



N° d'ordre : .....

**UNIVERSITE DE M'SILA**  
**FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE**

**Département d'Informatique**

**MEMOIRE de fin d'étude**

**Présenté pour l'obtention du diplôme de MASTER**

**Domaine : Mathématiques et Informatique**

**Filière : Informatique**

**Spécialité : Systèmes d'Informations Avancés**

**Par: Tabet Salah eddine.**

**SUJET**

**Classification des données non équilibrées**

**Encadreur : Brahimi Belkacem**

**Promotion : 2014 /2015**



# Dédicaces

Que ce travail témoigne de mes respects :

A mes parents :

Grâce à leurs tendres encouragements et leurs grands sacrifices, ils ont pu créer le climat affectueux et propice à la poursuite de mes études.

Et Surtout A ma mère qui est malheureusement décédée cette année et ne pourra pas assister à ma graduation.

Aucune dédicace ne pourrait exprimer mon respect, ma considération et mes profonds sentiments envers eux.

Je prie le bon Dieu de les bénir, de veiller sur eux, en espérant qu'ils seront toujours fiers de moi.

A ma sœur et à mon frère.

A tous mes professeurs :

Leur générosité et leur soutien m'oblige de leur témoigner mon profond respect et ma loyale considération.

A tous mes amis et mes collègues :

Ils vont trouver ici le témoignage d'une fidélité et d'une amitié infinie.

**Tabet Salah Eddine.**

# Remerciements

Avant de commencer la présentation de ce travail, je profite de l'occasion pour remercier toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce projet de fin d'études.

Je tiens à exprimer mes vifs remerciements pour mon grand et respectueux professeur, **M. BRAHIMI BELKACEM**, d'avoir accepté de m'encadrer pour mon projet de fin d'études, ainsi que pour son soutien et son encouragement.

Je tiens à exprimer ma profonde reconnaissance et toutes mes pensées de gratitude à **M. Benazzi**, qui m'a accompagné de près durant tout ce travail, pour sa disponibilité, pour la confiance qu'il a su m'accorder et les conseils précieux qu'il m'a prodigués tout au long de la réalisation de ce projet. Mes remerciements vont aussi à mes respectueux professeurs, et spécialement pour **Mr Mhenni Taher**, et **Mr Mokhtarri** pour leur soutien et leur remarques pendant toute la période du travail sur ce projet,

Mes remerciements vont aussi à tous les personnes qui m'ont soutenus jusqu'au bout, et qui n'ont pas cessé de me donner des conseils très importants en signe de reconnaissance.

# Table des matières

Dédicaces .....	I
Remerciments.....	II
Table des matières .....	III
Listes des Figures et tableaux.....	IV
Introduction Générale .....	1

## Chapitre 1

1-Introduction.....	2
2- Définition de la fouille de données.....	2
3- Processus du data mining.....	3
3.1- Définition et compréhension du problème.....	4
3.2 - Collecte des données .....	4
3.3 – Prétraitement.....	5
3.4- Estimation du modèle.....	5
4- Les tâches.....	5
4.1– Classification.....	6
4.2– L'estimation.....	6
4.3– Le groupement par similitude.....	7
4.4– L'analyse des clusters (segmentation).....	7
4.5– La description.....	7
5- Les Données.....	7
5.1 - Les données discrètes.....	8
5.2-Les données continues .....	8
5.3-Les dates .....	8
5.4-Les données textuelles.....	8
6- Les Méthodes.....	8
7- Conclusion .....	9

## Chapitre 2

1- Introduction .....	10
2- Définition .....	10
3-Les approches.....	10
3.1- Classification non supervisée.....	11
3.2-Classification supervisée.....	11
4- Les algorithmes de classification non supervisée .....	11
4.1 – K-moyennes.....	11
4.2- Règles d'association et motifs séquentiel.....	11
5-Les algorithmes de classification supervisée.....	12
5.1- Méthode de Bayes naïf .....	12
5.2- k plus proches voisins .....	13

5.3- Les arbres de décision .....	14
5.4 - Les réseaux de neurones .....	15
5.5 - machines à vecteurs de support.....	16
<b>6-Conclusion .....</b>	<b>17</b>
<b>Chapitre 3</b>	
<b>1- Introduction.....</b>	<b>18</b>
<b>2-Définition du Problème a étudié .....</b>	<b>18</b>
<b>3- SAMPLING METHODS.....</b>	<b>19</b>
3.1- Oversampling (sur-échantillonnage).....	19
3.2- Undersampling (sous-échantillonnage).....	20
3.3- HYBRID Sampling Algorithm .....	21
<b>4-Autres méthodes similaire .....</b>	<b>22</b>
<b>5-Etat de l'art et recherches précédentes .....</b>	<b>22</b>
<b>6-Choix du Domaine D'application .....</b>	<b>24</b>
6.1- Domaine Biologique.....	25
6.2- Le domaine de la médecine.....	25
<b>7-Methodologie du travail .....</b>	<b>25</b>
7.1- Sur échantillonnages (SMOTE) .....	25
7.2- Sous échantillonnages ( Under Sampling) .....	26
7.3- La combinaison des deux méthodes (SMOTE and SpreadSubSample).....	26
<b>8-Conclusion.....</b>	<b>26</b>
<b>Chapitre 4</b>	
<b>1-Introduction.....</b>	<b>27</b>
<b>2- Le Choix des algorithmes de classification .....</b>	<b>27</b>
<b>3-Choix et description des bases de données utilisées .....</b>	<b>27</b>
3.1- La Base de donnée 'Iris' .....	27
3.2- La Base de données Breast Cancer Wisconsin (Original).....	28
3.3-La BD du Centre Service de transfusion sanguine .....	29
<b>4- Les critères de mesure des performances des algorithmes .....</b>	<b>30</b>
4.1-Le rappel .....	31
4.2-La Précision .....	32
4.3-La F-mesure .....	32
<b>5- Méthode d'échantillonnage .....</b>	<b>32</b>
5.1- Utiliser l'ensemble d'apprentissage (using training set).....	32
5.2- Ensembles de testes fournis (supplied test set).....	33
5.3- La validation croisée ( cross validation) .....	33
5.4- Pourcentage scission (Pourcentage split).....	33
<b>6-Outil de travail .....</b>	<b>33</b>
<b>7-Expérimentations Sur La base de donné IRIS.....</b>	<b>34</b>
7.1-IRIS Base référence .....	34
7.1.1- OverSampling (SMOTE).....	34
7.1.2-UnderSampling ( SpreadSubSample).....	36
7.1.3-Hybryde (SMOTE and SpreadSubSample).....	36
7.2-IRIS base modifiée .....	37

