Université Hassan 1er

Centre d'Études Doctorales

Faculté des Sciences et Techniques

Settat

THÈSE DE DOCTORAT

Pour l'obtention du grade de Docteur en Informatique

Formation Doctorale: Mathématiques, Informatique et Applications

Spécialité: Informatique

Sous le thème

Machine learning pour la prédiction de l'employabilité au Maroc dans un environnement Big Data

Présentée par :

SAOUABI Mohamed

Soutenue le: 05/07/2021

A la Faculté des Sciences et Techniques de Settat devant le jury composé de :

Pr. LOUZAR Mohamed	**P.E.S**	**FST Settat**	Président
Pr. BENI-HSSANE Abderrahim	**P.E.S**	**FS El Jadida**	Rapporteur
Pr. HASNAOUI Lahcen	**P.H**	**EST Meknès**	Rapporteur
Pr. MARZOUK Abderrahim	**P.E.S**	**FST Settat**	Examinateur
Pr. MOUHSEN Ahmed	**P.E.S**	**FST Settat**	Examinateur
Pr. EZZATI Abdellah	**P.E.S**	**FST Settat**	Directeur de thèse

Année Universitaire: 2020/2021

Remerciements

Je tiens à remercier Dieu le tout puissant et miséricordieux, qui m'a donné la force et la patience d'accomplir ce modeste travail.

Je tiens à exprimer mes plus vifs remerciements, reconnaissances et ma gratitude à mon directeur de thèse Professeur Abdellah EZZATI pour son apport scientifique considérable ainsi que pour m'avoir donné de son temps et de son expérience sans parcimonie. Il me doit également la reconnaissance pour ses conseils et son soutien moral continu. Je le remercie infiniment pour avoir cru en moi, pour avoir su m'orienter et pour m'avoir soutenu généreusement à tous les stades de ce travail de thèse.

Je tiens à remercier très chaleureusement tous les membres du Jury, qui m'ont fait l'honneur d'accepter d'évaluer ce travail.

Je remercie le Professeur El Mostafa RAJAALLAH qui m'a beaucoup aidé tout au long mes années de recherche, qui a été là pour m'aider quand j'avais vraiment besoin, et qui m'a faciliter beaucoup de choses durant mes années d'études, grâce à son expérience dans le monde professionnel et d'enseignement supérieur, merci beaucoup professeur.
Je remercie également tous les membres des laboratoires LAVETE pour leur amitié, gentillesse et partages de connaissances.

Je réserve des remerciements et des sentiments de reconnaissance très particuliers à mes parents et mes sœurs pour leur confiance en moi, leurs encouragements, leur soutien permanent et inconditionnel et surtout pour m'avoir supporté dans les moments difficiles. Ce travail n'aurait jamais abouti sans vous, il vous est dédié à 100%, que vous trouverez ici l'expression de ma très grande affection et reconnaissance.

La présente étude n'aurait pas vue la lumière sans le bienveillant soutien de plusieurs personnes. Durant la période de préparation de ma thèse, j'ai rencontré des personnes qui ont contribué à ce travail et à qui j'adresse tous mes remerciements.

Enfin, je remercie l'ensemble des thésards de la FST-Settat avec qui j'ai eu l'occasion de discuter et d'échanger. Un grand merci aussi à tous mes amis et mes proches.

Dédicace

A mes chers :

Papa, Maman,

Mes sœurs : Laila, Houda, Imane

Résumé

L'ère des Big Data est aujourd'hui en pleine vigueur parce que le monde évolue, grâce aux progrès des technologies de la communication, les personnes et les objets sont de plus en plus interconnectés, pas seulement de temps en temps, mais presque tout le temps. Les gens utilisent de plus en plus les réseaux sociaux, des objets connectés tels que des Smartphones, des véhicules avec des capteurs de localisation, cela crée beaucoup de données que les outils et les technologies traditionnelles ne peuvent pas traiter et analyser.

Mais stocker cette quantité de données n'est pas le problème majeur; nous devons utiliser ces données pour extraire des informations utiles pouvant être utilisées par les décideurs. C'est pourquoi nous avons besoin du data mining ou l'exploration des données. Il s'agit d'utiliser des outils d'analyse de données pour découvrir de nouvelles connaissances inconnues, des relations cachées entre les vastes ensembles de données. Ces modèles cachés peuvent être utilisés pour prédire les comportements futurs. Ces outils peuvent gérer des algorithmes mathématiques, des modèles statistiques et des algorithmes de machine learning. Le data mining ne concerne pas seulement la collecte et la gestion de données, elle inclut l'analyse et la prédiction de données. Le data mining et le big data sont aujourd'hui utilisés dans de nombreux domaines, parmi lesquels l'employabilité. L'emploi est la principale forme d'intégration sociale, un facteur d'amélioration des conditions de vie et de prévention de la pauvreté. L'exploitation des données relatives à l'employabilité et l'extraction de la connaissance de ces données donnera aux décideurs une vue plus large des données et des opportunités d'amélioration dans ce secteur.

Trouver une opportunité de carrière pour un jeune demandeur d'emploi est un défi pour maintenir la cohésion sociale et la stabilité politique, afin de réduire la crise de confiance dans le système éducatif et pouvoir identifier les déterminants de l'intégration professionnelle des diplômés et partager et présenter les résultats qui seront utiles aux étudiants. Ces résultats seront en mesure d'aider les décideurs publics et les responsables de la formation à mieux évaluer la qualité du système de formation, et procéder aux réajustements nécessaires.

Notre thèse traite un sujet d'actualité: l'employabilité, les techniques de data mining et le big data. L'objectif de notre thèse est de proposer un modèle de prédiction de l'employabilité utilisant les techniques de data mining en utilisant Rapid Miner comme outil de data mining, ainsi de présenter les variables qui jouent un rôle important dans la prédiction de

l’employabilité des diplômés, ainsi que la proposition d’un système de prédiction de l’employabilité dans un environnement Big Data, en utilisant l'écosystème Hadoop.

Notre thèse est divisée en quatre chapitres, le premier chapitre est intitulé: Les techniques de Data mining et Big Data, le deuxième: Employabilité et Data mining, le troisième: Prédire l'employabilité à l'aide des techniques de Data mining et des algorithmes de machine learning, et le dernier chapitre: Proposition d'un système de prédiction de l’employabilité utilisant des techniques de data mining dans un environnement Big Data.

Dans le premier chapitre, nous avons présenté les techniques de data mining et de big data, tout en indiquant la nécessité des analyses de data mining et de leurs objectifs.

Dans le deuxième chapitre, nous avons d'abord présenté la situation de l'employabilité au Maroc à l'aide des graphes et des statistiques. Ensuite, nous avons présenté le rôle et la nécessité de ces technologies dans le domaine de l’employabilité, et pourquoi utiliser le data mining dans le domaine de l’employabilité, et quels sont les avantages de ces analyses.

Dans le troisième chapitre, nous avons présenté un modèle de prédiction d’employabilité utilisant des algorithmes de machine learning de classification, ainsi que les variables qui jouent un rôle important dans la prédiction de l’employabilité des diplômés. Mais avant, nous avions présenté une étude expérimentale comparant divers algorithmes de machine learning de classification sur des données d’employabilité au Maroc: arbre de décision, régression logistique et Naïve Bayes. Le but de cette expérience est de choisir d'abord l'algorithme le plus efficace qui correspond le mieux aux données d'employabilité présentant le meilleur modèle, puis de présenter le modèle d'employabilité et les variables jouant un rôle important dans la prédiction de l'employabilité des diplômés, et finalement, nous visualisons les résultats. Dans le quatrième chapitre, un système de prédiction de l'employabilité (EPS) a été présenté. Ce système traite et analyse les données dans l'écosystème Hadoop à l'aide des différentes technologies proposées par Hadoop. Nous avons présenté l'architecture détaillée du système que nous avons utilisé, ainsi que ses caractéristiques et les phases du processus, de manière détaillée, de la collecte des données jusqu’à la visualisation des résultats, et enfin une conclusion générale et perspectives.

Keywords: Data mining, Big Data, Employabilité, Prédiction, Classification, Hadoop, Rapid Miner.

Abstract

Quite simply, the big data era is in full force today because the world is changing, thanks to advances in communication technologies, people and things are increasingly interconnected and not just part of the time, but almost all the time. People are using more and more social networks, connected objects such as smartphones, vehicles with location sensors, it creates a lot of data -Big Data- that traditional tools and technologies cannot process and analyze.

But storing this amount of data is not the major problem; we need to use this data to extract useful information which can be used by decision makers. Which is why we need data mining, it's about using data analysis tools to discover new unknown knowledge, hidden relationships between the vast datasets we have, these hidden models can be used to predict future behaviors. These tools can handle mathematical algorithms, statistical models, and machine learning methods. Data mining is not just about collecting and managing data, it includes data analysis and prediction.

Data mining and big data are used in several domains today, and employability is one of them. Employment is the main form of social integration, a factor in improving living conditions and preventing risks of poverty and vulnerability and the most appropriate indicator for assessing the level of social cohesion in a country. Mining employability data will give decision makers a great view of the data and opportunities to make improvement in this sector.

Finding a career opportunity for a young job seeker is a challenge to maintain social cohesion and political stability, to reduce the crisis of confidence in the education system.

Identify the determinants of the professional integration of graduates and share and present the results that will be useful for students, and they will be able to help public decision-makers, and those responsible for training to better assess the quality of the training system, and to proceed to the necessary readjustments.

Our thesis deals with a topical subject: employability and data mining techniques and big data. The aim of our thesis is to propose a model of employability prediction using data mining techniques using Rapid Miner, and also present the variables that play an important role in the prediction of graduates' employability, and also proposition of a system in a Big Data environment using Hadoop ecosystem.

Our thesis is divided into four chapters, chapter 1: Big Data and Data Mining Techniques, Chapter 2: Employability and Data Mining, Chapter 3: Predicting Employability Using Data Mining Techniques, and Chapter 4: Proposition of an employability prediction system using data mining techniques in a Big Data environment.

In the first chapter, we presented the techniques of data mining and big data, while indicating the necessity of data mining analyzes and their objectives.

In the second chapter, we presented first, the situation of employability in Morocco using graphs and statistics. And then, we presented the role and the necessity of these technologies in the field of employability, and why using data mining in the field of employability, and what are the advantages of these analyzes.

In the third chapter, we presented an employability prediction model using classification algorithms, as well as variables that play an important role predicting the employability of graduates. But before, we presented an experimental study comparing various classification data mining algorithms on employability data in Morocco: decision tree, logistic regression and Naïve Bayes. The aim of this experiment is to choose first the most efficient algorithm that best fits the employability data that has the best model, and then presentation of the employability model and the variables playing an important role predicting employability of the graduates. And finally we visualize the results.

In the fourth chapter, an Employability Prediction System (EPS) was presented. This system processes and analyzes data in the Hadoop ecosystem using the various technologies proposed by Hadoop. We presented the general architecture of the system we used, as well as its characteristics and the phases of the process in details, from data collection to the visualization of the results, and finally general conclusion and perspectives.

Keywords: Data mining, Big Data, Employability, Prediction, Classification, Hadoop, Rapid Miner.

Table des matières

Liste des figures

Liste des tableaux :

Liste des acronymes

SI	Système d'information
DM	Data mining
DW	Data Warehouse
EPS	Système de prédiction de l'employabilité
HDFS	Hadoop Distributed File System
IdO	Internet des objets
IoI	Internet des objets industriel
KN	Kohonen Networks
NB	Naïves bayes
NoSQL	Not only SQL
RMS	Rapid Miner Studio
RL	Régression logistique
ROC	Receiver Operating Characteristic
SaaS	Software as a Service
SGBD	Système de Gestion de Base de Données
BI	Business intelligence
SQL	Langage de requête structurée
SQOOP	SQL to Hadoop

Introduction générale

De nos jours, une quantité énorme de données est collectée et stockée dans des bases de données partout dans le monde. Avec l'utilisation des réseaux sociaux, des objets connectés tels que les smart-phones, les voitures avec des capteurs GPS, les données sont générée presque tout le temps, on parle maintenant des bases de données contenant des zetta-octets de données dans les entreprises et les centres de recherche. En fait, la quantité de données disponibles pour l'analyse augmente d'année en année, que les technologies traditionnelles ne peuvent pas traiter cette quantité de données en raison de leur capacité de stockage limitée, de leur gestion rigide, de leur manque d'évolutivité et de leur flexibilité.

Des informations et des connaissances précieuses sont cachées dans de telles bases de données. Les technologies Big data vont permettre de gérer cette énorme quantité de données et ils vont permettre d'en extraire des informations utiles avec l'utilisation des algorithmes de machine learning de data mining. Le data mining implique une collecte et un stockage efficaces des données, ainsi que leur traitement.

Il s'agit d'utiliser des outils d'analyse de données pour découvrir de nouvelles connaissances inconnues, des relations cachées entre les vastes ensembles de données dont nous disposons, ces modèles cachés peuvent être utilisés pour prédire les comportements futurs. Ces outils peuvent gérer des algorithmes mathématiques, des modèles statistiques et des algorithmes d'apprentissage automatique. Le data mining ne concerne pas seulement la collecte et la gestion des données, il inclut l'analyse et la prédiction de données. Il existe de nombreux algorithmes de machine learning que nous pouvons utiliser afin de résoudre un problème particulier.

Le data mining a prouvé sa capacité à résoudre des problèmes concrets dans de nombreux domaines, et l'employabilité en est un. L'emploi est la principale forme d'intégration sociale, un facteur d'amélioration des conditions de vie et de prévention des risques de pauvreté et de vulnérabilité et l'indicateur le plus approprié pour évaluer le niveau de cohésion sociale dans un pays. Les diplômés sont confrontés chaque année à de véritables concours pour pouvoir trouver un travail et commencer une carrière professionnelle. Et pour définir brièvement l'employabilité, du point de vue du diplômé, c'est la capacité à obtenir un nouvel emploi. L'amélioration de l'employabilité des diplômés est une problématique importante et persistante. En appliquant les algorithmes de machine learning sur ces données

d’employabilité, les décideurs disposeront d’une vue imprenable sur ces données et auront la possibilité de proposer des solutions afin d’améliorer ce secteur.

Motivation

La numérisation croissante de nos activités, la capacité sans cesse accrue à stocker des données numériques, l’accumulation d’informations en tous genres qui en découle, génère un nouveau secteur d’activité qui a pour objet l’analyse de ces grandes quantités de données. Sont alors apparues de nouvelles approches, de nouvelles méthodes, de nouveaux savoirs, et de nouvelles manières de penser et de travailler. Ainsi, cette très grande quantité de données, et son traitement et le fouille de ces données, data mining, sous-tendent de profonds bouleversements, qui touchent à l’économie, au marketing, mais aussi à la recherche et aux savoirs, et qui sont maintenant utiliser dans diverses domaines, à savoir l’employabilité.

L’emploi constitue le défi des années à venir au Maroc, qui doit faire face à une pression plus aigüe sur le marché du travail. Aujourd’hui, les parcours d’insertion des jeunes diplômés sont d’une complexité croissante, les diplômés trouvent des difficultés de plus en plus pour intégrer le marché de travail. L’employabilité c’est la capacité d’obtenir un premier emploi, de combiner différents facteurs et de former un diplômé apte au travail. Mais elle ne se résume pas à une liste de compétences acquises par les diplômés. Certes, les compétences font partie des éléments de l'employabilité, mais d'autres éléments et d’autres facteurs sont inclus.

On peut voir le data mining comme une nécessité imposée par le besoin de valoriser les données et d’en extraire des relations cachées pour permettre une bonne prise de décision. L’utilisation du Big Data et du data mining permettra de clarifier la vue et de cerner les problèmes, et présenteront également des solutions telles que l’identification des déterminants responsables de l’insertion professionnelle des diplômés. Répondre aux problèmes de l’employabilité est de plus en plus difficile, plusieurs facteurs sont inclus, déterminer ces facteurs va offrir de nombreuses facilités et opportunités aux décideurs afin d'améliorer ce domaine.

Formalisation du problème de recherche

A l'instar des autres pays du monde, la crise de l'emploi, et plus particulièrement l'emploi des jeunes, constitue un sujet d'une importance capitale et d'une sensibilité extrême, et c'est le cas pour le Maroc.

La question de l'emploi n'est pas une chose facile à régler. Elle nous incite à déployer de grands efforts pour lutter contre le chômage. Il n'existe pas de recette à cette problématique. En effet, de plus en plus de diplômés trouvent des difficultés à s'insérer dans le marché de l'emploi, et nombreux sont ceux qui restent en chômage ou s'offrent un emploi dont les exigences en qualification ne correspondent pas à leur niveau de formation. Les problématiques liées au chômage nécessitent de nouvelles solutions pour améliorer la situation, avec l'évolution des nouvelles technologies.

Les technologies de communication ont progressé et ils ont changé la façon de vivre des gens, les personnes et les objets sont de plus en plus interconnectés, pas seulement de temps en temps, mais presque tout le temps. Les gens commencent à utiliser de nombreux objets connectés, tels que des voitures, des téléphones, ils s'impliquent de plus en plus dans les réseaux sociaux. Les données sont générées presque tout le temps, ce qui donne cette énorme quantité de données prêtes à être utilisées et stockées.

Mais les technologies traditionnelles ne peuvent pas traiter cette quantité de données en raison de leur capacité de stockage limitée, de leur gestion rigide, de leur manque d'évolutivité et de leur flexibilité, qui nous ont conduits à la révolution du Big Data. Mais le Big Data ne consiste pas simplement à stocker les données; l'extraction de la connaissance est le processus le plus important. Il est nécessaire d'analyser cette énorme quantité de données et d'en extraire des informations utiles. Par conséquent, des algorithmes de machine learning avancés sont nécessaires pour obtenir des résultats précis et des connaissances permettant de prévoir les observations futures afin d'aider les décideurs à prendre des décisions.

L'utilisation du Big Data et du data mining permettra de clarifier la vue et de cerner les problèmes, et apportera également des solutions telles que l'identification des déterminants responsables de l'insertion professionnelle des diplômés, que ce soit en raison du programme universitaire du diplômé ou du marché du travail, ou le domaine d'études choisi par les diplômés. Répondre à de telles questions pourrait être utile aux diplômés ainsi qu'aux

chercheurs et aux autorités publiques pour mieux évaluer le système qualité de la formation et procéder aux ajustements nécessaires.

Notre thèse traite un sujet d'actualité : l'employabilité et les techniques de data mining et big data. L'objectif de notre thèse est de proposer un modèle de prédiction d'employabilité utilisant des techniques et des algorithmes de data mining dans un environnement Big Data, de présenter également les variables jouant un rôle important dans la prédiction de l'employabilité des diplômés, ainsi que la proposition d'un système dans un environnement Big Data.

Organisation de la thèse

Notre thèse est divisée en quatre chapitres, chapitre 1: Techniques de Data mining et Big Data, chapitre 2: Employabilité et data mining, chapitre 3: Prédire l'employabilité à l'aide des techniques de data mining et des algorithmes de machine learning, et chapitre 4: Proposition d'un système de prédiction de l'employabilité utilisant des techniques de Data mining dans un environnement Big Data.

Dans le premier chapitre, on a présenté les techniques de data mining et du big data, tout en indiquant la nécessité des analyses du data mining et leurs objectifs.

Dans le deuxième chapitre, on a présenté premièrement en utilisant des graphes et des statistiques la situation de l'employabilité au Maroc. Et deuxièmement, le rôle et la nécessité de ces technologies dans le domaine de l'employabilité, et pourquoi utilisé le data mining dans le domaine de l'employabilité, et quelles sont les avantages de ces analyses.

Dans le troisième chapitre, on a présenté un modèle de prédiction de l'employabilité utilisant des algorithmes de machine learning de classification, ainsi que les variables jouant un rôle important dans la prédiction de l'employabilité des diplômés. Mais avant, on a présenté une étude expérimentale comparant divers algorithmes de data mining de classification sur des données d'employabilité au Maroc: arbre de décision, régression logistique et Naïve Bayes. L'objectif en premier est de choisir l'algorithme le plus efficace et le mieux adapté aux données d'employabilité qui présente le meilleur modèle. Et après cette étape la présentation du modèle d'employabilité et des variables jouant un rôle important dans la prédiction de l'employabilité des diplômés, et enfin la visualisation des résultats.

Dans le quatrième chapitre, on a présenté un système de prédiction de l'employabilité (EPS), dans un environnement Big Data. Ce système traite et analyse les données dans l'écosystème Hadoop à l'aide des différentes technologies proposées par Hadoop. Nous avons présenté l'architecture générale du système que nous avons utilisé, ainsi que ses caractéristiques et les phases du processus, depuis la collecte des données jusqu'à la visualisation des résultats.

Et enfin une conclusion générale et perspectives.

Chapitre 1 : Les techniques de Data mining et Big Data

« Vous pouvez avoir des données sans informations, mais vous ne pouvez pas avoir des informations sans données. »

Daniel KEYS Moran

Introduction

Aujourd'hui les données se développent beaucoup, tellement que les technologies traditionnelles et les systèmes de bases de données actuels ne peuvent pas gérer cette énorme quantité de données et la traiter efficacement. Le Big Data est intervenu pour résoudre ces problèmes et proposer des solutions permettant de traiter et d'analyser efficacement cette énorme quantité de données. La combinaison du Big Data et de l'exploration de données peut être très utile et offre aux décideurs des opportunités pour tirer parti des résultats et prédire l'avenir. Dans ce chapitre, on va présenter une compréhension du data mining, des algorithmes de machine learning et du big data.

I. Les techniques du Data mining

1. Définir le data mining

Bien que les méthodes de DM (Data mining) soient appliquées depuis les années 1960, le terme data mining est apparu vers 1990. Il existe de nombreuses définitions du data mining. Fayyad, G. Piatetsky-Shapiro [1] a défini le DM comme une tentative de trouver des modèles utiles dans les données.

Depuis les années quatre-vingt-dix, les statisticiens, les analystes de données et les experts en systèmes d'information utilisent indifféremment ces termes. Encyclopedia Britannica [2] définit le DM comme «le processus de découverte de modèles et de relations intéressants et utiles dans de grands volumes de données avec les sous-domaines de la modélisation prédictive, de la modélisation descriptive, de l'exploration de modèles et de l'extraction d'anomalies. L'objectif de DM consiste à extraire de nouvelles connaissances et à approfondir la compréhension d'un ensemble de données, qui peuvent continuer à être utilisées pour la prise de décision. Le DM chevauche d'autres disciplines, notamment les statistiques.

Il est tentant de se lancer directement dans le data mining [3], mais d'abord, il faut préparer les données. Cela implique d'examiner de plus près les attributs et les valeurs de données. Les données du monde réel sont généralement bruyantes, avec un volume énorme (on parle maintenant de giga-octets ou plus) et peuvent provenir de plusieurs sources. La connaissance

des données est utile pour le prétraitement des données, première tâche importante du processus d'exploration de données. La connaissance des statistiques concernant chaque attribut facilite la saisie des valeurs manquantes, des valeurs lisses et bruyantes et de la détection des valeurs aberrantes lors du prétraitement des données. La connaissance des attributs et des valeurs d'attribut peut également aider à résoudre les incohérences survenues lors de l'intégration des données. Tracer les mesures de la tendance centrale nous montre si les données sont symétriques ou asymétriques. Les graphiques, histogrammes et nuages de points sont d'autres représentations graphiques de descriptions statistiques de base. Celles-ci peuvent toutes être utiles lors du prétraitement des données et peuvent fournir des informations sur les domaines d'exploration.

2. Les objets de données et les types d'attributs

Les ensembles de données sont constitués d'objets de données [4]. Un objet de données représente une entité. Dans une base de données de ventes, les objets peuvent être des clients, des éléments de magasin et des ventes; dans une base de données médicale, les objets peuvent être des patients; dans une base de données universitaire, les objets peuvent être des étudiants, des professeurs et des cours. Les objets de données sont généralement décrits par des attributs [5]. Les objets de données peuvent également être appelés exemples, instances, points de données ou objets. Si les objets de données sont stockés dans une base de données, ils sont des nuplets de données. En d'autres termes, les lignes d'une base de données correspondent aux objets de données et les colonnes correspondent aux attributs. Dans cette section, nous définissons les attributs et examinons les différents types d'attributs.

2.1. Qu'est-ce qu'un attribut ?

Un attribut [6] est un champ de données représentant une caractéristique ou un élément d'un objet de données. Les noms, attributs, dimensions, entités et variables sont souvent utilisés de manière interchangeable dans la littérature. Le terme dimension est couramment utilisé dans l'entreposage de données. La littérature sur le machine learning utilise généralement le terme caractéristique, alors que les statisticiens préfèrent utiliser le terme variable. Les professionnels du data mining et des bases de données utilisent couramment le terme attribut. Les attributs décrivant un objet client, par exemple, peuvent inclure, l'identifiant, le nom et l'adresse du client. Les valeurs observées pour un attribut donné sont appelées observations. Un ensemble d'attributs utilisés pour décrire un objet donné est appelé un vecteur d'attribut

(ou vecteur de caractéristique). La distribution des données impliquant un attribut (ou variable) est appelée uni-variée. Une distribution à deux variables implique deux attributs, et ainsi de suite. Le type d'un attribut est déterminé par l'ensemble des valeurs possibles (nominales, binaires, ordinales ou numériques) que l'attribut peut avoir.

Maintenant on va présenter les différents types qu'un attribut peut avoir.

2.2. Attributs nominaux

Les valeurs d'un attribut nominal [7] sont des symboles ou des noms d'éléments. Chaque valeur représente un type de catégorie, de code ou d'état, et les attributs nominaux sont également appelés catégoriques. Les valeurs n'ont pas d'ordre significatif. Les valeurs sont également appelées énumérations. Supposons que la couleur des cheveux et l'état matrimonial soient deux attributs décrivant les objets personne. Les valeurs possibles pour la couleur des cheveux sont les suivantes: noir, brun, blond, rouge, auburn, gris et blanc. L'attribut état matrimonial peut prendre les valeurs : célibataire, marié, divorcé et veuf. La couleur des cheveux et l'état matrimonial sont des attributs nominaux. Un autre exemple d'attribut nominal est l'occupation, avec les valeurs enseignant, dentiste, programmeur, agriculteur, etc. Bien que les valeurs d'un attribut nominal soient des symboles ou des noms de choses, il est possible de représenter ces symboles ou «noms» avec des nombres. Avec la couleur des cheveux, par exemple, nous pouvons attribuer un code de 0 pour le noir, 1 pour le brun, etc. Un autre exemple est l'identifiant du client, avec des valeurs possibles entièrement numériques. Toutefois, dans de tels cas, les chiffres ne sont pas destinés à être utilisés de manière quantitative. Autrement dit, les opérations mathématiques sur les valeurs des attributs nominaux ne sont pas significatives. Il n'a aucun sens de soustraire un numéro d'identifiant client à un autre, contrairement à, par exemple, soustraire une valeur d'âge à un autre (où âge est un attribut numérique).

2.3. Attribut binaire

Un attribut binaire [8] est un attribut nominal ne comportant que deux catégories ou états: 0 ou 1, où 0 signifie généralement que l'attribut est absent et 1 signifie qu'il est présent. Les attributs binaires sont appelés booléens si les deux états correspondent à True et à False. Étant donné l'attribut fumeur décrivant un objet du patient, 1 indique que le patient fume, tandis que 0 indique que ce n'est pas le cas. De même, supposons que le patient subisse un test médical

qui a deux résultats possibles. L'attribut test médical est binaire, la valeur 1 indiquant que le résultat du test pour le patient est positif, tandis que la valeur 0 indique que le résultat est négatif. Un attribut binaire est symétrique si ses deux états ont la même valeur et le même poids; c'est-à-dire qu'il n'y a pas de préférence sur le résultat qui devrait être codé 0 ou 1. L'un de ces exemples pourrait être l'attribut genre ayant comme valeurs masculin et féminin.

2.4. Attribut ordinal

Un attribut ordinal [9] est un attribut avec des valeurs possibles ayant un ordre ou un classement significatif entre elles, mais la magnitude entre les valeurs successives n'est pas connue. Supposons que la taille de la boisson corresponde à celle des boissons disponibles dans un fast-food. Cet attribut nominal a trois valeurs possibles: petit, moyen et grand. Les valeurs ont une séquence significative. D'autres exemples d'attributs ordinaux comprennent le grade (par exemple, A +, A, A-, B +, etc.) et le rang professionnel. Les grades professionnels peuvent être énumérés dans un ordre séquentiel: par exemple, professeurs adjoints, professeurs associés, personnel spécialisé, caporal, sergent et grades militaires. Les attributs ordinaux sont utiles pour enregistrer des évaluations subjectives de valeurs impossibles à mesurer objectivement. Ainsi, les attributs ordinaux sont souvent utilisés dans les enquêtes pour les notations. Dans un sondage, les participants ont été invités à évaluer leur degré de satisfaction en tant que clients. La satisfaction de la clientèle comportait les catégories ordinales suivantes: 0: très insatisfait, 1: un peu insatisfait, 2: neutre, 3: satisfait et 4: très satisfait. Les attributs ordinaux peuvent également être obtenus à partir de la discrétisation des quantités numériques en divisant la plage de valeurs en un certain nombre de catégories. La tendance centrale d'un attribut ordinal peut être représentée par son mode et sa médiane (la valeur moyenne dans une séquence ordonnée), mais la moyenne ne peut pas être définie. Les attributs nominaux, binaires et ordinaux sont qualitatifs. C'est-à-dire qu'ils décrivent une caractéristique d'un objet sans donner une taille ou une quantité réelle. Les valeurs de ces attributs qualitatifs sont caractérisées par les mots. Si des nombres entiers sont utilisés, ils représentent des catégories, par opposition aux quantités mesurables (par exemple, 0 pour une petite taille de boisson, 1 pour une moyenne et 2 pour une grande).

2.5. Attributs Numériques

Un attribut numérique [10] est quantitatif; c'est-à-dire qu'il s'agit d'une quantité mesurable, représentée par des valeurs entières ou réelles. Les attributs numériques peuvent être mis à

l'échelle par intervalles ou par rapports. Les attributs à l'échelle d'intervalle sont mesurés sur une échelle d'unités de taille égale. Les valeurs des attributs d'échelle d'intervalle ont un ordre et peuvent être positives, nulles ou négatives. Ainsi, en plus de fournir un classement des valeurs, ces attributs nous permettent de comparer et de quantifier la différence entre les valeurs.

2.6. Attributs à l'échelle du ratio

Attributs à l'échelle du ratio [11] est un attribut numérique avec un point zéro inhérent. En d'autres termes, si une mesure est mise à l'échelle du rapport, nous pouvons parler d'une valeur comme étant un multiple (ou un rapport) d'une autre valeur. De plus, les valeurs sont ordonnées et nous pouvons également calculer la différence entre les valeurs, ainsi que la moyenne, la médiane et le mode. Exemple d'attributs à l'échelle des ratios comprennent les attributs de comptage tels que les années d'expérience (par exemple, les objets sont des employés) et le nombre de mots (par exemple, les objets sont des documents). Des exemples supplémentaires incluent des attributs permettant de mesurer le poids, la hauteur, les coordonnées de latitude et de longitude (par exemple, lorsqu'on regroupe des maisons).

2.7. Attributs discrets ou continus

Il y a plusieurs façons d'organiser les types d'attribut. Les types ne sont pas mutuellement exclusifs. Les algorithmes de classification élaborés à partir du domaine de l'apprentissage automatique parlent souvent d'attributs comme étant discrets ou continus. Chaque type peut être traité différemment. Un attribut discret [12] a un ensemble infini de valeurs, qui peuvent ou non être représentées par des entiers. Les attributs couleur de cheveux, fumeur, test médical et taille de boisson ont chacun un nombre fini de valeurs et sont donc discrets. Les attributs discrets peuvent avoir des valeurs numériques, telles que 0 et 1 pour les attributs binaires ou les valeurs 0 à 110 pour l'attribut âge. Un attribut est infini si l'ensemble des valeurs possibles est infini, mais les valeurs peuvent être mises en correspondance un à un avec des nombres naturels. Par exemple, l'attribut identifiant du client est infiniment comptable. Le nombre de clients peut atteindre l'infini, mais en réalité, l'ensemble de valeurs réel est comptable (les valeurs pouvant être mises en correspondance un à un avec l'ensemble des entiers). Les codes postaux sont un autre exemple. Si un attribut n'est pas discret, il est continu. Les termes attribut numérique et attribut continu sont souvent utilisés de manière interchangeable dans la littérature. En pratique, les valeurs réelles sont représentées à l'aide

d'un nombre fini de chiffres. Les attributs continus sont généralement représentés sous forme de variables à virgule flottante.

3. Les techniques de data mining :

Une grande variété de techniques et de méthodes sont couramment utilisées dans les applications de data mining. Le data mining implique souvent le regroupement, la construction de modèles de régression prédictive ou de classification, la sélection d'attributs et des caractéristiques.

Ces techniques peuvent être basées sur des statistiques, la théorie des probabilités, des réseaux bayésiens, des arbres de décision, des règles d'association, etc.

Dans cette section, nous présenterons les différentes techniques de data mining, mais en premier lieu, le graphe de la pyramide de sagesse de data mining [13].

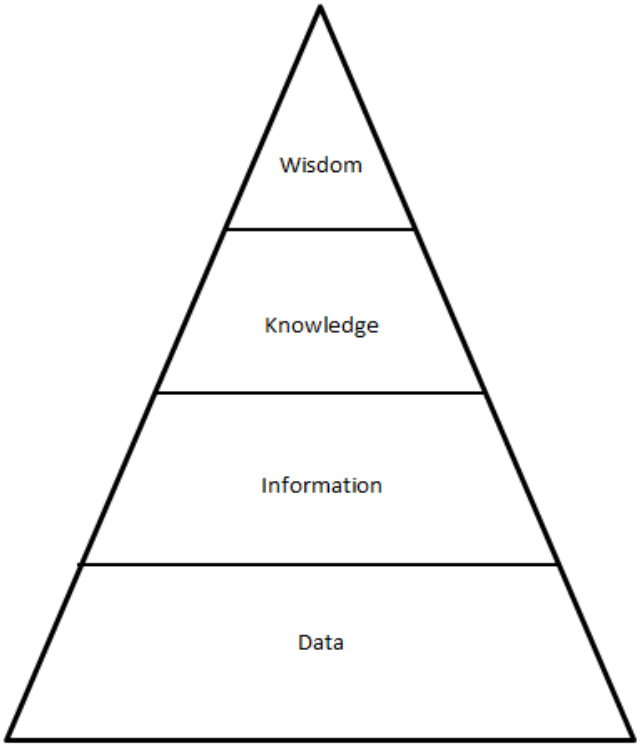

Figure 1. Graphe de la pyramide de sagesse de data mining

- **Les données**

– Les données sont des faits, des nombres, ou des textes pouvant être traités par un ordinateur. Aujourd'hui, les entreprises accumulent de vastes quantités de données sous différents formats, dans différentes quantités de données. Parmi ces données, on distingue :

Les données opérationnelles ou transactionnelles telles que les données de ventes, de coûts, d'inventaire, de tickets de caisse ou de comptabilité.

Les données non opérationnelles, telles que les ventes industrielles, les données prévisionnelles, les données macro-économiques.

Les métadonnées, à savoir les données concernant les données elles-mêmes, telles que les définitions d'un dictionnaire de données.

- **Information**

– Les patterns, associations et relations entre toutes ces données permettent d'obtenir des informations. Par exemple, l'analyse des données de transaction d'un point de vente permet de recueillir des informations sur les produits qui se vendent, et à quel moment ont lieu ces ventes.

- **Savoir**

– Les informations peuvent être converties en savoir pour prédire des tendances futures. Par exemple, l'information sur les ventes au détail d'un supermarché peut être analysée dans le cadre d'efforts promotionnels, pour acquérir un savoir au sujet des comportements d'acheteurs. Ainsi, un producteur peut déterminer quels produits doivent faire l'objet d'une promotion à l'aide du Data Mining.

