République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université d'Abbès Laghrour Khenchela
Faculté des Sciences et de la Technologie
Département des Mathématiques et d'Informatique

Mémoire de fin d'études

pour l'obtention du diplôme de Master en Informatique
Spécialité : Génie logiciel et systèmes distribués.

Modèles de classification pour la prédiction des résultats des étudiants de MI en premiere année universitaire

Présenté par :
BOUMAAZA Laid
ARAAR Mohamed Amine

Soutenu le 09/09/2020 devant le Jury déxamen :

Dr	**MAHDAOUI Rafik**	**Président**
Dr	**HAOUASSI Hichem**	**Rapporteur**
Dr	**MALIK Mohamed mahdi**	**Examinateur**

Promotion :2019/2020

Table des matières

Table des figures

Liste des tableaux

Remerciement

Nous aimerons, en premier lieu, remercier le bon Dieu tout puissant qui nous a donné la volonté et la force afin de réaliser ce travail.
Nous voudrions remercier grandement, Dr Hichem HAOUASSI, qui nous avoir dirigées. Il a toujours été disponible, à l'écoute de nos questions, et il s'est toujours intéressé à l'avancement de nous travaux. Pour tout cela merci.
Nous exprimons toute notre reconnaissance aux membres de jury de nous avoir fait l'honneur de participer à nos jury. Nous remerciements Merci a qui n'ont cessé d'être pour Nous des exemples de persévérance, de courage et de générosité, a vous, nos parents, nos soeur et nos frères.
Nous n'oublions pas de remercier tous nous collègues et amis sans exception.

Résumé

L'extraction des connaissances à partir des données est définie comme le processus d'analyse des données sous différentes perspectives et de découverte des modèles à partir des ensembles de données utiles pour prédire les résultats qui nous aident à prendre la bonne décision.
Comme ce processus passe par plusieurs étapes, de la (collecte des données. Nettoyage et transmission) à l'obtention des résultats, leur validation et leur interprétation, et enfin la fusion des connaissances acquises.
Il existe de nombreuses techniques d'analyse de données qui peuvent être utilisées pour extraire des modèles, parmi lesquelles des techniques de classification qui génèrent des modèles (phase d'apprentissage) utilisés pour prédire les données futures d'un objet (phase de prédiction) à partir de données provenant d'autres objets similaires.
Notre travail est lié à cet aspect. Nous avons d'abord collectée les données de 250 étudiants via un questionnaire constitué de 56 questions et nous avons ensuite générer plusieurs modèles qui permettent de classer les étudiants universitaires de première année en mathématiques et en informatique avec différentes performances selon l'approche utilisée.
Notre travail est concrétisé par une application développé sous Java et Weka qui permette de faire la saisie des nouvelles données, l'apprentissage pour la création des modèles de classification et de prédiction des classes de nouveaux étudiants via leurs données.
Mot clés :extraction de connaissances à partir de données, fouille de données éducatives, apprentissage automatique, classification, prédiction,Fouille de données.

Abstract

Knowledge extraction from data is defined as the process of analyzing data from different perspectives and discovering patterns from useful data sets to predict the results that help us make the right decision.
As this process goes through several stages, from (data collection. Cleaning and transmission) to obtaining the results, their validation and their interpretation, and finally the fusion of the acquired knowledge.
There are many data analysis techniques that can be used to extract models, including classification techniques that generate models (learning phase) used to predict future data for an object (prediction phase) from data from other similar objects.
Our work is linked to this aspect. We first collected data from 250 students via a questionnaire consisting of 56 questions and then we generated several models that allow ranking first-year university students in mathematics and computer science with different performances depending on the approach used.
Our work is concretized by an application developed under Java and Weka which allows to enter new data, learning to create classification models and class prediction of new students via their data.
Keywords : knowledge extraction from data, educational data mining, machine learning, classification, prediction, Data mining.

Chapitre 1

LA FOUILLE DE DONNÉE ÉDUCATIVE

Introduction

Durant ces dernières années, avec l'augmentation sans cesse de capacité de stockage des ordinateurs, les quantités de données collectées, dans divers domaines d'application de l'informatique, deviennent de plus en plus importantes, donc, nous parlons de Big Data, nous parlons de quantités inimaginables de données de nombreux types et sources de tailles énormes. Nous demandons ici à quel point ces données sont importantes à la lumière du fait qu'elles indiquent que les informations structurées pour ces données ne constituent qu'une petite fraction de 10% par rapport aux informations non organisées qui composent le reste. Cela a accru le besoin de développer des outils puissants pour l'analyse des données et l'extraction d'informations et de connaissances. Les méthodes conventionnelles et statistiques ne peuvent pas gérer cette énorme quantité, donc des outils intelligents sont utilisés pour traiter ces données. Ce chapitre a pour objet est de présenter dans un premier temps les concepts liées à la fouille de données. Dans un second temps, il présente les techniques de data mining qu'on peut utiliser pour l'extraction des connaissances à partir des données éducatives.

1.1 Définitions

1.1.1 Fouille de données (Data maining)

C'est une technologie qui vise à extraire des connaissances à partir d'énormes quantités de données, basées sur des algorithmes mathématiques qui constituent la base de l'extraction de données et dérivent de nombreuses sciences telles que les statistiques, les mathématiques, la logique, les sciences de l'apprentissage, l'intelligence artificielle et les systèmes experts. L'exploration de données est apparue à la fin des années 1980 et s'est révélée être une solution efficace pour analyser de grandes quantités de données et les convertir à partir d'informations accumulées et simplement incompréhensibles en informations précieuses qui pourraient être utiles pour une exploitation et une utilisation ultérieures.[1]

Une autre définition

La fouille de données est le processus d'analyse des données sous différentes perspectives et de découverte des déséquilibres, des modèles et des connexions dans les ensembles de données qui est perspicace et utile pour prédire les résultats qui vous aide à prendre une décision correcte.

Le processus d'exploration de données est basé sur les trois éléments de base suivant :

1.1.2 Données

Selon wikipedia : « Une donnée est une description élémentaire d'une réalité. C'est par exemple une observation ou une mesure ». Donc, 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[4] :

- 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.

Type de données

- Données quantitatives : Valeurs numériques et sommables, discrètes, continues.
- Données qualitatives :
  - ☐ Ordinales : Ex : petit, moyen, grand, trés grand.

- ☐ Nominales :(catégories ou modalités), Ex : féminin, masculin, célibataire, marié, divorcé, veuf...

1.1.3 Information

Selon wikipedia : Au sens étymologique, l'information est ce qui donne une forme à l'esprit. Elle vient du verbe latin « informare », qui signifie « donner forme à » ou « se former une idée de ». L'information est aussi une notion abstraite, mais d'un niveau d'abstraction supérieur à celui de la donnée. On peut dire pour simplifier que l'information est une donnée + un sens

1.1.4 Connaissance

Selon wikipedia : « La connaissance est une notion aux sens multiples à la fois utilisée dans le langage courant et objet d'étude poussée de la part des philosophes contemporains ». La connaissance est aussi une notion abstraite, d'un niveau d'abstraction supérieur à celui de l'information. La connaissance à la différence de l'information est partagée et s'appuie sur un référentiel collectif. Une connaissance est une information nouvelle, apprise par association d'informations de base, de règles, de raisonnement, d'expérience, d'expertise, etc.[6]

Processus de découverte des connaissances

Certaines personnes ne différencient pas l'exploration de données de la découverte de connaissances tandis que d'autres considèrent l'exploration de données comme une étape essentielle du processus de découverte de connaissances. Voici la liste des étapes impliquées dans le processus de découverte des connaissances :

- ***Nettoyage des données*** : Dans cette étape, le bruit et les données incohérentes sont supprimés.
- ***Intégration des données*** : Dans cette étape, plusieurs sources de données sont combinées.
- ***Sélection des données*** : Dans cette étape, les données pertinentes pour la tâche d'analyse sont extraites de la base de données.
- ***Transformation des données*** : Dans cette étape, les données sont transformées ou consolidées sous des formes appropriées pour l'exploration en effectuant des opérations de résumé ou d'agrégation.
- ***Exploration de données*** : Dans cette étape, des méthodes intelligentes sont appliquées afin d'extraire des modèles de données.
- ***Évaluation de modèle*** : Dans cette étape, les modèles de données sont évalués.
- ***Présentation des connaissances*** : Dans cette étape, les connaissances sont représentées.

1.1.5 Entrepôts de données

Il est utilisé dans les analyses temporelles, la découverte des connaissances et la prise de décision. Une énorme quantité de données peut être stockée qui peut être de différentes sources, par exemple plusieurs bases de données de plusieurs modèles, elles sont conçues pour extraire des données, les traiter, les représenter et les présenter de manière appropriée à ces fins.[2]

1.2 Processus d'Extraction des Connaissances à partir des Données (ECD)

Le processus d'extraction des connaissances à partir de données présenté en Figure[5]

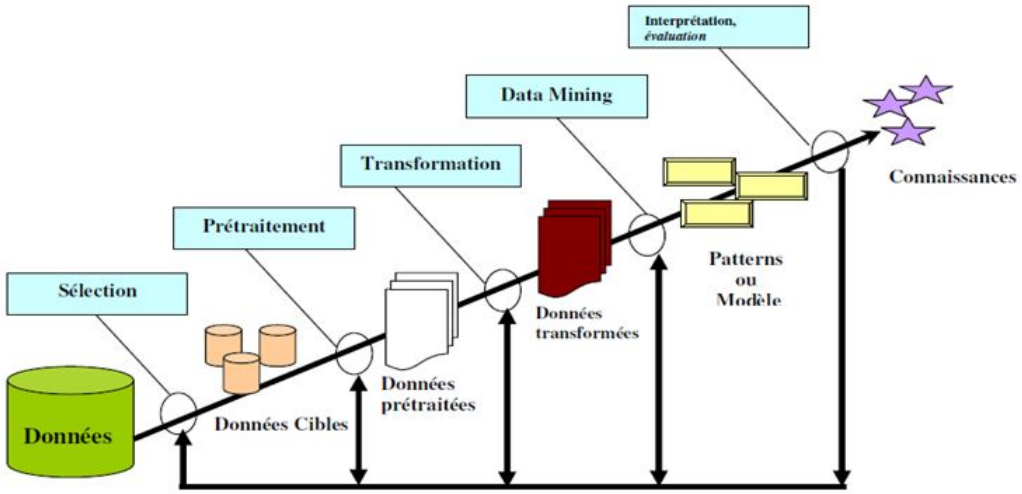


FIGURE 1.1 – Processus (ECD)

Pour extraire des connaissances des données, nous suivons plusieurs étapes :

- ♠ La sélection des données.
- ♠ Le prétraitement des données.
- ♠ La transformation de données.
- ♠ Fouille de données (Data Mining).
- ♠ Evaluationet interprétation des connaissances.

1. **La sélection des données** : L'objectif de l'extraction de connaissance est de déduire des nouvelles connaissances valides et utiles.Cet aspect est très important, mais on ne peut pas appliquer le processus d'ECD sur toutes les données qu'on a, donc le besoin est exprimé par l'utilisateur. Ce dernier fait la sélection des données selon l'objectif visé.

   Cette étape concerne donc le filtrage des données qui comprend deux opérations nécessaires :

   - ***La réduction de la dimensionnalité des données :***l'élimination d'attributs sans intérêt, ou ayant beaucoup de valeurs erronées ou manquantes.
   - ***La réduction de la taille des données :*** l'application des techniques du Data Mining est très coûteuse en terme de temps CPU et d'espace mémoire, c'est pour cela qu l'on ne peut pas l'appliquer sur la totalité des données.

2. **Le prétraitement des données :** Le rôle de cette étape est de préparer les données afin qu'elles soient de meilleurs qualités afin d'arriver à des résultats de qualité. Le prétraitement des données concerne, entre autres, le nettoyage des données, c'est-à-dire l'élimination du bruit, ainsi que le traitement des valeurs manquantes, ou erronées. Il faudrait alors définir les méthodes à utiliser pour le remplacement de ces valeurs.

3. **La transformation de données :** Cette étape consiste à préparer les données brutes, et à les convertir en données appropriées. La transformation se fait par attribut, c'est-à-dire toutes les valeurs d'un attribut doivent être transformées en un format unique. Formellement, un attribut « A » est transformé en « B » qui serait utilisable par la tâche de la fouille de données choisie.

4. **Fouille de données (Data Mining) :** Dans cette étape, des méthodes intelligentes sont utilisées afin d'extraire des modèles ou patterns. Cette étape est aussi désignée comme l'étape cœur du processus d'ECD . Il est clair que les étapes qui précèdent la fouille de données sont très importantes, car la qualité des modèles ou patterns extraites, ainsi que leur coût d'extraction sont liés directement à ces étapes.

5. **Evaluation et interprétation des connaissances :** Les modèles ou patterns extraits ne sont pas dans la plupart du temps exploitables. En effet, il est difficile d'avoir directement des connaissances valides et utiles, à ce point-là. Il existe, cependant, des méthodes d'évaluation des modèles extraits. Ces méthodes peuvent aussi aider à corriger les modèles, et à les ajuster aux données.

   Ces principales étapes de processus (ECD) visent à partir de données volumineuses l'extraction des connaissances, qui peuvent être exprimés sous forme d'un concept général qui enrichit le champ sémantique de l'usager

par rapport à une question qui le préoccupe. Elles peuvent prendre la forme (figure 1.1)

1.3 Tâches d'exploration de données

1.3.1 Classification et prédiction

La classification consiste à traiter un ensemble de modèles (ou fonctions) qui décrivent et distinguent les classes de donnée, afin de pouvoir utiliser le modèle pour prédire la classe d'objets dont l'étiquette de classe est inconnue. Le modèle dérivé est basé sur l'analyse d'un ensemble de données d'apprentissage (c'est-à-dire des objets de données dont l'étiquette de classe sont connus). La classification et la prévision peuvent devoir être précédées d'une analyse de pertinence qui tente d'identifier les hommages qui ne contribuent pas au processus de classification ou de prédiction. Ces attributs peuvent ensuite être exclus

1.3.2 Clustérisation

Contrairement à la classification et à la prédiction, qui analyse les objets de données nommés dans une classe, l'agrégation analyse les objets de données sans référence à une étiquette de classe connue. En général, les étiquettes de classe n'étaient pas présentes dans les données de formation simplement parce qu'elles n'étaient pas connues au début. L'agrégation peut être utilisée pour créer de telles étiquettes. Les objets sont regroupés selon le principe de maximiser la similitude entre les catégories et de réduire la similitude entre les classes. Autrement dit, des groupes d'objets sont formés de sorte que les objets d'un groupe sont très similaires aux autres groupes, mais sont très différents des autres objets du groupe.

1.3.3 Les règles d'association

Les règles d'association sont parmi les méthodes de recherche des implications (relation) entre les attributs d'une base de données. Les règles d'association ont été introduites par Agrawal et al. L'étymologie des règles d'associations, aussi connues sous le nom de l'analyse du panier de la ménagère (Market Basket Analysis), vient des travaux qui ont été réalisés à partir des données provenant des supermarchés. L'objectif de ces travaux est d'identifier les items ou les groupes d'items (Itemset), fréquemment achetés ensembles par un client lors d'une même transaction où son support est supérieur ou égal à un seuil minimum, (Minsupp), fixé par l'utilisateur.[5]

1.4 Applications d'exploration de données

Il existe de nombreux exemples d'applications d'exploration de données dans divers domaines, notamment les suivants :

1.4.1 Affaires et argent

L'exploration de données, par exemple, est utilisée pour anticiper la capacité des clients à rembourser leurs prêts financiers. Cela se fait en appliquant des algorithmes d'exploration de données aux enregistrements historiques des clients précédents, résultant en une forme ou un ensemble de règles qui déterminent si le client peut rembourser son prêt ou non.

1.4.2 La médecine

Une grande présence de l'exploration de données est les solutions que vous fournissez, telles que l'attente d'un patient atteint de certaines maladies en fonction de ses données et de son dossier médical, ou l'étendue des effets des médicaments et des médicaments sur les patients en fonction des dossiers des patients précédents.

1.4.3 Marketing et vente

Les techniques d'exploration de données aident les sociétés de marketing à créer des modèles basés sur des données historiques pour de nouvelles campagnes de marketing telles que le publipostage et les campagnes de marketing Internet. Grâce aux résultats, les sociétés de marketing disposeront d'une méthode appropriée pour vendre des produits rentables à des clients cibles. L'exploration de données aide les entreprises de vente au détail, grâce à l'analyse du panier de marché, à ce que le magasin dispose d'une gestion appropriée afin que les clients puissent toujours acheter ensemble les produits qu'ils achètent fréquemment. De plus, il aide les détaillants à offrir certaines remises pour des produits spécifiques susceptibles d'attirer plus de clients.

1.4.4 L'aspect éducatif

L'exploration de données est utilisée pour anticiper la performance académique des étudiants sur la base des données de performance des étudiants précédents. Il est également possible de prédire la performance des enseignants ou leur capacité à donner des cours spécifiques. De nombreuses études et recherches sont menées dans ce domaine, et les applications d'exploration de données sont appelées dans le domaine éducatif(Educational Data Mining). ***L'exploration de données a un rôle pivot dans tous les domaines scientifiques et pratiques Son existence est une nécessité des systèmes et des applications.***

1.5 Fouille de données éducatives (Educational Data Mining)

1.5.1 Définition

C'est un domaine moderne qui vise à développer des moyens d'explorer les types de données utiles obtenues à partir des environnements d'apprentissage et à utiliser ces méthodes pour mieux comprendre les élèves et les environnements dans lesquels ils apprennent. Les

principales utilisations de l'extraction de données éducatives comprennent la prévision des performances des élèves et l'étude du processus d'apprentissage afin de recommander des améliorations aux pratiques éducatives actuelles. Cela peut être considéré comme l'extraction de données éducatives, un domaine des sciences de l'éducation et de l'exploration de données, et l'analysedu processus d'apprentissage est un domaine connexe. Il aide à la prise de décision basée sur les données pour améliorer les pratiques et conditions d'enseignement actuelles.

1.5.2 Objectifs de la fouille de données éducatives

Les objectifs de la recherche rapportée sont de justifier les potentiels des algorithmes d'exploration de données dans le contexte de l'enseignement supérieur en développant un modèle d'exploration de données. Dans cette recherche, l'accent est mis sur l'extraction des connaissances cachées de la base de données des étudiants et la prévision des performances des étudiants en fonction des paramètres de dépendance. L'objectif de la recherche est de développer un modèle d'exploration de données avec des techniques de classification et de regroupement. La technique de classification classe les données en fonction de l'ensemble d'apprentissage et applique le modèle pour classer. L'EDM a été mis en œuvre ces dernières années pour atteindre un grand nombre d'objectifs éducatifs, notamment :

Objectifs pédagogiques

Ce sont les objectifs académiques qui aident à développer et à améliorer le niveau d'éducation d'un étudiant, mais aussi à guider les étudiants dans le cheminement scolaire en fonction de leur niveau et des résultats obtenus.

1.5.3 Tâches EDM

Les tâches EDM de base peuvent être mappées à des techniques d'exploration de données telles que : [3]

- ***Classification :***catégorisation des élèves pour déterminer les styles d'apprentissage et les préférences.
- ***La modélisation prédictive :*** prédit la performance d'un étudiant à l'examen semestriel.
- ***Regroupement :***regrouper des étudiants similaires en fonction des résultats scolaires pour l'apprentissage collaboratif.
- ***Exploration de modèles :***recherche de modèles.
- ***Analyse visuelle :*** raisonnement sur les processus éducatifs.

1.5.4 les méthodes

La fouille de données englobe plusieurs méthodes, qui permettent de créer des modèles ou patterns afin de les utiliser pour la découverte de connaissances. Nous nous concentrerons sur les méthodes d'explications et de prédictions. Ces méthodes peuvent être classifiées en trois catégories.

- Les méthodes de visualisation et de description.
- Les méthodes de classification et de structuration.
- Les méthodes d'explication et de prédictions.

les méthodes de visualisation et de description

Les méthodes de visualisation et de description, sont issues de la statistique descriptive et de l'analyse des données, ainsi que de la visualisation graphique. En effet, les méthodes de visualisation et de description sont fondées sur des graphiques, qui facilitent l'interprétation à l'utilisateur. La visualisation d'un graphique sert principalement à explorer les données, ou à confirmer des hypothèses.

Les méthodes de classification et de structuration

Les méthodes de classification et de structuration connus sous le nom classification automatique ou apprentissage non supervisé. Ces méthodes proviennent de l'analyse des données, de la reconnaissance des formes, de l'apprentissage automatique et du connexionnisme.

Les méthodes d'explication et de prédictions

Les méthodes d'explication et de prédictions ont pour objectif de relier un phénomène à expliquer à un phénomène explicatif, elles sont utilisées pour prévoir un comportement, ou bien pour classer de nouveaux cas dans des catégories prédéfinies. Ces méthodes sont issues de la statistique, de l'économétrie, de la reconnaissance de formes, de l'apprentissage automatique et du connexionnisme.

1.6 Outils de la fouille de données

1.6.1 Logiciels libres

Parmi les logiciels libres : KNIME et weka, ces deux logiciels sont décrits ci-dessous.

- ***KNIME*** **:**acronyme de Konstanz Information Miner, est un logiciel libre édité par un laboratoire de l'université de Constance dénommé Nycomed Chair for Bioinformatics and Information Mining. Il intègre notamment tous les modules d'analyse de Weka et permet de créer des scripts en langage R. KNIME s'exécute sur Linux, Windows et MacOS. Comme tous les logiciels libres, KNIME est extensible.