<b>8- Expérimentations sur La Base de données breast cancer wisconsin.....</b>	<b>37</b>
8.1-Oversampling (SMOTE) .....	37
8.2-UnderSampling (SpreadSubSample) .....	39
8.3-Hypride (SMOTE and SpreadSubSample).....	40
8.4- La Meilleur Méthode pour Breast Cancer .....	41
<b>9- Travail Avec La Base de données Blood Transfusion .....</b>	<b>42</b>
9.1-Oversampling ( SMOTE).....	42
9.2- Undersampling (SpreadSubSampling).....	43
9.3-Hybrid (SMOTE and SpreadSubSample ).....	44
9.4- Meilleur méthode pour Blood Transfusion.....	45
<b>10 - Conclusion .....</b>	<b>45</b>
<b>Conclusion Générale .....</b>	<b>47</b>
<b>Bibliographie.....</b>	<b>48</b>

# Listes Des Figures et Tableaux

## Liste des figures :

Figure1- un cas de données déséquilibrées (imbalanced data).....	1
Figure 1.1- Le Modèle de Fouille de Données.....	3
Figure 1.2 – Processus de data mining (CRISP-DM).....	4
Figure 2.1- A a un plus proche voisin B, B a de nombreux voisins proches autres que A..	14
Figure 2.2 - exemple d'arbre de décision .....	15
Figure.2.3 : Un perceptron multicouches : une couche d'entrée de 5 cellules, 2 couches cachées possédant respectivement 3 et 2 neurones, 1 couche de sortie à 1 neurone.....	16
Figure 4.1-L'interface du logiciel Weka.....	34
Figure 4.2 - le nombre d'instances après L'application de la hybride sur Breast Cancer.....	40
Figure 4.3- le nombre d'instances après L'application de l'hybride sur Blood Transfusion.....	44.

## Liste des tableaux :

Table 4.1- résultats avec et sans SMOTE avec 3 classifieur sur IRIS part 1.....	34
Table 4.2 - résultats avec et sans SMOTE avec 3 classifieur sur IRIS part 2.....	36
Table4.3 – comparaison de Base Iris avant et après modification (SMOTE test).....	37
Table 4.4- résultats de breast cancer part 1 .....	38
Table 4.5- résultats de breast cancer part 2.....	39
Table 4.6-Résultat Avec Undersampling pour Breast Cancer.....	40
Table 4.7- Résultat avec la méthode hybride pour breast cancer .....	41
Table 4.8-comparaison avec SMOTE.....	41
Table 4.9- comparaison avec Hybride.....	41
Table 4.10- comparaison avec spreadsubsample.....	41
Table 4.11- Les résultats avec SMOTE sur Blood Transfusion part 1 .....	42
Table 4.12- Les résultats avec SMOTE sur Blood Transfusion part 2.....	43
Table 4.13 –Undersampling avec la base Blood Transfusion.....	43
Table 4.14 - Résultat avec la méthode hybride pour Blood Transfusion.....	44
Table 4.15- comparaison de methode avec SMOTE.....	45
Table 4.16-undersamplig comparaison.....	45
Table 4.17-Hybrid comparaison.....	45

## **Introduction Générale**

# Introduction générale

La fouille de données ou le Data Mining en anglais est l'extraction de connaissances à partir de données, qui a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, Elle se propose d'utiliser un ensemble d'algorithmes issus de disciplines scientifiques diverses telles que les statistiques, l'intelligence artificielle ou l'informatique, pour construire des modèles à partir des données. La principale propriété de ce type de problème de classification est que les exemples d'une classe sont plus nombreux de manière significative qu'aux autres exemples [8], la classe minoritaire représente généralement le concept le plus important à étudier, et il est difficile de l'identifier, car il pourrait être associé à des cas exceptionnels et significatifs.

Le Principe obstacle des ensembles de données déséquilibrées est que les algorithmes d'apprentissage de classification standards sont souvent biaisés en faveur de la classe majoritaire (connu comme la classe «négative») et donc il ya un taux d'erreur de classification plus élevé pour les instances de classe minoritaire (appelé les exemples "positifs"), car la plupart des études sur le comportement de plusieurs classificateurs standards dans les domaines de déséquilibre ont montré que la perte significative de performance est principalement due à la distribution de classe biaisée. Au cours des dernières années, de nombreuses solutions ont été proposées pour résoudre ce problème, parmi ces derniers on peut citer : Les algorithmes de modification, l'apprentissage sensible aux coûts (Cost-sensitive learning), et l'échantillonnage (Data sampling).

Notre travail a pour but d'appliquer des algorithmes de classification choisis sur ce cas de déséquilibre de données, de comparer leurs performances (avec les propriétés Accuracy et F-mesure), et de tester leur robustesse, et aussi de comparer les différentes méthodes pour ajuster et équilibrer les données biaisées (échantillonnage, augmentation, combinaisons ....).

Dans la partie qui suit on expliquera en détaille la fouille de données, puis on entamera dans le deuxième chapitre la classification avec les algorithmes choisis, le troisième chapitre contiendra notre problématique des données déséquilibré, et les expérimentations et testes seront dans le dernier et principale chapitre de ce mémoire.

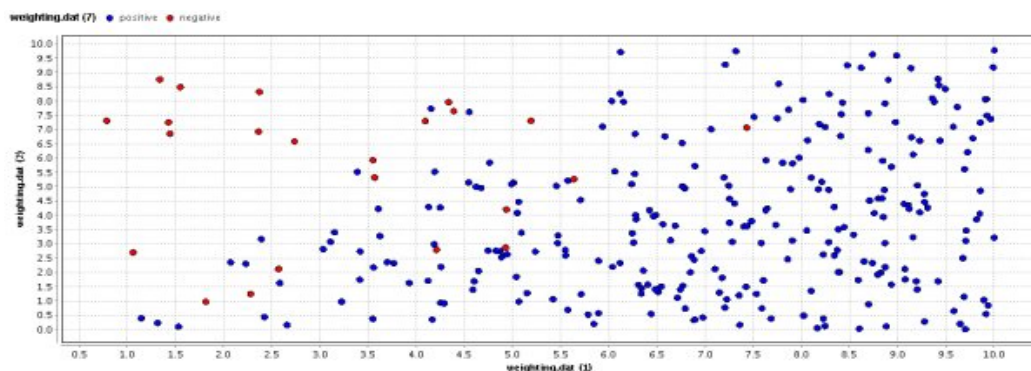


Figure1- Un cas de données déséquilibrées (imbalanced data) [29].

## **CHAPITRE 1**

### **FOUILLE DE DONNEES (DATA MINING)**

## **1-Introduction :**

L'extraction de données, ou bien l'extraction de l'information prédictive cachée dans de grandes bases de données, est une nouvelle technologie puissante avec un grand potentiel pour aider les entreprises à se concentrer sur l'information la plus importante dans leurs entrepôts de données. Dans ce chapitre on va résumer en brève la définition de la fouille de données, et donner un aperçu général sur le processus de cette dernière.

Et aussi donner un aperçu général sur le processus de la fouille de données, avec les type de données à fouiller, les méthodes qu'on peut utiliser, en expliquant aussi la tâche qu'on peut établir, comme l'estimation l'analyse des clusters, et la classification qui est la partie essentiel dans notre cas, qu'on va expliquer prochainement.

