LMSs permettent aux formateurs de dispenser, de planifier et d'organiser les cours en ligne facilement et efficacement [73]. Ils ont capable d'améliorer la qualité d'éducation en ligne et aussi de réaliser des classes virtuelles. L'adoption de LMS ouvre les portes aux enseignants et aux apprenants de communiquer et d'interagir à n'importe où et à tous les moments et sans obligation d'existence physique. A l'aide de LMS les outils nécessaires et le matériel essentiel qui permettent la gestion de l'environnement virtuel d'apprentissage et d'enseignement sont prêts et faciles de l'utiliser.

Dans l'environnement du LMS, on a trois types d'utilisateurs, les apprenants ou les étudiants, les enseignants et les administrateurs [72]. Les fonctions des administrateurs sont de trouver les solutions aux problèmes techniques à savoir, la supervision du fonctionnement du LMS et l'administration des

comptes des utilisateurs du LMS. Alors que les fonctions des enseignants sont de créer les cours, d'interagir avec les apprenants et les évaluer et la présentation du contenu. Et enfin les étudiants qui sont le centre d'intérêt du LMS, et les destinataires des cours. Ils peuvent communiquer et interagir d'une manière synchrone ou asynchrone avec leurs enseignants ou entre eux.

L'apprentissage mobile à cinq niveaux ou catégories des pratiques industrielles : Le niveau 0, les LMSs de ce niveau ne permettent pas l'apprentissage mobile, alors que les LMSs du niveau 1, à savoir Sakai, Moodle, sont construites graphiquement pour s'adapter aux appareils mobiles. Tandis que les LMSs du niveau 2, à savoir Mobile™, MOMO, MLE-Moodle, Black board, disposent des extensions mobiles. Et les LMSs du niveau 3, à savoir, Pushcast™, BlackBerry, sont mobiles autonomes .Et enfin les LMSs du niveau 4, à savoir, le cloud computing, mobiles innovants, qui contiennent les nouvelles fonctionnalités des appareils mobiles [74].

2.3 Historique des systèmes de gestion de l'apprentissage

Les LMSs ont apparu en 1960, mais ils ont été hors ligne du fait qu'ils n'étaient pas supportés par le web. Donc c'est un concept ancien. Parmi les LMSs on trouve, WebCT, Moodle, Blackboard, OPAL, eCollege, Desire2Learn et PLATO .La description de quelque LMSs est comme suit :

• PLATO : Programmed Logic for Automated Teaching Operations (PLATO) .Il a apparu en 1960, c'est le premier système d'apprentissage contrôler par ordinateur, il est inventé par l'Université de l'Illinois à Urbana-Champaign. Au début, les étudiants ont utilisé leurs ordinateurs personnels pour accéder à PLATO, et après, les cours sont devenus disponibles en ligne par PLATO et accessible de partout.

• Learning Manager : Il est inventé en 1980 sous le nom TLM (The Learning Manager), c'est un système de gestion d'apprentissage très populaire, avec des nouveaux outils de gestion, de développement, de supports d'apprentissage et de rapport. Il permet la gestion de l'apprentissage en ligne et il permet aussi aux étudiants et aux enseignants l'accès à distance à ses services.

• Projet Andrew : Il est créé en 1982 par l'Université Carnegie Mellon. Il permet la collaboration en ligne à l'aide d'un environnement unifié, dans le but de créer une plate-forme d'apprentissage assisté par ordinateur.

• EKKO : C'est une plateforme de conférence informatisé, créé par NKI. Elle a commencé la création des cours à distance en 1987, par NKI Distance Education.

• ATHENA : la première version est créée en 1983, puis il est amélioré par l'Institut Massachusetts de technologie en 1990. Ce système permet le partage d'article et sa rédaction et la communication avec tous les utilisateurs du campus, et aussi, il permet le suivi des étudiants. Il dispose d'un tableau noir, un simulateur, un enseignant, un manuel, et un laboratoire virtuel, c'est un environnement d'apprentissage complet.

• HyperCourseware : C'est un système d'apprentissage à distance et aussi en classe. Il est créé en 1990 par l'expert des ordinateurs Kent Norman au sein de l'Université du Maryland, pour rendre les ressources d'apprentissage et de l'enseignement accessible par voie électronique, à savoir, les discussions, les conférences, les documents et les manuels.

• WebCT : C'est un CMS créé au sein de l'Université de la Colombie-Britannique en 1996 par Murray Goldberg. Il est caractérisé par son système de messagerie, les forums de discussions, le chat en direct. C'est le CMS le plus populaire aux universités.

• Blackboard™: C'est un LMS commerciale et très populaire. Il est fondé en 1997. Il offre plusieurs plateformes, Blackboard Learn™ (qui est un LMS), Blackboard Connect™ (elle est créée pour les informations urgentes), Blackboard Collaborate™ (elle est créée pour l'enseignement synchrone à l'aide d'une salle de classe virtuelle), Blackboard Analytics™ (elle est créée pour simplifier l'accès aux informations importantes), Blackboard Mobile™ (c'est la version mobile du LMS Blackboard) et Blackboard Transact™ (elle est créée pour les achats sécurisés).

• Desire2Learn™ : Il est créé en 1999. Il dispose d'un environnement d'apprentissage en ligne. Il dispose aussi de six plateformes, ePortfolio, Learning Environment, Learning Repository, Analytics, Capture, et Mobile.

Après BlackboardTM, d'autres LMS sont apparus, et qui sont très efficaces à savoir, Modular Object-Oriented Dynamic Learning Environment (Moodle) et Sakai.

2.4 Systèmes de gestion de l'apprentissage mobile

Avec l'apparition des appareils mobiles, et leurs rôles importants de la simplicité d'accès à l'internet ainsi que son nombre d'utilisations exponentiel. Les créateurs des LMS ont rendu leurs produits compatibles avec les appareils mobiles. Parmi les entreprises qui ont pris en considération la compatibilité du LMS avec les appareils mobiles durant la phase de la conception de leur produit, pour créer une plateforme mobile. On trouve :

• Moodle Mobile (MOMO) : C'est un LMS libre qui contient toutes les fonctionnalités du Moodle, ainsi qu'il inclut d'autres fonctionnalités, telles que les objets d'apprentissage mobiles hors ligne (MLO), communauté mobile et mobile bloguer.

• Blackboard Mobile Learn™: Il permet grâce à sa compatibilité de son environnement d'apprentissage aux plateformes mobile, un accès au contenu sur la majorité des appareils mobiles. Parmi les plateformes mobiles qui sont compatibles avec les applications du Blackboard™, on trouve, Android™, iPhone™, Touch™ et iPod.

•Desire2Learn 2GO™ : Il est créé pour être compatible à Blackberry™, il est aussi compatible aux smartphones, à savoir IPhone.

2.5 Différence entre LMS, CMS et LCMS

2.5.1 Système de gestion de contenu

Un CMS est représenté par trois dimensions : le processus, le contenu et la technologie. Le processus regroupe toutes les activités qui régissent sur un ensemble d'entrée pour avoir et produire des résultats ou une sortie. À titre d'exemple : le partage des informations, le téléchargement des données et la publication des documents [75]. Alors que le contenu du CMS est de divers types, des

graphiques, des sons, des textes, des animations, des vidéos ou multimédia [76]. Enfin, la technologie qui est un outil d'exécution et de la mise en œuvre les traitements et les processus qui sont générés par les utilisateurs et aussi un outil pour contrôler et superviser le contenu à l'aide d'internet [77]. Les catégories du CMS sont diverses. Il y a CMS composantes, CMS Entreprise, cms web, gestion des enregistrements, gestion des documents,…, etc. Le CMS entreprise donne aux employés les outils, les méthodes, les processus et les stratégies qui ont besoin pour vérifier, d'accéder et de gérer les documents, les médias et les modèles. Alors que CMS composantes est un logiciel qui aide les utilisateurs à l'accès, la modification, l'enregistrement et la gestion des niveaux du contenu. Tant que CMS web permet le partage et la publication du contenu sur le web. Ce qui facilite aux créateurs du contenu de présenter, d'éditer et d'envoyer ce contenu sans avoir des compétences sur le langage HTML [78].

2.5.2 Système de gestion du contenu d'apprentissage

Le système de gestion de contenu d'apprentissage (LCMS), c'est le plus utilisé maintenant dans l'éducation. C'est un environnement multi-utilisateur. Il permet à l'éducateur de gérer, de stocker, de présenter, de créer et de réutiliser le contenu d'apprentissage au format numérique. Selon la méthode d'apprentissage, l'enseignant est capable de gérer et de présenter les devoir soit synchrone soit asynchrone à l'aide du système de gestion de contenu d'apprentissage. Les méthodes et les fonctions des LMS et aussi du CMS sont fournies par la majorité des LCMSs.

La norme SCORM (Shareable Content Object Reference Model), qui sert à contrôler la manière de communication entre les systèmes de gestion de l'apprentissage et le contenu d'apprentissage, cette norme est appliquée par la majorité des LCMS, LMS et CMS. D'après [79], SCORM est une liste des normes qui doit être respectée par les programmeurs des logiciels d'apprentissage en ligne, dont le but de rendre ce code utilisable et aussi exploitable avec d'autres logiciels d'apprentissage en ligne.
Le but majeur du LCMS, est la bonne gestion des actifs numériques qui sert au développement des logiciels et des outils d'apprentissage.

Ces systèmes permettent l'enregistrement des contenus et des travaux de chaque conférencier à l'aide des bases de données spécifiques. Elles sont appelées les objets de référentiel de contenu

d'apprentissage « learning content repository objects ». Elles servent à plusieurs taches, notamment, l'accessibilité de l'enseignant lui-même ou des autres enseignants au système pour développer ou mettre à jour le contenu, le travail en groupe des étudiants avec leurs enseignants à l'aide des outils de coopération, la gestion et la réalisation des quiz et des examens.

Les trois systèmes de gestion d'apprentissage LMS, CMS et LCMS semblent similaires, ils permettent de diffuser et de gérer du contenu, mais il y a une grande différence entre eux, bien qu'ils diffèrent par une seule lettre. On différencie entre les trois termes selon cinq critères :

2.5.3 Critères de différenciation entre LMS, CMS et LCMS.

- **La création du contenu**

La création du contenu n'est pas disponible dans LMS au contraire du CMS et LCMS, qui permet de créer le contenu d'apprentissage et aussi le partage en divers formats, qui ce soit des liens, des codes d'intégration ou d'exportation SCORM. Alors si vous utilisez LMS, il faut trouver un autre outil distinct pour créer des contenus d'apprentissage avant l'importation de ce dernier.

- **L'exportation SCORM**

L'exportation SCORM, fait une grande différence entre LMS, CMS et LCMS. Car les systèmes de gestion d'apprentissage comme LMS et LCMS sont capables de gérer ces fichiers qui contiennent des normes techniques à l'aide de leurs fonctionnalités d'apprentissage qu'ils disposent. Tandis que les systèmes de gestion du contenu n'ont pas les fonctionnalités d'apprentissage pour gérer les fichiers SCORM.

- **Les fonctionnalités d'apprentissage**

Depuis les noms des trois types des systèmes précédents, le mot « Learning » est absent dans le nom du CMS. Car, le contenu d'apprentissage n'est pas son but. C'est un outil de contenu général. Alors que LMSs prennent en charge l'apprentissage mobile, les quiz, la collaboration et aussi les outils qui permettent la planification de webinaire. Et aussi les LMSs présentent le progrès des étudiants, la réussite de l'étudiant ou l'échec et les cours qui sont terminés par l'étudiant. Ces informations ont fourni aussi par les LCMS avec plus de détails. Ils ont la possibilité de trouver les lacunes qui sont

liées aux connaissances de l'étudiant, ainsi que la prise des décisions qui concernent la mise à jour du contenu.

- **L'édition collaborative**

Les LCMS et CMS permettent la collaboration de plusieurs éditeurs sur le même contenu. Alors que les LMSs ne disposent pas cette fonctionnalité. Puisqu'ils n'ont pas les outils nécessaires à la création du contenu.

- **L'importation de contenu hérité**

Même si vous utilisez l'un des systèmes de gestion de contenu numérique, vous pouvez rencontrer des documents sous forme papier qui sont hérités à suivre. Pour cela, les LCMSs et les CMSs ont ajouté les fonctions nécessaires à l'importation pour numériser les documents sous forme de papier. Ces fonctions ont fourni par les systèmes de gestion de documents (DMS).

Le tableau suivant présente les différentes fonctionnalités des LMS, CMS et LCMS.

R : désigne robuste forte

L : désigne limitée

Les propriétés	Les fonctionnalités		
	LMS	CMS	LCMS
L'enregistrement par le système	R		L
La gestion du contenu		R	R
La bibliothèque en ligne pour la recherche du contenu réutilisable		R	R
La gestion des sessions	R		
La disponibilité du catalogue du cours	R		L
La gestion des étudiants/les apprenants	R		L
La gestion des savoirs et des	R		L

compétences			
Le suivie le contrôle de l'apprentissage en ligne	R		L
La création des examens, des devoirs et des évaluations	R		R
La création du contenu d'apprentissage		L	R
La collaboration et la synchronisation des outils d'apprentissage	L		R
L'intégration avec d'autres applications	R		
Le contenu ciblé à un apprenant		R	R

Table 2-1 Les propriétés des LMS, CMS et LCMS.

En bref, les différences majeures entre un LMS, un CMS et un LCMS sont distingués par la manière de manipulation du contenu et aussi par l'identification de l'utilisateur clé.

Les LMSs visent l'apprenant et aussi la manière dont l'étudiant utilise le contenu.

Les LCMS visent le formateur qui est chargé à la création du contenu d'apprentissage personnalisé. Ils sont très utilisés lorsqu'on a besoin des cours personnalisés.

Les CMSs ne visent aucun utilisateur spécifique, car ils sont des systèmes de stockage. Ils donnent la priorité au stockage et aussi l'organisation du contenu. C'est pour cette raison les CMSs peut-être un outil d'apprentissage, mais les LMSs sont les plateformes les plus adéquates, vu que leurs capacités importantes et leurs activités dynamiques.

2.6 Types des outils de LMS

Les outils du LMS sont regroupés en trois types : les outils de communication, les outils des compétences d'apprentissage, et les outils de productivité. La figure 2.1 suivante présente ces trois types du LMS avec des exemples.

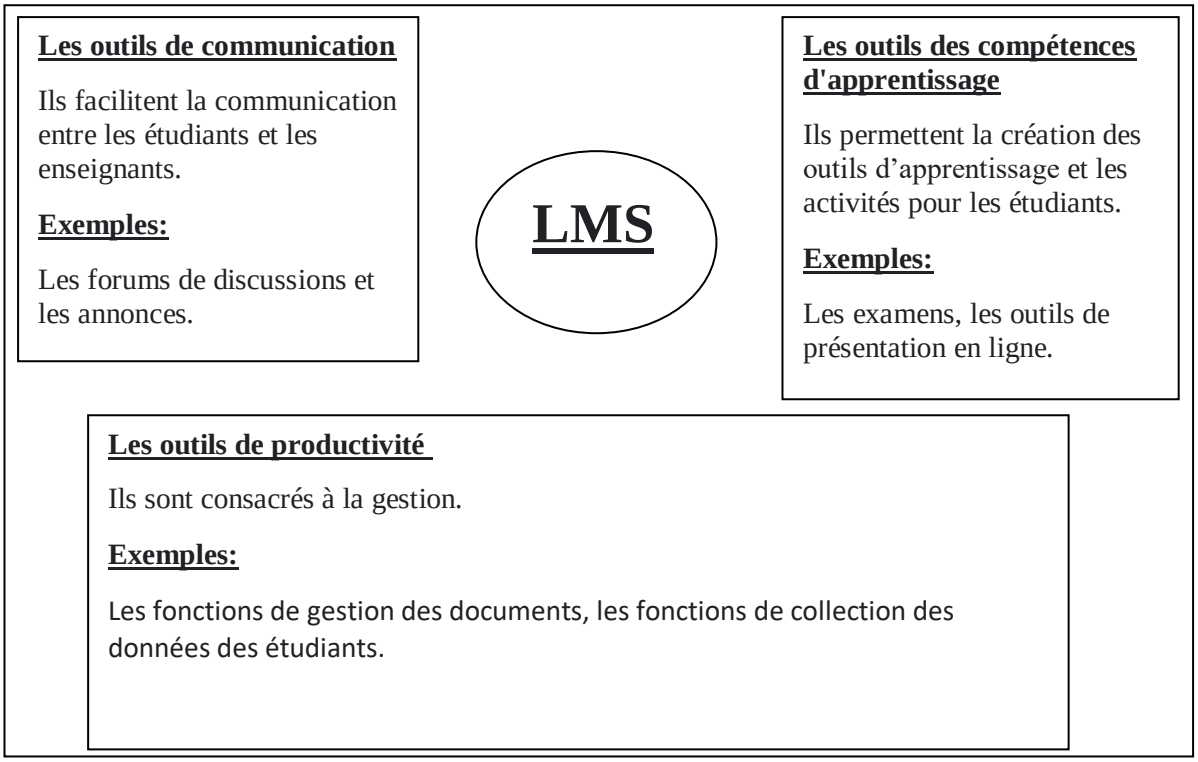

Figure 2.1 Les outils du LMS

D'après la littérature, les LMS sont divisés en trois groupes selon leurs fonctions :

Les outils de communication, les outils des compétences d'apprentissage, et les outils de productivité [80] [81] [82].

Pour le premier type, qui concerne les outils de communication, ils permettent la communication entre les étudiants et aussi entre les étudiants et les enseignants. Les discussions sont aussi des fonctions des outils de communication, car les étudiants et les enseignants peuvent communiquer ensemble, écrire et répondre aux questions et aux messages, et aussi voir les discussions des membres et les commentaires. Parmi les outils de communication on a « announcement », il permet l'annonce de toutes les nouvelles et les informations des cours et les activités à venir qui concernent les étudiants.

Le deuxième type, est les outils d'apprentissage et des compétences, ils regroupent les examens, les quiz, et les outils de présentation en ligne. Les modules de quiz ont des fonctions des bases de données des questions, un système de notation, la facilité et la simplicité des réponses et un moyen de faciliter les performances des étudiants. L'outil de présentation en ligne facilite le téléchargement sur

LMS lorsqu'il a une liaison avec d'autres sites web, par exemple YOUTUBE. Et pour les devoirs, l'enseignant met le devoir sur LMS, puis les étudiants ont la possibilité de répondre en ligne ou la possibilité de faire des modifications avant la date de soumission. Parmi les outils d'apprentissage des compétences on trouve « learning module » qui permet la création des outils d'apprentissage et les activités pour les étudiants.

Le dernier type est les outils de productivité, ils contiennent les systèmes de gestion des calendriers, des documents et des enquêtes. Parmi les fonctions du système de gestion des documents, on trouve la possibilité de chargement et de téléchargement des documents par les étudiants et aussi par les professeurs à l'aide de chaque ordinateur ou appareils mobiles connectés à l'internet. Il y a aussi les fonctions de collection des données des étudiants par exemple le nombre d'accès à LMS, les performances des étudiants. Et aussi les fonctions qui permettent l'accès aux notes de l'étudiant dans les examens, les quiz et les devoirs, et aussi le rapport de performance de l'étudiant. Les outils précédents sont présentés dans la figure 2.1.