- ***Weka*** **:**est un logiciel libre de fouille de données divelopé en java et créé par l'université de Waikato (Nouvelle-Zélande). C'est une collection d'algorithmes d'apprentissage automatique mis en place pour effectuer des tâches d'exploration de données. Les algorithmes peuvent soit être appliqués directement à un ensemble de données soit être appelés directement par un code Java. Weka contient des outils pour les prétraitements des données, la classification, la régression, le clustering, les règles d'association et la visualisation. Comme KNIME, weka est un logiciel open source.

- ***RapidMiner*** **:**est un logiciel open source 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, tous les algorithmes de data mining, évaluation des performances et visualisations diverses. C'est un logiciel puissant, il n'est pas facile à manipuler au premier abord, mais avec un peu de pratique, il permet de mettre en place rapidement une chaîne complète de traitement de données, de la saisie des données à leur classification.

1.6.2 Logiciels commerciaux

Les logiciels commerciaux sont édités par des sociétés bien connues sur le marché :

- ***KXEN*** **:**Analytic Framework est un logiciel commercial édité par la société KXEN basée en Californie et fondée en 1998. Les modules de KXEN Analytic Framework permettent la prédiction, la segmentation, les associations, la fouille de textes et l'analyse des réseaux sociaux.

- ***SAS Enterprise Miner*** **:**est un outil commercial édité par la société SAS Institute Inc. C'est un logiciel offrant toutes les facettes de l'exploration de données dont le processus est facilité par son interface homme-machine bien conçue.

- ***SPSS (Statistical Package for the Social Sciences)*** **:**est un logiciel de statistiques, édité par la filiale d'IBM du même nom, qui se décompose en plusieurs modules dont SPSS Modeler pour le Data mining, SPSS Amos pour les modèles d'équation structurelle et Predictive Analytics pour l'analyse prédictive.

- ***CORICO*** **:**est un logiciel commercial intégrant l' Iconographie des corrélations et les Interactions logiques, qui se prêtent bien à l'analyse multi relationnelle. Il intègre aussi une technique de modélisation prédictive fondée sur les modèles de régression multiple postulés et non postulés.

1.7 Conclusion

Le développement des sciences, de l'économie et des technologies de l'information et de la communication a accru la quantité de données numériques et, avec ces énormes quantités de données, les méthodes d'analyse traditionnelles ne sont plus en mesure de les gérer.
Par conséquent, pour résoudre ces problèmes, l'exploration de données a été fournie avec des outils et des logiciels qui aident à explorer la quantité considérable et croissante de données, car ces données sont produites au milieu de sites de réseautage social ; et diverses institutions telles que les banques et les compagnies d'assurance, il est nécessaire d'explorer cette quantité de données pour en bénéficier afin d'accéder à des connaissances qui nous aident à prendre des décisions.

Chapitre 2

CLASSIFICATION

Introduction

En 1749, le célèbre naturaliste et écrivain George Buffon a déclaré que la seule façon de créer une manière bénéfique et naturelle est de regrouper des choses similaires et d'en séparer des choses différentes. A travers les problèmes et les obstacles rencontrés lors du contrôle de l'information et de son intensité, les scientifiques ont prouvé que le processus général de classification dans le domaine des technologies de l'information consiste à l'appliquer aux données numériques (points, tableaux, images, sons, etc.) La performance générale des méthodes de classification, depuis 1749, a été la tradition et l'automatisation. A travers ce qui précède, nous présenterons dans ce chapitre la classification, ses méthodes, ses techniques, ses principales approches et ses domaines d'application... etc. Enfin, nous détaillerons ses principales approches en étudiant et analysant quelques algorithmes.

2.1 Définition

Le système de classification est un système qui est directement lié (et aussi appelé sources de terrain, et ce sont ces sources qui ont une relation directe avec le sujet de l'étude, et dans lesquelles les données sont collectées directement par le chercheur) ou indirectement (et aussi appelées sources historiques, Ce sont les sources qui contiennent des informations transférées directement des sources primaires (directement ou indirectement) dans de nombreux domaines, et il est courant pour plusieurs noms différents (groupement, division, classification ... etc) selon les choses que vous traitez et les objectifs à atteindre. Le terme "classification" est d'abord associé à la définition de ses racines. Il est dérivé du verbe "classer" qui définit un verbe dans plus d'un domaine ou d'une série de méthodes plutôt que d'une théorie unifiée. Nous appelons la classification, l'algorithme de classification des objets. Elle consiste à attribuer une catégorie à chaque objet (ou individu) à classer, à partir de données statistiques. Les méthodes d'apprentissage sont couramment utilisées et utilisées pour reconnaître les modèles.[16]
D'une manière générale en vertu de ces définitions, la classification se définit alors comme une méthode mathématique d'analyse de données.

2.2 Domaines d'application et exemples de problème de classification

La classification joue un rôle important dans presque tous les domaines de la science et de la technologie qui ont des utilisations multidimensionnelles [17]. Par exemple, nous considérons les sciences biologiques : botanique, zoologie, écologie, etc., qui utilisent le terme «taxonomie» pour désigner la taxonomie. En plus des sciences de la terre et de l'eau : fossiles, astronomie, géographie, étude de la pollution, bénéficiant grandement des classifications. Il existe d'autres avantages aux techniques de taxonomie en sciences technologiques : robotique, mathématiques, chimie, physique, ingénierie, etc ... et sans oublier les techniques Produits dérivés tels que sondages, marketing, etc. Ce dernier utilise parfois les mots «classification» et «hachage» pour définir la classification. Il faut également mentionner la médecine, l'économie et le génie agricole. . . Etc,Dans toutes ces disciplines. Nous pouvons considérer la classification comme une région spécifique, mais elle est souvent considérée comme un complément à d'autres méthodes statistiques. Il est largement utilisé pour interpréter des graphiques, analyser des facteurs ou définir des groupes d'objets homogènes.***Nous mentionnons un exemple de son utilisation :***

2.2.1 Exemples de problème de classification

Quant aux différents types de données et entrées aux techniques de classification, il est nécessaire de les présenter avant de traiter les méthodes de classification, et ce sont aussi les domaines que la classification peut cibler.

⊞ ***Reconnaissance de formes***

La reconnaissance de modèle (RdF) est avant tout une réduction systématique de l'information basée sur la catégorisation d'objets ou de modèles [17]. Par conséquent, nous considérons souvent la reconnaissance de formes comme un problème de classification, c'est-à-dire un problème de réglage de fonction qui affecte chaque prédicat prévisible de la catégorie pertinente. Par exemple ces motifs (formes) peuvent être une image (visage, empreinte digitale),Voice (reconnaissance vocale), et bien d'autres. Comme reconnaître les caractères manuscrits. La figure 1 représente la reconnaissance de caractères manuscrits.

FIGURE 2.1 – Reconnaissance de caractères manuscrits

⊞ ***Prédiction :***

La classification est très utile sur les documents Internet quelle que soit la nature du document (image, fichier, son.) .Elle peut être utilisée et classer les documents selon leur signification (web sémantique et moteurs de recherche.Par exemple, nous mentionnons le courrier électronique et le SPAM corrects.

2.3 Principe général et étapes de la classification

Le processus de classification comprend les étapes suivantes :[10]

- Représentation des individus.
- Définition d'une mesure de similarité appropriée aux données.
- Classification.
- Abstraction des données.
- Validation du résultat.

2.3.1 Représentation des individus

A pour but de déterminer les informations concernant les données : le nombre de classes désiré, le nombre d'individus disponibles, le nombre, le type et l'échelle des attributs de données. Ces informations sont utilisées dans l'algorithme de classification

2.3.2 La proximité des individus :

Est souvent mesurée par une fonction de distance entre chaque pair d'individus. De nombreuses mesures de proximité sont proposées, se basant sur la nature de données.

2.3.3 Classification

Est une phase de groupement des individus dans les classes. Plusieurs algorithmes de classification sont proposés. La différence entre eux est la manière dont ils groupent les individus telles que la méthode hiérarchique, la méthode de partition. . . , le type de données qu'ils traitent comme des données numériques, de catégorie, le flux de données. . . , la mesure de proximité des individus et des classes qu'ils utilisent, telle que le critère selon lequel on construit des classes.[10]

2.3.4 Abstraction des données

Est un processus d'extraction d'une représentation simple et compacte pour un jeu de données. Typiquement, une abstraction des données est une description compacte de chaque classe, souvent en termes de prototypes des classes ou d'individus représentatifs des classes comme le centre des classes.

2.3.5 Validation du résultat

Vise à déterminer si les classes fournies sont significatives en utilisant un critère spécifique d'optimalité. Cependant, un tel critère est souvent subjectif, donc il y a peu de manière standard pour valider la classification sauf dans certains domaines bien décrits à priori.

2.4 Type de méthode de classification

Le type d'une méthode de classification se décline généralement en 2 familles : le mode supervisé, le mode non supervisé. Si l'on dispose d'un ensemble de points étiquetés, on parlera de classification supervisée. Dans le cas contraire, on effectue une classification non supervisée appelée également classification automatique.[15]

2.4.1 Classification supervisé

Les classes d'appartenance des données sont connues. La recherche des frontières entre les classes peut être effectuée par la recherche des fonctions discriminantes [15].

2.4.2 Classification non supervisé (classification automatique)

La recherche d'une partition des données revient à regroupercelles-ci selon une certaine mesure de similarité ou de dissimilarité [15].

2.5 Méthodes de classification supervisée

La classification supervisée (classement ou classification inductive) a pour objectif « d'apprendre » par l'exemple. Elle cherche à expliquer et à prédire l'appartenance de documents à des classes connues a priori. Le «classement» est une méthode supervisée qui consiste à définir une fonction qui attribue une ou plusieurs classes à chaque donnée. Dans cette approche on suppose qu'un expert fournit auparavant les étiquettes pour chaque donnée, les étiquettes sont des classes d'appartenance. Ainsi c'est l'ensemble des techniques qui visent à deviner l'appartenance d'un individu à une classe en s'aidant uniquement des valeurs qu'il prend.[13]

2.5.1 Principe de la classification supervisée

La conception supervisée d'une structure de classeAest classée comme une division, qui à son tour est divisée en groupes limités par (a_i). Le fait que les objets G (s_i) soient classés de la même nature (des phonétique, caractères manuscrits, etc.). Sachant qu'elles étaient auparavant classées G ces classifications ont été effectuées par un "encadrant" dans les groupes A qui constituent un groupe d'apprentissage. Le système de classification supervisé sera conçu sur la base des exemples du superviseur (le groupe d'apprentissage où, pour tout exemple, nous connaissons l'intention de sa classe).[13] De tout cela, nous cherchons à prédire si l'objet "s_i" de la base de données, qui a un attribut du groupe "d", appartient ou n'appartient pas à la classe "a_j" entre N classes, nous l'exprimons comme suit :

$$G = (s_1, a_2), (s_2, a_4), (s_3, a_2).....(s_i, a_j)\ s_i \epsilon R^d, a_j \epsilon A$$

de sorte à minimiser les mauvais classements $\tau(x_i) \neq c_j$,La classification supervisée tente de chercher, à partir des données de de G , une fonction de décision τ qui va associer à tout nouveau élément s_i de test une classe a_j, puis on compare ce que nous a donné cette fonction avec la classe connai a priori de cet élément , donc l'objectif est de chercher à prédire la classe de toute nouvelle donnée.

2.5.2 Exemple

On pourrait donner l'exemple le plus connu : les superviseurs sont généralement les médecins afin de noter la classe des objets de l'ensemble d'apprentissage à partir des remarques constatées.Ou bien l'exemple d'un tableau où le dernier descripteur (maladie) représente la classes des exemples [13] :

Fievre	Douleur	Toux	Maladie
oui	Abdomen	non	Appendicite
non	Abdomen	oui	Appendicite
oui	gorge	non	rhume
oui	gorge	oui	rhume
non	gorge	oui	mal de gorge
oui	non	non	aucune
oui	non	oui	rhume
non	non	oui	refroidissement
non	non	non	aucune

FIGURE 2.2 – exemple de malade pour la classification

Ce tableau contient les données d'entraînement pour la classification, chacune étant classée comme «appendicite», «froide», «mal de gorge» et «aucune». Qu'est-ce qui permet de construire un modèle de classification permettant de prédire qu'une personne est malade et quelle est la cause de sa maladie ou qu'elle est malade.

2.5.3 Classification bayésienne

La classification bayésienne est basée sur la théorie bayésienne [19]. Il s'agit d'un simple classeur probabiliste linéaire, qui suppose que les descripteurs (attributs) qui décrivent les choses dans un ensemble d'apprentissage sont indépendants. L'approche bayésienne vise à réduire le risque d'erreur de classification.

Principe de l'algorithme

L'ensemble d'apprentissage «A» est connue, et on cherche à classer un nouveau document «d_{new}» Le Classifieur bayésien va choisir la classe «C_k» (chaque objet est étiqueté par sa classe «C_k») qui a la plus grande probabilité, on parle de règle MAP ***(maximum a posteriori)***[7] :

$$C_{MAP} = argmax_{C_k \in c} P(C_k \backslash d_{new}) = argmax_{C_k \in c} P(d_{new} \backslash C_k) \frac{P(d_{new} \backslash C_k) P(C_k)}{P(d_{new})} = argmax_{C_k \in c} P(C_k \backslash d_{new}) P(C_k)$$

Il faut estimer les probabilités $P(C_k)$ et $P(d_{new} \backslash C_k)$ à partir des données d'apprentissage. Les probabilités a priori des classes $P(C_k)$ peuvent être estimées facilement par :

$P(C_k) = \frac{nP(C_k)}{nA}$ (C_k)=nombre d'expression d'apprnetissage dans la classe C_k ***et*** nA =nombre totale de documents d'apprnetissage

Pour estimer les valeurs de $P(C_k \setminus d_new)$ puisque les descripteur(attributs) de d_new sont indépendants , alors on aura grace aux théories d'indépendance bayésienne entre les variables :

$P(d_{new} \setminus C_k) = P(f_1 \setminus C_k)P(f_2 \setminus C_k)P(f_3 \setminus C_k).....P(f_{n_F}) \setminus C_k)$

2.5.4 K plus proche voisin

Méthode k du plus proche voisin (k-NN ou k-PPv) : La méthode k du plus proche voisin (k-NN) est utilisée pour classer les points cibles en fonction de leur distance à un échantillon d'apprentissage La méthode k-PPv est une approche de classification automatique, et est une généralisation des méthodes de classification inductive.

Le principe général de la méthode k-NN et méthode k-PPV est le suivant :

*** K-NN**

Etant donné une base d'apprentissage correctement étiquetée, le classificateur k-NN détermine la classe d'un nouvel objet en lui affectant la classe majoritaire des (x) objets dans la base de données. Dans ce contexte, nous avons une base de données d'apprentissage constituée de (N) paires entrées-sorties. Pour estimer la sortie associée à une nouvelle entrée (x), la méthode des k voisins les plus proches consiste à prendre en compte les (k) échantillons d'apprentissage dont l'entrée sont les plus proches de la nouvelle entrée (x), selon une distance prédéterminée. Généralement, la détermination de la similitude est basée sur la distance euclidienne. L'algorithme illustre une prise de décision basée sur la recherche d'un ou plusieurs cas similaires. En effet, l'algorithme recherche les k voisins les plus proches du nouveau cas et prédit la réponse la plus fréquente en classant les points cibles en fonction de leur distance aux points dans la base d'apprentissage.[8]

*** K-PPV**

Cette technique étend la méthode des k plus proches voisins selon deux voies[9] :

1. Tout d'abord, un schéma de pondération des plus proches voisins est introduit en fonction de leur similarité avec la nouvelle observation à classer.

2. Basé sur le fait que le vote des plus proches voisins est équivalent au mode de la distribution de la classe, la seconde extension utilise la médiane ou la moyenne de cette distribution, si la variable cible est relative à une échelle ordinale ou de niveau plus élevé. Cette extension est fondée sur l'idée que les observations de l'échantillon d'apprentissage, qui sont particulièrement proches de la nouvelle observation (y,x), doivent avoir un poids plus élevé dans la décision que les voisins qui sont plus éloignés du couple (y,x).

2.5.5 Classification par arbre de décision

*** Définition d'un arbre de décision**

L'arbre de décision est un diagramme schématique des résultats utilisés pour extraire les données, est un ensemble de règles de classification basant leur décision sur des tests associes aux attributs, organises de manière arborescente[9].

Structure générale d'une Arbre de décision :

- ***Nœuds internes***(nœuds de décision) :étiquètes par des tests applicables a toute description d'une instance.
- ***Arcs issus d'un nœud interne*** :réponses possibles au test du nœud.
- ***Feuilles de l'arbre*** : étiquetées par une classe.
- ***Chaque nœud interne ou feuille*** est repéré par sa position (liste des numéros des arcs qui permettent d'y accéder en partant de la racine).
- une ***règle*** est générée pour chaque ***chemin*** de l'arbre (de la racine à une feuille).
- ***Les paires attribut-valeur*** d'un chemin forment une conjonction.
- ***Le nœud terminal*** représente la classe prédite.

Les règles sont généralement plus faciles à comprendre que les arbres.

Construction de l'arbre : La construction d'un arbre de décision a l'aide d'attributs nominal est un processus itératif utilisant les étapes suivantes :

- Décider si un nœud est terminal.
- Si un nœud n'est pas terminal, lui associer un test.
- Si un nœud est terminal, lui associer une classe.

*** Algorithmes de classification par arbre de décision**

1. Algorithme CART ***"Classification And Régression Tree'***

   CART proposés par Breiman, Friedman, Olshen et Stone (1984), permettent de construire de manière simple et rapide des classificateurs ou des répresseurs constants par morceaux à partir d'un échantillon d'apprentissage.

Avantage :

- utilisation de variables de tous type "Continues, discrètes, catégoriques".
- Traitement d'un grand nombre de variables explicatives.

Inconvénients :

- Arbres non optimaux.
- Nécessité d'un grand nombre d'individus.
- Temps de calculs importants.

2. Algorithme ID3 « Itératif Dichotomiser 3 »

   Est un algorithme inventé par Ross Quinlan utilisé en 1986 pour générer un arbre de décision à partir de l'ensemble de données.

   - ***ID3*** : est généralement utilisé dans la machine apprentissage et de traitement du langage naturel et ne prendre en compte que les attributs nominaux (qualitatifs) L'idée de cet algorithme est de choisir l'attribut qui sera placé en racine.
   - ***Algorithme C4.5*** : Est un algorithme de classification supervisé, publié par Ross Quinlan. En 1993 Il est basé sur l'algorithme ID3 auquel il apporte plusieurs améliorations.il prend en charge les attributs quantitatifs.
     Il se base sur une mesure de l'entropie dans l'échantillon d'apprentissage pour produire le modèle (graphe d'induction).

*** Classification à base de règles (classification associative)**

L'une des méthodes d'apprentissage supervisé est la classification associative. Cette dernière crée des règles d'association et les analyse pour les utiliser dans la classification. L'idée qui s'appuie sur elles est de trouver les règles solides qui existent dans la partie appelée "classe".
Diverses études ont mis en évidence les différents avantages de la classification associative et de la comparaison avec les arbres de décision, ce qui permet de maintenir des résultats clairs et de réduire le taux d'erreur.

*** Algorithmes de classification à base de règles**

1. Algorithme ZeroR [11] :
   ZeroR est le classifieur le plus simple, il se base sur la classe et ignore tous les attributs. Il prédit tout simplement la classe majoritaire. Il n'a aucune capacité de prédiction mais représente une base de comparaison pour les autres méthodes.

2. Algorithme CBA[11] :
   L'un des premiers algorithmes de classification associative est l'algorithme CBA (Classification-Based Association). Il utilise l'algorithme Apriori pour générer les règles d'association puis utilise une heuristique pour construire le classiffieur. Les règles sont ordonnées selon leurs supports et confidences. Si plusieurs règles ont la même partie gauche, la règle de la confidence la plus élevée est utilisée dans le classifieur. Pour classer un nouveau tuplet, la première règle le satisfaisant est utilisée. Le classifieur contient aussi une règle par défaut pour classer les tuplet dont une règles satisfaisante n'existe pas.

Support vecteur machine (SVM)

SVM a été initialement conçu pour les tâches de classification ou de reconnaissance de formes et est utilisé pour traiter les problèmes de régression linéaire. Dans ce contexte, nous préférons prêter attention à la classification supervisée. Dans ce contexte, l'abréviation SVM peut être traduite de manière appropriée par de grands écarts de marge.
Le principe théorique de : SVM contient deux points de base

- la transformation non linéaire (Φ) des exemples de l'espace d'entrée vers un espace dit de redescription de grande dimension muni d'un produit scalaire (espace de Hilbert).

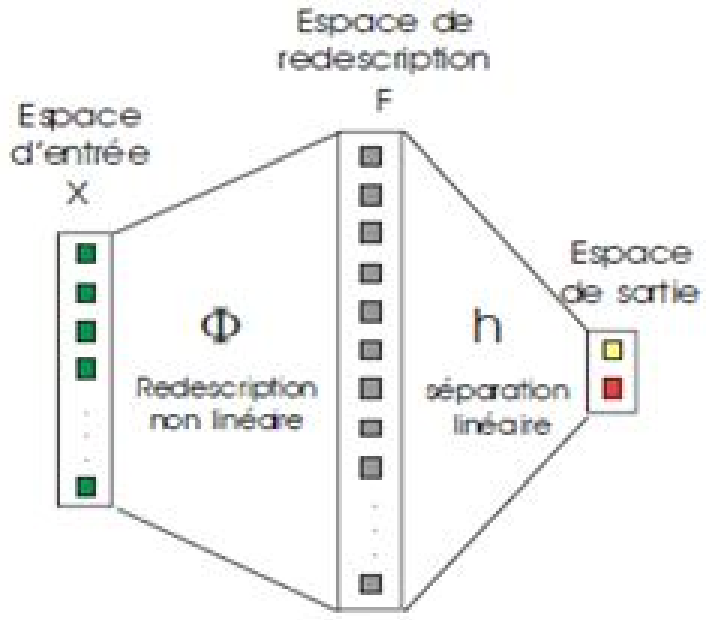


FIGURE 2.3 – la transformation non linéaire des données peut permettre une séparation linéaire des exemples dans un nouvel espace

- la détermination d'un hyperplan permettant une séparation linéaire Optimale dans cet espace de grande dimension.

