CRISP-DM Cross Industry Standard Process - Data Mining
REST Representational State Transfer
SVM Support Vector Machines
UAT Univesity Administrative Tools
CHAID Chi-square Automatic Interaction Detector
UNAF Union Nationale des Associations Familiales
RDC République Démocratique du Congo
ULPGL Université Libre des Pays des Grands Lacs
INAPS INapte A Poursuivre les études Supérieurs
RIP Reconnu d'Intérêt Pédagogique
UAT Univeristy Administrative Tool
CSV Commat Separted Values
SQL Structural Query Language
ANOVA Analyse of variance
EXETAT EXamen d'ETAT
DM Données Manquantes

Murhabazi Buzina Espoir Tech 2 GI ULPGL 2017

Sujet : DataMining appliqué à la prédiction de l'orientation des élèves finalistes du secondaire à l'université

INTRODUCTION

0.1 Problématique

Arrivés à la fin de leurs études secondaires la plupart des élèves finalistes du secondaire et, futurs universitaires sont confrontés au problème de choix de filières pour poursuivre leur études universitaires. Plusieurs options s'offrent à eux et ainsi ils se trouvent dans un embarras de choix, la plupart d'entre eux choisissent mal leurs orientations. C'est pour cela que nous remarquons qu'en première année d'université,le pourcentage d'échec ou d'abandon est très élevé suite à une mauvaise orientation des étudiants.[1]

En effet,améliorer le processus d'orientation des nouveaux étudiants à leur entré à l'université pourrait diminuer le taux d'échec et d'abandon en première année .

Il s'avère donc important de doter les universités des outils d'aide à la décision qui pourront leur permettre de bien orienter les étudiants avant d'entreprendre leurs études universitaires et ainsi leur permettre d'y tirer pleine satisfaction.

D'où nous nous sommes posés les questions suivantes :
- *Comment utiliser les techniques du DataMining pour doter les universités des outils d'aide à la décision les permettant de bien orienter les étudiants dès leur entrée à l'université ?*
- *Comment peut-on aider les élèves finalistes du secondaires à pouvoir faire le choix de leur filières à l'université ?*
- *Les techniques du DataMining peuvent - elles apporter leur contribution dans ce domaine ?*

0.2 Intérêts et Motivations du Sujet

Nous avons choisi de parler du sujet sur le DataMining à cause de notre passion pour les mathématiques et les sciences prédictives mais aussi car au 21

ème siècles l'immensité des données générées par les systèmes d'information ne cesse de croitre d'où il s'avère important de les analyser et apprendre de ces données !
Mais aussi l'intérêt scientifique de ce travail sera de fournir un outil de base à tous chercheur qui aimera aussi travail dans ce domaine dans les jours à venir.
Nous avons choisi de parler de l'orientation des étudiants car durant notre parcours universitaire nous avons constaté un fort taux d'abandon et d'échec due à une mauvaise orientation et ainsi nous voulions aider tant soit peut à résoudre ce problème.

0.3 Hypothèses de Travail

Une hypothèse est une supposition qui est faite en réponse à une question de recherche. [2]
Pour notre travail nous supposerons que les universités ainsi disposent d'une énorme quantité des données et qu'on peut les analyser enfin d'y découvrir des patterns cachés qui peuvent nous permettre de prédire l'orientation d'un nouvel étudiant sur base de ses résultats à l'école secondaire.

0.4 Objectifs du travail

L'objectif d'une recherche se divise en deux parties : l'objectif général concerne la contribution que les chercheurs espèrent apporter en étudiant un problème donné ; les objectifs opérationnels concernent les activités que les chercheurs comptent mener en vue d'atteindre l'objectif général. [2]
Pour notre travail l'objectif général sera de fournir un outil d'aide à la décision en nous basant sur les techniques du DataMining plus précisément les arbres de décisions et la technique du random forrest.
Cet outil se présentera sous forme de service web de type Representational State Transfer (REST) que les universités peuvent intégrer facilement dans leurs systèmes d'informations.

0.5 Méthodes et Techniques de Recherche

Pour atteindre notre objectif,notre travail utilisera la méthodologie de Cross Industry Standard Process - Data Mining (CRISP-DM) qui est le processus standard d'un projet DataMining, elle définit les étapes pour la

conduite d'un projet en le rendant plus efficace plus rapide et moins couteux. [3]
Cette méthodologie se base sur la technique documentaire qui consister à consulter les archives des universités et ainsi que le système d'information Univeristy Administrative Tool (UAT)en vue d'y collecter les données.
Elle utilise également les enquêtes ainsi que les interviews auprès des conseillers d'orientations d'universités pour savoir comment se déroule l'activité de l'orientation des étudiants.

0.6 Subdivision du travail

Hormis l'introduction et la conclusion ce travail sera subdivisé en cinq chapitres.

1. Le premier chapitre sera intitulé : Les généralités sur le DataMining.
2. Le Second sera intitulé : Analyse du domaine de l'orientation.
3. Le troisième sera intitulé : Présentation et exploitation des données obtenus.
4. Le quatrième sera intitulé : Conception ,développement et Présentation de notre Solution.
5. le Cinquième sera intitulé : Planification du Projet .

Chapitre 1

Généralités sur le DataMining

Nous sommes submergées des données et leur quantité augmente du jour au lendemain . Les ordinateurs, les smart-phones et de plus en plus des équipements connectées nous submergent et sont omniprésents dans notre vie et cela nous permet de générer des très grandes quantités des données , toutes nos décisions , nos choix dans les supermarchés,nos habitudes financières sont sauvegardées dans d'énormes bases des données.

L'internet est aussi submergé des informations c'est ce qui fait que chaque choix, chaque clic que nous faisons soit sauvegardé , ceci n'étant que des choix personnels mais qui ont d'innombrables contreparties dans le monde du commerce et de l'industrie . Mais paradoxalement on a pu constaté que plus les données augmentent et sont générées de moins en moins les personnes les comprennent et ainsi un immense fossé s'est créé entre le volume des données générées et la capacité de compréhension de celles ci. D'où l'importance de mettre en place des méthodes et techniques qui faciliterons l'analyse et l'obtention des informations considérables de ces données.

1.1 Définitions

Dans cette partie nous allons définir 3 termes qui portent souvent à confusion : L'intelligence Artificielle,Le Machine Learning ou apprentissage automatique, La fouille des données ou Le DataMining, nous tenterons de dégager à la fin les différences et les ressemblances entre ces termes.

1.1.1 L'intelligence Artificielle

L'intelligence artificielle(IA) est un domaine de l'informatique dédié à la création de matériel et de logiciels capables d'imiter la pensée humaine. Le

but principal de l'intelligence artificielle est de rendre les ordinateurs plus intelligents en produisant des logiciels permettant à un ordinateur d'émuler des fonctions du cerveau humain dans des applications définies. L'idée n'est pas de remplacer l'être humain mais de lui donner un outil plus puissant afin de l'aider à accomplir ses tâches.[4]
L'Intelligence artificielle est un domaine de l'Informatique qui a pour but de développer des machines (= ordinateurs) "intelligentes", c'est-à-dire capables de résoudre des problèmes pour lesquels les méthodes conventionnelles sont inefficaces et inapplicables. [4]

1.1.2 Le Machine Learning

Le Machine Learning est définie comme une branche de l'intelligence artificielle qui se penche sur la création des algorithmes qui peuvent apprendre et faire des prédictions à partir des données[5]

D'autres auteurs le définissent comme :
- Un type de l'intelligence artificielle qui donne aux machines la possibilité d'apprendre sans être explicitement programmer[6]
- La science qui consiste à la création des logicielles qui apprennent d'eux même en fonction des données.[6]

1.1.3 Le DataMining

Le DataMining ou la fouille de données est une technique consistant à rechercher et extraire de l'information (utile et inconnue) de gros volumes de données stockées dans des bases ou des entrepôts de données en utilisant les techniques du Machine Learning .[5]

Certains auteurs comme M. BATER[7] le définissent comme une discipline scientifique qui a pour but l'analyse exploratoires des grandes quantités des données , la découvertes des modèles utiles ,valides , inattendus ainsi que la connaissances compréhensibles dans celles ci. Outre la découvertes de l'information il englobe la collecte , le nettoyage et le traitement de ces données.

Le datamining peut aussi être défini comme un processus inductif, itératif et interactif de découverte dans les bases de données larges de modèles des données valides, nouveaux, utiles et compréhensibles.[8] - Itératif : nécessite plusieurs passes
- Interactif : l'utilisateur est dans la boucle du processus
- Valides : valables dans le futur
- Nouveaux : non prévisibles

- Utiles : permettent à l'utilisateur de prendre des décisions
- Compréhensibles : présentation simple

1.1.4 Différence entre le DataMining , le Machine Learning, ainsi que les statistiques . [6] [9] [10]

A première vue on pourrait constater que ces 3 disciplines n'ont aucune différence car tous traitent de la même question : comment apprendre des données ?, couvrent les mêmes matières et utilisent les mêmes techniques qui sont entre autres : la régression linéaire, la régression logistique , le réseau des neurones , le Support Vector Machine ,....

Mais en analysant de prêt on remarque qu'ils ont des différences surtout concernant les sujets et matières sur lesquels ils insistent, ou en d'autres termes quand et comment les méthodes qu'ils ont en commun sont utilisées.

- Les statistiques insistent sur les méthodes de la statistique référentielle,descriptives , multivariés (intervalle de confiance, test des hypothèses , estimation optimale) dans un problème à faible dimension (une petite quantité des données) et elles font des prédiction sur base de celle ci.[9]

- Le Machine Learning est beaucoup plus concentré vers le Génie logiciel il est beaucoup plus focalisé sur la construction des logicielles qui font des prédiction à partir des données apprennent de l'expérience. On dit qu'un programme apprend de l'expérience si ses performances sur les taches qu'il exécute augmentent avec l'expérience . ces performances sont évaluées à l'aide de certaines métriques que nous verront dans la suite.
Il nécessite l'étude des algorithmes qui peuvent extraire l'information utile dans des grosses quantités des données automatiquement , dans certaines mesures il s'inspirent des statistiques. [10]

-Le DataMining quand à lui s'inspire des techniques du Machine learning , des statistiques souvent avec un objectif bien fixé en vue de découvrir des patterns cachées dans des grosses volumes des données. - L'intelligence artificielle se focalise sur la création des agents intelligents , en pratique il consiste à programme un ordinateur pour qu'il simule le fonctionnement du cerveau humain . Qu'il exécute certaines taches comme un homme parmi ces taches figure la faculté d'apprendre sur base de l'expérience.

1.2 Différentes Techniques du Machine Learning [11]

Dans cette partie nous allons passer en revue quelques méthodes et algorithmes utilisées dans le machine learning. Notons que ces méthodes se divisent en 2 groupes : l'apprentissage supervisée et l'apprentissage non supervisé.Concernant l'apprentissage supervisé nous parlerons de la régression linéaire , la régression, le réseau des neurones, le SVM(Support Vector Machine) , les arbres de décision et nous finirons par le Random Forrest qui n'est qu'une amélioration des arbres de décisions, et pour l'apprentissage non supervisé nous allons parler de K-means.

1.2.1 Régression Linéaire

La régression est une technique consistant à prédire la sortie d'une observation en partant d'un certain nombre des variables en entrée.
On part d'un certain nombre des données d'apprentissage ou d'un training set avec :
m : nombres d'éléments de l'ensemble d'apprentissage
x_1 ,x_2,x_3 .... x_n : les variables d'entrées(qui peuvent varier de 1 (pour la régression linéaire avec une seule variable) à n pour plusieurs variables).
y : la variable de sortie on notera le (x_1,x_2,x_3 .... x_n, y) : un exemple d'apprentissage et
(x_1^i,x_2^i,x_3^i.....x_n^i,y^i) un exemple choisie le ème exemple d'apprentissage avec i comme index sur nos données.

Le but de la régression c'est de trouver la fonction qui permet de prédire la sortie en fonction de l'entrée cette fonction doit minimiser l'erreur entre les valeurs de la fonction aux points d'apprentissage et la valeur de sortie dans les données d'apprentissage .
Cette fonction on l'appelle hypothèse qui sera une droite ou un polynôme selon qu'on utilise la régression linéaire ou polynomiale.
L'hypothèse se présente de la manière suivante :

1 ère cas régression avec un seule variable :

Par exemple on tente de prédire le prix d'une maison en fonction de sa superficie :

$h_\theta(x) = \theta_0 + \theta_1 x$.

Cela signifie que y est une fonction linéaire de x avec θ_i des paramètres qu'on cherchera à déterminer . on peut le remarque sur la figure suivante ou notre hypothèse est une droite :

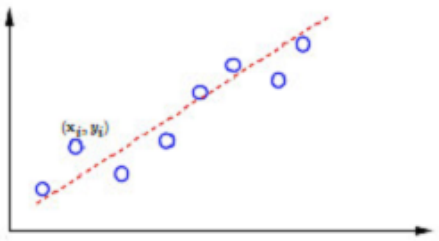


FIGURE 1.1 – Régression Linéaire avec une variable

2 ème cas régression avec plusieurs variables :

Pr exemple prédire le prix d'une maison cette fois en fonction de la superficie, du nombre des chambres ,et de l'année de construction

$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + ...\theta_n x_n$ Dans ce cas notre hypothèse sera en fonction des variables en entré un plan ou un hyperplan.

3 eme cas la regression polynomiale :

Dans ce cas au lieu d'utiliser une droite on utilise un polynôme à n degree qui es donnée par :

$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2^2 + \theta_3 x_3^3 + ...\theta_n x_n^n$ le travail restant sera de trouver

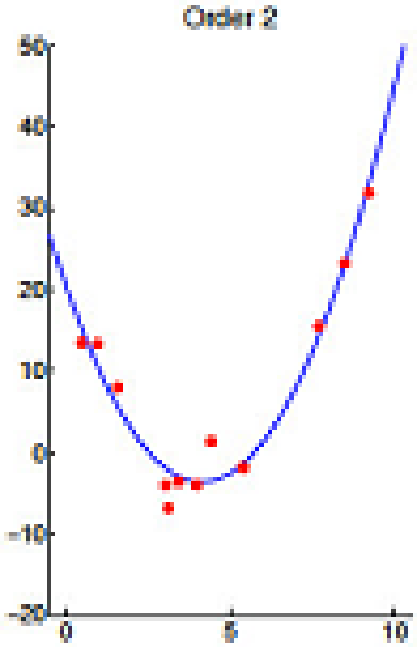


FIGURE 1.2 – Régression Polynomiale avec un polynôme de degré 2

les paramètres θ_i pour notre fonction $h_\theta(x)$ mais comment les trouver ?

Fonction Cout et calcul de L'erreur

On appelle l'erreur de prédiction , la valeur définie par :

$J(\theta_1, \theta_2, ...\theta_n) = \frac{1}{2m} \sum_{=1}^{m} \left[h_\theta\left(x^{(i)}\right) - y^{(i)}\right]^2$ avec m : le nombre des nos données d'apprentissage. Cette fonction représente l'erreur commise lors de la prédiction avec notre hypothèse par rapport à la valeur exacte. Les θ_i sont les valeurs qui minimisent cette erreur sur toutes les données de notre apprentissage.

Il existe différents techniques de minimisation de cette fonction entre autre :

- L'annulation de la dérivée première (trouver les valeurs θ_i qui annulent notre dérivée première) .Cette méthode à pour inconvénient le fait qu'il

convient pas pour les données d'apprentissage avec plusieurs attribues et plusieurs données.