La question qui se pose est : quelle est la plateforme idéale ou qui répond à notre besoin ?

Pour régler ce problème on a comparé les plateformes avec une méthode différente aux méthodes classiques.

2.7 Choix de la plateforme E-Learning

2.7.1 Critères utilisés

On a pris les critères à partir des études qui sont déjà faites et on a ajouté d'autres critères selon le but de notre étude qui concerne les performances des étudiants. Les critères utilisés pour la comparaison des systèmes de gestion d'apprentissage, Chacun d'eux contient trois sous critères. Ils sont comme suit :

- L'apprentissage adaptatif :

Il contient les sous-critères suivants : la personnalisation de l'interface du cours, les travaux pratiques par l'apprenant ainsi que l'adaptabilité du contenu selon le profil de l'apprenant.

• La collaboration technologique :

Ce critère prend en considération la possibilité de créer un groupe d'apprenants, des wikis, des forums et aussi la possibilité d'intégrer des outils de collaboration.

- La sécurité :

La sécurité est un critère très important, elle comprend : la validation de la fonction d'entrée, la fonction de cryptage des données, la couche du certificat SSL.

- Mobile:

Ce critère inclut l'existence d'une version Android et d'une version IPad pour ces plateformes E-Learning et qu'elle est Multiplateforme.

- L'évaluation :

Les notes sont les critères qui mesurent le niveau d'apprentissage des étudiants, elles comprennent : les évaluations en ligne, la fonction de notation en ligne, les QCM et les quiz.

- Les rapports d'analyse des apprentissages des étudiants :

Il comprend les statistiques sur les progrès des étudiants : le suivi des évaluations, le suivi des commentaires et le suivi des étudiants.

- La simplicité d'utilisation :

Ce critère prend en compte le temps d'installation et la simplicité de maintenance et de la mise à jour, ainsi que l'ergonomie et la compatibilité avec les différents navigateurs.

- Le prix :

Il présente les prix de la plate-forme en dollars américains.

2.7.2 Méthode utilisée :

Pour trouver la plateforme qui répond à nos besoins, qui permet le suivi des performances des apprenants, nous avons ordonné les plateformes suivantes selon cet objectif : Moodle, OpenedX, Sakai, Claroline, TalentLMS et Easy LMS, en utilisant la méthode MCDM (Multi-Criteria Decision Making). C'est la méthode de prise de décision multicritères. Cette méthode permet de prendre la décision avec la réduction de l'impact de l'incidence.

La première étape consiste à créer l'ensemble de données et pour le faire, nous avons comparé les six plates-formes du système de gestion de l'apprentissage par huit critères : l'apprentissage adaptatif,

la collaboration technologique, la sécurité, mobile, les évaluations, les rapports d'analyse de l'apprentissage, la facilité de l'utilisation et le prix. Et chaque critère contient trois sous-critères. Si la plate-forme a la fonctionnalité de la sous critère, nous lui attribuons 1, sinon nous lui attribuons 0. Ensuite, nous additionnons les points de ces sous-critères pour chaque critère de chaque plate-forme. Le tableau 2.2 présente une étude comparative de Moodle, OpenedX, Sakai, Claroline, TalentLMS et Easy LMS après l'application de la première étape.

Platforms	Moodle	Openedx	Sakai	Claroline	TalentLMS	Easy LMS
L'évaluation	3	2	1	1	2	2
L'apprentissage adaptatif	2	3	2	1	2	2
La collaboration technologique	2	2	3	2	2	2
La sécurité	3	3	2	3	3	2
Mobile	2	2	1	3	3	2
Le rapport d'analyse de l'apprentissage	3	2	2	1	1	2
La facilité de l'utilisation	1	2	2	2	2	3
Le Prix	0	0	0	0	429	250

Table 2-2 Tableau comparatif de Moodle, Open edX, Sakai, Claroline, TalentLMS et Easy LMS après l'application de la première étape.

La deuxième étape consiste à ordonner ces plates-formes, selon les données du tableau 2.2 et de prendre la première plate-forme qui a le score le plus élevé. Mais le problème est que la distribution des critères utilisés n'est pas la même, nous avons donc, nous avons pondéré les valeurs de chaque critère, et aussi nous voulons une plateforme avec des valeurs élevées en l'apprentissage adaptatif, la collaboration technologique, la sécurité, mobile, les évaluations, les rapports d'analyse de l'apprentissage, la facilité de l'utilisation et des valeurs bas pour le critère Prix. Qui se transforme en deux fonctions, maximise () et minimise (). Par conséquent, maximisez les valeurs des critères : l'apprentissage adaptatif, la collaboration technologique, la sécurité, mobile, les évaluations, les rapports d'analyse de l'apprentissage, la facilité de l'utilisation et minimisez les valeurs du Prix. Nous avons donc utilisé la méthode de prise de décision multicritères pour résoudre ce problème.

L'algorithme:

1- Définir les alternatives à classer

2- Définir les critères utilisés pour l'évaluation

3- Définir les poids de chaque critère qui représente l'importance de ce critère. Dans notre cas tous les poids =1.

4- Définir les critères à maximiser et les critères à minimiser.

5- Applique la fonction de maximisation et la fonction de minimisation.

6- Combine les scores par la somme ou le produit des valeurs des plateformes, puis décidez-vous en choisissant la plateforme qui a le score le plus élevé.

- Pour la fonction de maximisation :

F_j: représente la caractéristique à maximiser, et e1,..., en ($1 \leq i \leq n$) ses éléments dans le tableau 1.

Nous avons 3 méthodes : sum normalization, max normalization, max-min scaling.

sum normalization:

$$\max(e_i,F_j)=\frac{e_i}{\sum_i^n e_i} \quad (1)$$

max normalization:

$$\max(e_i,F_j)=\frac{e_i}{\max(F_j)} \quad (2)$$

max-min scaling:

$$\max(ei, F_j)=\frac{e_i-\min(Fj)}{\max(Fj)-\min(Fj)} \quad (3)$$

Pour la fonction de minimisation, nous avons 2 méthodes : la méthode 'inverse' et la méthode 'subtract':

La méthode 'inverse' :
$$\min(e_i, F_j)=\frac{1}{\max(e_i,Fj)} \quad (4)$$

La méthode 'subtract':

$$\min(e_i, F_j)= 1- \max(ei,Fj). \quad (5)$$

2.7.3 Les résultats

Le tableau 2.3 présente l'étude comparative de Moodle, OpenedX, Sakai, Claroline, TalentLMS et Easy LMS et la somme de toutes les valeurs de chaque fonctionnalité sauf le prix afin de les utiliser par la fonction de maximisation et en fonction de minimisation nous utilisons le Prix par la fonction min, c'est la fonction (5).

Les plateformes	Moodle	Openedx	Sakai	Claroline	TalentLMS	Easy LMS	Somme
L'évaluation	3	2	1	1	2	2	11
L'apprentissage adaptatif	2	3	2	1	2	2	12
La collaboration technologique	2	2	3	2	2	2	13
La sécurité	3	3	2	3	3	2	16
Mobile	2	2	1	3	3	2	13
Le rapport d'analyse de l'apprentissage	3	2	2	1	1	2	11
La facilité de l'utilisation	1	2	2	2	2	3	12
Le Prix	0	0	0	0	429	250	429

Table 2-3 Tableau comparatif de Moodle, Open edX, Sakai, Claroline, TalentLMS et Easy LMS.

Nous avons appliqué la méthode de prise de décision multicritères sur les données précédentes en utilisant la normalisation de la somme pour la fonction de maximisation et la soustraction pour la fonction de minimisation. Parce que nous ne pouvons pas diviser par 0 dans la fonction inverse pour le prix.

Ce tableau présente les résultats de la méthode de prise de décision multicritères.

Les plateformes	Moodle	Openedx	Sakai	Claroline	TalentLMS	Easy LMS
L'évaluation	0,27272727	0,18181818	0,09090909	0,09090909	0,18181818	0,18181818
L'apprentissage adaptatif	0,16666667	0,25	0,16666667	0,08333333	0,16666667	0,16666667
La collaboration technologique	0,15384615	0,15384615	0,23076923	0,15384615	0,15384615	0,15384615
La sécurité	0,1875	0,1875	0,125	0,1875	0,1875	0,125
Mobile	0,15384615	0,15384615	0,07692308	0,23076923	0,23076923	0,15384615
Le rapport d'analyse de l'apprentissage	0,27272727	0,18181818	0,18181818	0,09090909	0,09090909	0,18181818
La facilité de l'utilisation	0,08333333	0,16666667	0,16666667	0,16666667	0,16666667	0,25
Le Prix	1	1	1	1	0,36818851	0,63181149
Les scores	2,29064685	2,27549534	2,03875291	2,00393357	1,5463645	1,84480683

Table 2-4 Tableau comparatif de Moodle, Open edX, Sakai, Claroline, TalentLMS et Easy LMS en utilisant les scores.

Après avoir appliqué la fonction "subtract" sur le prix et la fonction "sum normalization" sur les critères, l'apprentissage adaptatif, la collaboration technologique, la sécurité, le mobile, l'analyse de l'apprentissage et la facilité d'"utilisation. Les scores sont calculés par la somme des résultats de chaque plateforme (tableau 2.2). D'après le tableau 2.4, la plateforme qui a le score le plus élevé est Moodle 2.29064685 puis Openedx, Sakai, Claroline, Easy LMS et enfin TalentLMS.

Le score de plateforme est calculé à l'aide de la formule suivante :

$$score(i) = \sum e_j \quad (6)$$

Avec :

i : La plateforme

j : Le critère

La figure 2.2 présente les scores de Moodle, Openedx, Sakai, Claroline, Easy LMS et TalentLMS.

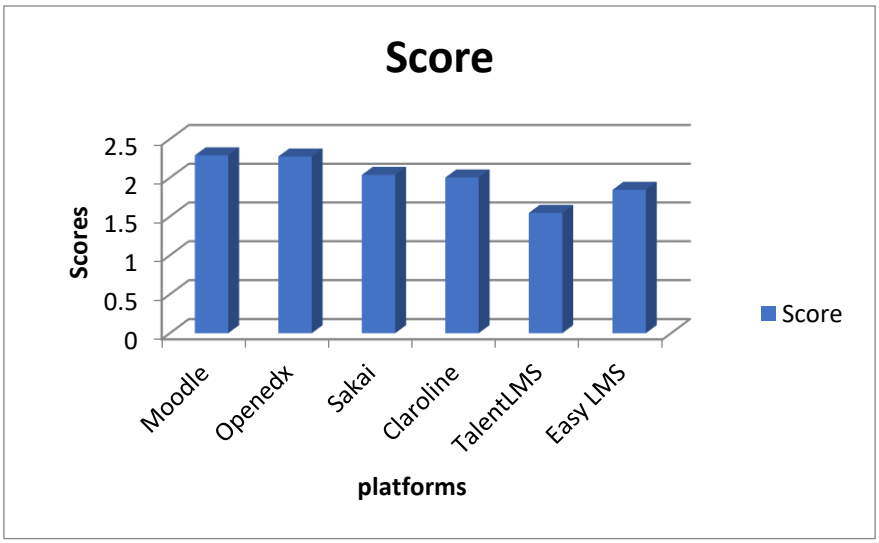

Figure 2.2 Les scores de Moodle, Openedx, Sakai, Claroline, Easy LMS et TalentLMS.

D'après la figure 2.2, il apparaît que moodle et Openedx ont des résultats similaires et les scores les plus élevés au-dessus de 2 points, mais TalentLMS a le score le plus bas inférieur à 1,5 point.

2.7.4 Conclusion

Dans notre étude, et afin d'aider les utilisateurs à choisir le LMS adapté à leur objectif, nous avons présenté une méthode de comparaison des plateformes LMS, en utilisant neuf critères, l'apprentissage adaptatif, la collaboration technologique, la sécurité, le mobile, l'analyse de l'apprentissage, la facilité d''utilisation et le prix . Ces critères ont contrôlé la sélection du meilleur système de gestion de l'apprentissage en fonction de son utilisation prévue. Mais la distribution de ces fonctionnalités n'est pas la même, nous avons utilisé donc l'algorithme de prise de décision multicritères pour résoudre ce problème. Pour tester cette méthode, nous avons comparé six systèmes de gestion de l'apprentissage : Moodle, Sakai, Claroline, TalentLMS, Easy LMS et OpenedX, afin d'étudier les performances des étudiants, où il est devenu clair que Moodle est le meilleur et le plus approprié pour étudier les performances des étudiants.

3. Chapitre III : Les systèmes d'orientation des étudiants à base de TOPSIS et AHP

3.1 Introduction

Après que les étudiants passent les examens du baccalauréat avec la révision et la préparation correspondantes, en plus de la pression psychologique des étudiants, ils se heurtent par une étape très importante, qui est la phase de sélection de la spécialité dans laquelle ils termineront leurs études. Mais le choix de cette dernière est très complexe parce qu'elle est reliée à plusieurs facteurs, à savoir : Les notes de l'étudiant, l'intérêt de l'étudiant, les qualités et les compétences de l'étudiant. Cette étape est très importante, car il influence l'avenir académique de l'étudiant. Un faux choix de spécialité implique l'échec scolaire et même un décrochage scolaire. La méthode utilisée pour l'orientation scolaire des étudiants est basée sur une étude du dossier de l'étudiant par des spécialistes de l'orientation scolaire. Cette méthode est très lente et aussi elle prend en considération seulement les notes de l'étudiant, tandis qu'il existe plusieurs facteurs qui influencent le choix de la spécialité. Donc pour prendre en considération plusieurs facteurs, cette méthode devienne plus lente. La solution est d'informatiser cette méthode d'orientation scolaire .Vu le nombre de critère qui influence le choix de la spécialité, on est arrivé à un problème de décision multi critère (MCDM) [83]. Pour cette raison on a réalisé deux systèmes d'orientation des étudiants qui utilisent deux méthodes de décision multicritères. Le premier système est basé sur une hybridation de la méthode TOPSIS et le gain d'information et le deuxième système est basé sur une hybridation de la méthode AHP et le gain d'information. Les deux systèmes utilisent la méthode SMOTE pour l'équilibrage du nombre des individus dans chaque classe. Dans le but d'améliorer la qualité de prédiction du système de recommandation de spécialité, notre méthode concidère le problème d'orientation des étudiants un problème de classement des spécialités.

Le plan de ce chapitre et comme suit : D'abord, nous avons commencé par la présentation de la méthodologie du travail de notre système et après nous avons présenté les méthodes utilisées (TOPSIS et AHP). Dans la partie suivante, nous avons présenté les résultats associés à chaque méthode et la

comparaison entre les deux systèmes .Et enfin nous avons terminé par une conclusion pour les deux systèmes réalisés.

3.2 La méthodologie

Les systèmes de recommandation de la spécialité sont très sensibles à la qualité des données d'apprentissages utilisés. Si les classes sont bien représentées par les individus, la précision du système sera augmentée. Pour cette raison on a commencé par l'équilibrage des données d'apprentissages. Puis on a passé au classement des spécialités, à ce niveau-là, on a réalisé deux systèmes. Un système basé sur l'hybridation de la méthode TOPSIS et le gain d'information et un autre système basé sur la méthode AHP et le gain d'information. Les deux systèmes ont la même architecture. Ils ne diffèrent que par la méthode utilisée dans le classement.

L'architecture du système de recommandation et comme le suivant :

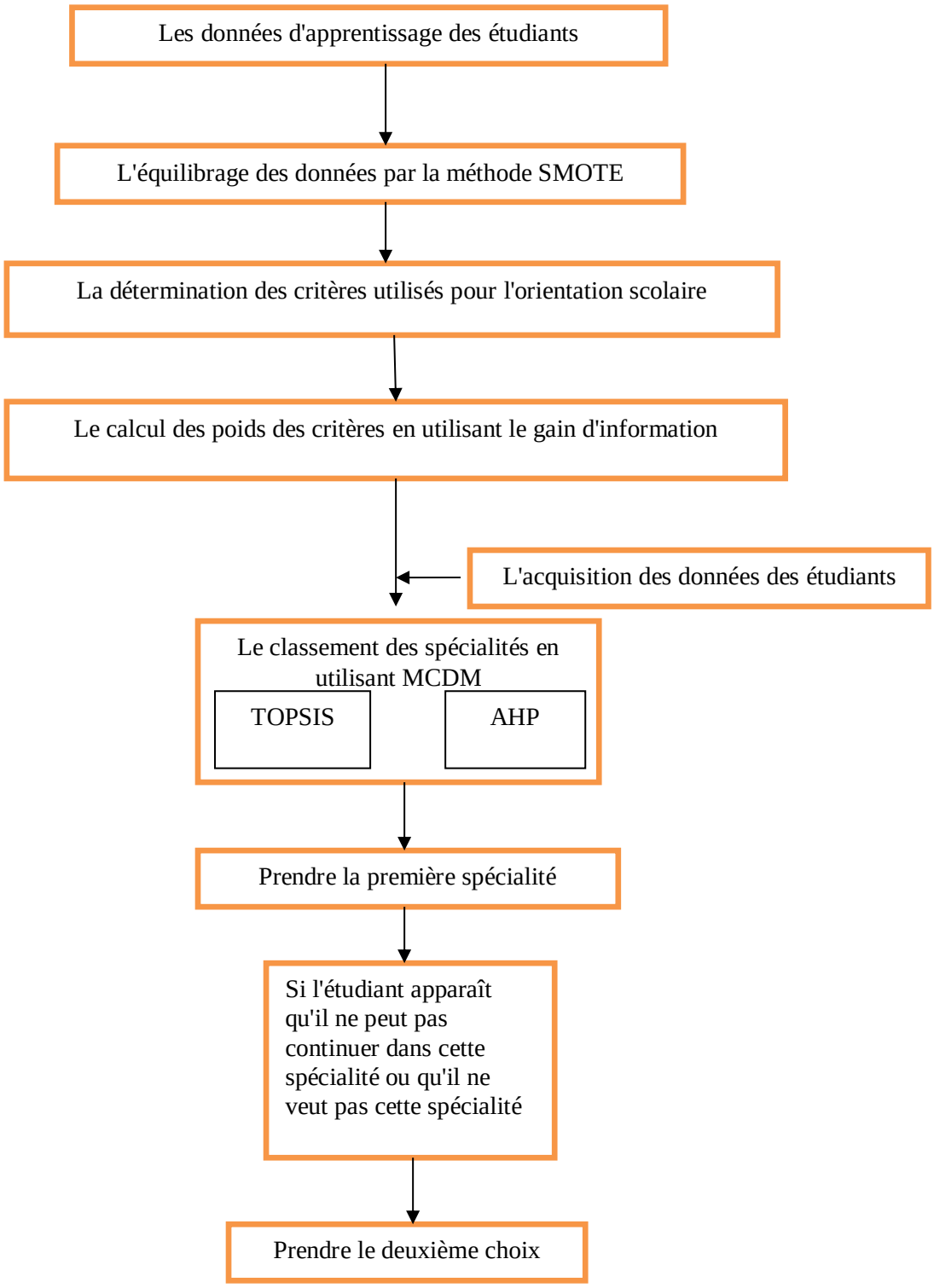

Figure 3.1 L'architecture du système de recommandation de la spécialité.