- **Sagesse**

– Placer les connaissances ou les savoirs dans un cadre pour les appliquer à des situations différentes pour pouvoir faire un changement et améliorer un domaine.

Maintenant on présente les techniques de data mining, ils sont principalement classées en deux grandes catégories : les techniques prédictives et les techniques descriptives.

3.1. Les techniques prédictives (supervisées)

Les techniques d'exploration de données supervisées [14] sont appropriées lorsqu'on a une valeur cible spécifique à prédire concernant les données. Les cibles peuvent avoir deux résultats possibles ou plus, ou même être une valeur numérique continue. On a deux grandes familles : classification et prédiction.

3.1.1. Classification :

La classification [15] est l'opération consistant à séparer diverses entités en plusieurs classes. Ces classes peuvent être définies par des règles métier, des limites de classe ou une fonction

mathématique. L'opération de classification peut être basée sur une relation entre une affectation de classe connue et des caractéristiques de l'entité à classer. Ce type de classification s'appelle supervisé. Si aucun exemple connu d'une classe n'est disponible, la classification n'est pas supervisée. L'approche de classification non supervisée la plus courante est la segmentation (clustering). Les applications les plus courantes de la technologie de regroupement concernent l'analyse d'affinité des produits de vente au détail (y compris l'analyse du panier de marché) et la détection de la fraude. L'exploration de données présente deux types généraux de problèmes de classification supervisée: la classification binaire (une seule variable cible) et les multiples classifications: plusieurs variables cibles. Un exemple d'analyses avec une seule variable cible est un modèle permettant d'identifier les répondants à probabilité élevée de lancer des campagnes de publipostage. Un exemple d'analyses avec plusieurs variables cibles est un modèle de diagnostic qui peut avoir plusieurs résultats possibles (grippe, angine de streptocoque, etc.).

Exemples des algorithmes de machine learning de classification:

- **L'arbre de décision:**

Un arbre de décision [16] est un algorithme qui divise un jeu de données en un certain nombre de segments en forme de branche. Les arbres de classification et de régression sont un autre algorithme plutôt ancien d'apprentissage automatique, introduit par Breiman [17]. Cet algorithme contient des implémentations modernes se retrouvent dans de nombreux outils d'analyse et fonctionnent souvent pour produire des modèles très prédictifs. Il reste l'une des applications d'arbre de décision les plus adaptables. Il peut être utilisé pour classer les ensembles de données en groupes et pour la prédiction de valeurs réelles, similaires à la régression linéaire. Une élaboration courante est l'algorithme des arbres boostés, qui est devenu très populaire pour la construction de modèles de classification.

- **La régression logistique: Binaire, Multinomiale**

La régression logistique [18] est utilisée dans la classification plutôt que dans la prédiction numérique. La RL est utilisée pour modéliser la relation non linéaire entre Y et les effets combinés des variables indépendantes. Cette relation est utilisée pour modéliser la probabilité d'occurrence d'un événement (une variable binaire, telle que oui / non ou 1/0), à l'aide de prédicteurs qualitatifs ou numériques. Cet algorithme a été largement utilisé dans les

entreprises pour prédire les événements liés à la vente d'un produit ou d'un groupe de produits spécifique ou tout événement ayant un résultat binaire.

Ci-dessous les avantages et les limitations [19] de la régression logistique :

- **Avantages de la régression logistique**

1. L'algorithme est très bien développé, permet l'interprétation des résidus et peut également être évalué avec la valeur R (coefficient de détermination), mais il est calculé en fonction des probabilités de la courbe logistique.
2. Il n'existe aucune hypothèse d'homogénéité de variance, contrairement à la régression linéaire.
3. Cela fonctionne sur les variables dépendantes binaires.
4. Autrefois, l'inconvénient d'être limité à l'analyse des nombres était un avantage, car les transformées numériques pouvaient être utilisées pour modifier les distributions de variables afin de les rendre plus conformes aux hypothèses de l'algorithme.
5. Elle représente une part importante de toute relation non linéaire entre la variable cible et les effets combinés des variables de prédiction, car cette réponse est définie avec une fonction de journalisation naturelle (une sorte de fonction exponentielle non linéaire).
6. Les arbres de décision recherchent des limites de décision rectangulaires dans l'espace des caractéristiques, tandis que la régression logistique permet de rechercher des lignes non rectangulaires pour séparer les catégories. Cet avantage facilite beaucoup l'interprétation des prédicteurs importants dans les classifications de régression logistique qu'un arbre de décision pour une valeur de classe cible donnée.
7. Certaines entreprises et industries qui se sont normalisées autour de mesures statistiques classiques (par exemple, l'agriculture) sont à l'aise avec son héritage de régression.

- **Limitations de la régression logistique**

Contrairement aux arbres de décision et aux réseaux de neurones, les variables prédictives sont supposées être indépendantes les unes des autres dans leur relation à la variable cible. Dans les applications métier, ce n'est presque jamais le cas. Il faut beaucoup plus de données que l'analyse discriminante, par exemple, pour créer des modèles stables. Une bonne règle empirique consiste à fournir 50 à 100 fois plus de lignes de données que de variables indépendantes.

- **Naïve Bayes:**

Naïves bayes (NB) ou la prédiction bayésienne [20] suivait davantage les schémas de la pensée humaine que l'analyse statistique classique ou même les algorithmes d'apprentissage automatique. L'inconvénient majeur de ce fait est que deux humains peuvent (et sont souvent en désaccord) dans les décisions qu'ils prennent à la suite de cette réflexion. Mais il existe bien d'autres situations dans lesquelles l'approche bayésienne de la vérité est beaucoup plus appropriée et peut même être meilleure lorsque nous devons faire face à la nécessité de classer une entité dans le monde autour de nous.

Les Bayésiens soutiennent que dans de nombreux cas, cette deuxième source de preuves est essentielle au bon classement de l'entité. Ils intègrent ces sources de preuves en les multipliant pour calculer la probabilité conditionnelle de la survenue d'un événement, sur la base de toutes les occurrences en compétition dans le passé. Pour les Bayésiens, la classification est un jugement reposant sur une probabilité conditionnelle. Dans de nombreuses situations de classification impliquant des attributs de données, nous en savons relativement peu sur l'entité que nous classons, et il peut être acceptable de considérer le processus de classification comme un jugement. Suivant cette logique, la classification bayésienne naïve est devenue une technique utile dans de nombreux domaines de l'exploration de données.

3.1.2. Prédiction :

La prédiction [21] dans l'exploration de données consiste à identifier des points de données uniquement sur la description d'une autre valeur de données associée. Ce n'est pas nécessairement lié aux événements futurs, mais les variables utilisées sont inconnues.

La prédiction dérive la relation entre une chose connue et une chose qu'on veut prédire pour pouvoir y référer ultérieurement.

Par exemple, un responsable marketing utilise les modèles de prédiction [22] dans l'exploration de données pour prédire le montant des dépenses d'un client lors d'une vente, de sorte que le montant de la vente à venir puisse être planifié en conséquence. La prédiction dans l'exploration de données est connue sous le nom de prédiction numérique. En règle générale, l'analyse de régression est utilisée pour la prédiction.

Exemple d'algorithme de machine learning prédictive:

La régression linéaire

La régression linéaire [23] a été proposée pour la première fois par Sir Francis Galton (1822-1911). Galton a inventé le terme régression [24] pour décrire l'observation selon laquelle la majorité des pères très grands avaient des fils plus petits et les pères très courts avaient des fils plus grands qu'eux. La tendance de cette progression en hauteur allait vers la taille moyenne. Ce phénomène a été appelé régression à la moyenne. Son analyse de cet effet est devenue simplement une régression. Les principaux objectifs de la régression linéaire sont les suivants:

- Déterminer s'il existe une relation entre une variable et une autre (ou un ensemble d'autres);
- Décrire la nature de cette relation;
- En quantifier l'exactitude;
- Evaluer contributions relatives de chaque variable, si plusieurs variables sont utilisées.

3.2. Les techniques descriptives (non-supervisées)

Les techniques de regroupement [25] ont pour but de détecter des sous-groupes similaires parmi un grand nombre de cas et d'attribuer ces observations aux groupes. Les groupes sont attribués à un numéro séquentiel pour les identifier dans les rapports de résultats. Un bon algorithme de clustering trouvera le nombre de clusters ainsi que les membres de chacun. Les groupes de cas doivent être beaucoup plus similaires les uns aux autres que les cas d'autres groupes. Un exemple typique d'analyse de groupe est une étude de marché dans laquelle un nombre important de variables liées au comportement du consommateur sont mesurées pour un grand échantillon de répondants. L'objet de l'étude est de détecter les «segments de marché», c'est-à-dire les groupes de répondants qui se ressemblent davantage entre eux qu'ils ne le sont pour les répondants d'autres groupes. La nécessité de déterminer en quoi ces groupes sont différents est tout aussi importante que l'identification de tels groupes.

3.2.1. Segmentation (Clustering):

Clustering [26] dans son sens le plus général, désigne la méthodologie de partitionnement d'éléments en groupes en fonction de certaines caractéristiques communes. L'analyse de cluster a été introduite pour la première fois en anthropologie par Driver et Kroeber en 1932.

De nos jours, le Clustering est devenu un instrument valable pour résoudre des problèmes complexes de science et de statistiques informatiques. En particulier, il est très utilisé dans le

data mining et efficace pour la découverte de modèle présentant un intérêt spécifique à partir de données afin de soutenir le processus de découverte de connaissances. Généralement, afin d'optimiser la solution de clustering, les données sont pré-traitées avant d'appliquer une approche de clustering. Ci-dessous les algorithmes les plus utilisés en Clustering.

- **K-means**

Le fonctionnement de base de k-means [27] est simple: étant donné un nombre fixe (k) de clusters, attribuez des observations à ces clusters de manière à ce que les moyennes entre les clusters (pour toutes les variables) soient aussi différentes que possible. La différence entre les observations est mesurée en termes de l'une des mesures de distance.

Pour les variables qualitatives, toutes les distances sont binaires (0 ou 1). La variable est affectée de la valeur 0 lorsque la catégorie d'une observation est identique à celle de la fréquence la plus élevée d'un cluster, sinon la valeur 1 est attribuée.

- **Kohonen Networks**

Le principe de Kohonen networks (KN) [28] est d'utiliser une forme de réseau de neurones dans laquelle il n'y a pas de variables dépendantes connues pour une utilisation en cluster non supervisé. Le réseau est formé en attribuant des centres de cluster à une couche radiale en soumettant de manière itérative des modèles de formation au réseau et en ajustant le centre d'unité radiale gagnant le plus proche et ses voisins en fonction du modèle de formation. L'opération résultante provoque l'auto-organisation des points de données en clusters. Un acronyme abrégé pour les réseaux Kohonen est une carte de fonctions auto-organisée.

Un réseau Kohonen présente les caractéristiques suivantes:

- Concurrence : Pour chaque modèle d'entrée, les neurones se font concurrence.
- La coopération : Le neurone gagnant détermine la localisation spatiale d'un voisinage topologique de neurones excités, fournissant ainsi la base d'une coopération entre les neurones.
- Adaptation synaptique : Les neurones excités ajustent leurs poids synaptiques pour améliorer leur réponse à un modèle d'entrée similaire.

3.2.2. Association:

La découverte des règles d'association [29] partage de nombreuses caractéristiques de l'induction de règles de classification. Les deux utilisent des règles pour caractériser les régularités dans un ensemble de données. Cependant, ces paradigmes de découverte à deux axes diffèrent considérablement par leur intention. Alors que l'induction des règles de classification se concentre sur l'acquisition d'une capacité à faire des prédictions, la découverte de règles d'association se concentre sur la fourniture d'informations à l'utilisateur. Plus précisément, il se concentre sur la détection et la caractérisation des relations inattendues entre des éléments de données. L'induction des règles de classification utilise généralement la recherche heuristique pour trouver un petit nombre de règles couvrant conjointement la majorité des données d'apprentissage. La règle d'association utilise généralement une recherche complète pour trouver un grand nombre de règles sans tenir compte toutes les données de formation. En raison de l'accent mis sur la découverte de petits jeux de règles, les systèmes d'induction de règles de classification effectuent souvent des choix entre des règles alternatives présentant des performances similaires. Bien que le système d'apprentissage informatisé puisse ne pas disposer d'une base permettant de distinguer ces règles, leur valeur pour l'utilisateur peut différer considérablement en raison de facteurs extérieurs à ceux représentés dans les données.

- **Apriori**

Apriori [30] est un algorithme classique d'apprentissage des règles d'association. Apriori est conçu pour fonctionner sur des bases de données contenant des transactions, par exemple, des collections d'articles achetés par des clients ou des détails sur la fréquentation d'un site Web.

Apriori utilise une approche ascendante, dans laquelle des sous-ensembles fréquents sont étendus élément par élément (étape appelée génération de candidats), et des groupes de candidats sont testés par rapport aux données. L'algorithme se termine lorsqu'aucune autre extension réussie n'est trouvée. Apriori utilise une recherche en largeur d'abord et une structure arborescente pour compter efficacement les ensembles d'éléments candidats. Il génère des ensembles d'éléments candidats de longueur k à partir d'ensembles d'éléments de longueur k - 1. Ensuite, il élague les candidats dont le sous-motif est peu fréquent. Après, il

analyse la base de données des transactions pour déterminer les ensembles d'articles fréquents parmi les candidats. Apriori, bien qu'historiquement significatif, souffre d'un certain nombre d'inefficacités ou de compromis, qui ont engendré d'autres algorithmes. La génération de candidats génère un grand nombre de sous-ensembles; l'algorithme tente de charger le plus grand nombre possible de candidats avant chaque balayage.

4. Processus général du data mining :

Dans cette section, on va présenter le processus général du data mining [31], en expliquant en détails chaque phase de ce processus.

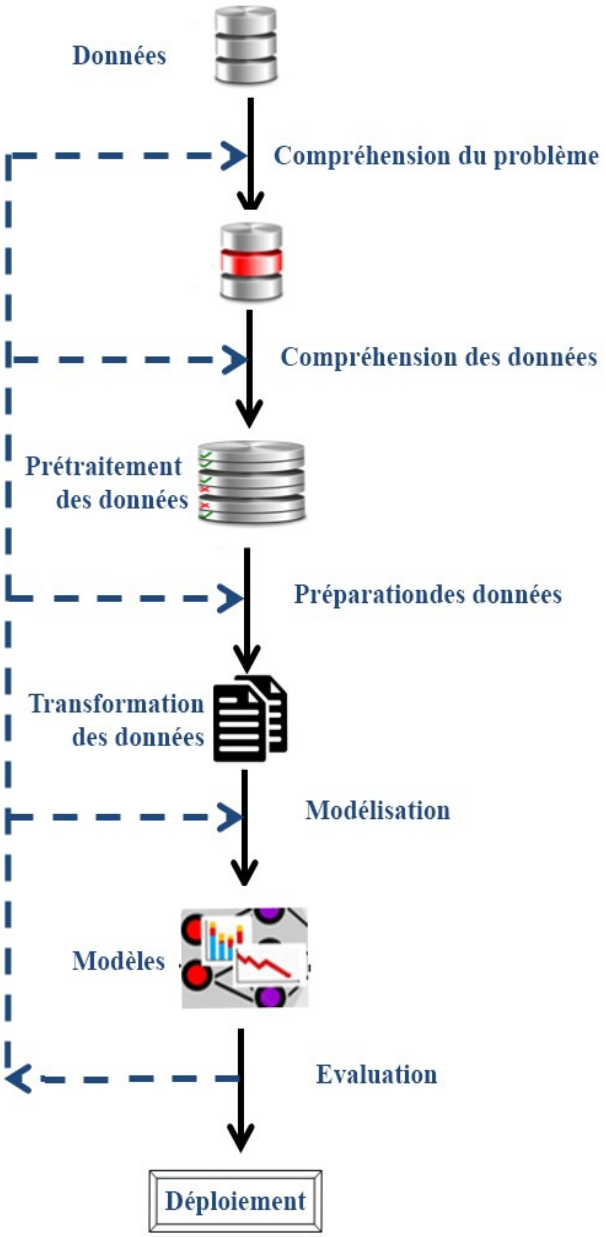

Figure 2. Le processus général du data mining

4.1. Compréhension du problème :

Dans cette phase, les objectifs, les ressources et les restrictions sont formulés en termes de problèmes d'exploration de données. Les objectifs et l'approche de la solution sont déterminés : Définir la problématique et collecter les données.

4.2. Compréhension des données :

Dans la deuxième étape, les données disponibles et les sources de données sont analysées. Outre l'évaluation des bases de données existantes, on vérifie comment les données peuvent être intégrées ou si de nouvelles données doivent être spécifiquement collectées.

4.3. Préparation des données :

La phase la plus intensive est généralement la préparation des données. Si nécessaire, des données externes sont identifiées et intégrées, et des données d'enquête sont collectées. Les données de différentes sources sont liées. Les valeurs manquantes sont remplacées à l'aide de diverses techniques, les données déviantes sont éliminées et les variables sont sélectionnées pour le modèle.

4.4. Modélisation :

Dans un processus itératif, différents modèles sont proposés, testés et optimisés. Les processus de modélisation et de préparation des données sont réalisés en coordination, en fonction du type de données. Des algorithmes de machine learning d'exploration de données sont choisis et appliqués afin de choisir le plus performant au cours de la phase d'évaluation.

4.5. Evaluation :

Les modèles sélectionnés sont appliqués aux données et évalué, c'est-à-dire testé pour leur capacité à répondre aux objectifs prédéfinis. Si les critères de contrôle ne sont pas remplis, le processus recommence depuis le début.

4.6. Déploiement :

Un modèle réussi est mis en œuvre. La forme de mise en œuvre dépend des conditions et de la nature de la mission. Le modèle peut être appliqué en livrant un script, en mode SaaS (Software as a Service) ou par une implémentation dans une infrastructure informatique.

D.Hand [32] a résumé quelques avertissements sur l'utilisation des outils d'exploration de données pour la découverte de modèles.

- **La qualité des données:**

Les méthodes d'exploration de données peuvent ne pas révéler explicitement une mauvaise qualité des données, et cette mauvaise qualité produira de mauvais modèles. Il est possible que de mauvaises données soutiennent la construction d'un modèle avec une prévisibilité relativement élevée.

- **Opportunité:**

De multiples opportunités peuvent transformer l'impossible en un événement très probable. Hand y fait référence comme étant le problème de la multiplicité ou la loi des grands nombres. Par exemple, les chances qu'une personne gagne à la loterie aux États-Unis sont

extrêmement faibles. Mais les chances de gagner deux fois aux États-Unis sont en réalité meilleures que jamais.

- **Interventions:**

Un des résultats inattendus d'un modèle d'exploration de données est que certaines modifications seront apportées pour l'invalider. Par exemple, l'élaboration de modèles de détection de la fraude peut conduire à l'adoption de mesures préventives efficaces à court terme. Mais peu de temps après, les fraudeurs peuvent évoluer dans leur comportement pour éviter ces interventions dans leurs opérations.

- **Séparabilité:**

Il est souvent difficile de séparer les informations intéressantes des informations non désirées d'un ensemble de données. De nombreux modèles peuvent exister dans un ensemble de données, mais seuls quelques-uns peuvent intéresser l'exploitation de données pour résoudre un problème donné. La définition de la variable cible est l'un des facteurs les plus importants qui déterminent le modèle que l'algorithme trouvera. Dans un but, la rétention d'un client peut être définie de manière très distincte en utilisant une variable telle que Close_date pour dériver la cible. Dans un autre cas, une baisse de 70% de l'activité des clients au cours des deux dernières périodes de facturation pourrait constituer le meilleur moyen de définir la variable cible. Le modèle trouvé par l'algorithme d'exploration de données pour le premier cas pourrait être très différent de celui du deuxième cas.

- **L'évidence:**

Certains modèles découverts dans un ensemble de données peuvent ne pas être utiles du tout car ils sont assez évidents, même sans analyse d'exploration de données. Par exemple, les fraudes par chèque surviennent le plus souvent chez les clients possédant un compte courant.

Conclusion

Data mining consiste à trier des ensembles de données afin d'identifier des modèles et d'établir des relations pour résoudre des problèmes grâce à l'analyse de données. Les outils de data mining permettent de prédire les tendances futures. Dans cette partie, on a listé les différents algorithmes de data mining, les types de problèmes de data mining ainsi que les algorithmes appropriés pour chaque problème, connaitre les différents algorithmes de data mining est nécessaire pour qu'on puisse faire le bon choix afin de l'appliquer sur les données et en

extraire la connaissance. Big Data nous offre aujourd'hui un important volume de données qui doit être analysé, parfois dans le but d'une prise de décision rapide, et de bénéficier de ces données dans le but d'améliorer un domaine. On va maintenant présenter la nécessité des technologies big data et leur rôle pour améliorer l'analyse des données et les défis auxquels elle répond.

II. Big Data

Tout simplement, la grande ère des données est en pleine vigueur aujourd'hui parce que le monde change. Grâce aux progrès des technologies de la communication [33], les gens et les choses sont de plus en plus interconnectés et pas seulement une partie du temps, mais tout le temps. Les gens utilisent de plus en plus les réseaux sociaux, les objets connectés tels que les Smartphones, les véhicules équipés de capteurs de localisation, ça à créer beaucoup de données volumineuses que les outils et les technologies traditionnelles ne peuvent pas traiter et analyser. Big Data offre des solutions, elle peut gérer une très grande quantité de données [34], structurées ou non, sur une variété de terminaux. Dans cette première partie, on va présenter une brève compréhension sur le Big Data et ses avantages.

1. Les bases de données traditionnelles et Big Data:

Les bases de données traditionnelles exigent que les données soient structurées dans un format précis pour qu'elles puissent être traitées normalement. Les personnes et les objets de nos jours génèrent des données plus que jamais en raison du développement des réseaux sociaux [35] et des technologies, on utilise de plus en plus des objets connectés [36] qui génèrent beaucoup de données, ces données doivent être analysées. Mais ces données sont également non structurées et ne peuvent pas être traitées par les bases de données traditionnelles. Nous présenterons ici les limites des bases de données traditionnelles et le besoin de Big Data [37] et de ce qu'elle apporte. La taille des données a énormément augmenté pour atteindre la plage de péta-octets et plus (un péta-octet = 1 024 téraoctets). Les bases de données traditionnelles [38] ont du mal à gérer de tels volumes de données. Pour résoudre ce problème, ils ont ajouté plus d'unités de traitement centrales ou de mémoire au système de gestion de la base de données afin de s'agrandir verticalement. La majorité des données sont présentées dans un format semi-structuré ou non structuré, issu des médias sociaux, de l'audio, de la vidéo, des textes et des courriels. Cependant, le deuxième problème est lié aux données non structurées.

Ils sont conçus et structurés pour s'adapter à la structure des données et aux données. En outre, les données volumineuses sont générées à une vitesse très élevée. Les bases de données traditionnelles manquent de vélocité élevée car ils sont conçus pour des données stables plutôt que pour une croissance rapide. Même si ils sont utilisés pour gérer et stocker des données volumineuses, il s'avérera très coûteux. De ce fait, l'incapacité des bases de données relationnelles à gérer les données volumineuses [39] a conduit à l'émergence de nouvelles technologies Big Data.

- **Relation de données**

Big Data contient une quantité énorme de données, ce qui rend la relation de base de données difficile à comprendre [40]. Cela affecte les éléments de données, ce qui rend également difficile la compréhension. Cependant, avec les données traditionnelles, il est facile de parcourir toutes les données et informations sans trop de difficultés. Cela aide également à comprendre facilement la relation entre les données et les éléments de données.

- **L'exactitude des données et la confidentialité**

Avec les données traditionnelles, il est difficile de maintenir la précision et la confidentialité [41] car la qualité des données est élevée et le stockage d'une telle quantité de données coûte cher. Cela affecte l'analyse des données, ce qui diminue également le résultat final en termes de précision et de confidentialité. Cependant, le Big data rend ce travail beaucoup plus simple et sans problème par rapport aux données traditionnelles. En outre, il fournit la haute précision et rend les résultats plus précis.

- **Différents types de données**

La base de données traditionnelle est principalement destinée à la structure rituelle, c'est-à-dire au stockage de données dans différents formats ou mélangés dans un fichier. Pour toute organisation, il est important de comprendre chaque problème et d'obtenir le meilleur aperçu des données pour mieux connaître la structure. Cependant, ce n'est pas possible avec les données traditionnelles. Big Data fournit de meilleurs détails et la structure des Big Data offre un meilleur accès aux données, ce qui contribue à améliorer le travail.

- **Taille de stockage de données**

La taille du stockage dans les données est importante. Dans les bases de données traditionnelles, il est impossible de stocker une grande quantité de données [42], la taille de stockage des données est limitée. Cependant, avec le Big Data, on peut facilement stocker

d’énormes données volumineuses. La base de données traditionnelle peut enregistrer des données en nombre de giga-octets ou en téraoctets. D’un autre côté, les Big Data peuvent économiser des centaines de téraoctets, de péta-octets et même plus. Cela permet également d’économiser le montant dépensé dans la base de données traditionnelle pour le stockage. En stockant des données volumineuses, on réduit les ressources et l'argent supplémentaires.

Alors qu'est-ce qui rend le Big Data meilleur et qu'est-ce qui le définit exactement?

Big Data et les données traditionnelles ne se différencient pas uniquement par la taille des données. Mais la façon dont les données peuvent être utilisée. Il existe différentes fonctionnalités qui rendent le Big Data meilleur en termes de traitement des données et de performances [43]. En clarifiant le concept, voici les principales fonctionnalités que le Big data peut fournir.

- **Mieux analyser**

Dans les bases de données traditionnelles, les données mettaient longtemps à être analysé correctement et à obtenir le résultat final, la qualité des données se dégrade. Mais avec Big Data, la performance et la méthode d'analyse deviennent avancées et facilement accessibles sans affecter la qualité.

- **Plus flexible**

Les Big Data sont flexibles [44] et faciles à manipuler sans aucune perturbation. Au paravant, les données ne pouvaient être sauvegardées que dans un type spécifique de structures de données. Cependant, ces jours-ci, un autre type de format est introduit. Le Big Data offre un meilleur accès aux données et l’organisation peut être modélisée en fonction des besoins.

- **Maintenir la qualité**

Afin que les données puissent être analysées rapidement et facilement, les Big Data n’affectent pas la qualité du travail. Pour toute organisation, la gestion de la qualité de ses données est un travail important à faire. Mais en raison du taux croissant de données, il est difficile de maintenir la norme. C’est pourquoi les Big Data facilitent le processus sans dégrader la qualité du contenu ainsi que des données.

- **Apprentissage automatique**

Dans le monde des données, l’apprentissage automatique prend de plus en plus d’importance. Les analyses Big data aident également à l'apprentissage automatique, alors que dans une base de données traditionnelle, l'utilisation de l’apprentissage automatique est rare [45].

- **Un moyen simple de stocker**

Avec le stockage traditionnel, les données sont utilisées pour stocker différents types de disques et de lecteurs. Aujourd'hui, cela peut être facilement fait à l'aide d'un logiciel qui rend ce travail indispensable. Cependant, il est difficile de stocker toutes sortes de données sur la plate-forme moderne [46], mais elles offrent alors l'option de transfert rapide.

2. L'explosion des données

Certes, le volume des données [47] produites dans les systèmes informatiques n'a jamais cessé de croitre. Mais pourquoi n'a-t-on commencé à en parler que lors de la dernière décennie.

Les objets intelligents [48] envahissent aujourd'hui, et envahiront d'avantage prochainement la vie quotidienne des entreprises et des particuliers. L'internet des objets est une notion complexe qui consiste dans le fait que tous les objets peuvent être connectés à Internet, et sont donc capables d'émettre de l'information et éventuellement de recevoir des commandes. L'Internet des objets (IdO) propose de créer une continuité entre le monde réel et le monde numérique : il donne une existence aux objets physiques dans le monde numérique. Des exemples concrets d'objets intelligents sont les smart phones, les tablettes, les montres connectées, les voitures connectées, les compteurs électriques intelligents, les chaussures connectées, etc. L'aspect extrêmement disruptif et innovateur de ces objets entraine une certaine réticence de la part des utilisateurs. Mais, tout comme les smart phones et les tablettes, leur usage continuera de se généraliser. Ces objets intelligents sont équipés de capteurs qui transmettent des données de manière régulière aux fabricants et aux fournisseurs, leur permettant de constituer une base de l'utilisation faite de leur produit. Ces données qui ne cessent d'exploser au fur et à mesure que l'usage de ces objets se banalise, et présentent un potentiel remarquable aussi bien pour les fournisseurs que pour les tiers.

L'explosion des données [49] s'accompagne de l'apparition de données dites « non-structurées », qu'il s'agisse de textes, de photos, de sons ou de vidéos. L'explosion du trafic sur Internet est fortement liée au développement de la vidéo, sous toutes ses formes (de la distraction à la communication, de YouTube à Facebook). Ces nouvelles formes de données non structurées posent des défis en termes de collecte, de stockage, d'indexation, de recherche et de manipulation pour les systèmes d'information classiques. Même si la majorité des

traitements [50] Big Data consiste à effectuer des opérations simples sur des très grands volumes de données, il faut pouvoir le faire très rapidement.

Plus de 2,5 exaoctets (2,5 milliards de giga-octets) de données sont générés [51] tous les jours.

Cet énorme volume de données provient de diverses sources :

- Il y a environ 5 milliards de téléphones mobiles dans le monde (dont 1,75 milliards Smartphones) ;
- Les utilisateurs YouTube téléchargent plus de 48 heures de vidéo par minute ;
- Les grands réseaux sociaux tels que Twitter et Facebook captent plus de 10 To de données par jour ;
- Il y a plus de 30 millions de capteurs en réseau dans le monde et chacun d'eux envoie des données en continu.

La quantité des données qui circule dans les systèmes d'informations (SI) est aujourd'hui considérable et ne cessera d'augmenter dans les années à venir. Le moteur principal de cette croissance est l'envahissement du digital de la vie humaine. Les technologies Big data sont là pour permettre à gérer les quantités importantes de données qui transitent dans son écosystème, et aussi d'explorer de nouvelles sources de données, dont celles caractérisées par la vélocité des flux qu'elle génère.

3. Qu'est-ce que le Big Data?

Si nous prenons une définition plus simple, nous pouvons dire que c'est un énorme volume de données qui ne peuvent pas être stockées et traitées avec l'approche et les outils traditionnels. Comme ces données peuvent contenir des informations précieuses, elles doivent être traitées dans un court laps de temps. Ces informations précieuses peuvent être utilisées pour effectuer des analyses prédictives, à des fins de marketing et à d'autres fins. Si nous utilisons l'approche traditionnelle, nous ne pourrons pas accomplir cette tâche dans les délais impartis, car la capacité de stockage et de traitement ne serait pas suffisante pour ces types de tâches. C'était une définition plus simple pour comprendre le concept de Big Data [52].

Le volume, la variété et la rapidité des informations disponibles [53] augmentent de manière exponentielle la complexité des informations à gérer. Le volume auquel les nouvelles données sont générées est stupéfiant. Nous vivons dans une époque où la quantité de données que nous attendons à générer dans le monde est mesurée en exaoctets et zettaoctets. De plus, la variété

de sources et de types de données générées s'agrandit aussi rapidement que de nouvelles technologies peuvent être créée. Les mesures de performance des moniteurs embarqués, des mesures de rendement au sol de la fabrication, de nombreux types d'appareils de soins de santé et du nombre croissant d'appareils d'énergie génèrent toutes des données.

Plus important encore, ils génèrent des données à un rythme rapide. La vitesse de génération, d'acquisition, de traitement et de sortie des données augmente de manière exponentielle à mesure que le nombre de sources et la variété des formats augmentent avec le temps. La révolution du Big Data a entraîné de profonds changements dans la capacité de traiter des événements complexes, de capturer des données transactionnelles en ligne, de développer des produits et services pour l'informatique mobile et de traiter de nombreux événements de données volumineux en temps quasi réel.

Nous présenterons ici quelques caractéristiques définissant le Big Data et ce qu'il apporte, ainsi que ce qui le distingue réellement des bases de données traditionnelles.

Voici quelques fonctionnalités qui définissent le Big Data:

a) Processus d'analyse hautement évolutifs

Les plates-formes Big Data telles que Hadoop et Spark [54] sont devenues populaires grâce en grande partie à leur capacité à évoluer. La quantité de données qu'ils peuvent analyser sans dégradation des performances est pratiquement illimitée. C'est ce qui distingue ces outils des méthodes classiques d'investigation des données, telles que les requêtes SQL de base. Ce dernier n'est pas évolutif, sauf s'il est intégré dans un cadre analytique plus vaste.

b) La flexibilité

Les données volumineuses sont des données flexibles [55]. Alors que dans le passé, toutes les données pouvaient avoir été stockées dans un type de base de données spécifique utilisant des structures de données cohérentes, les jeux de données actuels se présentent sous de nombreuses formes. Les stratégies d'analyse efficaces sont conçues pour être extrêmement flexibles et pour traiter tout type de données qui leur sont transmises. La transformation rapide des données est un élément essentiel du Big Data, tout comme la capacité de travailler avec des données non structurées.

c) Résultats en temps réel

Traditionnellement, les entreprises pouvaient se permettre d'attendre les résultats de l'analyse de données. Dans le monde du Big Data, toutefois, maximiser la valeur signifie obtenir des

informations en temps réel [56]. Après tout, lorsque nous utilisons le Big Data pour des tâches telles que la détection de fraude, les résultats reçus après coup n'ont que peu d'intérêt.

d) Applications d'apprentissage automatique

L'apprentissage automatique [57] n'est pas le seul moyen de tirer parti du Big Data. Il s'agit toutefois d'une application de plus en plus importante dans le monde du Big Data. Les cas d'utilisation de l'apprentissage automatique distinguent le Big Data des données traditionnelles, qui étaient très rarement utilisées pour l'automatisation de l'apprentissage automatique.

e) Systèmes de stockage évolutifs

Traditionnellement, les données étaient stockées sur des lecteurs de bande et de disque conventionnels. Aujourd'hui, les Big Data reposent souvent sur des systèmes de stockage évolutifs [58] définis par logiciel qui extraient les données du matériel de stockage sous-jacent. Bien entendu, tous les Big Data ne sont pas stockés sur des plates-formes de stockage modernes, ce qui explique pourquoi la possibilité de transférer rapidement des données entre un stockage traditionnel et un stockage de nouvelle génération reste importante pour les applications Big Data.

f) Qualité des données

La qualité des données [59] est importante dans n'importe quel contexte. Cependant, avec la complexité croissante du Big Data, il est devenu de plus en plus important d'assurer la qualité des données dans des ensembles de données et des opérations d'analyse complexes. Faire attention à la qualité des données est une caractéristique essentielle de tout flux de travail Big Data efficace.

Les données volumineuses ne représentent pas seulement la taille des données en termes de quantité de données. Il s'agit d'autres caractéristiques d'une source de données volumineuses. Ces aspects comprennent non seulement une augmentation du volume, mais a augmenté la vitesse et une plus grande variété. Bien entendu, ces facteurs entraînent également une complexité supplémentaire. Cela signifie que nous n'obtenons pas seulement beaucoup de données lorsque nous travaillons avec le Big Data. Elle vient également rapidement, dans des formats complexes, et provient de diverses sources. Dans la prochaine partie, nous décrirons les principales caractéristiques du Big Data.