- une autre méthode c'est la descente du gradient : celle ci consiste à effectuer plusieurs itérations sur les valeurs de θ_i jusqu'à trouver celle qui minimise l'erreur (jusqu'à ce qu'il converge vers zéro)

voici l'algorithme utilisée : 1

Algorithm 1 Algorithme de la descente du gradient

$\theta_i \leftarrow 0$

while J ne converge pas **do** ▷ faire pour chaque tuple

$\theta_i \leftarrow \theta_i - \alpha \frac{\partial}{\partial \theta_i} J(\theta_1, \theta_2, ...\theta_n)$

avec α : learning rate qui est compris entre [0,001 ;0,1] il sert à réguler notre algorithme

s'il est trop grand θ ne converge pas s'il est trop petit θ converge après plusieurs itérations. Dans la plupart des cas J converge après un nombre d'itérations élevé comme on peut le remarqué sur la figure suivante :

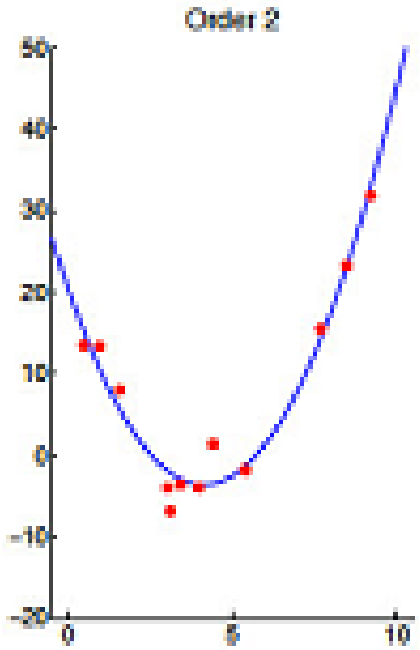


FIGURE 1.3 – Variation de J en fonction du nombre des itérations

NB : La Normalisation

Dans la pratique on peut avoir plusieurs données qui ne sont pas dans la même échelle

par ex : on cherche à prédire le prix d'une maison en fonction de la surface x_1 : compris entre 30 – 400 m^2, le nombres des chambres x_2 : 1-10,
Nous remarquons que nos 2 variables ne sont pas dans la même intervalle et cela peut présenter un problème de convergence et ainsi empêcher les θ_i de converger rapidement . pour limiter ce problème on effectue un normalisation elle consister à remplacer les x_i pour chaque tuple par :

$x_j^i = \frac{x_j^i - \mu(x_j)}{\sigma(x_j)}$ avec :

$\mu(x_j)$: la moyenne des termes x_j et $\sigma(x_j)$: la variance des termes x_j

1.2.2 Régression Logistique

La régression logistique est une technique de classification , elle consiste à appliquer une fonction h sur des éléments x_1 ,x_2,x_3 .... x_n en entrée et trouver une sortie *discrète* y.
Cette sortie nous permet de déterminer la classe du tuple , généralement y prend 2 valeurs 0 ou 1 et quelque fois -1 ou 1.
On utilise ce type de classification dans différents cas :

- Classification des emails : spams ou non spam
- Transaction en Ligne : Frauduleux ou pas
- Crédit bancaire : risquant ou non
- Tumeur : Bénigne ou Maligne

Représentation de l'hypothèse

Une première approche serait d'utiliser la régression linéaire avec la fonction $h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + ... \theta_n x_n$ pour effectuer la classification mais cela peut un problème car $h_\theta(x)$ peut prendre des valeur qui sont à l'extérieur de l'intervalle [0,1]. La meilleur solution serait de trouver une fonction qui ne prend que des valeurs comprises entre 0 et 1 .

Pour les problèmes de classification on utilise la fonction sigmoïde ,elle est définie par :

$$g(z) = \frac{1}{1+e^{-z}}$$

Et ainsi notre hypothèse sera :

$$h_\theta(x) = g(\theta^T x)$$

Avec $\theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + ... \theta_n x_n$. Voici comment se présente cette fonction :

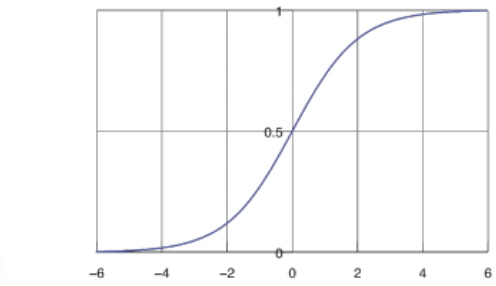


FIGURE 1.4 – Fonction Sigmoïde

On remarque que cette fonction ne prend que des valeurs comprise entre 0 et 1.

En d'autres termes $h_\theta(x)$ représente la probabilité que y=1 sur l'entrée x_n .

$$h_\theta(x) = P(y = 1|x; \theta)$$

La probabilité que y soit égale à 1 sachant x_n paramétré par θ.

On l'utilise pour effectuer une classification en supposant que :

- si $h_\theta(x) > 0.5$ la classe prédite est 1.
- si $h_\theta(x) < 0.5$ la classe prédite est 0.

On démontre que :

$$P(y = 1|x; \theta) + P(y = 0|x; \theta) = 1 \text{ et}$$
$$P(y = 0|x; \theta) = 1 - P(y = 1|x; \theta)$$

Frontière de Décision

Analysons notre fonction $g(z)$, on remarque qu'il est égale à 0.5 pour z = 0. Donc si z est positif $g(z)$ sera supérieur à 0.5, dans le cas contraire $g(z)$ sera inférieur à 0.5. Et ainsi avec notre hypothèse $h(\theta^T x) = 1$ si $\theta^T x > 0$ et $h(\theta^T x) = 0$ si $\theta^T x < 0$ On comprend qu'a partir du plan $\theta^T x$ qui divise l'espace en 2 partie on peut prédire la classe de chaque tuple sans problème .

Cette droite ou hyperplan s'appelle ***la frontière de décision*** . Voici comment elle se présente si on n'a que 2 attributs x_1 et x_2 :

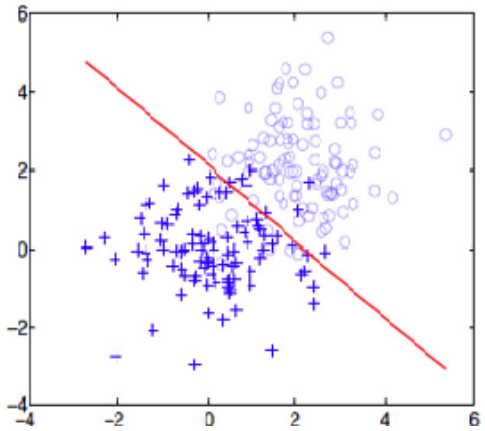


FIGURE 1.5 – Frontière de décision pour la Régression Logistique à 2 variables

La Fonction Coût : Calcul De L'erreur

On pourrait être tenter d'utiliser la même fonction coût que pour la régression linéaire

$$J(\theta_1, \theta_2, ...\theta_n) = \frac{1}{2m} \sum_{=1}^{m} \left[h_\theta\left(x^{(i)}\right) - y^{(i)}\right]^2$$

Mais cette fonction présente des problèmes de convergence à cause de la non linéarité de la fonction $h_\theta(x)$. Pour ce faire,on définit une nouvelle fonction de l'erreur :

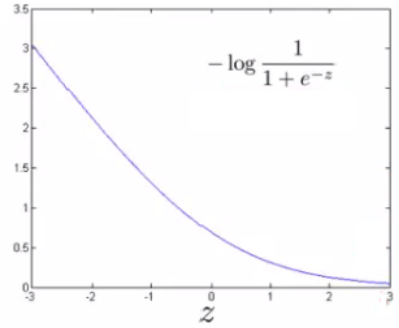
$$Err(h_\theta(x), y) = \begin{cases} -\log[1 - h_\theta(x)], & \text{si } y{=}0 \\ -\log[h_\theta(x)], & \text{si } y{=}1 \end{cases}$$

Analyse de l'erreur :

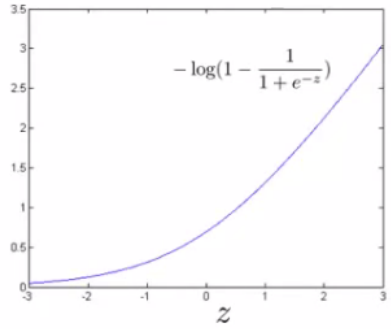
- si y= 1 et notre fonction $h_\theta(x) = 1$, alors l'erreur vaut 0 et augmente si $h_\theta(x)$ décroit.

- si y= 0 et notre fonction $h_\theta(x) = 0$, alors l'erreur vaut 0 et augmente si $h_\theta(x)$ croit .

On peut facilement le remarquer sur les figures suivantes :
Notre fonction cout peut prendre la forme compacte suivante :



(a) cas ou Y =1



(b) Y = 0

FIGURE 1.6 – Erreur pour la régression logistique

$$Err(h_\theta(x), y) = -y \log[h_\theta(x)] - (1 - y) \log[1 - h_\theta(x)]$$

Et Donc notre Fonction cout pour les θ sera :

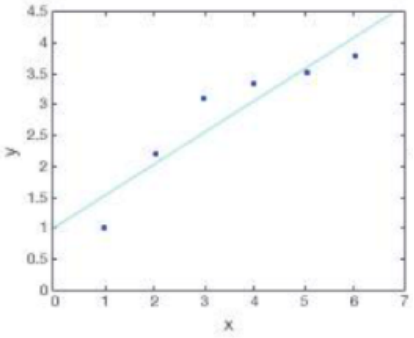
$$J(\theta) = \frac{1}{2m}[\sum_{=1}^{m} -y^{(i)} \log[h_\theta\left(x^{(i)}\right)] - (1 - y^{(i)}) \log[1 - h_\theta\left(x^{(i)}\right)]]$$

Avec $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$

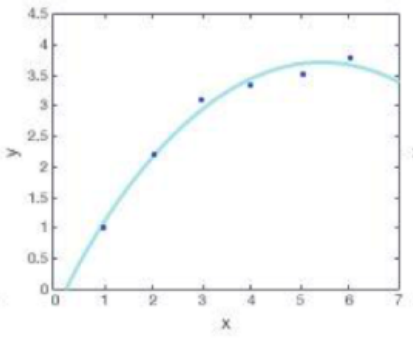
Pour le minimiser,on recours aux mêmes techniques que pour la régression linéaire et c'est le même algorithme de la descente du gradient qu'on utilise dans ce cas aussi.

Régulation : Phénomène de sur-apprentissage

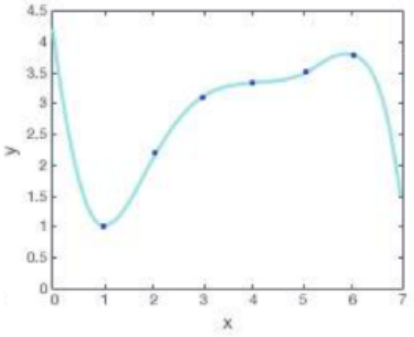
Nous venons de passer en revue quelques modèles pour effectuer la régression linéaire ainsi que la classification avec la régression logistique. Jetons un coup d'œil aux 3 images de la 'FIGURE 7
A la première figure (a) : avons la régression linéaire comme hypothèse mais



(a) $h_\theta(x) = \theta_0 + \theta_1 x$



(b) $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$



(c) $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

FIGURE 1.7 – a : Sous-apprentissage et c : Sur-apprentissage avec la régression linéaire

on remarque que cela n'est pas un bon modèle de prédiction car dans ce cas l'erreur est trop grande.on parle de *sous-apprentissage ou Underfitting*. A la deuxième figure notre hypothèse c'est un polynôme du second dégrée le modèle convient bien à notre ensemble d'apprentissage et dans ce cas l'erreur est moins élevée. Par contre sur la troisième cas l'hypothèse c'est un polynôme du quatrième dégrée qui convient parfaitement à notre ensemble d'apprentissage et l'erreur est nulle, mais c'est modèle n'est pas bon car il convient parfaitement sur notre ensemble d'apprentissage et n'est pas adaptable aux

nouvelles données, elle n'est pas une solution générale.Dans ce cas on parle de *Sur-Apprentissage ou Overfitting* .

On remarque ce même phénomène avec la régression logistique .

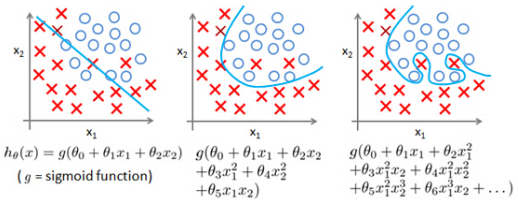


FIGURE 1.8 – Overfitting et Underfitting avec la régression logistique

Solutions aux problèmes de sur apprentissage

Signalons qu'une des cause de ce phénomène c'est le fait d'avoir peut des données avec plusieurs attributs .

Dans ce cas la première de solution consisterait à réduire le nombre des attribues , choisir celle qu'on doit retenir et laisser les autres,mais cette approche possède un inconvénient car on risquerait de perdre l'information apportée par ces attributs et cela pénaliserait notre modèle .

La seconde solution est celle qu'on appelle la *La régularisation* elle consiste à garde tous les attribues mais réduire l'ampleur des paramètres θ dans le calcul de l'erreur . On remarque que cette technique fonctionne très bien lorsqu'on dispose de plusieurs attribues qui contribue à la prédiction de la valeur y. Ainsi notre fonction de cout pour le calcul de l'erreur de la régression linéaire sera :

$$J(\theta_1, \theta_2, ...\theta_n) = \frac{1}{2m}[\sum_{=1}^{m} \left[h_\theta\left(x^{(i)}\right) - y^{(i)}\right]^2 + \lambda \sum_{=1}^{n} \theta_j^2]$$

Et Pour la régression logistique elle sera :

$$J(\theta) = \frac{1}{2m}[\sum_{=1}^{m} -y^{(i)} \log[h_\theta\left(x^{(i)}\right)] - (1 - y^{(i)}) \log[1 - h_\theta\left(x^{(i)}\right)] + \lambda \sum_{=1}^{n} \theta_j^2]$$

Avec λ le facteur de régularisation. C'est cette nouvelle fonction cout qu'on utilise pour l'algorithme de la descente du gradient régularisé.

1.2.3 Le Réseaux des Neurones

Pourquoi avons nous besoins du réseau des neurones ? Supposons que nous avons un problème complexe de classification, la première approche serait d'utiliser la classification logistique avec un polynôme de plusieurs termes :

- cela conviendrait pour problème avec 1, ou 2 attribut.

- supposons maintenant que nous avons des tuples avec plus de 1000 attributs ,

Ex : On veut prédire la chance d'une maison d'être vendu dans 6 mois : - Plusieurs facteurs entrent en jeu ,à peut près plus de 100 , on aura un polynôme avec les termes suivants $(x_1^2, x_1x_2, x_1x_3, x_1x_4, ...x_1x_{100})$, on aura a peut près 5000 termes et ce nombre augmentera énormément si on choisi de considérer les 3 ème degré. La régression logistique n'est pas vraiment approprié pour des problèmes de classification des données avec plusieurs attributs .

Par exemple dans le cas de la vision artificielle ou n peut aller de 2500 à 50000000 attributs. On remarque bien qu'une simple régression logistique n'est pas approprié pour ce cas.

Les Neurones et le cerveau[12]

Une des motivations qui ont poussé les scientifique à créer le réseaux des neurones c'est celui de répliquer le fonctionnement du cerveau humain . En effet le cerveaux humain dispose d'une capacité énorme d'apprentissage il s'apprend lui même comment apprendre et peut apprendre de n'importe quelle source des données,il contient environ 100 milliards de neurones. Voici à quoi ressemble à quoi ressemble un neurone dans le cerveau :

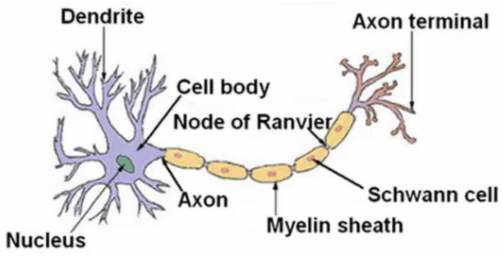


FIGURE 1.9 – Un Neurone Naturel

Les neurones reçoivent les signaux (impulsions électriques) par des extensions très ramifiées de leur corps cellulaire (les dendrites) et envoient l'information par de longs prolongements (les axones). Les impulsions électriques sont régénérées pendant le parcours le long de l'axone. La durée de chaque

impulsion est de l'ordre d'1 ms et son amplitude d'environ 100 mvolts. Les contacts entre deux neurones, de l'axone à une dendrite, se font par l'intermédiaire des synapses. Lorsqu'un potentiel d'action atteint la terminaison d'un axone, des neuromédiateurs sont libérés et se lient à des récepteurs post-synaptiques présents sur les dendrites. L'effet peut être excitateur ou inhibiteur. Chaque neurone intègre en permanence jusqu'à un millier de signaux synaptiques. Ces signaux n'opèrent pas de manière linéaire (effet de seuil).

Représentation d'un Neurone Artificiel

On pourrait définir un neurone ou perceptrons comme étant une unité de traitement qui reçoit des données en entrée sous forme vectorielle et produit une sortie réelle , cette sortie est fonction des entrées et des poids de connexions.