La première étape de notre système est le prétraitement des données d'apprentissage, dans laquelle on a fait appel à la méthode d'équilibrage des données. Car les données d'apprentissage sont

déséquilibrées. Pour régler ce problème on a utilisé la méthode SMOTE [84] (Synthetic Minority Over-sampling Technique) qui permet de créer des exemples artificiels.

3.3 La méthode SMOTE

SMOTE est une technique de sur échantillonnage. Cette technique augmente le nombre des instances de la classe minoritaire par des nouvelles instances par la méthode d'interpolation.

Les instances de la classe minoritaire qui se trouvent ensemble sont identifiées avant d'être utilisées pour former de nouvelles instances de classe minoritaire. Cette technique est capable de générer des instances synthétiques plutôt que de répliquer des instances de classe minoritaire. Par conséquent, il peut éviter le problème de sur-ajustement.

L'algorithme SMOTE

Début

D : Les données originales

M : Les instances de la classe minoritaire

Pour i ϵ M

Trouver le plus proche voisin j de i qui appartient à M.

diff=i-j.
Ecart=un nombre entre 0 et 1.
r=i+diff*écart
Ajouter r à D.
Fin pour

Fin

Le principe de la méthode SMOTE est présenté par la figure suivante :

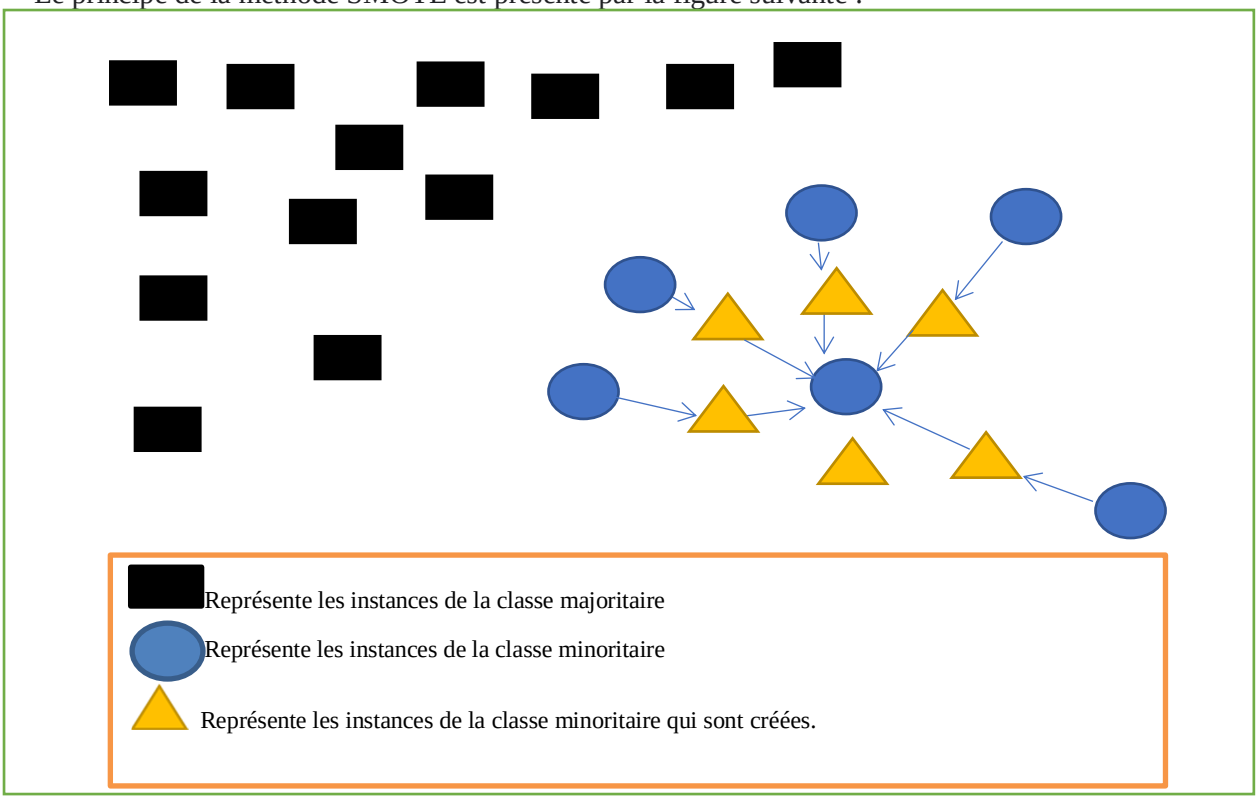

Figure 3.2 Le principe de la méthode SMOTE

La figure précédente présente le principe de la méthode SMOTE en utilisant le plus proche voisin. Après l'équilibrage des données par la méthode SMOTE, on passe à la définition des critères et les alternatives.

3.4 La définition des critères et les alternatives.

Premièrement, pour trouver les coefficients de pondération utilisés par TOPSIS et AHP, on a calculé le poids de chaque critère (attribut) en calculant le gain d'information de chaque attribut à partir la base de données d'apprentissage.

La formule de calcul de l'entropie :

$$\text{Entropie(D)}=\sum_{j=1}^{J}(-p_j)log_2(p_j) \quad (7)$$

Avec :

p_j est la proportion d'exemples de D ayant pour classe résultante j

La formule de calcul de gain d'information :

$$\text{Gain (D, A)} = \text{entropie(D)} - \sum_{v \epsilon V(A)} \frac{|D_v|}{|D|} \text{entropie}(D_v) \quad (8)$$

Avec :

v : Est la valeur de l'attribut A.

Pour les données numériques, elles sont découpées à des intervalles, pour devenir calculable par la formule du gain d'information.

Pour les alternatives, elles représentent les spécialités. On a cinq spécialités. Et pour chaque spécialité on a calculé ses critères, M1, M2, M3, M4, M5, M6 et M7, d'après le portfolio de l'étudiant. comme le suivant :

$$M_1 = \begin{cases} = \frac{0}{10} (si\ l'étudiant\ est\ redoublant\ en\ 2ème\ année\ du\ bacalauréat). \\ = \frac{5}{10} (si\ l'étudiant\ redoublant\ la\ 1ère\ année\ du\ bacalauréat). \\ = \frac{10}{10} (si\ l'étudiant\ n'est\ pas\ redoublant\ ni\ en\ 1ère\ année\ ni\ en\ 2ème\ année\ du\ baccalauréat). \end{cases}$$

$M_2 = \frac{n1+2n2}{3}$

n1 : La moyenne générale de la 1ère année du baccalauréat

n2 : La moyenne générale de la 2ème année du baccalauréat

$M_3 = \frac{examinrégionnal}{12} + \frac{moyennedescontrolecontinue}{36} + \frac{examinnationnal}{12}$

M_4= La note du jury des professeurs de la classe de la 2ème année du baccalauréat

M_5 : Le nombre des heures d'absence pour chaque matière

M_6 : La note du jury de chaque matière.

M_7 : La note du formulaire de chaque matière d'après le formulaire de mesure d'intérêts.

Le calcule de M_3 qui diffère selon la spécialité

La spécialité après l'obtention du baccalauréat : Mathématiques

La matière	Le coefficient selon la spécialité du baccalauréat
	Sciences mathématiques et sciences expérimentales
Mathématiques	4
Physique	3
La langue arabe	0,5
La langue française	1
La langue anglaise	0,5

Table 3-1 Les coefficients des matières pour la branche Mathématiques.

La spécialité après l'obtention du baccalauréat : Physique

La matière	Le coefficient selon la spécialité du baccalauréat
	Sciences mathématiques et sciences expérimentales
Mathématiques	3
Physique	4
La langue arabe	0,5
La langue française	1
La langue anglaise	0,5

Table 3-2 Les coefficients des matières pour la branche Physique.

La spécialité après l'obtention du baccalauréat : Biologie

La matière	**Le coefficient**		
	Science mathématiques	**Sciences expérimentales**	**Sciences agricultures**
Mathématiques	3,5	2,5	2,5
Physique	3,5	2,5	2,5
La langue arabe	0,5	0,5	0,5
La langue française	1	1	1
La langue anglaise	0,5	0,5	0,5
Science naturelle	-	2	1
Science des plantes	-	-	1

Table 3-3 Les coefficients des matières pour la branche Biologie.

La spécialité après l'obtention du baccalauréat : Economie

Matière	**Coefficient**	
	Sciences mathématiques and sciences expérimentales	**Economie**
Mathématiques	5,5	2,5
Economie général	-	1,5
Comptabilité	-	2
La langue arabe	0,5	0,5
La langue française	2	1,5
La langue anglaise	1	1

Table 3-4 Les coefficients des matières pour la branche Economie.

La spécialité après l'obtention du baccalauréat : Techniques

La matière	**Le coefficient selon la spécialité du baccalauréat**
	Technologie électrique et technologie mécanique
Mathématiques	3
Physique	2
La langue arabe	0,5
La langue française	1
La langue anglaise	0,5
Sciences d'ingénieur	2

Table 3-5 Les coefficients des matières pour la branche Technique.

Après le calcul de M_1, M_2, M_3, M_4, M_5, M_6 et M_7 on arrive à un tableau qui contient les spécialités à classer et les critères M_1, M_2, M_3, M_4, M_5, M_6 et M_7.

On remarque qu'on a plusieurs critères pour ordonner les spécialités selon les notes de l'étudiant et ses informations. Ces critères diffèrent en termes d'importance et aussi ils n'ont pas le même intervalle. Donc, pour régler ce problème on a fait recours aux méthodes de décision multi critères.

On a utilisé deux méthodes de prise de décision à critères multiples pour trouver la bonne alternative.

3.5 La méthode TOPSIS

La méthode TOPSIS (Technique of Order Preference Similarity to the Ideal Solution) [85] est une méthode très simple et très utilisée par les chercheurs. Elle permet d'ordonner les alternatives on se basant sur des critères favorables et aussi des critères défavorables. Son principe est basé sur la comparaison de la distance euclidienne de ces alternatives et la solution idéale et aussi à la solution anti-idéale.

La méthodologie de Topsis

La première étape est la construction de la matrice d'entrée qui contient les alternatives et les critères. Elle est sous forme d'Alternatives X Critères.

La deuxième étape : La normalisation de la matrice pour que les critères soient comparables et aussi pour éliminer les unités.

La normalisation des valeurs de la matrice est calculée comme suite (the vector normalization):

$$A_{ij}=\frac{b_{ij}}{\sqrt{\sum_i^m b_{ij}}}$$

i=1,2,….,m

j=1,2,….,n

m: Le nombre des alternatives

n: Le nombre de critères

Il existe d'autres méthodes de normalisation. Pour les critères qu'on veut maximiser, la normalisation est calculée comme suite :

$$A_{ij}=\frac{b_{ij}}{\max(b_j)}$$

Pour les critères qu'on veut minimiser, la normalisation est calculée comme suite :

$$A_{ij}=\frac{b_{ij}}{\min(b_j)}$$

Après la normalisation de la matrice, on passe à la multiplication des valeurs de la matrice par les poids des critères correspondants. Puis on calcule la meilleure solution M et la pire solution P.

Ensuite le calcul de la distance euclidienne des valeurs des alternatives et la meilleure solution et la pire solution.

Enfin on calcule la proximité de chaque alternative et on prend l'alternative idéale.

3.6 La méthode AHP

La deuxième méthode est AHP [86] qui est aussi une méthode de décision multi critères. Elle permet de trouver une solution au problème complexe, en se basant sur plusieurs critères. Le point fort de cette méthode est qu'elle fait une structuration des critères et donne aussi une solution très simple. Cette méthode repose sur quatre principes. La structuration hiérarchique, la structuration des priorités, la cohérence logique et une méthode semi-quantitative.

Les poids des critères sont très importants pour la décision, ils sont faits par les experts du domaine. La méthode la plus utilisée est L'échelle de Saaty. Elle est utilisée pour la comparaison. Elle contient neuf points comme le montre le tableau suivant :

Poid	Signification verbale
1	Faible importance
3	Importance modérée
5	Forte importance
7	Très forte importance
9	Importance absolue
2, 4, 6, 8	Ils sont utilisés pour les jugements intermédiaires.

Table 3-6 Les poids des critères et ses significations.

Cette méthode ne permet pas de trouver les poids des attributs lorsque l'importance de ces attributs est très proche. Pour cette raison on a utilisé le gain d'information pour trouver ces poids.

Les étapes de la méthode AHP

- La construction de la matrice de comparaison

$$M = \begin{pmatrix} b11 & \cdots & b1m \\ \vdots & \ddots & \vdots \\ bm1 & \cdots & bmm \end{pmatrix}$$

- la normalisation de la matrice, après le calcul de la moyenne géométrique de chaque ligne en utilisant la formule suivante :

$$Di=\frac{\sqrt[m]{\sum_{j=1}^{m} b_{ij}}}{\sum_{i=1}^{m} \sqrt[m]{\sum_{j=1}^{m} b_{ij}}}$$

Maintenant, on calcule λ_{max} .

On note :

$$MxD=N$$

$$\lambda_{max}=\frac{\sum_{i=1}^{m} N_i}{m}$$

Après λ_{max}, on passe au calcul de CI et CR.

$$C.I=\frac{\lambda max-m}{m-1}$$

$$C.R=\frac{C.I}{R.I}$$

R.I : index aléatoire

On répète les étapes précédentes jusqu'à on arrive à une valeur proche à la valeur souhaitée.

3.7 Les résultats expérimentaux

Pour comparer les résultats des deux modèles créés, on est basé sur la précision de la prédiction de la spécialité pour chaque spécialité et pour chaque modèle. Et aussi la précision totale pour chaque modèle. Ces critères de comparaison sont testés avant et après l'application de la méthode SMOTE. Puis on a comparé les deux modèles selon la complexité.

La figure 3.3 présente la précision de la prédiction de la spécialité pour chaque spécialité et aussi pour chaque modèle.

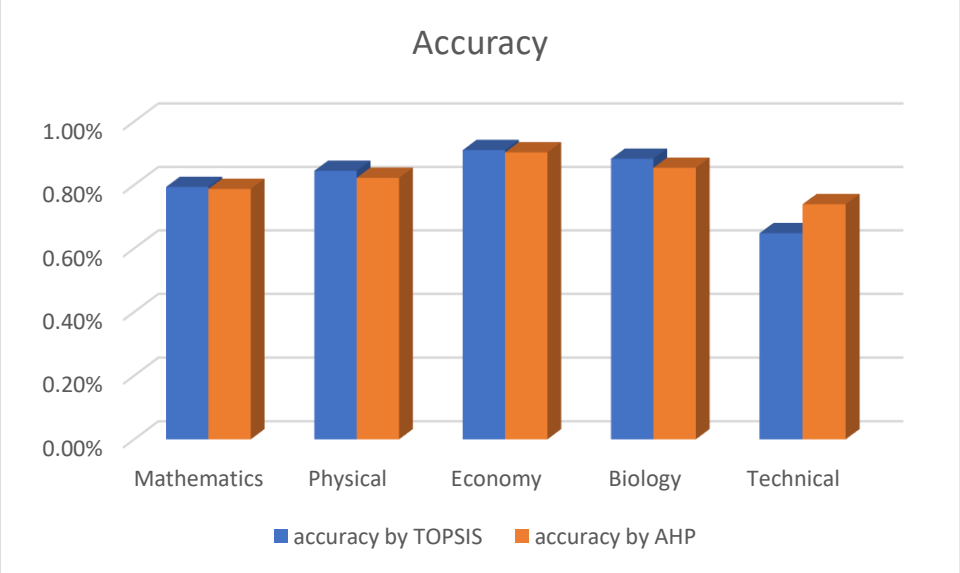

Figure 3.3 La précision de la prédiction de chaque spécialité par le modèle basé sur TOPSIS et le modèle basé sur AHP avant l'équilibrage des données.

La figure 3.3 présente une comparaison basée sur la précision pour chaque spécialité du modèle basé sur la méthode TOPSIS et du modèle basé sur la méthode AHP. Le modèle basé sur la méthode TOPSIS est plus précis que le modèle basé sur la méthode AHP pour la prédiction des spécialités, Mathématiques, Physique, Economie et Biologie. Alors que le modèle basé sur la méthode AHP est

plus précis pour la prédiction de la spécialité Technique. Et aussi on remarque que la précision de la prédiction de la spécialité des deux modèles est très haute pour la spécialité "Economie", alors que la précision de prédiction des deux modèles est basse pour la spécialité "Technique ".

La précision totale des deux modèles est présentée dans la figure 3.4.

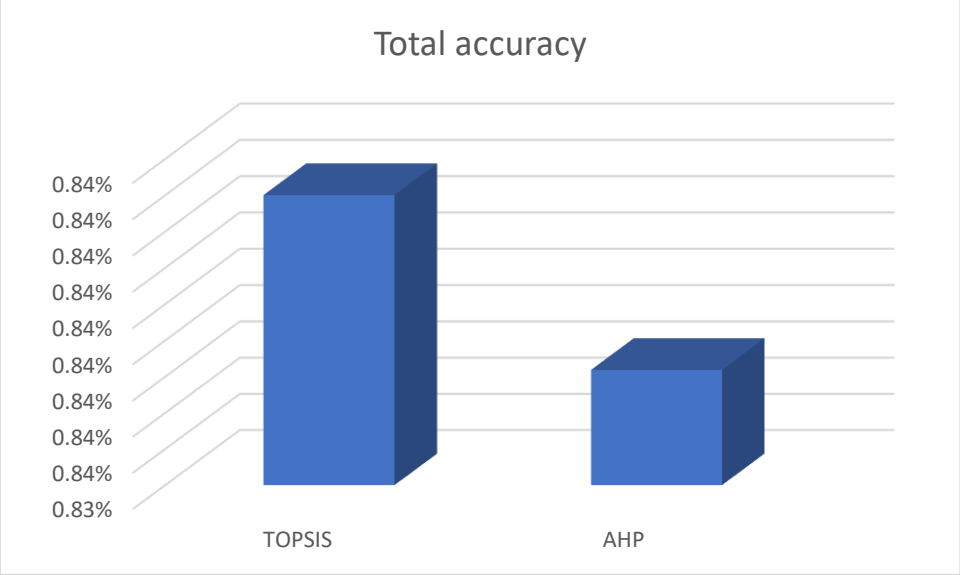

Figure 3.4 La précision totale des deux modèles avant l'équilibrage des données.

D'après la figure 3.4, la précision de la prédiction des spécialités du modèle basé sur la méthode TOPSIS et plus élevé que la précision de la prédiction du modèle basé sur la méthode AHP. Après l'application de la méthode SMOTE pour rééquilibrer les données, on a obtenu les résultats de la figure 3.5. Elle présente la précision de la prédiction de la spécialité pour chaque spécialité et aussi pour chaque modèle après l'application de la méthode d'équilibrage des données.