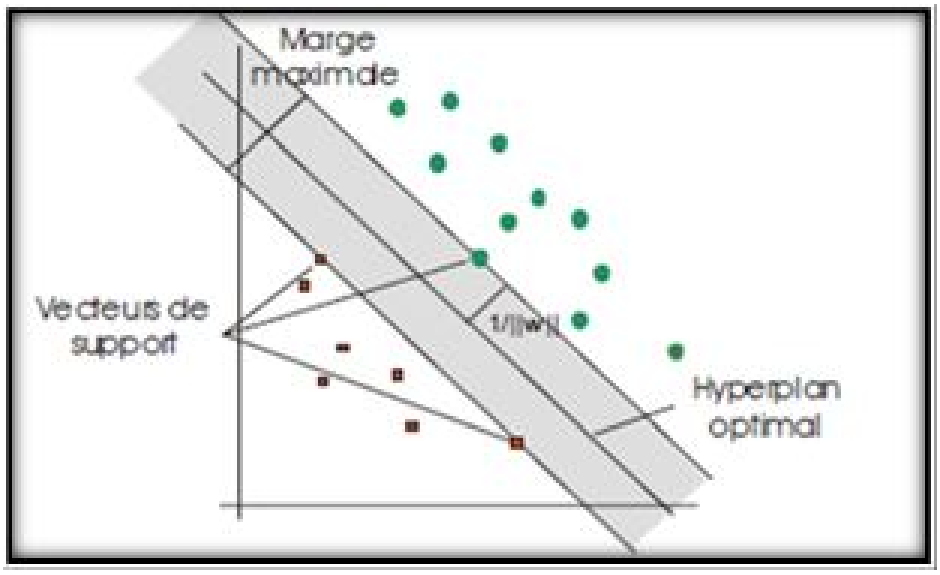


FIGURE 2.4 – l'hyperplan séparateur optimal est celui qui maximise la marge dans l'espace de redescription

SVM est une stratégie pour réduire les risques structurels, mais le problème est de trouver les limites de résolution qui séparent l'espace en deux régions à trouver l'hyperplan qui classifie correctement les données et qui se trouve le plus loin possible de tous les exemples. On dit qu'on veut maximiser la marge qui veut dire la distance du point le plus proche de l'hyperplan.

Dans le cas de la catégorisation des textes. les entrées sont des documents et les sorties sont des catégories. En considérant un classificateur binaire, on vaudra lui faire apprendre l'hyperplan qui sépare les documents appartenant à la catégorie et ceux qui n'en font pas partie [14].

Les SVM conviennent bien pour la classification de textes parce qu'une dimension élevée ne les affecte pas puisqu'ils se protègent contre le sur apprentissage. Autrement dit, il affirme que peu d'attributs sont totalement inutiles à la tâche de classification et que SVM permettent d'éviter une sélection agressive qui aurait comme résultat une perte d'information.

On peut se permettre de conserver plus d'attributs. Egalement une caractéristique des documents textuels est que lorsqu'ils sont représentés par des vecteur une majorité des entrées sont nulles.

Un aspect positif des SVM est qu'aucun ajustement des paramètres manuel n'est requis car ils ont l'habileté de trouver automatiquement des paramètres adéquats.

Réseaux de neurone

Un réseau de neurones est un assemblage de neurones connectés entre eux. Un réseau réalise une ou plusieurs fonctions algébriques de ses entrées, par composition des fonctions réalisées par chacun des neurones. La capacité de traitement de ce réseau est stockée sous forme de poids d'interconnexions obtenus par un processus d'apprentissage à partir d'un ensemble d'exemples d'apprentissage Il arrive souvent que les exemples de la base d'apprentissage comportent des valeurs approximatives ou bruitées. Si on oblige le réseau à répondre de façon quasi parfaite relativement à ces exemples, on peut obtenir un réseau qui est biaisé par des valeurs erronées[12].

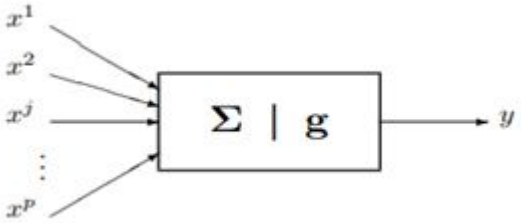


FIGURE 2.5 – Représentation d'un neurone

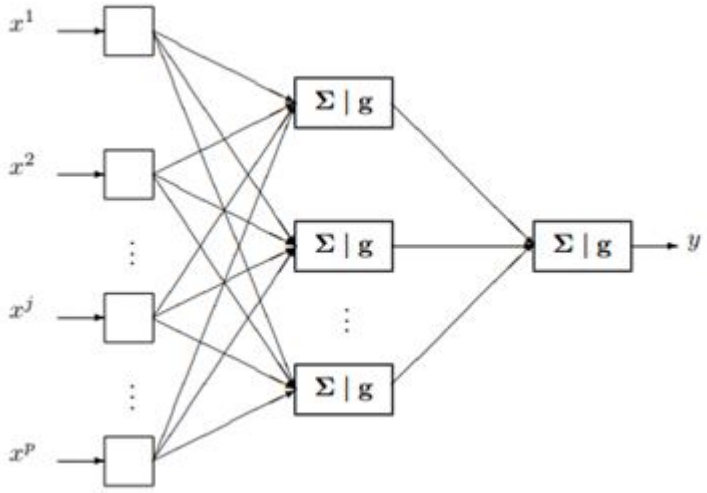


FIGURE 2.6 – Exemple de réseau de neurones

2.6 Conclusion

Nous avons traité de manière générale dans ce chapitre les concepts de classification et une vue générale et superficielle des principes généraux de classification et des méthodes de classification, la classification peut s'appuyer sur des méthodes supervisées et des concepts de proximité (voisins les plus proches) ou même des recherches dans les espaces de l'hypothèse (arbres de décision).
L'approche supervisée est utilisée pour de vastes zones, mais il n'y a pas de stratégie pour les exemples d'auto-apprentissage (c'est-à-dire l'apprentissage à partir d'une base sans connaissance préalable) et cela est reconnu par tous.

Chapitre 3

POPULATION DE L'ETUDE

Introduction

En science des données, résoudre des problèmes et répondre à des questions par l'analyse est une pratique standard qui a besoin d'un ensemble suffisant de données utilisé afin de donner certaine pertinence aux résultats d'analyse. Parmi les techniques d'analyse on trouve l'apprentissage automatique où le problème principal lors de la réalisation d'expériences est de trouver des données. Cela est d'autant plus difficile s'il faut trouver des données étiquetées. Certaines données peuvent provenir de systèmes de gestion de base de données relationnelle pour lesquels vous pouvez utiliser des outils comme SQL, Apache et d'autres outils pour collecter des données. Dans d'autres cas, les données peuvent provenir d'environnements Hadoop tels que HBase et d'autres – ou de flux temps réel comme Apache Kafka. Dans d'autre cas les données n'existe pas, donc, le chercheur doit les collectés directement à partir du terrain utilisant plusieurs techniques parmi elle le questionnement (notre cas).

3.1 Contexte de l'étude : Faculté des sciences et de la technologie

On a effectué notre étude au sein de la faculté des sciences et de la technologie (FST) d'université Abbas Laghrour Khenchela.
FST est composé de six départements où chaque Département gère un groupe des spécialités différentes. Comme suit :

1. Département de science de la matière.
2. Département de génie civil.
3. Département de génie industriel.
4. Département de génie mécanique.
5. Le Département des sciences et de la technologie.
6. Département de mathématiques et d'Informatique : qui est le département qui est le champs de notre étude, il est divisé en deux grandes spécialisations : Mathématiques et Informatique.

La Faculté des Sciences et de la Technologies est organisée selon l'organigramme présent dans la figure3-1 l'organisation administratives et pédagogique de la Faculté.

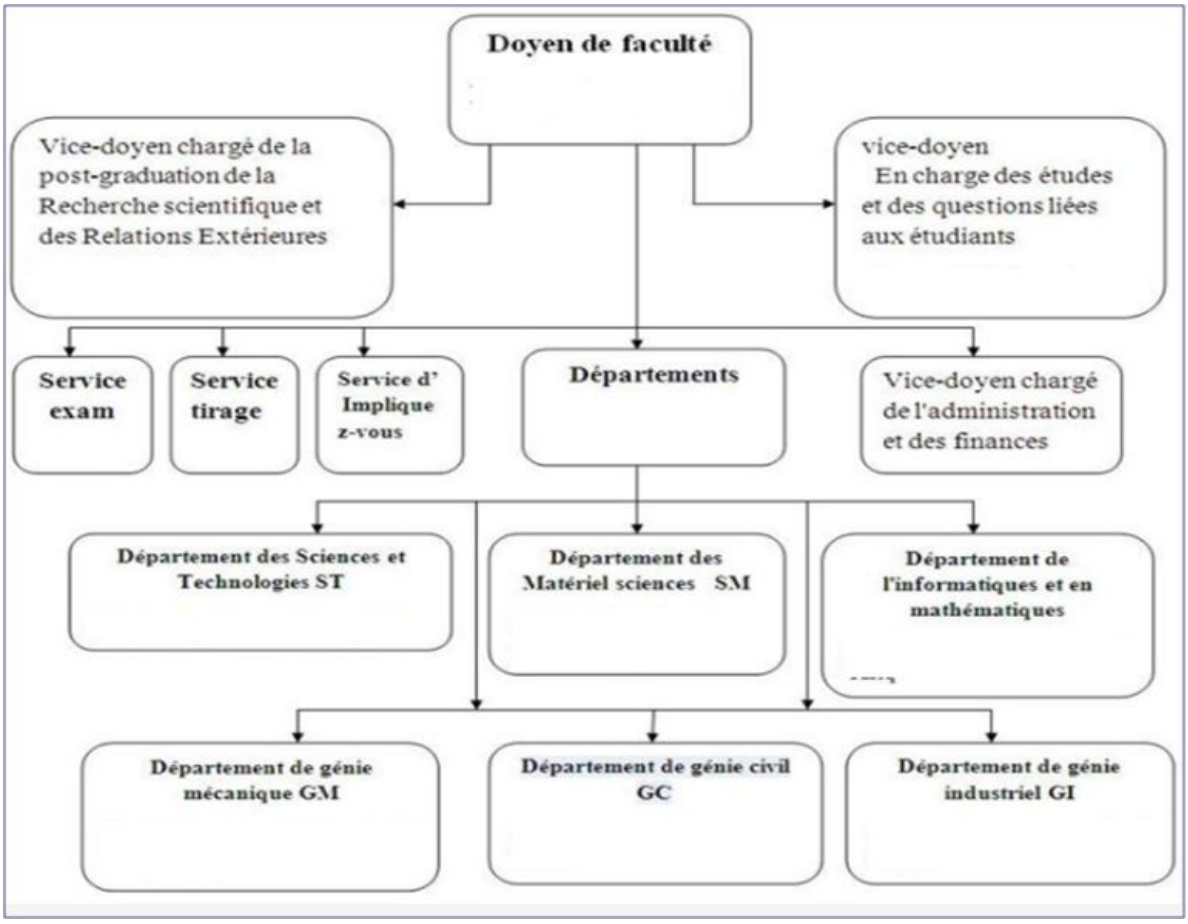


FIGURE 3.1 – Structure de la faculté des Sciences et de la technologie

3.2 Méthodologie

Pour accomplir notre objectif, et vu qu'il y a pas de données réelles dans l'établissement de l'étude, nous avons choisi de nous contacter directement la population de l'étude (étudiants) et d'obtenir les données qu'on à besoin pour les utilisées afin de générer des modèles de prédiction. Les données sont collectées via un questionnaire contenant des questions bien précises. Après la collection des réponses des étudiants sur le questionnaire on a conçue une petite application pour faire saisir toutes les réponses des étudiants afin de construire une base de données qui sera utilisé pour l'apprentissage et la génération des modèles de classification.

3.3 Collecte de données

La collecte des données consiste à enregistrer une ou plusieurs variables (longueur, durée, etc.) auprès des membres d'une population (unités de données).

3.3.1 Conception d'un questionnaire

Nous avons utilisé un questionnaire préparé au sein du laboratoire d'Informatique ***(ICOSI)*** et en le distribuant sur les étudiants qu'ils aient terminé leurs études ou poursuivent leurs études dans la spécialité. Le questionnaire contient un ensemble de questions réparties en six catégories comme suit :

A) ***Personnalité en première année*** :
Cette catégorie contient un ensemble de questions parmi eux :

- Le sexe de l'étudiant (homme ou femme).
- Quel âge avait son temps au moment de son entrée en première année de collège et sa situation familiale à cette époque, est-il célibataire, marié sans enfants, marié ou avec enfants ?
- Aussi, une question sur la résidence, si l'étudiant réside dans un wilaya ,daira, commune au village .
- En plus des questions sur la question de savoir si l'étudiant a des problèmes familiaux, une maladie chronique ou un handicap physique pouvant affecter ses résultats scolaires, il y a aussi des questions liées au niveau d'éducation des parents et de leurs professions.
  Exemple :
  ≈ ***Quel est votre état civil en première année ?***
  ☐ Célibataire.
  ☐ Marié(é) sans enfants Village.
  ☐ Marié(e) avec enfants.
  ≈ ***Avez-vous des problèmes familiaux en première année ?***
  ☐ Oui.
  ☐ Non.

B) ***Questions sur la vie quotidienne :***
Cette catégorie contient un ensemble de questions relatives à la vie quotidienne de l'étudiant, par exemple si l'étudiant fait un travail professionnel ou quel est le niveau de vie de l'étudiant (élevé, moyen ou bas).
L'étudiant possède-t-il son propre ordinateur et à quelle fréquence l'utilise-t-il ? Utilisez-le en ligne et découvrez où il se trouve (à la maison ou dans un cybercafé). Dans cette catégorie, l'étudiant détermine également comment il est arrivé à l'université (sur ses jambes, Dans le bus de l'étudiant ou dans une voiture privée). L'étudiant détermine également la distance entre sa résidence et l'université et le type de logement dans lequel il réside (la résidence des parents. Logement personnel ou universitaire).
Tout cela grâce aux options qui lui sont présentées dans le questionnaire qui lui a été

soumis.

Exemple :

≈ ***A quelle distance de l'université habitez-vous ?***

☐ Inférieure à 15 Km.

☐ Entre 15 et 20 Km.

☐ Entre 20 et 50 Km.

☐ Plus de 50 Km. .

≈ ***Votre niveau de vie est :***

☐ Elevé.

☐ Moyen.

☐ Faible.

C) ***Questions sur les études secondaires :*** Cette catégorie de questions se rapporte aux études au secondaire en posant un ensemble de questions dans ce domaine. Connaître les capacités de l'étudiant à l'échelle de la chimie, de la physique et des mathématiques, surtout en dernière année de lycée.
Aussi pour savoir si l'étudiant est redoublé en classe au lycée ou non, ainsi que pour suivre des cours spéciales pour l'aider à améliorer son niveau d'éducation.
Exemple :

≈ ***Lors de la Terminale, Quelle était votre performance en physique et chimie ?***

☐ Très bonne.

☐ Bonne.

☐ Moyenne.

☐ Passable.

☐ Faible.

D) Questions sur la situation pré-universitaire :
Le questionnaire contient un ensemble de questions liées à cet aspect, telles que la spécialité dans laquelle l'étudiant obtient le baccalauréat et combien était sa moyenne.
Exemple :

≈ ***Type du baccalauréat obtenu est ?***

☐ Mathématiques.

☐ Sciences expérimentales.

☐ Techniques Mathématiques.

☐ Autres Filière.

≈ ***Avez-vous aimé le domaine affecté ?***

☐ Oui.

☐ Non.

E) Les études de la première année universitaire à la première fois :
Cette catégorie de questions concerne les études en première année à l'université pour la première fois (c'est-à-dire avant la répétition de l'année) et comprend un ensemble de questions, parmi lesquelles nous mentionnons :
L'élève répond au domaine qu'il a étudié en première année. L'étudiant s'est-il adapté ou non au système universitaire ?, cette catégorie aborde également la relation de l'élève avec son professeur en raison de sa grande importance autour du niveau de l'étudiant, puisque l'enseignant est le cadre principal de l'étudiant.
Exemple :

≈ ***Quelle est votre domaine de 1ere année ?***

☐ MI.

☐ ST.

☐ SM.

☐ Autre.

≈ ***Assister vous TDs cours de la première années ?***

☐ Oui, toujours.

☐ Oui, rarement.

☐ Oui, Parfois.

☐ Non, Jamais.

F) Résultats de la première année universitaire (sans redoublement) :
Cette catégorie de questions concerne les résultats de l'étudiant obtenu au cours de sa première année universitaire pour la première fois.
Ce point concerne les résultats du premier et du deuxième semestre en plus des résultats finaux.
Exemple :

≈ ***Quelle est votre domaine de 1ere année ?***

☐ Admis la session normal.

☐ Admis après la session rattrapage.

☐ Admis avec dattes.

☐ Ajourné.

≈ ***Quels sont vos résultats finaux de la première année ?***

☐ Admis sans dattes.

☐ Admis avec dattes.

☐ Ajourné.

***Remarque :* Toutes les questions susmentionnées et dans toutes les catégories sont liées à la première année d'université pour l'étudiant pour la première fois (avant le redoublé de l'année).**

3.3.2 Saisie des données

Afin de faciliter le processus de collecte des données mentionnées dans le questionnaire expliqué précédemment, nous avons développé un petit logiciel qui permet à l'utilisateur de saisir des données dans une base de données via une interface graphique. Pour que chaque catégorie de questions on a la représentée par une interface graphique avec un ensemble de boutons qui permettent la saisie, la suppression ou la modification des données en cas de besoin. Après la saisie notre application génère un fichier avec une extension .csv qui peut être visualisé dans Excel et utiliser pour l'analyse et l'apprentissage.

L'interface graphique du logiciel de saisie des données

FIGURE 3.2 – Menu Principale

Cette interface principale contient un ensemble de boutons de sorte qu'une fois que vous cliquez sur un bouton, une interface liée à une catégorie de questions dans le questionnaire susmentionné s'ouvre.

Par exemple, en cliquant simplement sur le bouton "Personnalité l", l'interface suivante s'ouvre pour nous : L'interface de la figure 2 contient les différentes questions du questionnaire liées à la première catégorie de questions, qui sont des "questions sur la personnalité ". Chaque question contient au moins deux options où l'utilisateur clique sur l'option spécifique

FIGURE 3.3 – Interface pour les Questions de personnalité en première année

qui correspond au choix de l'étudiant dans le questionnaire.
Comme nous le voyons à la question 8, le choix 3 a été sélectionné.
L'interface contient aussi le numéro du questionnaire comme indiqué dans l'interface précédente, le numéro du questionnaire est 100.
En plus d'un groupe de boutons suivant :

◊ ***Le Bouton menu*** :Une fois que vous cliqué ce bouton, nous revenons à l'interface principale.

◊ ***Le Bouton sauvegarder*** :Une fois que nous avons cliqué sur ce bouton, après avoir répondu à toutes les questions, une boîte de dialogue apparaît pour nous qui indique à l'utilisateur que les données qu'il a choisies ont été enregistrées dans la base de données.

◊ ***Le bouton supprimer*** :Ce bouton permet de supprimer toutes les données relatives au numéro du questionnaire à supprimer, et ce en entrant le numéro du questionnaire dans la barre qui lui est affectée et en cliquant sur le bouton supprimer. Nous obtenons une boîte de dialogue indiquant que questionnaire a bien été supprimée comme suit.

Sortie de l'application

Une fois que toutes les données sont collectées et saisies, ces données sont enregistrées dans un fichier du format CSV (CommaSeparated Values), ce fichier est ouvert dans l'Excel pour inspection et préparation pour l'étape suivante, qui consiste à analyser ces données.

3.3.3 Analyse des données collectées

Le processus d'analyse des données de l'étude est de tire une vision globale sur ces données afin faciliter l'interprétation des résultats finaux.

Statistiques selon l'attribut « Sexe »

Les résultats de statistiques sur l'attribut sexe est présentés dans le tableau 1.

Sexe	***La fréquence***	***Pourcentage***
Male	105	42%
Female	145	58 %
total	***250***	***100%***

TABLE 3.1 – Statistiques sur le sexe des étudiants

Notre base d'apprentissage contient 250 attributs et représente le nombre total d'étudiants ce qui correspond au pourcentage 100%.
Là où nous constatons dans la caractéristique sexuelle que le pourcentage le plus élevé est des filles qui à était de 58%, avec un taux de 145 sur 250. Ensuite, le pourcentage d'hommes est de 42%, avec une fréquence de 105 sur 250.

Statistiques selon l'attribut « Age »

Le tableau 2 représente le rapport des âges des étudiants qui ont rejoint la première annéeé universitaire pour la première fois, où nous constatons que le pourcentage le plus élevé était pour les étudiants de 19 ans et ces pourcentages sont estimés à 57%, suivis par des étudiants qui ont dépassé l'âge de 20 ans de 24%. Nous notons que le pourcentage le plus faible concernait les étudiants de moins de 18 ans, estimé à 18,8%

Age	***Description***	***La fréquence***	***Pourcentage***
$\geqslant 20$	Il représente le pourcentage des étudiants de plu sou égal de 20 ans	60	24%
19	Il représente Le pourcentage d'étudiants âgés de 19 ans	143	57.2%
$\leqslant 18$	Il représente le pourcentage d'élèves dont l'âge est inférieur ou égal à 18 ans	47	18.8 %
total		***250***	***100%***

TABLE 3.2 – Statistiques sur l'âge des étudiants

Statistiques selon l'attribut « Résidences »

Dans tableau 3, nous notons que le pourcentage d'étudiants résidant dans la wilaya est le plus grand pourcentage, estimé à 45,2%, et le nombre d'étudiants pour ce ratio est estimé à 113 étudiants. Alors que le reste des étudiants sont répartis dans chacune des communes et des daïras alors qu'un petit nombre d'étudiants résidant dans les villages .