Il se caractérise par :

- les signaux en entrée $x_1, x_2, ..., x_n$,

il est toujours préférable d'ajouter le biais x_0 qui est égale à 1.,

- les coefficients synaptiques ou poids des connexions $\theta_{i0}, \theta_{i1}, ...\theta_{in}$,
- Une Fonction d'activation $h_\theta(x)$,
- l'état interne d'activation $a = h_\theta(x)$

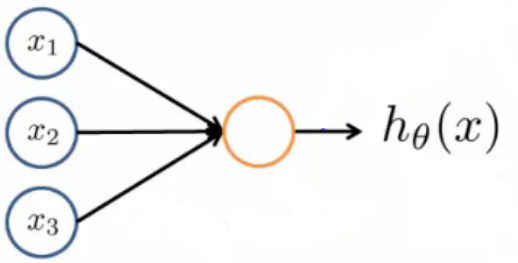


FIGURE 1.10 – un Perceptron ou Unité de traitement

D'une façon plus générale un réseau des neurones est un ensemble de plusieurs neurones liés entre eux pour effectuer un calcul se compose :

- D'une couche d'entré ou activation layer
- D'une couche de sortie ou output layer
- D'une ou plusieurs couches intermédiaires ou hidden layers .

Chaque couche dispose des perceptrons ou unités d'activation.

On note :

- a_i^j : l'unité d'activation i de la couche j

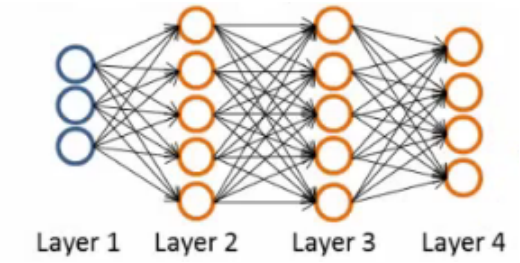


FIGURE 1.11 – un réseau des neurones complets avec 2 hidden layers

- θ^l : matrice des poids de connexions contrôlant le passage de la couche l à couche l+1.
La matrice θ^l est constituée des termes θ_{ij} avec :
- i l'indice de l'unité de la couche de destination
- j l'indice de l'unité de la couche d'origine .
Donc θ_{12} représente le poids de passage de l'unité 2 de la couche d'origine a l'unité 1 de la couche de destination.
Regardons encore une fois l'image suivante :

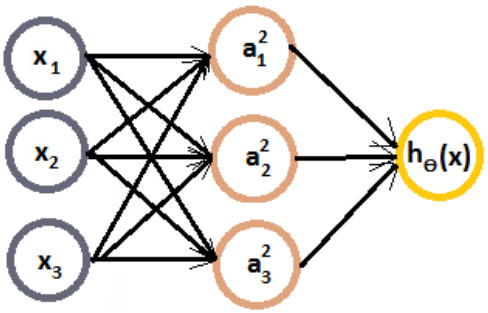


FIGURE 1.12 – Réseaux des neurones avec des unités d'activation

Nous avons :
$a_1^{(2)} = g(\theta_{10}^{(1)}x_0 + \theta_{11}^{(1)}x_1 + \theta_{12}^{(1)}x_2 + \theta_{13}^{(1)}x_3)$
$a_2^{(2)} = g(\theta_{20}^{(1)}x_0 + \theta_{21}^{(1)}x_1 + \theta_{22}^{(1)}x_2 + \theta_{23}^{(1)}x_3)$
$a_3^{(2)} = g(\theta_{30}^{(1)}x_0 + \theta_{31}^{(1)}x_1 + \theta_{32}^{(1)}x_2 + \theta_{33}^{(1)}x_3)$
$h_\theta(x) = a_1^{(3)} = g(\theta_{10}^{(2)}a_0^{(2)} + \theta_{11}^{(2)}a_1^{(2)} + \theta_{12}^{(2)}a_2^{(2)} + \theta_{13}^{(2)}a_3^{(2)})$

Chaque unité d'activation de la couche l+1 est obtenue en appliquant la fonction g (une sigmoïde) aux à la combinaison linéaire des unités de la couche précédente avec les éléments de la matrice θ^l.

Apprentissage par Réseau des neurones

Le réseau des neurones est un des plus puissant algorithme d'apprentissage , il cherche à trouver les paramètres du modèle d'apprentissage uniquement à partir de l'ensemble d'apprentissage .
Voici notre configuration :

- Notre ensemble d'apprentissage est : $(x^1, y^1), (x^2, y^2), (x^3, y^3), ...(x^n, y^m)$
- L nombre des couche de notre réseau . Pour notre cas L=4
- s_l nombre des unités dans la couche l

Donc en se référant à la FIGURE 11 On remarque que :

- L= 4
- $s_1 = 3$
- $s_2 = 5$
- $s_3 = 5$
- $s_4 = 4$

Avec le réseaux des neurones on peut faire aussi bien des classification binaires que des classifications avec plusieurs classe.
Pour la classification binaire la couche de sortie ne comprend qu'une seule unité k = 1 avec k le nombre des unités de la couche de sortie.
Pour une classification multi-classe k prend des valeurs supérieurs à 3,
la sortie $y \in R^K$ Pour 3 classes $y^1 = [1, 0, 0]$, $y^2 = [0, 1, 0]$, $y^3 = [0, 0, 1]$;

Fonction Cout

Rappelons que pour la régression logistique la fonction cout régularisée est la suivante :

$$J(\theta) = \frac{1}{2m}[\sum_{i=1}^{m} -y^{(i)} \log[h_\theta\left(x^{(i)}\right)] - (1 - y^{(i)}) \log[1 - h_\theta\left(x^{(i)}\right)] + \lambda \sum_{j=1}^{n} \theta_j^2]$$

Pour le réseau des neurones la fonction cout n'est que la généralisation de celui de la régression logistique mais au lieu d'une sortie nous avons K sorties .

$h_\theta(x) \in R^k$ et $(h_\theta(x))_k = k^{eme}$ sortie

$J(\theta) = \frac{1}{2m}[\sum_{i=1}^{m} \sum_{k=1}^{K} -y_k^{(i)} \log[(h_\theta\left(x^{(i)}\right))_k] - (1 - y_k^{(i)}) \log[1 - (h_\theta\left(x^{(i)}\right))_k] + \lambda \sum_{l=1}^{L-1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} (\theta_{ji}^{(l)})^2]$

Une fois la fonction cout connu le reste du travail consiste à entrainer notre réseau , donc de trouver :

- la valeur des unités d'activation a^l pour chaque couche
- Ainsi que celles de $\theta_{ji}^{(l)}$ qui minimisent la fonction cout .

Calcul des Valeurs des unités d'activation *Forward propagation*

Cette technique est celle qui permet de calculer de calculer les valeur des unités d'activation du réseau des neurones en partant de la couche d'entrée et propagé ces valeurs jusqu'à la couche de sortie d'où le nom de *Forward propagation.*
Nous avons déjà établie que :
- $a_1^{(2)} = g(\theta_{10}^{(1)}x_0 + \theta_{11}^{(1)}x_1 + \theta_{12}^{(1)}x_2 + \theta_{13}^{(1)}x_3)$
- $a_2^{(2)} = g(\theta_{20}^{(1)}x_0 + \theta_{21}^{(1)}x_1 + \theta_{22}^{(1)}x_2 + \theta_{23}^{(1)}x_3)$
- $a_3^{(2)} = g(\theta_{30}^{(1)}x_0 + \theta_{31}^{(1)}x_1 + \theta_{32}^{(1)}x_2 + \theta_{33}^{(1)}x_3)$

on définie :
$z_1^2 = \theta_{10}^{(1)}x_0 + \theta_{11}^{(1)}x_1 + \theta_{12}^{(1)}x_2 + \theta_{13}^{(1)}x_3$ et cela veut dire que :
$a_1^{(2)} = g(z_1^2)$
On peut ainsi définir les termes z_2^2 ,et z_3^2 qui sont les unités d'activation de la couche 2.
En utilisant les matrices nous pouvons écrire :

$$z^{(2)} = \begin{bmatrix} z_1^{(2)} \\ z_2^{(2)} \\ z_3^{(2)} \end{bmatrix} \text{ pour la 2ème couche et l'entrée peut être noté } x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Ainsi $z^{(2)} = \theta^{(1)}x$ et $a^{(2)} = g(z^{(2)})$
Pour la troisième couche écrit : $z^{(3)} = \theta^{(2)}a^2$ et $a^{(3)} = g(z^{(3)})$ et ainsi de suite jusqu'à la couche de sortie .
D'une façon générale on écrit : $z^{(l)} = \theta^{(l-1)}a^{l-1}$ et $a^{(l)} = g(z^{(l)})$ pour toutes les couches du réseau .

Minimisation de la fonction de cout : back propagation

Pour minimiser la fonction cout on doit calculer ses dérivées partielle par rapport aux éléments $\theta_{ji}^{(l)}$ ce qui n'est pas une tache facile . Le but est de trouver les $\theta_{ji}^{(l)}$ qui minimisent la fonction cout. On veut calculer $\frac{\partial}{\partial\theta_{ji}^{(l)}}J(\theta)$
Pour calculer les dérivées partielles on utilise la technique qu'on appelle *back propagation* ou *rétro-propagation du gradient* ,c'est une méthode qui permet de calculer le gradient de l'erreur pour chaque neurone d'un réseau de neurones, de la dernière couche vers la première.
Notons δ_j^l : l'erreur à l'unité j de la couche l.
- Pour la 4eme couche cette erreur n'est que la différence entre la valeur exacte et la valeur calculé par notre réseau des neurones.

D'ou : $\delta_j^4 = a_j^4 - y_j$
Pour calculer l'erreur de la couche l on utilise la couche suivante l+1 d'ou le

non de *rétro-propagation du gradient* Ainsi l'algorithme d'entrainement de notre algorithme sera le suivant : **??**

Algorithm 2 Algorithme de la rétro-propagation du gradient

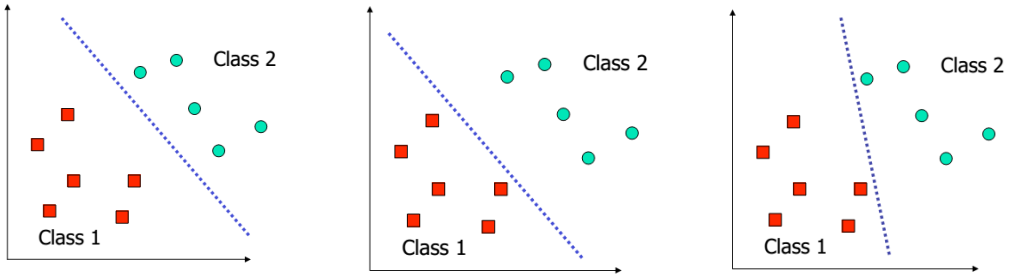
Require: $(x^1, y^1), (x^2, y^2), ...(x^n, y^m)$
  for $i = 1, 2...., s_{l+1}, j = 1, 2, ...s_l; l = 1, 2, ...L$ **do**
    $\theta_{ij}^{(l)} \leftarrow 0$ ▷ initialsation des valeurs des θ pour chaque couche à Zero
  for all tuple dans l'ensemble d'apprentissage **do**
    $a^1 \leftarrow x^i$ ▷ initialisation de la premiere couche avec les valeurs de l'entréé
    repeat
      for L =2,3.....L **do** ▷ Forward Propagation
        $z^{(l)} \leftarrow \theta^{(l-1)} a^{l-1}$
        $a^{(l)} \leftarrow g(z^{(l)})$
      $\delta^L \leftarrow a^{(L)} - y^{(i)}$ ▷ pour la dernière couche
      for L =L-1,....,2,1 **do** ▷ Backward Propagation
        $\delta^l \leftarrow (\theta^{(l)})^T \delta^{l+1} .* (a^{(l)} .* (1 - a^{(l)}))$
        $\frac{\partial J(\theta)}{\partial \theta^{(l)}} \leftarrow \delta^{l+1} . (a^l)^T$
        $\theta^{(l)} \leftarrow \theta^{(l)} + \frac{\partial J(\theta)}{\partial \theta^{(l)}}$ ▷ Descente d gradient
    until $J(\theta)$ converge vers 0
    if $j \neq 0$ **then**
      $D^l \leftarrow \frac{1}{2m} \theta^{(l)} + \lambda \theta^{(l)}$ ▷ Regularisation
    else
      $D^l \leftarrow \frac{1}{2m} \theta^{(l)}$ ▷ Au cas ou il n'ya pas de regularisation

1.2.4 Support Vector Machine - SVM [13]

Support Vector Machine est un classeur discriminant qui est défini par un hyperplan , étant donné un ensemble d'apprentissage l'algorithme de SVM permet de trouver un hyperplan qui va catégoriser les éléments de l'ensemble et d'autres nouveaux éléments en maximisant la marge entre les élements de l'ensemble d'apprentissage à la manière de la frontière de décision pour la régression logistique. On peut dire que SVN est une amélioration de la régression logistique .

Données Séparables linéairement

Si nous considérons le même problème de classification binaire que celui de la régression logistique sauf que les label de nos classe ici sont 1 et -1 au lieu de 0 et 1 donc ici y $\in$ {-1,1} , on constate que nous avons un large variétés des choix pour notre frontière de décision comme on peut le voir à la figure ci dessous : Nous remarquons que toutes ces frontières des décisions



(a) (b) (c)

FIGURE 1.13 – Différentes frontières de décision possibles

conviennent pour notre ensemble d'apprentissage.
Mais comment alors choisir la bonne ? ? celle qui est optimale ?
Remarquons que pour mieux généraliser les données la meilleur frontière de décision est celle qui doit être le plus éloigné que possible des nos données . Le but de SVM revient à trouver cette hyperplan qui constituerai notre frontière de décision.

Représentation de l'hypothèse Rappelons que pour la régression logistique notre hypothèse s'écrivait de la manière suivante :

$$h_\theta(x) = g(\theta^T x)$$

Avec $g(z)$ étant définit comme la fonction sigmoïde.
On a aussi fait remarquer qu'on utilise l'équation de la frontière de décision qui est $(\theta^T x)$ pour effectuer la classification. et de la on déduit que les données de la classe positif vérifierons l'équation $\theta^T x > 1$ et les autres $\theta^T x < -1$.
Ainsi la limite de la classe positif est l'hyperplan d'équation $\theta^T x = 1$ et celle des négatifs est $\theta^T x = -1$ si on tient compte de b comme distance avec l'origine ces droites s'écriront $\theta^T x + b = 1$ et $\theta^T x + b = -1$.

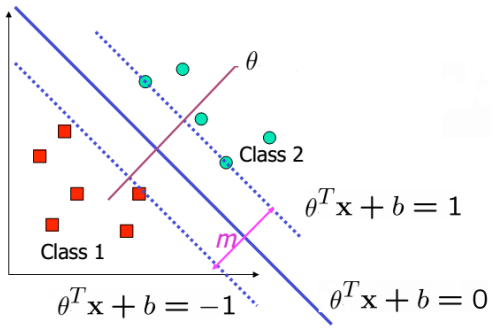


FIGURE 1.14 – Problème de SVM en 2 dimesions

Ainsi le but de SVM serait de trouver l'hyperplan qui maximise la marge ou la distance entre ces 2 hyperplans, et d'après la géométrie analytique cette distance se calcule de la manière suivante $d = \frac{2}{||\theta||}$ Notre problème se formulerait de la manière suivante :

$$\begin{cases} max \frac{2}{||\theta||}, \\ \text{Avec } \theta^T x + b > 1, & \text{si } y{=}1 \\ \text{et } \theta^T x + b < -1, & \text{si } y{=}\text{-}1 \end{cases}$$

Or maximiser $\frac{2}{||\theta||}$ revient simplement à minimiser $\frac{1}{2}\theta$ et nos 2 conditions peuvent se combiner en une seule qui est $y_i(\theta^T x_i + b) \geq 1 \ \forall i$
Ainsi le problème à résoudre devient le suivant :

$$\begin{cases} min \frac{1}{2}||\theta||, \\ y_i(\theta^T x_i + b) \geq 1, \quad \forall i \end{cases}$$

Ceci n'est rien d'autre qu'un problème d'optimisation sous contrainte, pour le résoudre on utilise la méthode de Lagrange .
Elle consiste à cherche le lagrangien de notre problème et égaliser son gradient à zéro.
le lagrangien est donné par :

$$L = \frac{1}{2}\theta^T\theta + \sum_{i=1}^{m} \alpha_i(1 - y_i(\theta^T x_i + b)) \text{ , car } ||\theta||^2 = \theta^T\theta$$

et $\nabla L = 0$
Ce qui nous donnent :
$\frac{\partial L}{\partial \theta} = \theta - \sum_{i=1}^{m} \alpha_i y_i x_i = 0$ ce qui donne $\theta = \sum_{i=1}^{m} \alpha_i y_i x_i$
$\frac{\partial L}{\partial b} = 0$ nous donne $\sum_{i=1}^{m} \alpha_i y_i = 0$
En mettant ces 2 expressions des L on obtient :
$L = \frac{-1}{2} \sum_{i=1}^{m} \sum_{i=j}^{m} \alpha_i \alpha_j y_i y_j X_i^T . X_j + \sum_{i=1}^{m} \alpha i$
,qui n'est qu'une fonction de α_i
Il est connu sous le nom d'un problème de dualité car si on connait α_i on connait θ L doit être maximiser maintenant , ainsi notre problème de dualité s'écrira :

$$\begin{cases} max \sum_{i=1}^{m} \alpha i - \frac{1}{2} \sum_{i=1}^{m} \sum_{i=j}^{m} \alpha_i \alpha_j y_i y_j X_i^T . X_j, \\ \text{Avec } \sum_{i=1}^{m} \alpha_i y_i = 0, \\ \text{et } \alpha_i \geq 0 \end{cases}$$

Avant de s'attaquer à la résolution de ce problème remarquons quelque éléments intéressants qu'il présente :
- la grande partie des valeurs de α_i sont nulles et aux valeurs non nulles correspondent des x_i qu'on appelle support Vector.
- Pour prédire la classe d'un nouveaux élément z nous avons qu'a verifer s'il se place au dessus ou en dessous de notre frontière de décision qui est donnée par $\theta^T z + b = \sum_{i=1}^{m} \alpha_i y_i (X_i^T . Z)$.
- Mais aussi il présente l'avantage que l'algorithme comprend le produit scalaire des tuples d'entrés , cet avantage sera utilisé lorsqu'on va traiter des frontières de décisions non linéairement séparable.