4. Caractéristiques de Big data:

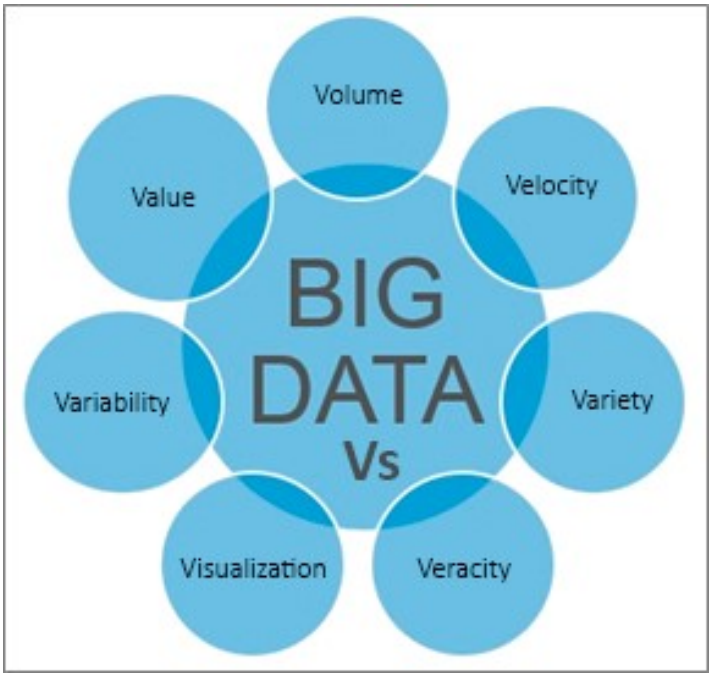

Figure 3. Les 7 V du Big Data

4.1. Volume:

Avec l'évolution des médias sociaux et des autres ressources Internet [60], les gens publient et téléchargent beaucoup de données comme des vidéos, photos, tweets, etc. Un téléphone portable contient de nombreux capteurs, qui génèrent des données pour chaque événement, qui nécessitent la collecte et l'analyse. Lorsque nous parlons de volume dans un contexte de données volumineuses, il s'agit d'une quantité de données énorme en ce qui concerne le système de traitement qui ne peut pas être collectée, stockée et traitée avec les approches traditionnelles. Ce sont des données qui sont déjà collectées et des données en continu qui sont générées en continu.

Nous prenons par exemple Facebook. Ils ont deux milliards d'utilisateurs, en 2016, actifs qui utilisent en permanence ce site pour partager leurs statuts, photos, vidéos, commenter leurs publications et bien d'autres activités. Selon les statistiques fournies par Facebook, 600 To de données sont stockées quotidiennement dans la base de données de Facebook.

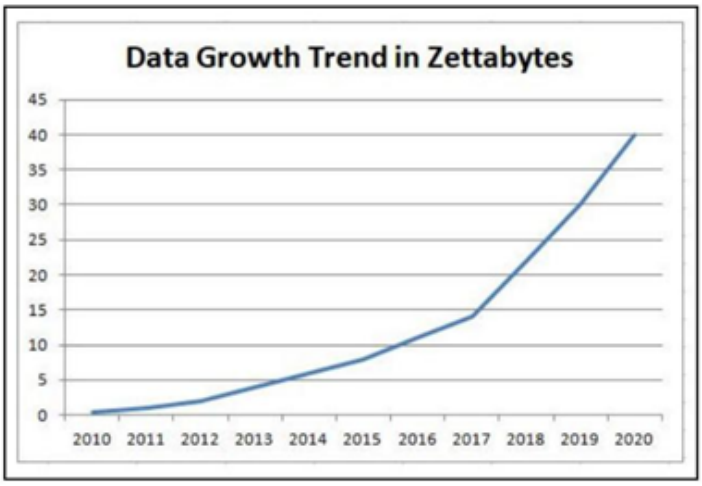

Figure 4. La croissance des données [61]

Une telle quantité de données était considérée comme un problème car le coût de stockage était très élevé. Mais maintenant, comme les coûts de stockage diminuent [62], ce n'est plus un problème. En outre, des solutions telles que Hadoop et différents algorithmes facilitant l'ingestion et le traitement de cette masse de données la rendent même pleine de ressources.

4.2. Vélocité:

La vélocité est la vitesse à laquelle les données sont générées, ou la rapidité avec laquelle elles arrivent. La quantité de données que Facebook, YouTube ou tout autre réseau social reçoit chaque jour. Ils doivent être stockés, traités d'une manière ou d'une autre, pour pouvoir être récupérés. Si nous prenons l'exemple des tendances des médias sociaux, plus de données signifie plus d'informations révélatrices sur des groupes de personnes sur différents territoires.

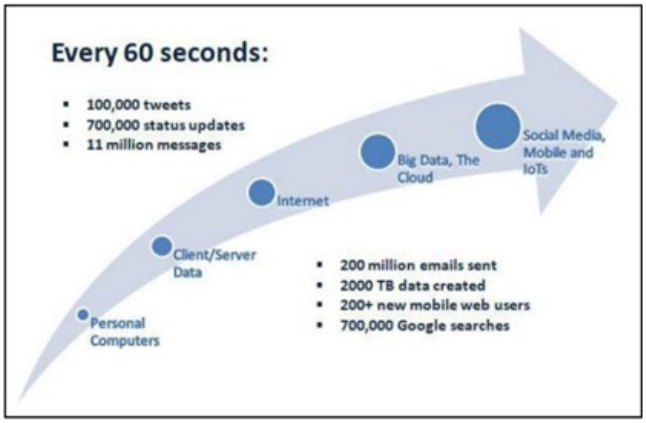

Figure 5. La vitesse à laquelle les données sont générées [63]

Le graphe ci-dessus indique la quantité de données générées par les utilisateurs sur les sites Web de réseaux sociaux populaires. Cela nous donne une image de la fréquence à laquelle les données sont générées en fonction des activités de ces utilisateurs. Ceci est juste un aperçu de ce qui se passe. Une autre dimension de la vitesse est la période de temps pendant laquelle les données ont un sens et sont précieuses.

4.3. Variété

Dans cette section, nous étudions la classification des données. Il peut s'agir de données structurées, semi-structurées ou non structurées. Les données structurées sont préférées pour les informations ayant un schéma prédéfini ou un modèle de données avec des colonnes prédéfinies, des types de données, etc. Les données semi-structurées sont une forme de données structurées qui n'obéissent pas à la structure formelle des modèles de données associés aux bases de données relationnelles ou à d'autres formes de tables de données, mais qui contiennent néanmoins des balises ou d'autres marqueurs permettant de séparer les éléments sémantiques et d'imposer des hiérarchies d'enregistrements et de champs dans les données. Alors que les données non structurées ne présentent aucune de ces caractéristiques. Celles-ci comprennent une longue liste de données telles que des documents, des courriels, des SMS, des vidéos, des images fixes, du son, des graphiques, ainsi que la sortie de tout type de données générées par une machine à partir de capteurs, de périphériques, de journaux de machines et de téléphones portables GPS signaux, et plus encore.

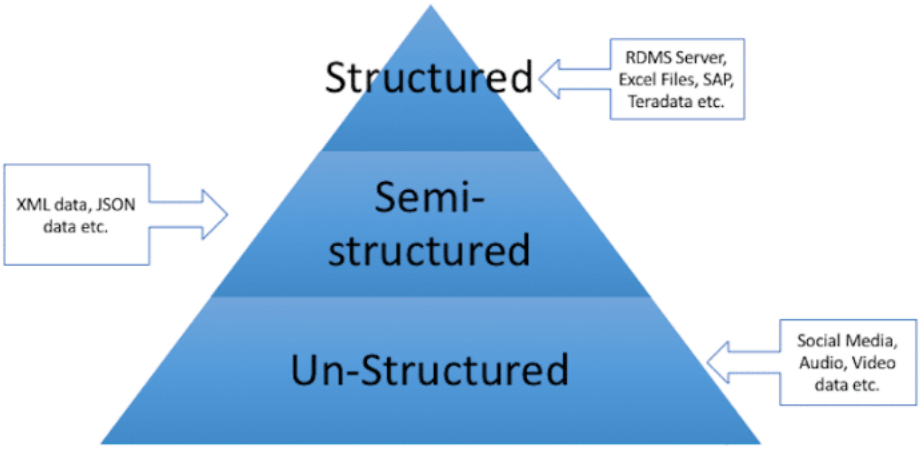

Figure 6. La classification des types de données [64]

Prenons un exemple, 30 milliards de contenus sont partagés sur Facebook chaque mois. 400 millions de Tweets sont envoyés par jour. Chaque mois, 4 milliards d'heures de vidéos sont visionnées sur YouTube. Ce sont tous des exemples de données non structurées générées qui doivent être traitées, soit pour une meilleure expérience utilisateur, soit pour générer des revenus pour les entreprises elles-mêmes.

4.4. Véracité

Ce vecteur traite de l'incertitude des données [65]. Cela peut être dû à la mauvaise qualité des données ou au bruit dans les données. C'est le comportement humain que nous ne faisons pas confiance aux informations fournies. Nous pouvons considérer d'une manière que la vitesse et la variété dépendent des données épurées avant l'analyse et la prise de décision, alors que la véracité est le contraire de ces caractéristiques, car elle découle de l'incertitude des données.
Le principal défi est que nous n'avons pas le temps de nettoyer les données en continu ou à grande vitesse pour éliminer les incertitudes. Des données telles que des données d'événements sont générées par des machines et des capteurs et si nous attendons de les nettoyer et de les traiter en premier, ces données risquent de perdre de la valeur. La véracité concerne l'incertitude et la confiance que nous avons dans les données, mais lorsque nous l'utilisons dans le contexte du Big Data, il se peut que nous devions redéfinir les données de confiance avec une définition différente.

4.5. Valeur :

Big Data analyse d'énormes ensembles de données pour révéler les tendances sous-jacentes et aider les décideurs à prévoir les problèmes potentiels. Connaître la performance future des équipements utilisés en cours de fonctionnement et identifier les défaillances avant qu'elles ne surviennent peut donner à la société un avantage concurrentiel et une valeur ajoutée.
Les données seules ne valent rien. Les données valent quand elles sont intégrées et analysées dans de nombreux points de vue différents et c'est ce qui génère de la valeur en donnant la capacité de prendre des décisions efficaces, efficientes et précises pour proposer des opportunités d'améliorations.

4.6. Variabilité

La variabilité est différente de la variété.

La variabilité vise principalement à bien comprendre et interpréter les significations correctes des données brutes qui dépendent de leur contexte.

4.7. Visualisation

La visualisation fait référence à la manière dont les données sont présentées à la direction pour sa prise de décision. Les données peuvent être présentées de différentes manières, telles que de longs fichiers Excel avec des lignes et des colonnes de données, des documents Word, des graphiques, etc.

Mais, l'utilisation de tableaux et de graphiques pour visualiser de grandes quantités de données complexes est beaucoup plus efficace pour transmettre une signification que les feuilles de calcul et les rapports chargés de nombres et de formules.

Quel que soit le format, les données doivent être facilement lisibles, compréhensibles et accessibles. C'est pourquoi la visualisation des données est importante.

5. Classification des types de données Big Data:

Les données peuvent en gros être classées comme structurées, non structurées ou semi-structurées. Bien que ces distinctions aient toujours existé, la classification des données dans ces catégories est devenue plus importante avec l'avènement du Big Data.

5.1. Structuré

Les données structurées [66], comme leur nom l'indique, désignent des ensembles de données ayant une structure organisationnelle définie, telle que des fichiers Microsoft Excel ou CSV. En termes de base de données, les données doivent pouvoir être représentées à l'aide d'un schéma.

Les bases de données commerciales telles que Teradata, Greenplum, ainsi que Redis, Cassandra et Hive dans le domaine open source sont des exemples de technologies permettant de gérer et d'interroger des données structurées.

5.2. Non structuré

Les données non structurées [67] sont constituées de tout jeu de données ne possédant pas de schéma d'organisation prédéfini. Les mots parlés, la musique, les vidéos et même les livres seraient considérés comme non structurés. Cela ne signifie nullement que le contenu n'a pas d'organisation. En effet, un livre par exemple a une table des matières, des chapitres, des sous-chapitres, il suit une organisation définie. Cependant, il serait vain de représenter chaque mot et chaque phrase comme faisant partie d'un ensemble strict de règles. Une phrase peut être composée de mots, de chiffres, de signes de ponctuation, etc. Et ne possède pas de type de données prédéfini, contrairement aux feuilles de calcul. Pour être structuré, le livre devrait avoir un ensemble exact de caractéristiques dans chaque phrase, ce qui serait déraisonnable et peu pratique. Les données des médias sociaux, telles que les publications sur Twitter, les messages de Facebook et les photos sur les médias sociaux, sont des exemples de données non structurées.

5.3. Semi-structuré

Les données semi-structurées [68] font référence à des données contenant à la fois les éléments d'un schéma d'organisation et des aspects arbitraires. Un agenda téléphonique personnel avec des colonnes pour le nom, l'adresse, le numéro de téléphone et les notes pourrait être considéré comme un ensemble de données semi-structurées. L'utilisateur peut ne pas connaître les adresses de toutes les personnes et, par conséquent, certaines entrées peuvent ne comporter qu'un numéro de téléphone et inversement. De même, la colonne de notes peut contenir des informations descriptives supplémentaires. C'est un champ arbitraire qui permet à l'utilisateur d'ajouter des informations complémentaires. Les colonnes pour nom, adresse et numéro de téléphone peuvent donc être considérées comme structurées en ce sens qu'elles peuvent être présentées sous forme de tableau, tandis que la section des notes n'est pas structurée en ce sens qu'elle peut contenir un ensemble arbitraire d'informations descriptives qui ne peuvent pas être traitées et représentés dans les autres colonnes du journal. Les données semi-structurées sont généralement représentées par des formats, tels que JSON, pouvant encapsuler des associations structurées, ainsi que des associations sans schéma et arbitraires, généralement à l'aide de paires clé-valeur. Un exemple plus courant pourrait être les messages électroniques, qui comportent à la fois une partie structurée, telle que le nom de l'expéditeur, l'heure de réception du message, et bientôt commune à tous les messages électroniques et une

partie non structurée représentée par le corps ou le contenu de l'email. Les plates-formes telles que MongoDB et CouchDB [69] sont généralement utilisées pour stocker et interroger des ensembles de données semi-structurés.

6. Sources des données Big Data:

La technologie actuelle nous permet de collecter des données à un rythme effréné, à la fois en termes de volume et de variété. Différentes sources génèrent des données, mais dans le contexte du Big Data, les sources principales [70] sont les suivantes:

- Réseaux sociaux: on peut dire que les réseaux sociaux qui se sont multipliés au cours des 5 à 10 dernières années constituent la principale source de toutes les méga-données que nous connaissons aujourd'hui. Il s'agit dans l'ensemble de données non structurées représentées par des millions d'affichages dans les médias sociaux et d'autres données générées seconde par seconde par le biais d'interactions entre les utilisateurs sur le Web dans le monde entier. L'augmentation de l'accès à Internet à travers le monde est un acte auto-réalisateur pour la croissance des données dans les réseaux sociaux.
- Médias: en grande partie à cause de la croissance des réseaux sociaux, les médias représentent des millions, sinon des milliards, de téléchargements audio et visuels quotidiens. Les vidéos, la musique et les images téléchargées sur les médias sociaux sont d'excellents exemples de médias dont le volume ne cesse de croître de manière effrénée.
- Entrepôts de données (Data Warehouse): les entreprises investissent depuis longtemps dans des installations de stockage de données spécialisées, communément appelées entrepôts de données. Un DW est essentiellement un ensemble de données historiques que les entreprises souhaitent conserver et cataloguer pour les retrouver facilement, que ce soit pour un usage interne ou réglementaire. Au fur et à mesure que les industries se tournent vers la pratique du stockage de données sur des plates-formes telles que Hadoop et NoSQL, de plus en plus d'entreprises transfert des données de leurs entrepôts de données préexistants vers certaines des technologies les plus récentes. Les e-mails d'entreprise, les enregistrements comptables, les bases de données et les documents internes sont quelques exemples de données DW qui sont

maintenant déchargées sur les technologies Hadoop ou d'autres technologies Big Data qui exploitent plusieurs nœuds pour fournir une plate-forme hautement disponible et tolérante aux pannes.

- Capteurs: un phénomène plus récent dans l'espace des méga-données a été la collecte de données à partir de dispositifs de capteurs. Si les capteurs ont toujours existé et que des industries telles que le pétrole et le gaz utilisent des capteurs de forage pour mesurer sur des plates-formes pétrolières depuis de nombreuses décennies, l'avènement des dispositifs connectés, également connus sous le nom d'Internet of Things, tels que Fitbit et Apple Watch, signifie qu'une personne pourrait transmettre des données au même rythme que quelques plates-formes pétrolières il y a seulement 10 ans.

Les appareils portables peuvent collecter des centaines de mesures d'un individu à un moment donné. Bien que le secteur ne soit pas encore un gros problème de données, l'industrie évoluant sans cesse, les données relatives aux capteurs vont probablement ressembler davantage au type de données spontanées générées sur le Web par le biais d'activités de réseau social.

7. L'utilisation des techniques de data mining et big data

La numérisation croissante de nos activités, la capacité sans cesse accrue à stocker des données numériques, l'accumulation d'informations en tous genres qui en découle, génère un nouveau secteur d'activité qui a pour objet l'analyse de ces grandes quantités de données. Sont alors apparues de nouvelles approches, de nouvelles méthodes, de nouveaux savoirs et de nouvelles manières de penser et de travailler. Ainsi, cette très grande quantité de données et son traitement sous-tendent de profonds bouleversements, qui touchent à l'économie, au marketing, mais aussi à la recherche et aux savoirs.

Aujourd'hui, de plus en plus de données sont stockées sur les serveurs par les entreprises : les données sur la production, les ventes, les données clients, les données comptables, etc. En fait, en 2016, 90 % des données mondiales ont été générées lors de seules années 2014 et 2015. L'utilisation croissante des réseaux sociaux et des objets connectés (IoT) participent activement à cette collecte massive des données sur les comportements des utilisateurs.

Parce que les données en disent beaucoup sur les préférences des clients, elles constituent des enjeux commerciaux et marketing pour l'entreprise. La data mining et les algorithmes de machine learning permettent d'extraire et d'analyser les données pour en tirer des informations pertinentes simplement, les outils et les algorithmes de data mining traitent l'immense quantité de données pour en faire ressortir des tendances, des modèles, des corrélations.

Les données volumineuses et l'exploration de données sont deux choses différentes. Les deux concernent l'utilisation de grands ensembles de données pour gérer la collecte ou la création de rapports destinés aux entreprises ou à d'autres destinataires. Cependant, les deux termes sont utilisés pour deux éléments différents de ce type d'opération.

Big data est un terme pour un grand ensemble de données. Les ensembles de données volumineuses sont ceux qui dépassent le type simple de bases de données et d'architectures de traitement de données qui étaient utilisées dans les temps anciens, lorsque les données massives étaient plus coûteuses et moins réalisables. Par exemple, des ensembles de données trop volumineux pour être facilement manipulés dans une feuille de calcul Microsoft Excel peuvent être appelés ensembles de données volumineuses.

L'exploration de données fait référence à l'activité consistant à parcourir des ensembles de données volumineuses pour rechercher des informations pertinentes. L'idée est que les entreprises collectent d'énormes quantités de données pouvant être homogènes ou collectées automatiquement. Les décideurs doivent avoir accès à des données plus petites et plus spécifiques provenant de ces grands ensembles. Ils utilisent les algorithmes de machine learning pour découvrir les informations qui informeront le leadership et aideront à tracer la voie à suivre pour une entreprise.

Conclusion

Big Data ne consiste pas simplement à stocker les données. Trouver les connaissances cachées à partir des données collectées est en réalité l'objectif principal des technologies Big Data et des analyses. Il est clair que les applications de datamining et des algorithmes de machine learning couvrent un très large spectre. Big Data nous offre aujourd'hui un important volume de données qui doit être analysé, dans le but d'une prise de décision rapide, et de bénéficier de ces données dans le but d'améliorer un domaine. Dans ce chapitre, on a exploré

l’évolution des données et la nécessité des technologies big data et son rôle concernant la gestion et le traitement des données volumineuses. Après on a présenté les techniques et les algorithmes de machine learning utilisés pour extraire la connaissance des données. Dans le chapitre suivant, nous allons présenter le rôle et la nécessité de ces technologies et de ces algorithmes dans le domaine de l’employabilité.

Chapitre 2 : Employabilité et data mining

Introduction

Au fil des ans, l'employabilité a été abordée sous différentes dimensions. Alors que Hillage et Pollard (1998) [71] ont défini l'employabilité comme étant l'acquisition et la conservation d'un travail satisfaisant, Fugate et al. (2004) [72] ont proposé à l'employabilité une nouvelle direction de l'adaptabilité proactive incluant les dimensions de l'identité de carrière, de l'adaptabilité personnelle et des compétences en réseautage social (capital social et humain). Harvey (2001) [73] a affirmé que la plupart des définitions de l'employabilité développaient cinq caractéristiques: le type d'emploi, le moment choisi, les caractéristiques relatives au recrutement, l'acquisition de compétences supplémentaires et les compétences relatives à l'employabilité. Dans ce chapitre, on va présenter une définition globale sur l'employabilité, on va présenter aussi l'état actuel de l'employabilité au Maroc, et on va préciser le rôle du data mining et sa nécessité pour proposer des solutions et améliorer ce domaine.

I. Travaux liés

De nombreux domaines tirent maintenant parti de la puissante utilisation de l'exploration de données pour améliorer un domaine particulier. Nous présentons ci-dessous quelques travaux utilisant des techniques et des algorithmes de machine learning pour la prédiction de l'employabilité.

Parneet Kaur, Manpreet Singh and Gurpreet Singh Josan [74] ont proposé une expérience visant à identifier les élèves avec un lent apprentissage dans le domaine de l'éducation à l'aide de techniques de classification d'exploration de données utilisant des données du monde réel recueillies dans des écoles secondaires dans le but de proposer des solutions et assurer une bonne employabilité pour ces élèves. En utilisant Weka, ils ont appliqué plusieurs algorithmes tels que Multilayer Perception, Naïve Bayes, SMO, J48 et RepTree afin de trouver le meilleur modèle de classificateur. Les résultats ont montré que l'algorithme de perception multicouche est le meilleur classificateur avec une précision de 75% du modèle. À l'avenir, l'intégration des techniques d'exploration de données avec les techniques de SGBD et d'Elearning est fusionnée sur différents jeux de données afin de rechercher la précision et les prédictions des résultats souhaités. De plus, les outils de GED sont faciles à comprendre et interfacés avec diverses techniques. Les éducateurs n'ayant aucune expertise dans l'exploration de données peuvent également s'impliquer dans ces domaines. De nouveaux facteurs peuvent également

être appliqués pour améliorer les performances, l'apprentissage et les capacités de rétention de l'élève.

Tripti Mishra, Dharminder Kumar and Sangeeta Gupta [75] ont mené une enquête pour expliquer que l'enseignement supérieur est devenu un domaine de recherche passionnant et que la prédiction de l'employabilité en utilisant les techniques de data mining est bénéfique pour les établissements. Ensuite, ils ont discuté les travaux effectués dans les deux domaines de prédiction. Ils ont conclus que la prédiction de l'employabilité a beaucoup progressé.

Mohd tajul rizal and yuhanis Yusof [76] ont proposé une expérience de prédiction de l'emploi des diplômés, leur expérience comprenait l'utilisation de cinq techniques d'extraction de données, à savoir Naïve Bayes, la régression logistique, le perceptron multicouche, k-Nearest-Neighbor et l'arbre de décision. Les résultats ont montré que la régression logistique est le meilleur classificateur pour l'ensemble de données. Le modèle de classification produit sera bénéfique pour la direction du collège, car il fournira un aperçu de la qualité des diplômés qu'ils produisent et de la manière dont leur programme peut être amélioré afin de pouvoir les aider à trouver un emploi.

II. L'employabilité au Maroc

Les employeurs attachent beaucoup d'importance à la recherche des candidats dotés des compétences requises pour leurs organisations. Selon le secteur de carrière et la profession dans laquelle le diplômé a choisi de travailler, des compétences, des aptitudes et des connaissances très spécifiques sont nécessaires pour effectuer le travail. L'employabilité concerne les chômeurs qui n'ont pas d'emploi actuellement et qui sont à la recherche d'un emploi selon Mohamed SOUALI [77]. Cependant, l'employabilité reste un concept contesté du point de vue de son utilisation théorique et politique et a été utilisée au cours du siècle dernier comme concept à la fois d'offre et de demande de main-d'œuvre.

Certains chercheurs et décideurs adoptent une approche étroitement définie du côté de l'offre, tandis que d'autres adoptent une perspective plus large sur l'employabilité. La vision plus large met l'accent sur l'employabilité des individus en termes de leur capacité à accéder à un nouvel emploi sur le marché du travail. Par conséquent, l'approche globale intègre des facteurs tels que la recherche d'emploi et les conditions de la demande de travail, qui déterminent si une personne peut réellement trouver ou changer un emploi, ainsi que

l'ensemble des compétences et des attributs liés à l'employabilité qui sont au centre des concepts étroits d'aptitude à l'emploi axés sur l'offre.

Plusieurs définitions de l'employabilité ne semble pas donner l'image exacte puisque l'employabilité inclus d'autres facteurs qui l'influence. Définir les facteurs qui influencent l'employabilité peut donner des solutions et des opportunités aux décideurs pour l'améliorer.

1. La situation de l'employabilité au Maroc en 2016

Dans cette section et en se basant sur les statistiques du Haut-Commissariat au Plan (HCP), nous allons présenter la situation de l'employabilité au Maroc en 2016 [78], ça va nous donner une image bien claire sur la situation de l'employabilité au Maroc.

La situation de l'employabilité en 2016 a continué à être marquée par la persistance à la baisse des taux d'activité et d'emploi. Ainsi, avec 11.747.000 personnes, la population active âgée de 15 ans et plus a baissé, entre les années 2015 et 2016, de 0,7% au niveau national. La population en âge d'activité s'est accrue, quant à elle, de 1,5%. De ce fait, le taux d'activité est passé de 47,4% à 46,4%, marquant une diminution de 1 point. Le taux d'emploi a, quant à lui, reculé de 0,8 point pourcentage au niveau national, passant de 42,8% à 42%.

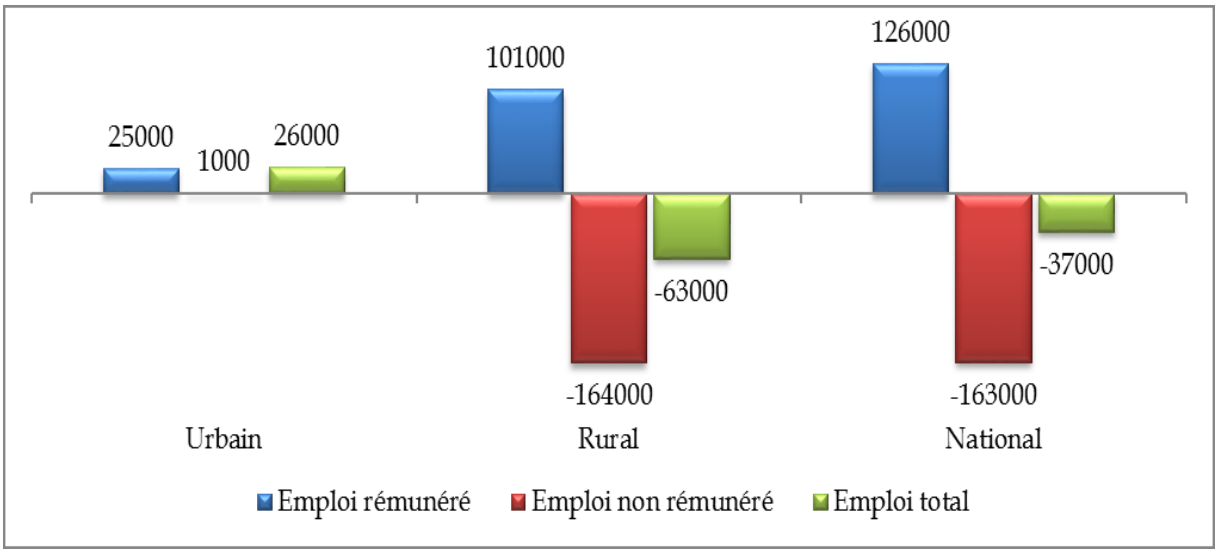

Figure 12. Création nette d'emplois, entre 2015 et 2016, selon le milieu de résidence

Le nombre de chômeurs est passé, entre 2015 et 2016, de 1.148.000 à 1.105.000 personnes. Le taux de chômage est ainsi passé, entre les deux périodes, de 9,7% à 9,4% au niveau

national. Dans ces conditions, la baisse du taux de chômage est l'expression d'un recul du volume de chômage (-3,7%) plus important que celui de l'emploi (-0,4%).

Les baisses les plus importantes du taux de chômage ont été relevées parmi les adultes âgés de 35 à 44 ans (-0,7 point) et les détenteurs d'un diplôme (-0,4 point) qui restent, cependant, respectivement 135.000 et 854.000 chômeurs. Le taux de chômage des jeunes âgés de 15 à 24 ans a enregistré une hausse de 1,7 point pourcentage au niveau national.

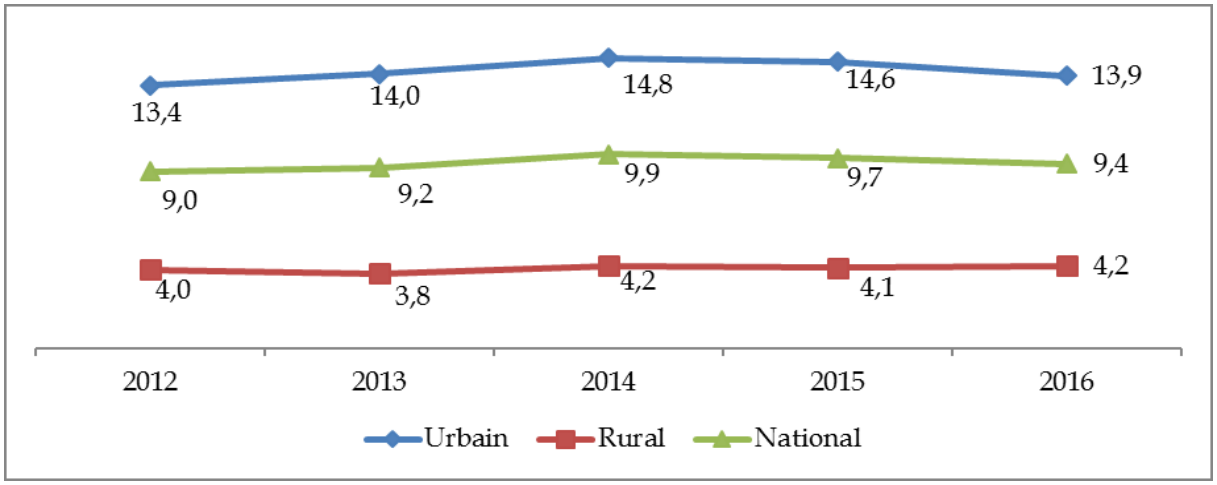

Figure 13. Evolution du taux de chômage par milieu de résidence (en %)

Cinq régions regroupent près des trois quarts de l'emploi.

Cinq régions se sont accaparées près des trois quarts (72,4%) de l'effectif global de l'emploi. Il s'agit de Casablanca-Settat (22,4%), de Marrakech-Safi (13,8%), de Rabat-Salé-Kénitra (13,5%), de Fès-Meknès (11,6%) et de Tanger-Tétouan-Al Hoceima (11,1%). Les autres régions affichent des parts variant entre 0,8% pour Eddakhla-Oued Eddahab et 7,3% pour Béni Mellal-Khénifra.

Sur un autre plan, le poids en termes d'emploi au niveau des régions d'Eddakhla- Oued Eddahab, de Casablanca-Settat, de Tanger-Tetouan-Al Hoceima et de Marrakech–Safi est plus important que le poids démographique, avec un écart relatif plus prononcé pour la région d'Eddakhla-Oued Eddahab (1,6 fois).

Pour le reste des régions, la part dans le volume global de l'emploi reste inférieure à celle dans la population totale. La région de Laâyoune-Sakia El Hamra connaît l'emploi relativement le moins représenté en comparaison avec son poids démographique (0,8 fois).

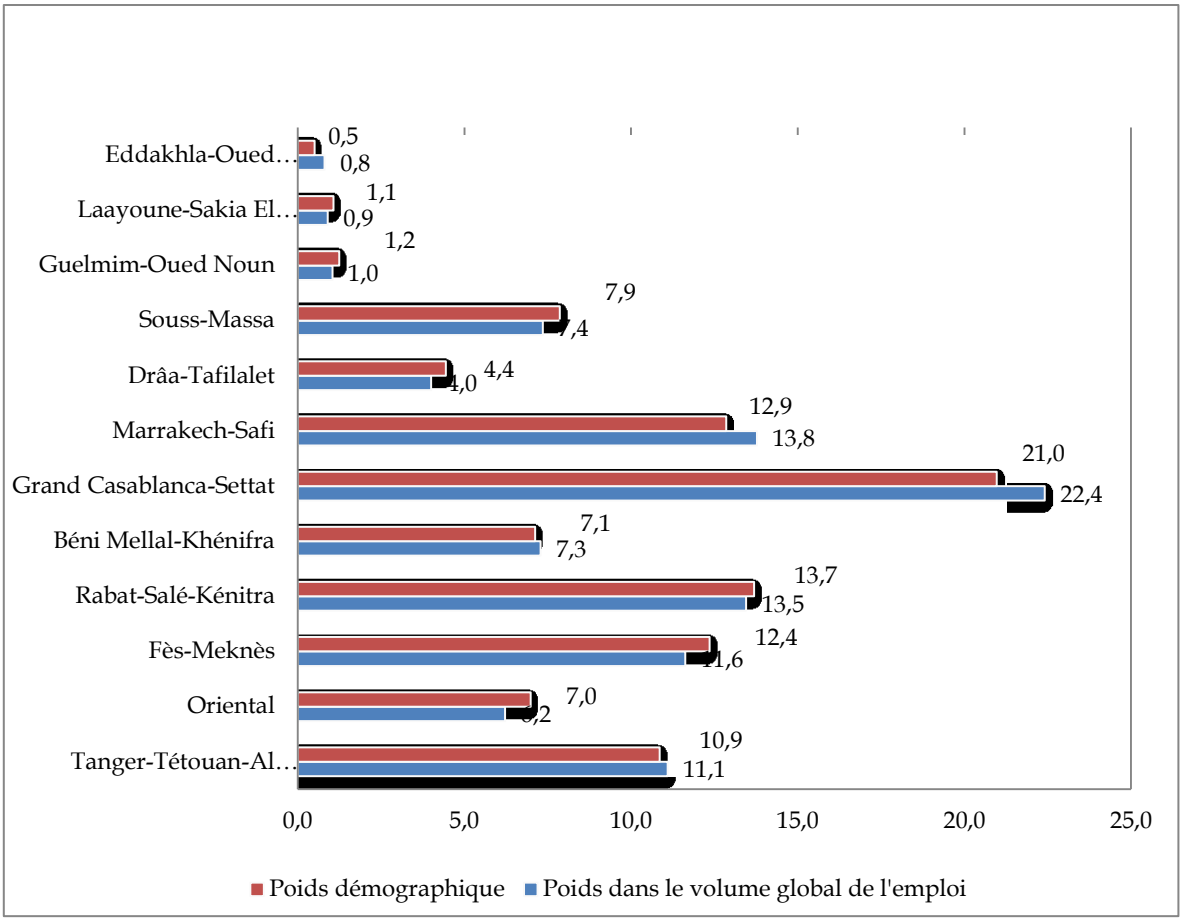

Figure 14. Poids dans le volume global de l'emploi et poids démographique par région (en %)

6 régions sur 12 abritent la quasi-totalité des chômeurs

En 2016, plus de huit chômeurs sur dix (82,8%) sont concentrés dans six régions du Royaume. La région de Casablanca-Settat vient en première position avec 25,1%, suivie de Rabat-Salé-Kénitra (17,5%), l'Oriental (11,3%), Fès-Meknès (10,8%), Marrakech-Safi (9,4%) et, enfin, Tanger-Tétouan-Al Hoceima (8,7%).