Age	***Description***	***La fréquence***	***Pourcentage***
DAI	Le pourcentage d'étudiants résidant dans daïra	65	26%
WIL	Le pourcentage d'étudiants résidant dans wilaya	113	45.2%
COM	Le pourcentage d'étudiants résidant dans commun	64	25.6%
VIL	Le pourcentage d'étudiants résidant dans village	8	3.2%
total		***250***	***100%***

TABLE 3.3 – Statistiques sur la résidence des étudiantes

Statistiques selon l'attribut « Niveau de vie» des étudiants

Cet attribut représente le niveau de vie des étudiants, nous constatons que la majorité des étudiants ont un niveau de vie moyen, avec un taux estimé à 89.2%. Alors que ceux dont le niveau de vie est bas sont beaucoup plus faibles par rapport au pourcentage précédent.

Niveau de vie	***Description***	***La fréquence***	***Pourcentage***
MED	Le pourcentage des étudiants dont le niveau de vie est moyen	223	89.2%
HIG	Le pourcentage d'élèves dont le niveau de vie est élevé	19	7.6%
LOW	Le pourcentage d'élèves dont le niveau de vie est faible	8	3.2%
total		***250***	***100%***

TABLE 3.4 – Niveau de vie des étudiants

Statistiques selon l'attribut « Posséder PC»

Ce tableau représente les changements dans le pourcentage et le nombre d'étudiants utilisant l'ordinateur pour leurs études, où nous avons constaté que la plupart des étudiants utilisent parfois l'ordinateur et un pourcentage moyen du nombre d'étudiants qui utilisent

l'ordinateur chaque jour. Alors qu'il y a un pourcentage très faible des étudiants qui n'utilisent pas l'ordinateur.

Posséder PC	***Description***	***La fréquence***	***Pourcentage***
YSO	Il représente le pourcentage des étudiants qui utilisent un ordinateur tous les jours	94	37.6%
YID	Il représente le pourcentage d'élèves qui utilisent parfois l'ordinateur	133	53.2%
NNE	Il représente le pourcentage d'élèves qui n'utilisent jamais d'ordinateur	23	9.2%
total		***250***	***100%***

TABLE 3.5 – Le pourcentage des étudiants qui utilisent un ordinateur

Statistiques selon l'attribut « Utilisation d'Internet»

En ce qui concerne l'utilisation d'Internet par les étudiants, nous remarquons que les résultats sont divisés comme suit.
50% concerne les étudiants qui utilisent quotidiennement Internet dans le cybercafé, qui est le pourcentage le plus élevé, car le nombre d'étudiants dans ce cas est estimé à 125 étudiants. Les 50% restants sont divisés en 4 groupes et comprennent l'utilisation quotidienne d'Internet à domicile, ce pourcentage étant estimé à 17%, soit le deuxième pourcentage le plus élevé. Alors que nous constatons que le nombre d'étudiants qui n'utilisent jamais Internet est de 11.6% Quant au pourcentage le plus bas, c'est le nombre d'étudiants qui utilisent parfois Internet et dans le cybercafé, où leur taux était estimé à 4.6% et ils sont 16 étudiants.

utilisant Internet	***Description***	***La fréquence***	***Pourcentage***
YSH	Il représente le pourcentage des étudiants qui utilisent Internet toujours et à domicile	44	17.6%
YIH	Le pourcentage d'étudiants qui utilisent Internet toujours et dans le cybercafé	125	50%
YSC	Il représente le pourcentage d'étudiants qui utilisent parfois Internet et à la maison	36	14.4%
YIC	Il représente le pourcentage d'étudiants qui utilisent parfois Internet et dans le cybercafé	16	6.4%
NNE	Il représente le pourcentage des étudiants qui n'utilisent jamais Internet	29	11.6%
total		***250***	***100%***

TABLE 3.6 – Le pourcentage et le nombre d'étudiants utilisant Internet

Statistiques selon l'attribut « performance en mathématique»

Ce point concerne la performance des étudiants en mathématiques au baccalauréat
Là où nous notons que la majorité des étudiants qui étudient les sciences et la technologie en général peuvent obtenir des résultats très satisfaisants, en mathématiques d'environ 59%, alors que nous constatons que les pourcentages des étudiants avec des résultats moyens et des étudiants avec des résultats acceptables dans les pourcentages raisonnables.
Cependant, les pourcentages des étudiants ayant obtenu de très bons résultats et les pourcentages des étudiants ayant obtenu de mauvais résultats étaient identiques à 4% comme montre le tableau 7.

perf en mathé	***Description***	***La fréquence***	***Pourcentage***
PAS	Il représente le pourcentage des é tudiants qui obtiennent un point passable	60	24%
GOO	Il représente le pourcentage des étudiants qui obtiennent un bon point	147	58.8%
MEY	Il représente le pourcentage des étudiants qui obtiennent un point moyen	23	9.2%
VGO	Il représente le pourcentage des étudiants qui obtiennent un très bon point	10	4%
LOW	Il représente le pourcentage des étudiants qui obtiennent un point faible	10	4%
total		***250***	***100%***

TABLE 3.7 – Performance desétudiants en mathématiques

Statistiques selon l'attribut « Résultats scolaires»

À ce stade, nous discutons des résultats scolaires au niveau secondaire pour les étudiants selon les résultats du tableau 8. Où l'on constate que la majorité des étudiants de notre base d'apprentissage ont obtenir des résultats satisfaisants avec un pourcentage de 62% et des résultats très satisfaisants avec un pourcentage de 27,6%. Quant au pourcentage des étudiants ayant obtenu des résultats insatisfaisants, il s'agit d'un faible pourcentage par rapport aux taux précédents. Alors que le pourcentage des étudiants ayant de très mauvais résultats est quasi inexistant.

Résultats scolaires	***Description***	***La fréquence***	***Pourcentage***
UNS	Le pourcentage des étudiants dont les résultats scolaires sont Très satisfaisants	69	27.6%
SAT	Le pourcentage des étudiants dont les résultats scolaires sont satisfaisants	155	62%
VSA	Le pourcentage des étudiants dont les résultats scolaires sont insatisfaisants	24	9.6%
VUS	Le pourcentage des étudiants dont les résultats scolaires sont Très insatisfaisants	2	0.8%
total		***250***	***100%***

TABLE 3.8 – Performance desétudiants en mathématiques

Statistiques selon l'attribut « Résultat du baccalauréat»

Le tableau 9 présente les pourcentages des résultats des étudiants au baccalauréat. Ces résultats peuvent être classés en trois catégories :
La première catégorie : c'est le pourcentage d'étudiants qui obtiennent une moyenne de baccalauréat entre 10 et 12 sur 20, et ce pourcentage est estimé à 55,6%. L'équivalent de 139 étudiants, ce qui est la plus grande valeur.
Deuxième catégorie : elle est liée au pourcentage d'étudiants qui obtiennent un taux de baccalauréat supérieur à 14 sur 20 et cette valeur est estimée à 6,4%, ce qui est la plus petite valeur. Où le nombre d'étudiants dans cette proportion est estimé à 16 étudiants.
Quant à la dernière catégorie, elle concerne le pourcentage d'étudiants recevant un baccalauréat allant de 12 à 14 sur vingt pourcentages, estimé à 38%. Le nombre d'étudiants est estimé à 95 étudiants.

Résul du bacc	***Description***	***La fréquence***	***Pourcentage***
10_12	Le pourcentage des étudiants ayant réussi le baccalauréat se situe entre 10 et 12	139	55.6%
>14	Le pourcentage des étudiants ayant réussi le baccalauréat est supérieur à 14	16	6.4%
12_14	Le pourcentage des étudiants ayant réussi le baccalauréat se situe entre 12 et 14	95	38%
total		***250***	***100%***

TABLE 3.9 – Résultat du baccalauréat

Statistiques selon l'attribut « le choix de domaine»

Tous les étudiants doit choisissent leur domaine d'études à l'université après l'obtention du baccalauréat. Ce tableau affiche les résultats selon est ce que le domaine dans lequel l'étudiant a étudié est de son choix ou non.
On note que la majorité des étudiants ont étudié leur domaine d'études, qui les ont eux-mêmes choisis parmi leurs trois choix lors de leur inscription à l'université, avec environ 69%.

le choix de domaine	***Description***	***La fréquence***	***Pourcentage***
YES	Le pourcentage d'étudiants qui ont choisi leur domaine d'études fait partie des spécialisations proposées	172	68.8%
NO	Le pourcentage d'étudiants qui n'ont pas choisi leur domaine d'études fait partie des spécialisations proposées	78	31.5%
total		***250***	***100%***

TABLE 3.10 – Le domaine choisi par l'étudiant pour étudier en première année d'université.

Statistiques selon l'attribut « Difficultés en première année»

La majorité des étudiants de première année ont rencontré des difficultés et des obstacles majeurs pour étudier à l'université, par exemple en changeant les programmes d'enseignement, la langue dans laquelle ils sont enseignés, etc.
Ce tableau montre les pourcentages d'étudiants qui ont eu des difficultés à étudier en première année d'université. D'où le pourcentage de ces étudiants est 67.2% . Quant au reste, il concerne les étudiants qui n'ont rencontré aucune difficulté lors de leur première année universitaire.

Diffic en 1ér année	***Description***	***La fréquence***	***Pourcentage***
YES	Le pourcentage des étudiants a yant rencontré des difficultés en première année	168	67.2%
NNO	Pourcentage des étudiants n'ayant pas rencontré de difficultés en première année	82	32.8%
total		***250***	***100%***

TABLE 3.11 – Le pourcentage d'étudiants ayant rencontré des difficultés, en première année universitaire.

Statistiques selon l'attribut « Universite_Like»

Ce tableau présente les pourcentages d'étudiants qui aime l'université et qui se joignent parfois à eux et aux étudiants qui détestent l'université, ce qui affecte leurs résultats et leurs performances tout au long de leur carrière universitaire. Où nous constatons que la

majorité des étudiants n'aiment pas l'université et sont absents de leurs études, alors que le pourcentage d'étudiants qui vont à l'université a toujours été un pourcentage significatif par rapport au précédent.

Universite Like	***Description***	***La fréquence***	***Pourcentage***
NO	Le pourcentage d'étudiants qui ne vont pas à l'université	121	48.4%
YID	Le pourcentage d'étudiants qui fréquentent toujours l'université	27	10.8%
JDU	Représente le pourcentage d'étudiants qui détestent l'université	38	15.2%
YSO	Le pourcentage d'étudiants qui fréquentent parfois l'université	64	25.6%
total		***250***	***100%***

TABLE 3.12 – Les pourcentages d'étudiants qui fréquentent toujours l'université et qui détestent l'université.

Statistiques selon l'attribut « Révisez cours»

La révision des cours a toujours été un catalyseur majeur pour la réussite et la supériorité des étudiants dans leur parcours académique à tous les niveaux d'études. Le tableau 13 montre le pourcentage des étudiants revoyant leurs leçons. Le pourcentage le plus élevé des étudiants qui ont révisé leurs cours à la maison était que le pourcentage était estimé à 48%. Le deuxième pourcentage était pour les des étudiants qui avaient révisé leurs cours à la bibliothèque et était estimé à 25,2%. Le pourcentage des étudiants qui n'ont pas révisé leurs cours était de 17,2%, ce qui est un pourcentage significatif.

Révisez cours	***Description***	***La fréquence***	***Pourcentage***
NO	Le pourcentage d'étudiants qui ne révisent pas leurs leçons	43	17.2%
YAH	Le pourcentage d'étudiants révisant leurs cours à la maison	120	48%
YAL	Le pourcentage d'étudiants révisant leurs cours à la bibliothèque	63	25.2%
YOTH	Le pourcentage d'étudiants revoyant leurs cours ailleurs	24	9.6%
total		***250***	***100%***

TABLE 3.13 – Révisez cours

Statistiques selon l'attribut « Assister cours»

Ce tableau présente les résultats des étudiants qui assistent à leurs cours, où nous notons que le pourcentage le plus élevé concernait les étudiants qui assistent à leurs cours à un taux de 55,2%, ce qui équivaut 138 étudiants.
Alors que le pourcentage le plus faible d'élèves qui n'assistent jamais à leurs cours était de 12%, ce qui équivaut à 30 étudiants.sur un total de 250 étudiants.

Assister cours	***Description***	***La fréquence***	***Pourcentage***
YSO	Le pourcentage d'étudiants qui a ssistent parfois à leurs cours	138	55.2%
YID	Le pourcentage d'étudiants qui assistent toujours leurs cours	82	32.8%
NNE	Le pourcentage d'étudiants qui n'assistent jamais à leurs cours	30	12%
total		***250***	***100%***

TABLE 3.14 – Assister cours

Statistiques selon l'attribut « AssisterTDs»

Le tableau 15 fournit des résultats pour les étudiants assistent aux leurs TDs, où nous notons que le pourcentage le plus élevé était pour les étudiants qui assistent les TDs à un taux de 79,2%, ce qui équivaut à 198 étudiants.

Assister TDs	***Description***	***La fréquence***	***Pourcentage***
YID	Le pourcentage d'étudiants qui assistent toujours leurs TDs	198	79.2%
NNE	Le pourcentage d'étudiants qui n'assistent jamais à leurs TDs	13	5.2%
YSO	Le pourcentage d'étudiants qui assistent parfois à leurs TDs	36	14.4%
YON	Le pourcentage d'étudiants qui assistent rarement à leurs TDs	3	1.2%
total		***250***	***100%***

TABLE 3.15 – AssisterTDs

Alors que le pourcentage le plus faible d'étudiants qui assistent rarement les TDs est de 1,2%, ce qui équivaut à 3 étudiants sur un total de 250 étudiants, ce qui est un très faible pourcentage.
Alors que nous avons constaté que le pourcentage d'étudiants qui n'assistent pas TDS est de 5,2%, ce qui équivaut à 13 étudiants. Alors que le pourcentage d'étudiants assistent parfois les TDs était important, il était estimé à 14,4%.

Statistiques selon l'attribut « Révisez avec collègues »

La révision avec des collègues a toujours été la clé de la réussite et du développement des étudiants Et cela passe par l'échange d'idées entre eux et leur coopération pour comprendre les différents problèmes académiques, mais il y a un groupe d'étudiants qui préfèrent se réviser eux-mêmes plutôt que de les revoir avec leurs collègues pour de nombreuses raisons.
Dans ce tableau, nous montrons les différents pourcentages d'étudiants qui révisent leurs leçons avec leurs collègues et les étudiants qui préfèrent réviser eux-mêmes. Le pourcentage le plus élevé concerne les étudiants qui révisent parfois leurs leçons avec des collègues, et ce pourcentage est estimé à 46,8%. Alors que le pourcentage d'étudiants qui préféraient l'évaluation seule était de 10%, ce qui est un faible pourcentage par rapport au premier.
Pour les étudiants qui révisent toujours leurs cours avec leurs collègues, leur taux est de 23,2%. C'est une proportion acceptable.

Rév avec collèg	***Description***	***La fréquence***	***Pourcentage***
YSO	Le pourcentage d'étudiants qui revoient parfois leurs leçons avec leurs collègues	117	46.8%
YON	Le pourcentage d'étudiants qui revoient rarement leurs leçons avec leurs collègues	50	20%
NNE	Le pourcentage d'étudiants qui ne revoient jamais leurs leçons avec leurs collègues	25	10%
YID	Le pourcentage d'étudiants qui revoient toujours leurs leçons avec leurs collègues	58	23.2%
total		***250***	***100%***

TABLE 3.16 – Révisez avec collègues

Statistiques selon l'attribut « Utiliser la bibliothèque»

La bibliothèque a toujours été connue comme un moyen d'aider les étudiants à améliorer leur niveau d'éducation en utilisant des livres, des mémoires et diverses références qui y sont disponibles.
Le tableau 17 contient le nombre et le pourcentage d'élèves qui se rendent à la bibliothèque et à leur convenance. Lorsque nous constatons que la majorité des étudiants utilisent rarement la bibliothèque, un pourcentage d'environ 43%.
Alors que nous avons constaté que les étudiants fréquentant toujours à la bibliothèque sont très peu nombreux par rapport au premier, car ce pourcentage ne dépasse pas 8,4%.

Utili biblio	***Description***	***La fréquence***	***Pourcentage***
NNE	Le pourcentage d'étudiants qui n'utilisent jamais la bibliothèque	76	30.4%
YON	Le pourcentage d'étudiants qui utilisent rarement la bibliothèque	107	42.8%
YSO	Le pourcentage d'étudiants qui utilisent parfois la bibliothèque	46	18.4%
YID	Le pourcentage d'étudiants qui utilisent toujours la bibliothèque	21	8.4%
total		***250***	***100%***

TABLE 3.17 – Utilisation de la bibliothèque

Statistiques selon l'attribut « Absences»

Le tableau 18 montre les pourcentages et le nombre d'étudiants absents de leurs études. Là où nous le voyons dans ce tableau, environ 49% des étudiants manquent rarement. Quant aux étudiants qui sont définitivement absents de leurs études, leur taux est de 4%, soit 10 étudiants
Alors que nous avons constaté que 24,4% des étudiants ne manquent jamais leurs cours, et c'est une petite pourcentage par rapport à ce que contient notre base éducative.

Absences	***Description***	***La fréquence***	***Pourcentage***
YSO	Le pourcentage d'étudiants qui manquent parfois leurs études	57	22.8%
YON	Le pourcentage d'étudiants qui manquent rarement leurs études	122	48.8%
NNE	Le pourcentage d'étudiants qui ne manquent jamais leurs études	61	24.4%
YID	Le pourcentage d'étudiants qui manquent toujours leurs études	10	4%
total		***250***	***100%***

TABLE 3.18 – Absences

Statistiques selon l'attribut « Résultats du premier Semestre»

Ce tableau présente les résultats du premier semestre de la première année d'études universitaires en pour la première fois, où nous constatons que le pourcentage le plus élevé concernait les étudiants qui ont réussi le premier semestre à la session normale est de 54%.
Alors que le pourcentage d'étudiants n'ayant pas réussies au premier semestre était presque de 15%.
Le pourcentage le plus faible est pour les étudiants qui ont réussi le premier semestre avec des dettes était de 9,2%.

résu du	***Description*** *1ér Seme*	***La fréquence***	***Pourcentage***
AAR	Le pourcentage d'étudiants admis après la session rattrapage	55	22%
ASN	Le pourcentage d'étudiants admis a la session normale	135	54%
AAD	Le pourcentage d'étudiantsadmis avec dettes	23	9.2%
AJO	Le pourcentage d'étudiants ajourné	37	14.8%
total		***250***	***100%***

TABLE 3.19 – Les résultats du premier Semestre

Statistiques selon l'attribut « Les résultats du deuxième Semestre»

Ce tableau présente les résultats du deuxième semestre de la première année d'études universitaires pour la première fois, où nous constatons que le pourcentage le plus élevé d'étudiants qui ont réussi le deuxième semestre du session normale était de 51,6%.
Le pourcentage d'étudiants qui n'ont pas réussi au deuxième semestre était aussi presque de 15%.
Alors que le pourcentage le plus faible pour les étudiants qui ont réussi le deuxième semestre avec des dettes était de 6,8%.

résultat du 2ème Semestre	***La fréquence***	***Pourcentage***
AAR	67	26.8%
ASN	129	51.6%
AAD	17	6.8%
AJO	37	14.8%
total	***250***	***100%***

TABLE 3.20 – Les résultats du deuxième Semestre

Statistiques selon l'attribut « Les résultats finals»

Le tableau 21 présente les résultats finaux de la première année d'études universitaires pour la première fois, où nous constatons que le pourcentage le plus élevé était pour les étudiants qui ont réussi sans les dettes à 62%, ce qui est un bon pourcentage. Les étudiants qui ont passé la première année avec des dettes étaient de 21,6%, ce qui est un pourcentage

significatif aussi. Alors que le pourcentage d'étudiants qui ont échoué en première année de l'université était de 16,4%, soit 41 étudiants.

Les résultats finals	***La fréquence***	***Pourcentage***
AAD	54	21.6%
ASD	155	62%
AJO	41	16.4%
total	***250***	***100%***

TABLE 3.21 – Les résultats finals

3.4 Conclusion

Ces statistiques visent à établir une étude sur les différentes données collectées dans la base de données relatives aux étudiants de première année universitaire. Ceci est avant de générer un modèle de classification des résultats des étudiants et d'évaluation de leurs performances académiques dans le chapitre suivant. Ces modèles permettant de prédire les performances des nouveaux étudiants dans la même discipline.

Chapitre 4

IMPLEMENTATION ET REALISATION

Introduction

L objectif principal de l'implémentation de l'application de datamining après une série de plusieurs étapes dans le processus de développement est de développer des modèles de classification des étudiants universitaires de première année en Mathématiques et Informatique utilisant les différents algorithmes offert par l'outil Weka. Ces modèles nous permettent de prédire la performance de nouveaux étudiants dans la même discipline en utilisant ses informations personnelles.
Dans les chapitres précédents, nous avons fourni les concepts nécessaires à la conception et à la mise en œuvre de la classification et de la prédiction. Ce chapitre donne un aperçu de notre système et les outils utilisés pour le développement. Dans ce chapitre nous allons présenter en premier temps, l'environnement du développement avec les différentes bibliothèques utilisées, ainsi que les techniques de datamining utilisées pour implémenter ce type de système. Ensuite, on va présenter notre application de classification développé illustré par quelque résultat obtenu et enfin, nous terminons ce chapitre par une conclusion.