Données non Séparables linéairement : Notion de Kernel [14]

Dans la plupart des cas surtout dans la pratique les données ne peuvent pas toujours être classer avec une droite ou un hyperplan comme on peut le voir à la FIGURE15 :
Une propriété stipule que les données qui ne sont pas linéairement séparable dans un espace de dimension n le seront dans un espace de dimension m avec $m \geq n$ [**8**]. il suffit juste de faire un mapping des tuples x_i dans R^n vers R^m en utilisant une fonction $\phi(x)$. Voyons un exemple à la FIGURE16

Définition : Kernel Un Kernel est une fonction qui permet de calculer le produit scalaire du mapping d'un tuple dans un espace de dimension supérieur que celui dans lequel il est définie . on a :

$$K(x, z) = \phi^T(x) . \phi(z)$$

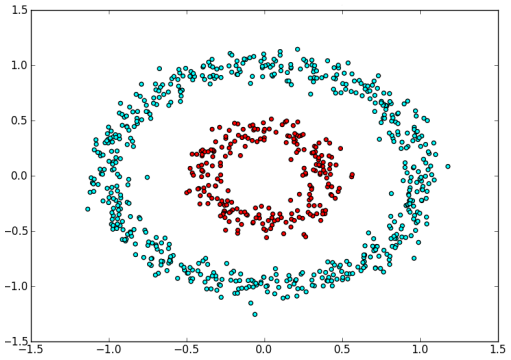


FIGURE 1.15 – Données non séparable linéairement

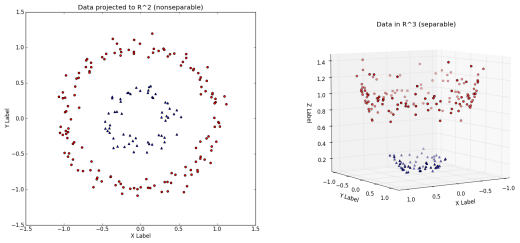


FIGURE 1.16 – Données non séparable linéairement dans R^2 mais qui le devient dans R^3 avec $\phi(x_1, x_2) = (x_1, x_2, x_1^2 + x_2^2)$

Mais on n'est pas obligé de connaitre $\phi(x)$ pour calculer $K(x, z)$ il existe des Kernel bien définie qui permettent de calculer le produit scalaire sans pour autant connaitre ϕ et ainsi faciliter les calcul .
En voici quelque uns les plus utilisé :
- Le Kernel Polynomiale de degré d : $K(x, z) = (x^T z + 1)^d$
- Le Kernel Gaussien : $K(x, z) = \exp(\frac{-||x-z||^2}{2\sigma^2})$
- Le Kernel sigmoïde :$K(x, z) = \tanh(kx^T y + \theta)$

Ainsi dans notre problème d'optimisation on peut juste remplacer le produit scalaire de x_i et x_j et ensuite palier aux problème des frontières de décisions non séparable linéairement.
On choisie un Kernel en fonction des caractéristiques de l'ensemble d'apprentissage .
Le problème sera :

$$\begin{cases} max \sum_{i=1}^{m} \alpha i - \frac{1}{2} \sum_{i=1}^{m} \sum_{i=j}^{m} \alpha_i \alpha_j y_i y_j \phi^T(x).\phi(y), \\ \text{Avec } \sum_{i=1}^{m} \alpha_i y_i = 0, \\ \text{et } \alpha_i \geq 0 \end{cases}$$

Et en introduisant le Kernel on a :

$$\begin{cases} max \sum_{i=1}^{m} \alpha i - \frac{1}{2} \sum_{i=1}^{m} \sum_{i=j}^{m} \alpha_i \alpha_j y_i y_j K(x, y), \\ \text{Avec } \sum_{i=1}^{m} \alpha_i y_i = 0, \\ \text{et } \alpha_i \geq 0 \end{cases}$$

Et ainsi notre frontière de décision sera : $\theta^T z + b = \sum_{i=1}^{m} \alpha_i y_i K(x, y) + b$.
Pour résoudre ce problème on utilise la méthode de SMO ou Sequential Minimal Optimization qui est la technique la plus utilisée pour la résolution de ces genres des problèmes , il est implémenté dans la plupart des package d'apprentissage automatique .

1.2.5 Arbres des Decisons [15] [16]

Définitions

Une arbre de décision est un ensemble des règles de classification et de régression basant leurs décisions sur des tests associé aux attributs ,organisées de manière arborescente .
C'est une représentation d'une procédure d'apprentissage.
Voici quelques termes ou notions liées aux arbres de décisions :
Noeud Principale ou Root Node : c'est l'attribut qui se situe au premier niveau et sur base de celui- ci s'effectue notre prédiction.
Spliting ou segmentation : c'est un processus consistant à diviser un nœud en 2 ou plusieurs sous nœuds sur base d'un test sur un attribue.
Nœud Interne ou Nœud de décision : c'est un nœud étiqueté par un test qui peut être appliqué à toute description d'un individu de la population.
Une feuille : c'est le nœud où on ne peut plus diviser ou segmenter l'arbre , il est étiqueté par une classe.

Construction de l'arbre et Algorithme

Mesure de la pureté des feuilles Pour Construire les nœuds de l'arbre,les choix des « questions les plus discriminantes » peuvent se faire selon plusieurs critères : l'algorithme CART utilise l'indice de Gini, l'algorithme C4.5 utilise l'entropie, l'algorithme Chi-square Automatic Interaction Detector (CHAID) utilise le test de khi carré, et plusieurs autres algorithmes.Dans cette partie nous ne parlerons que de deux premiers. Ces deux outils mathématiques visent à évaluer la « pureté » de chaque feuille : lorsque l'on se situe à un nœud donné de l'arbre, le but est de créer deux feuilles qui soient plus homogènes que le nœud qui les précède. Il faut donc disposer d'un moyen de mesurer cette homogénéité, ou « pureté ». Grâce à cela, à chaque nœud, le split est construit de manière à maximiser le gain d'information apporté par une question donnée sur la connaissance de la variable réponse.

Entropie

L'entropie est une fonction mathématique créée par Claude Shannon en 1948, pour des questions initialement liées à la théorie du signal.
En Machine Learning c'est la mesure du désordre ou de l'inégalité de répartition pour le choix d'un test à une position de l'arbre.
Notons par E notre ensemble d'apprentissage divisé en classes $\omega_1, \omega_2, ..., \omega_k$
- L'entropie de la distribution des classes = quantité moyenne d'information

nécessaire pour identifier la classe d'un exemple de E.

$$H(E) = -\sum_{j=1}^{k} P(\omega_k) \log_2(\omega_k)$$

où $P(\omega_k)$ est la probabilité a priori de la classe ω_k
- Soit un test T(portant sur une variable X) ayant m alternatives possibles qui divisent En sous-ensembles E_j , caractérisé par une entropie $H(E_j)$.
- L'entropie de la partition résultante, c'est-à-dire l'entropie conditionnelle de E étant donné T, est définie comme l'entropie moyenne des sous-ensembles :

$$H(E|T) = \sum_{j=1}^{j} P(E_j)H(E_j)$$

- Le gain d'information apporté par le test Test donc :

$$Gain(E, T) = H(E) - H(E|T)$$

L'algorithme de l'arbre de decisison se base sur ces propriétés.

L'indice de Gini

L'indice (ou coefficient) de Gini est une mesure, comprise entre 0 et 1, de la dispersion d'une distribution. Il est très souvent utilisé en économie ou en sociologie afin de mesurer les inégalités sociales au sein d'un pays. Dans ce contexte, plus le coefficient est proche de 1 et plus la société est inégalitaire. il se calcule de la manière suivante :

$$Gini = \sum_{i \neq j} p(\omega_i)p(\omega_j)$$

Avec $p(\omega_i)$ la probabilité de la classe ω_i .
De la même façon que l'entropie on calcul le gain d'information avec l'indice de gini.

Algorithme

En général, on décide qu'un nœud est terminal lorsque tous les exemples associés à ce nœud, ou du moins la plupart d'entre eux sont dans la même classe, ou encore, s'il n'y a plus d'attributs non utilisés dans la branche correspondante.

En général, on attribue au nœud la classe majoritaire (éventuellement calculée à l'aide d'une fonction de coût lorsque les erreurs de prédiction ne sont pas équivalentes). Lorsque plusieurs classes sont en concurrence, on peut choisir la classe la plus représentée dans l'ensemble de l'échantillon, ou en choisir une au hasard.

Algorithm 3 Pseudo code de l'arbre de décision

Require: Training set E

Initialiser l'arbre courant 'a l'arbre vide ; la racine est le nœud courant
while On n'as pas encore obtenu l'arbre **do**
Décider si le nœud courant est terminale
if le nœud est terminal **then**
lui affecter une classe
else
Sélectionner un test et créer autant de nouveaux nœuds fils qu'il y a de réponses possibles au test
Passer au nœud suivant non exploré s'il en existe

1.2.6 Élagage de l'arbre ou *Prunning*

L'objectif de cette étape est de supprimer les parties de l'arbre qui ne semblent pas performantes pour prédire la classe de nouveaux cas et les remplacées par un nœud terminal (associé à la classe majoritaire).
Le processus est remplacées par un nœud terminal (associé à la classe majoritaire).
il existe différentes façons d'estimer le taux d'erreur entre autre :
– sur base de nouveaux exemples disponibles ;
– via une validation croisée ;
– sur base d'une estimation statistique, ex : borne supérieure d'un intervalle de confiance construit sur un modèle binomial .

1.2.7 Avantages et inconvénients des arbres de décisions

Avantages

- Interprétabilité : chaque élément du modèle est facile à comprendre et à analyser pour un humain, et peut donner de l'information sur les données. Ceci est surtout vrai pour les petits arbres. À cause de l'instabilité des arbres (par exemple si on varie le choix des variables ou des données), il faut utiliser des techniques spécialisées pour determiner quelles variables sont réellement importantes. On peut faire correspondre un arbre de décision à un ensemble de règles SI-ALORS (en introduisant des nouvelles classes correspondant aux noeuds cachés de l'arbre).
- Flexibilité : peut être utilisé sur des données de n'importe quel type (dont évidemment les variables continues et discrètes) pour lesquelles un ensemble fini (pas trop grand) de questions possibles pour la partition peut être défini

(en principe cela peut être appliqué à n'importe quel type de structures de données).

Inconvénients

Un des inconvénients principaux des méthodes d'apprentissage par arbres de décision est leur instabilité. Sur des données réelles, il s'en faut souvent de peu qu'un attribut soit choisi plutôt qu'un autre et le choix d'un attribut-test, surtout s'il est près de la racine, influence grandement le reste de la construction. La conséquence de cette instabilité est que les algorithmes d'apprentissage par arbres de décision ont une variance importante, qui nuit à la qualité de l'apprentissage. Des méthodes comme le Bagging(pour Bootstrap Aggregating) ou les Random Forests(qui consiste à utiliser plusieurs arbres et utiliser le vote classe faite par chaque arbre pour classer une instance) [17] permettent dans une certaine mesure de remédier à ce problème.

Random Forrest

1.2.8 Segmentation Par K-means

La Segmentation est une technique d'apprentissage non supervisé, Les classes des données ne sont pas connus en avance . La segmentation consiste à grouper les données dans des clusters ou segments selon leurs ressemblance , et leurs similarités. L'objectif de la segmentation est la maximisation des distances inter-cluster et la minimisation de la distance intra-cluster . L'unité de mesure de la ressemblance entre les données c'est la distance .
Il existes plusieurs métriques pour définir la distance entre 2 points x_i et y_i :
- La distance euclidienne $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- La distance de Manhattan $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$
- La distance de Minkowski $d(x, y) = \sqrt[q]{\sum_{i=1}^{n} |x_i - y_i|^p}$

Supposons que nous disposons de m points x_i que nous voulons segmenter en k clusters , l'objectif est d'assigner un cluster à chaque point .
La methode K-means consiste à trouver les positions μ_k k = 1,2...K qui minimisent la distance distance du point aux cluster . L'algorithme K-means consiste à résoudre le roblème suivant :

$$\min \sum_{i=1}^{k} \sum_{\underline{X} \in c_i} d(\underline{x}, \mu_k)$$

ou c_i est l'ensemble des points appartenant au cluster i

Algorithme K-means

1.3 Méthodologie de mise en place du processus de data mining[3]

. Au début des années 90, l'intérêt croissant pour le data mining a mis en lumière l'absence d'une méthodologie pour la mise en place d'un processus de découverte de connaissances, applicable quelle que soit l'industrie visée ou l'outil utilisé. De ce besoin est née l'initiative CRISP-DM. A partir du

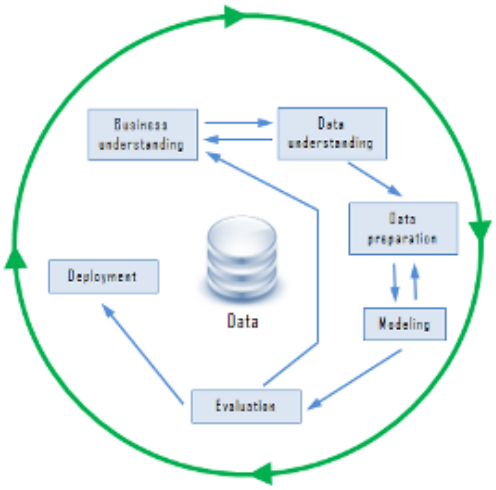


FIGURE 1.17 – Data mining process model

processus de découverte de connaissances utilisé dans les premiers projets de data mining ,CRISP-DM a défini et validé une méthodologie potentiellement applicable dans tous les secteurs de l'industrie. Elle permet de rendre les projets de data mining à grande échelle plus rapides, moins coûteux, plus fiables et surtout améliorer leur gestion. Cette méthodologie ne vise pas que les grands projets car même les petits projets de découverte de connaissances peuvent tirer profit de son utilisation. Essentiellement, cette méthodologie fournit un aperçu du cycle de vie d'un projet de data mining. Elle identifie clairement les principales phases de ce processus au travers de tâches et des relations entre ces tâches. Même si le modèle ne le spécifie pas explicitement, il y a des relations possibles entre toutes les tâches en fonction des objectifs d'analyse et des données qui sont analysées. Les six phases importantes du processus sont :

1. *La compréhension du problème métier* : concerne la définition du problème d'analyse sur la base des objectifs métiers qui en sont à l'origine.
2. *La compréhension des données* :cette phase vise à déterminer précisément les données à analyser et à identifier la qualité des données .

3. *La préparation des données* : couvre les activités liées à la construction de l'ensemble précis des données à analyser à partir des données brutes. Ceci inclue le nettoyage des données, la sélection d'attributs, le choix des observations, etc.
4. *La modélisation* : est la phase consistant dans le paramétrage et le test de différentes techniques de data mining sur les données choisies, dans l'objectif d'optimisation du modèle ou des connaissances obtenues par ces techniques.
5. *L'évaluation* :vise à vérifier le modèle ou les connaissances obtenues afin de s'assurer qu'ils répondent aux objectifs formulés au début du processus. Elle contribue aussi à la décision de déploiement du modèle ou, au contraire, de la nécessité que ce dernier soit reconstruit ou amélioré.
6. *Le déploiement* : est l'étape finale du processus de découverte de connaissances. Son objectif est de mettre la connaissance obtenue par la modélisation dans une forme adaptée et l'intégrer au processus de prise de décision. Le déploiement peut aller, selon les objectifs, de la simple génération d'un rapport décrivant les connaissances obtenue jusqu'à la mise en place d'une application spécifique permettant l'utilisation du modèle obtenu pour la prédiction des valeurs inconnues d'un paramètre d'intérêt.