## **2- Définition de la fouille de données :**

La fouille de données est un domaine qui est apparu avec l'explosion des quantités d'informations stockées, avec le progrès important des vitesses de traitement et des supports de stockage. La fouille de données vise à découvrir, dans les grandes quantités de données, les informations précieuses qui peuvent aider à comprendre les données ou à prédire le comportement des données futures. La fouille de données utilise depuis son apparition plusieurs outils de statistiques et d'intelligence artificielle pour atteindre ses objectifs.

La fouille de données s'intègre dans le processus d'extraction des connaissances à partir des données ECD ou (KDD : Knowledge Discovery from Data en anglais). Ce domaine en pleine expansion est souvent appelé le data mining. La fouille de données est souvent définie comme étant le processus de découverte des nouvelles connaissances en examinant de larges quantités de données (stockées dans des entrepôts) en utilisant les technologies de reconnaissance de formes de même que les techniques statistiques et mathématiques. Ces connaissances, qu'on ignore au début, peuvent être des corrélations, des patterns ou des tendances générales de ces données. La science et l'ingénierie modernes sont basées sur l'idée d'analyser les problèmes pour comprendre leurs principes et leur développer les modèles mathématiques adéquats. Les données expérimentales sont utilisées par la suite pour vérifier la correction du système ou l'estimation de quelques paramètres difficile à la modélisation mathématiques. Cependant, dans la majorité des cas, les systèmes n'ont pas de principes compris ou qui sont trop complexes pour la modélisation mathématique. Avec le développement des ordinateurs, on a pu rassembler une très grande quantité de données à propos de ces systèmes. La fouille de données vise à exploiter ces données pour extraire des modèles en estimant les relations entre les variables (entrées et sorties) de ses systèmes. En effet, chaque jour nos banques, nos hôpitaux, nos institutions scientifiques, nos magasins, ... produisent et enregistrent des milliards et des milliards de données.

La fouille de données représente tout le processus utilisant les techniques informatiques (y compris les plus récentes) pour extraire les connaissances utiles dans ces données. Actuellement, La fouille de données utilise divers outils manuels et automatiques : on commence par la description des données, résumer leurs attributs statistiques (moyennes, variances, covariance,...), les visualiser en utilisant les courbes, les graphes, les diagrammes, et enfin rechercher les liens significatifs potentiels entre les variables (tel que les valeurs qui se répètent ensemble). Mais la description des données toute seule ne fournit pas un plan d'action. On doit bâtir un modèle de prédiction basé sur les informations découvertes, puis tester ce modèle sur des données autres que celles originales [9].

### Data Mining Model

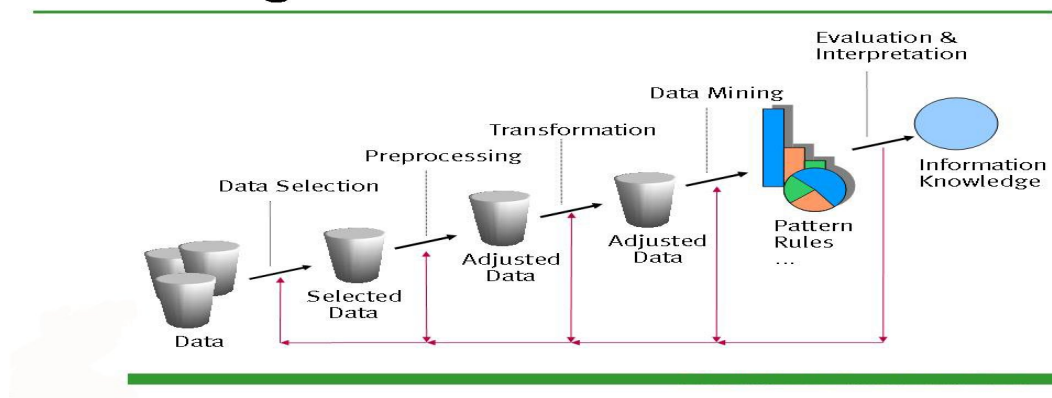


Figure 1.1- Le Modèle de Fouille de Données [30].

### 3- Processus du data mining :

Il est très important de comprendre que le data mining n'est pas seulement le problème de découverte de modèles dans un ensemble de donnée. Ce n'est qu'une seule étape dans tout un processus suivi par les scientifiques, les ingénieurs ou toute autre personne qui cherche à extraire les connaissances à partir des données. En 1996 un groupe d'analystes définit le data mining comme étant un processus composé de cinq étapes sous le standard CRISP-DM (Cross-Industry Standard Process for Data Mining) comme schématisé ci-dessous selon [9] :

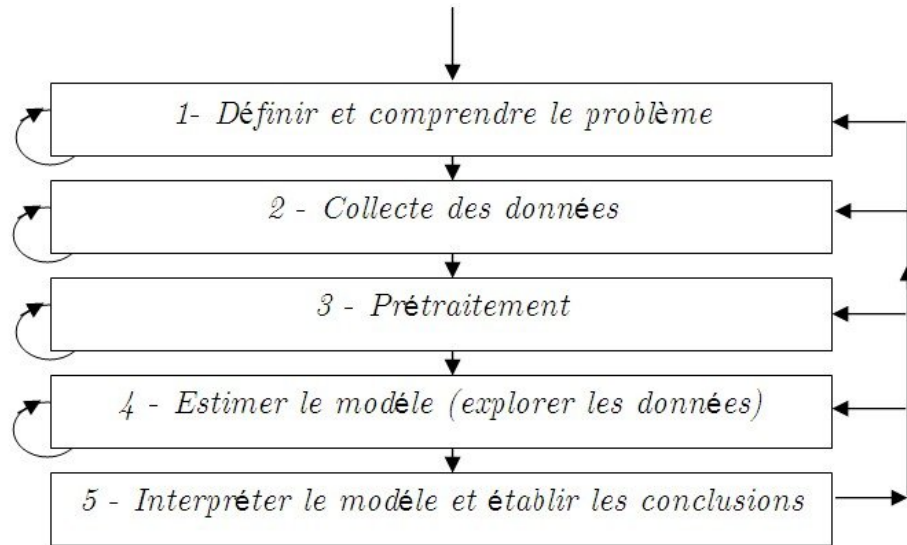


Figure 1.2 – Processus de data mining (CRISP-DM) [9].

Ce processus, composé de cinq étapes, n'est pas linéaire, on peut avoir besoin de revenir à des étapes précédentes pour corriger ou ajouter des données. Par exemple, on peut découvrir à l'étape d'exploration (4) de nouvelles données qui nécessitent d'être ajoutées aux données initiales à l'étape de collection (2). Décrivons maintenant ces étapes :

### 3.1- Définition et compréhension du problème :

Dans la plus part des cas, il est indispensable de comprendre la signification des données et le domaine à explorer. Sans cette compréhension, aucun algorithme ne va donner un résultat fiable. En effet, Avec la compréhension du problème, on peut préparer les données nécessaires à l'exploration et interpréter correctement les résultats obtenus. Généralement, le data mining est effectué dans un domaine particulier (banques, médecine, biologie, marketing, ...etc) où la connaissance et l'expérience dans ce domaine jouent un rôle très important dans la définition du problème, l'orientation de l'exploration et l'explication des résultats obtenus. Une bonne compréhension du problème comporte une mesure des résultats de l'exploration, et éventuellement une justification de son coût. C'est-à-dire, pouvoir évaluer les résultats obtenus et convaincre l'utilisateur de leur rentabilité.