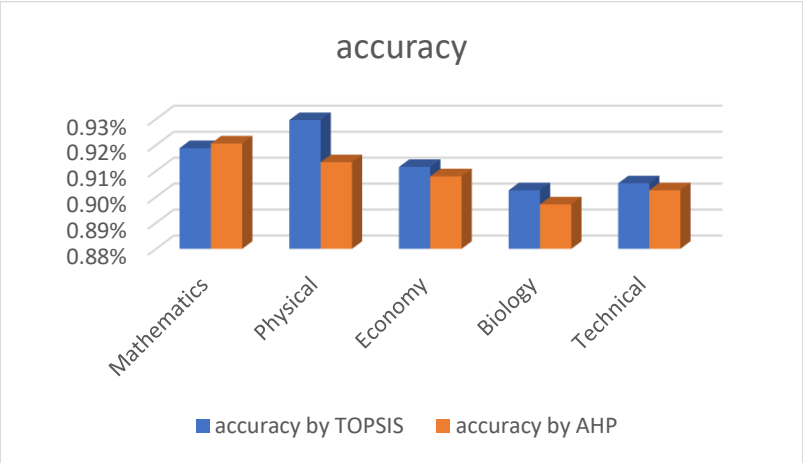

Figure 3.5 La précision de la prédiction de chaque spécialité par le modèle basé sur TOPSIS et le modèle basé sur AHP après l'équilibrage des données.

D’après la figure 3.5, la précision de la prédiction des spécialités pour chaque spécialité des deux modèles est augmentée. Mais, après l’application de la méthode SMOTE, la précision de la prédiction de la spécialité "Mathématiques" du modèle basé sur TOPSIS est devenue inférieure à la précision de l’autre modèle, et aussi la précision de la spécialité "Technique" du modèle basé sur TOPSIS est devenue plus élevée que l’autre modèle.

Les résultats de la précision totale des deux modèles sont présentés dans la figure 3.6.

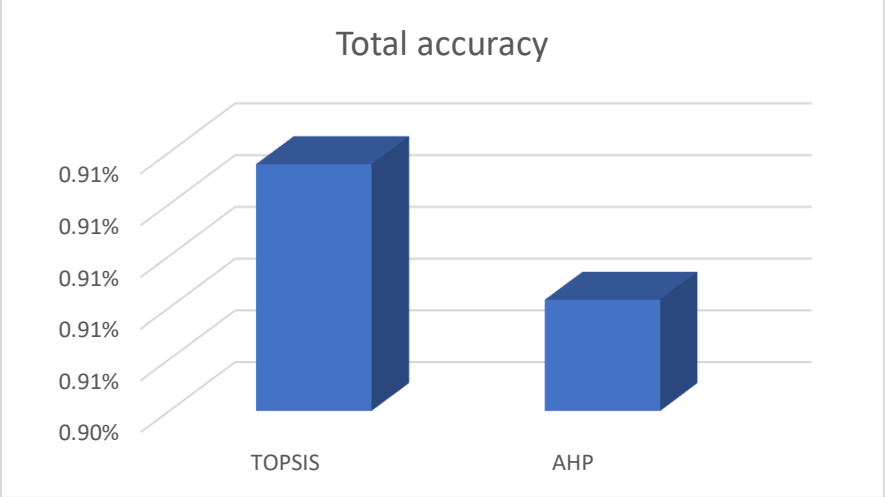

Figure 3.6 La précision totale des deux modèles après l’utilisation de la méthode SMOTE.

La figure 3.6 montre que la précision totale des deux modèles est augmentée après l’utilisation de la méthode SMOTE. Et que le modèle basé sur la méthode TOPSIS et plus précis que le modèle basé sur la méthode AHP.

Pour comparer les deux modèles à base de la complexité. On a utilisé les formules suivantes pour calculer la complexité des deux modèles.

D’après [87] la complexité de TOPSIS est calculée par cette formule :

$C=mn+mn+m(n+1)+m(n+1)+m=4mn+3m.$

Avec :

n : Le nombre de critères

m : Le nombre des alternatives

La complexité de la méthode AH P est calculée par cette formule :

$C= n(n+1)+m(n+1)+mn.$

Les résultats obtenus après le calcul de la complexité des deux méthodes selon le nombre des alternatives sont présentés dans la figure 3.7.

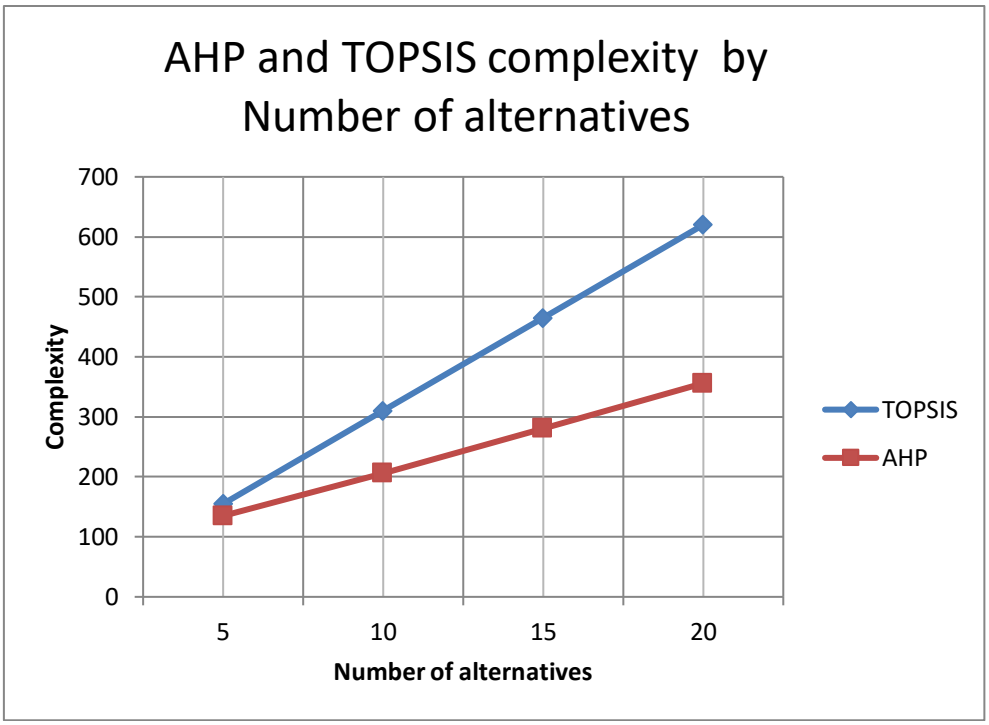

Figure 3.7 La complexité de la méthode TOPSIS et la méthode AHP par le nombre des alternatives.

D'après la figure 3.7, la complexité de la méthode TOPSIS en fonction du nombre des alternatives est augmentée de plus que la complexité de la méthode AHP.

Les résultats obtenus après le calcul de la complexité des deux méthodes selon le nombre des critères sont présentés dans la figure suivante.

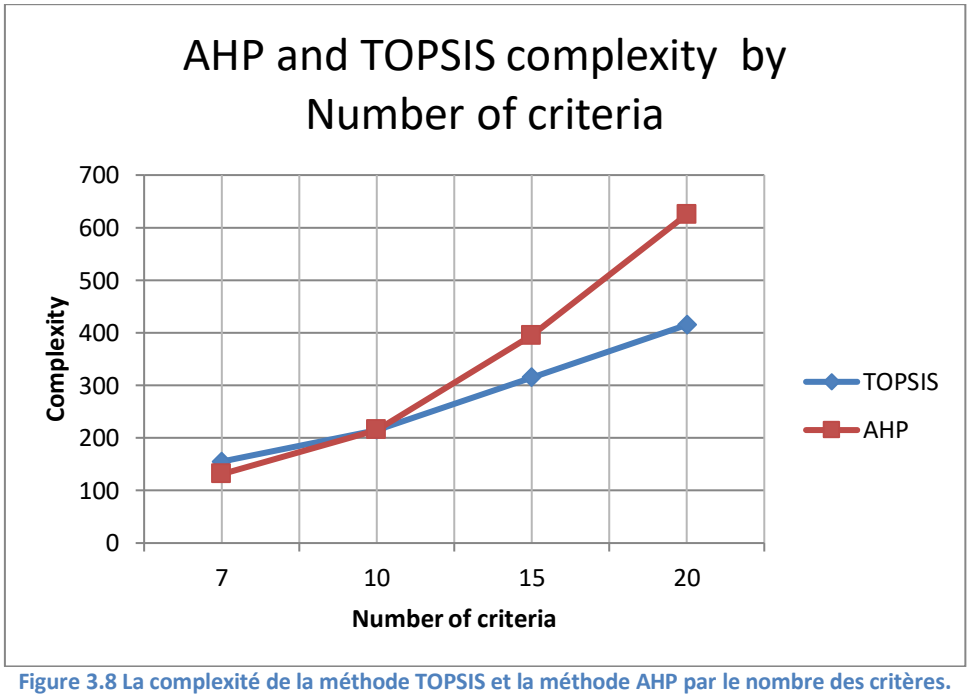

Figure 3.8 La complexité de la méthode TOPSIS et la méthode AHP par le nombre des critères.

D’après la figure 3.8, la complexité de la méthode AHP en fonction du nombre des alternatives est augmentée de plus que la complexité de la méthode TOPSIS.

On remarque que la complexité de la méthode TOPSIS est plus élevée que la complexité de la méthode AHP.

3.8 Conclusion

Le système réalisé est très usuel car, il permet l’orientation des étudiants par la proposition de la spécialité adéquate à leurs compétences et à leurs profils, et aussi il permet la réorientation au cas où l’étudiant a rencontré des défis scolaires reliés à la matière enseignée. Dans ce cas, on prend la deuxième spécialité proposée par le système pour lutter contre l’échec scolaire ou le décrochage scolaire.

Les résultats de la comparaison des deux systèmes réalisés, montrent que le système d’orientation et de réorientation basé sur la méthode TOPSIS est plus précis que le système basé sur la méthode AHP. Mais ce dernier est plus rapide que le système basé sur la méthode TOPSIS. La précision des deux systèmes est augmentée après l’utilisation de la méthode SMOTE. Car, l’inégalité du nombre des individus de chaque classe (spécialité) dans la base de données d’apprentissage influence l’efficacité du système, ce qui signifie l’augmentation de la précision des deux systèmes après l’application de la méthode SMOTE.

4. Chapitre IV : L'orientation scolaire des étudiants à l'aide du Big Data

4.1 Introduction

L'orientation de l'étudiant est un processus important et difficile car, pour prendre la décision concernant l'être humain est très compliqué, mais c'est un processus qui dépend principalement des notes de l'apprenant en premier lieu et du désir au second degré. Par exemple, si l'étudiant aime une spécialité, mais il n'a pas les capacités suffisantes pour lui, il ne pourra pas suivre le rythme du programme d'études de cette spécialité. C'est pourquoi l'orientation pédagogique est considérée comme une étape cruciale dans le cursus de chaque élève, plus particulièrement des lycéens. Malheureusement, les élèves du secondaire se retrouvent toujours confrontés à ce problème d'orientation, car lorsqu'ils sont au secondaire, ils ne peuvent pas encore décider de leurs choix d'orientation, ce qui empêche chacun de choisir directement sa spécialité de préférence, ce qui le rend frustré et qui peut conduire à l'abandon des études [88].

De multiples facteurs influencent l'orientation des étudiants, principalement des données sociales qui n'influencent pas les résultats scolaires et sont loin de prendre en compte les caractéristiques de tous les étudiants. Deuxièmement, les familles savent ou entendent que, selon la spécialité du baccalauréat obtenu, les possibilités de poursuite d'études et d'accès à l'enseignement supérieur et à l'insertion professionnelle diffèrent, surtout lorsque le marché du travail est tendu. Pour résoudre ce problème, nous essayons de créer un système pour aider l'étudiant à y parvenir. Compte tenu du nombre d'étudiants et du besoin de temps, nous avons décidé d'utiliser le Big Data.

Dans la littérature, certains travaux connexes ont comparé les méthodes d'apprentissage automatique telles que Sunita B. Aher et Lobo L.M.R.J. [89], ils ont comparé sept algorithmes de classification : Naïve Bays, Simple Cart, ZeroR, J48, Table de Décision, ADTree et Random Forest, en utilisant Weka, puis ils ont constaté qu'ADTree fonctionne mieux pour la base de données Moodle. Alors que Seyed Reza Pakiz et Abolfazl Gandomi [90] ont comparé quatre algorithmes de classification, mais en utilisant le modèle Mapreduce avec des modèles traditionnels, ils ont conclu

que les algorithmes de classification basés sur le modèle Mapreduce fonctionnent mieux dans les grands ensembles de données. De même, Wael Etaiwi, Mariam Biltawi et Ghazi Naymat [91] ont comparé deux classificateurs d'apprentissage automatique, Naïve Bayes et Support Vector Machine (SVM) en utilisant MLlib, d'Apache Spark, et ils ont conclu que Naïve Bayes est plus puissant que SVM pour Big Data. Dans un autre article, Amine Rghioui, Jaime Lloret et Abedlmajid Oumnad [92] comparent J48, Bayes Net, ZeroR et Naïve Bayes en utilisant des données de santé, et ils ont conclu que j48 est meilleur que l'autre algorithme de classification avec une précision de 99,21%, une autre étude [93] comparent Naïve Bayes, les réseaux de neurones et les k-plus proches voisins, par précision et le temps de la classification. Ils constatent que l'algorithme de Naïve Bayes fonctionne mieux, mais dans l'apprentissage en ligne [94][95][96] et précisément dans l'orientation des étudiants, il n'y a pas d'études qui pourraient faciliter ce processus. Pour cela, nous essayons d'utiliser le Big Data pour aider les étudiants dans leur orientation.

Ce chapitre est structuré comme suit. Le premier sous chapitre présente l'architecture du système d'orientation basée sur la technologie Big Data. Alors que le deuxième chapitre présente le système de l'orientation scolaire des étudiants à l'aide du Big Data sous Map Reduce et sous Weka dans le troisième sous chapitre. Le quatrième sous chapitre est concacré aux résultats obtenus, et enfin une conclusion du chapitre.

4.2 L'architecture du système d'orientation basée sur Big Data

La figure suivante présente l'architecture générale du système de recommandation

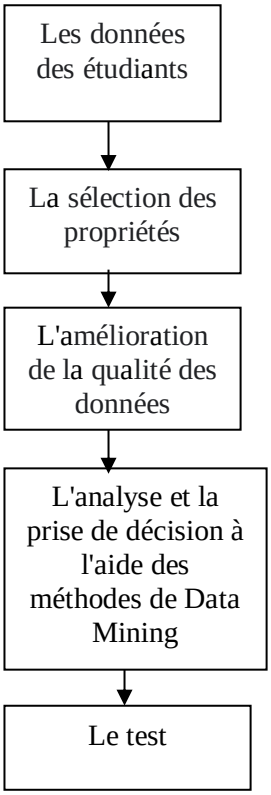

Figure 4.1 L'architecture du système de recommandation

La première étape est l'acquisition des données. Dans cette étape on utilisé une base de données qui s'appelle OULAD d'après la plateforme Kaggle. Cette base de données contient un ensemble de données sur le comportement et les performances des étudiants. Il contient des informations sur 22 cours, 32 593 étudiants, leurs résultats d'évaluation et les journaux de leurs interactions avec le VLE représentés par des résumés quotidiens des clics des étudiants (10 655 280 entrées). La deuxième étape est la sélection des caractéristiques à l'aide des méthodes de sélection des attributs. Et après, on passe au nettoyage et la transformation des données et après on passe à l'étape la plus importante, qui est l'analyse de données et la prise de décision par l'application des méthodes de Data Mining. Et enfin le test du système de recommandation.

Dans ce chapitre on va présenter deux modèles de recommandation de spécialité. Les deux modèles sont basés sur la technologie du Big Data.

4.3 L'orientation scolaire des étudiants à l'aide du Big Data sous Map Reduce

Les méthodes d'orientation des étudiants sont basées sur les notes des étudiants seulement. Alors que l'orientation des étudiants est liée à plusieurs facteurs : les notes des étudiants durant les années du collège et du lycée pour que le modèle soit très robuste, le nombre d'absences non justifiées de chaque matière et aussi la note de participation en classe pour mesurer l'intéresse et les inclinaisons des étudiants, les jugements du jury du conseil départemental, le nombre des années redoublantes.

Vu au nombre progressif des étudiants et des spécialités et aussi le nombre de facteurs, les systèmes d'orientations existant n'ont pas la capacité de stocker et de traiter ces données massives ainsi que les étudiants veulent la décision en temps réel. Notre solution est d'utiliser la technologie Big Data.

Dans la première méthode on a utilisé le modèle suivant :

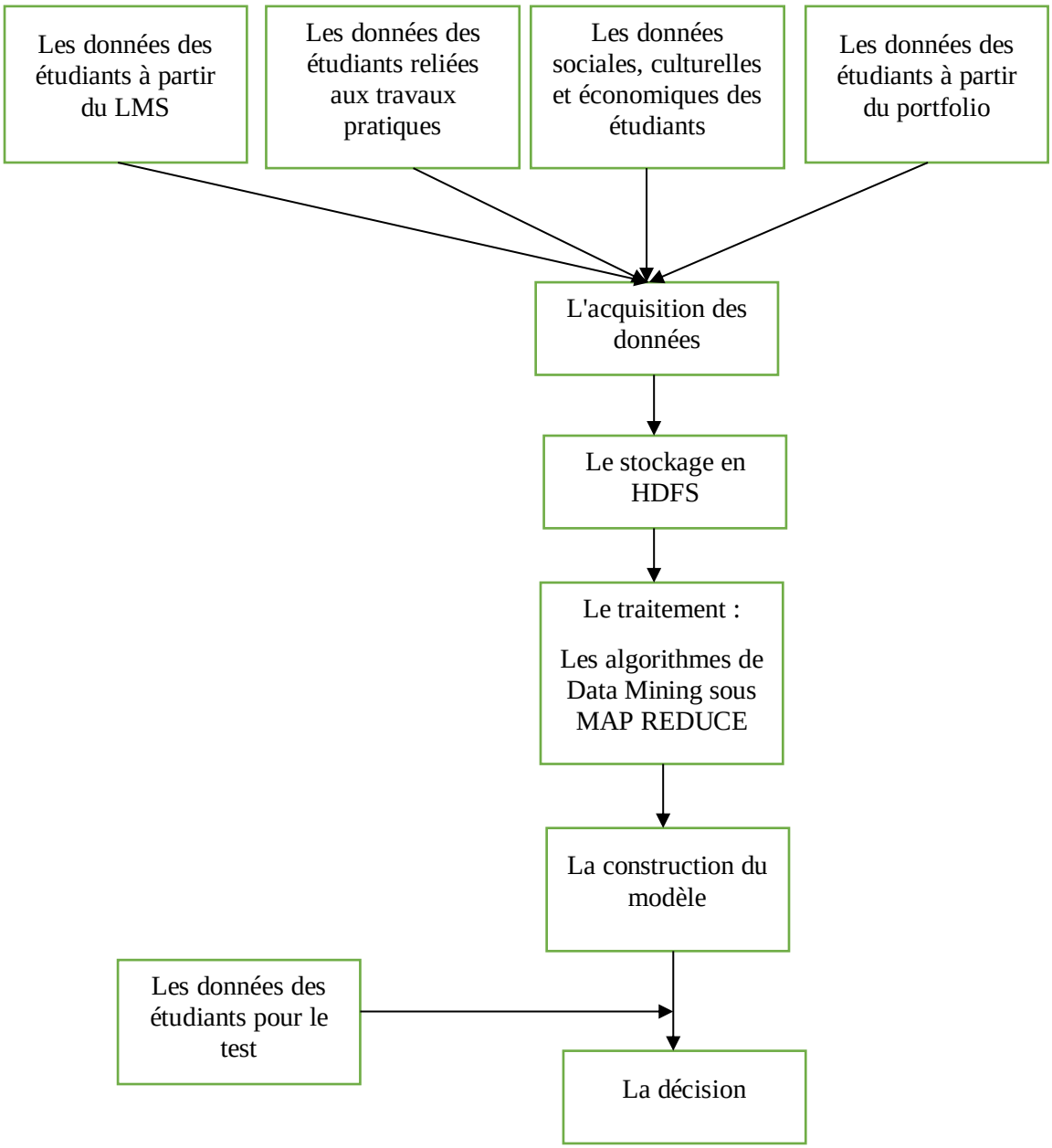

Figure 4.2 L'architecture du système de recommandation de spécialité utilisée.