Chapitre 2

Analyse du domaine de l'orientation [18] [19] [20] [21] [22]

2.1 Généralités sur l'orientation des étudiants

Union Nationale des Associations Familiales (UNAF) considère l'orientation comme un véritable parcours, qui comporte des étapes certes, mais qui doit s'inscrire dans la durée. Cette inscription dans la durée nécessite que des passerelles multiplient entre les différentes filières afin de permettre les réorientations, même en cours d'année. Ces passerelles devraient par ailleurs permettre aux jeunes d'aller aussi loin qu'ils le désirent dans leurs études, quel que soit leur choix initial. Ainsi l'orientation pourra être dédramatisée et aucune formation choisie ne sera une « voie sans issue ». Ce serait aussi un moyen de lutter contre la déscolarisation qui provient souvent d'un choix d'orientation qui ne convient pas et qui pousse les jeunes à arrêter complètement leur scolarité au lieu de se réoriente

En effet, le terme "orientation" recouvre deux activités que la langue anglaise distingue : le processus qui répartit les élèves dans différentes voies de formation, filières et options ("students distribution") ; l'aide aux individus dans le choix de leur avenir scolaire et professionnel ("vocational guidance", "school and career counseling").

Face à la crise, le diplôme protège du chômage et favorise l'accès à la formation continue. Le rapport « Formations et emplois » de l'Insee (2013), fait preuve qu'environ 800 000 élèves sont inscrits en troisième et 700 000 sont candidats au bac. Il s'avère important de savoir si chacun trouvera-t-il la place qui lui convient pour la suite ? Ce n'est pas si facile, car la plupart

des filières imposent une sélection à l'entrée, en fonction du nombre de places disponibles

La difficulté face à un choix d'orientation, légitime à l'adolescence, peut être handicapante par la suite, d'où la nécessité d'une réelle éducation au choix, à l'orientation qui prennent en compte les trois volets : connaissance de soi, connaissance des métiers et de l'environnement professionnel, connaissance des formations. Cependant, au final, faute d'avoir trouvé sa voie ou acquis le niveau nécessaire, un nombre important d'élèves quitte chaque année le système éducatif sans avoir obtenu de qualification.

A ce propos, le rapport haut conseil d'éducation(2008) stipule que chaque année 120 000 jeunes quittent le système de formation initiale sans diplôme (surtout dans la voie professionnelle). A l'université, le décrochage concerne 80 000 jeunes par an (1/4 de l'ensemble des sortants de l'université d'une année). Les services qui s'occupent de l'orientation sont éparpillés, il existe un accès inégal à ces services, les ressources humaines et matérielles limitées influent sur la qualité des services offerts Pourtant, l'objectif des universités vise à offrir une bonne démarche d'orientation aux étudiants. Elles ont comme défi d'élaborer pour chaque étudiant une démarche d'orientation stimulante, cohérente, réaliste et adaptée à sa dynamique personnelle, d'établir une véritable culture d'orientation en impliquant, d'une part, tout le personnel de l'université, les amis et la famille de l'élève ainsi que les partenaires de la communauté et, d'autre part, en unissant les forces du réseau d'intervenants qui soutiennent les étudiants sur le plan personnel et professionnel pour faciliter leur insertion dans notre société

L'orientation au collège et au lycée est déterminée par les notes qui répartissent les élèves de manière hiérarchisée, et non sur les appétences et potentialités personnelles. (Au collège plus de 4 élèves sur 10 estiment donc que leur orientation est subie plus que choisie), elle peine à encourager les jeunes issus des milieux les moins favorisés, elle est trop irréversible c'est-à-dire absence de passerelles et de réorientation entre voie professionnelle et voie technologique et générale mais aussi elle est conditionnée par l'offre de formation de proximité et cette offre de formation est mal répartie sur le territoire. On constate également la mobilité des élèves qui est faible, d'où l'offre de formation ne s'adapte que lentement aux nécessités économiques ce qui conduisent les jeunes à s'engager parfois dans des filières sans perspective.

Ainsi, actuellement en République Démocratique du Congo (RDC), aucune structure d'orientation scolaire et professionnelle n'existe et le pouvoir public n'y pense même pas. Avec la prolifération des Instituts Supérieurs et des Universités publics ou privés, créées anarchiquement par le pouvoir public depuis des décennies, les élèves finalistes de l'enseignement secondaire s'orientent en considérant, en fin de compte, l'implantation géographique de ces établissements. Le système de prime qui a été instauré depuis les années 1992 complique encore davantage le système d'orientation des élèves. Les enfants issus des catégories sociales défavorisées embrassent, pour la plupart, les études dans les institutions publiques qui coûtent moins cher comparativement aux institutions privées. De ce fait, beaucoup d'élèves effectuent le choix des études aléatoirement sans tenir compte des aptitudes intellectuelles ou des aspirations professionnelles futures. Les jeunes sont donc abandonnés à leur triste sort. Cette situation étant considérée comme étant normale, personne n'en fait nullement allusion.

2.2 Présentation du service de l'orientation des étudiants

Qui intervient dans le processus d'orientation ?

L'élève en quête d'orientation n'est pas laissé à lui-même. Outre sa famille, il peut requérir l'aide de ses parents, de ses enseignants, du directeur d'établissement et des conseillers.

1. ***les parents :***
   Parmi les acteurs de l'orientation, les parents sont bien entendu les premiers concernés. Si le rôle des parents est irremplaçable, c'est avant tout parce qu'ils sont probablement seuls capable d'aimer l'enfant de manière inconditionnelle, indépendamment de ses résultats scolaires et autres performances. Ils sont les seuls à continuer à croire en l'enfant, lorsque les autres désespèrent. Et l'une des fonctions parentales est justement de construire l'estime de soi de leur enfant. Ils interviennent pour trois objectifs qui sont : - améliorer l'orientation afin qu'elle soit choisie plutôt que subie - renforcer les relations école-famille - prévenir le décrochage scolaire Les jeunes confrontés aux choix d'orientation ont besoin d'une écoute attentive de la part de leurs parents. Les parents ne sont pas forcément aux faits des filières de formation et ne connaissent pas tous les métiers mais ils peuvent être au jour le jour un soutien pour leur enfant qui s'interroge.

2. ***Les enseignants***
   Dont l'action même d'enseigner est partie prenante de l'orientation et qui doivent aussi y exercer une mission précise. Quoiqu'i en soit, en matière d'orientation, les enseignants exercent une influence déterminante par la seule mise œuvre de leurs démarches pédagogiques. Ils disposent, par les outils et les méthodes qu'ils utilisent ainsi que par leur comportement, des moyens d'agir sur certains éléments constitutifs de la maturation des choix de l'élève, particulièrement sur l'image qu'il se forge de lui. Les enseignants aident l'élève à prendre conscience de ses potentialités et guide ses choix d'orientation.
3. ***Le chef d'établissement***
   Le chef d'établissement joue aussi un rôle fondamental dans la préparation de ses élèves à l'orientation. Il est responsable du dispositif d'orientation mis en place dans son établissement. C'est sous sa responsabilité que le programme d'information et d'orientation s'élabore et s'applique. Il est le garant du respect des réglementations en vigueur, mais au-delà il imprime selon ses convictions son expérience et les caractéristiques des populations dont il a la charge, le mouvement qui induit généralement à une véritable politique de l'information et de l'orientation. Il a un rôle particulièrement important dans la mise en œuvre des procédures d'orientations, dans le déroulement du dialogue avec les parents et dans le suivi des décisions.
4. ***Le conseiller d'orientation***
   Les tâches du conseiller d'orientation dans une école s'avèrent multiples et difficiles à accomplir par lui seul car c'est un travail qui requiert la collaboration de différents acteurs comme nous l'avons signalé ci-dessus. L'une de ses missions consiste à fournit les informations. Le processus de cette action d'information est de : - Mettre l'élève en état de réceptivité ; - Proposer des informations en rapport avec les intérêts connus ; - Adapter le message au niveau de l'élève et pour cela, il faut donner des propositions supplémentaires et faire des synthèses nécessaires avec la participation de l'élève. Le but à atteindre est de mettre l'élève en état de réaliser une self-orientation. De plus, le conseiller a le devoir de fournir aux parents des informations nécessaires pour qu'ils puissent assurer leurs responsabilités vis-à-vis de leurs enfants.

2.3 Etude des procédures d'orientation

L'orientation et les formations proposées aux élèves tiennent compte de leurs aspirations, de leurs aptitudes et des perspectives professionnelles liées

aux besoins prévisibles de la société, de l'économie et de l'aménagement du territoire.
Dans ce cadre, les élèves ou les étudiants élaborent leur projet d'orientation scolaire et professionnelle avec l'aide des parents, des enseignants, des personnels d'orientation et des autres professionnels compétents, etc. Le chef d'établissement, décideur final, en l'état actuel du système, c'est lui qui prend la décision finale d'orientation, sur proposition des membres du conseil de classe.

- **Un conseil personnalisé : ,**
  Cette méthode s'appuie sur l'échange et le dialogue et permet de déceler des potentiels sans intervenir de manière directive dans les réponses et les choix du candidat. A l'aide de tests validés par des psychologues d'une part, et d'entretiens non directifs d'autre part, le conseiller définit des profils de métiers et d'études en adéquation avec les motivations, les aptitudes et la capacité de travail. Il tient compte des débouchés réels offerts par le marché du travail. Cette méthode permet de s'adapter et de respecter la personnalité de chacun en construisant ensemble un projet d'orientation.
- **Le bilan d'orientation : ,**

  Destiné à tous les élèves à partir de la 3ème, ainsi qu'aux étudiants en recherche de réorientation, ce bilan permet de prendre les bonnes décisions d'orientation, en toute connaissance de cause. Le bilan d'orientation se déroule en 3 phases : Phase 1 : Exploration Ce premier rendez-vous permet de recueillir vos motivations et vos aspirations et d'évaluer vos aptitudes, etc. Cette phase s'appuie sur : des entretiens non directifs l'évaluation des aptitudes de l'étudiant ; et un test concernant ses motivations, sa personnalité, ses centres d'intérêt ...
  Phase 2 : Analyse et Synthèse Le conseiller effectue une synthèse des résultats obtenus à l'aide d'un logiciel développé au Canada et utilisé depuis plus de 30 ans, etc. Ce logiciel a reçu le label Reconnu d'Intérêt Pédagogique (RIP) du ministère de l'Education Nationale, il est remis à jour chaque année. Il bâtit ensuite un projet d'orientation cohérent.
  Phase 3 : Restitution des résultats Lors du deuxième rendez-vous, le conseiller propose à l'étudiant une sélection de métiers et de formations. Un rapport écrit contenant des fiches métiers (descriptif, niveau de formation à atteindre, perspectives d'emploi ...) et des informations sur les formations à suivre vous est remis.

2.4 Etudes des documents utilisées

- Dossiers scolaires et psychologiques
  Les données de toute étude pédagogique et psychologique en orientation sont contenues dans le dossier scolaire et le dossier psychologique établis pour chaque élève à un moment de sa scolarité, puis complétés au cours de ses études à l'aide des informations scolaires, familiales, psychologiques, qui permettront au conseiller psychologue de suivre son évolution et de donner un conseil d'orientation au moment du choix scolaire ou professionnel. Son but est de :
  - Connaitre l'élève au cours de son évolution ;
  - Etudier les facteurs de réussite ou d'échec qui sont apparus au cours de cette évolution ;
  - En rechercher les causes et trouver les remèdes s'il le faut Son contenu est constitué par les données quantitatives et qualitatives.
    1. Les données quantitatives : sont des résultats obtenus :
       - D'une part aux examens écrits ou oraux.
       - D'autre part aux tests d'aptitudes et de connaissances.
    2. Les données qualitatives :
       sont recueillies par les tests de personnalité et les entretiens. Elles sont complétées par les comptes rendus des conseils de classe, de délibérations....
- Analyse du dossier psychologique
  Afin d'obtenir la vue la plus objective possible de l'enfant, les différents éléments des dossiers sont analysés de façon à permettre des confrontations et de recoupements
  1. les données physiologiques :
     vision, audition, motricité, fatigabilité, déficience, etc. sont obtenues grâce à l'étude :
     - De la fiche médicale,
     - Des questionnaires (élèves-parents-enseignants),
     - Des examens psychologiques.
  2. les données mentales :
     efficience, mémoire (formes), intelligence (niveau, attitudes, facteurs spécifiques..,), aptitudes particulières (scientifiques, littéraires, artistiques...) sont apportées par :

— La fiche scolaire,

— Les questionnaires (élèves-parents),

— L'examen psychologique.

3. les données scolaires (niveau d'acquisition scolaire...) sont fournies par :

— La fiche scolaire ou bulletin,

— Les tests de connaissances scolaires.

4. le données caractérielles :
activité ou passivité, sens de l'effort, émotivité, affectivité, attitude devant le travail, attitude à la maison, à l'école, attitude devant la réussite, devant l'échec, attitude envers les parents, enseignants, élèves...sont apportées par :

— Les questionnaires,

— Les entretiens,

— Les épreuves psychologiques particulières (les tests projectifs).

Notons également que quand le dialogue entre l'équipe éducative et la famille n'a pas permis d'aboutir à une décision commune, les parents peuvent engager une procédure d'appel regroupant la commission d'appel. Cette dernière est présidée par le directeur des services académiques, ayant l'objectif de réexaminer la décision d'orientation en fonction des notes de l'élève, de ses capacités, de ses difficultés, de ses projets. Son point de vue est extérieur. Ainsi, les membres de la commission (chefs d'établissement, professeurs, conseillers d'orientation...) ne sont pas directement impliqués dans l'histoire scolaire de l'élève.

2.5 L'orientation à l' Université Libre des Pays des Grands Lacs (ULPGL) /GOMA [23]

Abordant la question de l'orientation à l'université libre des pays des grands lacs, l'orientation est assurée par l'apparitorat central et la direction de scolarité.

Dès son arrivé à l'université, on présente au candidat un papier sur lequel toutes les facultés sont mentionnées ainsi que les orientations, mais en plus les exigences pour chaque faculté. A part cela, l'étudiant doit écrire une lettre

manuscrite, et doit compléter une fiche de demande d'inscription, et déposer son dossier contenant : son diplôme d'Etat ou son équivalent, une attestation de naissance, une attestation de bonne vie et mœurs, une attestation de célibat, une attestation de nationalité, une attestation de résidence (pour les étrangers) et un certificat d'aptitude physique. Après avoir fournis tous ces éléments, la direction de scolarité va siéger et analyser le dossier du candidat avant de le soumettre à un examen écrit. Un examen est organisé pour les étudiants ayant obtenu moins de soixante pourcent. Ce concours est organisé en deux moments. Le premier concours se déroule au mois de septembre et le deuxième concours avant la rentrée académique pour permettre à ceux qui ont raté le premier examen de se rattraper et être éligible. Les résultats obtenus à ce concours sont affichés à la valve et la commission chargé des inscriptions, dirigé par le secrétaire général académique décide si le récipiendaire est retenu, rejette ou réorienté dans une autre faculté compte tenu de ses résultats obtenus.

Concernant l'orientation proprement dite, la direction de scolarité explique aux étudiants comment fonctionne chaque faculté de l'université, les exigences, les débouchés sur le milieu professionnel et chaque étudiant choisi sa faculté compte tenu de ses choix, ses gouts, ses aspirations, ses motivations, ses objectifs et sa vision sans pour autant le contraindre. Cependant la contrainte intervient au cas où, l'étudiant a choisi une faculté qui ne correspond pas à sa capacité intellectuelle, mais en plus si l'étudiant échoue ou produit un résultat moins satisfaisant à la fin d'une année académique, la direction de scolarité décide de le mettre en observation, après les examens du premier semestre ils peuvent décider de le réorienter dans une autre faculté où il peut s'en sortir mieux. Si cette orientation ne produit pas des résultats satisfaisants la commission décide du refus, ou du renvoi de l'étudiant de l'université car il est INapte A Poursuivre les études Supérieurs (INAPS) ou il a épuisé toutes les possibilités qui lui ont étés offertes.