### 3.2-Collecte des données :

Dans cette étape, on s'intéresse à la manière dont les données sont générées et collectées. D'après la définition du problème et des objectifs du data mining, on peut avoir une idée sur les données qui doivent être utilisées. Ces données n'ont pas toujours le même format et la même structure. On peut avoir des textes, des bases de données, des pages web, ...etc. Parfois, on est amené à prendre une copie d'un système d'information en cours d'exécution, puis ramasser les données de sources

éventuellement hétérogènes (fichiers, bases de données relationnelles, temporelles, ...). Quelques traitements ne nécessitent qu'une partie des données, on doit alors sélectionner les données adéquates. Généralement les données sont subdivisées en deux parties : une utilisée pour construire un modèle et l'autre pour le tester. On prend par exemple une partie importante (suffisante pour l'analyse) des données (80 %) à partir de laquelle on construit un modèle qui prédit les données futures. Pour valider ce modèle, on le teste sur la partie restante (20 %) dont on connaît le comportement.

### **3.3-Prétraitement :**

Les données collectées doivent être "préparées". Avant tout, elles doivent être nettoyées puisqu'elles peuvent contenir plusieurs types d'anomalies : des données peuvent être omises à cause des erreurs de frappe ou à causes des erreurs dues au système lui-même. Des données peuvent être incohérentes c-à-d qui sortent des intervalles permis, on doit les écarter ou les normaliser. Parfois on est obligé à faire des transformations sur les données pour unifier leur poids. Le prétraitement comporte aussi la réduction des données [9] qui permet de réduire le nombre d'attributs pour accélérer les calculs et représenter les données sous un format optimal pour l'exploration. Plusieurs techniques de visualisation des données telles que les courbes, les diagrammes, les graphes,... etc, peuvent aider à la sélection et le nettoyage des données. Une fois les données collectées, nettoyées et prétraitées on les appelle entrepôt de données (Data Warehouse).

### **3.4- Estimation du modèle :**

Dans cette étape, on doit choisir la bonne technique pour extraire les connaissances (exploration) des données. Des techniques telles que les réseaux de neurones, les arbres de décision, les réseaux bayésiens, le clustering, ... sont utilisées. Généralement, l'implémentation se base sur plusieurs de ces techniques, puis on choisit le bon résultat. Dans le chapitre suivant on va détailler les différentes techniques utilisées dans l'exploration des données et l'estimation du modèle.

### **3.5-Interprétation du modèle et établissement des conclusions :**

Généralement, l'objectif du data mining est d'aider à la prise de décision en fournissant des modèles compréhensibles aux utilisateurs. En effet, les utilisateurs ne demandent pas des pages et des pages de chiffres, mais des interprétations des modèles obtenus. Les expériences montrent que les modèles simples sont plus compréhensibles mais moins précis, alors que ceux complexes sont plus précis mais difficiles à interpréter.

## **4- Les tâches :**

Nous allons présenter, dans cette partie, les tâches, i.e. les problèmes que l'on cherche à résoudre et quelques-unes des techniques disponibles pour résoudre ces tâches.

On dispose de données structurées. Les objets sont représentés par des enregistrements (ou descriptions) qui sont constitués d'un ensemble de champs (ou attributs) prenant leurs valeurs dans un domaine. On peut mettre en évidence différentes problématiques. Les termes employés pouvant varier d'une discipline à l'autre (parfois même dans une même discipline selon le domaine d'application), nous définissons notre vocabulaire avec la description associée de la tâche.

Beaucoup de problèmes intellectuels, économiques ou même commerciaux peuvent être exprimés en termes des six tâches suivantes :

- La classification.
- L'estimation.
- Le groupement par similitude (règles d'association).
- L'analyse des clusters.
- La description.

Les trois premières tâches sont des exemples de la fouille supervisée de données dont le but est d'utiliser les données disponibles pour créer un modèle décrivant une variable particulière prise comme but en termes de ces données. Le groupement par similitude et l'analyse des clusters sont des tâches non-supervisées où le but est d'établir un certain rapport entre toutes La description appartient à ces deux catégories de tâches, elle est vue comme une tâche supervisée et non-supervisée en même temps.

#### **4.1– Classification :**

La classification est la tâche la plus connue de la fouille de données qui semble être une tâche humaine primordiale. Afin de comprendre notre vie quotidienne, nous sommes constamment obligés à classer, catégoriser et évaluer. La classification consiste à étudier les caractéristiques d'un nouvel objet pour l'attribuer à une classe prédéfinie, elle sera expliquée en détail dans le prochain chapitre.

#### **4.2– L'estimation :**

L'estimation est similaire à la classification à part que la variable de sortie est numérique plutôt que catégorique. En fonction des autres champs de l'enregistrement l'estimation consiste à compléter une valeur manquante dans un champ particulier.

Des exemples de tâche d'estimation sont :

- noter un candidat à un prêt ; cette estimation peut être utilisée pour attribuer un prêt (classification), par exemple, en fixant un seuil d'attribution,
- estimer les revenus d'un client.

#### **4.3– Le groupement par similitude (Analyse des associations et de motifs séquentiels) :**

Le groupement par similitude consiste à déterminer quels attributs "vont ensemble". La tâche la plus répandue dans le monde du business, est celle appelée l'analyse d'affinité ou l'analyse du panier du marché, elle permet de rechercher des associations pour mesurer la relation entre deux ou plusieurs attributs. Les règles d'associations sont, généralement, de la forme "Si <antécédent>, alors <conséquent>".

#### **4.4– L'analyse des clusters (segmentation) :**

Le clustering (ou la segmentation) est le regroupement d'enregistrements ou des observations en classes d'objets similaires. Un cluster est une collection d'enregistrements similaires l'un à l'autre, et différents de ceux existants dans les autres clusters. La différence entre le clustering et la classification est que dans le clustering il n'y a pas de variables sortantes. La tâche de clustering ne classe pas, n'estime pas, ne prévoit pas la valeur d'une variable sortantes. Au lieu de cela, les algorithmes de clustering visent à segmenter la totalité de données en des sous groupes relative- ment homogènes. Ils maximisent l'homogénéité à l'intérieur de chaque groupe et la minimisent entre les différents groupes.

#### **4.5– La description :**

Parfois le but de la fouille est simplement de décrire ce qui se passe sur une Base de Données compliquée en expliquant les relations existantes dans les données pour en premier lieu comprendre le mieux possible les individus, les produit et les processus présents dans cette base. Une bonne description d'un comportement implique souvent une bonne explication de celui-ci. Comme exemple, "les femmes supportent le changement plus que les hommes", peut provoquer beaucoup d'intérêt et promouvoir les études de la part des journalistes, sociologues, économistes et les spécialistes en politique .