4.1 La conception du système de prédiction

4.1.1 Processus global de notre système

Après la phase de collecte et de formatage des données, l'étape suivante selon le processus d'extraction des données éducative (voir chapitre 1) est l'étape d'extraction et d'utilisation de connaissances où les méthodes de classification sont utilisées pour prédire les résultats des étudiants de la première année universitaire en Mathématiques et Informatique, ce qui permet de classer le nouvel étudiant et de prédire son rendement scolaire sur la base d'un ensemble de données. Recueilli auprès des données des étudiants des dernières années (voir chapitre 3).

figure suivante montre le processus global de notre système :

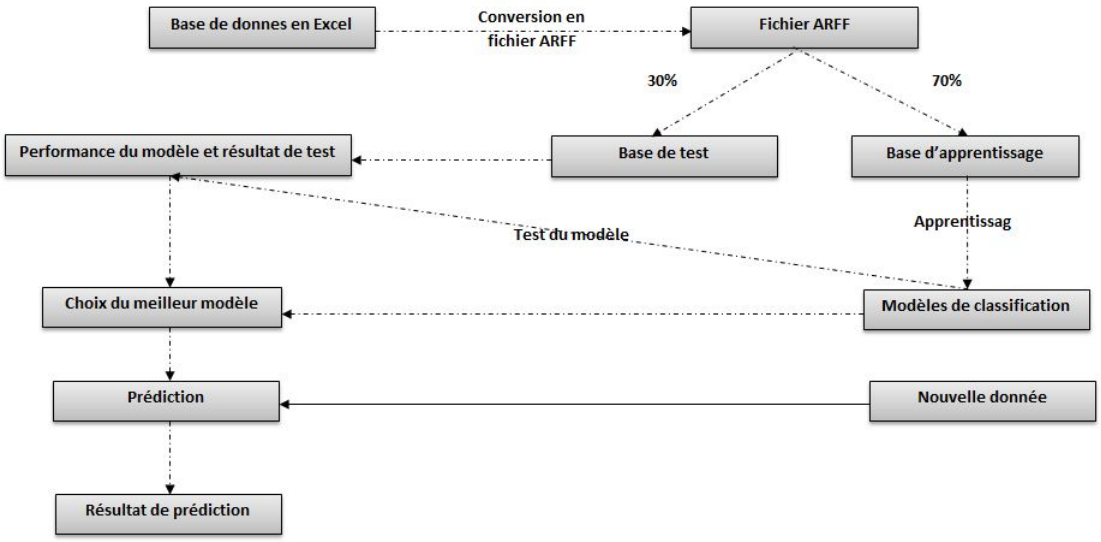


FIGURE 4.1 – Processus global de notre système

4.1.2 Environnement et outils de mise en œuvre

Notre système était développé sur un ordinateur de processeur Core i3 avec une RAM de 4 Go sous Windows 10, mais on peut implémenter ce système sur n'importe quel machine grâce au virtuel machine de Java.

Java

Java est un langage de programmation orienté objet, inventé par James Gosling en 1992 .alors qu'il travaillait dans les laboratoires Sun Microsystems - pour l'utiliser comme esprit pensant utilisé pour alimenter des dispositifs d'application intelligents tels que la télévision

interactive, et Java était un développement de C + +, Et à sa naissance, son créateur l'a appelé "chêne", c'est-à-dire le chêne, qui est l'arbre qu'il a vu depuis la fenêtre de son bureau alors qu'il travaillait dans les laboratoires de Sun Microsystems, puis le nom a été changé en Java, et ce nom (inhabituellement dans les langages de programmation de dénomination) n'est pas les premières lettres de Mots d'une phrase ou d'une expression particulière avec une signification spécifique, et pour vous C'est juste un nom développé par les développeurs de cette langue pour rivaliser avec d'autres noms.
Il est fourni avec un ensemble d'outils (le JDK Java Développement Kit) et un ensemble de packages : ensemble de classes. Ces différentes classes de base couvrent beaucoup de domaine (entrées/sorties, interface graphique, réseau, etc.) Cette richesse en "librairies standards" explique sûrement en partie le succès de Java. Le langage lui-même se trouve dans le package java. Lang .
Java est donc :

- Un langage de programmation orienté objets.
- Une architecture de machine virtuelle.
- Un ensemble d'outils.

Ses avantages : Le langage Java a des caractéristiques spéciales, ce qui en fait le langage de programmation le plus excitant. Ce qui le distingue sont les suivants :

- Facilité.
- Facilité d'accès.
- Transférable et exécutable.
- Jeux d'écriture et utilitaires.
- Créez des programmes avec une interface utilisateur graphique.
- Conception logicielle qui tire parti de tous les avantages d'Internet.

Le langage Java fournit un environnement interactif via le World Wide Web et est donc utilisé pour écrire des programmes éducatifs pour Internet via un logiciel de simulation informatique pour les expériences scientifiques et un logiciel de classe virtuelle pour le e-learning et l'enseignement à distance. L'efficacité de Java ne se limite pas au Web uniquement. Elle nous permet également de créer des programmes pour un usage personnel et professionnel. Ces programmes sont mis en œuvre à travers un certain nombre de programmes qui facilitent l'écriture de commandes telles que le programme NetBeans et Eclipse .

NetBeans

La programmation peut se faire pour des exemples simples avec le compilateur java, mais pour avoir plus de confort il est préférable d'utiliser un environnement de développement

intégré ou IDE, comme Eclipse ou NetBeans. Dans notre projet nous avons utilisé NetBeans.

NetBeans est un environnement de développement intégré (IDE) pour Java, placé en open source par Sun en juin 2000 sous licence CDDL (Common Développement and Distribution License). En plus de Java, NetBeans permet également de supporter différents autres langages, comme Python, C, C++, XML et HTML. Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages web).NetBeans est disponible sous Windows, Linux, Solaris (sur x86 et SPARC), Mac OS X et Open VMS. NetBeans est lui-même développé en Java, ce qui peut le rendre assez lent et gourmand en ressources mémoires.[22]

Version utilisée La version utilisée pour le développement de notre application est la version IDE 8.0.2

l'outil WEKA

WEKA (WAIKATO Environment for Knowledge Analysis) est un outil de fouille de données (licence GNU) développé en Java. [21] Il a été créé à l'université de Waikato, en Nouvelle-Zélande, par un groupe de chercheurs issus de l'apprentissage automatique, de la reconnaissance de formes et de la fouille de données.

WEKA permet de prétraiter des données (onglet Preprocess dans l'interface graphique), faire de la classification supervisée (Classify) et non-supervisée (Cluster), des régressions (Select Attributes), rechercher des règles d'association (Associate), et de visualiser différentes représentations graphiques des données (Visualize).[20]
Il s'agit d'un logiciel « open source » gratuit dédié à la classification et à la fouille de données. Il s'adresse à deux types de publics. D'un côté, il présente une interface graphique, le rendant ainsi accessible à une utilisation de type « chargé d'études » sur des données réelles. De l'autre, du fait que le code source est librement disponible et l'architecture interne très simplifiée, il se prête à une utilisation de chercheurs qui veulent avant tout expérimenter de nouvelles techniques en améliorant celles déjà implémentées ou en introduisant de nouvelles.

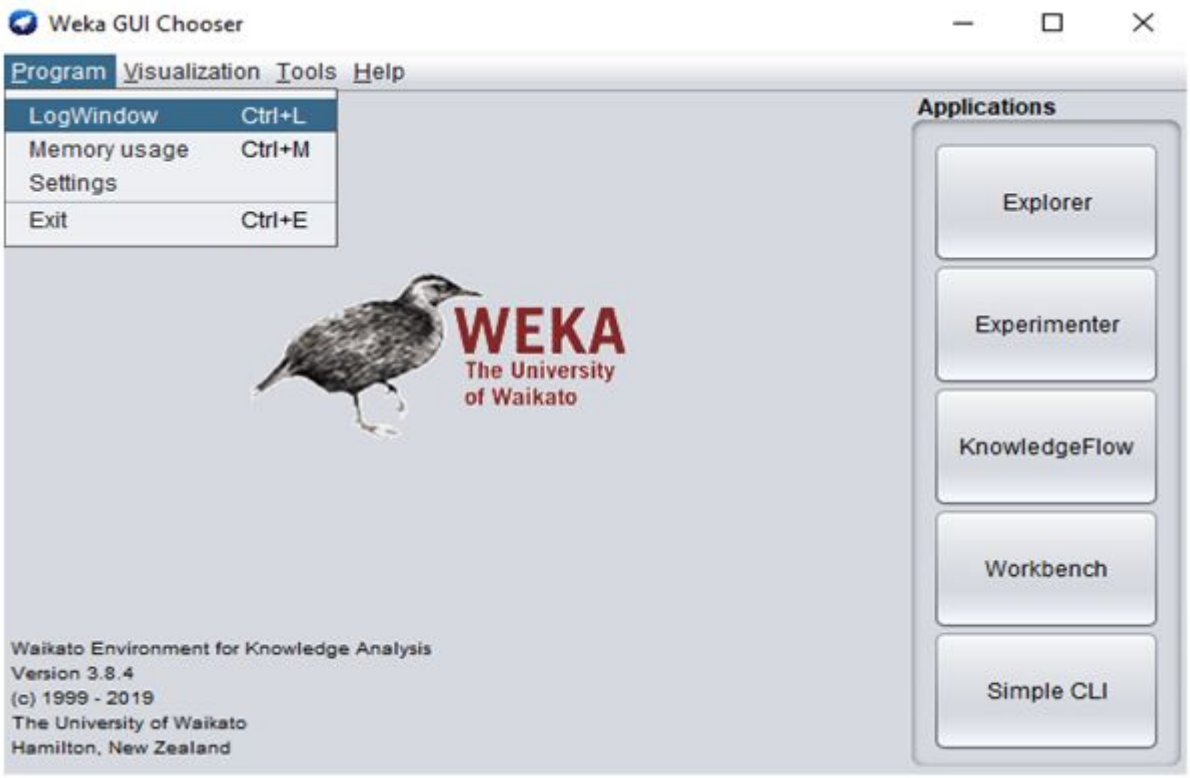


FIGURE 4.2 – Interface graphique de WEKA

Caractéristiques principales

⊙ Plus de 49 outils de prétraitement de données.

⊙ Plus de 76 algorithmes de classification régression regroupés en 07 familles.

⊙ Plus de 8 algorithmes de "clustering".

⊙ Plus de 15 évaluateurs d'attributs et plus de 10 algorithmes pour la sélection d'attribut.

⊙ 3 algorithmes de recherche de règles d'association.

⊙ 3 interfaces graphiques GUI.

⊙ « Explorer» (explorateur d'analyse de données).

⊙ « Expérimenter » (environnement expérimental).

⊙ «KnowledgeFlow » (le nouveau modèle de processus avec interface).

Structure de données

WEKA traite des données contenues dans des fichiers respectant le format ARFF Attribute-Relation File Format. Il s'agit de fichiers de type texte, décrivant des ensembles de "tuples" caractérisés par un certain nombre d'attributs communs.

Format d'un fichier ARFF (Attribute-Relation File Format) WEKA utilise (entre autres) le format de fichier arff pour enregistrer les données. Un fichier arff est composé d'une liste d'exemples définis par leurs valeurs d'attributs. Un fichier arff comprend toujours trois types d'informations : un nom pour la base de données, des attributs et des données. La chaîne de caractères @RELATION permet de donner un nom à la base de données. Par exemple, dans le cas du fichier Data.arff, le nom donné est Data. @RELATION Data. ***La chaîne de caractères @ATTRIBUTE permet de définir un attribut. Un attribut peut être de 4 types :***

- réel (NUMERIC ou REAL).
- Nominal (valeurs-possible) ***par exemple :*** @attribute Sexe FEM,MAL signifie que l'attribut Sexe peut avoir comme valeur soit Sexe-FEM ou soit Sexe-Mal.
- Chaîne de caractère (STRING).
- Date (date [<date-format>] @data : suivi d'une instance par ligne. Les valeurs d'instance sont séparées par une virgule.

Remarque préliminaire : le caractère « % » marque les lignes de commentaires. WEKA propose un éditeur de fichier arff (tools $\longrightarrow$ arffViewer) permettant de visualiser les fichiers arff sous la forme d'un tableau, et éventuellement de les modifier.

Exemple de fichier ARFF

@relation Data.
@attribute Sexe FEM,MAL.
@attribute Age >=20,19,'<=18 '.
@attribute Civil_ State SIG.
@attribute Residence DAI,WIL,COM,VIL.
@attribute Fam_ Problems NO,YES.
@data.
FEM , >=20 , SIG , DAI , NO , NO.
FEM , 19 , SIG , DAI , NO , NO.
MAL , 19 , SIG , DAI , NO , NO.
MAL , 19 , SIG , COM , YES ,NO.

4.1.3 Vue globale sur l'application

Objectif

Pour mettre en place notre application, nous avons réalisé une application sous l'environnement NetBeans qui utilise les algorithmes offert par le Weka via une interface facile à utilisé. Donc, notre application appel les packages et les différentes classes nécessaire de Weka dans un programme Java afin d'assurer les fonctionnalités de d'apprentissage, de test et de prédiction fixés par notre étude.

L'application a pour but est :

- D'appliquer et comparer certains algorithmes de classification pour construire le modèle de classification.
- Évaluation de performances et choit du meilleur modèle qui donne des bonnes performances sur les données de notre étude.
- D'utiliser un modèle pour la prédiction des résultats d'un étudiant via ces informations.

4.2 Expérimentations

Comme cité dans la section de la présentation de l'outil WEKA, les classificateurs sont regroupés par famille nous avons essayé d'appliquer au moins un classificateur de chaque famille, voire deux classifieurs, sur notre corpus sans faire aucun prétraitement.

4.2.1 Résultats de classification

1. **J48** :Le résultat obtenu est un Arbre de décision généré par l'algorithme J48 pour la classification des étudiants selon les résultats du premier semestre.
   WEKA nous fournit aussi une visualisation graphique de l'arbre de décision de classification. Cela est obtenu utilisant Weka et en sélectionnant "Visualisation de l'arbre".
   L'arbre de décision obtenu est décrit dans la Figure suivante. L'arbre de la figure suivante est considéré comme un modèle de classification de l'attribut R_ Sem1 (résultat du premier semestre).
   - **classification de l'attribut R_ Sem1 (résultat du premier semestre)**

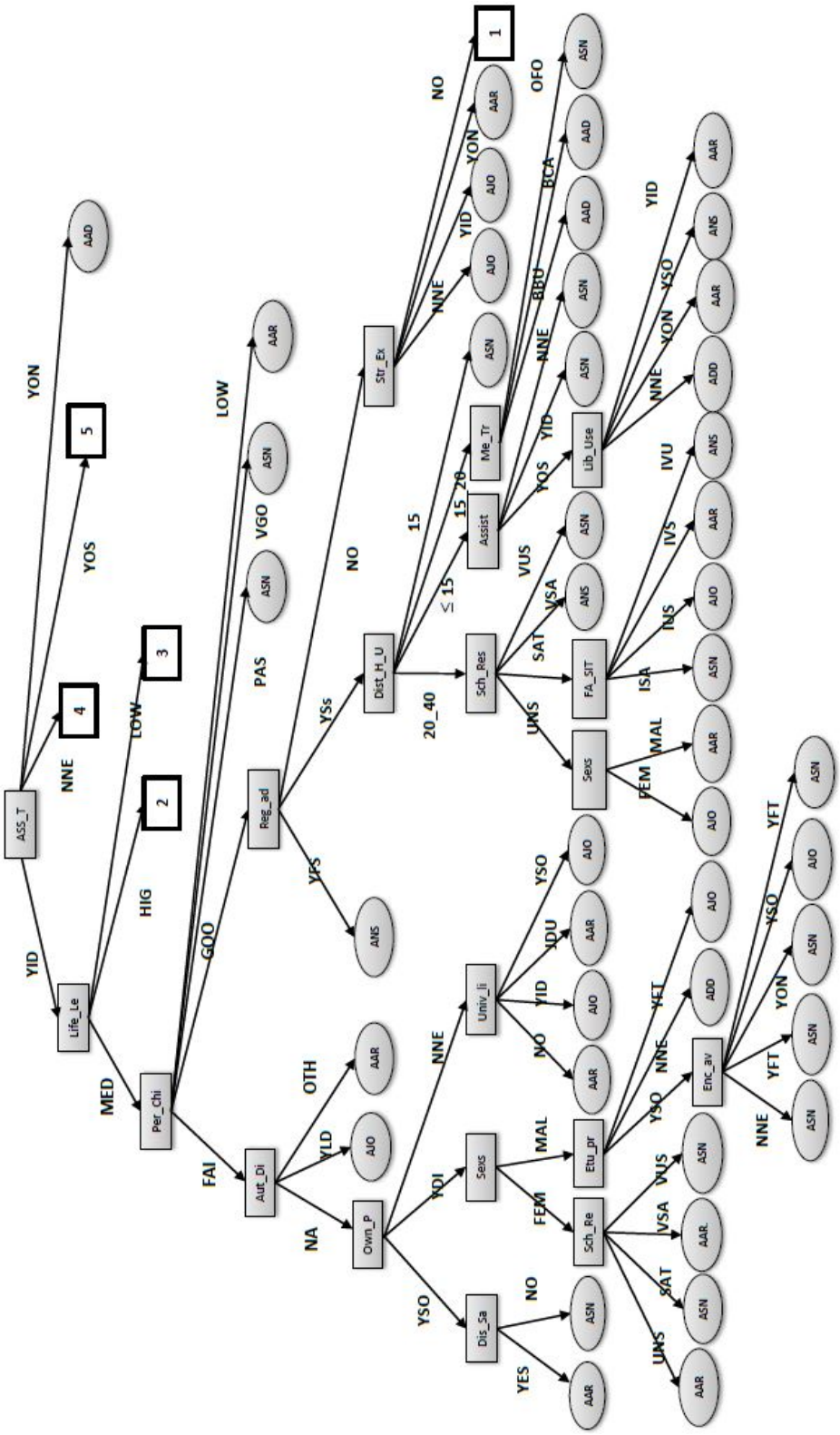
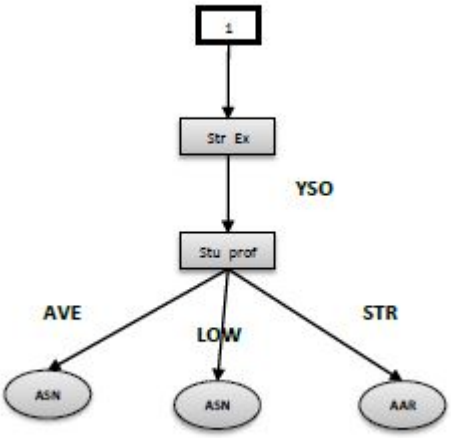
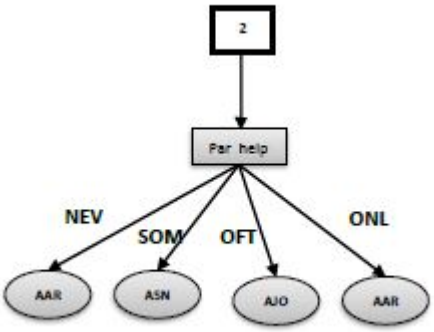
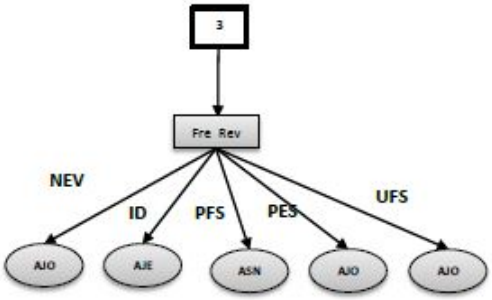
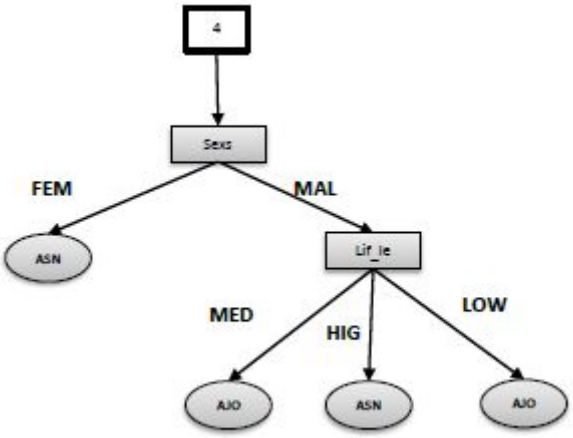


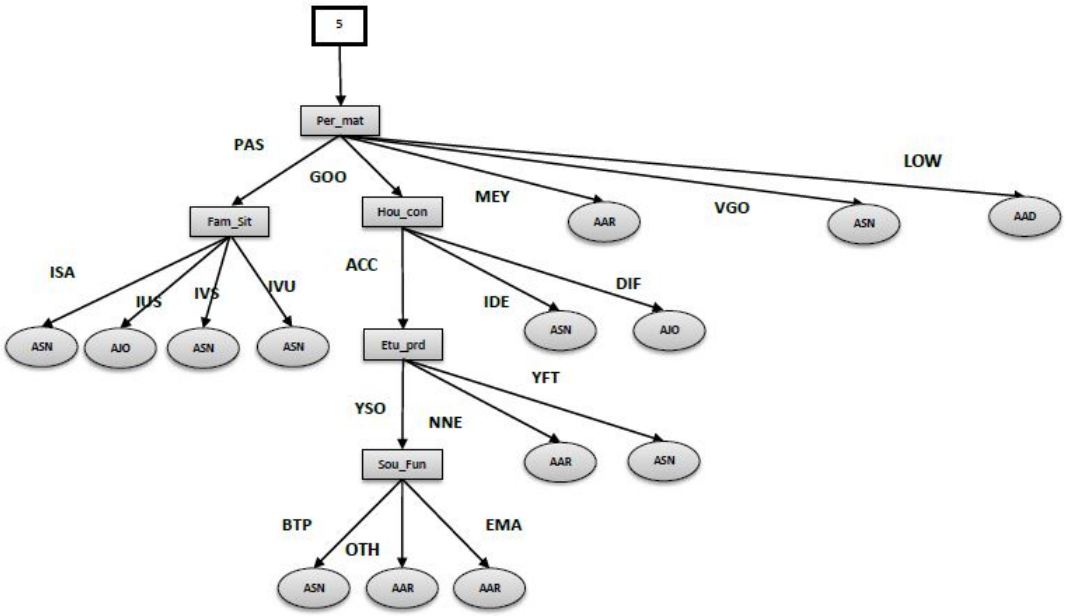
FIGURE 4.3 – Arbre de décision selon l'attribut R_ Sem1 utilisant l'algorithme J48


1
Str_Ex
YSO
Stu_prof
AVE
LOW
STR
ASN
ASN
AAR


2
Par_help
NEV
SOM
OFT
ONL
AAR
ASN
AJO
AAR


3
Fre_Rev
NEV
ID
PFS
PES
UFS
AJO
AJE
ASN
AJO
AJO


4
Sexe
FEM
MAL
ASN
Lif_le
MED
HIG
LOW
AJO
ASN
AJO



	Nombre d'instances	***Pourcentage***
Instances correctement classées	32	42.6667%
Instances incorrectement classées	43	57.3333%
total	***75***	***100%***

TABLE 4.1 – Résultats de classification par l'algorithme J48 selon l'attribut « Résultat du premier semestre »

(a) ***confusion matrix :***

a	b	c	d	$<--classified$
2	9	1	1	$\|a = AAR$
10	29	1	5	$\|b = ASN$
3	3	0	1	$\|c = AAD$
1	7	1	1	$\|d = AJO$

TABLE 4.2 – Matrice de confusion de la classification selon l'attribut R_Sem1 utilisant l'algorithme J48

(b) ***de l'arbre de la figure 4-3*** :

On a remarqué dans l'arbre de la figure 4-3 qu'il ya huit niveaux et 23 attributs présent dans l'arbre (Assist_td ,Life_Level ,Sexe ,Performance_Mat ,Performance_Chi ,Parents_Help ,Freq_Review_Colleag ,Auther_Dip ,Regime_adapt ,Etude_prd ,Own_PC ,Dist_Hous_Univ ,Stress_Exam ,Source_ Funding , Disciplinary_Sanctions ,Universite_Like ,school_Results ,Assist ,Means_Trasport ,Student_Profession ,Family_Situation ,Library_Use ,enc_ava)ces attributs sont considérés comme les attributs les plus important et qui influencent la performance des résultats des de premier semestre en première année.