Chapitre 3

Présentation et Exploitation des Données Obtenues

Dans ce chapitre nous allons exploités les données mise en notre disposition par les autorités de l'ULPGL. Celles-ci sont issues du système d'information UAT et pour des raisons de confidentialité nous n'avons pas eu accès a toute la base des données nous avons juste fais une requête des donnes dont nous avons besoin pour notre étude et l'administrateur a exécuté une requete vers sa base des données et nous a fourni les données dont nous avions besoin pour l'étude sous forme d'un fichier Commat Separted Values (CSV). Comme souligné dans le chapitre premier ce chapitre se basera sur la méthodologie CRISP-DM elle sera subdivisé en différentes sections :

— L'exploration et la préparation des données

— Construction du modèle de Prédiction

— L'amélioration du modèle [24]

3.1 Exploration et la préparation des donnés

Les spécialités affirment que 70-80 % du temps consacré à un projet DataMining est alloué à la phase de l'exploration et la préparation des données [25] , il n' ya pas des raccourcis pour cette phase et si on l'a pas bien effectué nous risquons de nous retrouver entrain d'améliorer l'exactitude de notre algorithme mais en vain nous serons toujours obligées de retourner à cette phase et toutes ces techniques de l'exploration des données pourrons nous venir en aide .
Les Étapes de la phase d'exploration et la préparation des donnes sont mentionnées sur la figure suivante :

En bref l'exploration des données consiste à se plonger dans le passée pour prédire l'avenir .Souvenons nous que la qualité de notre entré détermine la qualité de notre sortie, ces phases nous permettent d'améliorer la qualité de notre entré en vue d'avoir une bonne sortie. Voici les étapes de cette phases :

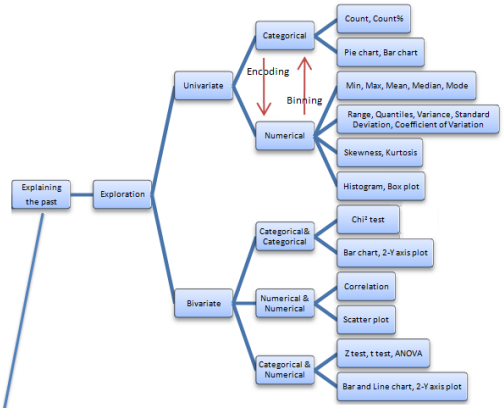


FIGURE 3.1 – Étapes de la phase d'exploration des données

— Identification des variables
— Statistique Descriptive
— Analyse Bi-varié
— Traitement des valeur manquantes
— Traitement des déviations ou valeurs aberrantes (*Outliers*)
— Transformation des variables
— Création des nouvelles variables

Comme nous l'avons soulignées dans le chapitre 1 ce processus est un processus itératif et incrémentai nous exécuterons cette phase 2 a 5 fois ou plus en vue d'avoir un bon modèle.

3.1.1 Identification des Données et des Variables

Comme soulignées dans la phase d'introduction les données mise à notre disposition sont sous format CSV et nous allons utilisé la librairie pandas de python pour faire l'analyse , nous utiliserons aussi d'autres libraires qui nous permettrons de faire les statistiques ainsi que les visualisations : Nous remarquons que les données sont stocké dans un e structure de type matricielle appelé Dataframe. [26] Un DataFrame selon la documentation officielle de

pandas est une structure des données bidimensionnel avec des colonnes des données des différentes types . Il peut être comparée à une feuille de calcul Excel ou une table dans Structural Query Language (SQL). Notre ensemble d'apprentissage de départ est une matrice de 9606 lignes et 22 colonnes .
Chaque ligne comprend les informations d'un étudiant pour une année académique et voici la description des nos colonnes.

1. IDENTIFICATION : contient une identification unique et anonyme d'un étudiant les noms et les matricules réels des étudiants on éé cachées pour des raisons de confidentialités
2. BIRTHDAY : contient la date de naissance de chaque étudiant
3. NAME : contient le sexe de chaque étudiant
4. DIPLOMTYPE : le type de diplôme
5. DIPLOMMENTION : mention de diplôme
6. DIPLOMPERCENTAGE : le pourcentage au diplôme
7. DIPLOMSECTION : la section du diplôme
8. DIPLOMOPTION : l'option
9. DIPLOMPLACE : l'endroit d'obtention du diplôme
10. SCHOOL : l'école de provenance
11. SCHOOLPROVINCE : la province de provenance
12. SCHOOLCODE : code de l'école
13. SCHOOLSTATUS : le statuts de l'école (privée , publique , conventionné ,..)
14. ACADYEAR : l'année académique
15. PERC1 : pourcentage en première session
16. MENT1 : mention en première session
17. PERC2 : pourcentage en seconde session
18. MENT2 : mention en seconde session
19. FAC : la faculté de l'étudiant
20. OPT : l'option choisie par l'étudiant
21. PROM : la promotion de l'étudiant

Comme nous pouvons le constater les colonnes 1-13 regorgent les informations que chaque étudiant donné à son inscription , ils constituerons nos variables d'entres les restes seront utilisées pour constituer notre variable de sortie.

Nous l'avons aussi signales que chaque ligne comprend les information d'un étudiant pour une année académique . Pour mener bien notre analyse nous allons grouper les information de chaque étudiant en une ligne

nos données seront groupé selon les variables d'entrées ensuite les donnés de sorties seront groupes selon une fonction d'agrégation prédéfinie. Nous allons premièrement faire une analyse uni-varié sur les données en entrées ! Soulignons que nous avons décidé de supprimer certaines colonnes n'ayant pas des informations importantes car contenant plus de 90 % des valeurs manquantes il s'agit entre autres des colonnes suivantes : 'DIPLOMDATE','DIPLOMMENTION','DIPLOMPLACE','SCHOOLCODE'.

Après suppression les colonnes en entrée deviennent les 10 premiers colonnes.

Pour une première approche nous allons grouper notre ensemble en fonction des données en entré et ensuite écrire une fonction qui va grouper les donnes de sortie la fonction de qui regroupe les données en sortie ne fait que grouper les mention , et pourcentage pour une année académique d'un étudiant dans une liste.
Après groupement en fonction des matricules nous venons de remarquer que notre ensemble comprend 4715 lignes et 18 colonnes et c'est sera notre ensemble pour notre étude ,cet ensemble est subdivisé en variables d'entré et variables de sortie !
Voici un aperçu des nos données en entrée ainsi que les données en sorties au tableau la figure la figure 3.1

Nous venons de finir avec la présentation des nos données nous allons maintenant débuter avec la phase d'analyse promptement dite des donnés que nous avons en entré et ensuite nous effectuerons une analyse des données en sortie en enfin analyse les donnes des sortie combinées à celles des données en entrée

3.1.2 Préparation des données

Cette phase consistera aux traitement des valeurs manquantes et valeurs aberrantes mais aussi au nettoyage des données mal orthographiées lors de la saisie des données pour certaines attribues. Commençons par Expliquer les raisons de la présence des valeur manquantes , ainsi que les données aberrantes et la manière de les traiter

Malgré la quantité croissante de données, les problématiques de données manquantes et des valeurs extrêmes restent très répandues dans les problèmes statistiques et nécessitent une approche particulière. Ignorer les données manquantes et les valeurs extrêmes peut entraîner, outre une perte de précision,

TABLE 3.1 – Bref aperçu de notre ensemble d'apprentissage

	SCHOOLSTATUS	SCHOOL_RIGHT	OPTION_RIGHT	SCHOOLPROVINCE	DIPLOMPERCENTAGE
45	protestant	zanner	commmerciale et adm	NORD-KIVU	61,000000
215	publique	chemchem	pedagogie	MANIEMA	51,000000
343	catholique	kambali	bio-chimie	NORD-KIVU	62,000000
356	catholique	mwanga/ uvira	latin philo	SUD-KIVU	51,000000
429	protestant	maendeleo	inconnu	NORD-KIVU	56,876522
474	protestant	maendeleo	latin philo	NORD-KIVU	56,000000
644	publique	ngoma	math-physique	NORD-KIVU	68,000000
645	publique	ngoma	pedagogie	NORD-KIVU	59,000000

	ID	ACADYEAR	PERC1	MENT1	PERC2	MENT2	FAC	PROM
0	45	[2013-2014]	[nan]	[AA]	[nan]	[nan]	FPSE	[L2]
1	215	[2012-2013]	[nan]	[ADM]	[63.0999984741]	[S]	FD	[L2]
2	343	[2015-2016]	[nan]	[AA]	[52.2000007629]	[A]	FSEG	[G2]
3	356	[2015-2016]	[nan]	[ADSTM]	[59.9000015259]	[S]	FSEG	[L2]
4	429	[2013-2014]	[nan]	[AA]	[nan]	[A]	FD	[G1]
5	474	[2014-2015, 2015-2016]	[nan, 62.5]	[AA, S]	[nan, nan]	[nan, nan]	FD	[G3, G3]
6	644	[2013-2014, 2014-2015]	[60.4000015259, 68.0]	[S, S]	[nan, nan]	[nan, nan]	FD	[L1, L2]
7	645	[2014-2015, 2015-2016]	[nan, nan]	[AA, AA]	[61.4000015259, nan]	[S, NAF]	FD	[L1, L2]

de forts biais dans les modèles d'analyse et comme signaler dans l'introduction de ce chapitre peuvent augmenter l'erreur de prédiction.

Valeurs Manquantes

Cause et types des Données Manquantes [27] Les Données Manquantes (DM) ont de multiples causes. Il peut être impossible de contacter une personne sélectionnée pour faire partie d'une enquête (non-réponse totale) ou un répondant peut refuser de répondre à une ou plusieurs questions (non-réponse partielle). Une mauvaise saisie de l'information peut également générer des DM. Finalement,des DM peuvent aussi être causées par l'existence de données aberrantes qui doivent être supprimées avant d'effectuer des analyses. Selon les cause on classe les données manquantes selon différentes types il existe plusieurs types de données manquantes, ils peuvent être introduit lors de :

1. l'extraction des données : il peut être possible qu'on ait des problèmes lors de l'extraction des données dans ce cas il es préférable de vérifier les données avec les donateurs c'est comme par exemple pour notre étude dans une première approche nous n'avons pas réussie les données des étudiants ayant passer en première session certaines fonctions peuvent aussi créer des données manquantes comme c'est le cas de notre fonction qui calcule les dates ' ces genres des données peuvent facilement être détecter.
2. la collecte des données ; c'est le cas le plus courant et il est plus facile de le détecter.

Dans ce cas la classification la plus couramment utilisée ayant été proposée par Little et Rubin [27] :

— Missing completely at random (MCAR) (complètement aléatoire)
— Missing at random" (MAR) (aléatoire)
— Missing not at random" (MNAR) (non aléatoire)

Les DM sont MCAR lorsque la probabilité de non réponse pour une variable ne dépend pas de celle-ci, mais uniquement de paramètres extérieurs, indépendants de cette variable. Cela veut dire qu'il n'est pas possible de définir un profil des individus ayant des DM et que la probabilité des DM est uniforme. De manière générale, ce type de DM est très rare.

Les DM sont dites MAR lorsque la probabilité de non-réponse peut dépendre des observations mais pas des DM, par exemple s'il existe une différence de non-réponse entre les hommes et les femmes concernant la question

du revenu, mais que parmi les hommes entre eux ou parmi les femmes entre-elles, la probabilité d'avoir des non-réponses est identique quel que soit le niveau du revenu.

Finalement, les DM sont de type MNAR lorsque la probabilité de non-réponse est liée aux valeurs prises par la variable ayant des DM.comme par exemple pour notre cas les étudiant n'ayant pas passé tous leurs examens à une session n'ont pas des pourcentage à cette session et on pour mention AA assimiléé aux ajournées .

Méthodes d'imputations Il existe 8 à 9 méthodes de traitement des données manquantes largement répandues à l'heure actuelle, y compris des méthodes connues pour être peu performantes mais cependant toujours utilisées.

1. Analyse des cas complets (CC)
2. Imputation par la moyenne (MEAN) : c'est la méthode que nous utiliserons par défaut
3. Imputation par la médiane (MED)
4. Imputation par régression simple (REG)
5. Imputation multiple par Markov Chain Monte-Carlo (MCMC)
6. Imputation par le plus proche voisin (KNN)
7. Imputation multiple par un algorithme basé sur le bootstrap, approchant des résultats de l'algorithme EM (EM)
8. Imputation multiple par "Predictive Mean Matching" (PMM)

Toutes ces techniques existe sont implémentées dans les librairies que nous utilisons.
Dans ce travail selon le cas nous allons faire une imputation par la moyenne ou par analyse des cas complets mais soulignons qu'il faut bien réfléchir avant d'utiliser une imputation par la moyenne car il a une mauvaise influence sur la variance

Valeurs aberrantes ou extrêmes ou outliers

Une valeur aberrante est une valeur qui diffère de façon significative de la tendance globale des autres observations quand on observe un ensemble de données ayant des caractéristiques communes. Par exemple dans l'analyse des pourcentage du diplôme à l'EXETAT nous avons trouvé des diplôme qui ne sont pas dans l'intervalle 0 à 50 % des étudiant avec des diplômes de 634% ou des diplômes de 0%

Voici quelque remarques à considérer pour les valeurs aberrantes :

1. Les valeurs aberrantes ne sont pas forcément erronées. Dans certains cas, la valeur aberrante doit être acceptée comme une indication intéressante. par exemple après analyse on trouve une étudiant âge de 60 ans !
2. Il ne faut pas adopter une attitude radicale de rejet, ou d'inclusion systématique des valeurs aberrantes. Le rejet systématique peut entraîner la perte d'informations réelles Le rejet des valeurs aberrantes a des conséquences statistiques non négligeables car l'analyse est ensuite faite sur un échantillon censuré qui n'est plus aléatoire.

En fonction des circonstances, il existe des méthodes, dites robustes, qui prennent en compte toutes les données mais minimisent l'influence des valeurs aberrantes. L'apparition de valeurs aberrantes est due à diverses sources de natures différentes, d'où la complexité de l'examen des valeurs aberrantes.

Pour détecter les valeurs manquantes nous avons utiliser les techniques suivantes :

a) Contrôle sur le domaine des valeurs : Exemple : Pour la variable « DIPLOMPERCENTAGE », une borne maximale (100 %) est connue et la valeur minimale est de 50 . Les valeurs supérieures à 100 et inférieur à 50 sont considérés comme aberrantes.

b) Détection graphique : Pour détecter la présence de valeurs aberrantes On a utilisé :

— Boxplot

— diagramme de dispersion des observations classées en fonction de leur rang

Traitement des valeurs aberrantes :

1. Les valeurs aberrantes pouvant provenir d'erreurs de saisie, on le traite séparément en étudiant cas par cas.c'est cette technique que nous allons utilisé pour certaines valeur du pourcentage de diplôme.
2. On les rejette et on applique ensuite une des méthodes d'imputation (moyenne,médiane. . .) vues pour les valeurs manquantes.
3. On adopte des méthodes qui diminuent leur impact au cours des analyses statistiques :la médiane

Voici Un aperçu des nos colonnes avec données manquantes et données aberrantes ce tableau décrit.

Ce tableau décrit toutes les informations possibles sur les données continues et de prime à bord nous sommes à mesure de constater certaines incohérences sur les diplôme pourcentage qui on un maximum de 6053 et un

TABLE 3.2 – Statistiques Des Données avec Valeurs manquantes et abberantes

	IDENTIFICATION	BIRTHDAY	DIPLOMPERCENTAGE
count	4038.000000	4038.000000	4038.000000
mean	8792.137692	26.072022	57.749876
std	2333.542658	3.957584	95.088746
min	215.000000	20.000000	0.000000
25perc	7310.250000	24.000000	52.000000
50perc	9181.500000	25.000000	55.000000
75perc	10540.750000	27.000000	60.000000
max	12360.000000	58.000000	6053.000000

minimum de 0 qui est vraiment impossible car le diplôme en RDC dois être compris entre 50 et 100 % !Nous allons visualisé ces incohérence de plus prêt avec des box-plots.