### **5- Les Données :**

Ce sont les valeurs des champs des enregistrements des tables de l'entrepôt. Ces données possèdent un type qu'il est important de préciser. En effet, la plupart des méthodes sont sensibles aux données manipulées. Par exemple, certaines méthodes sont mises en défaut par les données continues alors que d'autres peuvent être sensibles à la présence de données discrètes.

### **5.1-Les données discrètes :**

- les données binaires ou logiques : 0 ou 1 ; oui ou non ; vrai ou faux. ce sont des données telles que le sexe, être bon client, ...
- les données énumératives : ce sont des données discrètes pour lesquelles il n'existe pas d'ordre défini a priori. Par exemple : la catégorie socioprofessionnelle, la couleur, ...
- les données énumératives ordonnées : les réponses à une enquête d'opinion (1: très satisfait ; 2 : satisfait ; ...), les données issues de la discrétisation de données continues (1 : solde moyen < 2000 ; 2 :  $2000 \leq \text{solde moyen} < 5000$  ; ...)

### **5.2-Les données continues :**

Ce sont les données entières ou réelles : l'âge, le revenu moyen, ... mais aussi les données pouvant prendre un grand nombre de valeurs ordonnées.

### **5.3-Les dates :**

Sont souvent problématiques car mémorisées selon des formats différents selon les systèmes et les logiciels. Pour les applications en fouille de données, il est fréquent de les transformer en données continues ou en données énumératives ordonnées. On transforme une date de naissance en âge entier ou en une variable énumérative ordonnée correspondant à des tranches d'âge.

### **5.4-Les données textuelles :**

Ne sont pas considérées dans notre cas. Cependant, un texte peut, pour certaines applications, être résumé comme un n-uplet constitué du nombre d'occurrences dans le texte de mots clés d'un dictionnaire prédéfini.

## **6- Les Méthodes :**

Nous ne présentons que certaines méthodes qui viennent compléter les outils classiques que sont : les requêtes SQL, les requêtes d'analyse croisée, les outils de visualisation, la statistique descriptive et l'analyse des données. Les méthodes choisies qui seront détaillées dans les sections suivantes sont :

- un algorithme pour la segmentation,
- les règles d'association,
- les plus proches voisins (raisonnement à partir de cas),
- les arbres de décision,
- les réseaux de neurones,
- les algorithmes génétiques.

Des méthodes importantes ne sont pas étudiées dans ce cours. Citons la programmation logique inductive et les machines à vecteurs de support (SVM for Support Vector Machine).

De plus, pour les méthodes proposées dans cette recherche, il existe de nombreuses variantes, non présentées, qui peuvent s'appliquer à des tâches différentes. En tout état de cause, un fait important communément admis est que :

*<<Il n'existe pas de méthode supérieure à toutes les autres>>.*

Par conséquent, à tout jeu de données et tout problème correspond une ou plusieurs méthodes. Le choix se fera en fonction

- de la tâche à résoudre,
- de la nature et de la disponibilité des données,
- des connaissances et des compétences disponibles,
- de la finalité du modèle construit.
- de l'environnement de l'entreprise.

Dans le chapitre suivant des exemples seront définis.

## **7- Conclusion :**

L'exploration de données est un processus 'aide à la décision', dans lequel nous cherchons des modèles d'informations dans les données. Dans ce chapitre nous avons entamé les différentes processus et étapes concernant la fouille de données, les types de données existantes, et aussi les différentes tâches qui se résument à l'estimation et la classification de données, qui est une partie essentielle à s'approfondir dedans. Et aussi des petites entrées sur les méthodes d'analyse de données qui sont proposées dans ce mémoire, il existe de nombreuses variantes, qui peuvent s'appliquer à des tâches différentes. Comme les classifieurs célèbres dans la fouille de données, les arbres de décision ou le plus proches voisin par exemple, plus de détail sera entamé dans le chapitre suivant.

## **CHAPITRE 2**

### **LA CLASSIFICATION SUPERVISEE**

## **1- Introduction :**

La classification est la tâche la plus importante dans la fouille de données. Elle consiste à attribuer automatiquement des exemples à des classes bien déterminées. Mais avant d'appliquer une méthode de classification, une étape préalable est nécessaire pour préparer la base de données (nettoyage, remplissage, ...), pour ensuite appliquer une classification soit supervisée avec différentes méthodes ou non supervisée avec d'autres méthodes.

En effet, dans la pratique, les données souffrent de plusieurs situations (redondance, manque, biais, ...), nous allons choisir les trois algorithmes de classification les plus utilisés pour la classification supervisée de ces genres de données, et définir leur méthode de travail.

## **2-Définition :**

La classification est un outil puissant d'exploration des données. Elle est parmi les tâches les plus importantes du data mining ou fouille de données. Au-delà de ces aspects descriptifs, la classification est aussi un intermédiaire pour d'autres objectifs. Par exemple, les classes peuvent définir des sous-populations où des modèles s'adapteront mieux qu'aux populations générales. La classification est aussi utilisée conjointement aux techniques factorielles pour une exploration multidimensionnelle plus efficace.

Voici les exemples de cas où la tâche d'analyse des données est de Classification:

\* Un agent de prêt bancaire veut analyser les données afin de savoir quel client (prêt demandeur) sont à risque ou qui sont sans danger.

\* Un directeur du marketing à une entreprise a besoin d'analyser de deviner un client avec un profil donné va acheter un nouvel ordinateur. Dans les deux exemples ci-dessus un modèle ou un classificateur est conçu pour prédire les étiquettes catégoriques. Ces étiquettes sont à risque ou danger pour les données de demande de prêt et oui ou non des données de marketing.

Il ya deux formes d'analyse de données qui peuvent être utilisés pour les modèles d'extrait décrivant les classes importantes ou de prédire les tendances futures de données. Ces deux formes sont les suivantes:

- La classification - La prédiction.

Ces données d'analyse nous aident à fournir une meilleure compréhension des grandes données. Par exemple, nous pouvons construire un modèle de classification pour classer les demandes de prêt bancaire que soit sûr ou risqué, ou un modèle de prédiction pour prédire les dépenses en dollars de clients potentiels sur le matériel informatique donné leur revenu et la profession.

## **3-Les approches :**

La classification est une technique utilisée pour regrouper les objets dans des classes d'objets telles que:

- Les classes sont homogènes (les objets dans une classe ont des caractéristiques semblables)
- Chaque classe a des caractéristiques propres qui la différencient des autres classes.

On distingue deux types d'approches de classification ou d'apprentissage :

### 3.1- Classification non supervisé :

Quand le système ou l'opérateur ne disposent que d'exemples, mais non d'étiquettes, et que le nombre de classes et leur nature n'ont pas été prédéterminés, on parle d'apprentissage non supervisé. Aucun expert n'est requis. L'algorithme doit découvrir par lui-même la structure plus ou moins *cachée* des données [21].