Les données des étudiants sont hétérogènes et aussi parvient des sources multiples, ce qui rend l'utilisation du Big Data très important. Ces données sont acquises et stockées en HDFS qui permet le stockage réparti et la tolérance contre les pannes par la redondance des données ainsi que la vitesse de stockage. Après cette étape, ces données sont utilisées pour créer un modèle d'orientation des étudiants à l'aide des algorithmes de classification. Une fois le modèle d'orientation est construit, on peut l'utiliser pour le test et enfin la décision (la spécialité adéquate à l'étudiant).

Les algorithmes de classification qui sont utilisés dans ce modèle, sont : Les réseaux de neurones, l'algorithme k-plus proche voisins et les réseaux bayésiens.

4.3.1 Les réseaux de neurones

Un réseau de neurones est un algorithme de classification supervisé. L'idée générale des réseaux de neurones est inspirée du fonctionnement des neurones des êtres vivants. Il est constitué de trois couches : une couche d'entrée de neurones, une couche ou deux ou trois couches cachées et une couche de sortie. Voilà un exemple de réseau de neurones.

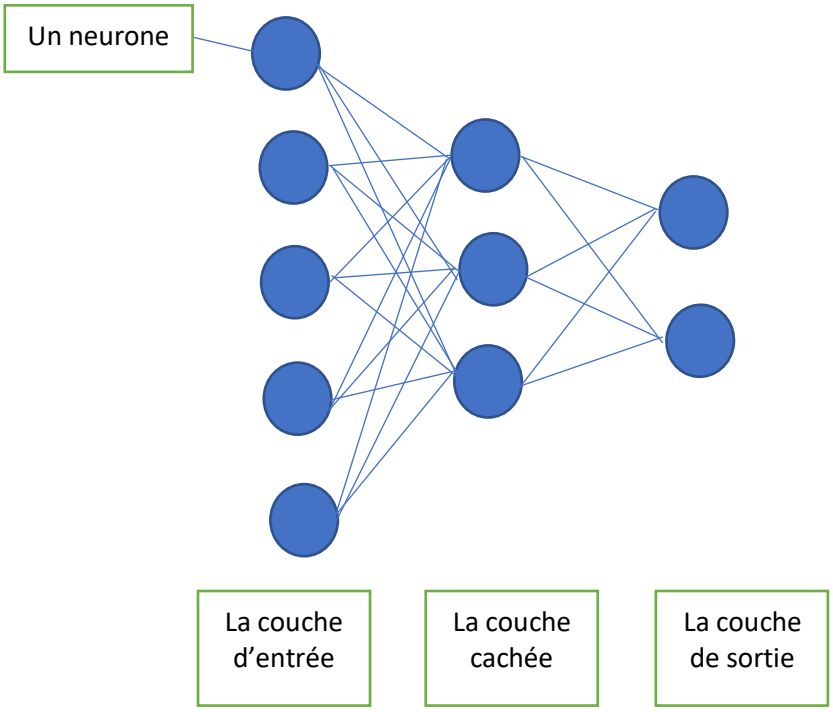

Figure 4.3 L'architecture du réseau de neurones.

Le nombre de neurones de la couche d'entrée est le nombre de propriétés et le nombre de neurones de la couche de sortie est le nombre de classes de sortie. Chaque relation entre les neurones a un poids. La formule générale utilisée par les réseaux de neurones est la suivante :

$$S_i = \sigma(\sum_{j=1}^{N} W_{i,j} x_j + \theta_i^{Cid}) \quad (8)$$

Avec :

S_i: La sortie du neurone i de la couche cachée.

σ : La fonction d'activation.

$W_{i,j}$: Le poids

N : Le nombre de neurones d'entrée.

x_j : Les entrées du neurone d'entrée

θ_i^{Cid} : Le seuil

La fonction d'activation σ est définie comme suit :

$$\sigma(t) = \frac{1}{1+\exp(-t)} \quad (9)$$

L'algorithme du réseau de neurones utilisé

Entrée :
T,D
Sortie :
C
p mappeurs, 1 reducteurs
1- La construction de rétro-propagation avec p entrées entrées, l sorties, s neurones dans la couche cachée par chaque mappeur.
2 - Initialise ω_{ij},θ_{2j}
3 -$\forall$ t ϵ T, ti ={a_1, a_2,..., a_n}
Entrée ai→pi, neurone j en couche cachée
Ijs = $\sum_{i=1}^{p} a_i.\omega_{ij} + \theta_j$
Ljs = $1/(1 + e\text{^}Ijs)$
4- Entrée lj → sortie i, neurone j en couche de sortie
Ijl = $\sum_{i=1}^{s} L_{js}.\omega_{ij} + \theta_j$
Ljl = $1/(1 + e\text{^}Ijl)$
5 Pour chaque sortie
ERjl= Ljl (1- Ljl)($cible_j$- Ljl)
6- Pour chaque couche cachée
ERjs = Ljs(1- Ljs)$\sum_{i=1}^{l} ER_i.\omega_{il}$
7 – Mise à jour
ω_{ij}= ω_{ij} + µ . ERj .Lj

$\theta_j = \theta_j + \mu$. ERj
repète 3,4,5,6,7
jusqu'à min (E[e^2])= min (E$\left[(\text{cible}_j - \text{Ljl})^2\right]$)

8- Divise T
9- Chaque mappeur execute (3),(4)
10- Les sorties du mappeur (tj,Lj)
11- Recueille des réducteurs
repète (9),(10),(11)
jusqu'à T est traversé
12- Les sorties des réducteur C
Fin

Figure 4.4 L'algorithme du réseau de neurones utilisé.

4.3.2 L'algorithme k-plus proches voisins.

L'algorithme k-plus proches voisins ou KNN (K-nearest neighbors) est un algorithme de classification. Selon son nom, il calcule la distance entre l'instance à classifier et chaque individu de la base d'apprentissage puis, il cherche les k individus qui ont la distance minimale. Il est aussi utilisé pour la régression. Le calcule de distance est relié aux types de données utilisées, dans notre cas on a des données quantitatives.

On peut utiliser les distances suivantes :

La distance Euclidienne : $\sqrt{\sum_{j=1}^{n}(x_j - y_j)^2}$ (9)

La distance de Manhattan : $\sum_{j=1}^{n}|x_j - y_j|$ (10)

La distance Minkowski : $\sqrt[d]{|x_j - y_j|^d}$ (11)

La distance de Canberra : $\sum_{j=1}^{n}\frac{|x_j - y_j|}{|x_j| + |y_j|}$ (12)

La distance de Tchebychev : $\sup_{1 \le j \le n}|x_j - y_j|$ (13)

L'algorithme du knn utilisé :

Map du knn

L'entrée:

- Dtrain = (d1train, ..., drtrain);// les données d'apprentissage ;
- LC train = (lc1train, ..., lc rtrain);// les classes des données d'apprentissage:
- Dtest= (d1test, ..., dttest);//les données du test
- Liste l;

L'algorithme

Pour i du 1 à t faire

//Un test par chaque exécution du Map

/Enregistrement de la distance de tous les individus de données d'apprentissage*/

Pour j de 1 à r faire

Calculer la distance euclidienne entre $D_j$train et $D_i$test en utilisant l'équation (9)

dj ← d($d_i$test , $d_j$train);

l.add(d);

Fin map();

Reducer();

Reducer

Int l, k;// k et l: entier;

- Calculer la classe lc_itest du $i^{ème}$ exemple, qui est vaut la classe de son plus proche voisin :

-Trouver l'index du ppv de $d_i$test :

ind_ ppvi ← arg minnj=1 dj

- trouver la classe du ppv de $d_i$test

(où $d^{train}_{ind\_ppvi}$):

$lc_i^{test}=lc^{train}_{ind\_ppvi}$

Fin

La sortie

les classes des données de test LCtest= (lc_1^{test}, ..., lc_n^{test}).

Figure 4.5 L'algorithme knn utilisé.

4.3.3 L'algorithme Naïve Bayes

L'algorithme Naïve Bayes est un algorithme basé sur la probabilité bayésienne et l'hypothèse de l'indépendance forte. C'est-à-dire que la probabilité d'un attribut n'est pas liée à la probabilité de l'autre. L'algorithme Naïve Bayes utilise le théorème de bayés suivante :

$$P(c|f1, f2,..., fn)=\frac{p(c)p(f1,f2,...,fn|c)}{p(f1,f2,...,fn)}$$

Avec :

n: Le nombre des attributs

Donc le nombre des hypothèses indépendantes est (2n)!

Généralement l'algorithme Naïve Bayes donne des bons résultats. Mais, il existe des facteurs qui influencent les résultats de cet algorithme. Selon [97], il y a trois facteurs d'erreur de l'algorithme Naïve Bayes. Les données d'apprentissages bruitées, le biais lorsque la taille des regroupements de données d'apprentissage est grande, et la variance, lorsque la taille des regroupements de données d'apprentissage est petite.

L'algorithme Naive Bayes utilisé :

Map :

L'entrée : Les données d'apprentissage (Les attributs et les classes)

La sortie : Une paire (clé, valeur)

Charger les données d'apprentissage qui sont sous forme des lignes.

Créer une clé pour chaque individu (classe, nom_attibut, valeur_attribut)

Régler la valeur de toutes les valeurs des attributs sur 1

La sortie (clé, valeur)

Reduce :

L'entrée : Les clés, et L : la liste de toutes les valeurs correspondent à chaque clé.

La sortie : (clé', valeur')

somme=0 ;

Pour chaque valeur en L :

Pour j=1 jusqu'à n :

somme=somme+1 ;

La sortie (clé', valeur')

clé' : Les clés ;

valeur' : le nombre des occurrences d'un string dans la sortie des mappeurs.

Figure 4.6 L'algorithme Naïve Bayes utilisé.

4.4 L'orientation scolaire des étudiants à l'aide du Big Data sous WEKA.

Dans la deuxième méthode on a utilisé l'outil SMO de Weka.

Weka est un outil d'exploration des données sous licence publique GNU [98]. Il est open source et aussi il est écrit en java. Il est créé par l'université WAIKATO en Nouvelle-Zélande. Il fournit aux

utilisateurs un ensemble d'algorithmes et des méthodes, à savoir : Le filtrage, la classification, les règles d'associations, le groupement, la régression et la visualisation.

Weka a plusieurs avantages :

- Il est libre.
- Il est portable, car il est implémenté dans le langage de programmation Java et fonctionne sur toutes les architectures.
- Il fournit un large ensemble de techniques de modélisation et aussi de prétraitement des données.
- Il est très facile à utiliser, car il fournit une interface graphique à ses utilisateurs.
- La 3ème version de weka, permet aussi le traitement du big data grâce à l'outil MOA.

MOA (Massive Online Analysis) est un Framework open source créé par java, il est aussi portable et extensible avec les nouveautés. Il permet la gestion des flux de données massifs, évolutifs et infinis. Il est conçu aussi pour l'implémentation des algorithmes pour l'apprentissage en ligne, en utilisant des données évolutives. Ce qui rend MOA très utilisable pour résoudre les problèmes du monde réel. Comme il permet de comparer les algorithmes de classification simples ou multi-étiquettes, ainsi que les algorithmes de clustring

On a utilisé les algorithmes de classification sous Weka.

4.4.1 Les machines à vecteurs de support (SVM)

Les machines à vecteurs de support (SVM) sont des techniques très connues en reconnaissance des formes. Elles sont utilisées pour l'apprentissage automatique, en utilisant une méthode basée sur les statistiques afin de résoudre le problème de fonction d'approximation. Elles ont créé par Vapnik.

Pour séparer les classes, les méthodes antérieures utilisent un hyperplan. Mais le problème est que parfois, on rencontre un problème avec des classes non séparables. Et pour régler ce problème, les SVM cherchent un hyperplan dans une autre dimension supérieure à l'aide d'une fonction noyau, pour rendre ces classes linéairement séparables. Parfois des erreurs parviennent en classification et cela est dû à la sélection d'une fonction noyau inadéquate. Mais, le problème est de trouver le bon emplacement de l'hyperplan de décision et de sélectionner les frontières linéaires, pour diviser les deux

classes linéairement par un hyperplan avec une marge maximale, afin d'améliorer les résultats de la classification.

L'ensemble du support de vecteur du SVM est défini par :

$$V= \{(e_i, s_i) \mid e_i \in IR^d , s_i \in \{-1,1\}\}^n_{i=1} \quad (17)$$

Avec :

e_i : Est un individu des données d'apprentissage caractérisé par d attributs.

s_i : Est la sortie correspond à l'individu e_i .

La formule de Vapnik :

$$F(x) = \sum_{j=1}^{m} w_j x_j + b$$

L'algorithme SVM sous map reduce

- La phase map :

V=ø // V est un vecteur vide ;

Tant que $B^i \neq B^{i-1}$, faire // Tant que l'hypothèse actuelle B^i est différente de l'hypothèse précédente B^{i-1}

Pour mϵM // M est le nombre de map ou le nombre des machines.

$T_m^i \leftarrow T_m^i \cup V$ // L'échantillon des données à la machine m.

Fin pour

Fin tant que

- La phase reduce

Tant que $B^i \neq B^{i-1}$ faire

Pour mϵM

/*les données fusionnées sont utilisées pour l'entrainement, afin d'avoir les vecteurs de support et les hypothèses.*/

$V_m, B^i \leftarrow$ SVM (T_m).

Fin pour

Pour mϵM

$V \leftarrow V \cup V_m$.

Fin pour

Fin tant que

Figure 4.7 L'algorithme SVM utilisé.

Avec :

B^i : Est l'hypothèse construite à la $i^{ème}$ itération

M : Est le nombre de machines utilisées en MapReduce.

T_m : Les données d'apprentissage de la machine m.

V_m : Les vecteurs de support qui sont créés par la $m^{ème}$ itération.

V: Le vecteur de support final.

4.4.2 L'algorithme Random Forest Tree (Les fôrets aléatoires)

L'algorithme du foret aléatoire est un algorithme très utilisé à la prédiction. Il utilise des données d'apprentissage échantillonnées pour chaque arbre. Ces données sont fournies par la méthode bagging

(Bootstrap). Les propriétés de la division sont choisies par Random Forest d'une manière semi-aléatoire. Pour trouver le bon nombre de divisions, le nombre de prédicateurs est choisi d'une manière aléatoire. C'est pour cela dit arbre de décision aléatoire. La seule propriété déterminée est la meilleure division.

4.4.3 L'algorithme Bagging

Bagging ou "bootstrap aggregating", c'est une méthode de combiner des arbres de décision ou d'autres algorithmes de classification. Pour améliorer les algorithmes de classification, le bagging fait appel à l'algorithme principal (dans notre cas, c'est Random Forest) dans une série d'instructions. Il est similaire à l'algorithme boosting. La seule différence entre Bagging et Boosting est la façon d'appeler l'algorithme de classification principale. Car l'algorithme principal fait leur traitement d'apprentissage en utilisant un autre ensemble de données généré par "Bootstrape réplicate " à chaque tour. Les éléments de cet ensemble sont générés d'une manière aléatoire à partir de l'ensemble de données d'apprentissage avec la même dimension de ce dernier. Il peut contient une réplication des éléments ou non. Après J tours, la classification se fait par le vote de chaque classification généré, et après l'algorithme prend la classe majoritaire.

Le pseudo-code de l'algorithme de Bagging :

L'entrée : T= {(d_1,m_1) ;(d_2,m_2);...;(d_n,m_n)} // Les données d'apprentissage

Pour j de 1 à J

Sélectionner aléatoirement n échantillon à partir de T ;

Créer l'ensemble de données pour la réplication de bagging D ;

sj ← Rf(D)

S= majorité (s1, s2,..., sJ) ;

La sortie : S

Figure 4.8 Le pseudo-code de l'algorithme de Bagging.

L'algorithme Random Forest sous Map Reduce

Map j

J est le nombre de sous-ensembles.

L'entrée :

Les données d'apprentissage T ;

Les attributs correspond P (Le choix du sous ensemble se fait aléatoire) ;

p est le sous-ensemble d'attributs.

La sortie : Les arbres de décision.

- Déterminer D (D : Le nombre du cluster)

- L'initialisation du T = $y_j j \epsilon \{1, \dots, \text{nombre des éléments du sous} - \text{ensemble}\}$;

Puis l'appelle de l'algorithme Bagging pour la génération des échantillons aléatoires.

- La création d'arbre pour chaque échantillon.

Tant que $i \leq p$
Pour chaque attribut candidat AC, AC_i faire :

- Calculer Max(ACi), i* = argmax (AC), puis diviser les sur le Max(ACi);

Fin pour

Fin tant que

Retourne "arbre de décision" par AC

$\text{Reduce}_j \; j \epsilon \{1, \dots, \text{nombre des éléments du sous} - \text{ensemble}\}$

L'entrée :

L'ensemble des arbres de décision,

Données de test T*, P vecteurs V_i, $V_i \epsilon$ T*

La sortie : le résultat du vote.

La comparaison du Vi avec les nœuds des arbres de décisions générées.

Prends le résultat majoritaire Ri.

Figure 4.9 L'algorithme utilisé d'Arbres de décision sous Map Reduce

4.5 Les résultats expérimentaux

Nous avons évalué ces algorithmes par deux critères, le temps d'exécution des algorithmes et le taux de classification des algorithmes.

Les résultats du modèle de l'orientation des étudiants à base des algorithmes de classification :

La comparaison des algorithmes par le temps d'exécution :

Cette figure présente le temps d'exécution du réseau de neurones, le temps d'exécution de K plus proches voisins et le temps d'exécution de Naïve Bayes.

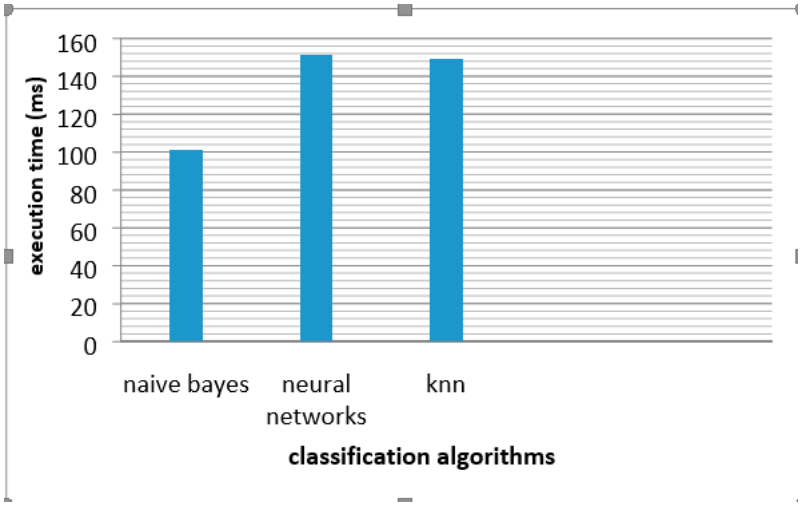

Figure 4.10 Le temps d'exécution du réseau de neurones, le temps d'exécution des K plus proches voisins et le temps d'exécution de Naïve Bayes.