D'un autre côté, cinq régions du Royaume se caractérisent par le fait que leur contribution au volume du chômage est plus importante que leur poids démographique en termes de population en âge d'activité. La région de l'Oriental vient en premier lieu avec un écart absolu de 4,3 points, suivie de Casablanca-Settat avec 4,2 points, Rabat-Salé-Kénitra (3,8 points) et dans des degrés moindres, les régions de Guelmim-Oued Noun (0,7 point) et d'Eddakhla-Oued Eddahab (0,1 point).

En revanche, au niveau des autres régions, la contribution au volume du chômage est en deçà du poids démographique. A ce titre, il convient de remarquer que Marrakech-Safi constitue la

région la plus favorable en termes d'accès au marché du travail où la contribution au volume du chômage est inférieure à son poids démographique de 3,4 points.

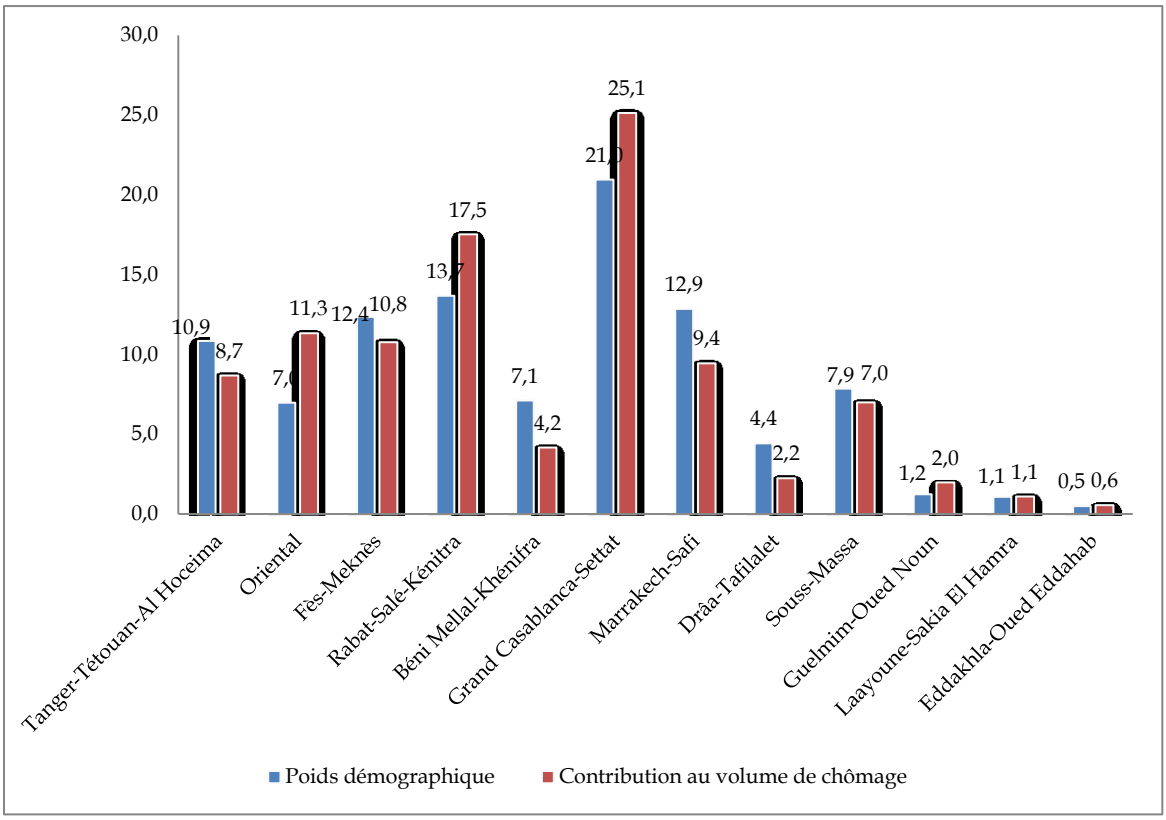

Figure 15. Contribution au volume global de chômage et poids démographique selon les régions.

Le chômage des jeunes est très préoccupant dans toutes les régions.

Au niveau national, les jeunes âgés de 15 à 29 ans représentent la catégorie la plus touchée par le chômage avec un taux de 23,5% pour les hommes et 32,1% pour les femmes. La prévalence du chômage diminue au fur et à mesure que l'âge croît; le taux correspondant s'établit à 7,1% pour les 30-44 ans et 2,5% pour les 45 ans et plus. Au niveau régional, la manifestation du chômage la plus forte est enregistrée parmi les jeunes âgés de 15 à 29 ans dans la région de Guelmim-Oued Noun avec un taux de 43,9%. La situation du chômage des jeunes dans cette région est encore plus alarmante parmi les femmes âgées de 15 à 29 ans avec un taux de 59,4%. Ce dernier taux, avec celui enregistré dans la région de Laayoune-Sakia El Hamra (61,7%), représentent les taux de chômage les plus élevés et expriment le fait que six femmes sur dix âgés de 15 à 29 ans sont en situation de chômage dans ces deux

régions. En revanche, la prévalence la plus faible du chômage des jeunes est observée dans la région d'Eddakhla-Oued Eddahab avec un taux de 9,4%, suivie de Drâa-Tafilalet (14,7%). Concernant la catégorie des 45 ans et plus, le chômage atteint 7,1% dans la région de l'Oriental, ce qui correspond à environ trois fois le niveau national (2,5%). Ce taux devient quasiment nul au niveau des deux régions de Laayoune-Sakia El Hamra et de Tanger-Tetouan-Al Hoceima.

En plus, concernant les diplômés de l'enseignement général, plus le diplôme est élevé plus le taux de chômage est faible.

En revanche, les taux de chômage augmentent avec les diplômes de la formation professionnelle et d'une intensité qui dépasse celle des diplômes d'enseignement général, exception faite du diplôme de technicien et de cadre moyen (Tableau 1, Figure 16 et 17). Mais les taux de chômage de ces deux catégories sont supérieurs à ceux de la population sans diplôme (11,2%).

Tableau 1. Taux de chômage selon les diplômes d'enseignement général et de formation professionnelle

Grand groupe de diplômes	**Taux de chômage (%)**
Sans diplôme	11,2
Diplômes d'enseignement général	**19,7**
Primaire	20,2
Secondaire collégial	22,4
Secondaire qualifiant	19,8
BTS/CPGE	17,1
DEUG	15,1
Licence	18,9
DEA/DES/Master	15,9
Ingénieur/Cadre supérieur	7,7
Doctorat	3,9
Diplômes de formation professionnelle	**25,5**
Technicien spécialisé	26,4
Technicien/Cadre moyen	21,5

Qualification professionnelle	28,4
Spécialisation professionnelle	26,5
Initiation professionnelle	21,3
Total	16,2

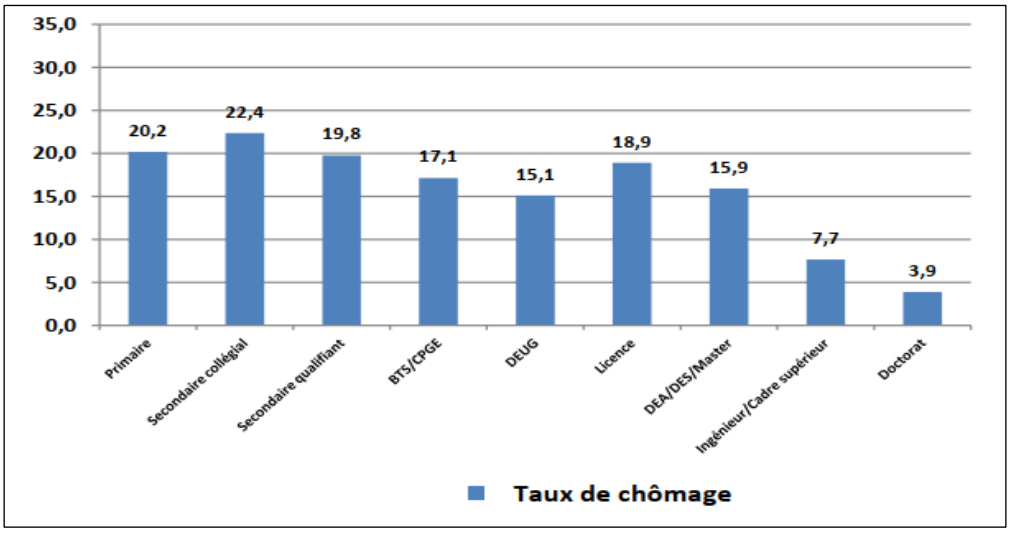

Figure 16. Taux de chômage selon les diplômes d'enseignement général

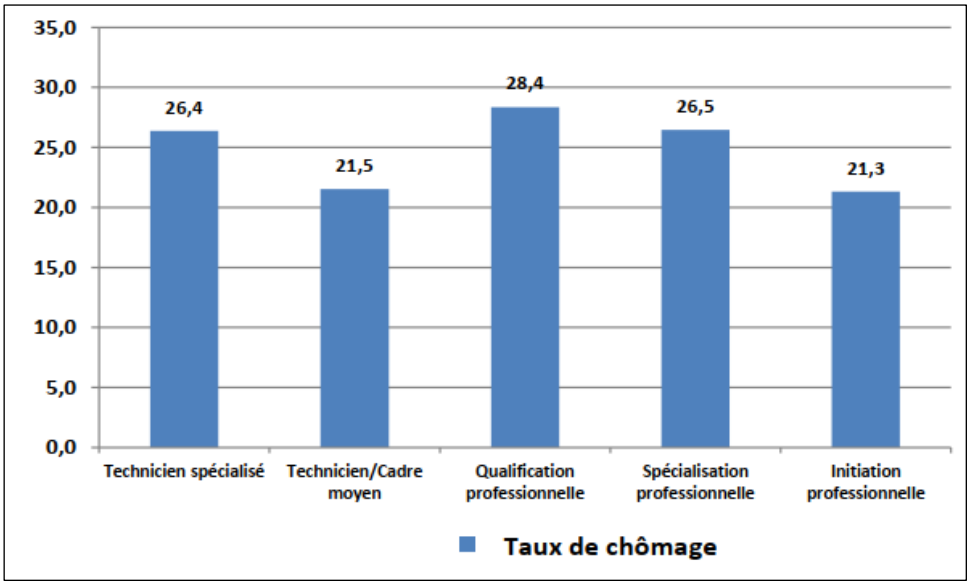

Figure 17. Taux de chômage selon les diplômes de formation professionnelle

2. Présentation des données utilisées

Après avoir présenté les statistiques que le Haut-Commissariat au Plan (HCP) à partager en 2016 sur la situation de l'employabilité au Maroc, maintenant et en se basant sur une enquête sur l'employabilité conduite par l'université Hassan 1er en 2016 en partenariat avec le Bureau

national de l'évaluation (NEO) du Conseil supérieur de l'éducation, de la formation et de la recherche scientifique, on va présenter et visualiser les données collectées qu'on a utilisé pour avoir une idée sur la situation de l'employabilité dans différentes universités au Maroc.

Le graphe ci-dessous visualise la tranche d'âge des diplômés qui n'ont pas trouvé un emploi.

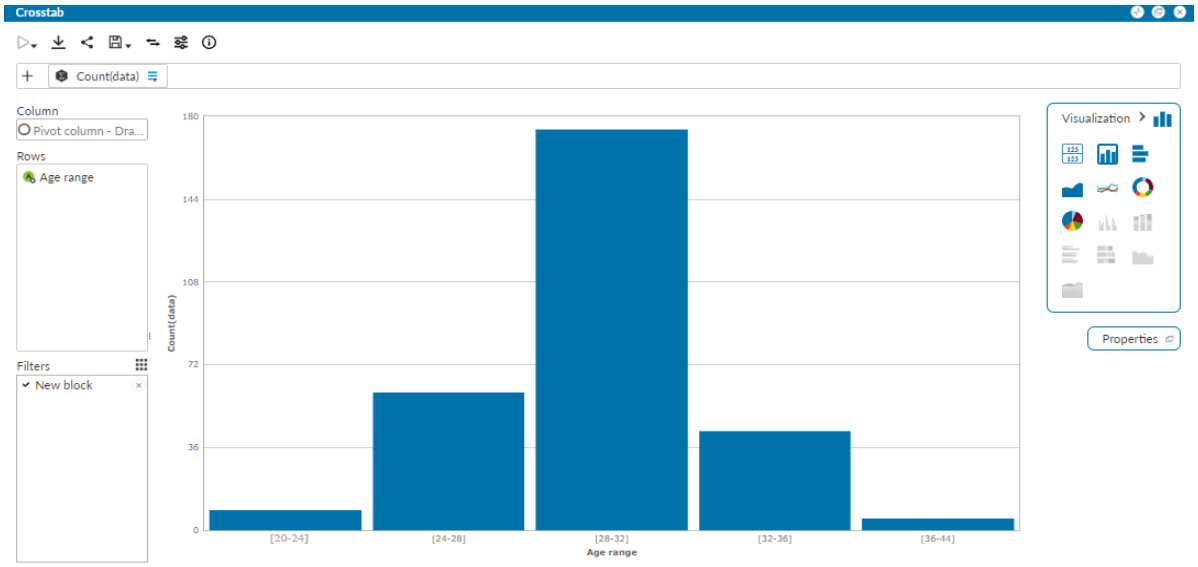

Figure 18. Les diplômés qui n'ont pas trouvé un emploi regroupés par tranche d'âge

Le graphe ci-dessous représente l'effectif des diplômés regroupés par diplôme en se limitant aux diplômés qui ont trouvé un emploi.

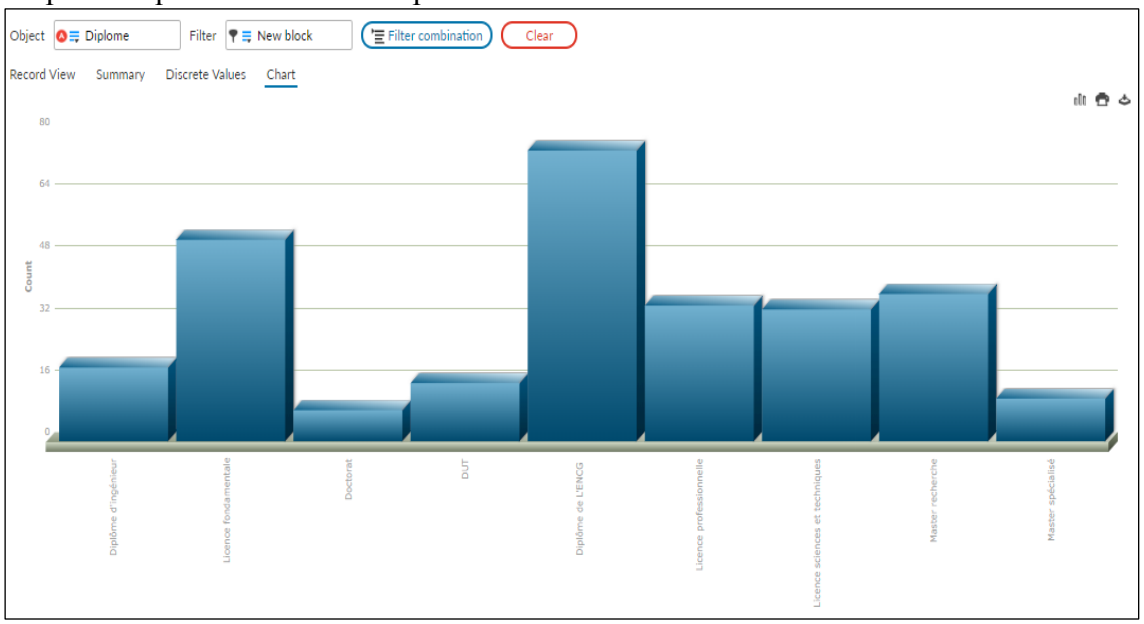

Figure 19. Les diplômés qui ont trouvés un emploi regroupés par « Diplôme »

D’après ce graphe on remarque que les diplômés ayant un diplôme de l’ENCG ont plus de chance de trouver un emploi par rapport aux autres diplômes.

Le graphe ci-dessous représente l’effectif des diplômés qui ont trouvé un emploi ainsi qui ne l’ont pas, regroupé par établissement.

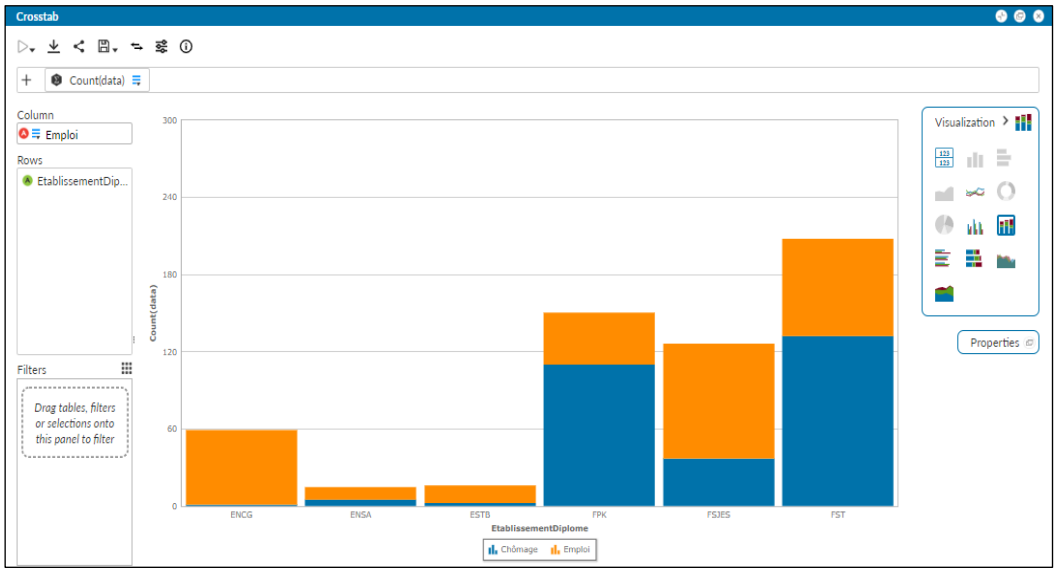

Figure 20. Les diplômés qui ont trouvé un emploi ainsi qui ne l’ont pas, regroupé par établissement

Le graphe ci-dessous représente l’effectif des diplômés qui ont trouvé un emploi ainsi qui ne l’ont pas, regroupé par diplôme.

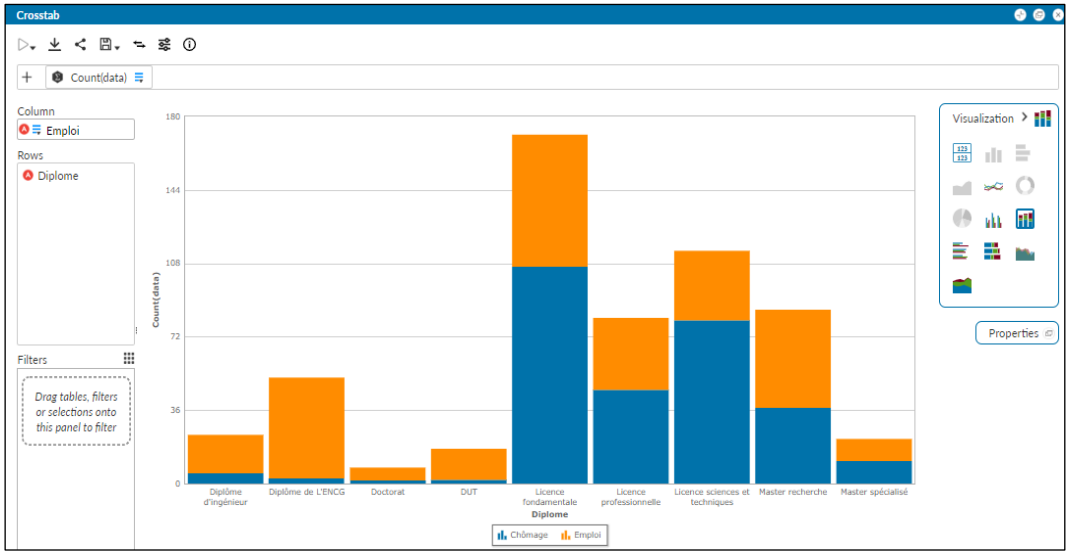

Figure 21. Les diplômés qui ont trouvé un emploi ainsi qui ne l’ont pas, regroupé par diplôme.

Le graphe ci-dessous représente les diplômés qui ont déjà passé un stage ou non, regroupés par établissement, en se limitant aux diplômés qui n'ont pas trouvé un emploi.

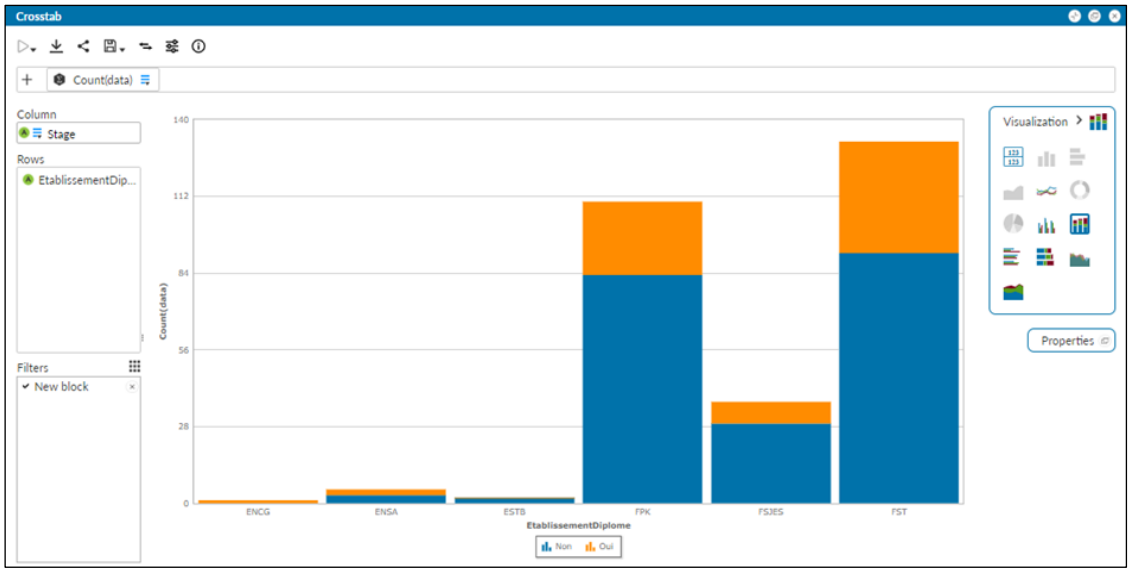

Figure 22. Les diplômés qui ont déjà passé un stage ou non, regroupés par établissement, en se limitant aux diplômés qui n'ont pas trouvé un emploi

Le graphe ci-dessous représente l'effectif des diplômés qui ont trouvé un emploi, en se limitant à ceux qui ont passé un stage ou non.

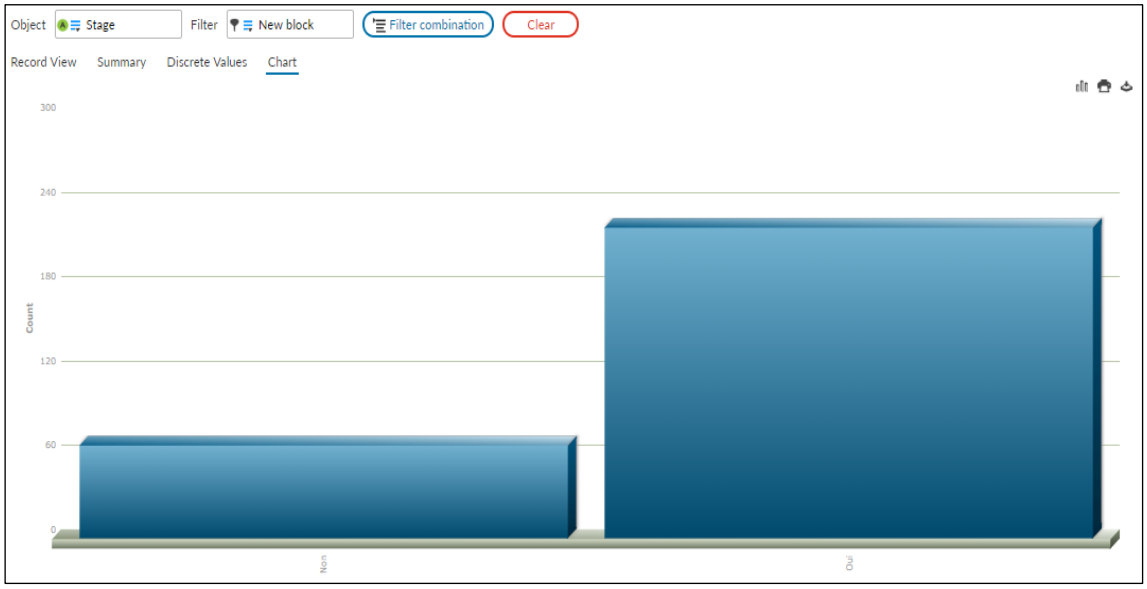

Figure 23. Les diplômés qui ont trouvé un emploi, en se limitant à ceux qui ont passé un stage ou non

L’objectif de l’enquête mené est de pouvoir proposer des solutions pour l’amélioration de l’employabilité et donner des opportunités aux décideurs pour qu’ils puissent faire un changement pour le meilleur. La présentation des données en utilisant des graphes peut nous donner une image un peu plus claire, mais ce n’est pas suffisant, prendre des décisions en se basant sur des graphes peut donner des résultats imprécis, le data mining et les algorithmes de machine learning ont prouvé leur capacité et leur exactitude dans plusieurs domaines, dans le but de proposer des solutions pour résoudre les problèmes.

Dans la partie suivante, on va préciser le rôle du data mining et des algorithmes de machine learning et leur grand potentiel pour pouvoir proposer des solutions pour l’amélioration de l’employabilité.

III. Le rôle du data mining sur l’employabilité

Les chercheurs de l'enseignement supérieur commencent à explorer le potentiel du data mining en analysant des données dans le but de fournir un service de qualité et les besoins de leurs diplômés [79]. L’emploi est la principale forme d’intégration sociale, un facteur d’amélioration des conditions de vie et de prévention des risques de pauvreté et de vulnérabilité, et l’indicateur le plus approprié pour évaluer le niveau de cohésion sociale dans un pays.

L’amélioration de l’employabilité des diplômés est une problématique importante et persistante. Elle représente un grave problème pour les diplômés; ils font chaque année face à de véritables concours pour assurer leur employabilité, l'insertion professionnelle est de plus en plus difficile. Il existe de nombreuses possibles explications et causes à cet égard, des facteurs qui prennent une grande responsabilité, par exemple la performance économique du pays, la structure de l'économie, le système éducatif, ou peut-être les domaines d'études universitaires qui font de l'insertion professionnelle un peu difficile.

L'utilisation du data mining et des algorithmes de machine learning permettra de clarifier la vue et de cerner les problèmes, tout en présentant des solutions telles que l'identification des déterminants responsables de l'insertion professionnelle des diplômés. C’est peut-être à cause du programme scolaire du diplômé, ou peut-être du marché du travail, ou du domaine

d’études choisi par les diplômés. Répondre à de telles questions pourrait être utile aux diplômés ainsi qu'aux chercheurs et aux autorités publiques pour mieux évaluer le système et la qualité de la formation et procéder aux ajustements nécessaires.

Un certain nombre de mesures d’orientation, d’information et de soutien doivent être adoptées pour offrir aux diplômés de bonnes solutions leur permettant de s’orienter facilement et de s’infiltrer dans le monde du travail.

Les résultats des algorithmes de machine learning sur les données de l’employabilité peuvent donner l'occasion de prendre des mesures efficaces pour améliorer l'intégration professionnelle des diplômés, ainsi que de valoriser les ressources humaines et le développement économique et social du pays. Clarifier le débat sur la problématique de l'employabilité et de l'insertion professionnelle des diplômés, et proposer un certain nombre de recommandations concernant les mesures éventuelles à prendre et les différentes ressources à mobiliser afin de renforcer l'employabilité et l'intégration professionnelle des diplômés au Maroc. Et c’est précisément le but de l’application du datamining sur les données d’employabilité.

Conclusion

Dans ce chapitre, nous avons défini l'employabilité et ce qu’elle signifie en général, après avoir présenté quelques travaux précédents sur l’application de l’exploration de données sur l’employabilité et les résultats obtenus, après on a présenté sous forme de graphes la situation de l'employabilité au Maroc en se basant sur les données d’une enquête menée par l'université Hassan 1er en 2016,et enfin nous avons expliqué pourquoi l'exploration de données est importante et comment elle va aider les décideurs à améliorer l’employabilité.

Chapitre 3 : Prédire l'employabilité à l'aide des techniques de Data mining et des algorithmes de machine learning

Introduction

L'emploi est la principale forme d'intégration sociale, un facteur d'amélioration des conditions de vie et de prévention des risques de pauvreté et de vulnérabilité, et l'indicateur le plus approprié pour évaluer le niveau de cohésion sociale dans un pays. L'exploitation des données relatives à l'employabilité donnera aux décideurs une excellente vue des données et des opportunités d'amélioration dans ce secteur. Des algorithmes avancés de data mining sont nécessaires pour obtenir des résultats précis et des connaissances permettant de prédire les observations futures.

L'amélioration de l'employabilité des diplômés est une problématique importante et persistante. L'employabilité représente un grave problème pour les diplômés; ils font chaque année face à de véritables concours pour assurer leur employabilité, l'insertion professionnelle est de plus en plus difficile. C'est pourquoi nous utilisons le data mining afin de proposer des solutions et des perspectives de prédiction future.

Dans ce chapitre, on va présenter un modèle de prédiction de l'employabilité utilisant des algorithmes de machine learning de classification, ainsi que les variables jouant un rôle important dans la prédiction de l'employabilité des diplômés. Mais avant, on va présenter une étude expérimentale comparant divers algorithmes de machine learning de classification sur des données d'employabilité au Maroc: arbre de décision, régression logistique et Naïve Bayes. L'objectif de notre expérience en premier est de choisir l'algorithme le plus efficace et le mieux adapté aux données d'employabilité, et après cette étape la présentation du modèle d'employabilité et des variables jouant un rôle important dans la prédiction de l'employabilité des diplômés, et la visualisation des résultats.

I. Travaux liés

Des travaux précédents ont été réalisés pour explorer et comparer les algorithmes de classification de data mining, pour choisir le plus efficace et présenter par la suite le modèle de l'algorithme le plus performants. Quelques travaux sont listés ci-dessous.

M.venkatadri, Lokanatha et C. Reddy [80] ont présenté une étude comparative de différentes algorithmes de classifications de data mining avec leurs limites, et ont également évalué leurs performances à l'aide des analyses expérimentales basées sur des données générées, bien que

les données collectées réelles soient meilleures pour l'extraction des données et la réalisation des modèles créés. Préparer les données au début sera difficile et prendra beaucoup de temps, mais les résultats seront réels et pourront être utilisés dans la vie réelle. D'après leur résultat, ils ont observé qu'il existait une relation directe entre le temps d'exécution lors de la construction du modèle arborescent et le volume des enregistrements de données, ainsi qu'une relation indirecte entre le temps d'exécution lors de la construction du modèle et la taille d'attribut des ensembles de données. Au cours de leur expérience, ils ont concluent que les algorithmes SPRINT et Random Forest ont une bonne précision de classification par rapport aux autres algorithmes comparés.

Pooja Thakar, Anil Mehta et Manisha [81] ont proposé une étude empirique comparant des algorithmes de classification variés sur deux ensembles de données d'étudiants MCA (Masters in Computer Applications) collectés dans divers collèges affiliés d'une université d'État réputée en Inde. La prédiction et l'analyse en temps opportun des performances des étudiants peuvent aider la direction, les enseignants et les étudiants à travailler pour obtenir de meilleurs résultats et de meilleures perspectives d'emploi. Les résultats ont montrés que l'algorithme d'arbre de décision présente de meilleurs résultats avec une précision de modèle de 84.5%.

Muskan Kukreja, Stephen Albert Johnston et Phillip Stafford [82] ont utilisé plusieurs algorithmes de classification pour analyser les données. Ils ont constaté que Naïve Bayes est beaucoup plus utile que d'autres méthodes largement utilisées en raison de sa simplicité, de sa robustesse, de sa rapidité et de sa précision. Ils ont travaillé sur les données d'employabilité. Il est donc nécessaire de recourir à de différents algorithmes de classification pour déterminer celle qui convient le mieux à ce type de données, car chaque donnée a sa propre spécificité, et la performance d'un algorithme peut être différente, cela dépend du type des données.

Mohd tajul rizal, yuhanis Yusof [83] ont proposé une expérience de prédiction de l'emploi des diplômés; leur expérience comprenait l'utilisation de cinq techniques de data mining, à savoir Naive Bayes, la régression logistique, le perceptron multicouche, k-nearest neighbor et l'arbre de décision. Les résultats ont montré que la régression logistique est le meilleur classificateur pour l'ensemble de données.

KENO C. PIAD [84] a proposé un modèle d'exploration de données permettant de prédire l'employabilité des diplômés en informatique, en déterminant si les diplômés en informatique se retrouveraient dans une profession liée ou non à l'informatique. L'objectif de l'étude est

également déterminer les attributs dominants à l'aide d'algorithmes de data mining dans le cadre de l'apprentissage supervisé et comparer leur précision. Parmi les techniques de classification utilisées pour comparer l'exactitude, les algorithmes Naive Bayes, J48, Simple Cart, Régression logistique et Chaid. Les résultats ont montré que 3 facteurs significatifs ayant un effet direct sur l'employabilité informatique, comprennent IT_Core, IT_Professional et Gender.

Azziaty Abdul Rahman, Kian Lam Tan et Chen Kim Lim [85] ont proposé une expérience pour déterminer le meilleur modèle pouvant être utilisé pour prédire le statut d'emploi d'institutions publiques récemment diplômées : employées ou non, six mois après l'obtention de leur diplôme. K-nearest neighbor, Naive Bayes, arbre de décision, réseau de neurones, régression logistique et machines à vecteurs de support ont été comparés à l'aide du jeu de données de formation de l'étude Tracer afin de déterminer la plus grande précision, puis utilisé comme modèle prédictif. Rapid Miner en tant qu'outil de data mining a été utilisé.

Parneet Kaur,Manpreet Singh et Gurpreet Singh Josan [86] ont proposé une expérience visant à identifier les élèves qui ont le potentiel de trouver un emploi à l'aide de techniques de classification d'exploration de données utilisant des données du monde réel recueillies dans des écoles secondaires. Ils ont appliqué plusieurs algorithmes tels que Multilayer Perception, Naïve Bayes, SMO, J48 et REPTree afin de trouver le meilleur modèle de classification. Les résultats ont montré que l'algorithme de perception multicouche est le meilleur classificateur avec une précision de 75% du modèle.

Rosna Awang Hashim, Lim Hock Eam, Bidin Yatim, Tengku Faekah Tengku Ariffin, Ainol Madziah Zubairi, Haniza Yon et Omar Osman [87] ont proposé l'élaboration d'un modèle de prédiction pour l'identification précoce de l'employabilité des diplômés en Malaisie. Ils ont présenté dans leur travail comment un modèle de prédiction pourrait être construit pour la détection et l'identification précoce de l'état d'employabilité des diplômés. En se basant sur les résultats obtenus, les déterminants ayant une influence positive sur l'employabilité des diplômés sont le sexe, l'âge, la MPC, le type de diplôme, l'université et la maîtrise de la langue anglaise.

II. Prédire l'employabilité à l'aide des techniques de Data mining:

Dans cette expérience, nous avons utilisé RMS (Rapid Miner Studio) Educational Version 8.1.000. C'est un logiciel open source et gratuit [88] dédié au data mining. Il contient de nombreux outils pour traiter des données : lecture de différents formats d'entrée, préparation et nettoyage des données, statistiques, algorithmes de data mining, évaluation des performances et visualisations diverses de données. Il propose l'implémentation des algorithmes d'apprentissage automatique, et il inclut également une extension Weka permettant d'implémenter des algorithmes conçus pour l'outil d'extraction de Weka.