### 3.2-Classification supervisé :

Si les *classes* sont prédéterminées et les *exemples* connus, le système apprend à classer selon un *modèle* de classement, on parle alors d'apprentissage supervisé. Un expert (ou *oracle*) doit préalablement étiqueter des exemples. Le processus se passe en deux phases. Lors de la première phase, il s'agit de déterminer un modèle des données étiquetées. La seconde phase consiste à prédire l'étiquette d'une nouvelle donnée, connaissant le modèle préalablement appris. Parfois il est préférable d'associer une donnée non pas à une classe unique, mais une probabilité d'appartenance à chacune des classes prédéterminées [21].

## 4- Les algorithmes de classification non supervisée :

Ce n'est pas notre intérêt dans notre cas ici, mais cette classification comporte l'algorithme des K-moyennes (K-means en anglais), et règles d'associations.

### 4.1 – K-moyennes :

D'après Wikipédia [20] on a que le partitionnement en *k*-moyennes (ou *k*-means en anglais) est une méthode de partitionnement de données et un problème d'optimisation combinatoire. Étant donnés des points et un entier *k*, le problème est de diviser les points en *k* partitions, souvent appelés *clusters*, de façon à minimiser une certaine fonction. On considère la distance d'un point à la moyenne des points de son cluster ; la fonction à minimiser est la somme des carrés de ces distances.

Il existe une heuristique classique pour ce problème, souvent appelée *méthodes des k-moyennes*, utilisée pour la plupart des applications. Le problème est aussi étudié comme un problème d'optimisation classique, avec par exemple des algorithmes d'approximation.

Les nuées dynamiques sont une généralisation de ce principe, pour laquelle chaque partition est représentée par un noyau pouvant être plus complexe qu'une moyenne.

Comme inconvenant la méthode des *k*-moyennes et ses variantes résolvent une tâche dite non supervisée, c'est-à-dire qu'elle ne nécessite aucune information sur les données. La segmentation peut être utile pour découvrir une structure cachée qui permettra d'améliorer les résultats de méthodes d'apprentissage supervisé (classification, estimation, prédiction).

### 4.2- Règles d'association et motifs séquentiel :

La recherche des règles d'association est une méthode populaire étudiée d'une manière approfondie dont le but est de découvrir des relations ayant un intérêt pour le statisticien entre deux ou plusieurs variables stockées dans de très importantes bases de données.

On peut utiliser par exemple les règles d'associations dont le but est de découvrir des

similitudes entre des produits dans des données saisies sur une grande échelle dans les systèmes informatiques des points de ventes des chaînes de supermarchés.

Une telle information peut être utilisée comme base pour prendre des décisions marketing telles que par exemple des promotions ou des emplacements bien choisis pour les produits associés, les règles d'association sont employées aujourd'hui dans plusieurs domaines incluant celui de la fouille du web, de la détection d'intrusion et de la bio-informatique.[21].

Un attrait principal de la méthode est la clarté des résultats produits. En effet, le résultat de la méthode est un ensemble de *règles d'association*. Des exemples de règles d'association sont :

- si un client achète des plantes alors il achète du terreau,
- si un client achète du poisson et du citron alors il achète du vin blanc.
- si un client achète une télévision, il achètera un magnétoscope dans un an.

Ces règles sont intuitivement faciles à interpréter car elles montrent comment des produits ou des services se situent les uns par rapport aux autres.

Parmi les Algorithmes qui peut utiliser les règles d'associations je cite : L'algorithme le plus connue Apriori, Eclat, et Fp- growth, Opus et beaucoup d'autres.

Comme critique on peut dire que les règles d'association sont faciles à interpréter. Elles sont faciles à utiliser pour des utilisations concrètes, et aussi que la méthode est plus efficace pour les articles fréquents. Pour les articles rares, on peut restreindre la forme des règles choisies ou faire varier le support minimum.

## **5-Les algorithmes de classification supervisée :**

On présente ici plusieurs types d'algorithmes:

### **5.1- Méthode de Bayes naïf :**

Le classifieur bayésien naïf est une méthode d'apprentissage supervisé qui repose sur une hypothèse simplificatrice forte : les descripteurs ( $X_j$ ) sont deux à deux indépendants conditionnellement aux valeurs de la variable à prédire ( $Y$ ) [19]. Pourtant, malgré cela, il se révèle robuste et efficace. Ses performances sont comparables aux autres techniques d'apprentissage.

La classification naïve bayésienne est un type de classification Bayésienne probabiliste simple basée sur le théorème de Bayes, l'estimation des paramètres pour les modèles bayésiens naïfs repose sur le maximum de vraisemblance. L'avantage du classifieur bayésien naïf est qu'il requiert relativement peu de données d'entraînement pour estimer les paramètres nécessaires à la classification.

Soit  $X$  un échantillon de données («preuves»): étiquette de classe est inconnue, et Soit  $H$  une hypothèse que  $X$  appartient à la classe  $C$ , la Classification consiste à déterminer  $P(H | X)$ , la probabilité que l'hypothèse est donné l'échantillon de données observées  $X$ ,  $P(H)$  (de probabilité a priori), la probabilité initiale. Par exemple,  $X$  va acheter l'ordinateur, indépendamment de l'âge, le revenu, ...

$P(X)$ : la probabilité que les données de l'échantillon est observé

$P(X | H)$  (probabilité a posteriori), la probabilité d'observer l'échantillon  $X$ , étant donné que l'hypothèse est titulaire. Par exemple, étant donné que  $X$  va acheter un ordinateur, le prob. que  $X$  est 31..40, le revenu moyen.

Compte tenu des données de formation  $X$ , probabilité a posteriori d'une hypothèse  $H$ ,  $P(H | X)$ , suit le théorème de Bayes selon [28] :

$$P(H | \mathbf{X}) = \frac{P(\mathbf{X} | H)P(H)}{P(\mathbf{X})} \quad (1)$$

Officieusement, cela peut être écrit comme :

posteriori = probabilité x avant / preuves

Prédit  $X$  appartient à  $C_2$  ssi la probabilité  $P(C_i | X)$  est le plus élevé parmi tous les  $P(C_k | X)$  pour toutes les classes de  $k$

Difficulté pratique: exiger des connaissances initial de nombreuses probabilités, coût de calcul importante.

Soit  $D$  un ensemble de tuples et leurs étiquettes de classe associés de formation, et chaque tuple est représenté par un vecteur d'attribut  $nD$   $X = (x_1, x_2, \dots, x_n)$

Supposons qu'il existe  $m$  catégories  $C_1, C_2, \dots, C_m$ .

La classification est de tirer le maximum de posteriori, à savoir le maximum  $P(C_i | X)$

Ceci peut être obtenu d'après le théorème de Bayes

$$P(C_i | \mathbf{X}) = \frac{P(\mathbf{X} | C_i)P(C_i)}{P(\mathbf{X})} \quad (2)$$

Étant donné que  $P(\mathbf{X})$  est constante pour toutes les classes, seulement doit être maximisée  $P(C_i | \mathbf{X}) = P(\mathbf{X} | C_i)P(C_i)$  (3)

La dérivation du classifieur naïve bayes sera comme suite :

$$P(\mathbf{X} | C_i) = \prod_{k=1}^n P(x_k | C_i) = P(x_1 | C_i) \times P(x_2 | C_i) \times \dots \times P(x_n | C_i) \quad (4)$$

## 5.2- k plus proches voisins :

La méthode des plus proches voisins (PPV en bref, nearest neighbor en anglais ou kNN pour abrégée) est une méthode dédiée à la classification qui peut être étendue à des tâches d'estimation. La méthode PPV est une méthode de raisonnement à partir de cas. Elle part de l'idée de prendre des décisions en recherchant un ou des cas similaires déjà résolus en mémoire. Contrairement aux autres méthodes de classification, il n'y a pas d'étape d'apprentissage consistant en la construction d'un modèle à partir d'un échantillon d'apprentissage. C'est l'échantillon d'apprentissage, associé à une fonction de distance et d'une fonction du choix de la classe en fonction des classes voisins les plus proches, qui constitue le modèle.