D'après la figure 4.10, le temps d'exécution des réseaux de neurones est très long par rapport aux autres algorithmes, mais il est proche du temps d'exécution de k-plus proche voisins.

La figure suivante présente le taux de classification pour les trois algorithmes : Les réseaux de neurones, K-plus proches voisins et Naïve Bayes.

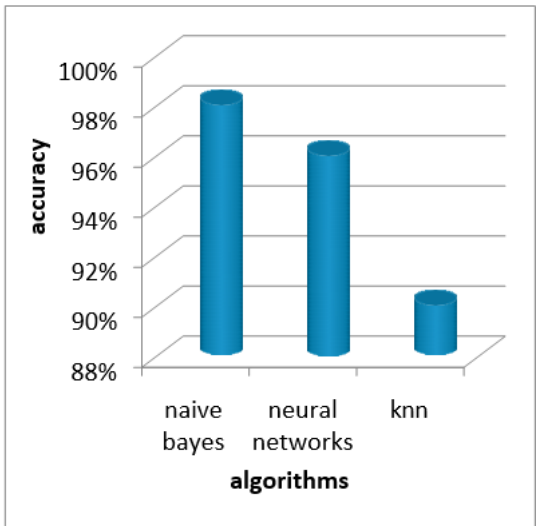

Figure 4.11 Le taux de classification des algorithmes de classification, Réseau de neurones, K plus proches voisins et Naïve Bayes.

A partir de cette figure, le taux de classification de Naïve Bayes est très élevé par rapport aux autres algorithmes.

Les résultats du deuxième modèle basé sur l'outil MOA de WEKA :

La figure suivante présente le temps d'exécution des algorithmes, Réseau de neurones, l'algorithme Naïve Bayes, SVM et Random Forest.

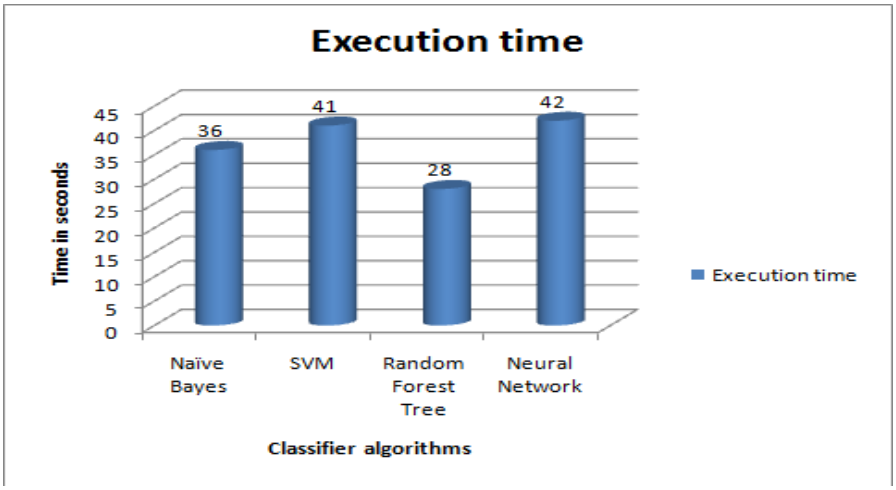

Figure 4.12 Le temps d'exécution de SVM, Random Forest, Naïve Bayes et Neural Network

La figure 4.12 illustre le temps de traitement des données pour les algorithmes de classification. Le classificateur Random Forest est l'algorithme le plus rapide. Alors que l'algorithme de réseau de neurones est le plus lent.

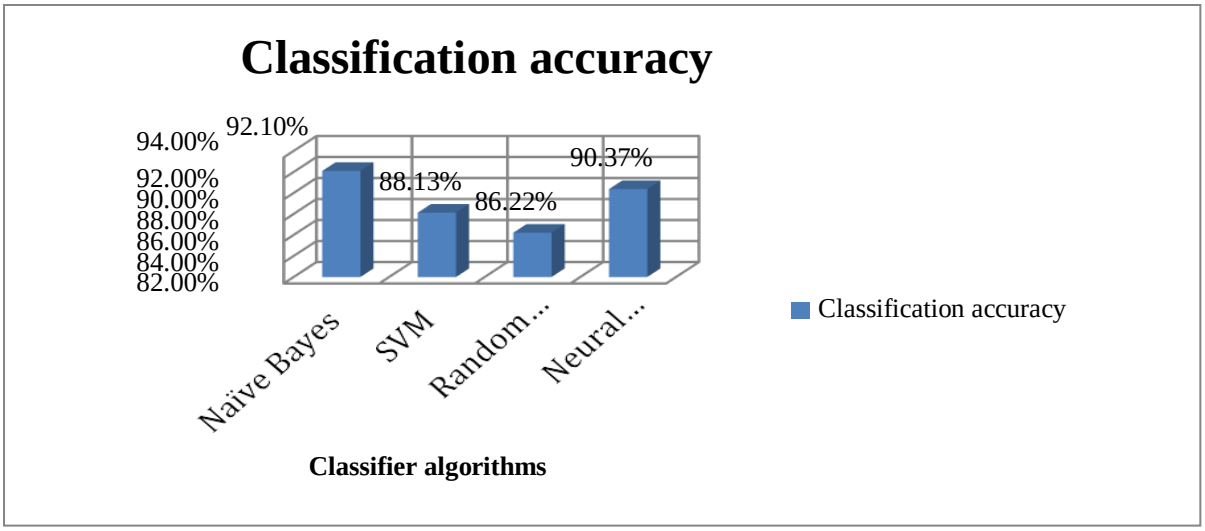

Figure 4.13 Le taux de classification du SVM, Random Forest Tree, Naïve Bayes et Réseau de neurones.

Comme le montre la figure 4.13, le classificateur Naïve Bayes est le plus précis, avec un taux de classification de 92,10 %, puis Réseau de neurones avec 90,37 %, SVM donne une précision de 88,13 %, suivi de Random Forest donne une précision de 86,22 %.

Le classificateur Naïve Bayes est le plus précis parmi tous les classificateurs utilisés. Et aussi, on voit que le temps d'exécution de Naïve Bayes est adéquat pour cet usage.

4.6 Conclusion

Les deux modèles se basent sur la technologie du Big Data, ce qui offre un stockage distribué et aussi un traitement parallèle des données des étudiants. Ces deux éléments minimisent le temps d'exécution et la tolérance contre les pannes. Pour le premier modèle, qui se base sur Hadoop avec HDFS et MapReduce, et après avoir comparé Naïve Bayes, les Réseaux de neurones et les K- plus proches voisins, en utilisant le temps d'exécution et le taux de classification (la précision), il s'est avéré que Bayes est le plus approprié à l'orientation scolaire des étudiants, et cela est nécessaire pour que nous puissions prendre la décision avec une haute qualité.

Dans le deuxième modèle, nous comparons quatre algorithmes de classification, pour trouver le bon algorithme pour l'orientation des étudiants, en utilisant les notes des étudiants et aussi le nombre d'absences pour chaque matière. Ces quatre algorithmes de classification sont Neural Network, Naïve Bayes, SVM et Random Forest. Nous utilisons Weka avec le package MOA, pour tester les résultats. Après le test, nous constatons que Naïve Bayes est le meilleur pour l'orientation des étudiants.

5. Chapitre V : La prédiction de la réussite ou l'échec des étudiants

5.1 Introduction

Data mining est un outil analytique qui contient un ensemble de méthodes et des algorithmes complexes et aussi sophistiqués, qui sont utilisés pour extraire des informations utiles à partir des données analysées. Il est utilisé presque dans tous les domaines, à savoir : La santé, le marketing, l'astronomie, les assurances, l'économie, la finance, les Ressources Humaines, la pharmaceutique, l'industrie et ces dernières années, il est appliqué dans l'éducation. Il a pris le nom d'Educational Data Mining, par ce qu'il est appliqué sur les données éducatives. Il a fait un saut quantique et une révolution dans le domaine de l'éducation. Les méthodes d'EDM permettent de la prédiction, Le clustering, l'extraction de relation, la découverte de modèles et la visualisation des données. Il est utilisé pour des finalités distinctes : Pour évaluer les apprenants, développer les modèles des apprenants et détecter le désengagement des apprenants.

Dans ce chapitre on a concentré sur le succès académique des étudiants.

La réussite de l'étudiant est un indicateur important pour mesurer la qualité de l'éducation et la réussite de l'établissement. Un étudiant réussi, signifie que l'étudiant est accompli son programme et valide tous les semestres. Le succès académique des étudiants est défini comme un groupe de métrique qui mesure l'engagement, l'achèvement des cours et aussi l'apprentissage. Il existe aussi plusieurs définitions de la réussite de l'étudiant, qui sont présentées en [99], elles ont réuni que la réussite de l'étudiant est relié à l'engagement dans les activités scolaires, les compétences, la satisfaction, l'acquisition de connaissances et la persévérance. Mais pour mesurer la réussite académique de l'étudiant, on utilise "Grade Point Average" GPA, ou "Cumulative Grade Point Average" CGPA, qui exprime et mesure les performances académiques des étudiants.

Dans ce chapitre on va présenter une méthode pour la prédiction de la réussite des étudiants universitaires, par ce que l'étudiant fait face à de nombreux changements, tant dans les méthodes

d'enseignement que dans les méthodes d'évaluation. Alors que cet étudiant besoin de l'aide pour réussir.

Ce chapitre est divisé en cinq parties, la première partie, contient les méthodes de sélection des attributs, la deuxième partie concerne la technologie Big Data, la troisième partie présente les algorithmes de classification utilisés, la quatrième partie présente le modèle crée et la cinquième partie contient une description des données utilisées et les résultats obtenus, et enfin la conclusion.

5.2 Les méthodes de sélection des propriétés

Dans le but d'augmenter le taux de classification et aussi d'améliorer la classification des données, on a appliqué les méthodes de sélection des propriétés [100], pour sélectionner les propriétés significatives qui portent des informations utiles à notre prédiction. Ces méthodes réduisent le nombre des propriétés par la suppression des propriétés qui n'ont pas un grand effet sur les résultats de la prédiction. Avec la réduction des propriétés, la taille des données sera réduite aussi, ce qui minimise le temps d'exécution.

Il existe trois types de méthodes de sélection de caractéristique. Ces méthodes diffèrent du point de vue de ses interactions avec l'algorithme de classification. On distingue :

-La méthode de filtrage qui enlève les caractéristiques redondantes et aussi cherche les propriétés non pertinentes en utilisant une métrique statistique, et après elle calcule un score pour chaque caractéristique et les ordonne selon ce score.

La métrique :

- La **corrélation de Pearson, qui est calculé par** la covariance de deux variables divisée par le produit de leurs écarts-types**.**
- **Le test du khi-deux** qui mesure l'écart entre les valeurs attendues et les résultats obtenus

- Les méthodes enveloppes (wrapper methods) font une recherche d'un sous-ensemble dans l'espace des propriétés, ces méthodes sont basées sur les algorithmes de classification pour sélectionner le meilleur sous-ensemble des propriétés.

- Enfin, les méthodes embarquées (Embedded methods), qui combinent les deux types précédents, la sélectionne dess sous-ensemble de propriétés se fait durant la phase d'apprentissage et aussi elles utilisent un algorithme de classification sans la phase de validation.

Dans ce chapitre, on a comparé trois méthodes de sélection de variables, MRMR, et les algorithmes de classification, J48 et SMO.

5.3.1 L'algorithme MRMR

L'algorithme de Minimal Redundancy Maximal Relevance [101] est un algorithme de filtrage basé sur le calcul de la corrélation et l'information mutuelle pour minimiser la redondance entre les caractéristiques et aussi maximiser la pertinence.

$$R(m)=\frac{1}{|P|^2}\sum_{m,n\epsilon P} I(m,n) \quad (20)$$

$$P(i)=\frac{1}{|P|^2}\sum_{m,n\epsilon P} I(m,X) \quad (21)$$

|P| : La taille des caractéristiques.

i : La caractéristique.

I(m, n) : L'information mutuelle entre la m^{ème} caractéristique et la n ^{ème} caractéristique.

I(m, X) : L'information mutuelle entre la m^{ème} caractéristique et les étiquettes de la classe X.

Pour calculer le score d'une caractéristique on utilise la formule suivante :

$$\text{Score (j) =Pertinence (j) - Redondance (j)} \quad (22)$$

5.3.2 La sélection de fonctionnalités basée sur l'algorithme J48

La deuxième méthode est une méthode de sélection de fonctionnalités qui utilise l'algorithme j48 [102] pour valider le sous-ensemble de propriétés.

5.3.3 La sélection de fonctionnalités basée sur l'algorithme SMO

La troisième méthode est une méthode de sélection de fonctionnalités qui utilise l'algorithme SMO [103] qui est une nouvelle version de l'algorithme SVM pour valider le sous-ensemble de propriétés.

Vu à la taille progressive des données des étudiants et la nécessité de traitement en temps réel, les méthodes basiques de stockage et de traitement ne sont pas suffisantes. Ce qui nous a poussés à utiliser la technologie Big Data, pour rendre le stockage et le traitement distribués.

5.4 Les algorithmes de classification utilisés

Pour prédire le succès des étudiants, il existe de nombreuses méthodes de Data mining. Pour prédire le succès des étudiants on a basé sur les algorithmes de classification, K-Nearest Neighbours (KNN), C4.5 et l'algorithme SVM. Ces algorithmes sont exécutés sous Hadoop par Map Reduce.

L'algorithme C4.5 [104] est un algorithme de classification supervisé, parce qu'il utilise des échantillons d'apprentissage. Il permet la génération des arbres de décision pour prendre la décision. L'algorithme C4.5 utilise les données discrètes et aussi les données continues au contraire de l'algorithme ID3 qui utilise les données discrètes seulement. L'algorithme C4.5 fait un élagage des arbres de décision après la construction de ces derniers. Il est souvent utilisé en cas des données incomplètes.

L'algorithme C4.5 sous Map Reduce

Map 1
L'entrée : <p1, v1>
p1 : Le numéro de la ligne
v1 : Les enregistrements
- Extraire l'étiquette et l'attribut de la classe
p2 : L'attribut et l'étiquette de la classe
v2 = 1
La sortie (p2, v2)
Reduce 1
L'entrée : < p2, Liste(v2)>
- Compter la fréquence de l'attribut avec la classe Label
p3 = l'étiquette de la classe avec l'attribut ;
v3 = la fréquence ;
La sortie (p3, v3) ;
Map 2
L'entrée : p3 ,V3
- Calculer entropie (p3) et gain-information (p3) et split-info (p3).
- P4 = les attributs ; V4 = le gain d'information, l'entropie, split-info
La sortie (P4 , V4);
Reduce 2
- Calculer le gain d'information ratio de chaque attribut;
- p5 = Trouver le nœud de décision; v5 = Le gain d'information ratio;
- La sortie (P5 , V5) ;
MAP 3
- Calculer l'identifiant du nœud pour l'attribut le plus élevé
- P6 = id du nœud;
- V6 = Éléments (les valeurs d'attributs)
Jusqu'à ce que toutes ces données soient classées, rappelez ce processus pour créer des branches non feuilles.
Reduce 3
- Créer l'arbre.

Figure 5.1 L'algorithme c4.5 sous MapReduce.

L'algorithme C4.5 commence par la sélection de la racine, puis les branches sont divisées pour chaque cas de l'attribut ainsi de suite jusqu'à ce que les cas de la branche ont la même classe.

Pour trouver l'attribut racine, l'algorithme C4.5 utilise le ratio du gain de tous les attributs, puis il prend l'attribut qui a le ratio de gain le plus élevé. Le ratio de gain est calculé par la formule suivante :

$$\text{Gain-ratio (T,F)} = \text{l'entropie (T)} - \sum_{j=1}^{m} \frac{|T_j|}{T} * \text{l'entropie } (T_j).$$

T : les données d'entraînement.
F : L'attribut.
m : Le nombre d'éléments de F.
| Tj |: le nombre de cas dans la j^{ème} partition.
|T| : Le nombre de cas dans T

$$\text{L'entropie (T)} = \sum_{j=1}^{m} -d_i * log_2 d_i$$

d_i: La proportion de Ti à T

Mais, le stockage et le traitement des données massives qui sont générées par les étudiants (les données qui sont déjà présentées) augmentent le temps d'exécution et rendent le service plus lent, ce qui nous faisons appel à la technologie Big Data, pour rendre ces algorithmes parallèles afin de minimiser le temps d'exécution.

5.5 Les données utilisées

Il existe un certain nombre de facteurs qui influencent la prédiction de la réussite d'un élève. Lorsque nous avons compté les facteurs selon leur occurrence dans la littérature, dans le premier ordre, nous avons trouvé, le rendement scolaire, les activités scolaires de l'étudiant en E-Learning, la démographie d'étudiant, les attributs psychologiques et aussi l'environnement d'étudiant.

Ces facteurs contiennent une combinaison d'attributs, comme le montre cette figure.

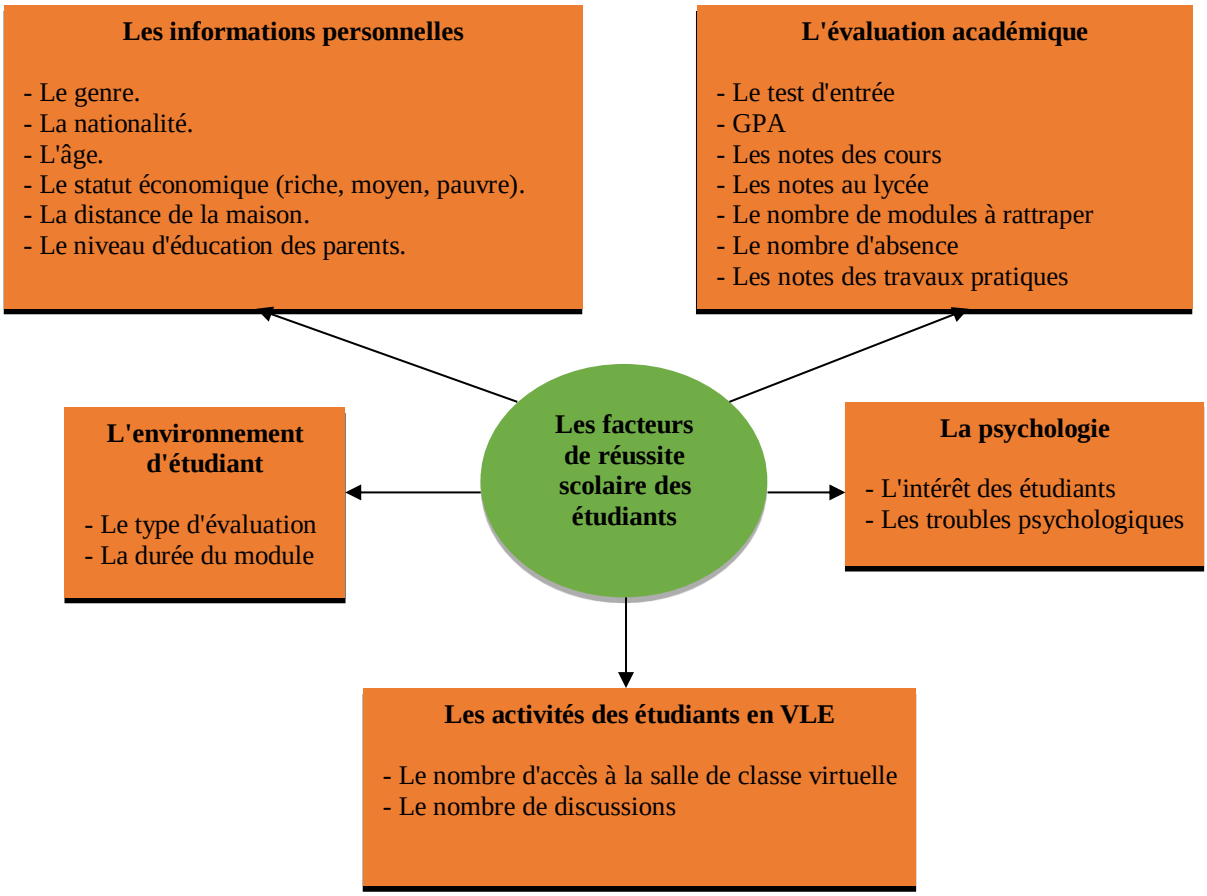

Figure 5.2 Les attributs utilisés pour la prédiction de réussite de l'étudiant.