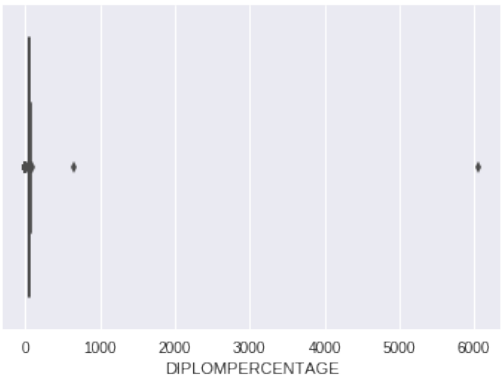


FIGURE 3.2 – Box Plot Attribue Diplôme Pourcentage

Au vu de ces courbes nous remarquons que l'attribue diplôme pourcentage dispose de beaucoup des déviations. Nous avons pu corriger au cas par cas en remplaçant les valeurs aberrantes par leurs valeurs exactes et d'autres par la moyenne comme il n'était pas nombreux

Données mal orthographiées

Dans notre première analyse nous avons remarqué que certaines attribues ont des valeurs très désorganisées et vraiment dispersé et ce qui a une mauvaise influence sur le calcul de l'entropie et ainsi sur les algorithmes du

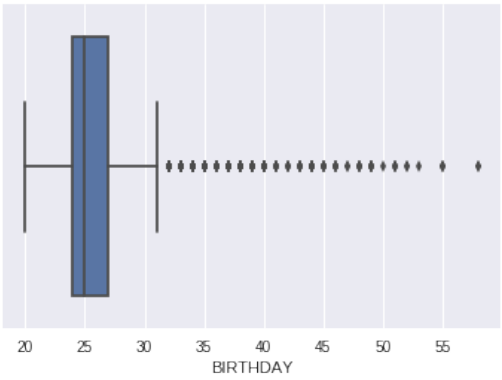


FIGURE 3.3 – Box Plot Attribue Age

Machine Learning . Nous pouvons aisément constater que ce problèmes est du à des fautes d'orthographes commise lors de la phase de saisie des données et ainsi pour continues nous devons essayer de corriger ces erreurs et bien organisé les données .

Par exemple pour la variable diplôme section on a pu voir les valeurs suivantes : 'TECSC', 'Techniqe', 'TECHN IQUE',technique','TCH' qui sont saisie pour la même et unique section 'techniques' mais avec différentes erreurs d'orthographe.
Pour l'attribut SCHOOL nous avons obtenus les valeurs suivantes : 'matanoia', 'metanoia', 'metonoina', '%etanoia', 'meta' pour la seule école 'metanoia' Nous avons aussi remarqué ces genres d'incohérence pour les autres attribues catégorielles. Après analyse nous avons pu détecter qu'il étaient du à des erreurs d'orthographe. Ces genres d'erreur de notation ont pour conséquence le fait qu'il font augmenter l'entropie de nos colonnes et ainsi pénalisent nos algorithmes surtout lorsqu'on travaille avec les arbres de décisions .Nous avons procédé à un nettoyage automatique qui a consisté en un groupement des valeurs proches en utilisant la distance de Leveinstein :[28]combinée au clustering par l'algorithme d'affinity propagation,et ainsi qu'un nettoyage manuelle pour arranger les données à la fin de cette phase nous avons obtenus des données moyennement propres et bien nettoyer avec un entropie faible.

3.1.3 Analyse des données

Cette phase comprend une analyse statistique bi-varié et uni-varié nous visualiserons les résultat à l'aide des graphiques . Dans cette partie nous utiliserons beaucoup plus la statistique descriptives et la statistique inférentielle Comme nous pouvons le remarquer notre ensemble d'apprentissage comprend

à la fois des données numériques (continues) ainsi que des données discrètes catégories. voici comment nous allons procéder

Analyse Uni-Varié

Dans cette partie nous allons effectuer les statistiques descriptives pour chaque variable .

1. ***Variable Numériques ou continues :*** Pour les données continues nous allons essayer de comprendre la tendance et la dispersion des nos variables .les métriques utilisées sont sur la figure suivante : En bref

Central Tendency	Measure of Dispersion	Visualization Methods
Mean	Range	Histogram
Median	Quartile	Box Plot
Mode	IQR	
Min	Variance	
Max	Standard Deviation	
	Skewness and Kurtosis	

FIGURE 3.4 – Techniques d'exploration des données en entré

   nous allons examiner le moyenne , le mode , l'écart-type et la variance , nous conterons aussi les variables nous allons faire les visualisations avec des box-plot ! cette étape nous sera aussi utile dans le traitement des valeur manquantes et des valeurs aberrantes !

2. ***Variable catégorielle ou quantitative :*** Pour les données discrètes nous allons utiliser les tables des fréquences pour comprendre la distribution de chaque catégorie nous pour aussi voir le pourcentage de chaque catégorie , les histogrammes et bar char seront utilisées.

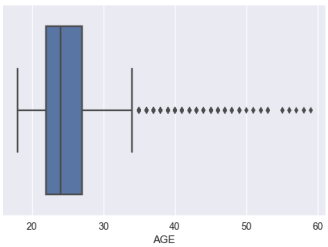
Commençons par l'analyse des attributs numériques age et pourcentage à l' EXamen d'ETAT (EXETAT)

L'AGE et Le Diplôme Pourcentage Voici un bref aperçu des statistiques descriptives de ses variables à la la figure la figure 3.3

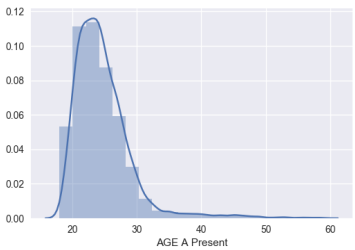
AGE Nous avons calculer l'age en se basant sur la date de naissance et la date d'aujourd'hui.signalons que cet attribue disposait des valeurs manquantes au début et on était imputer par la moyenne , comme nous pouvons le voir à la la figure la figure 3.3 nos étudiant disposent en moyenne d'un age de 24 ans avec une variance de 4,5 le moins âgé a 18 ans et le plus âgé a 59 ans. On peut aisément que l'age suit une distribution normale .

TABLE 3.3 – Statistiques Des Données

	DIPLOMPERCENTAGE	AGE
count	4715,000000	4715,000000
mean	56,878914	24,732768
std	5,756663	4,621602
min	50,000000	18,000000
25%	52,000000	22,000000
50%	56,000000	24,000000
75%	60,000000	27,000000
max	86,000000	59,000000



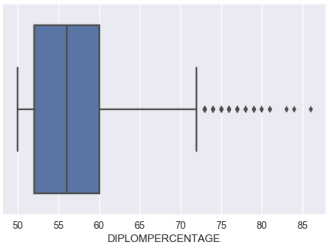
(a) Box-Plo det l'age



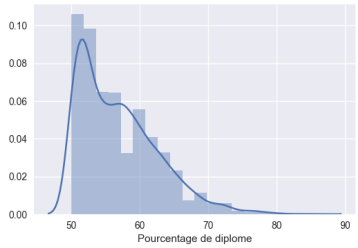
(b) distribution de l'age

FIGURE 3.5 – Graphiques de l'attribue age

Pourcentage du diplôme Après nettoyage et contrairement à la la figure la figure 3.2 on a constater qu'après nettoyage la moyenne est de 56,8 % avec unn écart type de 5,7 le minimum de 50 % et un maximum de 86% . les graphiques représentent les informations sur l'age son à la la figure la figure 3.6a et la figure la figure **??**



(a) Box-Plot du Pourcentage du diplome



(b) distribution du Pourcentage

FIGURE 3.6 – Graphiques de l'attribue du Pourcentage à l'examen d'état

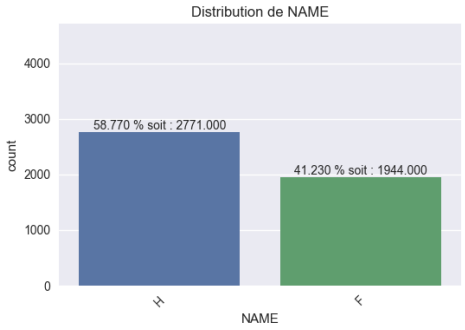


FIGURE 3.7 – Distribution du sexe des étudiants

Le Sexe des étudiants La la figure la figure 3.7 nous donne la répartition de sexes dans notre ensemble d'apprentissage on peut aisément constater qu'il n'est pas si déséquilibre que ça !, le genre est vraiment respecté avec 41% dés nouveaux étudiant étant de sexe féminin.

le type d'école Dans la la figure la figure 3.8 nous pouvons remarquer aisément que 29% des étudiants proviennent des écoles dites protestantes , 27% viennent des écoles catholiques , 11% des écoles privé , 15 des écoles

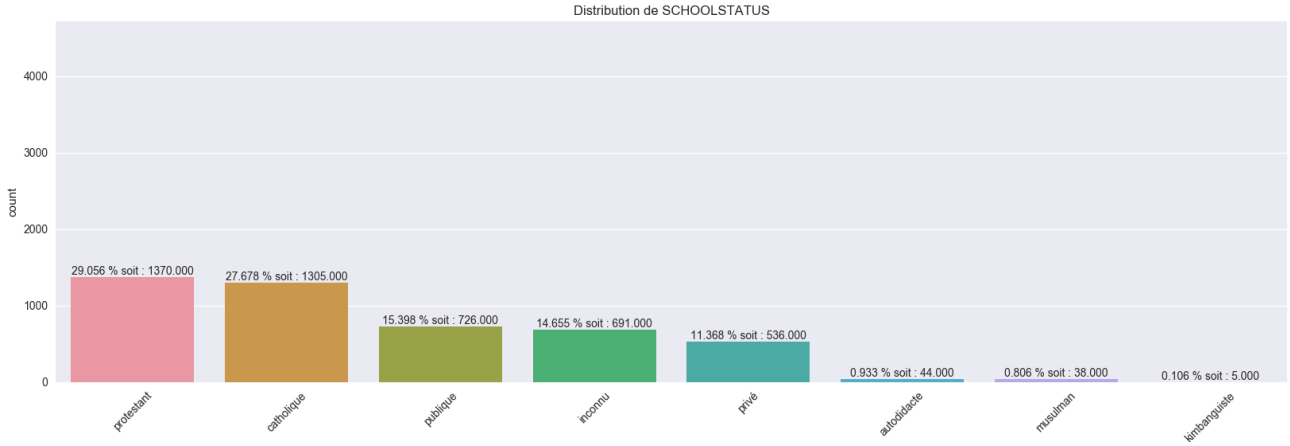


FIGURE 3.8 – Différente statut de l'école de provenance

publiques mais aussi il ya des étudiants venant des autodidactes , ceux provenant des écoles musulmanes et kibanguistes mais en proportion vraiment négligeable.

l'Option de provenance Nos étudiant Proviennent des 33 écoles différentes :

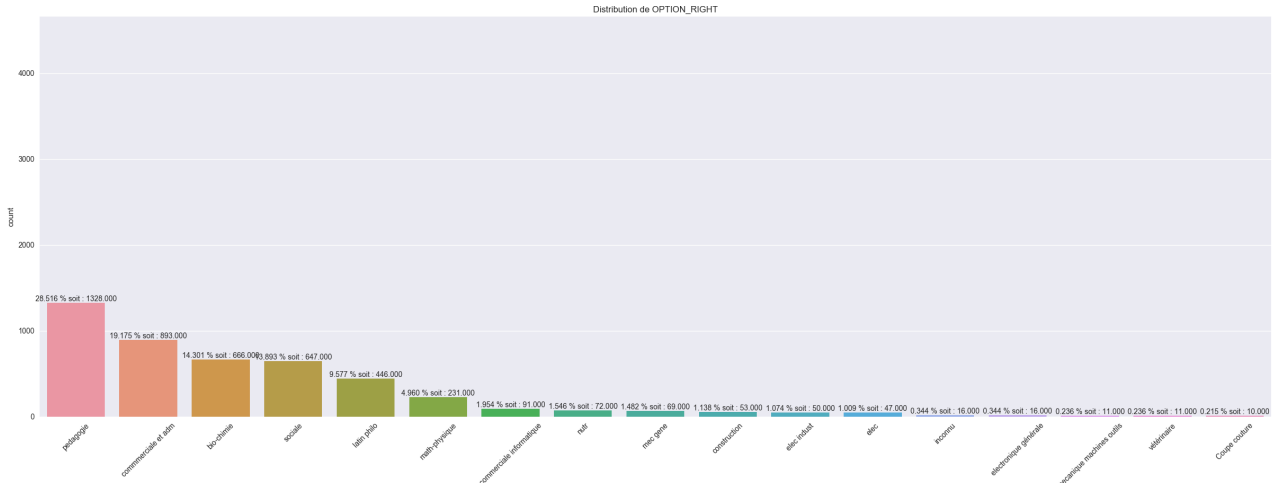


FIGURE 3.9 – l'option suivie à l'école secondaire

Sur la la figure la figure 3.9 nous pouvons remarquer que la majeure partie des étudiants de notre études proviennent de la section pédagogique

avec environ 28% ensuite vienne la section commerciale et administrative avec 19% , suivent sociale avec 13%, scientifique bio-chimie avec 14 % ensuite viennent autres différentes options avec des valeurs inférieurs à 5%.

Attribut School cette attribue comprend les valeurs de l'école de provenance des nos finaliste combiné avec l'attribue SCHOOLSTATUS il joue un rôle important dans notre étude. nous pouvons aisément remarquer que les élèves proviennent de 594 écoles différentes sur la figure 3.10 nous allons visualiser les écoles les plus représentées. Nous pouvons remarquer aisément

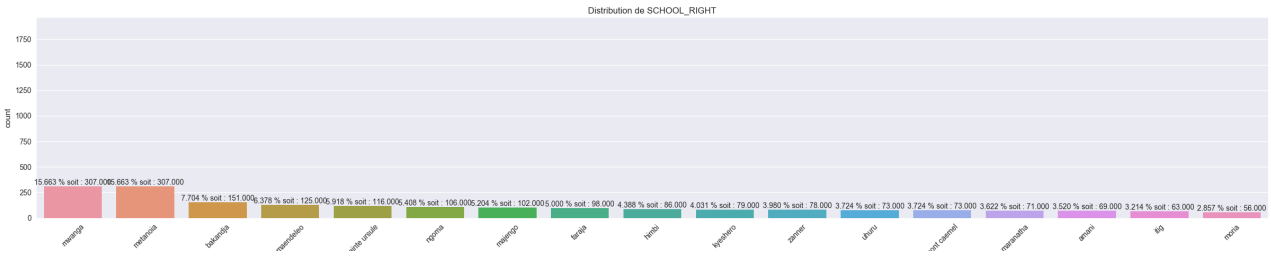


FIGURE 3.10 – les écoles de provenances

que le top 10 des école de provenance est constituer de grandes écoles de la ville de Goma avec l'institut metanoia et le collège mwanga en tête de liste avec l'institut mwanga et metanoia en tete de liste avec 15% chacun ensuite vienne l'institut bakanja avec 6% ensuite vienne maendelo, le lycée sainte ursule et l'institut de Goma avec 6%, 5% et 5 % respectivement et d'autres école se partagent le reste de 50%.

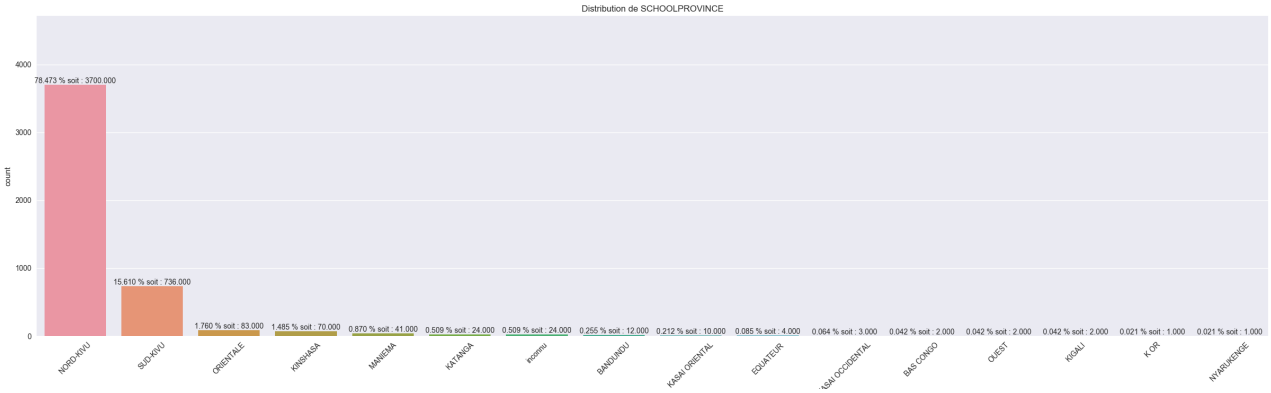


FIGURE 3.11 – les provinces de provenances des étudiants

Attribue SCHOOL Province A la figure 3.11 nous pouvons aisément constater que 75 % des étudiants de l'ULPGL proviennent de la province du nord kivu mais il ya une autre catégorie provenant du Sud Kivu soit 15% l'autre partie provient des autres provinces de la RDC.