1. Choix de l'outil du data mining : Rapid Miner

Kalpana Rangra et K.L. Bansal [89] ont présenté une étude comparative des outils de data mining, afin de présenter les avantages et les inconvénients de chacun, et les résultats de cette comparaison ont montré, comme présenté dans le Tableau 2, que le choix dépend de la nature de l'expérience souhaitée. Par exemple, Weka est très robuste avec des fonctionnalités intégrées et offre des fonctionnalités supplémentaires. Rapid Miner et Orange sont destinés aux utilisateurs avancés, en particulier dans les sciences exactes, car ils nécessitent des compétences de programmation supplémentaires et une prise en charge limitée de la visualisation.

Tableau 2. Description des outils de data mining

Outil	**Type**	**Avantages**	**Limitations**
Rapid Miner	Analyse statistique, data mining, analyse prédictive.	Visualisation, statistique, sélection d'attribut, détection des valeurs aberrantes, optimisation des paramètres.	Nécessite une connaissance approfondie de la gestion des bases de données et des techniques de data mining.
Orange	Apprentissage automatique, Data mining, Visualisation de données.	Meilleur débogueur, scripts les plus courts, statistiques médiocres, convient aux experts debutants.	Grande installation, capacités de reporting limitées.

R	Calcul statistique.	Purement statistique.	Moins spécialisé pour l'exploration de données, nécessite la connaissance du langage de tableau.
Weka	Apprentissage Machine.	Facilité d'utilisation, peut être étendu dans RM.	Documentation médiocre, statistiques classiques faibles, optimisation des paramètres médiocres, lecteur csv faible.

Rapid Miner est l'outil présentant l'indépendance de la limitation du langage et fournit une analyse prédictive et des capacités statistiques. Il est donc facile à utiliser et à mettre en œuvre sur pratiquement tous les systèmes. Il intègre également des algorithmes de tous les autres outils mentionnés, et il est utilisé pour le Big Data.

Rapid Miner a été nommé l'un des leaders du Magic Quadrant 2019 de Gartner pour les plates-formes de science des données et d'apprentissage automatique pour la sixième année consécutive, selon Gartner [90].

Selon le rapport de Gartner en 2019 [91], les leaders devraient conduire la transformation du marché. Ils ont les scores combinés les plus élevés pour la capacité à exécuter et la complétude de la vision. Ils se débrouillent bien et sont préparés pour l'avenir avec une vision claire et une compréhension approfondie du contexte plus large des affaires numériques. Ils ont de solides partenaires de distribution, une présence dans plusieurs régions, des performances financières constantes, une large plate-forme de support et un bon support client. Rapid Miner est une plate-forme logicielle pour les équipes d'analyse qui allie préparation des données, apprentissage automatique et déploiement de modèles prédictifs.

Les caractéristiques [92] de RapidMiner:

- **Simplicité sophistiquée:**

Des fonctionnalités telles que le modèle automatique, des fonctionnalités d'analyse étendues telles que Turbo Prep et une interface utilisateur au-dessus de la moyenne font de Rapid Miner Studio un favori des scientifiques. Les utilisateurs plus avancés apprécient la richesse des fonctionnalités de Rapid Miner, notamment la possibilité d'accéder aux fonctionnalités

open source et de les réutiliser, ce qui augmente leur productivité et leur permet de créer et de gérer un grand nombre de modèles.

- **Fonctions avancées:**

La facilité d'utilisation n'exclut pas la présence de pouvoir. Au-delà de l'apprentissage en profondeur et de la prise en charge des GPU, la plate-forme Rapid Miner inclut désormais une fonctionnalité d'augmentation des données et des fonctionnalités améliorées pour les séries chronologiques. La société s'est également concentrée aussi sur l'interprétation, tant du point de vue du modèle que du processus d'analyse. En plus d'aider à expliquer les comportements des modèles, fournir une plus grande transparence au niveau du processus, du développement au déploiement (en définissant clairement les étapes du pipeline analytique et en fournissant la logique analytique reliant ces étapes), permet une plus grande collaboration entre les rôles.

- **Plate-forme cohérente de bout en bout:**

Les clients de référence ont formulé de nombreux commentaires complémentaires sur la cohérence de l'expérience utilisateur de Rapid Miner. Les éléments contribuant au continuum incluent: Rapid Miner Studio pour le développement de modèles, Rapid Miner Server pour le partage, la collaboration, le déploiement et la maintenance de modèles, et RapidMiner Real-Time Scoring introduit en 2018 pour fournir un moteur d'exécution de modèle à faible temps de latence.

Dans la section suivante, nous allons présenter nos premières contributions, elles consistent à trouver le meilleur algorithme pour la prédiction de l'employabilité au Maroc, en appliquant plusieurs algorithmes de classification et en choisissant le meilleur modèle en se basant sur des métriques d'évaluation de modèle. Et aussi de présenter le modèle de prédiction en définissant les variables qui ont l'impact le plus sur l'employabilité, en présentant le processus détaillé du data mining qu'on a suivi.

2. Résultats expérimentaux:

Dans cette section, nous allons présenter les résultats obtenus. En premier temps, nous avons comparé divers algorithmes de data mining de classification, pour classer les diplômés en «Working» et «NotWorking», tels que l'arbre de décision, la régression logistique et Naïve Bayes, puis nous avons choisi l'algorithme le plus efficace et le plus adapté aux données d'employabilité, après nous allons présenter les variables qui ont impact sur l'employabilité

en se basant sur le modèle de l'algorithme le plus performant en utilisant des métriques d'évaluation de modèle. Nous avons utilisé Rapid Miner Studio Educational Version 8.1.000.

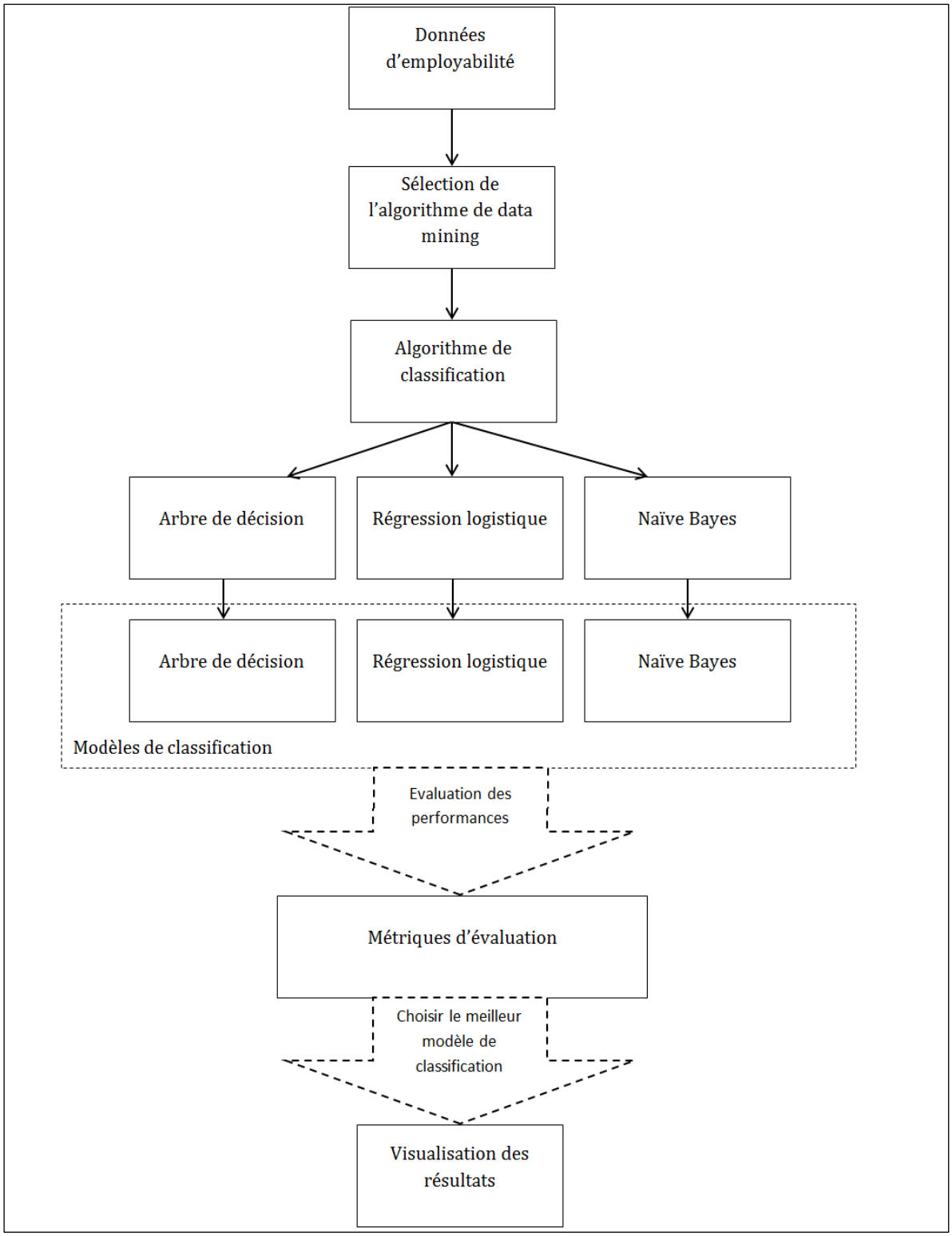

Figure 24. Architecture présentant les différentes étapes suivies depuis la sélection et l'implémentation des algorithmes jusqu'à la visualisation des résultats

Ce graphe présente les différentes étapes suivies dans ce chapitre afin de présenter le modèle de prédiction de l'employabilité; nous avons commencé par la collecte des données que nous avons utilisées et la manière dont elles ont été collectées. Puis nous avons implémenté plusieurs algorithmes de classification, et en utilisant des métriques d'évaluation de modèle, on a choisi l'algorithme le plus adapté et le plus efficace présentant les meilleures performances. Après on a présenté les résultats en présentant le modèle et les variables qui jouent un rôle important dans la prédiction de l'employabilité.

2.1. La collecte des données:

Les données utilisées dans cette étude sont recueillies à partir d'une enquête sur l'employabilité conduite par l'université Hassan 1er en 2016 en partenariat avec le Bureau national de l'évaluation (NEO), sous l'égide du Conseil supérieur de l'éducation, de la formation et de la recherche scientifique. Les données sont volumineuses, multi variées, incomplètes, hétérogènes et déséquilibrées. Dans la prochaine phase, nous préparerons les données et nous les nettoierons afin de pouvoir appliquer les algorithmes de classification: Arbre de décision, Régression logistique et Naïve Bayes.

2.2. Préparation des données :

La préparation des données pour qu'elles soient prêtes pour l'implémentation des algorithmes de data mining est une phase cruciale et constitue la partie la plus fastidieuse du processus. Il est donc nécessaire de préparer les données. Les données contiennent 1752 lignes et 22 attributs. Ces données doivent toutefois être nettoyées, on doit supprimer les données non pertinentes, telles que le nom, le numéro de téléphone, l'email, etc. Transformer également des données et créer de nouveaux attributs, des attributs calculés par exemple.

On commence par **la sélection de données**, nous sélectionnons les données dont nous avons besoin pour répondre au problème en question, nous explorons donc les données et nous ne sélectionnons que les données que nous souhaitons utiliser.

Après vient l'étape : **la suppression des colonnes.** Nous supprimons toutes les colonnes non pertinentes qui n'ont rien à voir avec notre problématique, telles que le nom, le téléphone, l'email, etc.

Remplacer les valeurs manquantes : Les valeurs manquantes conduisent à une génération de modèle inexacte, ce qui entraîne des résultats erronés sur lesquels nous ne pouvons pas

compter. Nous avons donc résolu ce problème par le biais de techniques d'imputation, ou éliminé complètement les instances si nécessaire.

Maintenant **la réorganisation des attributs**, nous réorganisons les attributs de l'ensemble de données, uniquement pour une meilleure organisation des données.

Les données finales contiennent 1208 instances de 13 attributs. La liste et la description des attributs figurent ci-dessous dans le Tableau 3.

Tableau 3. Les attributs avec description

No.	Attributs	Type	Description
1.	Gender	Binaire	Sexe des diplômés (homme, femme).
2.	Diploma	Nominal	Type de diplôme.
3.	Field	Nominal	Filière d'étude.
4.	Grade	Ordinal	Quelle mention au baccalauréat
5.	University	Nominal	Quelle université?
6.	PracticeLevel	Ordinal	Le niveau de pratique dans son domaine d'études.
7.	InformaticLevel	Ordinal	Le niveau de l'informatique du diplômé.
8.	FrenchLevel	Ordinal	Le niveau de français du diplômé.
9.	EnglishLevel	Ordinal	Le niveau d'anglais du diplômé.
10.	BaccalaureateSerie	Nominal	L'option du baccalauréat.
11.	TrainingPeriod	Binaire	Le diplômé a-t-il fait des stages (oui ou non)?
12.	TheoricalLevel	Ordinal	Le niveau de théorie dans son domaine d'études
13.	Employability	Binaire	Est-ce que le diplômé travaille ou ne travaille pas.

2.3. Phase de modélisation: Implémentation des algorithmes de classification

La phase de modélisation et la phase d'évaluation peuvent être répétées plusieurs fois afin de pouvoir modifier les paramètres jusqu'à l'obtention de valeurs optimales.

La première étape de la phase de modélisation est **la définition de la cible**.

Maintenant, nous définissons la cible, ce qui signifie que la variable que nous voulons prédire: l'employabilité est notre variable, avec deux classes, **working** et **notWorking**.

L'étape suivante consiste à répondre à la question suivante : **Quel modèle devrions-nous appliquer?**

En fonction du type de données dont nous disposons et du type de variable que nous voulons prédire, nous choisissons la méthode qu'on va appliquer au cours de cette phase.

Nous allons travailler avec l'une des méthodes supervisées, puisque nous avons une cible, qui est la variable à prédire, l'employabilité, avec deux classes: **working** et **notWorking**.

Le type de notre variable à prédire est qualitatif, contenant deux classes, ce qui signifie que le type de la variable est binaire; nous allons travailler avec une technique de classification.

On va appliquer les algorithmes de classification, Arbre de décision, régression logistique et Naïve Bayes classés parmi les dix premiers algorithmes de classification [93]. Après, on va évaluer chaque modèle et, enfin, choisir quel algorithme qui présente le modèle le plus précis en évaluant les performances de ces modèles au cours de la phase suivante.

Divisé les données est la dernière étape dans cette phase avant d'appliquer les algorithmes, nous allons diviser les données en deux ensembles, les données d'apprentissage à 80% et les données de test à 20%. Dans les données d'apprentissage, nous construisons les modèles et les appliquons ensuite aux données de test. Comme les données de test n'ont jamais été vues par le modèle, les performances sont donc un bon guide pour savoir ce qui sera vu lorsque le modèle est appliqué à des données non vues par le modèle.

Maintenant, nous avons implémenté les algorithmes de classification, Arbre de décision, Régression logistique et Naïve Bayes, est ci-dessous les résultats obtenus.

Tableau 4. La distribution des classes

Classe	**Distribution**
Working	586 (49%)
Not working	622 (51%)

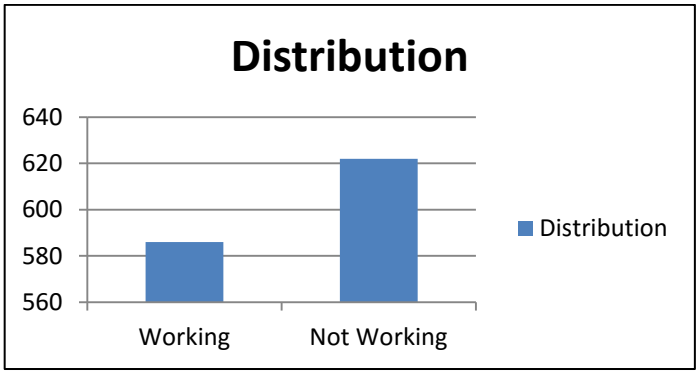

Figure 25. Graphe de la distribution des classes

Nous avons utilisé différentes métriques [94] pour comparer les modèles obtenus. Les différentes métriques utilisées : Précision, Taux d'erreur, Recall, kappa statistics, F measure, ROC (receiver operating characteristic) et le temps pour construire le modèle, Voici une description ci-dessous des différents métriques que nous avons utilisés.

2.4. Métriques d'évaluation des performances :

Tableau 5. Matrice de confusion pour la classification binaire

		Prédite	
		Classe positive	**Classe négative**
Réelle	**Classe positive**	**Vrais positifs (VP)**	Faux négatifs (FN)
	Classe négative	Faux positifs (FP)	**Vrais négatifs (VN)**

VP: le nombre de «vrais positifs», instances positives correctement identifiées par l'algorithme.

FP: le nombre de «faux positifs», instances négatives mal identifiées par l'algorithme.

FN: le nombre de «faux négatifs», instances positives mal identifiées par l'algorithme.

VN: le nombre de «vrais négatifs», instances négatives correctement identifiées par l'algorithme.

2.4.1. Précision / Accuracy (P):

En général, la précision mesure le rapport des prévisions correctes sur le nombre total d'instances évaluées.

Formule :

$$\frac{vp + vn}{vp + fn + fp + vn}$$

2.4.2. Taux d'erreur:

Le taux d'erreur mesure le rapport des prévisions incorrectes sur le nombre total d'instances évaluées.

Formule:

$$\frac{fp + fn}{vp + fp + vn + fn}$$

2.4.3. Rappel / Recall (Rec) :

Recall calcule en fait le nombre de positifs actuels capturés par notre modèle en le qualifiant de positif (de vrai positif).

Formule:

$$\frac{vp}{vp + fn}$$

Kappa:

Le coefficient de Kappa est toujours compris entre -1 et 1 (accord maximal).

Formule :

$$R = ((VP + FN)(VP + FP) + (VN + FP)(VN + FN))/n2$$

$$Kappa = \frac{P - R}{1 - R}$$

En se basant sur l'interprétation de Cohen [95] pour les résultats Kappa on utilise le barème suivant pour interpréter la valeur de Kappa obtenue :

Tableau 6. Coefficient de kappa et interprétation

Kappa	**Accord**
<0	Grand désaccord / Très mauvais
0.00 – 0.20	Accord très faible / Mauvais
0.21 – 0.60	Accord faible / Modéré
0.61 – 0.80	Accord satisfaisant / Bon
0.81 – 1.00	Accord excellent / Excellent

2.5. Résultats:

Après avoir appliqué les algorithmes Arbre de décision, Régression logistique et Naïve Bayes, le tableau 7 ci-dessous décrit les résultats de cette expérience.

Tableau 7: Résultats de comparaison des performances des classificateurs

Algorithme / Métrique	Arbre de décision	Régression logistique	Naïve Bayes
Précision	81.70 %	80.79 %	78.23 %
Taux d'erreur	18.30 %	19.21 %	21.77 %
Recall	92.92 %	80.13 %	90.20 %
Kappa	0.631	0.616	0.561
F-measure	0.84	0.81	0.81
Temps (ms)	390	344	47

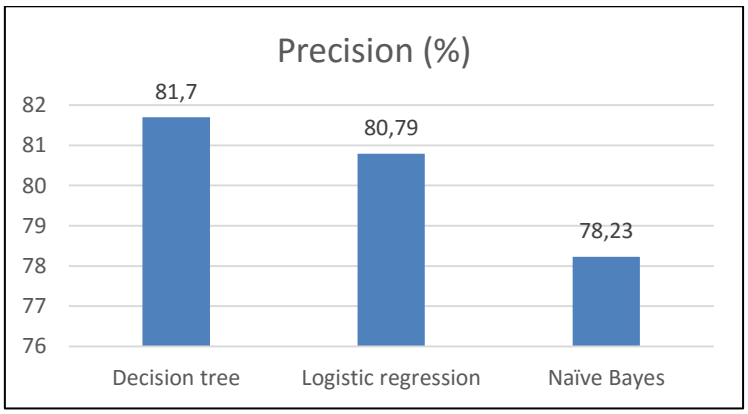

Figure 26. Graphe de la précision de prédiction des classificateurs

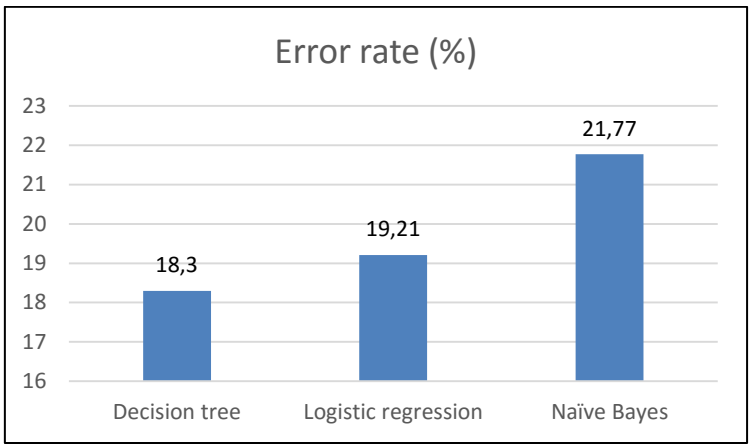

Figure 27: Le taux d'erreur de classification des modèles

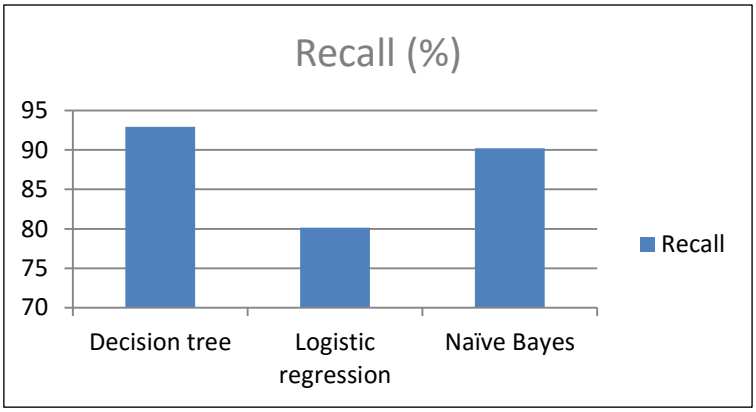

Figure 28: Le taux de Recall de classification des modèles

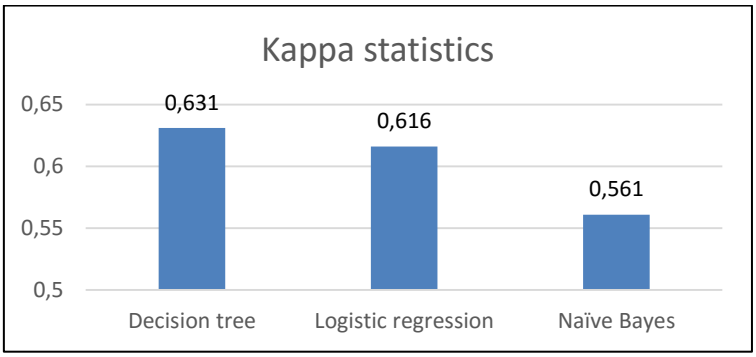

Figure 29. Graphe de la statistique de Kappa

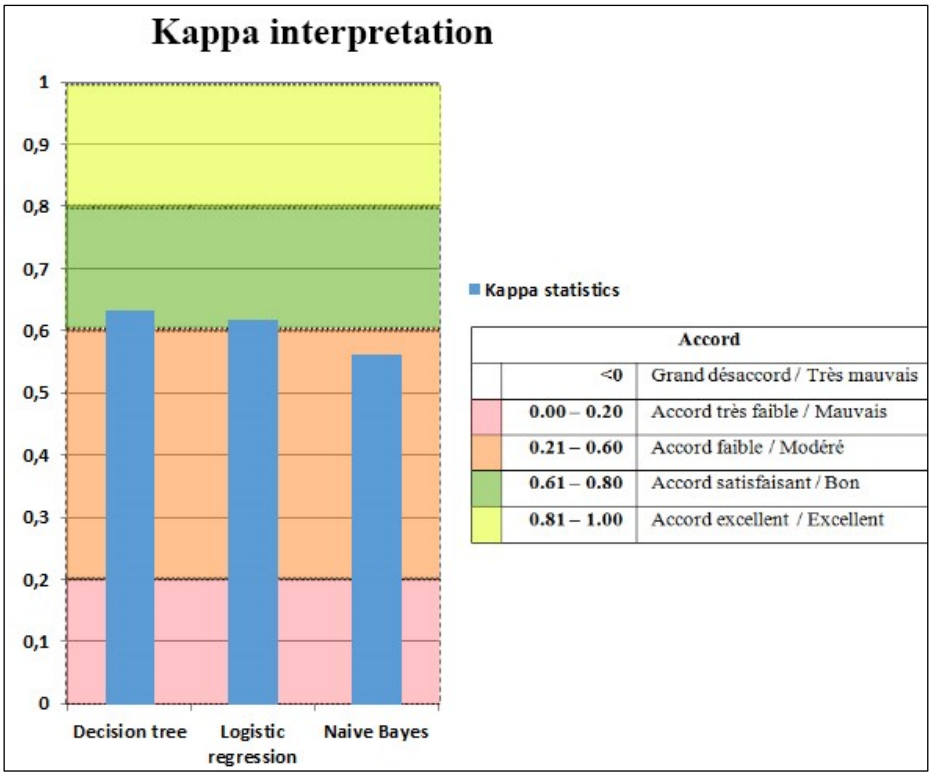

Figure 30. L'interprétation de Kappa

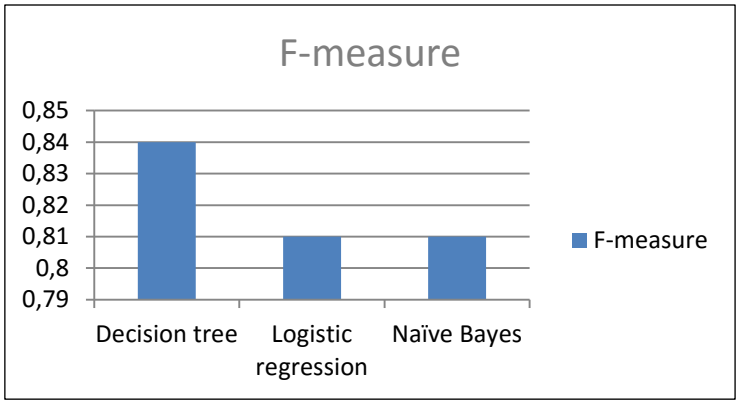

Figure 31. Graphe de f-measure des classificateurs

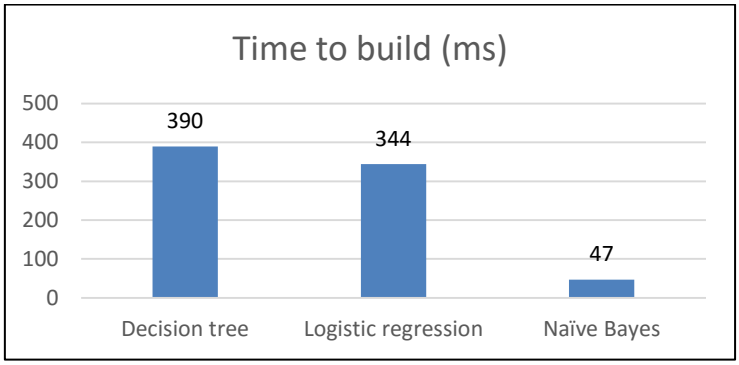

Figure 32. Graphe représentant le temps de construction des modèles (ms)

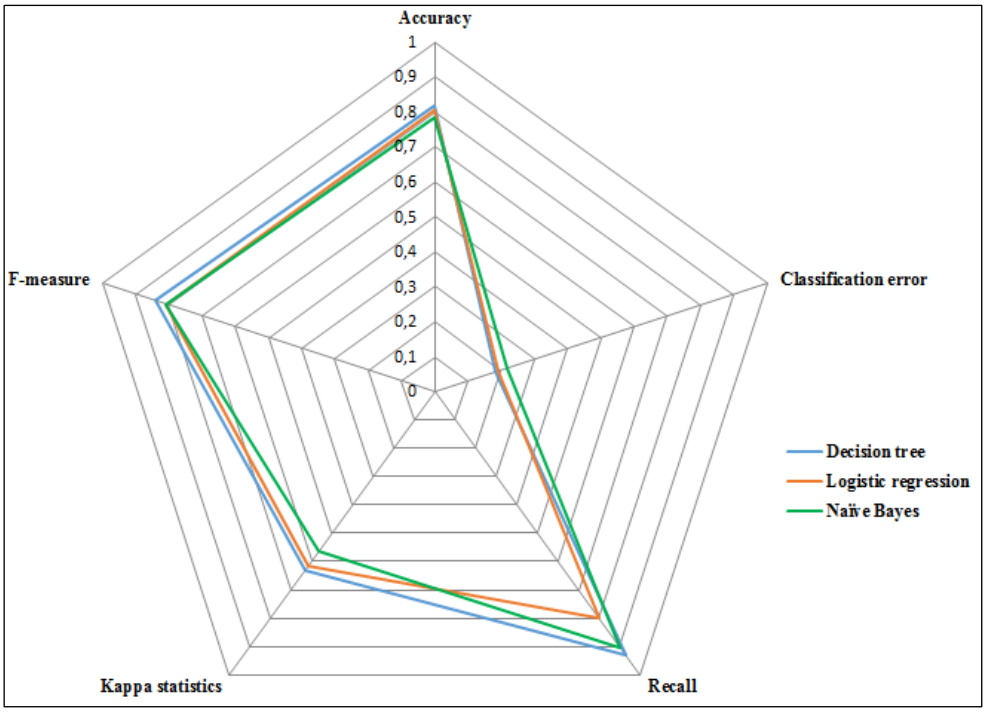

Figure 33. Un affichage radial comparant les trois algorithmes utilisant les différentes métriques d'évaluation

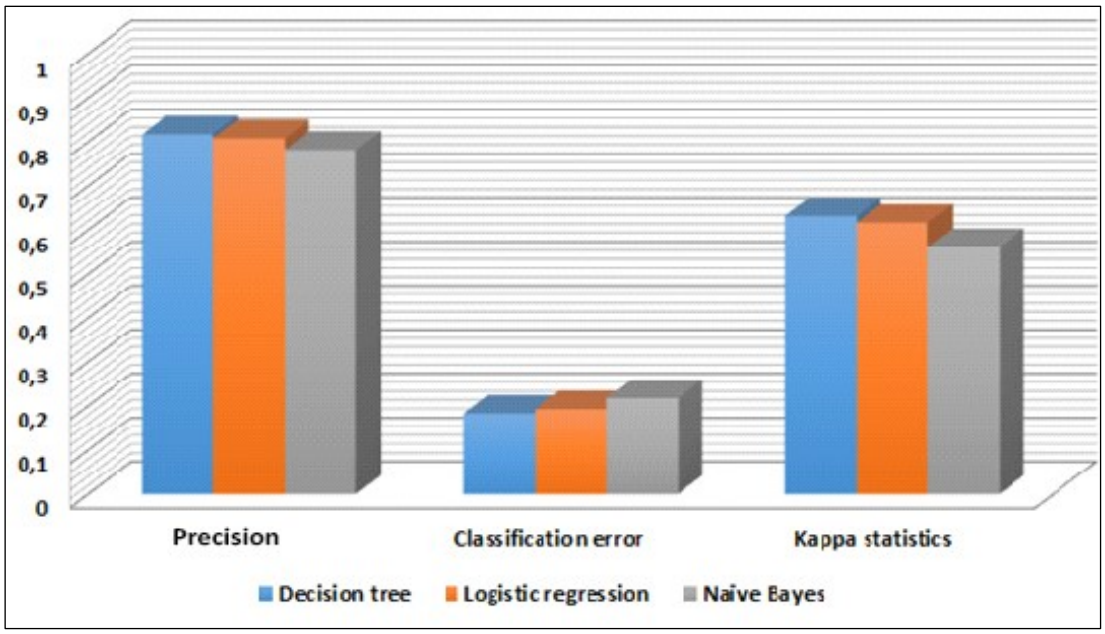

Figure 34. Précision, erreur de classification et kappa statistique

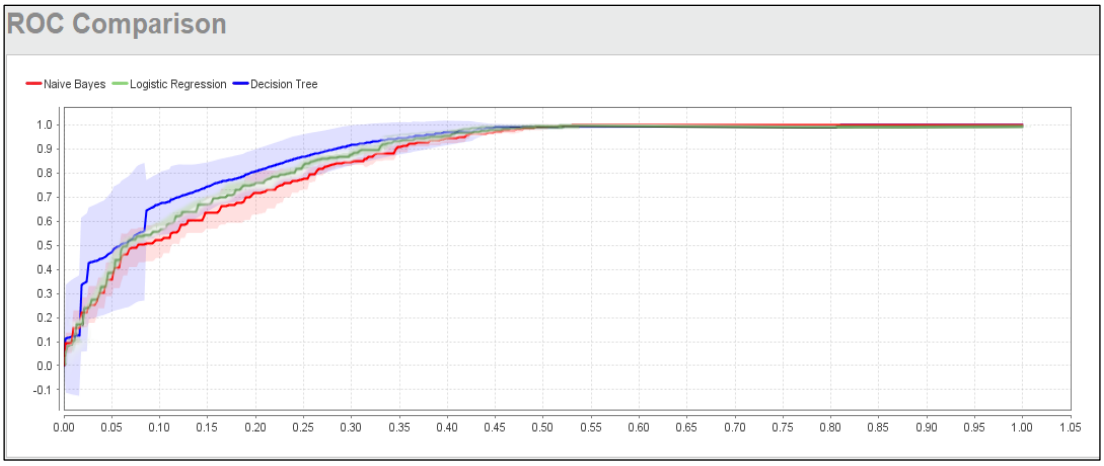

Figure 35. Comparaison des performances des trois classificateurs en utilisant la courbe de ROC

Dans notre expérience de prédiction utilisant les algorithmes de classification Arbre de décision, Régression logistique et Naïve Bayes, nous avons utilisé différentes métriques pour comparer et choisir l'algorithme le plus efficace pour les données d'employabilité, telles que la précision, le taux d'erreur, Recall, F-measure, la statistique de Kappa, ROC et temps pour construire le modèle.

La précision représente les pourcentages d'instances correctement classées par l'algorithme, en se basant sur les résultats. La figure 26 montre que la précision du modèle prédite par

l’arbre de décision (81,70%) est plus précise que la régression logistique (80,79) et du Naïve Bayes avec (78,23%), ce qui signifie également que le taux d’erreur de l'arbre de décision (18,30%) est inférieur à celle de la régression logistique (19,21%) et du Naïve Bayes (21,77%), qui ont plus d'instances non classées. Aussi, la statistique de Kappa a montré que le modèle d’arbre de décision (0,631) était meilleur que la régression logistique (0,616) et Naïve Bayes (0,561). En se basant sur l'interprétation de Cohen pour les résultats Kappa, les modèles d'arbre de décision (0.63) et de régression logistique (0.61) sont bons et représentent un accord satisfaisant et le modèle de Naïve Bayes (0.56) est modéré et représente un accord faible.

Une autre métrique importante est la F-measure, elle permet de rechercher un équilibre entre précision et rappel. Les résultats montrent à nouveau que l'arbre de décision est classé correctement avec un taux F-measure de 0,84, par rapport à la régression logistique avec 0,81 et Naïve Bayes avec 0,81. Comme nous pouvons le constater, l’arbre de décision a la valeur la plus élevée pour la mesure F, ce qui garantit que la précision et le rappel sont raisonnablement élevés.