Dans chaque niveau on trouve un ou plusieurs attributs, et les attributs sont présent dans les différents niveaux selon leurs importance où le premier niveau est le plus important suivi par le deuxième niveau, ... etc.

Donc les 23 attributs sont triés selon leur importance dans le tableau 4-3. Selon ce tableau on conclu que l'assistance au classe est l'attribut le plus important qui influence les résultats des étudiants en première année.

Niveau	***Identifients de l'attributs***	***Attributs***
1	students assist their td	Assist_ td
2	Student's life level	Life_Level
	sexs	sexs
	Student's performance in mathematics	Performance_Mat
3	Student's performance in physics and chemistry	Performance_Chi
	Parents help	Parents help
	Frequency of review with colleagues	Freq_Review_Colleag
	Student's life level	Life_Level
4	another diploma	Auther_Dip
	Adapt to the university system	Regime_adapt
	Practicing recreational activities during the study	Etude_prd
5	The student have a PC, how often it use it	Own_PC
	The distance between housing and university	Dist_Hous_Univ
	Stress during exams	Stress_Exam
	Source of funding	Source_Funding

Niveau	***Identifients de l'attributs***	***Attributs***
6	Disciplinary_Sanctions	Disciplinary_Sanct
	Sexe	Sexe
	The student likes going to university	Universite_Like
	School results	school_ Results
	Course assistant	Assist
	Means of transportation to the university	Means_Trasport
	The student exercise a professional activity or no	Student_Profession
7	School results	school_Results
	Practicing recreational activities during the study	Etude_prd
	Sexe	Sexe
	Family situation	Family_Situation
	Use of library	Library_Use
8	make discussions at home	enc_ava

TABLE 4.3 – Liste des attributs influencent la performance des R_sem1

- **classification de l'attribut R_ Sem2)**

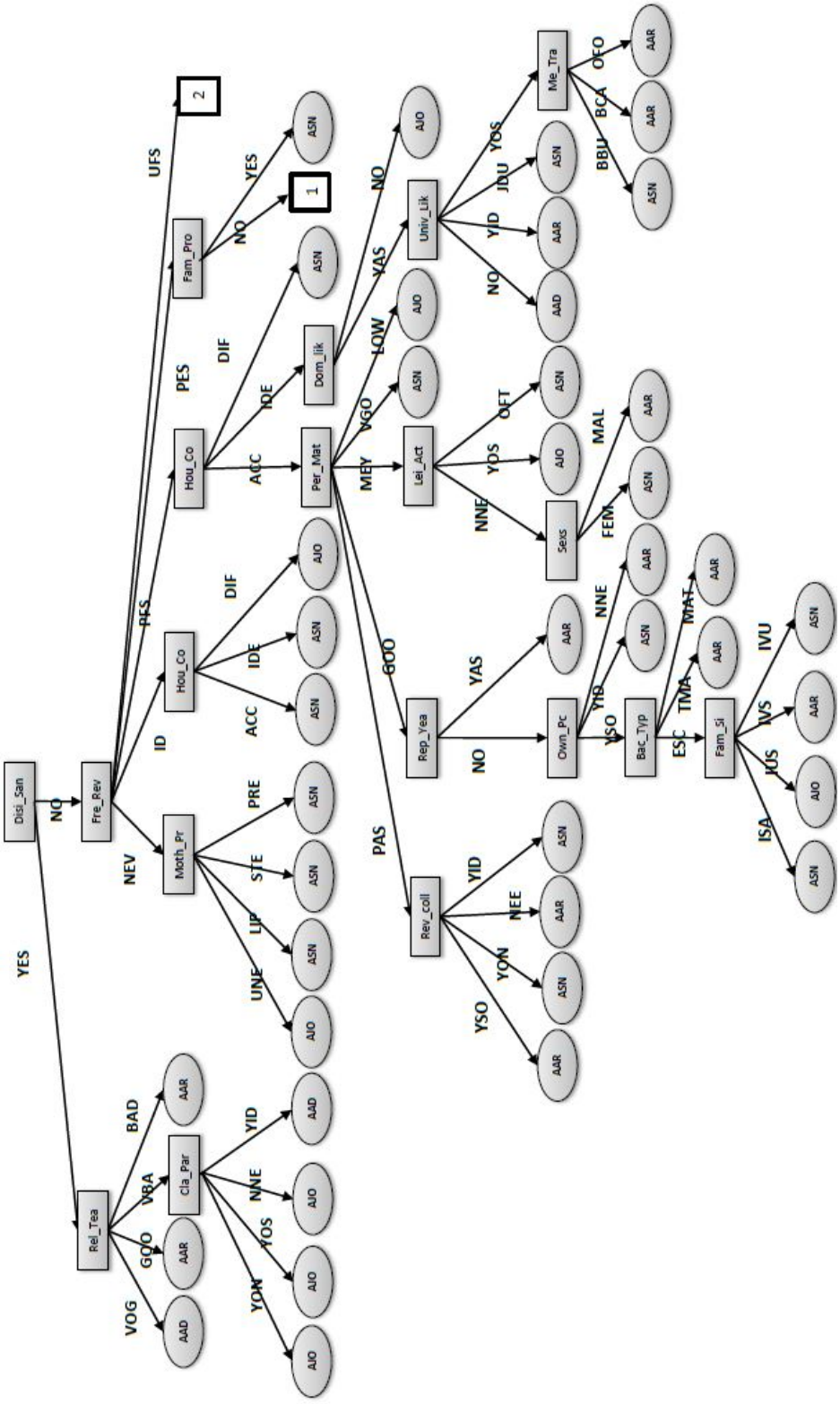
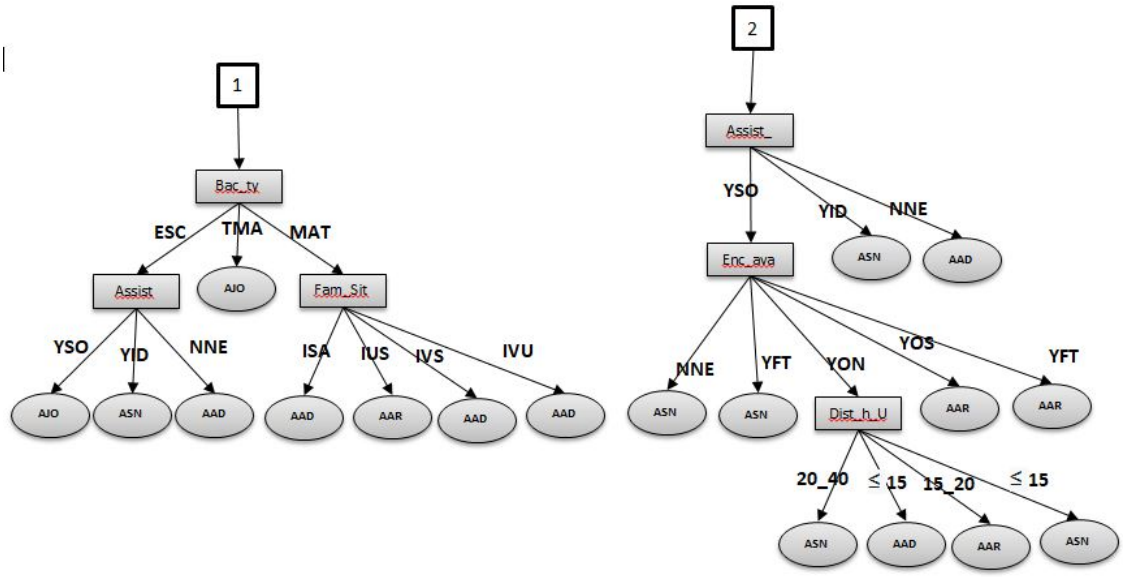


FIGURE 4.4 – Arbre de décision selon l'attribut R_Sem2 utilisant l'algorithme J48



	Nombre d'instances	***Pourcentage***
Instances correctement classées	30	40%
Instances incorrectement classées	45	60%
total	***75***	***100%***

TABLE 4.4 – Résultats de classification par l'algorithme J48 selon l'attribut « Résultat du premier semestre »

a) ***confusion matrix :***

a	b	c	d	$<--classifiedas$
1	0	3	1	$\|a = AAD$
1	23	11	5	$\|b = ASN$
0	12	6	3	$\|c = AAR$
0	3	6	0	$\|d = AJO$

TABLE 4.5 – Matrice de confusion du classification selon l'attribut R_Sem2 utilisant l'algorithme J48

b) ***Analyse de l'arbre de la figure 4-4 :***

On a remarqué dans l'arbre de la figure 6 qu'il ya huit niveaux et 21 attributs présent dans l'arbre (Disciplinary_Sanct, Relation_Teacher, Freq_Review_Colleag, Class_ Participation, Mother_Profession, Housing_Conditions, Fam_Problems , Assist, Performance_Mat, Dom_Like, Bac_Type , enc_ava, Review_Colleag , Repeat_Year , Leisure_Activity, Universite_Like, Family_Situation, Dist_Hous_Univ, Own_PC, Sexe, Means_Trasport) ces attributs sont considérés comme les attributs les plus important et qui influencent la performance des résultats de deuxième semestre en première année.

Dans chaque niveau on trouve un ou plusieurs attributs, et les attributs sont présent dans les différents niveaux selon leurs importance où le premier niveau est le plus important suivi par le deuxième niveau, ... etc.

Donc les 21 attributs sont triés selon leur importance dans le tableau 4-6. Selon ce tableau on conclu que Sanctions disciplinaires est l'attribut le plus important qui influence les résultats des étudiants en première année.

Niveau	***Identifients de l'attributs***	***Attributs***
1	Disciplinary_ Sanctions	Disciplinary_ Sanct
2	Relations with teachers	Relation_Teacher
	Frequency of review with colleagues	Freq_Review_Colleag
3	Class participation	Class_Participation
	Mother profession	Mother_Profession
	Housing conditions	Housing_Conditions
	Family problems	Fam_Problems
	Course assistant	Assist
4	Student's performance in mathematics	Performance_Mat
	Did you like the affected domain	Dom_Like
	Baccalaureate type	Bac_Type
	make discussions at home	enc_ava
5	Review with colleagues	Review_Colleag
	The student repeated a year in secondary	Repeat_Year
	The student exercises a leisure activity	Leisure_Activity
	The student likes going to university	Universite_Like
	Course assistant	Assist
	Family Situation	Family_Situation
	The distance between housing and university	Dist_Hous_Univ

Niveau	***Identifients de l'attributs***	***Attributs***
6	The student have a PC, how often it use it Sexe Means of transportation to the university	Own_PC Sexe Means_Trasport
7	Baccalaureate type	Bac_Type
8	Family Situation	Family_Situation

TABLE 4.6 – Liste des attributs influencent la performance des résultats R_sem_2

- **classification de l'attribut R_final)**

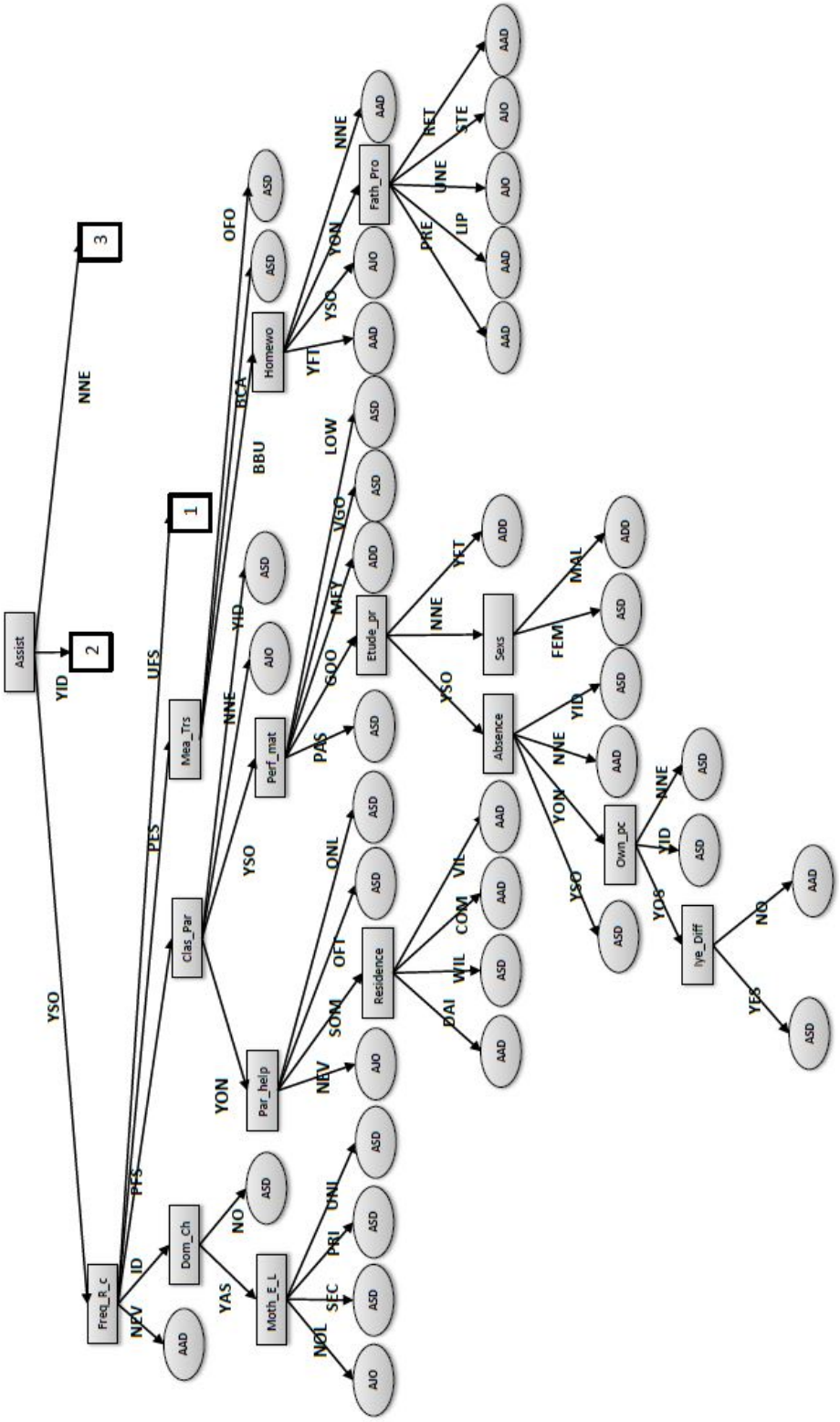
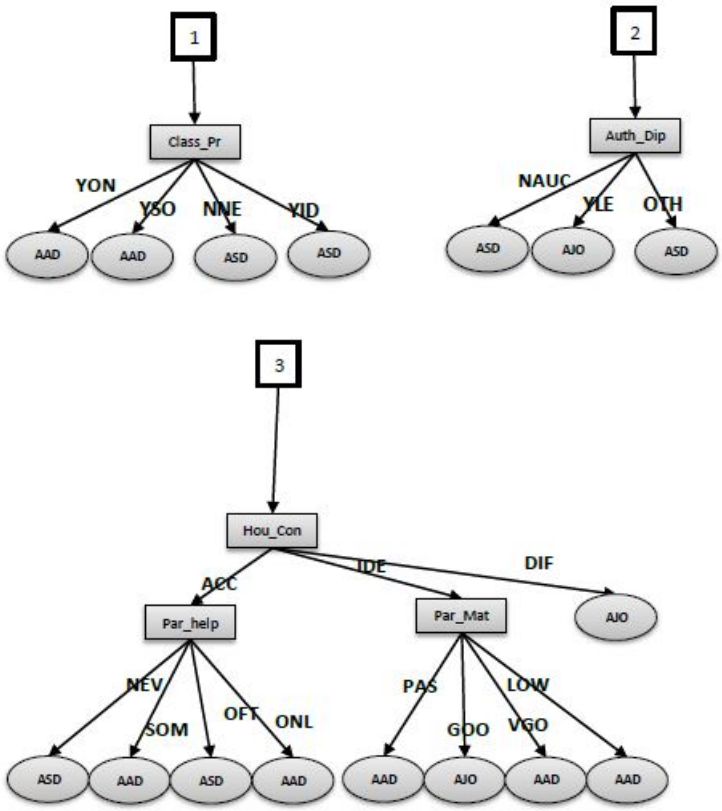


FIGURE 4.5 – Arbre de décision selon l'attribut R_final utilisant l'algorithme J48



	Nombre d'instances	***Pourcentage***
Instances correctement classées	40	53.3333 %
Instances incorrectement classées	35	46.6667 %
total	***75***	***100%***

TABLE 4.7 – Résultats de classification par l'algorithme J48 selon l'attribut «Résultat final»

a) ***confusion matrix :***

$$\begin{array}{ccc|l} a & b & c & <--classifiedas \\ 4 & 10 & 1 & |a = AAR \\ 9 & 35 & 4 & |b = ASD \\ 3 & 8 & 1 & |c = AJO \end{array}$$

TABLE 4.8 – Matrice de confusion de la classification selon l'attribut Résultat final utilisant l'algorithme J48.

b) ***Analyse de l'arbre de la figure4-5 :***

On a remarqué dans l'arbre de la figure 6 qu'il ya huit niveaux et 18 attributs présent dans l'arbre (Assist, Review_Colleag, Auther_Dip, Housing_Conditions, Dom_ Chois , Class_Participation , Means_Trasport , Parents_Help , Performance_Mat , Mother_Educ_Level , Homework , Residence , Etude_prd , Father_Profession , Absence , Sexe , Own_PC, 1Year_Difficult) ces attributs sont considérés comme les attributs les plus important et qui influencent la performance des résultats final des étudiants en première année.

Dans chaque niveau on trouve un ou plusieurs attributs, et les attributs sont présent dans les différents niveaux selon leurs importance où le premier niveau est le plus important suivi par le deuxième niveau, ... etc.

Donc les 18 attributs sont triés selon leur importance dans le tableau 4-9. Selon ce tableau on conclu que l'assistance au classe est l'attribut le plus important qui influence les résultats des étudiants en première année.