Attribue Faculté Pour finaliser l'analyser uni-varié des nos données en entrée nous allons jeter un coup d'œil sur la colonne Fac qui contient la faculté choisie par l'étudiant. Voici comment elle se présente a la figure 3.12 : Nous remarquons la distribution des valeurs pour l'attribue faculté des

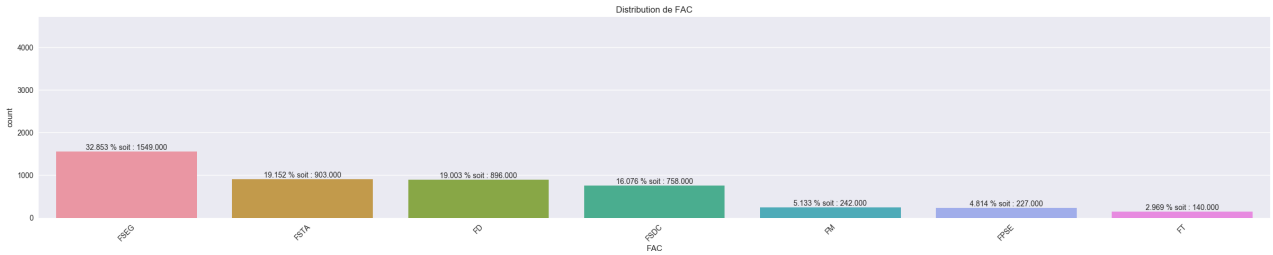


FIGURE 3.12 – les Facultés

étudiants : - FSGEG : 32,8% , FSTA : 19,153% , FD :19% -,FSDC :16% ,FM : 5% ,FPSE : 4% ,FT :3%

Analyse Bi-variée

C'est une technique d'analyse statistique des données, consistant à découvrir la relation pouvant exister entre 2 variables dans le but de tester l'hypothèse d'association et de causalité entre 2 variables !Par exemple dans notre analyser nous allons essayer de voir la relation existant entre le choix de la faculté et le pourcentage du diplôme à l'examen d'état. Elle se déroule en 4 étapes : [29]

- Définition de la nature des relations
- identification et direction des relations
- Détermination si la relation est important du point de vue statistique(Intervalle de confiance)
- Détermination si la relation est important du point de vue statistique(Intervalle de confiance)

Nous effectuerons cette analyse à 3 niveau :

1. ***Variables Continues et catégorielle ou quantitatives :*** Pour effectuer cette analyse nous utiliserons le test Analyse of variance (ANOVA),

c'est une technique statistique permettant de comparer les moyennes de plus de deux populations. Son but est en fait de procéder à une sorte de généralisation de la comparaison des moyennes ou de la comparaison des pourcentages lorsqu'il y a plus de deux valeurs à comparer. Il s'agit aussi de l'équivalent, pour des variables qualitatives de la régression linéaire.

Cette technique est utile en sciences sociales dans l'analyse de certaines données, organisées en blocs de même taille. Il s'agit dans ce cas le plus souvent d'analyse de la variance à un seul facteur. Les analyses à deux facteurs sont en revanche fréquentes dans l'exploitation d'enquêtes d'usage psychologique.

Les données de notre étude répondent bel et bien à l'usage de la technique de l'Analyse de la variance. Ainsi, à travers les pages qui suivent, nous analysons l'effet des variables Age, pourcentage à l'examen d'état, sur l'option choisie par l'étudiant . Par la même occasion nous nous permettons de comparer des histogrammes de chaque groupe en fonction des mêmes variables ; ainsi que les divers seuils de significative.

Pour cette analyse l'hypothèse nulle est du type : H0 : les moyennes sont égales dans toutes les catégories. et son hypothèse alternatif est H1 : au moins une moyenne est différente des autres..

2. ***Variable catégorielle ou catégorielle :*** Pour ces types des données nous allons effectué le test de chi-carré : Le chi-carré est un test statistique conçu pour déterminer si la différence entre deux distributions de fréquences est attribuable à l'erreur d'échantillonnage (le hasard) ou est suffisamment grande pour être statistiquement significative. Ho - est, comme son nom l'indique, une hypothèse qui postule qu'il n'y a pas de différence entre les fréquences ou les proportions des deux groupes elle est considéré comme hypothèse nulle. Si la différence entre les deux distributions est réduite, l'hypothèse nulle sera acceptée. Si la différence est grande, l'hypothèse nulle sera rejetée. Dans ce dernier cas, on parlera d'une différence statistiquement significative parce que l'écart entre les deux distributions est trop important pour être expliqué par le hasard seulement : une différence réelle existe donc.
3. ***Variable continues et continues :*** Pour les variables continues on utilise cherche la corrélation et pour notre travail nous allons utilisée le coefficient de corrélation de Pearson : Les coefficients de corrélation

permettent de donner une mesure synthétique de l'intensité de la relation entre deux caractères et de son sens lorsque cette relation est monotone. Le coefficient de corrélation de Pearson permet d'analyser les relations linéaires et le coefficient de corrélation de Spearman les relations non-linéaires monotones. Il existe d'autres coefficients pour les relations non-linéaires et non-monotones. Signalons que python dispose des multiples librairies pour effectuer ces genres d'analyse et nous les utiliserons dans la suite

Passons maintenant à l'analyser des proprement dite Dans cette partie nous allons nous allons effectué une analyse bivariée entre les attribues en entrée et les attribues en sortie , pour une première approche nous allons faire une analyse entre la faculté choisie et les différentes variables d'entres de notre ensemble d'apprentissage dans la seconde approche nous essayerons de le faire la même chose pour le autres variables de sortie. voici le différentes combinaisons que nous allons effectuer :

1. FAC-Diplome Province : Pour voir la relation entre la faculté et la province d'origine de l'étudiant
2. FAC-DIPLOMPOURCENTAGE : Pour voir la relation entre la faculté et la pourcentage obtenu à l'EXETAT
3. FAC-AGE : Pour voir la relation entre l'age de l'étudiant et la FAC
4. FAC-DIPLOMEOPTION : Pour voir la relation entre la faculté et l'option du diplôme
5. FAC-SEXE : Pour la relation avec le sexe des étudiants
6. FAC-SCHOOL : pour la relation entre l'école de provenance la fac
7. FAC-SCHOOL-STATUTS : pour la relation entre le statuts de l'école de la FAC

Analyse des Variables continues Vs Variables Numériques Comme ces une des colonnes dispose des variables continues nous allons utiliser le test A-NOVA ce test nous permettra de savoir si la moyenne de l'age des étudiants est la même pour chaque faculté : cela constituera notre hypothèse nulle ,on va cherche la probabilité p est on décidera sur base de cette valeur ! si elle es inférieur à 0.05 on rejettera l'hypothèse nulle.

AGE comme la valeur de PR est inférieur à 0.05 on peut conclure que la moyenne de l'age n'est pas la même au sein de chaque faculté Pour prouver le rejet de notre hypothèse nulle on peut remarquer que les facultés de Médecine et celui de technologie on une moyenne d'age de 21 et 23 respectivement et les

TABLE 3.4 – Analyse ANOVA Age - faculté

	df	sum_sq	mean_sq	F	PR(>F)
C(FAC)	6.0	14857.002658	2476.167110	135.823792	3.721673e-159
Residual	4708.0	85830.284723	18.230732	NaN	NaN

TABLE 3.5 – Analyse répartition de la moyenne de l'age et pourcentage dans les facultés

FAC	DIPLOMPERCENTAGE	AGE
FD	56.151372	24.771205
FM	59.434420	21.487603
FPSE	56.202099	28.224670
FSDC	55.329211	25.974934
FSEG	56.832564	24.323434
FSTA	58.941594	23.307863
FT	53.814286	31.428571

facultés de psychologie et celui de théologie on une moyenne d'age respective de 28 et 31 ans. Pour plus de détails on peut visualiser les détails sur la figure **??**

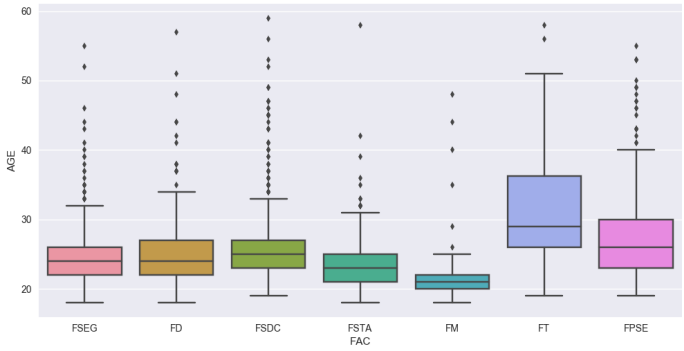


FIGURE 3.13 – distribution des ages au sein des différentes facultés

Pourcentage Également ici on rejette aussi notre hypothèse nulle qui stipulait que la moyenne du pourcentage du diplôme est la même au sein de chaque faculté . Pour prouver le rejet de notre hypothèse nulle on peut re-

marquer que les facultés de Médecine et celui de technologie on une moyenne de pourcentage de 59% chacun et celui de théologie à une moyenne de 53%. Pour Plus des détails sur les distributions statistique

TABLE 3.6 – Analyse ANOVA pourcentage - faculté

	df	sum_sq	mean_sq	F	PR(>F)
C(FAC)	6.0	9139.202728	1523.200455	48.757704	2.256165e-58
Residual	4708.0	147078.865397	31.240201	NaN	NaN

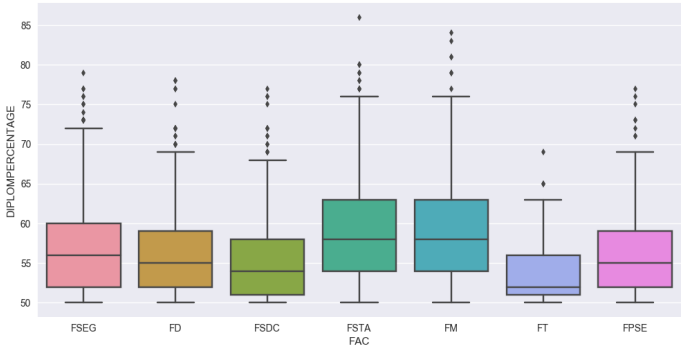


FIGURE 3.14 – distribution des pourcentage au sein des différentes facultés

Variables Catégorielles et Catégorielles Pour ces genres de relation nous avons effectuer le test chi-carré , Dans le tableau 3.7 nous pouvons visualiser les différentes valeurs et effectuer nos conclusions sur base de celle si.
Notre hypothèse nulle H0 est du type : Il n'ya pas de relation entre la variable considéré et la faculté. Voici Les conclusion que nous prenons sur base de notre test statistique :

1. FAC-PROVINCE : lorsque nous comparons ces 2 valeurs nous remarquons que notre valeur de chi2 est supérieur à notre valeur critique et tombe dans la région de rejet de notre hypothèse nulle et ainsi celle est rejetée et ainsi nous concluons avec 95% de certitude que notre hypothèse alternative est vrai : il ya une relation entre la province et la faculté en d'autre terme ayant un nombre des étudiant venant d'une province nous pouvons prédire la faculté choisie.

TABLE 3.7 – test chi carrée avec des différentes variables avec la faculté

Variable	Dégrée De Lib.	Valeur Critique T	Valeur de Chi2	Décision sur H0
Province	90	113	256	Rejet
Option	174	205	4155	Rejet
SchoolStatus	42	58	304	Rejet
SCHOOL	3558	3697	6389	Rejet
SEXE	6	12,5	445	Rejet

2. FAC-SEXE : l ya une relation de dépendance entre le sexe et la faculté ce qui est logique car par exemple en faculté de technologie et théologie on trouve moins des hommes que des femmes.
3. FAC-OPTION : sur base de ce fait on fait la même conclusion que pour la province : il ya une relation de dépendance entre la section du diplôme et la faculté ce qui est logique car les étudiant en général se base sur leur option pour choisir leur faculté
4. FAC-SCHOOLSTATUS : il ya une relation de dépendance entre la section du diplôme et la faculté ce qui est logique car les étudiant en général se base sur leur option pour choisir leur option

Bibliographie

[1] E. B. OMAR S. IMANE. "Prediction approach for improving students orientation in university : case of sidi mohammed ben abdelah university". In : *International Journal of Computer Science and Applications* (2015), p. 108.

[2] I. Perrier R. R. TREMBLAY. *Savoir plus : outils et méthodes de travail intellectuel*. Les Éditions de la Chenelière inc, 2006.

[3] RITHME CONSULTING. *Méthodologie de mise en place du processus de data mining*. 2017. URL : http://www.rithme.eu/?m=resources\&p=dmmethod\&lang=fr (visité le 29/04/2017).

[4] K. MULENDA. "cours d'intelligence artificielle et systèmes experts". inédit FSTA ULPGL. juin 2016.

[5] Mark A. Hall IAN H. WITTEN Eibe Frank. *Data Mining Practical Machine Learning Tools and Techniques*. USA : Morgan Kaufmann Publishers, 2011, p. 1.

[6] DHAKSHAYANI N. *What is the difference between data mining, artificial intelligence and machine learning ?* Sous la dir. de QUORA. Answer To quora Question. URL : \url{https://www.quora.com/What-is-the-difference-between-data-mining-artificial-intelligence-and-machine-learning}.

[7] M. BATER. *Learning Data mining with R*. Mumbai : Pack publish, 2015, p. 10.

[8] Djazia CHAMI. "Une plate forme orientée agent pour le data mining". Thèse de doct. Université de Batna 2, 2010.

[9] SHARP SIGHT LABS. *What is the difference between machine learning,statistics,and datamining ?* Sous la dir. de SHARP SIGHT LABS. http://www.sharpsightlabs.com/differenceÂŋmachineÂŋlearningÂŋstatisticsÂŋda ?utm_content=bufferf9474\&utm_medium=social. Accessed : 2016-05-16.

[10] Gongsun L. *What is the difference between data mining, statistics, machine learning and AI ?* Sous la dir. de StackExchange. Answer To StackExchange Question. URL : \url{https://stats.stackexchange.com/questions/5026/what-is-the-difference-between-data-mining-statistics-machine-learning-and-ai}.

[11] Andrew Ng. *Stanford CS229 -Machine Learning -Ng*. https://www.coursera.org/learn/machine-learning. Accessed : 2016-11-23 au 2016-12-05. 2011.

[12] Bishop C. *Neural networks for pattern recognition*. Oxford : Clarendon Press, 1995.

[13] Alexandre KOWALCZYK. *SVM Tutorial*. http://www.svm-tutorial.com/. Accessed : 2016-12-6. 2014.

[14] Eric Kim. *Everything You Wanted to Know about the Kernel Trick*. http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html. Accessed : 2016-12-09. 2011.

[15] M.Saerens C. Decaestecker. *Les arbres de décision (decision trees)*. Belgique : inédit ULB, 2006, p. 1.

[16] ANALYTICS VIDHYA CONTENT TEAM. *A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python) MACHINE LEARNING PYTHON R*. 2016. URL : https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/ (visité le 27/03/2017).

[17] Leo Breiman. "Random forests". In : t. 45. 1. Springer, 2001, p. 5–32.

[18] Laure J. *Guide Pratique Parents*. Paris, 2015, p. 10.

[19] Bruno Racine. et al s. *Rapport haut conseil d'éducation*. Paris : inédit, 2008.

[20] ANNIE TARDIF. *L'orientation scolaire et professionnelle pour la réussite des élèves handicapés ou en difficulté*. Québec : inédit, 2008.

[21] J. Guichard et M. Huteau. *Psychologie de l'orientation*. Paris : Dunod, 2009.

[22] P ANDREANI F. et LARTIQUE. *L'Orientation des élèves*. Paris : A. Colin, 2006.

[23] A. DELVAUX J.P. et BRUYNEEL. *L'orientation à l'entrée des universités congolaises*. Kinshasa : Univ. de Kinshasa, 1971.

[24] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly Media, Inc., March 2017.

[25] SUNIL RAY. *Guide to Data Exploration*. 2016. URL : https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration (visité le 27/05/2017).

[26] Fabian PEDREGOSA et al. "Scikit-learn : Machine learning in Python". In : *Journal of Machine Learning Research* 12 (2011), p. 2825–2830.

[27] Rubin D.B. LITTLE R.J.A. *Statistical Analysis with Missing Data*. John Wiley, 1987.

[28] NIlesh SH. *Fuzzy Group By, Grouping Similar Words*. Sous la dir. de Stack EXCHANGE. Answer To StackOverflow question. URL : https://stackoverflow.com/questions/11535483/fuzzy-group-by-grouping-similar-words.

[29] William BECKER. "Uncertainty propagation through large nonlinear models". Thèse de doct. University of Sheffield, 2011.