Une autre métrique importante est Recall, elle indique 92,92% pour l’arbre de décision, 90,20% pour Naïve Bayes et 80,13% pour la régression logistique. En termes de temps, Naïve Bayes n'a pris que 47 ms pour construire le modèle, tandis que l'arbre de décision a pris 390 ms et la régression logistique 344 ms. Nous avons également utilisé la courbe de ROC pour comparer les trois modèles. La figure 33 montre clairement que le modèle de l’arbre de décision est plus précis que la régression logistique et Naïve Bayes. Plus la courbe est proche du bord gauche et du bord supérieur de l'espace ROC, plus le modèle est précis.

D’après les résultats obtenus comparant les trois algorithmes de classification, Arbre de décision, Régression logistique et Naïve Bayes, les résultats ont montré que le modèle d’Arbre de décision est le meilleur par rapport à la Régression logistique et Naïve Bayes dans toutes les métriques, Précision, Taux d’erreur, kappa, F-measure, Recall, et ROC, sauf le temps pour construire le modèle, Naïve Bayes était le plus rapide. Nous allons maintenant présenter le modèle d’arbre de décision (C4.5) pour prédire l’employabilité et présenter les résultats selon le modèle d'arbre de décision appliqué. Les variables qui jouent un rôle important dans la prédiction de l'employabilité des diplômés et qui sont identifiées par l’algorithme comme des prédicteurs ayant un effet direct sur l’employabilité sont les

suivantes : University, elle représente l'université ou le diplômé a eu son diplôme. Diploma, le diplôme du diplômé (Master, doctorat, etc.). Grade, représente la mention que le diplômé a eu au baccalauréat, Training-Period, elle représente si le diplômé a effectué un stage ou non, et French-Level¸ elle représente le niveau du français du diplômé.

Tableau 8. Matrice de confusion du modèle d'arbre de décision

Matrice de confusion	**Working**	**Not Working**
Working	**409**	44
Not Working	177	**578**

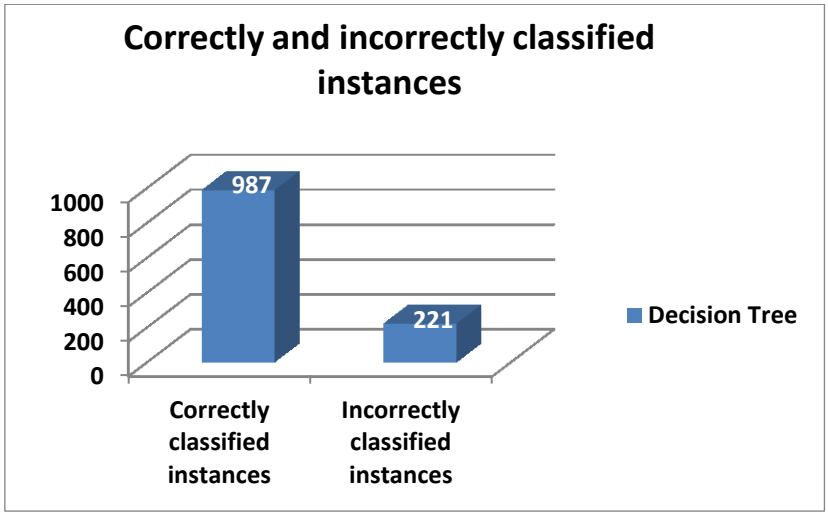

Figure 36. Les instances correctement classées et incorrectement classées par le classificateur arbre de décision

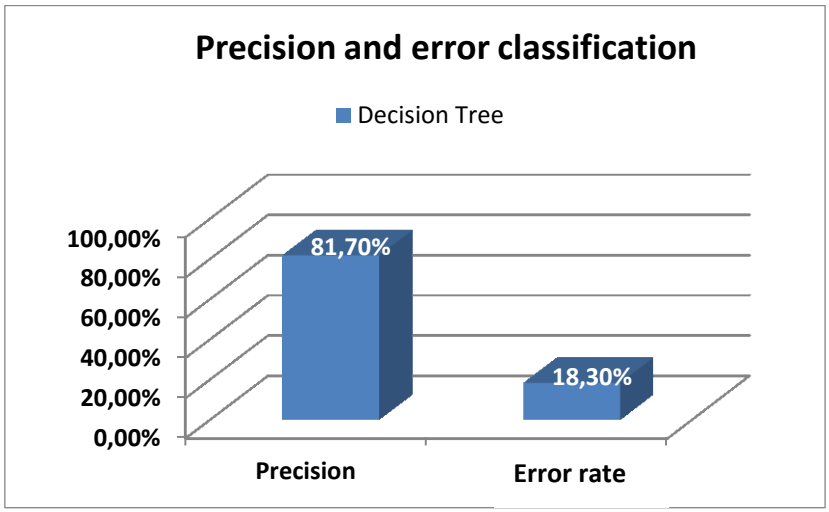

Figure 37. Précision et taux d'erreur de l'arbre de décision

2.6. Modèle développé par l'algorithme d'arbre de décision

Ci-dessous le modèle d'arbre de décision, ces règles décrivent et donnent un aperçu clair des attributs qui affectent l'employabilité des diplômés.

University = ENCG: Working {Working=143, NotWorking=30}
University = ENSA
| | Grade = Good: Working {Working=15, NotWorking=2}
| | Grade = Passable: NotWorking {Working=3, NotWorking=16}
| Gender = Male: Working {Working=20, NotWorking=3}
University = ESTB
| EnglishLevel = High
| | TrainingPeriod = No: NotWorking {Working=2, NotWorking=11}
| | TrainingPeriod = Yes
| | | FrenchLevel = Low: NotWorking {Working=1, NotWorking=3}
| | | FrenchLevel = Medium: Working {Working=3, NotWorking=1}
| | | FrenchLevel = High
| | | | InformaticLevel = Very Good: Working {Working=8, NotWorking=1}
| | | | InformaticLevel = Medium: Working {Working=6, NotWorking=0}
| | | | InformaticLevel = Excellent: Working {Working=13, NotWorking=0}
| EnglishLevel = Excellent: Working {Working=5, NotWorking=0}
| EnglishLevel = Low: NotWorking {Working=0, NotWorking=3}
| EnglishLevel = Medium: Working {Working=3, NotWorking=0}
University = FPK: NotWorking {Working=100, NotWorking=212}
University = FSJES: Working {Working=161, NotWorking=105}
University = FST
| Grade = Very Good
| | Diploma = Diplôme d'ingénieur: Working {Working=10, NotWorking=0}
| | Diploma = Doctorat: Working {Working=6, NotWorking=0}
| | Diploma = Licence professionnelle: NotWorking {Working=5, NotWorking=13}
| | Diploma = Licence sciences et techniques
| | | FrenchLevel = Medium: Working {Working=10, NotWorking= 3}
| | | FrenchLevel = Low: NotWorking {Working=0, NotWorking=13}
| | | FrenchLevel = Excellent: Working {Working=16, NotWorking=0}
| | | FrenchLevel = Vey Low
| | | | EnglishLevel = High
| | | | | PracticeLevel = Low: NotWorking {Working=0, NotWorking=10}
| | | | | PracticeLevel = Very Good
| | | | | | Gender = Female: Working {Working=12, NotWorking=0}
| | | | | | Gender = Male: Working {Working=4, NotWorking=0}
| | | | | PracticeLevel = Medium
| | | | | | InformaticLevel = Medium: Working {Working=6, NotWorking=1}

| | | | | | InformaticLevel = Very Good: Working {Working=7, NotWorking=0}
| | | | | PracticeLevel = Excellent
| | | | | | BaccalaureateSerie = SVT: Working {Working=7, NotWorking=0}
| | | | | | BaccalaureateSerie = Sciences et technologie électrique: Working {Working=2, NotWorking=0}
| | | | EnglishLevel = Excellent
| | | | | Field = Gestion: NotWorking {Working=0, NotWorking=12}
| | | | | Field = Génie Electrique: Working {Working=4, NotWorking=0}
| | | | | Field = Génie Mécanique: NotWorking {Working=0, NotWorking=4}
| | | | | Field = Management Logistique et Transport: Working {Working=2, NotWorking=0}
| | | | | Field = Protection de l'environnement: Working {Working=2, NotWorking=0}
| | | | | Field = Techniques d'Analyse et Contrôle de Qualité: NotWorking {Working=0, NotWorking=5}
| | | | EnglishLevel = Medium: Working {Working=10, NotWorking=1}
| | Diploma = Master
| | | PracticeLevel = Very Good: Working {Working=14, NotWorking=7}
| | | PracticeLevel = Low: NotWorking {Working=0, NotWorking=12}
| | | PracticeLevel = Medium: Working {Working=11, NotWorking=6}
| | | PracticeLevel = Excellent: Working {Working=4, NotWorking=0}
| | Diploma = Master: Working {Working=6, NotWorking=1}
| Grade = Good: Working {Working=43, NotWorking=24}
| Grade = Passable: NotWorking {Working=21, NotWorking=64}
| Grade = Excellent: Working {Working=37, NotWorking=4}

2.7. Analyse prescriptive :

L'analyse prescriptive [96] est liée à l'analyse prédictive. Alors que l'analyse prédictive aide à modéliser et à prévoir ce qui pourrait se produire, l'analyse prescriptive cherche à déterminer la meilleure solution ou le meilleur résultat parmi divers choix, compte tenu des paramètres connus.

3 niveaux de « Data ANALYTICS » :

- Analyse descriptive : Extraire des connaissances à partir des données.
- Analyse prédictive : Construire des modèles pour prévoir le futur.
- Analyse prescriptive : Assister la prise de décision. Optimisation de la prise de décision, une approche opérationnelle de la décision.

L'analyse prescriptive va au-delà de la connaissance. L'objectif principal est de «prescrire» un certain nombre d'actions différentes possibles qui vont guider vers une solution. En résumé,

ces analyses visent à fournir des conseils. L'analyse prescriptive tente de quantifier l'effet des décisions futures afin de conseiller sur les résultats possibles avant que les décisions ne soient réellement prises. Au mieux, l'analyse prescriptive prédit non seulement ce qui se passera, mais aussi pourquoi cela se produira en fournissant des recommandations concernant les actions qui tireront parti des prédictions.

Ces analyses vont au-delà de l'analyse prédictive en recommandant un ou plusieurs plans d'action possibles. Essentiellement, ces analyses permettent d'évaluer un certain nombre de résultats possibles en fonction de leurs actions. L'analyse prescriptive utilise une combinaison de techniques et d'outils tels que les règles métier, les algorithmes d'apprentissage automatique et les procédures de modélisation informatique. Ces techniques sont appliquées sur de nombreux ensembles de données différents, y compris les données historiques et transactionnelles, les flux de données en temps réel et les méga-données.

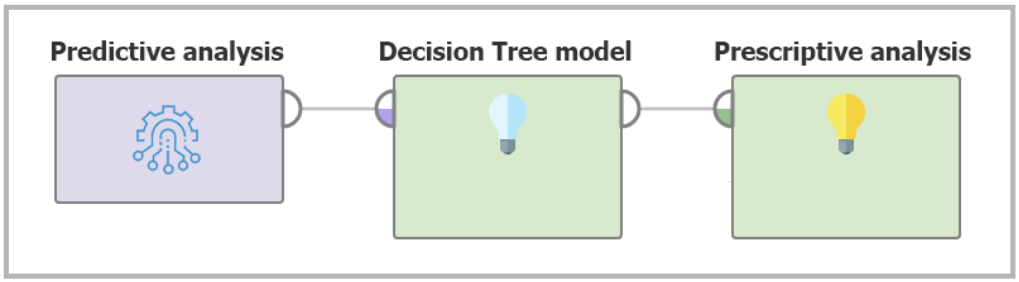

Figure 38. Analyse prédictive vers une analyse prescriptive

Dans la modélisation prédictive, un modèle est utilisé pour prédire un résultat, en fonction d'une entrée. Cet opérateur inverse cette procédure, en commençant par un modèle et une sortie souhaitée, et en prescrivant une entrée optimisée pour atteindre le résultat souhaité, ci-dessus une figure décrivant le processus de l'analyse prescriptive.

Résultats de l'analyse prescriptive:

...	predict...	confidence(Emploi)	...	ConnaissancesInformatiques	Diplome	Etablisseme...	Filiere	MentionBac
1	Emploi	1	0	Moyen	Diplôme d'ingénieur	ENSA	Mathématiques et Applications	Très bien

NiveauAnglais	NiveauFranç...	Pratique	SerieBac	Sexe	Stage	Theorie
Moyen	Excellent	Fort	Sciences math..	Homme	Oui	Moyen

Figure 39. Résultat de l'analyse prescriptive

On a utilisé une méthode d'optimisation évolutive, basée sur le modèle. On a le choix d'utiliser l'une des cibles suivantes:

- Minimiser la confiance pour une classe ;
- Maximiser la confiance pour une classe ;
- Se rapprocher le plus possible d'une certaine confiance pour une classe ;
- Minimiser la prédiction de régression ;
- Maximiser la prédiction de régression ;
- Se rapprocher le plus possible d'une certaine prédiction de régression.

On a maximisé la confiance de la classe « Working », pour pouvoir trouvé les valeurs optimales en se basant sur le modèle d'arbre de décision développé.

On a utilisé l'opérateur Prescriptive Analytics pour rechercher les valeurs d'attributs optimales qui optimisent les chances de trouver un travail.
Les valeurs de paramètres par défaut fourniront des résultats raisonnables sans aller aux extrêmes. Mais nous avons effectué des réglages importants. Nous avons défini qu'il s'agissait d'un problème de classification et que nous voulions maximiser la prédiction de "Oui", qui correspond à Working, ça veut dire qu'on va chercher le profil le plus performant qui a le plus de chance pour trouver un travail.

Conclusion

Déterminer les facteurs qui influencent l'employabilité des diplômés donnera aux décideurs une vue claire et des opportunités pour améliorer ce secteur.
Le type de données joue un rôle important dans le choix de l'algorithme de data mining approprié que nous souhaitons appliquer. Dans ce chapitre, on a déterminé quel algorithme qui convenait le mieux aux données de l'employabilité et qui présentait le modèle de prédiction le plus performant. Nous avons appliqué trois algorithmes de classification ; Arbre de décision, Régression logistique et Naïve Bayes, et les résultats ont montré que l'arbre de décision est le meilleur et le plus adapté à la prédiction de l'employabilité par rapport à la régression logistique et à Naïve Bayes. En fait, l'arbre de décision était meilleur dans toutes les métriques utilisées pour évaluer les performances des modèles, la précision, le taux

d’erreur, la statistique de kappa, F-measure, Recall, et la courbe de ROC, à l'exception du temps pour construire le modèle, Naïve Bayes était plus rapide. Après, on a présenté en détail le modèle de l'algorithme d’Arbre de décision. Nous avons ensuite présenté les variables qui jouent un rôle important dans la prédiction de l’employabilité des diplômés.

Nous vivons maintenant dans une ère de Big Data, où les données sont générées chaque jour en gros volumes, en raison de l'utilisation de technologies, dans le chapitre suivant, nous présentons un système intelligent en appliquant le processus du data mining pour la prédiction de l’employabilité dans un environnement Big Data, utilisant l'écosystème Hadoop, on va présenter l’architecture qu’on a utilisé en détails, ainsi que les phases suivies.

Chapitre 4 : Proposition d'un système de prédiction de l'employabilité utilisant des techniques de data mining dans un environnement Big data

Introduction

Les technologies traditionnelles et les systèmes de bases de données actuels ne peuvent pas gérer cette énorme quantité de données et la traiter efficacement. Le Big Data est intervenu pour résoudre ces problèmes et proposer des solutions permettant de traiter et d'analyser efficacement cette énorme quantité de données. La combinaison du Big Data et de data mining peut être très utile et offre aux décideurs des opportunités pour tirer à partir des résultats des solutions pour prédire l'avenir. Dans le chapitre précédent, nous avons présenté un modèle de prédiction de l'employabilité utilisant des algorithmes de machine learning de classification, ainsi que les variables qui jouent un rôle important dans la prédiction de l'employabilité des diplômés, en présentant toutes les phases depuis la définition du problème et la collecte des données jusqu'à la présentation des résultats et du modèle obtenu. Mais cette opération a été manuelle, nous avons pris les données de l'employabilité sous format Excel, et on les a utilisées dans le processus de data mining pour la prédiction.

Nous présentons dans ce chapitre un système de prédiction de l'employabilité (EPS), qui va rendre cette opération dynamique, et en traitant les données dans un environnement Big Data. Ce système traite et analyse les données dans l'écosystème Hadoop à l'aide des différentes technologies proposées par Hadoop. Nous avons présenté l'architecture générale du système que nous avons utilisé, ainsi que ses caractéristiques et les phases du processus, depuis la collecte des données jusqu'à la visualisation des résultats.

I. Big Data, Data mining ET Employabilité

De nos jours, les personnes et les objets sont constamment interconnectés, grâce aux progrès des technologies de la communication. L'utilisation des appareils connectés intelligents, tels que les voitures équipées de capteurs de localisation, les smart-phones et les réseaux sociaux, génère une grande quantité des données que les systèmes de base de données traditionnels ne peuvent pas gérer. Big Data offre des solutions, il peut gérer une très grande quantité de données, structurées ou non, sur une variété de terminaux. Mais le plus gros défi pour le Big Data n'est pas simplement de stocker ces données, mais d'explorer ces données et de les exploiter afin d'extraire des informations et des connaissances précieuses pour des actions futures. C'est pourquoi le data mining est le processus le plus important. Il permet de prendre

des mesures proactives en fonction des modèles générés et des connaissances acquises afin de répondre aux questions problématiques et de prédire l'avenir.

De nombreuses institutions et industries utilisent le Big Data et le data mining pour améliorer leurs institutions et améliorer leurs conditions de vie. L'employabilité représente un problème sérieux pour les diplômés. Ils sont confrontés chaque année à une grande concurrence en matière d'insertion professionnelle, qui est de plus en plus difficile. Il existe de nombreuses explications et causes à cet égard, par exemple la faible performance économique du pays, la structure de l'économie et son système éducatif assume une grande responsabilité, ou peut-être les domaines d'études universitaires qui rendent l'insertion professionnelle un peu difficile.

L'utilisation du Big Data et du data mining permettra de clarifier la vue et de cerner les problèmes, et apportera également des solutions telles que l'identification des déterminants responsables de l'insertion professionnelle des diplômés, que ce soit en raison du programme scolaire du diplômé ou du marché du travail, ou le domaine d'études choisi par les diplômés. Répondre à de telles questions pourrait être utile aux diplômés ainsi qu'aux chercheurs et aux autorités publiques pour mieux évaluer le système éducatif et la qualité de la formation et procéder aux ajustements nécessaires.

II. Le système proposé: Système de prédiction de l'employabilité (EPS)

1. L'environnement de travail

On a utilisé cloudera-quickstart-vm.5.8.0-0-vmware pour l'environnement Big Data Hadoop. Cloudera [97] est la distribution la plus complète, testée et populaire d'Apache Hadoop et des projets associés. Elle fournit les éléments clés de Hadoop, stockage évolutif et traitement distribuée, avec une interface utilisateur Web. Cloudera est une source ouverte sous licence Apache, la version de Hadoop installé sur Cloudera qu'on a utilisé est : Hadoop 2.6.0-cdh5.8.0 comme présenté dans les figures 36 et 37 ci-dessous.

Figure 40. L'environnement de travail: Cloudera

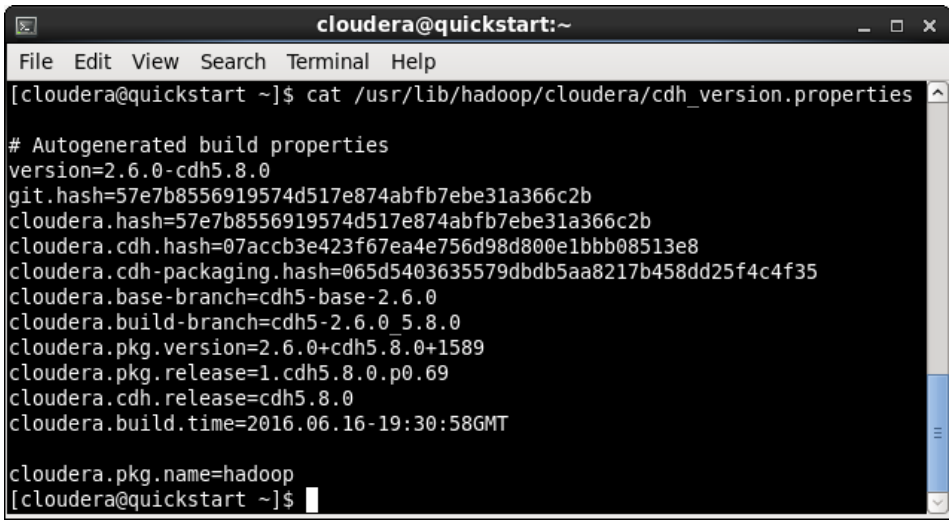

Figure 41. La version de Hadoop

2. Outils de développement

2.1. Php :

Hypertext Preprocessor [98], plus connu sous son code PHP, est un langage de programmation libre, principalement utilisé pour produire des pages Web dynamique via un serveur HTTP, mais qui peut également être interprété n'importe quel langage interprété de façon locale. PHP est un langage impératif orienté objet.

2.2. Mysql :

MySQL [99] est un système de gestion de bases de données relationnelles. Il est distribué sous une double licence GPL et propriétaire.

9. L’écosystème Hadoop

Hadoop [100] est l’un des moteurs de la révolution du Big Data. Il permet le traitement distribué de grands ensembles de données sur des clusters de serveurs standard. Il est conçu pour passer d’un serveur unique à des milliers de machines avec un degré de tolérance aux pannes très élevé. Plutôt que de compter sur du matériel haut de gamme, la résilience de ces clusters provient de la capacité du logiciel à détecter et à gérer les pannes.

Hadoop a changé les données et la dynamique de l'informatique à grande échelle en permettant de construire des capacités informatiques élevées à un coût très bas.

Les caractéristiques [101] les plus importantes de Hadoop:

- **Évolutif :**

De nouveaux nœuds peuvent être ajoutés si nécessaire sans avoir besoin de changer les formats de données, le mode de chargement des données, l'écriture des tâches ou les applications sur le dessus.

- **Rentable :**

Hadoop apporte un traitement massivement parallèle à des serveurs standard. Il en résulte une économie significative sur le coût par péta-octet de stockage, ce qui rend abordable la modélisation de jeux de données complets.

- **Flexible :**

Hadoop est sans schéma et peut absorber tout type de données, structurées ou non, à partir de plusieurs sources. Les données provenant de sources multiples peuvent être jointes et agrégées de manière arbitraire, permettant ainsi des analyses plus approfondies que n'importe quel système.

- **Tolérance de panne :**

Si le cluster perd un nœud, le système est capable de rediriger le travail vers un autre emplacement des données et poursuit le traitement.

Les types de données [102] générés et stockés, par exemple, si les données codent des informations vidéo, images, audio ou texte, numériques, ci-dessous les différent types de données que Hadoop peut gérer :

- **Vidéo:**

Communications et médias, gouvernement, éducation, santé, etc.

- **Audio:**

Communications et médias, gouvernement, éducation, etc.

- **Image:**

Soins de santé, médias, etc.

- **Texte, nombres:**

Services bancaires, assurances, services de valeurs mobilières et d'investissement, commerce de détail, commerce de gros, services professionnels, soins de santé, transports, communication et médias, services publics, gouvernement, etc.

Hadoop consiste en deux grandes parties principales [103]:

- Stockage des données : **HDFS** (**H**adoop **D**istributed **F**ile **S**ystem)
- Traitement des données : **MapReduce**

9.1. HDFS

Hadoop est un Framework Apache Open Source qui prend en charge le traitement de grands ensembles de données dans un environnement informatique distribué. Il se compose de deux parties principales: HDFS [104] qui nous permet de stocker des données en parallèle dans un cluster et MapReduce qui traite les données en parallèle. HDFS peut gérer une grande quantité de données et offre un accès facile et rapide aux données, en parallèle et de manière tolérante aux pannes.

Le système de fichiers distribué Hadoop (HDFS) est un système de fichiers qui couvre tous les nœuds d'un cluster pour le stockage de données. Il relie les systèmes de fichiers sur de nombreux nœuds locaux pour les transformer en un seul et même gros système de fichiers.

Dans les modèles de systèmes de base de données distribués, les données sont logiquement intégrées. Le stockage et le traitement des données sont physiquement répartis sur plusieurs nœuds dans un environnement de cluster.

Les avantages du système de bases de données distribuées incluent l'évolutivité matérielle, la réplication des données entre les nœuds du cluster, les transactions simultanées, la disponibilité en cas de défaillance de nœud, la base de données sera toujours en ligne et les améliorations de performances dues au matériel distribué.

HDFS est basé sur une architecture Master/slave:

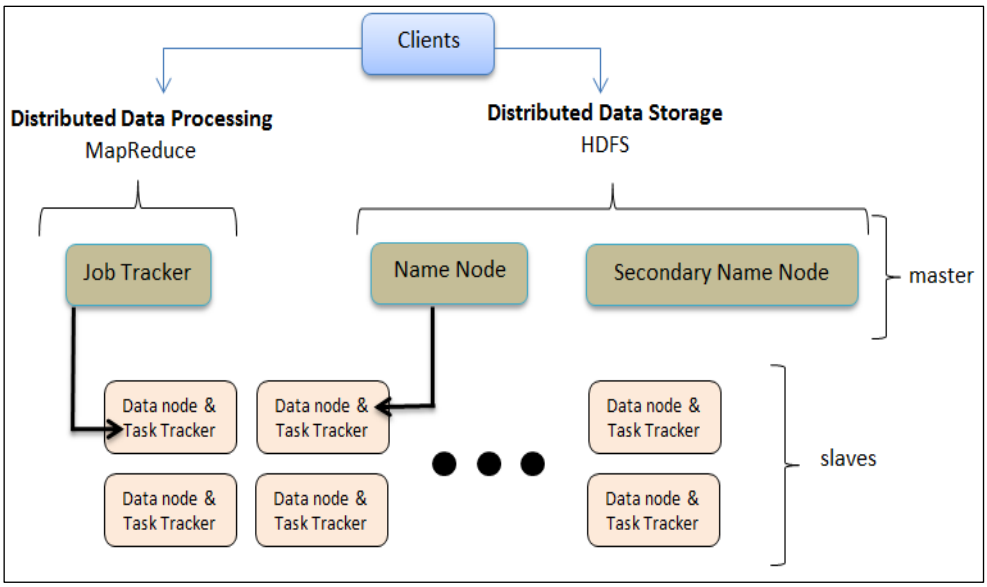

Figure 42. Architecture de HDFS

Les Master Nodes supervisent les deux cœurs de Hadoop: stocker des données en parallèle dans HDFS et traiter toutes ces données à l'aide de MapReduce. Name Node supervise et coordonne la fonction de stockage de données dans HDFS, tandis que Job Tracker supervise et coordonne le traitement parallèle des données à l'aide de MapReduce.

Les Slave Nodes constituent la grande majorité des machines et effectuent tout le travail de stockage des données et d'exécution des calculs. Les Slave Nodes exécutent des nœuds de données et le démon Task Tracker qui reçoit des instructions des Master Nodes. Le démon Data Node est un esclave de Name Node et le démon Track Tracker est un esclave de Job Tracker.

9.2. MapReduce

MapReduce [105] est un paradigme de programmation permettant le traitement de données distribuées sur plusieurs serveurs d'un cluster. Le MapReduce est devenu dominant en matière de traitement par lots. L'idée principale est de diviser les jeux de données d'entrée en morceaux traités de manière complètement parallèle.

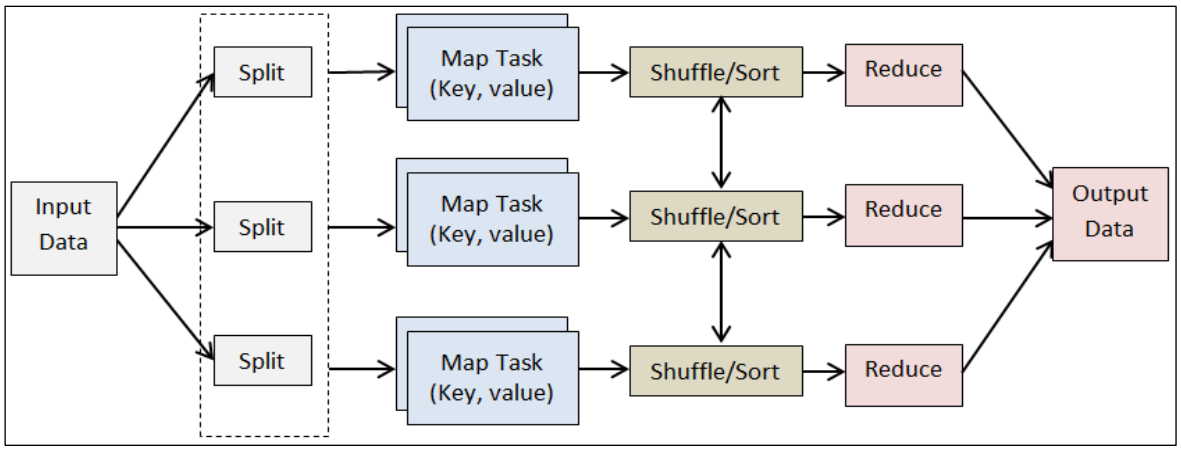

Figure 43. Architecture de travail de MapReduce

Il comprend deux fonctions principales, Map et Reduce. Le mappeur traite les données et crée plusieurs petits morceaux de données de paires clé / valeur qui sont traitées en parallèle. Les sorties de la fonction Mappeur seront triées et remaniées, ce qui nous mènera à la phase suivante, la phase de réduction et ici les données sont agrégées pour renvoyer le résultat.

Dans MapReduce, il est facile de faire évoluer le traitement des données sur plusieurs nœuds informatiques. Dans ce modèle, les primitives de traitement de données sont appelées mappeurs et réducteurs. L'écriture de l'application dans le formulaire MapReduce n'est pas si facile, mais une fois écrite, elle permet de faire évoluer l'application pour qu'elle s'exécute sur des milliers, voire des dizaines de milliers de machines d'un cluster.

Phase de mappage, les données d'entrée sont stockées sous la forme de fichiers dans le système de fichiers Hadoop (HDFS). Ce fichier d'entrée est ensuite transmis ligne par ligne à la fonction mappeur. Ces données sont ensuite traitées par le mappeur et plusieurs petits morceaux de données sont créés.

Les nœuds de travail de la phase aléatoire redistribuent les données en fonction des clés de sortie (produites par la fonction Map), de sorte que toutes les données appartenant à une clé se trouvent sur le même nœud de travail. Les nœuds de réduction de phase Worker traitent chaque groupe de données de sortie, par clé, en parallèle. Le réducteur prend la sortie du mappeur et du processus. Il génère ainsi un nouvel ensemble de sorties qui seront stockées dans HDFS.

9.3. Hbase:

Apache HBase [106] fournit un accès aléatoire en temps réel aux données de Hadoop. Il a été créé pour héberger de très grandes tables, ce qui en fait un excellent choix pour stocker des données multi-structurées ou fragmentées. HBase est accessible via des interfaces de programmation d'applications (API) telles que Thrift, Java et REST (Representational State Transfer). Ces API ne possèdent pas leurs propres langages de requête ou de script. Par défaut, HBase dépend entièrement d'une instance de ZooKeeper. Les caractéristiques d'Hbase sont la tolérance aux pannes et la rapidité par rapport aux autres technologies.

9.4. Apache HIVE

Apache HIVE [107] est un système de stockage de données proposé par Hadoop utilisant le langage HIVEQL [108] pour interroger les données stockées dans Hadoop et faciliter les requêtes ad hoc, l'agrégation et l'analyse de grands volumes de données stockées dans les systèmes de fichiers distribués Hadoop. L'apprentissage de HIVEQL est simple pour les utilisateurs familiarisés avec le langage SQL. Le traitement des données stockées dans HDFS a besoin de Map Reduce. La programmation d'une réduction de carte n'est pas simple. HIVE peut également convertir les requêtes HIVEQL en travaux MapReduce exécutables sur Apache Tez, qui est un cadre d'exécution sur Hadoop.

9.5. Apache IMPALA

En plus d'Apache HIVE, impala [109] propose également une syntaxe SQL permettant d'envoyer des requêtes SQL interactives directement sur les données Apache Hadoop stockées sur HDFS. Il fournit une plate-forme unifiée et il est familier pour les requêtes par lots ou en temps réel. Apache impala offre de nombreux avantages. Outre l'interface SQL bien connue, il est également possible d'interroger une grande quantité de données sur Hadoop. Également la possibilité d'échanger des données entre les tables Impala et HIVE pour la lecture et l'écriture, offrant une analyse simple des données produites par Hive.

9.6. Apache SOLR

Apache SOLR [110] permet à l'utilisateur Hadoop d'explorer et de découvrir les données stockées dans HDFS, de les visualiser et d'offrir une recherche dynamique de tableaux de bord. Il est optimisé pour les gros volumes de données, ce qui le rend capable d'intégrer de manière intelligente une grande quantité de données à l'utilisateur, en quelques millisecondes.

9.7. HUE

HUE [111] offre une interface interactive pour analyser les données stockées dans Hadoop, offrant de nombreuses fonctionnalités telles qu'un éditeur pour les requêtes HIVE, des requêtes Impala, un navigateur pour le concepteur de jobs, une interface OOZIE pour planifier des travaux, etc. HUE offre un autre moyen d'interagir avec Hadoop sans l'invite de commande pour la plupart des activités Hadoop. HUE est développé par Cloudera.

9.8. Mahout

Apache Mahout [112] est un projet open source principalement utilisé dans la production d'algorithmes d'apprentissage machine évolutifs. Il met en œuvre des techniques d'apprentissage automatique populaires telles que: recommandation, classification, mise en cluster. Il est divisé en quatre groupes principaux: collectif, filtrage, catégorisation, regroupement et extraction de modèles fréquents parallèles. La bibliothèque Mahout appartient au sous-ensemble qui peut être exécuté en mode distribué et peut être exécuté par MapReduce.

9.9. Spark

Apache Spark [113] est une technologie de calcul en cluster ultra-rapide, conçue pour un calcul rapide. Il est basé sur Hadoop MapReduce et étend le modèle MapReduce afin de l'utiliser efficacement pour plusieurs types de calcul, ce qui inclut les requêtes interactives et le traitement de flux. La principale caractéristique de Spark est son traitement en mémoire qui augmente la vitesse de traitement d'une application.

9.10. Apache OOZIE

Apache OOZIE [114] est un planificateur de flux de travail utilisé pour planifier et gérer les travaux Hadoop, prenant en charge différents types de travaux Hadoop (HIVE, SQOOP, Java MapReduce, etc.). Toutes sortes de programmes impliqués dans le cluster Hadoop peuvent être organisés dans un ordre d'exécution spécifique à l'aide de OOZIE, offrant également un mécanisme permettant d'exécuter des travaux à un moment donné, selon un calendrier prédéfini.

9.11. Apache SQOOP

Apache SQOOP [115] joue un rôle important dans l'écosystème Hadoop. Comme nous le savons, les applications et les sites Web fonctionnent généralement avec des bases de données relationnelles, ce qui en fait l'une des sources les plus importantes générant des données volumineuses. Apache SQOOP fournit une interaction entre les bases de données relationnelles et HDFS. Il a été conçu pour transférer des données entre HDFS et des bases de données relationnelles telles que MySQL, Oracle, TeraData, SQLITE, etc.

3. L'architecture globale du système

Dans cette section, nous présentons l'architecture globale de notre système, en commençant par l'insertion des données dans le système, jusqu'à ce que les données soient prêtes pour le data mining. Nous allons décrire l'architecture utilisée en détail et les phases dans les sections suivantes de ce chapitre.