On a ajouté les données des travaux pratiques, le nombre du jour d'absence de l'étudiant, la distance entre la maison et l'établissement scolaire et aussi le nombre de modules rattrapés comme des attributs qui vont donner une grande valeur à la prédiction de succès des étudiants.

5.6 Le modèle utilisé

Les étapes du modèle utilisé sont les suivantes :

Les sources de données étudiant : LMS, le dossier universitaire de l'étudiant.

L'extraction de données.

Le prétraitement (nettoyage des données, la discrétisation).

La sélection de fonctionnalités.

Le stockage en HDFS.

La classification.

La création de modèles.

L'évaluation.

La figure suivante présente l'architecture du système.

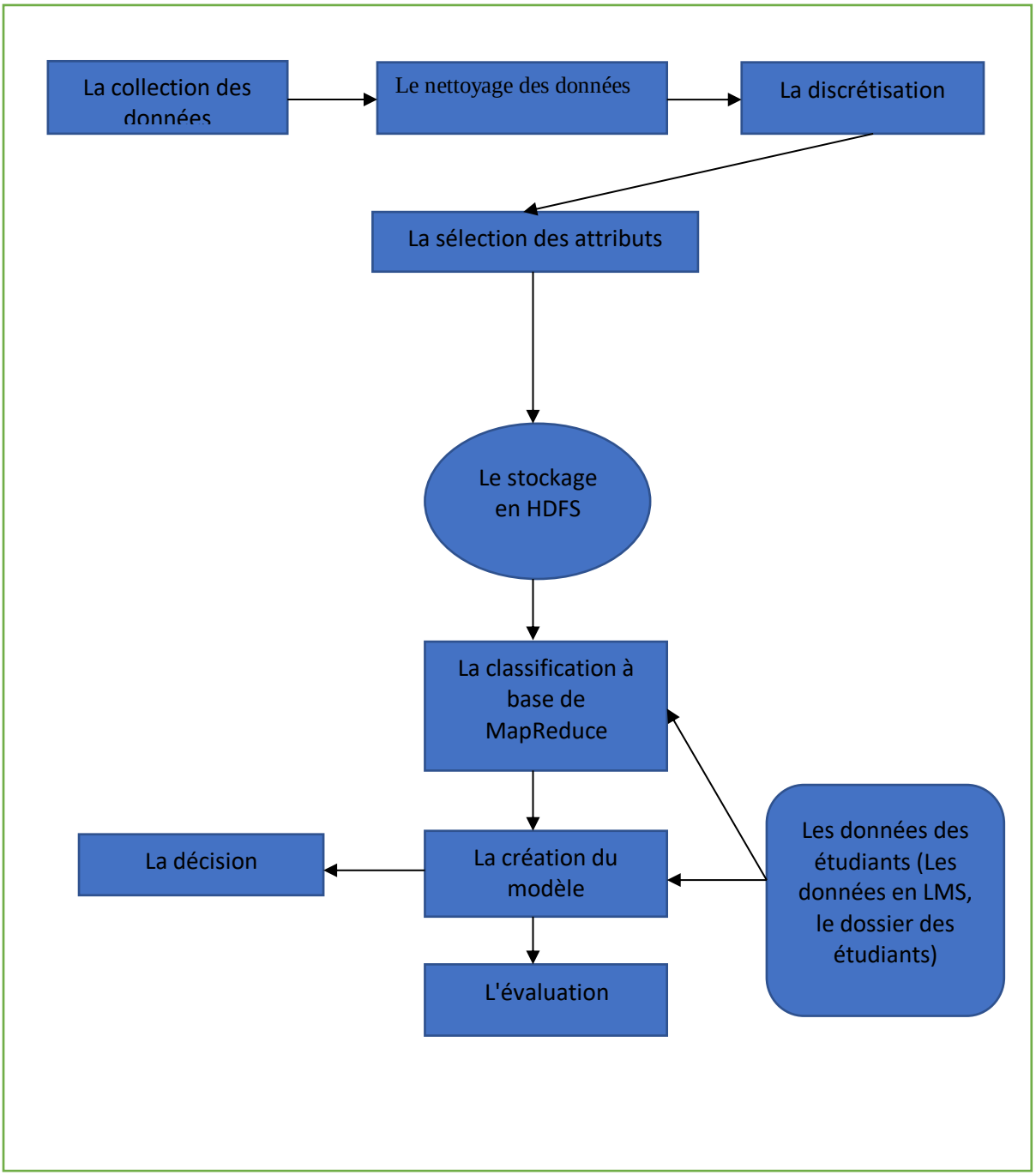

Figure 5.3 L'architecture du système utilisé pour la prédiction de réussite de l'étudiant.

Ce modèle aide les administrateurs et les professeurs à prédire la réussite ou l'échec des étudiants et aussi d'intervenir, pour aider les étudiants qui sont menacés d'échec scolaire par les outils possibles.

Ce system commence par la collection des données des étudiants pour construire une base de données pour la phase d'apprentissage et de test. Ces données sont passées à la phase de nettoyage pour supprimer les données redondantes et les données incomplètes, et après la discrétisation des données continues, afin de les rendre utilisables par les algorithmes de classification. Les trois phases précédentes constituent l'étape de prétraitement. L'étape suivante est la sélection des attributs ou la sélection des variables. Cette étape permet de réduire la taille des données et la suppression des attributs non significatifs ou redondants ou non pertinents. Ce qui augmente la qualité de prédiction et minimise le temps d'exécution. Dans cette étape on a comparé quatre méthodes de sélection des attributs, la méthode de MRMR et deux méthodes de sélection des attributs de la catégorie enveloppe, qui sont basés sur les algorithmes de classification J48 et SVM. Maintenant les données sont prêtes à utiliser, elles sont stockées d'une manière distribuée dans HDFS, afin de l'exploiter pour l'étape de classification. Dans cette étape, on a comparé trois algorithmes de classification binaire, car on a deux classes, la classe « à succès » et la classe « échec ». Après la classification on a pris l'algorithme le plus performant pour l'utiliser dans le modèle. Enfin on a construit un système de recommandation qui permet la prédiction du succès des étudiants.

5.7 Les résultats expérimentaux

Pour choisir le modèle adéquat à la prédiction du succès des étudiants, on a testé plusieurs modèles qui se diffèrent selon les attributs utilisés, les méthodes de sélection des attributs et les algorithmes de classification utilisés.

On a passé par la matrice de confusion, pour calculer les mesures de performance, qui nous a permis d'évaluer les modèles de prédiction du succès ou l'échec des étudiants.

Les mesures utilisées sont : Le rappel, la précision, F-Mesure, le taux de classification et la spécificité.

Le rappel "Recall", il est aussi appelé "sensitivity", il représente la capacité de détecter le succès des étudiants par ce modèle.

$$\text{Le rappel} = \frac{TP}{TP+FN} \quad (22)$$

La précision, elle représente la capacité de ce modèle de ne détecter que les étudiants réellement réussis.

$$\text{La précision} = \frac{TP}{TP+FP} \quad (23)$$

F-Mesure, c'est un critère d'évaluation qui est lié à la fois au rappel et à la précision, c'est la moyenne harmonique.

$$\text{F−mesure}=2\text{x}\frac{\text{Précision X}R\text{appel}}{\text{Précision}+R\text{appel}}=\frac{2TP}{2TP+\text{FP}+FN} \quad (24)$$

Le taux de classification, c'est la proportion de réussite des étudiants correctement prédits.

$$\text{Le taux de classification} = \frac{TP+TN}{TP+\text{TN}+\text{FP}+FN} \quad (25)$$

La spécificité, elle représente la capacité de ce modèle de détecter tous les étudiants échoués.

$$\text{La spécificité} = \frac{TN}{\text{FP}+\text{TN}} \quad (26)$$

Avec :

TP : Vrai positif "True positives".

TN : Vrai négatif "True négatives".

FP : Faux positif "False positives".

FN : Faux négatif "False négatives".

Positive signifie la classe réussite et négative signifie la classe échec, vrai représente une classification correcte et faux représente une classification non correcte.

La qualité de la prédiction augmente lorsque les mesures précédentes augmentent ou proche à 1.

On a appliqué ces mesures sur les modèles qui utilisent la méthode MRMR pour la sélection des attributs.

Le tableau suivant présente les résultats obtenus.

Les algorithmes	Le rappel	La précision	F-Mesure	Le taux de classification	La spécificité
SVM	82.73	83.65	83.19	83.00	83.26
KNN	80.08	82.82	81.42	80.99	81.99
C4.5	78.35	83.24	80.72	80.00	81.89

Table 5-1 Les résultats obtenus par la méthode MRMR.

Ce tableau présente : le rappel, la précision, F-Mesure, le taux de classification et la spécificité de chaque modèle de prédiction du succès des étudiants avec les algorithmes de classification utilisés.

Selon les résultats du tableau on remarque que toutes les mesures sont entre 80% et 84% sauves le rappel de l'algorithme C4.5 qui est inférieur à 79%. Le taux de reconnaissance du SVM et aussi les autres mesures du même algorithme sont les plus élevés, suivis du KNN et enfin C4.5. Ce qui rend l'algorithme SVM le plus performant en utilisant la méthode MRMR pour la sélection des attributs.

Pour les résultats des mesures de performance des modèles qui utilisent l'algorithme j48 pour la sélection des attributs on a obtenu les résultats qui sont présentés dans le tableau suivant.

Les algorithmes	Le rappel	La précision	F-Mesure	Le taux de classification	La spécificité
SVM	81.33	88.51	84.76	84.00	87.23
KNN	82.13	84.60	83.34	83.00	83.92
C4.5	83.08	90.63	86.69	86.00	89.55

Table 5-2 Les résultats obtenus par la méthode basée sur l'algorithme J48.

Selon le tableau précédent, le modèle qui utilise l'algorithme C4.5 a un taux de classification très élevé (86%) par rapport aux autres modèles, ceci est dû à l'utilisation de l'algorithme j48 qui est une implémentation de l'algorithme C4.5 à la phase de sélection des attributs, avec une précision qui dépasse 90%. Et au deuxième rang on trouve le modèle basé sur l'algorithme SVM à la phase de classification avec un taux de classification de 84%, et en troisième rang on a le modèle qui utilise l'algorithme KNN à la phase de classification avec un taux de classification de 83%.

Après les résultats de la mesure de performances des modèles, qui sont basés sur l'algorithme j48. Le tableau suivant présente les résultats des critères de performances qui correspondent aux modèles qui sont basés sur l'algorithme SMO à la phase classification.

Les algorithmes	Le rappel	La précision	F-Mesure	Le taux de classification	La spécificité
SVM	87.09	87.82	87.45	87.32	87.57
KNN	83.00	88.07	85.45	84.92	87.12
C4.5	85.00	87.87	86.41	86.10	87.29

Table 5-3 Les résultats obtenus par la méthode basée sur l'algorithme SMO.

Lors de l'utilisation de l'algorithme SMO à la phase de sélection des attributs on a remarqué que le modèle basé sur l'algorithme SVM a le taux de classification le plus élevé (87.32%) suivis par le modèle basé sur l'algorithme C4.5 puis le modèle basé sur l'algorithme KNN. L'augmentation du taux de classification du modèle basé sur l'algorithme SVM est due à l'utilisation de l'algorithme SMO à la

phase de sélection des attributs, car l'algorithme SMO est une implémentation plus simple de l'algorithme SVM.

D'après les résultats des tableaux précédents on a remarqué que le modèle basé sur l'algorithme SVM et l'algorithme SMO est le plus performant car, le taux de classification de ce modèle et aussi sa capacité de détecter le succès ou l'échec des étudiants sont les plus élevés. Selon le taux de classification on remarque aussi que les modèles qui sont basés sur l'algorithme SMO ont le taux de classification maximale par rapport aux autres algorithmes (KNN et C4.5).

Ces mesures de performance seulement ne sont pas suffisantes. Car le choix de modèle dépend aussi du temps d'exécution. Le graphique suivant présent le temps d'exécution de chaque modèle. On a pris le temps d'exécution des modèles basés sur l'algorithme SMO. Car, il a donné une précision maximale pour tous les algorithmes de classification.

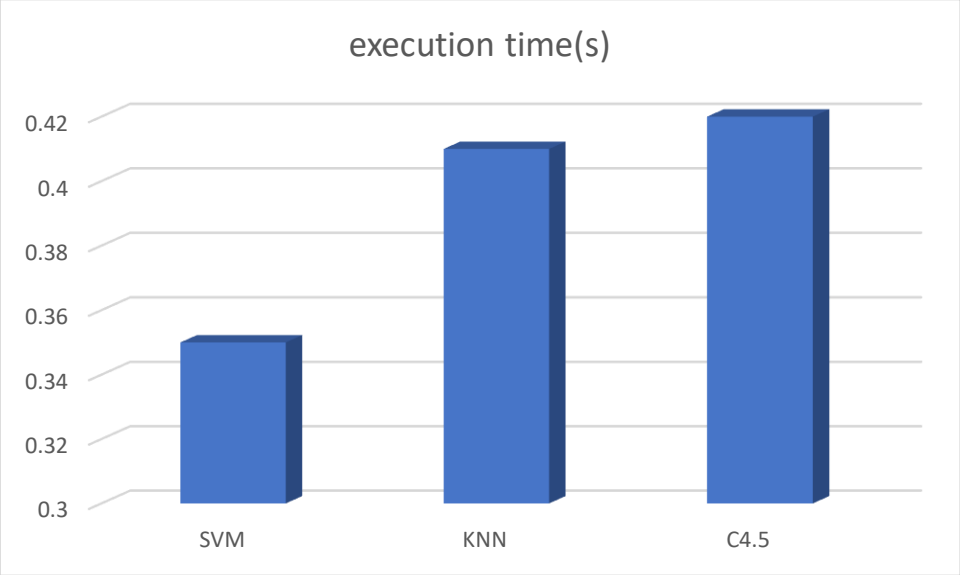

Figure 5.4 Le temps d'exécution des algorithmes de classification SVM, KNN et C4.5.

Le graphique précédent présente le temps d'exécution des algorithmes de classification SVM, KNN et C4.5.

D'après le graphique, on remarque que le temps d'exécution de l'algorithme SVM est le minimal suivi du KNN et enfin l'algorithme C4.5.

5.8 Conclusion

La prédiction du succès des étudiants est une méthode très utile car, elle permet de diminuer le taux de l'échec scolaire par l'intervention des professeurs et tous les acteurs dans l'enseignement au cas où un étudiant est (découvris) déclenché par le système comme un étudiant menacé par l'échec scolaire. Les résultats de la comparaison des modèles par les mesures de performances et par le temps d'exécution prouvent que le modèle basé sur l'algorithme SMO à la phase de sélection d'attributs et l'algorithme SVM à la phase de classification est le meilleur modèle parmi les modèles construits, il a le taux de classification le plus élevé (87,32%) et le temps d'exécution le plus bas.

Après la sélection des attributs, on a trouvé que les facteurs qui influencent plus la prédiction du succès des étudiants sont premièrement tous les attributs de l'évaluation académique suivis du statut économique, le niveau d'éducation des parents, la distance de la maison, l'intérêt des élèves, les troubles psychologiques et le nombre d'accès à la salle de classe virtuelle. Les autres attributs n'ont pas une valeur à ajouter pour la prédiction du succès des étudiants.

Conclusion générale et perspectives

Dans cette thèse, on a comparé les plateformes de l'enseignement à distance à l'aide de la méthode de prise de décision multicritères (MCDM). Le but de cette comparaison est de trouver la plateforme adéquate au suivi des activités des étudiants et l'évolution de ses performances et ses capacités. Après la sélection de la plateforme de E-learning, on a utilisé pour l'orientation scolaire et la prédiction de réussite des étudiants. On a trouvé que Moodle est la plateforme la plus adéquate à cet objectif. Puis, on a comparé deux modèles de l'orientation. Le premier modèle est basé sur une hybridation de la méthode TOPSIS et le gain d'information, par contre, le deuxième modèle est basé sur une hybridation de la méthode AHP et le gain d'information. Dans les deux modèles on a utilisé la méthode SMOTE pour l'équilibrage des données. Les résultats montrent que le système d'orientation et de réorientation basé sur la méthode TOPSIS, est plus précis que le système basé sur la méthode AHP, par contre, le modèle basé sur la méthode AHP est plus rapide que le modèle basé sur la méthode TOPSIS.

Pour prendre en considération les facteurs qui influencent l'orientation scolaire des étudiants, les modèles précédents sont limités à un certain nombre de facteurs, et aussi au nombre des étudiants qui est en progression. Ce que nous a poussé à utiliser la technologie Big Data, pour rendre le stockage des données distribués à l'aide du HDFS et aussi le traitement parallèle à l'aide de MapReduce. Pour le traitement des données et la prise de décision, on a comparé les algorithmes de classification. Les résultats montrent que l'algorithme Naïve Bayes est le plus précis.

L'objectif principal des méthodes de l'enseignement et de l'apprentissage est la réussite des étudiants. C'est pourquoi on a utilisé la technologie Big Data pour la prédiction de la réussite ou l'échec des étudiants. On a utilisé les algorithmes de la sélection des attributs pour trouver les facteurs les plus influents à la réussite ou l'échec de l'étudiant. Puis on a appliqué les algorithmes de classification pour la prise de décision. Les résultats montrent que le modèle basé sur l'algorithme SMO à la phase de sélection d'attributs et l'algorithme SVM à la phase de classification est le meilleur modèle.

Nous allons désormais élargir ce système pour prendre en considération M-Learning. Nous allons aussi utiliser les agents mobiles pour bénéficier de ses avantages (l'efficacité, la fiabilité, la portabilité et la flexibilité).