Niveau	***Identifients de l'attributs***	***Attributs***
1	Course assistant	Assist
2	Review with colleagues another diploma Housing conditions	Review_Colleag Auther_Dip Housing_Conditions
3	Your first-year domain is among the top three choices Class participation Means of transportation to the university Parents help Student's performance in mathematics	Dom_ Choise Class_Participation Means_Trasport Parents_Help Performance_Mat
4	Mother's educational level Parents help Student's performance in mathematics At home, the student does his homework requested by the teachers the first year of university	Mother_Educ_Level Parents_Help Performance_Mat Homework
5	Residence Practicing recreational activities during the study Father profession	Residence Etude_prd Father_Profession
6	Absence during lessons Sexe	Absence Sexe
7	The student have a PC, how often it use it	Own_PC
8	The student has difficulties in the first year	Year_Difficult

TABLE 4.9 – Liste des attributs influencent la performance des résultats final des étudiants

2. **l'algorithme One R** :

	Nombre d'instances	***Pourcentage***
Instances correctement classées	34	45.3333 %
Instances incorrectement classées	41	54.6667 %
total	***75***	***100%***

TABLE 4.10 – Résultats de classification par l'algorithme One R selon l'attribut « Résultat du premier semestre »

a) ***confusion matrix :***

a	b	c	d	$<--classifiedas$
2	9	0	2	$\|a = AAR$
11	31	0	3	$\|b = ASN$
1	6	0	0	$\|c = ADD$
1	8	0	1	$\|d = AJO$

TABLE 4.11 – Matrice de confusion de la classification selon l'attribut R_Sem2 utilisant l'algorithme One R

	Nombre d'instances	***Pourcentage***
Instances correctement classées	32	42.6667 %
Instances incorrectement classées	43	57.3333 %
total	***75***	***100%***

TABLE 4.12 – Résultats de classification par l'algorithme One R selon l'attribut « Résultat du deuxième semestre »

a) ***confusion matrix :***

a	b	c	d	$<--classifiedas$
0	1	4	2	$\|a = AAR$
0	28	12	0	$\|b = ASN$
1	17	4	0	$\|c = AAR$
1	3	6	1	$\|d = AJO$

TABLE 4.13 – Matrice de confusion de la classification selon l'attribut R_Sem1 utilisant l'algorithme One R

	Nombre d'instances	***Pourcentage***
Instances correctement classées	49	65.3333 %
Instances incorrectement classées	26	34.6667 %
total	***75***	***100%***

TABLE 4.14 – Résultats de classification par l'algorithme One R selon l'attribut « Résultats final »

a) ***confusion matrix :***

a	b	c	$<--classifiedas$
7	8	0	$\|a = AAD$
5	42	1	$\|b = ASD$
4	8	0	$\|c = AJO$

TABLE 4.15 – Matrice de confusion de la classification selon l'attribut Résultats final utilisant l'algorithme One R

3. **l'algorithme IBK** :

	Nombre d'instances	***Pourcentage***
Instances correctement classées	39	52 %
Instances incorrectement classées	36	48 %
total	***75***	***100%***

TABLE 4.16 – Résultats de classification par l'algorithme IBK selon l'attribut « Résultats du premier semestre »

a) ***confusion matrix :***

a	b	c	d	$<--classifiedas$
8	2	2	1	$\|a = AAR$
13	27	3	2	$\|b = ASN$
4	2	0	1	$\|c = ADD$
2	2	2	4	$\|d = AJO$

TABLE 4.17 – Matrice de confusion de la classification selon l'attribut R_Sem1 utilisant l'algorithme IBK

	Nombre d'instances	***Pourcentage***
Instances correctement classées	42	56 %
Instances incorrectement classées	33	44 %
total	***75***	***100%***

TABLE 4.18 – Résultats de classification par l'algorithme IBK selon l'attribut « Résultats du deuxième semestre »

a) ***confusion matrix :***

a	b	c	d	$<--classifiedas$
1	3	0	1	$\|a = AAR$
1	33	4	2	$\|b = ASN$
0	10	7	4	$\|c = AAR$
0	5	3	1	$\|d = AJO$

TABLE 4.19 – Matrice de confusion de la classification selon selon l'attribut R_Sem2 utilisant l'algorithme IBK

	Nombre d'instances	***Pourcentage***
Instances correctement classées	47	62.6667 %
Instances incorrectement classées	28	37.3333 %
total	***75***	***100%***

TABLE 4.20 – Résultats de classification par l'algorithme IBK selon l'attribut « Résultats final»

a) ***confusion matrix :***

$$\begin{array}{cccc} a & b & c & <--classifiedas \\ 6 & 7 & 2 & |a = AAD \\ 8 & 37 & 3 & |b = ASD \\ 6 & 2 & 4 & |c = AJO \end{array}$$

TABLE 4.21 – Matrice de confusion de la classification selon selon l'attribut Résultats final utilisant l'algorithme IBK

4. **l'algorithme Naive base** :

	Nombre d'instances	***Pourcentage***
Instances correctement classées	34	45.3333 %
Instances incorrectement classées	41	54.6667 %
total	***75***	***100%***

TABLE 4.22 – Résultats de classification par l'algorithme Naive base selon l'attribut « Résultats du premier semestre »

a) ***confusion matrix :***

$$\begin{array}{ccccc} a & b & c & d & <--classifiedas \\ 5 & 4 & 3 & 1 & |a = AAR \\ 17 & 26 & 1 & 1 & |b = ASN \\ 2 & 3 & 2 & 0 & |c = ADD \\ 5 & 3 & 1 & 1 & |d = AJO \end{array}$$

TABLE 4.23 – Matrice de confusion de la classification selon l'attribut R_Sem1 utilisant l'algorithme Naive base

	Nombre d'instances	***Pourcentage***
Instances correctement classées	39	52 %
Instances incorrectement classées	36	48 %
total	***75***	***100%***

TABLE 4.24 – Résultats de classification par l'algorithme Naive base selon l'attribut « Résultats du deuxième semestre »

a) ***confusion matrix :***

a	b	c	d	$<--classifiedas$
1	1	3	0	$\|a = AAR$
2	27	11	0	$\|b = ASN$
2	8	10	1	$\|c = AAR$
2	2	4	1	$\|d = AJO$

TABLE 4.25 – Matrice de confusion de la classification selon selon l'attribut R_Sem2 utilisant l'algorithme Naive base

	Nombre d'instances	***Pourcentage***
Instances correctement classées	49	65.3333 %
Instances incorrectement classées	26	34.6667 %
total	***75***	***100%***

TABLE 4.26 – Résultats de classification par l'algorithme Naive base selon l'attribut « Résultats final»

a) ***confusion matrix :***

a	b	c	$<--classifiedas$
9	6	0	$\|a = AAD$
9	38	1	$\|b = ASD$
5	5	2	$\|c = AJO$

TABLE 4.27 – Matrice de confusion de la classification selon selon l'attribut Résultats final utilisant l'algorithme Naive base

5. **l'algorithme SMO** :

	Nombre d'instances	***Pourcentage***
Instances correctement classées	25	33.3333 %
Instances incorrectement classées	50	66.6667 %
total	***75***	***100%***

TABLE 4.28 – Résultats de classification par l'algorithme SMO selon l'attribut « Résultats du premier semestre »

a) ***confusion matrix :***

a	b	c	d	$<--classifiedas$
6	2	2	3	$\|a = AAR$
25	17	0	3	$\|b = ASN$
3	4	0	0	$\|c = ADD$
2	5	1	2	$\|d = AJO$

TABLE 4.29 – Matrice de confusion de la classification selon l'attribut R_Sem1 utilisant l'algorithme SMO

	Nombre d'instances	***Pourcentage***
Instances correctement classées	35	46.6667 %
Instances incorrectement classées	40	53.3333 %
total	***75***	***100%***

TABLE 4.30 – Résultats de classification par l'algorithme SMO selon l'attribut « Résultats du deuxième semestre »

a) ***confusion matrix :***

a	b	c	d	$<--classifiedas$
1	2	2	0	$\|a = AAR$
5	26	9	0	$\|b = ASN$
2	8	8	3	$\|c = AAR$
3	3	3	0	$\|d = AJO$

TABLE 4.31 – Matrice de confusion de la classification selon selon l'attribut R_Sem2 utilisant l'algorithme SMO

	Nombre d'instances	***Pourcentage***
Instances correctement classées	42	56 %
Instances incorrectement classées	33	44 %
total	***75***	***100%***

TABLE 4.32 – Résultats de classification par l'algorithme SMO selon l'attribut « Résultats final»

a) ***confusion matrix :***

$$\begin{array}{cccc} a & b & c & <--classifiedas \\ 4 & 10 & 1 & |a = AAD \\ 11 & 34 & 3 & |b = ASD \\ 3 & 5 & 4 & |c = AJO \end{array}$$

TABLE 4.33 – Matrice de confusion de la classification selon selon l'attribut Résultats final utilisant l'algorithme SMO

6. **l'algorithme Bayes Net** :

	Nombre d'instances	***Pourcentage***
Instances correctement classées	33	44 %
Instances incorrectement classées	42	56 %
total	***75***	***100%***

TABLE 4.34 – Résultats de classification par l'algorithme Bayes Net selon l'attribut « Résultats du premier semestre »

a) ***confusion matrix :***

$$\begin{array}{cccc} a & b & c & d & <--classifiedas \\ 5 & 4 & 3 & 1 & |a = AAR \\ 17 & 25 & 1 & 2 & |b = ASN \\ 2 & 3 & 2 & 0 & |c = ADD \\ 4 & 3 & 2 & 1 & |d = AJO \end{array}$$

TABLE 4.35 – Matrice de confusion de la classification selon l'attribut R_Sem1 utilisant l'algorithme Bayes Net

	Nombre d'instances	***Pourcentage***
Instances correctement classées	40	53.3333 %
Instances incorrectement classées	35	46.6667 %
total	***75***	***100%***

TABLE 4.36 – Résultats de classification par l'algorithme Bayes Net selon l'attribut « Résultats du deuxième semestre »

a) ***confusion matrix :***

a	b	c	d	$<--classifiedas$
0	5	0	0	$\|a = AAR$
0	40	0	0	$\|b = ASN$
0	21	0	0	$\|c = AAR$
0	9	0	0	$\|d = AJO$

TABLE 4.37 – Matrice de confusion de la classification selon selon l'attribut R_Sem2 utilisant l'algorithme Bayes Net

	Nombre d'instances	***Pourcentage***
Instances correctement classées	49	65.3333 %
Instances incorrectement classées	26	34.6667 %
total	***75***	***100%***

TABLE 4.38 – Résultats de classification par l'algorithme Bayes Net selon l'attribut « Résultats final»

a) ***confusion matrix :***

a	b	c	$<--classifiedas$
9	6	0	$\|a = AAD$
9	38	1	$\|b = ASD$
6	4	2	$\|c = AJO$

TABLE 4.39 – Matrice de confusion de la classification selon selon l'attribut Résultats final utilisant l'algorithme Bayes Net

7. **l'algorithme Multi ClassClassifier** :

	Nombre d'instances	***Pourcentage***
Instances correctement classées	27	36 %
Instances incorrectement classées	48	64 %
total	***75***	***100%***

TABLE 4.40 – Résultats de classification par l'algorithme Multi ClassClassifier selon l'attribut « Résultats du premier semestre »

a) ***confusion matrix :***

a	b	c	d	$<--classifiedas$
6	4	1	2	$\|a = AAR$
22	18	3	2	$\|b = ASN$
2	3	1	1	$\|c = ADD$
4	3	1	3	$\|d = AJO$

TABLE 4.41 – Matrice de confusion de la classification selon l'attribut R_Sem1 utilisant l'algorithme Multi ClassClassifier

	Nombre d'instances	***Pourcentage***
Instances correctement classées	35	46.6667 %
Instances incorrectement classées	40	53.3333 %
total	***75***	***100%***

TABLE 4.42 – Résultats de classification par l'algorithme Multi ClassClassifier selon l'attribut « Résultats du deuxième semestre »

a) ***confusion matrix :***

a	b	c	d	$<--classifiedas$
1	2	2	0	$\|a = AAR$
5	24	9	2	$\|b = ASN$
1	8	9	3	$\|c = AAR$
2	4	2	1	$\|d = AJO$

TABLE 4.43 – Matrice de confusion de la classification selon selon l'attribut R_Sem2 utilisant l'algorithme Multi ClassClassifier

	Nombre d'instances	***Pourcentage***
Instances correctement classées	34	45.3333 %
Instances incorrectement classées	41	54.6667 %
total	***75***	***100%***

TABLE 4.44 – Résultats de classification par l'algorithme Multi ClassClassifier selon l'attribut « Résultats final»

a) ***confusion matrix :***

$$\begin{array}{cccl} a & b & c & <--classifiedas \\ 6 & 7 & 2 & |a = AAD \\ 13 & 27 & 8 & |b = ASD \\ 8 & 3 & 1 & |c = AJO \end{array}$$

TABLE 4.45 – Matrice de confusion de la classification selon selon l'attribut Résultats final utilisant l'algorithme Multi ClassClassifier

8. **l'algorithme RANDOM FOREST** :

	Nombre d'instances	***Pourcentage***
Instances correctement classées	44	58.6667 %
Instances incorrectement classées	31	41.3333 %
total	***75***	***100%***

TABLE 4.46 – Résultats de classification par l'algorithme RANDOM FOREST selon l'attribut « Résultats du premier semestre »

a) ***confusion matrix :***

$$\begin{array}{cccccl} a & b & c & d & <--classifiedas \\ 5 & 8 & 0 & 0 & |a = AAR \\ 6 & 39 & 0 & 0 & |b = ASN \\ 2 & 5 & 0 & 0 & |c = ADD \\ 4 & 8 & 0 & 0 & |d = AJO \end{array}$$

TABLE 4.47 – Matrice de confusion de la classification selon l'attribut R_Sem1 utilisant l'algorithme RANDOM FOREST

	Nombre d'instances	***Pourcentage***
Instances correctement classées	41	54.6667 %
Instances incorrectement classées	34	45.3333 %
total	***75***	***100%***

TABLE 4.48 – Résultats de classification par l'algorithme RANDOM FOREST selon l'attribut « Résultats du deuxième semestre »

a) ***confusion matrix :***

a	b	c	d	$<--classifiedas$
0	2	3	0	$\|a = AAR$
0	34	6	0	$\|b = ASN$
0	14	7	0	$\|c = AAR$
0	4	5	0	$\|d = AJO$

TABLE 4.49 – Matrice de confusion de la classification selon selon l'attribut R_Sem2 utilisant l'algorithme RANDOM FOREST

	Nombre d'instances	***Pourcentage***
Instances correctement classées	48	64 %
Instances incorrectement classées	27	36 %
total	***75***	***100%***

TABLE 4.50 – Résultats de classification par l'algorithme RANDOM FOREST selon l'attribut « Résultats final»

a) ***confusion matrix :***

a	b	c	$<--classifiedas$
2	13	0	$\|a = AAD$
2	46	0	$\|b = ASD$
2	10	0	$\|c = AJO$

TABLE 4.51 – Matrice de confusion de la classification selon selon l'attribut Résultats final utilisant l'algorithme RANDOM FOREST

4.2.2 Comparaison de résultats

Nous avons fait une comparaison de résultats obtenues par l'application des huit algorithmes (J48, One R, IBK, Naive base , SMO , Bayes Net, Multi ClassClassifier , RANDOM FOREST) pour voir quel est l'algorithme qui donne le meilleur modèle prédictif qui nous aide à prédure les résultats de nouveau étudiants sans connaitre leur classes, pour la comparaison nous avons choisi les deux critères les plus important qui sont :

- Instances correctement classées
- Instances incorrectement classées

Les tableau 4-52, 4-53 et 4-54 récapitulent les résultats des différents algorithmes pour bien faire la comparaison selon les attributs « Résultat du premier semestre », « Résultat du deuxième semestre » et « Résultat final » respectivement.

Critère / *Algo*	***Instanances correctement classées***	***Instances incorrectement classées***
J48	42.6667 %	57.3333%
One R	45.3333 %	54.6667 %
IBK	52 %	48%
Naive base	45.3333 %	54.6667 %
SMO	33.3333 %	66.6667 %
Bayes Net	44%	56%
Multi ClassClassifier	36 %	64%
RANDOM FOREST	58.6667 %	41.3333 %

TABLE 4.52 – Comparaison des résultats de classification de l'attribut R_Sem1.

Critère / *Algo*	***Instanances correctement classées***	***Instances incorrectement classées***
J48	40 %	60%
One R	42.6667 %	57.3333 %
IBK	56 %	44%
Naive base	52 %	48 %
SMO	46.6667 %	53.3333 %
Bayes Net	53.3333 %	46.6667 %
Multi ClassClassifier	46.6667 %	53.3333 %
RANDOM FOREST	54.6667 %	45.3333 %

TABLE 4.53 – Comparaison des résultats de classification de l'attribut R_Sem2.

Critère / *Algo*	***Instanances correctement classées***	***Instances incorrectement classées***
J48	53.3333 %	46.6667%
One R	65.3333 %	34.6667 %
IBK	62.6667 %	37.3333 %
Naive base	65.3333 %	34.6667 %
SMO	56 %	44 %
Bayes Net	65.3333 %	34.6667 %
Multi ClassClassifier	45.3333 %	54.6667 %
RANDOM FOREST	64 %	36 %

TABLE 4.54 – Comparaison des résultats de classification de l'attribut Résultat final

Après la comparaison précédente nous constatons que l'algorithme RANDOM FOREST est le meilleur parmi ces trois algorithmes où il classifier 58 % des exemples correctement selon l'attribut « Résultat Sem1 ». Pour l'attribut « Résultat Sem2 »on a constaté que l'algorithme IBK donne le meilleur tau de classification, alors que les algorithmes One R , Naive base et Bayes Net donne les meilleur taux de classification selon l'attribut « Résultat final ». Où ils classifier 65 % des exemples correctement.
Pour la classification des étudiants selon les trois attributs des résultats étudiés plus ce que sa matrice de confusion été une matrice diagonale ça nous confirme l'efficacité de cet algorithme et pour cela nous avons considérer les cinq arbres comme des modèles de prédiction des résultats des étudiants.

4.3 L'application de classification

4.3.1 Description de l'application

1. Sélection du fichier d'apprentissage (Fichier ARFF)
2. Apprentissage et génération du meilleur modèle de classification
   - Application des algorithmes de classification
   - Evaluation des modèles générés
   - Sélection du meilleur modèle
3. Utilisation du meilleur modèle pour une prédiction
   - Visualisation du résultat de prédiction

4.4 Les interfaces de l'application développée

Nous présentons dans cette section des captures d'écran de l'application développée.

4.4.1 L'Interface d'accueil

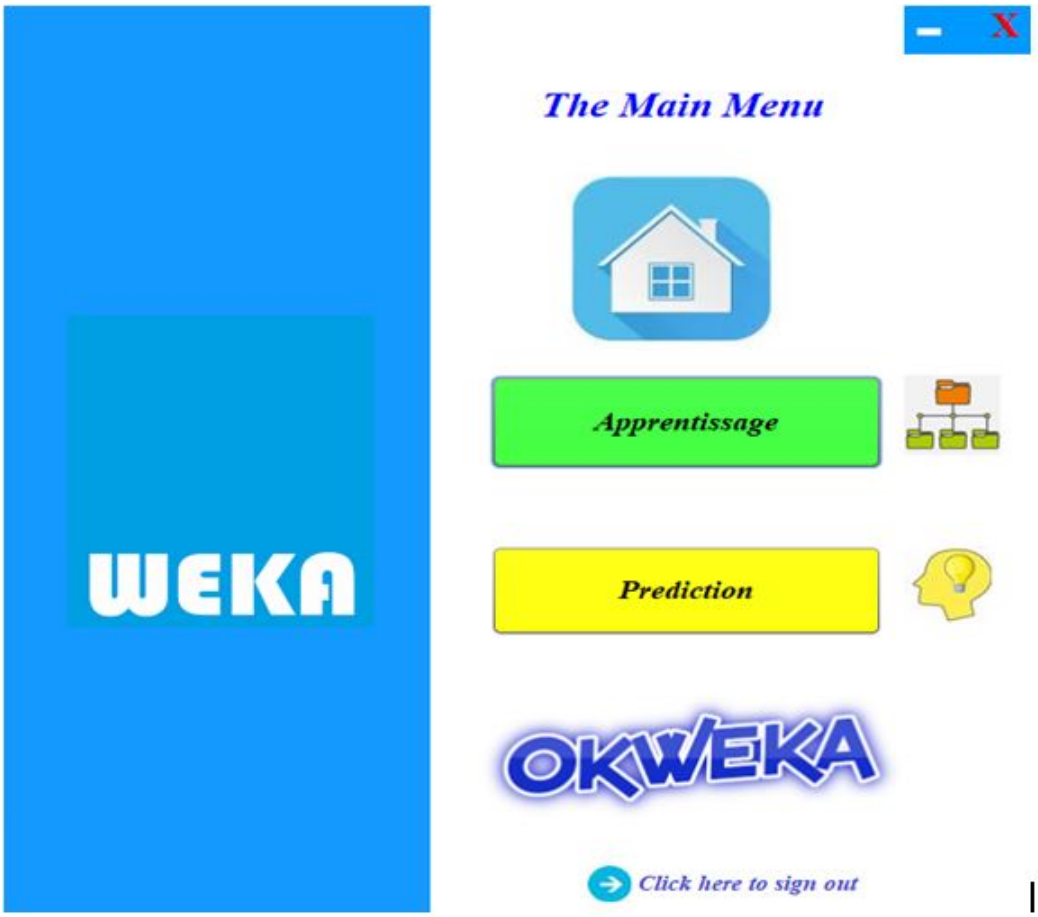


FIGURE 4.6 – Fenêtre principale de l'application

Cette fenêtre contient deux boutons principaux :

- Le bouton « Apprentissage » permet de sélectionner la base d'apprentissage et de lancer l'apprentissage pour construire le modèle de classification.
- Le bouton « Prédiction » implémente la tâche de prédiction des résultats d'un nouvel étudiant.

4.4.2 Sélection de la base d'apprentissage

La fenêtre de sélection de fichiers ARFF pour sélectionner la base de données d'un flux au cours de la première année Pour les étudiants en informatique et mathématiques, comme illustré dans la figure suivante.