Toutes les instances correspondent à des points dans l'espace nD  
 Le voisin le plus proche sont définis en termes de distance euclidienne,  $\text{dist}(X_1, X_2)$ ,  
 la fonction cible pourrait être discrète- ou réel évalué.  
 Pour Discret-évaluée, k-NN renvoie la valeur la plus courante parmi les exemples de formation de k plus proches de  $X_q$ .  
 Diagramme de Voronoï: la surface de décision induite par 1-NN pour un ensemble typique d'exemples de formation.

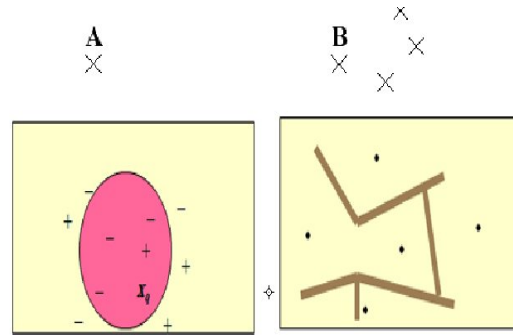


Figure.2.1- A a un plus proche voisin B, B a de nombreux voisins proches autres que A

Le k-NN pour la prédiction de valeur réelle pour un tuple donné inconnue Renvoie les valeurs moyennes des k plus proches voisins.

La distance pondérée du plus proche voisin algorithm, affecte un Poids de la contribution de chacun des voisins de k en fonction de leur distance à la requête  $x_q$ , il donne le plus de poids aux plus proches voisins

$$w \equiv \frac{1}{d(x_q, x_i)^2} \quad (5)$$

Le KNN est Robuste aux données bruitées en faisant la moyenne des k plus proches voisins, par contre la distance entre voisins pourrait être dominé par les attributs non pertinents Pour éviter ça, il faut avoir les axes étirer ou l'élimination des attributs les moins pertinents.

### 5.3- Les arbres de décision :

Les arbres de décision sont l'une des structures de données majeures de l'apprentissage statistique. Leur fonctionnement repose sur des heuristiques qui, tout en satisfaisant l'intuition, donnent des résultats remarquables en pratique (notamment lorsqu'ils sont utilisés en « forêts aléatoires »). Leur structure arborescente les rend également lisibles par un être humain, contrairement à d'autres approches où le prédicteur construit est une « boîte noire » [18].

Un arbre de décision est une représentation graphique d'une procédure de classification. Les nœuds internes de l'arbre sont des tests sur les champs ou attributs, les feuilles sont les classes. Lorsque les tests sont binaires, le fils gauche correspond à une réponse positive au test et le fils droit à une réponse négative. Un exemple d'arbre de décision est:

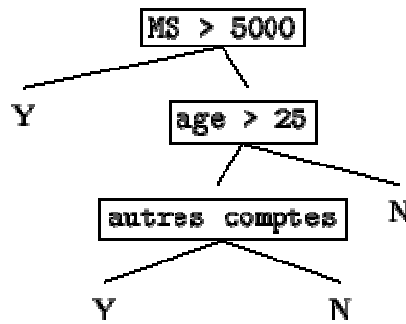


Figure 2.2 - exemple d'arbre de décision [24].

MS est la moyenne des soldes du compte courant, autres comptes est un champ binaire qui vaut oui si le client dispose d'autres comptes, la classe Y indique un a priori favorable pour l'attribution d'un prêt.

Pour classer un enregistrement, il suffit de descendre dans l'arbre selon les réponses aux différents tests pour l'enregistrement considéré. On peut déjà remarquer quelques propriétés importantes des arbres de décision :

- la procédure de classification associée est compréhensible par tout utilisateur,
- la classe associée à un enregistrement particulier peut être justifiée,
- les attributs apparaissant dans l'arbre sont les attributs pertinents pour le problème de classification considéré.

On peut également remarquer qu'un arbre de décision est un système de règles. Il est immédiat de transformer l'arbre de la Figure 5 en :

Si MS > 5000 alors Y  
 Si MS ≤ 5000 et age > 25 et autres comptes=oui alors Y  
 Si MS ≤ 5000 et age > 25 et autres comptes=oui alors Y  
 Si MS ≤ 5000 et age ≤ 25 alors N

Un arbre de décision est facile à interpréter et est la représentation graphique d'un ensemble de règles. Si la taille de l'arbre est importante, il est difficile d'appréhender l'arbre dans sa globalité. Cependant, les outils actuels permettent une navigation aisée dans l'arbre (parcourir une branche, développer un nœud, élaguer une branche), et aussi que l'algorithme peut prendre en compte tous les types d'attributs et les valeurs manquantes. Il est robuste au bruit.

#### 5.4 - Les réseaux de neurones :

Les réseaux de neurones sont des outils très utilisés pour la classification, l'estimation, la prédiction et la segmentation. Ils sont issus de modèles biologiques, sont constitués d'unités élémentaires (les neurones) organisées selon une architecture. Nous nous limitons dans ce paragraphe aux réseaux de neurones dédiés aux tâches d'estimation et classification que sont les *Perceptrons multicouches (PMC)*. Ceux-ci obtiennent de bonnes performances, en particulier, pour la reconnaissance de formes et sont donc bien adaptés pour des problèmes

comprenant des variables continues éventuellement bruitées. Le principal désavantage est qu'un réseau est défini par une architecture et un grand ensemble de paramètres réels (les coefficients synaptiques), le pouvoir explicatif est faible : on parle parfois de < boîte noire >.

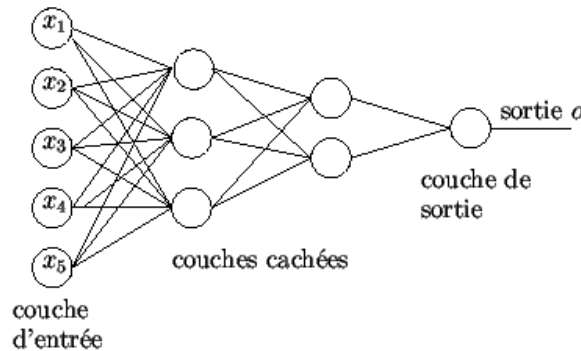


Figure.2.3 : Un perceptron multicouche [24].