Les références

1. Barth, M., Adomßent, M., Fischer, D., Richter, S., & Rieckmann, M. "Learning to change universities from within: A service-learning perspective on promoting sustainable consumption in higher education". Journal of Cleaner Production, 62,72–81. (2014).
2. Devedzic, V. "Introduction to Web-Based Education". Semantic web and education (Vol. 11).1-28. (2006).
3. Dutta, B., "Semantic web based e-learning". DRTC Conference on ICT for Digital Learning Environment 11th – 13th January, Bangalore (2006).
4. A. Khan, R., Qudrat-Ullah, H. "Technology Adoption". Adoption of LMS in Higher Educational Institutions of the Middle East, 7–12. (2021).
5. Valentine, D., "Distance learning: Promises, problems, and possibilities" . Online Journal of Distance Learning Administration, 5(3). (2002).
6. Nagy, J. T., "Using learning management systems in business and economics studies in Hungarian higher education" . Education and Information Technologies, 21(4), 897–917. (2017).
7. Matheos, K., Daniel, B. K., & McCalla, G. L., "Dimensions for blended learning technology: Learners' perspectives" . Journal of Learning Design, 1(1), 56–76. (2005).
8. Doneva, R., Nikolaj, K., & Totkov, G., "Towards mobile university campuses" . International Conference on Computer Systems and Technologies – CompSysTech. (2006).
9. Georgiev, T., Georgieva, E., & Smrikarov, A., " M-learning-a new stage of e-learning" . International Conference on Computer Systems and Technologies-CompSysTech, 28. (2004).
10. Guri-Rosenblit, S. "Distance education'and e-learning : Not the same thing" . Higher education, 49(4), 467-493. (2005).
11. Sujit, K. B., Marguerite, W. and Paul,B.́l. "E-learning, M-learning and D-learning: Conceptual definition and comparative analysis" . E-Learning and Digital Media, Vol. 15(4) 191–216. (2018).
12. Vuopala, E., Hyvo, P., J. , S., "Interaction forms in successful collaborative learning in virtual learning environments". Active Learning in Higher Education. 17. 1-14. (2015).
13. Kim M, "Processes of emotional experiences in online discussions: Emotional changes throughinteracting with other students". The Korean Journal of Educational Psychology 22(4): 697–722. (2008).
14. Jan, P. T., Lu, H. P., Chou, T. C. "The adoption of e-learning: An institutional theory perspective". Turkish Online Journal of Educational Technology-TOJET, 11(3), 326-343. (2012).
15. Trinder J, "Mobile technologies and systems". Mobile Learning: A Handbook for Educators and Trainers.USA: Taylor & Francis, pp. 7–24. (2005).
16. Walker K, "Introduction: Mapping the landscape of mobile learning". A Report of a New Workshop by the Kaleidoscope Network of ExcellenceMobile Learning Initiative. UK: University of Nottingham, pp. 5–6. (2007)
17. Dörnyei, Z., "Attitudes, orientations, and motivations in language learning: Advances in theory, research, and applications". Language Learning, 53(1), 332. (2003).
18. Ahmadi, N. , Motallebzadeh, K., Fatemi, M. "The Effect of Cooperative Learning Strategies on Iranian Intermediate Students' Writing Achievement". Open Access Library Journal, Vol.1 No.9. (2014).
19. Yi, Z., "The Instructor's Roles in Distance Education for Library and Information Science". Chinese Librarianship: an International Electronic Journal, vol(34). (2012).
20. Ularu, E. G., Puican, F. C., Apostu, A., et al. "Perspectives on big data and big data analytics". Database Systems Journal, vol. 3, no 4, p. 3-14. (2012).
21. Zakir, J., Seymour, T., BERG, K. "Big Data Analytics". Issues in Information Systems, vol. 16, no 2. (2015).
22. Uma, N. , et al , " A light weight encryption over big data in information stockpiling on cloud". Indonesian Journal of Electrical Engineering and Computer Science Vol. 17, No. 1, January, pp. 389-397. (2020).
23. Bedi, P., Jindal, V., & Gautam, A., "Beginning with big data simp lified". International Conference on Data Mining and Intelligence Computing (ICDMIC)1-7. (2014).
24. Demchenko, Y.,Gruengard, E., Klous, S., "Instructional Model for Building Effective Big Data Curricula for Online and Camp us Education". IEEE 6th International Conference on Cloud Computing Technology and Science, p p .935–941. (2014).
25. Demchenko, Y., Grosso, P., De Laat, C., & M embrey , P., "Addressing big data issues in Scientific Data Infrastructure". International Conference on Collaboration Technologies and Systems (CTS)(p p . 48–55. (2013).

26. Ward, J. S., Barker, A. “Undefined by data: a survey of big data definitions”. arXiv preprint arXiv:1309.5821. (2013).
27. De Mauro, A., Greco, M., & Grimaldi, M. “What is big data? A consensual definition and a review of key research topics”. AIP conference proceedings (Vol. 1644, No. 1, pp. 97-104). (2015).
28. Oguntimilehin A., Ademola E.O.‘‘A Review of Big DataManagement, Benefits and Challenges’’. Journal of Emerging Trendsin Computing and Information Sciences, vol-5, pp-433-437, June. (2014).
29. Lee, J. G., & Kang, M. “Geospatial big data: challenges and opportunities”. Big Data Research, 2(2), 74-81. (2015).
30. Zicari, R. V. “Big data: Challenges and opportunities’’ . Big data computing, 564, 103. (2014).
31. Stanly Wilson , et al ,"Twitter data analysis using hadoop ecosystems and apache zeppelin". Indonesian Journal of Electrical Engineering and Computer Science Vol. 16, No. 3, December, pp. 1490~1498. (2019).
32. Manjula, K., Meenakshi S., S. “Optimized Approach (SPCA) for Load Balancing in Distributed HDFS Cluster”. sn comput. sci. 1- 102. (2020).
33. Agarwal, S., Panda, A., Mozafari, B., Madden, S., Stoica, I., Panda, A., Milner, H., Madden,S., Stoica, I., Mozafari, B., Madden, S., Stoica, I., Berkeley, U.C. “BlinkDB: queries with bounded errors and bounded response times on very large data”. Proceedings of ACM EuroSys, Prague. (2013).
34. Győrödi, C., Győrödi, R., Pecherle, G., & Olah, A. “A comparative study: MongoDB vs. MySQL”. 13th International Conference on Engineering of Modern Electric Systems (EMES) (pp. 1-6). IEEE. (2015).
35. Herrera, V. M., Khoshgoftaar, T. M., Villanustre, F., & Furht, B. “Random forest implementation and optimization for Big Data analytics on LexisNexis’s high performance computing cluster platform”. Journal of Big Data, 6(1), 1-36. (2019).
36. Sagiroglu, S., Sinanc, D.: “Big data: a review”. International Conference on Digital Object Identifier, 42-47. (2013).
37. Malviya, A., Udhani, A., & Soni, S. “R-tool: Data analytic framework for big data”. Symposium on Colossal Data Analysis and Networking (CDAN) (pp. 1-5). IEEE. (2016).
38. Fernandes, D., & Bernardino, J. “Graph Databases Comparison: AllegroGraph, ArangoDB, InfiniteGraph, Neo4J, and OrientDB”. In Data (pp. 373-380). (2018).
39. Pulla, V. S. V., Varol, C., & Al, M. “Open source data quality tools: Revisited”. Information Technology: New Generations (pp. 893-902). Springer, Cham. (2016).
40. Târnăveanu, D. “Pentaho business analytics: a business intelligence open source alternative”. Database Systems Journal, 3(3), 23-34. (2012).
41. Chawla, G., Bamal, S., & Khatana, R. “Big data analytics for data visualization: Review of techniques”. International Journal of Computer Applications, 182(21), 37-40. (2018).
42. Aziz, K., Zaidouni, D. & Bellafkih, M. “Leveraging resource management for efficient performance of Apache Spark”. J Big Data 6-78. (2019).
43. Shaw S., Vermeulen A.F., Gupta A., Kjerrumgaard D. “Introducing Hive”. Practical Hive. 23–35. (2016).
44. Singh, R., Kaur, P.J. “Analyzing performance of Apache Tez and MapReduce with hadoop multinode cluster on Amazon cloud”. J Big Data 3, 19. (2016).
45. Vohra D. “Using Apache HBase”. Pro Docker. 141–150. (2016).
46. Wadkar S., Siddalingaiah M., “Apache Ambari”. Pro Apache Hadoop. 399–401. (2014).
47. Vohra D., “Apache Kafka”. Practical Hadoop Ecosystem. 339–347. (2016).
48. Ismail, A., Truong, HL. & Kastner, W. “Manufacturing process data analysis pipelines: a requirements analysis and survey”. J Big Data 6, 1. (2019).
49. Demirbaga, U. “HTwitt: a hadoop-based platform for analysis and visualization of streaming Twitter data”. Neural Comput & Applic. (2021).
50. Fikri, N., Rida, M., Abghour, N. et al. “An adaptive and real-time based architecture for financial data integration”. J Big Data 6, 97. (2019).
51. Dahdouh, K., Dakkak, A., Oughdir, L. et al. “Large-scale e-learning recommender system based on Spark and Hadoop”. J Big Data 6, 2. (2019).
52. Ahmed, N., Barczak, A.L.C., Susnjak, T. et al. “A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using HiBench”. J Big Data 7, 110. (2020).
53. Noor M.Alqudah, Hisham M Jammal, Omar Saleh,Yousef Khader,Nail ObeidatandJumana Alqudah. “Perception and experience of academic Jordanian ophthalmologists with E-Learning for undergraduate course during the COVID-19 pandemic”. Annals of Medicine and Surgery,59, 44-47. (2020).

54. Abdulaziz A., Harun C., Alex K., Firoz A., Hamed A., "Utilization of Learning Management Systems (LMSs) in higher education system: A case review for Saudi Arabia",Energy Procedia 160, 731–737. (2019).
55. Mark N., "A comparison of two online learning systems", Journal of Open, Flexible and Distance Learning, 20(1), 19–32. (2016).
56. Tania A. and Sergio L. M., "Comparison from the levels of accessibility on LMS platforms that supports the online learning system". 8th International Conference on Education and New Learning Technologies, (2016).
57. Harry D., Fitriana D., Michael S., Surajiyo and Musa J., "Comparison of Learning Management System Moodle, Edmodo and Jejak Bali". Advances in Social Science, Education and Humanities Research, 422, (2020).
58. Petra P., Blanka K., Martin K. "Selected E-Learning Systems and Their Comparison". International Symposium on Educational Technology (ISET). (2019).
59. Murshitha, S. M., Wickramarachchi, A. P. R., "A study of students' perspectives on the adoption of LMS at the University of Kelaniya". Journal of Management, 9(1), 16. (2016).
60. Nagy, J. T., "Using learning management systems in business and economics studies in Hungarian higher education". Education and Information Technologies, 21(4), 897–917. (2016).
61. Szabo, M., & Flesher, K., "CMI theory and practice: Historical roots of learning management systems". Educational Technology, 32, 58–59. (2002).
62. Shiau, W. L., & Chau, P. Y. K., "Understanding behavioral intention to use a cloud computing classroom: A multiple model comparison approach". Information & Management, 53(3), 355–365. (2015).
63. Chu, L. F., Erlendson, M. J., Sun, J. S., Clemenson, A. M., Martin, P., & Eng, R. L., "Information technology and its role in anaesthesia training and continuing medical education". Best Practice & Research. Clinical Anaesthesiology, 26(1), 33–53. (2012).
64. Gilhooly, K., "Making e-learning effective: Industry trend or event". Computerworld, 35(29), 52–53. (2001).
65. Tortora, G., Sebillo, M., Vitiello, G., & D'Ambrosio, P., "A multilevel Learning management system". 14th International Conference on Software Engineering and Knowledge Engineering. 541–547. (2002).
66. Sharma, S. A. T., Paul, A., Gillies, D., Conway, C., Nesbitt, S., Ripstein, I. R. A., Mcconnell, K., "Learning/Curriculum Management Systems (LCMS): Emergence of a new wave in medical education". Learning, 11(13). (2011).
67. Rapuano, S., & Zoino, F., "A learning management system including laboratory experiments on measurement instrumentation". Instrumentation and Measurement, IEEE Transactions On, 55(5), 1757–1766. (2006).
68. Bogarín, A., Cerezo, R., & Romero, C., "Discovering learning processes using Inductive Miner: A case study with Learning Management Systems (LMSs)". Psicothema, 30(3), 322–329. (2018).
69. Nichols, M., "A theory for eLearning". Educational Technology & Society, 6(2), 1–10. (2003).
70. Jordan, M. M., & Duckett, N. D., "Universities confront 'Tech disruption': Perceptions of student engagement online using two learning management systems". The Journal of Public and Professional Sociology. 10(1). (2018).
71. Alharbi, S., & Drew, S., "Using the technology acceptance model in understanding academics' behavioural intention to use learning management systems". International Journal of Advanced Computer Science and Applications, 5, 143–155. (2014).
72. Avgeriou, P., Papasalouros, A., Retalis, S., & Skordalakis, M., "Towards a pattern language for learning management systems". Educational Technology & Society, 6(2), 11–24. (2003).
73. Jafari, A., McGee, P., & Carmean, C., "Managing courses defining learning: What faculty, students, and administrators want". Educause Review, 4(4), 50–51. (2006).
74. Antonenko, P. D., Derakhshan, N., & Mendez, J. P., "Pedagogy 2 go: Student and faculty perspectives on the features of mobile learning management systems". International Journal of Mobile Learning and Organisation, 7(3–4), 197–209. (2013).
75. S. Wichadee, "Factors Related to Faculty Members' Attitude and Adoption of a Learning Management System,". Turkish Online Journal of Educational Technology. 14(4), 53–61. (2015).
76. I. Lurie., A Web Content Management Blueprint". Planning for a Content-Rich, Successful Web Site.14, 2007 (2002).
77. A. Nawaz, G. M. Kundi, "Sustained Technical Support: Issue and Prospects for E-Learning in HEIs," Malaysian Journal of Distance Education., 12(2), 61–77. (2010).
78. S. Mohorovicic et al., "Using Web Content Management Systems in University E-Commerce Courses". International Journal of Emerging Technologies in Learning.,5(2), 38–43. (2010).
79. Shen, Y. H. "Design of Digital Network Shared Learning Platform based on SCORM Standard". International Journal of Emerging Technologies in Learning, 13(7). (2018).

80. S. Wichadee, “Factors Related to Faculty Members’ Attitude and Adoption of a Learning Management System,”. Turkish Online Journal of Educational Technology., 14(4), 53–61. (2015).
81. Al-Sharhan, S., Al-Hunaiyyan, A., Alhajri, R., & Al-Huwail, N. “Utilization of learning management system (LMS) among instructors and students”. Advances in Electronics Engineering 15-23. Springer, Singapore. (2020).
82. N. Srichanyachon, “Efl Learners’ Perceptions of Using LMS”. Turkish Online Journal of Educational Technology. 13(4), 30–35. (2014).
83. Siksnelyte B., Indre, Z., Edmundas K., et Streimikiene, D., "Multi-criteria decision-making (MCDM) for the assessment of renewable energy technologies in a household: A review". Energies 13(5), 1164. (2020).
84. Shuja, M., Mittal, S., Zaman, M. "Effective prediction of type ii diabetes mellitus using data mining classifiers and SMOTE". Advances in computing and intelligent systems. Springer, Singapore, 195-211. (2020).
85. Meshram, S., Gajbhiye, A., Ehsan, MESHRAM, C., et al. "Application of SAW and TOPSIS in prioritizing watersheds". Water Resources Management, 34(2), 715-732. (2020).
86. Liu, Y. E., Claudia M., EARL, C. "A review of fuzzy AHP methods for decision-making with subjective judgements". Expert Systems with Applications, 113738. (2020).
87. Atef M. G., Husam K. ,Ali A. ,Syed H. M., Lotfi H. "Assessment and Comparison of Various MCDM Approaches in theSelection of Manufacturing Process”. Advances in Materials Science and Engineering, 16. (2020).
88. M. Phanupong, et al., “Prediction of Student droput using personel profile and data mining aproch”. Intelligent and Evolutionary Systems, 143-155. (2016).
89. B.Sunita et al., "Selecting the Best Supervised Learning Algorithmfor Recommending the Course in E-Learning System". International Journal of Computer, 41, 42-49. (2012).
90. Seyed Reza Pakize, et al., "Comparative Study of Classification Algorithms Based On MapReduce Model". International Journal of Innovative Research in Advanced Engineering, 1(7). (2014).
91. Etaiwi, Maria et al., " Evaluation of classification algorithms for banking customer’s behavior under Apache Spark Data Processing System". 4th International Symposium on Emerging Information, Communication and Networks Procedia Computer Science, 113, 559–564. (2017).
92. Amine Rghioui et al., " Big Data Classification and Internet of Things in Healthcare ". International Journal of E-Health and Medical Communications, 11, 20-37. (2020).
93. F.Ouatik , et al., " Comparative study of MapReduce classification algorithms for students orientation".," Procedia Computer Science, 170, 1192-1197. (2020).
94. F. Ouatik, et al., “The EOLES project remote labs across the Mediterranean: an example of a successful experience”. International Conference on Smart Digital Environment, (2017).
95. R. Conjin, et al., “Predicting Student Performance from LMS Data: A Comparison of 17 Blended Courses Using Moodle LMS,” IEEE Transactions on Learning Technology, 10(1),17-29. (2017).
96. Sfenrianto S., et al., “E-Learning Effectiveness Analysis in Developing Countries: East Nusa Tenggara, Indonesia Perspective”. Bulletin of Electrical Engineering and Informatics, 7(3). (2018).
97. Nivedita N., Dharaskar, R. “An effective approach to network intrusion detection system using genetic algorithm”. International Journal of Computer Applications, 1(2). (2010).
98. Bin Othman M.F., Yau T.M.S., “Comparison of Different Classification Techniques Using WEKA for Breast Cancer”. International Conference on Biomedical Engineering, 15, 520-523. (2007).
99. Melguizo, T., Martorell, P., Swanson, E., Chi, W. E., Park, E., & Kezar, A. “Expanding student success: The impact of a comprehensive college transition program on psychosocial outcomes”. Journal of Research on Educational Effectiveness, 14(4), 835-860. (2021).
100. Meekins, R., Adams, S., Farinholt, K. et al. “ROC with Cost Pareto Frontier Feature Selection Using Search Methods”. Data-Enabled Discov. Appl. 4(6). (2020).
101. Radovic, M., Ghalwash, M., Filipovic, N. et al. “Minimum redundancy maximum relevance feature selection approach for temporal gene expression data”. BMC Bioinformatics, 18(9). (2017).
102. Singh, J., Singh, G. & Singh, R. “Optimization of sentiment analysis using machine learning classifiers”. Hum. Cent. Comput. Inf. 7(32). (2017).
103. Ghosh, M., Sanyal, G. “An ensemble approach to stabilize the features for multi-domain sentiment analysis using supervised machine learning”. J Big Data 5(44). (2018).
104. Padillo, F., Luna, J. & Ventura, S. “Evaluating associative classification algorithms for Big Data”. Big Data Anal 4(2). (2019).