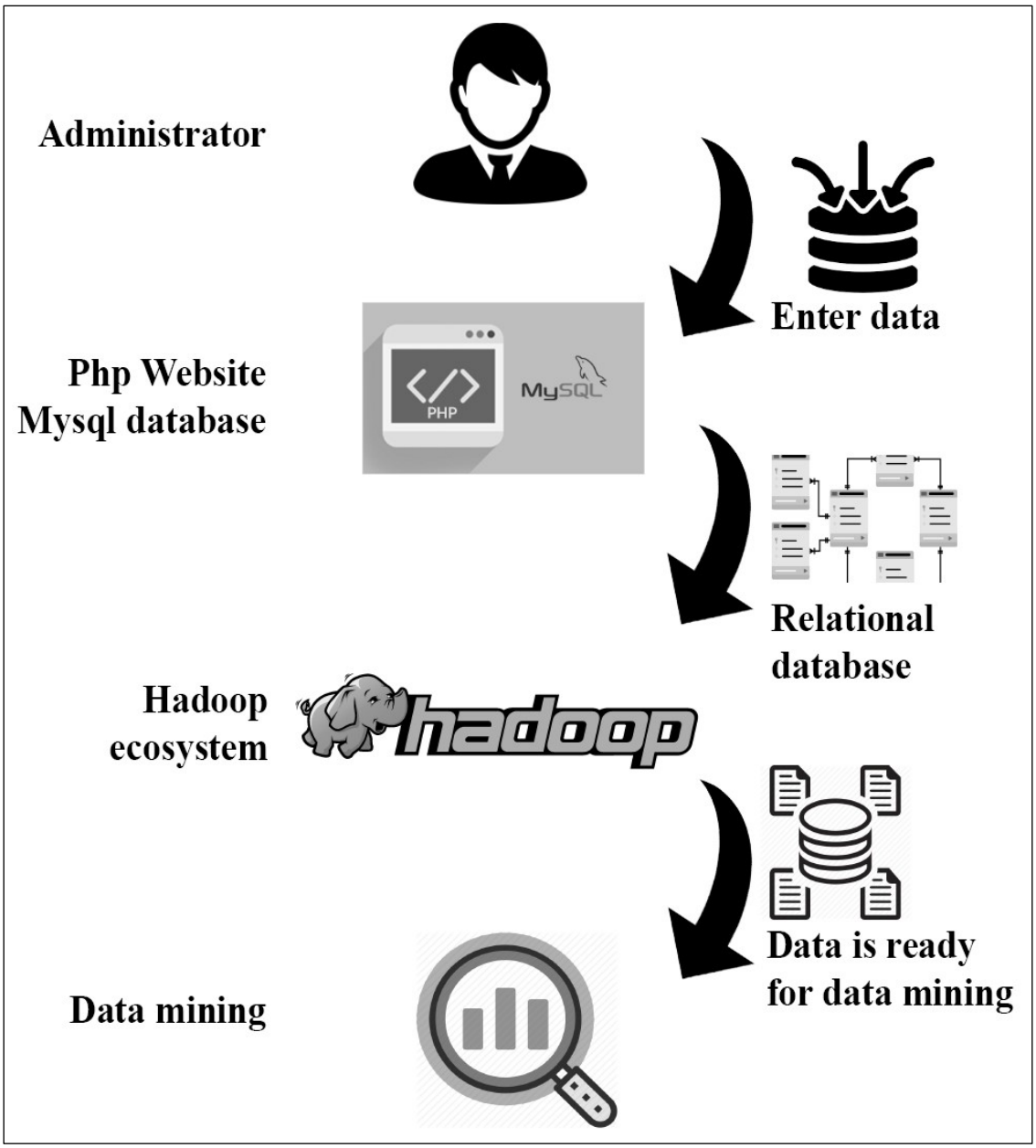

Figure 44. L'architecture globale du système

4. Les caractéristiques du système

Notre système de prédiction de l'employabilité (EPS) offre de nombreuses opportunités aux décideurs afin d'améliorer le domaine de l'employabilité. Dans le chapitre précédent, nous avons présenté un modèle de prédiction de l'employabilité utilisant des algorithmes de classification, ainsi que les variables qui jouent un rôle important dans la prédiction de l'employabilité des diplômés, en présentant toutes les phases depuis la définition du problème et la collecte des données jusqu'à la présentation des résultats et du modèle obtenu. Mais cette opération a été manuelle, nous avons pris les données de l'employabilité sous format Excel, et

on les a utilisées dans le processus de data mining pour la prédiction. Notre système de prédiction de l'employabilité (EPS), va rendre cette opération dynamique, et en traitant les données dans un environnement Big Data, ci-dessous les caractéristiques de notre système en détails.

Le système sera capable de:

- Fournir à la base de données les données nécessaires (baccalauréat, domaine, diplôme, université, diplômés, etc.);
- Transférer les données de la base de données MYSQL dans l'écosystème Hadoop;
- Interroger et traiter les données dans Hadoop en utilisant les différentes technologies de l'écosystème Hadoop;
- Visualiser les données et créer des tableaux de bord dans Hadoop afin de comprendre et visualiser les données;
- Appliquer le processus de data mining pour extraire des informations précieuses à partir des données;
- Partager les résultats et les graphiques générés par les analyses de data mining;
- Proposer un modèle de prédiction de l'employabilité;
- Proposer les variables qui ont le plus d'impact sur l'employabilité;
- Prédire si les diplômés seront classés comme « Working » ou « NotWorking ».

5. L'architecture du système utilisé pour la prédiction de l'employabilité

Nous allons maintenant présenter l'architecture utilisée de notre système. Le système est basé sur l'écosystème Hadoop. Hadoop propose de nombreux outils et technologies facilitant le traitement et l'analyse des données. Mais, choisir la bonne technologie pour la bonne tâche n'est pas aussi facile qu'il semble, nous présentons ici l'architecture du système, ainsi que tous les outils et technologies utilisés, en expliquant chaque technologie de côté et son fonctionnement.

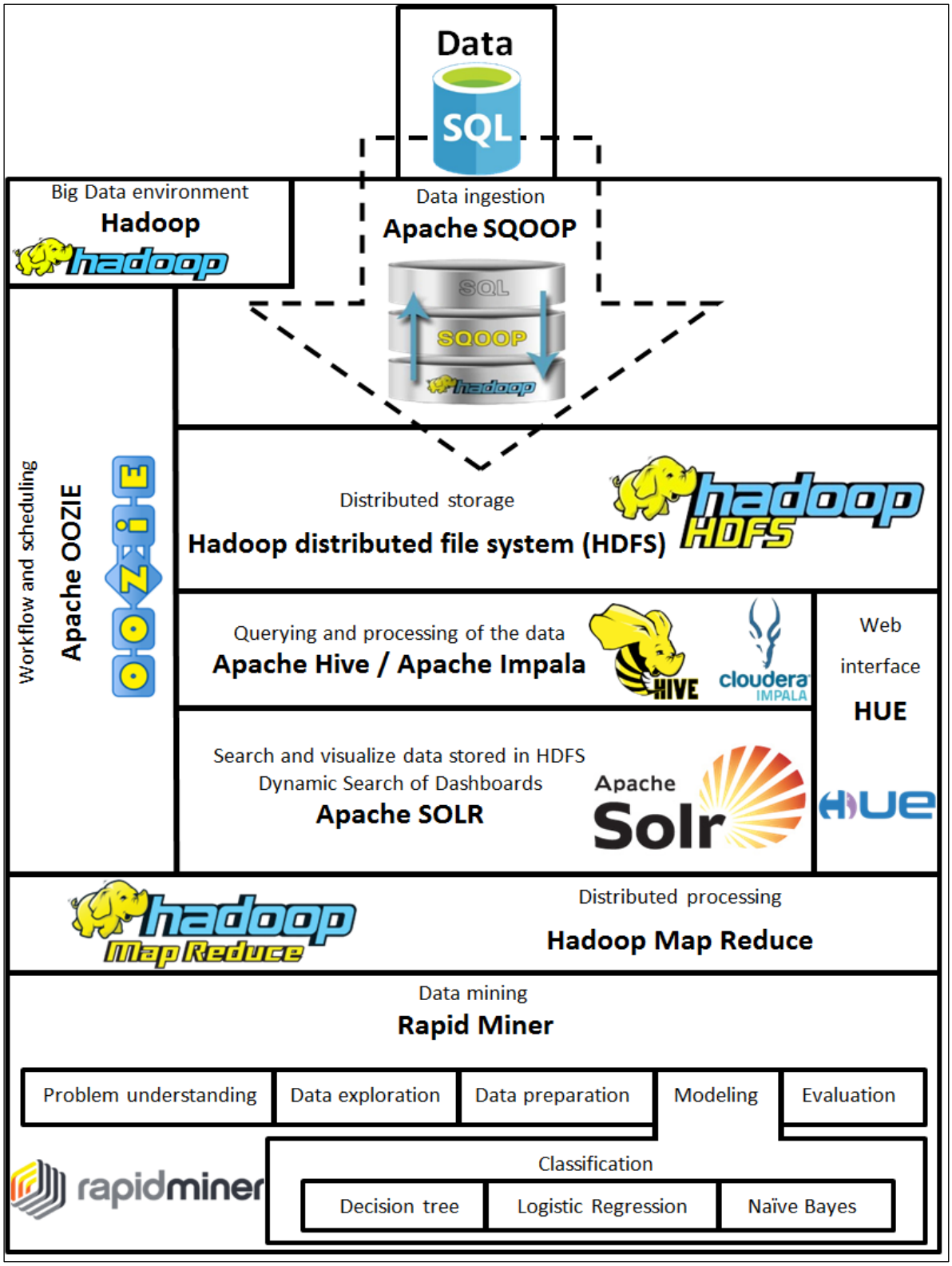

Figure 45. L'architecture générale du système: Système de prédiction de l'employabilité (EPS)

6. Les phases du processus de notre système

Dans cette partie, nous présenterons les différentes phases de notre système en détail, en décrivant chaque phase séparément. Nous présentons tout d'abord le workflow proposé et en dessous détaillant les phases dans la figure 47.

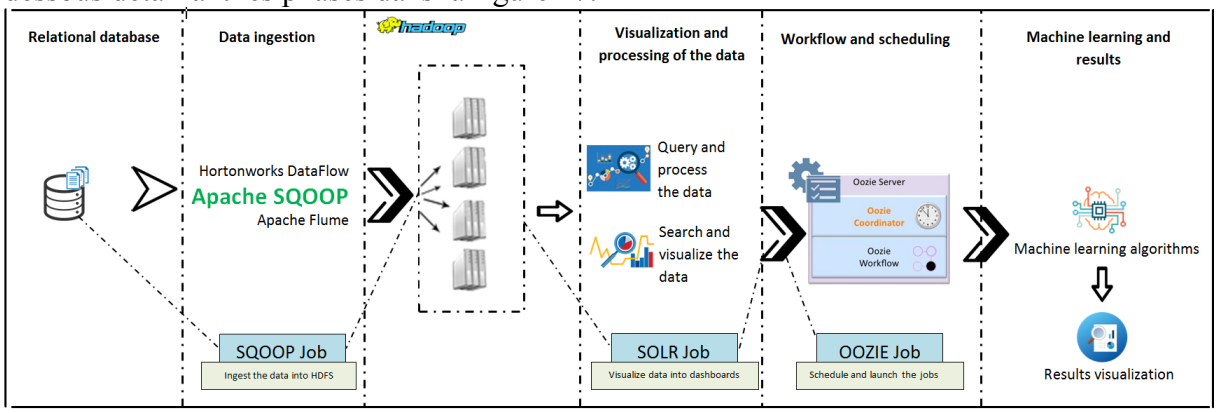

Figure 46. Proposition d'un workflow pour l'intégration et l'application des algorithmes de machine learning dans un environnement big data

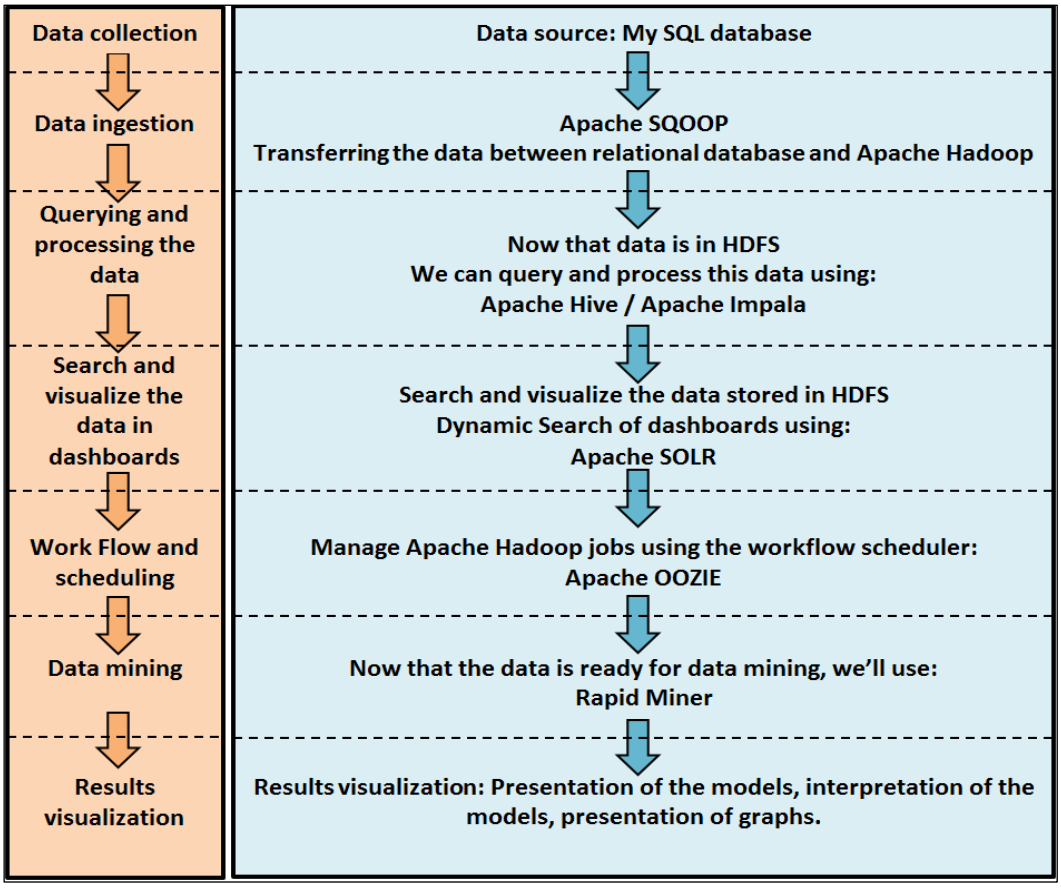

Figure 47. Système de prédiction de l'employabilité (EPS): Le processus

6.1. La collecte des données

Dans cette phase, nous collectons les données, les données utilisées dans cette étude sont recueillies à partir d'une enquête sur l'employabilité conduite par l'université Hassan 1er en 2016 en partenariat avec le Bureau national de l'évaluation (NEO), sous l'égide du Conseil supérieur de l'éducation, de la formation et de la recherche scientifique, ces données vont être insérer en utilisant un site web avec MYSQL comme système de gestion de base de données relationnelle (SGBDR). Dans la phase suivante, nous intégrerons les données de MYSQL dans le système de fichiers distribués Hadoop (HDFS) à l'aide d'Apache SQOOP. La figure ci-dessous représente l'interface dans laquelle les données vont être insérer dans la base de données relationnelle MYSQL.

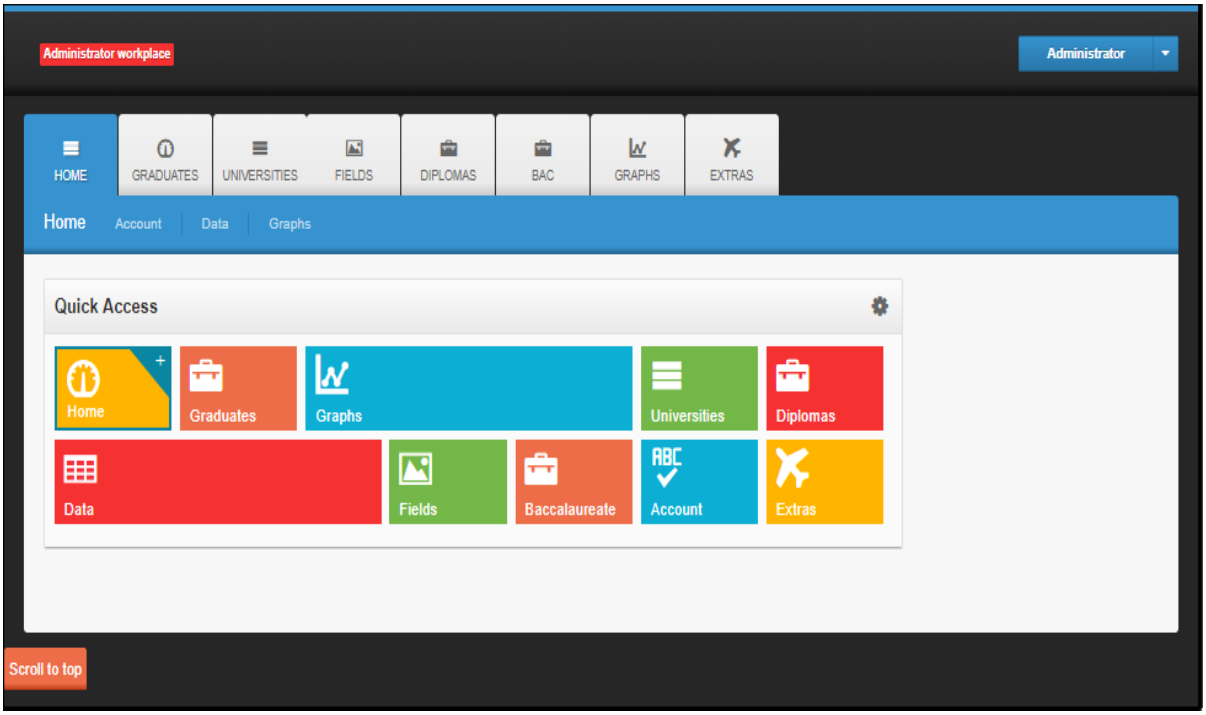

Figure 48. L'interface utilisateur du système

6.2. Ingestion de données

L'un des avantages de l'écosystème Hadoop, il permet l'ingestion de données à partir de différentes sources, il y a différentes outils, cela dépend du type des données, les données dont nous disposons sont dans un format de base de données relationnelles, donc on va utiliser SQOOP, qui signifie SQL en HADOOP.

Dans cette phase, nous devons mettre les données dans HDFS en utilisant SQOOP. Tout d'abord, nous créons un job SQOOP, puis nous l'avons lancé comme présenté dans les figures 44 et 45 ci-dessous.

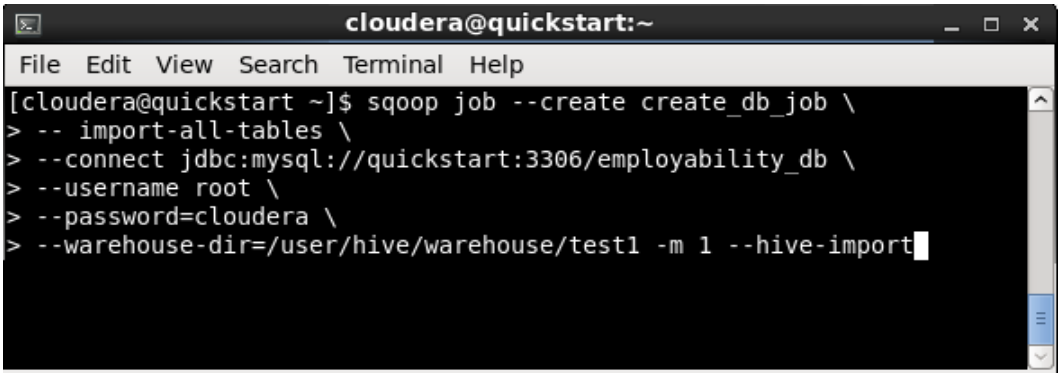

Figure 49. La création du job de SQOOP

Et ensuite, nous lançons le job en exécutant la commande suivante:

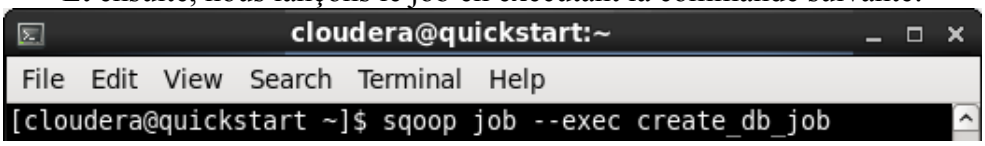

Figure 50. L'exécution du job de SQOOP

Ce graphe représente le succès de l'exécution du job MapReduce.

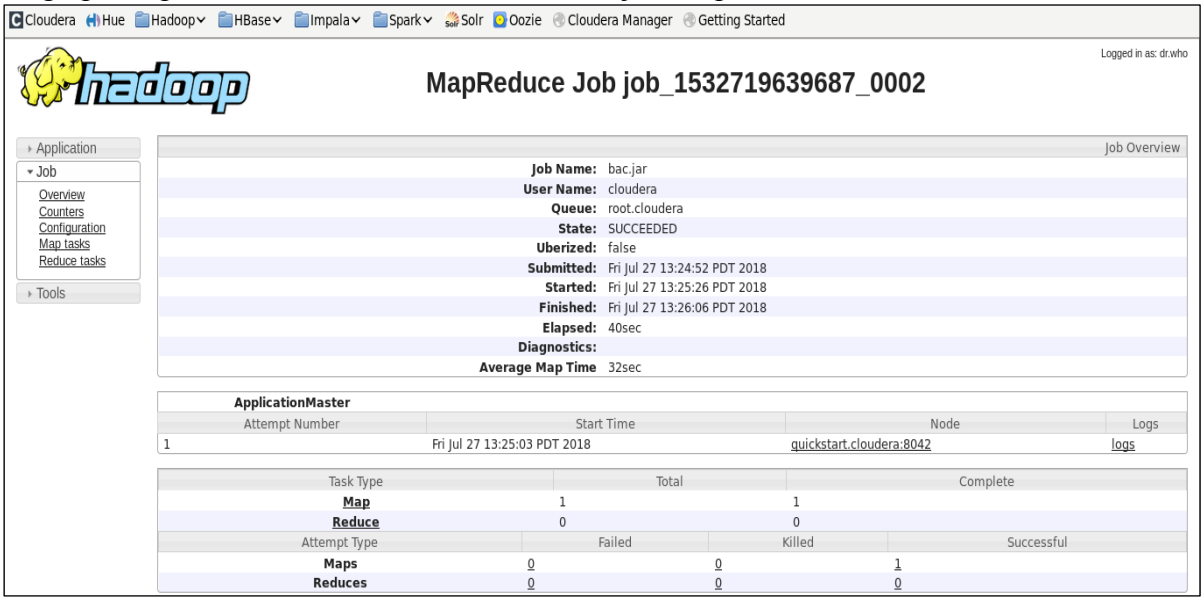

Figure 51. Le succès de l'exécution du job MapReduce

Après l'exécution du job, nous nous assurons que les tables de notre base de données se trouvent bien dans le système de fichier de Hadoop HDFS, comme indiqué dans le graphe ci-dessous.

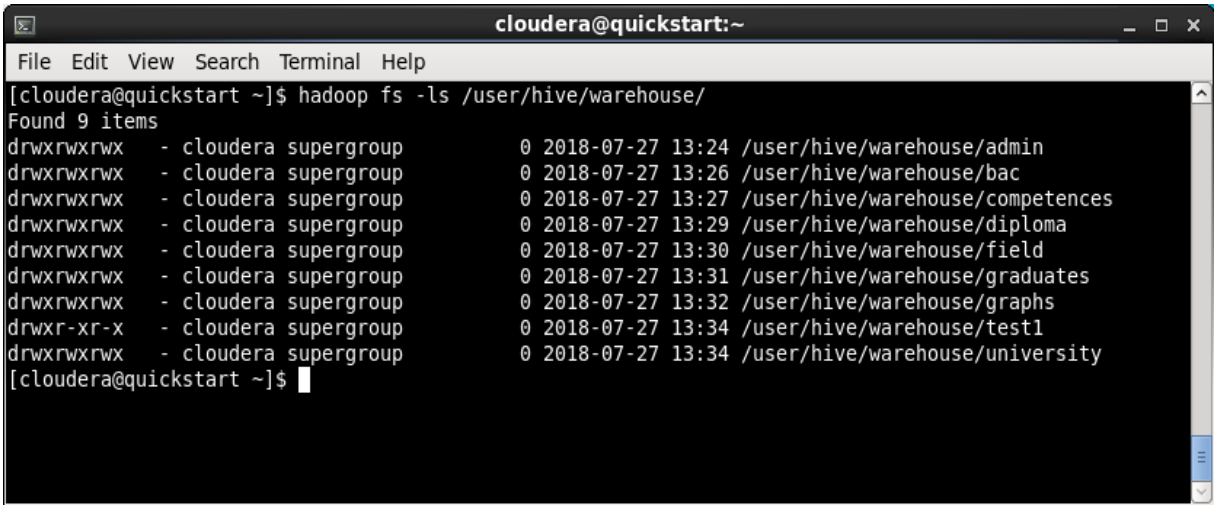

Figure 52. Les tables de la base de données dans le système de fichier de Hadoop HDFS

Maintenant, les données sont transférées de MYSQL et stockées dans HDFS, elles sont prêtes à être interrogées et traitées dans l'écosystème Hadoop.

6.3. Interrogation et traitement des données

Hadoop propose de nombreux outils et technologies facilitant le traitement des données. Nos données sont maintenant dans HDFS, nous pouvons traiter ces données et même envoyer des requêtes SQL sur Hadoop, afin de voir nos données et de rechercher les problèmes de nos données. Nous pouvons utiliser Hive Shell ou l'interface HUE comme présenté dans les figures ci-dessous.

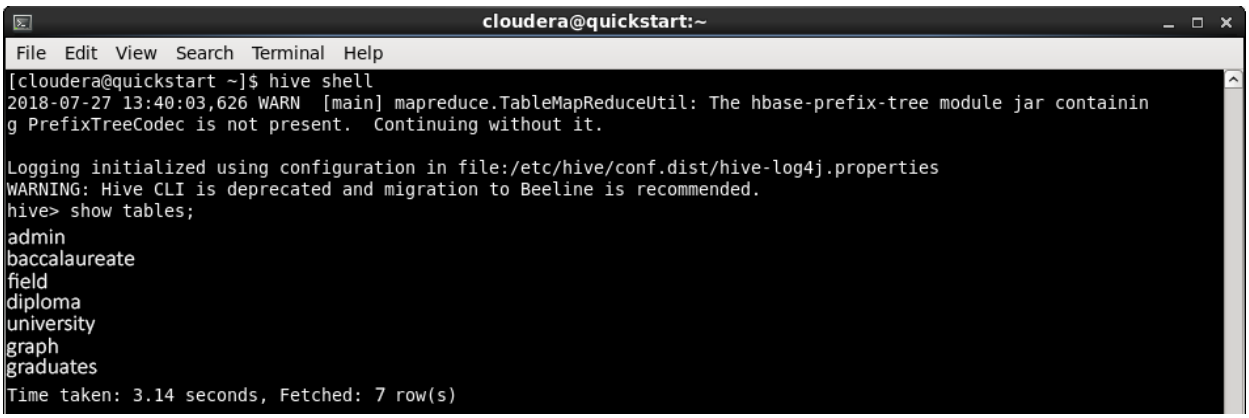

Figure 53. Liste des tables de la base de données importées dans HDFS par SQOOP

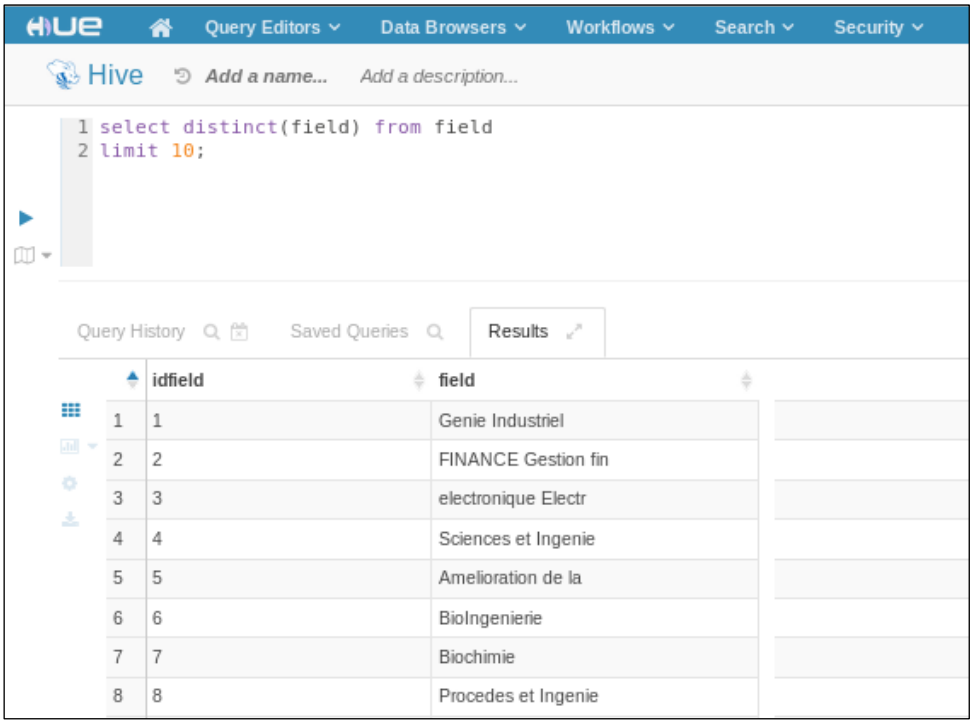

Figure 54. Interrogation de la base de données avec HIVE utilisant l'interface de HUE

6.4. Rechercher et visualiser les données dans des tableaux de bord

Au cours de cette phase, nous allons explorer et découvrir les données stockées dans HDFS à l'aide des graphes et de tableaux de bord avec Apache SOLR. Nous avons utilisé l'interface HUE qui facilite et accélère la création des tableaux de bord. Ici, nous avons présenté dans la figure 51 un exemple de tableau de bord créé visualisant les données dans différents graphes, mais auparavant, Apache Solr offre une phase de préparation des données comme décrit dans la figure 50 ci-dessous.

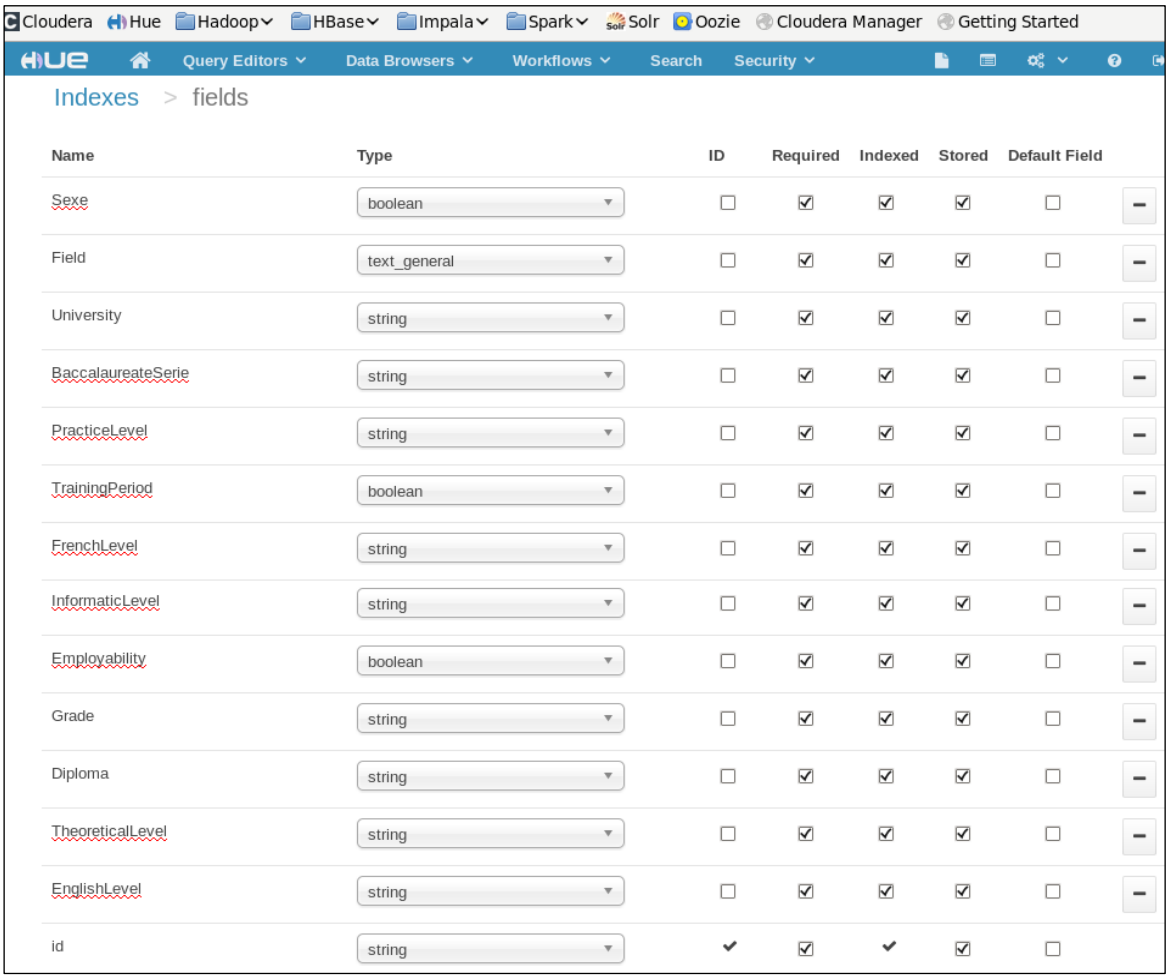

Figure 55. La phase de préparation des données d'Apache SOLR

Figure 56. Tableau de bord visualisant les données sur Apache SOLR

6.5. Flux de travail et planification

Au cours de cette phase, nous allons planifier l'exécution des jobs créés à l'aide d'Apache OOZIE. Nous prendrons par exemple SQOOP, qui est la première étape pour importer les données de la base de données MYSQL dans Hadoop HDFS. Supposons que nous voulions lancer le job SQOOP chaque semaine le lundi à 06h30, voici un exemple ci-dessous dans la figure 52 présentant la création du coordinateur en donnant le nom du coordinateur, le nom du job à exécuter ainsi que le moment de l'exécution, puis nous présentons le succès de l'exécution du job dans la figure 53.