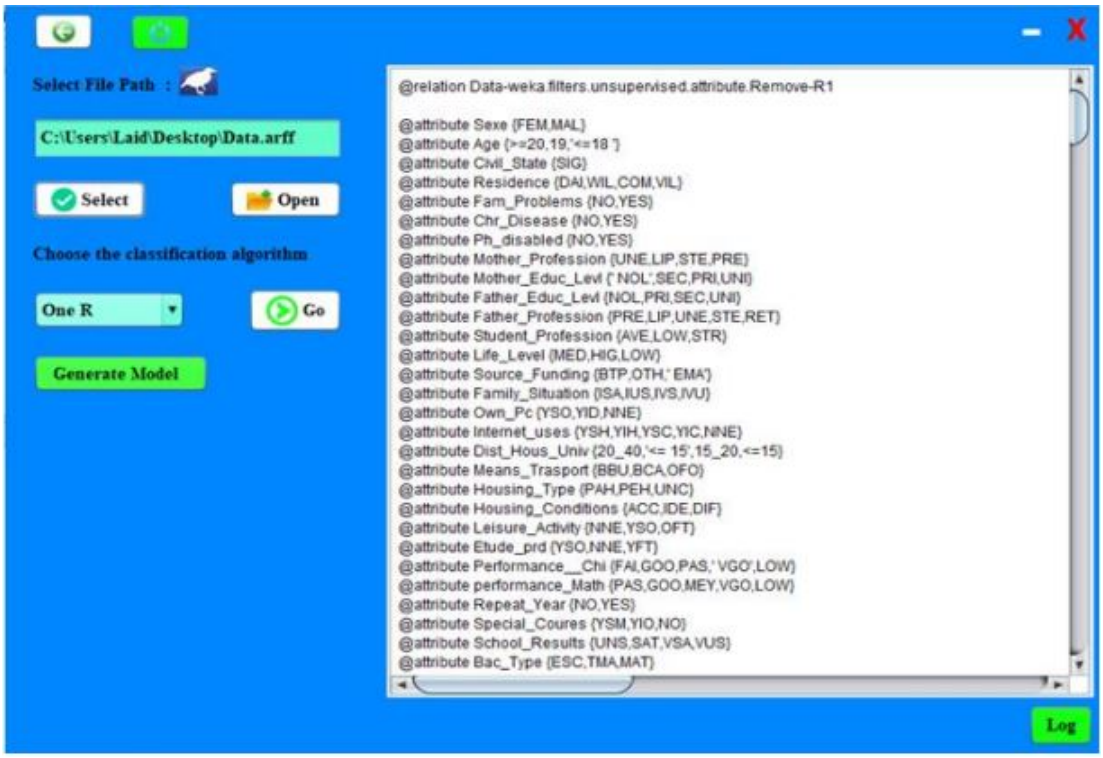


FIGURE 4.7 – Sélection de la base d'apprentissage

Observation :
Le bouton « Open » vous permet d'afficher le fichier dans son éditeur principal, que ce soit dans le programme WEKA si le fichier avec une extension «.arff» ou dans l'Excel si le fichier avec une extension « .csv ».

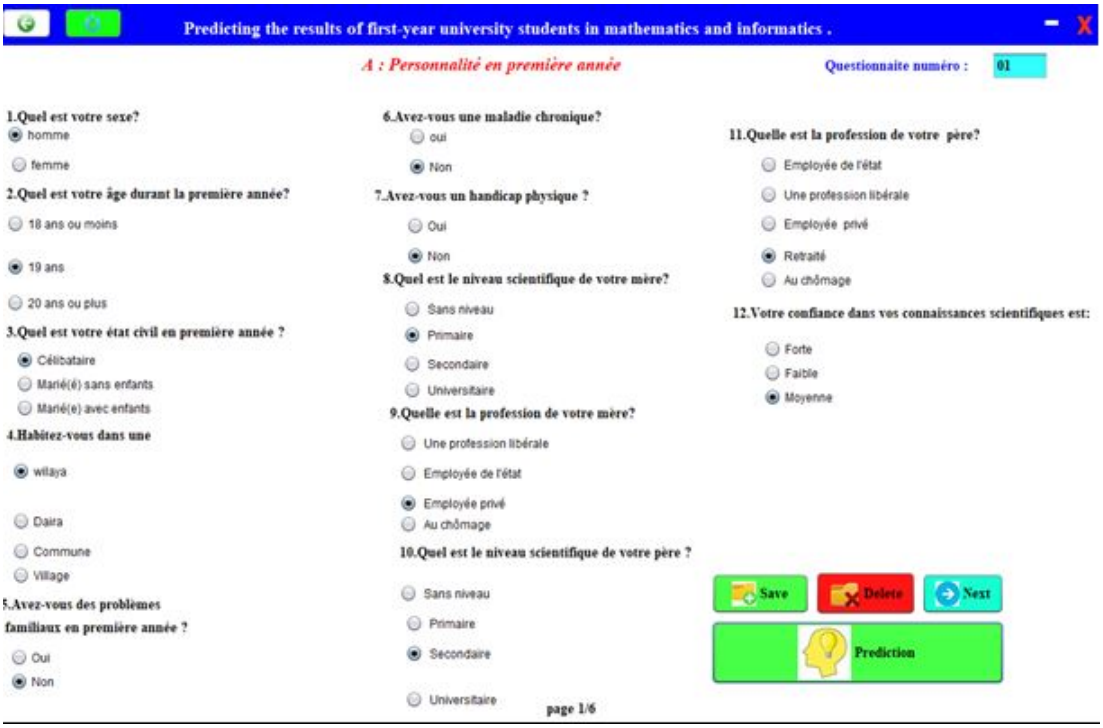


FIGURE 4.8 – Interface de prédiction des résultats d'un étudiant

Cette interface est l'interface pour prédire les résultats du nouvel étudiant , en entrant les informations de l'étudiant avec une
L'application des algorithmes utilisés pour prédire les performances d'étudiant.

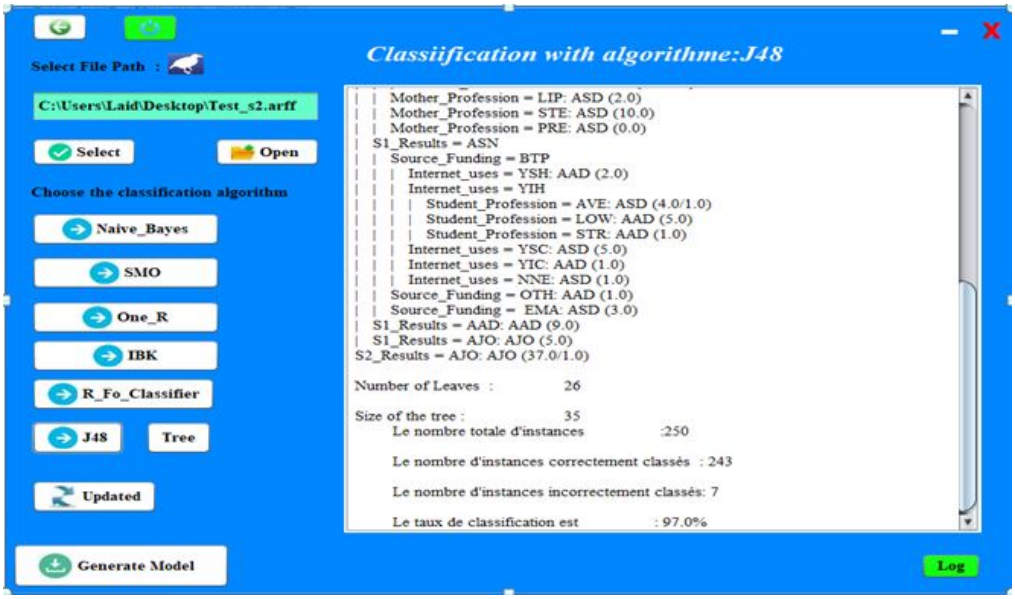


FIGURE 4.9 – exemple de appliqué l'algorithme IBK

4.5 Conclusion

Dans ce chapitre, nous avons présenté les résultats obtenus grâce aux expériences que nous avons menées pour tester notre système développé pour classer les résultats des élèves et pour évaluer les classifications fournies par les outils d'exploration de données Weka selon des critères spécifiques.
Sur la base des résultats des expériences, nous avons remarqué que l'acquisition de connaissances nous permet de prédire les résultats du nouvel étudiant en première année de collège en mathématiques et informatique.
Selon les différentes données collectées sur les étudiants. Enfin, on peut dire que c'est à travers l'étude que nous avons menée dans ce domaine.
On peut anticiper les résultats des nouveaux étudiants en première année universitaire de spécialisation en mathématiques et informatique.

Conclusion Générale

Notre projet de fin d'études fait partie d'un projet consistant à concevoir et produire une application de classification qui permet de prédire les résultats des étudiants en mathématiques et informatique en première année universitaire. Pour cela, nous avons présenté des concepts liés à l'extraction de données éducatives, qui représentent un ensemble de techniques d'exploration de données qui permettent l'extraction de connaissances à partir d'une base de données créer à partir des réponses des étudiants sur des questions posées via un questionnaire.

Nous avons créé des modèles de classification des étudiants en utilisant des techniques de classification offertes par l'outil Weka, nous avons évalué ensuite les performances des modèles de classification crées pour voir leurs qualité.

Selon les résultats d'évaluation, nous avons constaté que le plus grand pourcentage de classification était obtenu par l'algorithme " Naive_bayse " et " One R" avec un taux de 65% alors que nous avons constaté que le deuxième pourcentage était obtenu par l'algorithme " IBK" par un taux de 62 %. Alors que le modèle généré par l'algorithme " Multi ClassClassifier" donne le plus faible taux avec un pourcentage de 45%. L'algorithme " j48 " et " SMO " donnent respectivement les taux 53 % et 56 %.

Nous espérons améliorer notre étude par l'élargissement de la base d'apprentissage par les étudiants des années à venir et des autres domaines dans la première année de l'université, afin de populariser l'application. Et d'un autre coté de proposer et d'appliquer d'autres algorithmes de classification afin d »améliorer la qualité des modèles de prédiction générés.

ANNEXE

N	***les attributs***	***code attrib***	***valeur attri***	***code et valeur attrib***
1	The sex of the student	Sex	(Male,Female)	Male, Female
2	Age	Age	(18 or younger, 19, 20 or older)	<=18,19,>=20
3	Civil state	Civil_State	Single, Married Without Children, Married With Children)	(SIG, MWC, MWH)
4	Residence	Residence	(Wilaya, Daïra, Commune, Village)	(WIL,DAI,COM ,VIL)
5	Family problems	Fam_Problems	(Yes, No)	(YES,NO)
6	Chronic disease	Chr_Disease	(Yes, No)	(YES,NO)
7	Physically disabled	Ph_disabled	(Yes, No)	(YES,NO)
8	Mother's educational level	Mother_ Educ_Level	(No Level, Primary, Secondary , University)	(NOL, PRI, SEC , UNI)
9	Mother_ Profession	Mother profession	(Liberal Profession, State Employee, Private Employee , Unemployed)	(LIP,STE,PRE ,UNE)
10	Father's educational level	Father_ Educ_Level	(No Level, Primary, Secondary, University)	(NOL,PRI,SEC, UNI)

N	***les attributs***	***code attrib***	***valeur attri***	***code et valeur attrib***
11	Father profession	Father_ Profession	(Liberal Profession, State Employee, Private Employee, Retirement , Unemployed)	(STE,LIP,PRE, RET,UNE)
12	confidence in your scientific knowledge	Confidence_ scientific	(Strong, Low , Averag)	(STR,LOW,AVE)
13	Life level	Life_Level	(High, Medium, Low)	(HIG,MED,LOW)
14	Source of funding	Source_ Funding	(By the Parents, Employed) Activity, Other	(BTP,EMA,OTH)
15	Family situation	Family_ Situation	(is Very Satisfactory, is Satisfactory, is Unsatisfactory, is Very Unsatisfactory)	(IVS,ISA,IUS, IVU)
16	The student have a PC, how often it use it	Own_PC	(Yes Every Day, Yes Sometimes , No Never)	(YID,YSO,NNE)
17	The student uses the Internet , how often he uses it and where	Internet_uses	(Yes Every Day at Home, Yes Every Day at the Cyber cafe, Yes Sometimes at Home, Yes, Sometimes at the Cyber cafe, No Never)	(YIH,YIC,YSH, YSC,NNE)
18	The distance between housing and university	Dist_Hous _Univ	(Less than 20KM , Between 20 and 50KM, Greater than 20 KM)	(<15, 15_20, 20_40,>40)

N	***les attributs***	***code attrib***	***valeur attri***	***code et valeur attrib***
19	Means of transportation to the university	Means_Trasport	(On Foot, By Bus, By Car)	(OFO,BBU,BCA)
20	Housing type	Housing_Type	(University Campus, Personal House, Parents' House)	(UNC,PAH,PEH)
21	Housing conditions	Housing_ Conditions	(Idéales, Acceptables , Difficiles)	(IDE,ACC,DIF)
22	The student exercises a leisure activity	Leisure_ Activity	(No Never, Yes,) Sometimes,Often	(NNE,YSO,OFT)
23	exercise leisure activities during the study period	Etude_prd	(No never, Yes sometimes, Yes, often)	(NNE, YSO, YFT)
24	Student's performance in physics and chemistry	Performance _Chi	(Good, Average, Fair, Poor)	(VGO,GOO,FAI ,PAS,LOW)
25	Student's performance in mathematics	Performance _Mat	(Good, Average, Fair, Poor)	(VGO,GOO,FAI ,PAS,LOW)
26	The student repeated a year in secondary	Repeat _Year	(Yes, No)	(YES,NO)
27	the student does special courses in terminale	Special _Courses	(Yes in Math, Yes in Physics , yes in , Science in Other , No)	(YSM,YIO,NO)
28	School results	School _Results	(Very Satisfactory , Satisfactory, Unsatisfactory)	(VSA,SAT,UNS ,VUS)

N	***les attributs***	***code attrib***	***valeur attri***	***code et valeur attrib***
29	Baccalaureate type	Bac_Type	(Mathematics, Experimental Sciences, Technical Math, Other Speciality)	(MAT,ESC,TMA ,OTHS)
30	Baccalaureate result	Bac_Everage and 12, between 12 and 14, Superior to 14)	(between 10	(10_12, 12_14,>14)
31	another diploma	Auther _Dip	(No None, Yes, license, Yes, Master, Yes, Classic system , Other)	(NAUC, YLE, YMA, YCS, OTH)
32	People who influence your choice of field for year the first academic	Choice _Field _Infl	(None, My Family, My Loved ones, Other)	(AUC, MAF, MRE , OTH)
33	Your first-year domain is among the top three choices	Dom_ Choise	(Yes, No)	(YES,NO)
34	Did you like the affected domain	Dom_Like	(Yes, No)	(YES,NO)
35	first year domain	Dom_adapt	(Mathematics and informatics ,science and technology,Material Sciences,Other)	(MI,ST,SM, OTH)

N	***les attributs***	***code attrib***	***valeur attri***	***code et valeur attrib***
36	the student is adapted to the university regime	Regime _adapt	(Yes in the First Quarter, Yes in the Second Quarter, Yes in, No)	(YFS,YSs,NO)
37	The student has difficulties in the first year	1Year _Difficult	(Yes, No)	(YES,NO)
38	Parents help	Parents _Help	(Yes Often, Yes Sometimes, Yes Rarely, Never)	(OFT,SOM,ONL ,NEV)
39	At home, the student does his homework requested by the teachers of the first year of university	Homework	(Yes Often, Yes Sometimes, Yes Rarely, Never)	(YFT,YSO,YON ,NNE)
40	The student likes going to university	Universite_Like	(Yes always, I hate the university Yes sometimes,No)	(YID,JDU,YSO ,NO)
41	the student revise his courses	Courses _Review	(Yes at Home, Yes at the Library, Yes Others, No)	(YAH,YAL, YOTH,NO)
42	Relations with teachers	Relation _Teacher	(Very Good, Good, Bad, Very Bad)	(GOO,VGO,BAD ,VBA)
43	Parents' encouragement	Parents _Encourag	(Yes Often, Yes Sometimes, Yes Rarely , Never)	(YFT,YSO,YON ,NNE)
44	The student make discussions at home with his family which encourages him	enc_ava	(Yes Often, Yes Sometimes, Yes Rarely, Never)	(YFT,YSO,YON ,NNE)

N	***les attributs***	***code attrib***	***valeur attri***	***code et valeur attrib***
45	Do you attend the first years	Assist	(Yes always, Yes sometimes, Yes, only ,No never)	(YID,YSO,YON ,NNE)
46	Prepare TDs during the first year	Assist_Td	(Yes always, Yes sometimes ,Yes, only ,No never)	(YID,YSO,YON ,NNE)
47	Review with colleagues	Review _colleag	(Yes Often, Yes Sometimes, Yes Rarely, Never)	(YID,YSO,YON ,NNE)
48	Frequency of review with colleagues	Freq_ Review_ Colleag	(Every Day, Several Times a Week, About Once a Week, During Exam Period Only, Never)	(ID,PFS,UFS ,PES,NEV)
49	Class participation	Class_ Participation	(Yes Often, Yes Sometimes, Yes Rarely, Never)	(YID,YSO,YON ,NNE)
50	Use of library	Library_Use	(Yes Often, Yes Sometimes, Yes Rarely, Never)	(YID,YSO,YON ,NNE)
51	Absence during lessons	Absence	(Yes Often, Yes Sometimes, Yes Rarely, Never)	(YID,YSO,YON ,NNE)
52	Stress during exams	Stress _Exam	(Yes Often, Yes Sometimes, Yes Rarely, Never)	(YID,YSO,YON ,NNE)
53	The student sanctioned by	Disciplinary _Sanctions	(Yes, No)	(YES,NO)

N	***les attributs***	***code attrib***	***valeur attri***	***code et valeur attrib***
54	Results in S1	S1_Results	(Admis a la session normale,Admis aprés la session rattrapage,Admis avec dettes,Ajourné)	(ASN,AAR,AAD ,AJO)
55	Results in S2	S2_Results	(Admis a la session normale,Admis aprés la session rattrapage, Admis avec dettes ,Ajourné)	(ASN,AAR,AAD ,AJO)
56	Final result	Final_Results	(Admis sans dattes,Admis avec dattes,Ajourné)	(ASD,AAD,AJO)

Bibliographie

Bibliographie

[1] ZIANE, A, ARAHMANE, S 2018 *L'intelligence artificielle pour l'informatique affective : Reconaissance de la valence des emotions par apprentissage profond Doctoral dissertatio*

[2] Peña-Ayala, A. (2014). *data mining : A survey and a data mining-based analysis of recent works. Expert systems with applications, 41(4), 1432-1462.*

[3] Kamath, R. S., Kamat, R. K. (2016). *data mining with R and rattle. River Publishers.*

[4] Stéphane, T. (2012). *Data mining et statistique décisionnelle : l'intelligence des données. Editions Technip.*

[5] Mr : Abdelouahab ATTIA ,20/04/2010 *Présenté à la Faculté des Sciences de l'Ingénieur Département d'Informatique Pour l'Obtention du Diplôme de MAGISTER ÉCOLE DOCTORALE(STIC)*

[6] *https ://www.tutorialspoint.com/data_ mining/dm_ knowledge_ discovery.htm, LEARN DATA MINING,(data pattern evaluatio)*

[7] Jalam, R. (2003). *Apprentissage automatique et catégorisation de textes multilingues. PhD Tesis, Université Lumiere Lyon, 2.*

[8] Uskov, V. L., Bakken, J. P., Howlett, R. J., Jain, L. C. (Eds.). (2017). *universities : concepts, systems, and technologies (Vol. 70). Springer.*

[9] Mathieu-Dupas, E. (2010). *Algorithme des k plus proches voisins pondérés et application en diagnostic.*

[10] ABDERRAHMANE LESHOB MA1 2013 *COMME EXIGENCE PARTIELLE DU DOCTORAT EN INFORMATIQUE*

[11] Dr. Abdelhamid Djeffal. *Cours Fouille de données avancée.Classification*

[12] Touzet, C. (1992). *les réseaux de neurones artificiels, introduction au connexionnisme.*

[13] Alaoui Abdiya 2011/2012 *Application des techniques des métaheuristiques pour l'optimisation de la tâche de la classificationde la fouille de données UNIVERSITE DES SCIENCES ET DE LA TECHNOLOGIE D'ORAN Mohamed BoudiafThème*

[14] Touina, H. (2018). *Classification automatique de textes (Doctoral dissertation, FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE DEPARTEMENT D'INFORMATIQUE).*

[15] Hasan, M., Boris, F. (2006). *Machines à vecteurs de support ou séparateurs à vastes marges. Rapport technique, Versailles St Quentin, France. Cité, 64.*

[16] Koudri Mohammed] 2011 *mémoire fin d'étudeDjillaliLiabesUniversity , Departement of computer science.*

[17] [Michael , al (2007) *: W.B. Michael and MaluCastellanos, survey of textminingclustering, classification and retrieval Survey of TextMining : Clustering, Classification, and Retrieval , Second Edition, Springer*

[18] NACERI, D., MENNOUR, H. (2015). *Implémentation et Génération d'un Ensemble Bagging dans la Plateforme Weka.*

[19] DE LAVERGNE, C., Pauline, L. S. *Dispositifs d'apprentissage hybrides et complexité.*

[20] BOUHADJLA, H. (2017). *Développement d'un Web-Robot pour Analyser les Clés Publiques et Certificats.*

[21] , G. Germain Lacasse.*http ://www.techno-Science.net/ ?onglet=glossaire& Définition=5346.*

[22] ,Iqbal, M., Abid, M. M., Waheed, U., Alam Kazmi, S. H. (2017). Classification of Malicious Web Pages through a J48 Decision Tree, aNaïve *Bayes, a RBF Network and a Random Forest Classifier forWebSpam Detection. International Journal of u-and e-Service, Science and Technology, 10(4), 51-72.*