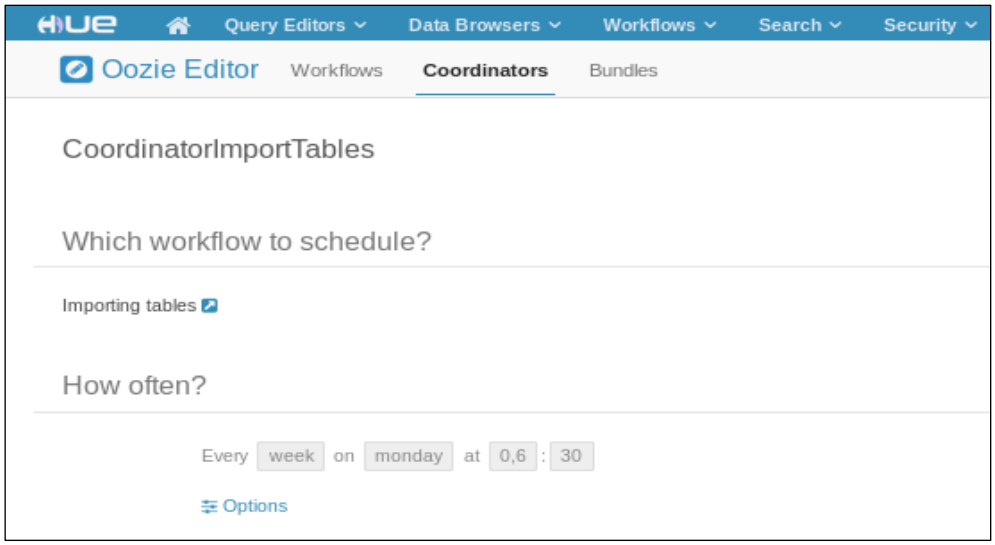

Figure 57. Création d'un coordinateur utilisant Apache OOZIE

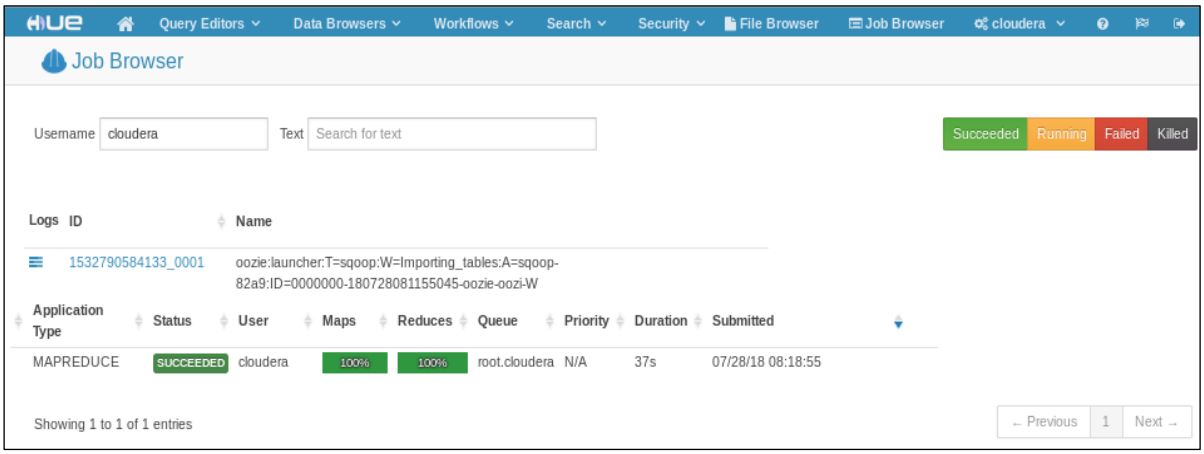

Figure 58. Succession de l'exécution du job d'OOZIE

6.6. Data mining

La phase la plus importante du processus est le data mining. Au cours de cette phase, nous appliquerons le processus de data mining à l'aide de Rapid Miner Studio Educational Version 8.1.000. Le processus comprend les phases suivantes: compréhension du problème et collecte des données, exploration ou compréhension des données, préparation, modélisation et évaluation des données.

Une fois toutes ces phases terminées, nous pourrons extraire les connaissances et visualiser les résultats lors de la prochaine phase du processus.

6.7. Visualisation des résultats

Il s'agit de la phase finale du processus, où la présentation et l'interprétation des modèles, la présentation des graphes, les variables qui ont un impact sur la variable objective, ces résultats aideront les décideurs et leur donneront la possibilité de prendre des mesures proactives de l'avenir.

Conclusion

Dans ce chapitre, nous avons présenté un système intelligent utilisant les techniques de data mining sur les données d'employabilité, dans un environnement Big Data, dans lequel nous avons utilisé l'écosystème Hadoop dans la distribution Cloudera, et pour le data mining, nous avons utilisé Rapid Miner Studio Educational Version 8.1.000. Nous avons d'abord présenté les caractéristiques du système, puis l'architecture générale du système, les outils et les technologies utilisées, et enfin le processus du système présentant les phases détaillées à partir de la collecte des données jusqu'à la visualisation des résultats.

Conclusion générale et perspectives

Notre thèse traite un sujet d'actualité : l'employabilité et les techniques de data mining et big data. La question de l'employabilité est un problème majeur dans plusieurs pays, et l'utilisation de techniques de data mining et big data va permettre de proposer des solutions afin d'améliorer ce domaine.

L'utilisation du data mining et des algorithmes de machine learning permettra de clarifier la vue et de cerner les problèmes, tout en présentant des solutions telles que l'identification des déterminants responsables de l'insertion professionnelle des diplômés. C'est peut-être à cause du programme scolaire du diplômé, ou peut-être du marché du travail, ou du domaine d'études choisi par les diplômés. Répondre à de telles questions pourrait être utile aux diplômés ainsi qu'aux chercheurs et aux autorités publiques pour mieux évaluer le système et la qualité de la formation et procéder aux ajustements nécessaires.

Un certain nombre de mesures d'orientation, d'information et de soutien doivent être adoptées pour offrir aux diplômés de bonnes solutions leur permettant de s'orienter facilement et de s'infiltrer dans le monde du travail.

Les résultats du data mining sur les données de l'employabilité peuvent permettre de prendre des mesures efficaces pour améliorer l'intégration professionnelle des diplômés, ainsi que de valoriser les ressources humaines et le développement économique et social du pays. Clarifier le débat sur la problématique de l'employabilité et de l'insertion professionnelle des diplômés, et proposer un certain nombre de recommandations concernant les mesures éventuelles à prendre et les différentes ressources à mobiliser afin de renforcer l'employabilité et l'intégration professionnelle des diplômés au Maroc. Et c'est précisément le but de l'application du data mining sur des données d'employabilité.

Après avoir présenté les techniques de data mining et du big data, tout en indiquant la nécessité des analyses du data mining, des algorithmes de machine learning et leurs objectifs, nous avons par la suite présenté premièrement en utilisant des graphes et des statistiques la situation de l'employabilité au Maroc. Et deuxièmement, le rôle et la nécessité de ces technologies dans le domaine de l'employabilité, et pourquoi utilisé le data mining dans le domaine de l'employabilité, et quelles sont les avantages de ces analyses. Après, on a présenté une étude expérimentale comparant divers algorithmes de machine learning de classification sur des données d'employabilité au Maroc: arbre de décision, régression logistique et Naïve

Bayes. L'objectif en premier est de choisir l'algorithme le plus efficace et le mieux adapté aux données d'employabilité qui présente le meilleur modèle. Et après cette étape la présentation du modèle d'employabilité et des variables jouant un rôle important dans la prédiction de l'employabilité des diplômés, et enfin la visualisation des résultats. Et finalement, on a présenté un système de prédiction de l'employabilité (EPS), dans un environnement Big Data. Finalement, on a conclu que l'application des techniques de data mining est très prometteuse dans le domaine de l'employabilité, elle va permettre aux futurs diplômés de bénéficier du progrès technologique pour assurer leur employabilité, on a conclu aussi qu'une énorme quantité de données ne veut pas dire nécessairement qu'on développera un bon modèle de prédiction. La qualité de prédiction dépend du modèle de prédiction certes, mais avec des données précises bien collectées et bien nettoyées pour assurer des résultats réels et prendre des décisions logiques et concrètes. On vit maintenant dans l'ère du big data, ignorer ce domaine sera comme prendre un pas en arrière. On a pu appliquer les techniques du data mining et les algorithmes de machine learning dans un environnement big data, on a utilisé l'écosystème Hadoop, cela nous a permet de bénéficier des avantages de Hadoop en terme de traitement et d'analyse des données.

Comme perspectives aux travaux présentés ici :

- Nous comptons principalement améliorer la précision du modèle en introduisant d'autres données pour que le modèle puisse apprendre davantage sur ces données et prédire l'employabilité avec plus de précision ;
- L'amélioration de la qualité des données et améliorer les performances des phases de prétraitement (Nettoyage, Filtrage, ...) ;
- Analyse et traitement des données d'employabilité utilisant la fouille des données par Deep Learning.
- On souhaite aussi étendre la base de données utilisée et prendre en compte d'autres variables qui peuvent avoir un impact sur l'employabilité ;
- A la fin, on souhaite que les travaux réalisés soient pris en considération et soient appliquer dans le domaine de l'employabilité au Maroc afin de pouvoir améliorer ce domaine et que les futurs diplômés puissent bénéficier des améliorations.

Liste des publications

1. SAOUABI Mohamed et EZZATI Abdellah, Data Mining Approach for Employability prediction in Morocco, **Springer Books Series: Embedded Systems and artificial Intelligence Proceedings of ESAI 2019 of 1st International Conference on Embedded Systems and Artificial Intelligence ESAI'19 in** ISBN: 978-981-15-0947-6, 2-3 Mai 2019, Faculté de médecine et de pharmacie de fes, Fes, Maroc.
2. SAOUABI Mohamed et EZZATI Abdellah, A data mining process using classification techniques for employability prediction, **Indonesian Journal of Electrical Engineering and Computer Science** ISSN: 2502-4752, Volume 14, (2), (2019), 1025-1029.
3. SAOUABI Mohamed et EZZATI Abdellah, Prediction model for employability in Morocco using data mining techniques, **Journal of Engineering and Applied Sciences** ISSN: 1816-949X, Volume 14, (5), (2019), 1690-1694.
4. SAOUABI Mohamed et EZZATI Abdellah, Proposition of an employability prediction system using data mining techniques in a big data environment, **International Journal of Mathematics and Computer Science** ISSN: 1814-0432, Volume 14, (2), (2019), 411-424.
5. SAOUABI Mohamed et EZZATI Abdellah, Data mining classification Algorithms, **International Journal of Mathematics and Computer Science** ISSN: 1814-0432, Volume 15, (1), (2020), 389-394.
6. SAOUABI Mohamed et EZZATI Abdellah, Data Mining Techniques for Predicting Employability in Morocco**, International Journal of Engineering & Technology** ISSN: 2227-524X, Volume 7, (4.32), (2018), 17-20.
7. SAOUABI Mohamed et EZZATI Abdellah, Data Mining Techniques for Predicting Employability in Morocco, **International Conference on Communication, Management and Information Technology ICCMIT'18**, 2-4 Avril 2018, Universidad Politecnica de Madrid, Madrid, Espagne.
8. SAOUABI Mohamed et EZZATI Abdellah, A comparative between Hadoop MapReduce and Apache Spark on HDFS, **ACM Digital Library Proceedings of the 1st International Conference on Internet of Things and Machine Learning IML '17**, Article no 14, 17-18 Octobre 2017, Liverpool John Moores University, Liverpool, Angleterre.

Références

[1] Fayyad, U, Piatetsky-Shapiro, G. & Smyth, P, from data mining to knowledge discovery in databases, AI magazine, Volume 17, (3), 1-37.

[2] Christopher Clifton, Data mining COMPUTER SCIENCE, https://www.britannica.com/technology/data-mining, Encyclopedia britannica, (2017).

[3] Akhtar, Syed Muhammad Fahad, Big Data Architect's Handbook, Packt Publishing, ISBN: 978-1-78883-582-4, (2018), pages: 476.

[4] Jiawei Han, Shojiro Nishio, Hiroyuki Kawano, Wei Wang, Generalization-based data mining in object-oriented databases using an object cube model, Data & Knowledge Engineering, Volume 25, (1,2), 55-97.

[5] Michel J.Anzanello, Flavio S.Fogliatto, Karina Rossini, Data mining-based method for identifying discriminant attributes in sensory profiling, Food Quality and Preference, Volume 22, (1), 139-148.

[6] Mansi Gera, Shivani Goel, Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity, International Journal of Computer Applications, Volume 113, (18), (2015), 22-29.

[7] Dasgupta, Nataraj, Practical Big Data Analytics, Packt Publishing, ISBN: 978-1-78355-439-3, (2018), pages: 402.

[8] TANLEY WASSERMAN AND SHEILA O'LEARY WEAVER, Statistical Analysis of Binary Relational Data: Parameter Estimation, JOURNAL OF MATHEMATICAL PSYCHOLOGY, 29, 406-427.

[9] Eibe Frank, Mark Hall, a Simple Approach to Ordinal Classification, EMCL '01 Proceedings of the 12th European Conference on Machine Learning - Springer, 145-156.

[10] Takeshi Fukuda, Yasuhiko Morimoto, Shinichi Morishita, Takeshi Tokuyama, Mining Optimized Association Rules for Numeric Attributes, Journal of Computer and System Sciences, Volume 58, (1), Pages 1-12.

[11] Alex A, Freitas, Understanding the Crucial Role of Attribute Interaction in Data Mining, Artificial Intelligence Review, Volume 16, (3), 177–199,

[12] Kochetov Vadim, Overview of different approaches to solving problems of Data Mining, 8th Annual International Conference on Biologically Inspired Cognitive

Architectures Procedia Computer Science, Volume 123, (2018), 234–239.

[13] Murray E. Jennex, Big Data, the Internet of Things, and the Revised Knowledge Pyramid, The DATA BASE for Advances in Information Systems, Volume 48, Number 4, (2017), 69-79.

[14] Nisbet, Robert, Handbook of Statistical Analysis and Data Mining, Elsevier Science, ISBN: 9780080912035, pages: 864.

[15] Tariq O. Fadl Elsid, Mirghani. A. Eltahir, Data Mining: Classification Techniques of Students'DatabaseA Case Study of the Nile Valley University, North Sudan, International Journal of Computer Trends and Technology IJCTT, Volume 16, (5), (2014), 192-203.

[16] Alexandru Topîrceanu, Gabriela Grosseck, Decision tree learning used for the classification of student archetypes in online courses, Procedia Computer Science, Volume 112, (2017), 51-60.

[17] Breiman, Chapter: Classification Algorithms and Regression Trees, the University of California, San Diego Medical Center, 246-280.

[18] Shaoyan Zhang, Christos Tjortjis, Xiaojun Zeng, Hong Qiao, Iain Buchan, John Keane, Comparing data mining methods with logistic regression in childhood obesity prediction, Information Systems Frontiers, Volume 11, (4), 449-460.

[19] Jianing Fang, Why Logistic Regression Analyses Are More Reliable Than Multiple Regression Analyses, Journal of Business and Economics, Volume 4, (7), 620-633.

[20] PISOTE A, BHUYAR V, REVIEW ARTICLE ON OPINION MINING USING NAÏVE BAYES CLASSIFIER, Advances in Computational Research, Volume 7, (1), (2015), 259-261.

[21] Parneet Kaur, Manpreet Singh, Gurpreet Singh Josan, Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector, Procedia Computer Science, Volume 57, (2015), 500-508.

[22] Shital H. Bhojani, Nirav Bhatt, REVIEW OF LITERATURE OF DATA MINING TECHNIQUES FOR CROPYIELD PREDICTION, INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH, Volume-6, (12), (2017), 357-358.

[23] NANHAY SINGH, RAM SHRINGAR RAW, CHAUHAN R.K, DATA MINING WITH REGRESSION TECHNIQUE, Journal of Information Systems and Communication, Volume 3, (1), 199-202.

[24] Sir Francis Galton, RÉGRESSION LINÉAIRE, Statistique Numérique et Analyse de

Données, 1-4.

[25] Shivangi Bhardwaj, Data Mining Clustering Techniques A Review, International Journal of Computer Science and Mobile Computing, Volume 6, (5), (2017), 183-186.

[26] Alessia Amelio, Andrea Tagarelli, Data Mining: Clustering, In book: Encyclopedia of Bioinformatics and Computational BiologyPublisher: Elsevier, (2017), 1-22.

[27] Noam Slonim, Ehud Aharoni, Koby Crammer, Hartigan's K-Means Versus Lloyd's K-Means – Is It Time for a Change?, Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, 1677-1684.

[28] James Malone, Kenneth McGarry, Stefan Wermter and Chris Bowerman, Data Mining using Rule Extraction fromKohonen Self-Organising Maps, Neural Computing and Applications, Volume 15, (1), 9-17.

[29] CARLOS FERNANDEZ-BASSO, M. DOLORES RUIZ, MARIA J. MARTIN-BAUTISTA, EXTRACTION OF ASSOCIATION RULES USING BIG DATA TECHNOLOGIES, International Journal of Design & Nature and Ecodynamics, Volume 11, (3), (2016), 178–185.

[30] Pratibha Mandave, Megha Mane, Sharada Pati, APRIORI algorithm and improved approach with illustration, International Journal of Latest Trends in Engineering and Technology IJLTET, Volume 3, (2), 108-113.

[31] Vikas Gupta, Prof. Devanand, A survey on Data Mining: Tools, Techniques, Applications, Trends and Issues, International Journal of Scientific & Engineering Research, Volume 4, (3), 1-14.

[32] Hand, D., Mannila, H., & Smyth, P.. Principles of Data Mining. Cambridge, MA and London, England:The MIT Press: A Bradford Book.

[33] Anurag Agrahari, D.T.V. Dharmaji Rao, A Review paper on Big Data: Technologies, Tools and Trends, International Research Journal of Engineering and Technology IRJET, Volume 4, (10), (2017), 640-649.

[34] Nana Kwame Gyamfi, Big Data Analytics: Survey Paper, Conference Proceedings: Dialogue on Sustainability and Environmental Management, (2017), 101-111.

[35] M. Brian Blake, Iman Saleh, Social-Network-Sourced Big Data Analytics, Published by the IEEE Computer Society IEEE INTERNET COMPUTING, 62-69.

[36] Ejaz Ahmed, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Imran Khan, Abdelmuttlib Ibrahim Abdalla Ahmed, Muhammad Imran, Athanasios V.Vasilakos, The role of big data analytics in Internet of Things, Elsevier: computer networks,

Volume 129, (2), (2017), 459-471.

[37] Nada Elgendy, Ahmed Elragal, Big Data Analytics: A Literature Review Paper, Industrial Conference on Data Mining Springer International Publishing Switzerland, (2014), 214-227.

[38] Yasin N. Silva, Isadora Almeida, Michell Queiroz, SQL: From Traditional Databases to Big Data, Proceedings of the sixteenth SIGCSE technical symposium on Computer science education, (2016), 1-6.

[39] D. P. Acharjya, Kauser Ahmed P, A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools, International Journal of Advanced Computer Science and Applications, Volume 7, (2), (2016), 511-518.

[40] S.Vijayarani, S.Sharmila, RESEARCH IN BIG DATA – AN OVERVIEW, Informatics Engineering, an International Journal (IEIJ), Volume 4, (3), (2016), 1-20.

[41] Hanlu Chen, Zheng Yan, Security and Privacy in Big Data Lifetime: A Review, International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage, (2016), 1-13.

[42] Rajeev Agrawal, Christopher Nyamful, Challenges of big data storage and management, Global Journal of Information Technology, Volume 6, (1),(2016), 01-10.

[43] Raquel Mello, Jose Eduardo Mendonca Xavier, Roberto Antonio Martins, Use of Big Data Analytics in Performance Measurement Systems, Proceedings of the 2015 Industrial and Systems Engineering Research Conference, 1-9.

[44] Nigel Franciscus, Precomputing architecture for flexible and efficient big data analytics, Vietnam Journal of Computer Science, Volume 5, (2), (2018), 133–142.

[45] Alexandra L'Heureux, Katarina Grolinger, Hany F. ElYamany, Miriam A. M. Capretz, Machine Learning with Big Data: Challenges and Approaches, IEEE Access, (2017), 1-22.

[46] Salisu Musa Borodo, Siti Mariyam Shamsuddin, Shafaatu nnur Hasan, Big Data Platforms and Techniques, Indonesian Journal of Electrical Engineering and Computer Science, Volume 1, (1), (2016), 191-200.

[47] Harshawardhan S. Bhosale, Devendra P. Gadekar, A Review Paper on Big Data and Hadoop, International Journal of Scientific and Research Publications, Volume 4, (10), (2014), 1-7.

[48] Mubashir Hussain, Jatinder Manhas, ARTIFICIAL INTELLIGENCE FOR BIG DATA: POTENTIAL AND RELEVANCE, International Academy of Engineering

and Medical Research, Volume 1, (1), (2016), 1-5.

[49] Kaiser J. Giri, Towseef A.Lone, Big Data -Overview and Challenges Kaiser, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, (6), (2014), 525-529.

[50] Duygu Sinanc Terzi, Umut Demirezen, and Seref Sagiroglu, EVALUATIONS OF BIG DATA PROCESSING, Services Transactions on Big Data, Volume 3, (1), (2016), 44-53.

[51] Mugdha Ghotkar, Priyanka Rokde, Big Data: How it is Generated and its Importance, National Conference on Recent Trends in Computer Science and Information Technology, (2016), 1-5.

[52] Cheikh Kacfah Emani, Nadine Cullot, Christophe Nicolle, Understandable Big Data: A survey, ScienceDirect, computer science review, Volume 17, (2015), 70-81.

[53] Alexandru Adrian TOLE, Big Data Challenges, Database Systems Journal, Volume 4, (3), 31-40.

[54] Abdul Ghaffar Shoro & Tariq Rahim Soomro, Big Data Analysis: Spark Perspective, Global Journal of Computer Science and Technology: Software & Data Engineering, Volume 15, (1), (2015), 6-14.

[55] R. Siva Ram Prasad, and Chittineni Aruna, Scalable and Flexible Big Data Analytic Framework (SFBAF) For Big Data Processing and Knowledge Extraction, International Conference on Engineering Technologies and Big Data Analytics ETBDA'2016, (2016), 51-55.

[56] Kamalika Dutta, Manasi Jayapal, Big Data Analytics for Real Time Systems, Conference: Big Data Analytics Seminar, (2015), 1-13.

[57] Abdelladim Hadioui, Nour-eddine El Faddouli, Machine Learning Based On Big Data Extraction of Massive Educational Knowledge, Machine Learning Based On Big Data, Volume 12, (11), (2017), 151-167.

[58] Himanshu Shekhar, Manoj Sharma, A Framework for Big Data Analytics as a Scalable Systems, Special Conference Issue: National Conference on Cloud Computing & Big Data, (2017), 72-82.

[59] Li Cai, Yangyong Zhu, The Challenges of Data Quality and Data Quality Assessment in the Big Data Era, Data Science Journal, Volume 14, (2), 1-10.

[60] K.R.Kundhavai, S.Sridevi, IoT and Big Data- The Current and Future Technologies: A Review, International Journal of Computer Science and Mobile Computing, Volume 5,

(1), (2016), 10-14.

[61] C. Lakshmi, V. V. Nagendra Kumar, Survey Paper on Big Data, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, (8), (2016), 368-381.

[62] R.Devakunchari, Analysis on big data over the years, International Journal of Scientific and Research Publications, Volume 4, (1), (2014), 1-7.

[63] Awodele .O, Izang A.A, Kuyoro S.O, Osisanwo F.Y, Big Data and Cloud Computing Issues, International Journal of Computer Applications, Volume 133, (12), (2016), 14-19.

[64] Hamid Bagheri, Abdusalam Abdullah Shaltooki, Big Data: Challenges, Opportunities and Cloud Based Solutions, International Journal of Electrical and Computer Engineering, Volume 5, (2), (2015), 340-343.

[65] Shashi Shekhar, Michael R. Evans, Viswanath Gunturi, KwangSoo Yang, Spatial Big-Data Challenges Intersecting Mobility and Cloud Computing, NSF Workshop on Social Networks and Mobility in the Cloud, 1-6.

[66] Alexandru Adrian TOLE, Big Data Challenges, Database Systems Journal, volume 4, (3), 31-40.

[67] Abu Bakar Munir, Siti Hajar Mohd Yasin, Firdaus Muhammad-Sukki, Big Data: Big Challenges to Privacy and Data Protection, International Journal of Social, Education, Economics and Management Engineering, Volume 9, (1), (2015), 355-363.

[68] ANCA D CHIRITA, the Rise of Big Data, Durham University, (2014), 1-26.

[69] Rakesh Kumar, Bhanu Bhushan Parashar, Sakshi Gupta, Yougeshwary Sharma, Neha Gupta, Apache Hadoop, NoSQL and NewSQL Solutions of Big Data, International Journal of Advance Foundation and Research in Science & Engineering IJAFRSE, Volume 1, (6), (2014), 28-36.

[70] Eiman Al Nuaimi, Hind Al Neyadi, Nader Mohamed, Jameela Al-Jaroodi, Applications of big data to smart cities, Journal of Internet Services and Applications, (2015), 1-15.

[71] Ibrahim Abaker Targio Hashem, Victor Chang , Nor Badrul Anuar, Kayode Adewole, Ibrar Yaqoob, Abdullah Gani, Ejaz Ahmed, Haruna Chiroma,The Role of Big Data in Smart City, International Journal of Information Management, (2016), 1-20.

[72] Fadila Bentayeb, Entrepôts et analyse en ligne de données complexes centrés utilisateur, Databases Université Lumiére - Lyon II.

[73] Ansari Faheem, Arif V Shaikh, Massive Analyzing of Data Processing using Hadoop Eco System Based on Cloud Environment, International Conference on Business and Culture with Changing Technology In Emerging Market, Volume: 1, (2016), 1-6.

[74] Zinayida Petrushyna, Alexandra Chueva, Ralf Klamma, Joachim Lanfermann, A Near Real-Time Application for Twitter Data Analysis, e-dynamics web intelligence, (2015).

[75] Arul Murugan, Anguraj, Boopathi, Big Data: Privacy and Inconsistency Issues, International Journal of Research in Engineering and Technology, Volume 3, Special Issue 7, (2014), 812-815.

[76] Enrico Giacinto Caldarola, Antonio Maria Rinaldi, Big Data Visualization Tools: A SurveyThe new paradigms, methodologies and tools for large data sets visualization, 6th International Conference on Data Science, Technology and Applications, (2017), 1-10.

[77] Anurag Agrahari, V. Dharmaji Rao, A Review paper on Big Data: Technologies, Tools and Trends, International Research Journal of Engineering and Technology, Volume 04, (10), (2017), 640-649.

[78] M. Usman Nisar, Arash Fard, John A. Mille, Techniques for Graph Analytics on Big Data, IEEE International Congress on Big Data, 1-8.

[79] Brijesh B. Mehta, Udai Pratap Rao, Nikhil Kumar, Towards privacy preserving big data analytics, 2016 Sixth International Conference on Advanced Computing & Communication Technologies, (2016), 28-35.

[80] Nancy Victor, Daphne Lopez, Privacy models for big data: a survey, International Journal of Big Data Intelligence, Volume 3, (1), (2016), 61-75.

[81] Venketesh Palanisamy Ramkumar Thirunavukarasu, Implications of big data analytics in developing healthcare frameworks – A review, Journal of King Saud University - Computer and Information Sciences, (2017), 1-11.

[82] Ari Banerjee, Big Data & advanced analytics in telecom a multi-billion dollar revenue opportunity, Heavy reading, 1-24.

[83] Andrianambinina Marius, Développement d'un moteur de recommandation, Université de la Réunion UFR Sciences et Technologies, (2017), 1-45.

[84] Savita Kumari, Impact of big data and social media on society, Gjra - Global Journal for Research Analysis, Volume 5, (3), (2016), 437-439.

[85] C. K. M. LeeYi CaoKam K.H. NgKam K.H. Ng, Big Data Analytics for Predictive

Maintenance Strategies, in book: Supply Chain Management in the Big Data Era, (2017), 50-74.

[86] Eddy Bajic, Oussama Hajlaoui, APPORTS DES PARADIGMES SOCIAUX DANS L'INTERNET DES OBJETS INDUSTRIEL: VERS DES OBJETS COMMUNICANTS INDUSTRIELS SOCIAUX, 12e Conférence Internationale de Modélisation, Optimisation et SIMulation, (2018), 1-8.

[87] Maged N Kamel Boulos, Najeeb M Al-Shorbaji, On the Internet of Things, smart cities and the WHO Healthy Cities, INTERNATIONAL JOURNAL OF HEALTH GEOGRAPHICS, Volume 13, (10), (2017), 1-6.

[88] Edward Curry, Schahram Dustdar, Quan Z. Sheng, Amit Sheth, Smart cities – enabling services and applications, Journal of Internet Services and Applications, Volume 7, (6), (2016), 1-3.

[71] Jim Hillage, Emma Pollard, Employability: Developing a framework for policy analysis, Institute for Employment Studies, 1-4.

[72] Mel Fugate,a, Angelo J. Kinicki,b, Blake E. Ashforth, Employability: A psycho-social construct, its dimensions, and applications, Journal of Vocational Behavior, Volume 65, (1), 14-38.

[73] LEE HARVEY, Defining and Measuring Employability, Quality in Higher Education, Volume 7, (2), 97-109.

[74] Parneet Kaur, Manpreet Singh, Gurpreet Singh Josan, Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector, Procedia Computer Science, Volume 57, (2015), 500-508.

[75] Tripti Mishra1, Dharminder Kumar, Sangeeta Gupta, Students' Performance and Employability Prediction through Data Mining: A Survey, Indian Journal of Science and Technology, Volume 10, (24), (2017), 1-6.

[76] Mohd tajul rizal, yuhanis Yusof, Application of data mining in forecasting graduates employment, Journal of Engineering and Applied Sciences, Volume 12, (16), (2017), 4202-4207.

[77] SOUALI, Mohamed, Le Maroc In : Enseignement supérieur et marché du travail dans le monde arabe, Presses de l'Ifpo.

[78] Haut-Commissariat au Plan, Direction de la Statistique, Enquête nationale sur l'emploi, http://www.hcp.ma, (2016).

[79] Monika Goyal, Rajan Vohra, Applications of Data Mining in Higher Education, IJCSI

International Journal of Computer Science Issues, Volume 9, (2), 113-120.

[80] Venkatadri.M, Lokanatha C. Reddy, A Comparative Study On Decision Treeclassification Algorithms In Data Mining, International Journal Of Computer Applications Inengineering, Technology And Sciences, Volume 2, (2), 24-29.

[81] Pooja Thakar, Anil Mehta, Manisha,Role of Secondary Attributes to Boost the Prediction Accuracy of Students' Employability Via Data, International Journal of Advanced Computer Science and Applications,Volume 6, (11), (2015), 84-90.

[82] Muskan Kukreja, Stephen Albert Johnston, Phillip Stafford, Comparative study of classification algorithms for immunosignaturing data, BMC Bioinformatics, Volume 13, (139), 1-25.

[83] Mohd tajul rizal, yuhanis Yusof, Application of data mining in forecasting graduates employment, Journal of Engineering and Applied Sciences, Volume 12, (16), (2017), 4202-4207.

[84] Keno C. Piad, Determining The Dominant Attributes Of Information Technology Graduates Employability Prediction Using Data Mining Classification Techniques, Journal of Theoretical and Applied Information Technology, Volume 96, (12), (2018), 3780-3790.

[85] Azziaty Abdul Rahman, Kian Lam Tan, Chen Kim Lim, Supervised and Unsupervised Learning in Data Mining for Employment Prediction of Fresh Graduate Students, International journal of telecommunication, electronic and computer engineering, Volume 9, (2), 155-161.

[86] Parneet Kaura,Manpreet Singhb,Gurpreet Singh Josan, Classification and prediction based data mining algorithms to predict slow learners in education sector, 3rd International Conference on Recent Trends in Computing, Procedia Computer Science, Volume 57, (2015), 500 – 508.

[87] Rosna Awang Hashim, Lim Hock Eam, Bidin Yatim, Tengku Faekah Tengku Ariffin, Ainol Madziah Zubairi, Haniza Yon et Omar Osman, ESTIMATING A PREDICTION MODEL FOR THE EARLY IDENTIFICATION OF LOW EMPLOYABILITY GRADUATES IN MALAYSIA, The Singapore Economic Review, Vol. 60, No. 2 (2015), 1-22.

[88] Rapid Miner, https://rapidminer.com, (2019).

[89] Kalpana Rangra, K. L. Bansal, Comparative Study of Data Mining Tools, International Journal of Advanced Research in Computer Science and Software Engineering,

Volume 4, (6), (2014), 216-223.

[90] Gartner Magic Quadrant, https://www.gartner.com/en/research/methodologies/magic-quadrants-research.

[91] Carlie Iodine, Peter Krensky, Gartner Magic Quadrant for Data Science and Machine-Learning Platforms, (2019).

[92] Magic Quadrant for Data Science and Machine Learning Platforms.

[93] Firas Mohammed Ali, El-Bahlul Emhemed Fgee, Zakaria Suliman Zubi, PREDICTING PERFORMANCE OF CLASSIFICATION ALGORITHMS, International Journal of Computer Engineering and Technology, Volume 6, Issue 2, February (2015), pp. 19-28.

[94] Hossin, M, M.Sulaiman, A REVIEWON EVALUATION METRICS FOR DATA CLASSIFICATION EVALUATIONS, International Journal of Data Mining & Knowledge Management Process, Volume5, (2), (2015), 1-11.

[95] Anthony J. Viera, MD; Joanne M. Garrett, PhD, Understanding Interobserver Agreement: The Kappa Statistic, Research Series, Volume 7, (5), (2015), 360-363.

[96] Katerina Lepenioti, Alexandros Bousdekis, Dimitris Apostolou, Gregoris Mentzas, Prescriptive analytics: Literature review and research challenges, International Journal of Information Management, Volume (50), (2020), 57-70.

[97] Cloudera, https://www.cloudera.com/, 2019.

[98] Php, http://php.net/, (2019).

[99] MySQL, https://www.mysql.com/fr/, (2019).

[100] Rahul Beakta, Big DataAnd Hadoop: A Review Paper, RIEECE, Volume 2, (2), (2015), 13-15.

[101] IBM, what is Hadoop, http://www-01.ibm.com/software/data/infosphere/hadoop.

[102] McKinsey Global Institute, Big Data: The Next Frontierfor Innovation, Competition,and Productivity,McKinsey & Co, 1-20.

[103] Nikolaos Samaras, Sotiris Madamas, A review of the Hadoop ecosystem exploring the TFOCS optimization solver utilizing the data processing engine of Apache Spark, 5th International Symposium & 27th National Conference on Operational Research, (2016), 1-5.

[104] Fredrick Romanus Ishengoma, HDFS: Erasure Coding Based Hadoop Distributed File System, INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, Volume 2, (8), 1-8.

[105] R.Thangaselvi, S.Ananthbabu, R.Aruna, An efficient Mapreduce scheduling algorithm in hadoop, International Journal of Engineering Research & Science IJOER, Volume 1, (9), (2015), 102-108.

[106] Neseeba P.B, Dr. Zahid Ansari, Performance Analysis of Hbase, International Journal of Latest Technology in Engineering, Management & Applied Science IJLTEMAS, Volume 6, (10), (2017), 10-14.

[107] Rotsnarani Sethy, Santosh Kumar Dash, and Mrutyunjaya Panda, Performance Comparison Between Apache Hiveand Oracle SQL for Big Data Analytics, Springer International Publishing Proceedings of the Eighth International Conference on SoftComputing and Pattern Recognition (2016), 130-141.

[108] Rakesh Kumar, Neha Gupta, Shilpi Charu, Somya Bansal, Kusum Yadav, Comparison of SQL with HiveQL, International Journal for Research in Technological Studies, Volume 1, (9), (2014), 2348-1439.

[109] Prof. Pramod Patil, Amit Patange, Impala: Open Source, Native Analytic Database for Apache Hadoop - A Review Paper, International Journal of Science and Research, Volume 4, (5), (2015), 1445-1448.

[110] Subhani shaik, Nalamothu Naga Malleswara Rao, A Conceptual Review of Elastic Search – Survey Paper, International Journal for Research in Applied Science & Engineering Technology IJRASET, Volume 5, (11), (2017), 1703-1710.

[111] Priya P. Sharma, Chandrakant P. Navdeti, Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution, International Journal of Computer Science and Information Technologies, Volume 5, (2), (2014), 2126-2131.

[112] Sebastian Schelter, Sean Owen, Collaborative Filtering with Apache Mahout, Recommender Systems Challenge 2012 in conjunction with the ACM Conference on Recommender Systems, 1-2.

[113] Abhishek Bhattacharya, Shefali Bhatnagar, Big Data and Apache Spark: International Journal of Engineering Research & Science IJOER, Volume 2, (5), (2016), 206-210.

[114] Poonam S. Patil, Rajesh. N. Phursule, Survey Paper on Big Data Processing and Hadoop Components, International Journal of Science and Research, Volume 3, (10), (2014), 585-590.

[115] Varsha B.Bobade, Survey Paper on Big Data and Hadoop, International Research Journal of Engineering and Technology IRJET, Volume 3, (01), (2016), 861-863.