Ici dans la figure 2.3 on a une couche d'entrée de 5 cellules, 2 couches cachées possédant respectivement 3 et 2 neurones, 1 couche de sortie à 1 neurone

Le résultat de l'apprentissage est un réseau constitué de cellules organisées selon une architecture, définies par une fonction d'activation et un très grand nombre de poids à valeurs réelles. Ces poids sont difficilement interprétables. Pour un vecteur d'entrée, il est difficile d'expliquer le pourquoi de la sortie calculée.

Le réseau étant construit, le calcul d'une sortie à partir d'un vecteur d'entrée est un calcul très rapide [24].

### 5.5 - machines à vecteurs de support:

Du Site Officiel [25] on trouve que Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais *Support Vector Machine*, SVM) sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination note et de régression. Les SVM sont une généralisation des classifieurs linéaires. Les SVM ont rapidement été adoptés pour leur capacité à travailler avec des données de grandes dimensions, le faible nombre d'hyperparamètres, leurs garanties théoriques, et leurs bons résultats en pratique .

Les SVM ont été appliqués à de très nombreux domaines (bio-informatique, recherche d'information, vision par ordinateur, finance...). Selon les données, la performance des machines à vecteurs de support est de même ordre, ou même supérieure, à celle d'un réseau de neurones ou d'un modèle de mélanges gaussiens.

Elles sont des classificateurs qui reposent sur deux idées clés, qui permettent de traiter des problèmes de discrimination non-linéaire, et de reformuler le problème de classement comme un problème d'optimisation quadratique.

## **6-Conclusion :**

Dans ce chapitre on a résumé en générale la définition de la classification qui est la plus répandue des techniques descriptives, elle consiste à construire un modèle à partir d'exemples dont les classes sont connues à l'avance.

On a entamé aussi la définition de l'apprentissage automatique des données et ces types, qui se résument, à l'apprentissage supervisée qui inclue beaucoup de méthodes et algorithmes de classification nous avons donné un aperçu que pour les plus connues (la méthode bayésienne, le plus proche voisin, les arbres de décision, SVM ...), et en autre partie l'apprentissage non supervisée qui dans ce cas ne nous concerne pas dans cette étude. Reste à étudier un cas de problèmes avec les données (biais, manque...) sur lesquelles on appliquer l'apprentissage supervisé, dans ce qui suit le cas de déséquilibre de données et ça relation avec la classification vont être abordé.

## **CHAPITRE 3**

### **LE PROBLEME DE LA CLASSIFICATION DE DONNEES NON EQUILIBREE (IMBALANCED DATA)**

## **1- Introduction :**

La Classification est un problème qui a été bien étudié dans l'apprentissage automatique récemment. Diverses techniques de classification tels que les arbres de décision, réseaux de neurones, le plus proche voisin et l'induction de règles ont été mis au point et appliqué avec succès à de nombreux domaines. Beaucoup de ces algorithmes de classifications standards actuelles, supposent généralement que l'ensemble d'apprentissage (training exemples en anglais) est répartis équitablement entre les différentes classes. Cependant d'après les chercheurs dans les années précédentes, qui ont remarqué que les ensembles de données déséquilibrées apparaissent souvent dans de nombreuses applications pratiques. Dans un ensemble de données non équilibrée, la classe majoritaire est représentée par une grande partie de tous les exemples, tandis que l'autre, la classe minoritaire, a seulement un faible pourcentage de tous les exemples. Pour un problème de classification multi-classe, il peut y avoir plusieurs classes minoritaires. Dans certaines applications, toutes les classes sont les minorités et ce qui est particulièrement vrai pour de nombreuses applications d'extraction d'information.

Des études ont montré que dans de nombreuses applications, la distribution des classes déséquilibrées conduisent à de mauvaises performances des algorithmes de classification standard. Ces algorithmes de classification génèrent des classificateurs qui maximisent la précision globale de classification. Lorsqu'on traite les données déséquilibrées directement, on aura toujours de mauvaises performances.

Les méthodes comprennent le redimensionnement des ensembles de données de formation (resizing of training data sets), en ajustant les coûts de mauvaise classification (adjusting misclassification costs), et l'apprentissage basé sur la reconnaissance (apprentissage de la minorité classe). La première méthode (Sampling technique or resizing training data set) est notre intérêt dans ce chapitre, cette méthode consiste à redimensionner la quantité de données et jouer avec le nombre d'échantillons pour aboutir à des tests plus performants.

## **2- Définition du Problème à étudié :**

Au cours des dernières années des changements majeurs et des évolutions ont été faites sur la classification des données.

Cette dernière devient difficile en raison de la taille illimitée et le déséquilibre nature des données. Le problème de déséquilibre de classe devient le plus grand problème dans l'exploration de données.

Le problème de déséquilibre appelé en anglais (Problem of Imbalanced Data ) se produit lorsque l'une des classes ayant un échantillon plus que les autres classes, une grande différence de nombre ou de pourcentages de données par rapport à tout l'ensemble .

Ce problème est un défi relativement nouveau qui a attiré une attention croissante à la fois académique et industriel.

Le plus et la majorité des Algorithmes sont plus concentrés sur la classification de l'échantillon majeure tout en ignorant ou une mauvaise classification échantillon minoritaire.

Les échantillons des minorités sont ceux qui se produisent rarement mais très important. Il existe différentes méthodes disponibles pour la classification de l'ensemble de données déséquilibrées (Imbalanced Data) qui est divisé en trois catégories principales, l'approche algorithmique, l'approche prétraitement de données (preprocessing data approach) qui nous concerne et l'approche de sélection de fonction. Chacun de cette technique à ses propres avantages et inconvénients.

Dans notre approche on va définir et aboutir à ce qui donne en finale la bonne direction pour la recherche dans le problème de déséquilibre de classe, et nous présenterons une analyse de l'évolution de la recherche dans l'apprentissage à partir de données déséquilibrées. Notre objectif est de fournir une étude comparative entre les algorithmes de classification et des méthodes d'amélioration qui vont être définis par suite.

### **3- SAMPLING METHODS (méthodes d'échantillonnage) :**

C'est une méthode facile qui a pour but d'équilibrer les classes, qui consiste principalement au rééchantillonnage de l'ensemble de données d'origine, soit par sur-échantillonnage (over-sampling ) de la classe minoritaire ou par le sous-échantillonnage (under-sampling ) de la classe majoritaire, jusqu'à ce que les classes seront à peu près également représentés. Les deux stratégies peuvent être appliquées dans n'importe quel apprentissage, car ils agissent comme une phase de prétraitement, ce qui permet le système d'apprentissage de recevoir les instances de formation, comme si ils appartenait à un ensemble de données bien équilibré.

Ainsi, tout biais du système vers la classe majoritaire en raison de la différente proportion d'exemples par classe on peut s'attendre à une suppression.

Selon [27] dans le domaine, on suggère que l'utilité des méthodes de rééchantillonnages dépend d'un certain nombre de facteurs, y compris le rapport entre les exemples positifs et négatifs, d'autres caractéristiques des données, et la nature de les classer. Cependant, les méthodes de rééchantillonnages ont montré aussi des inconvénients. Parmi eux le sous-échantillonnage qui peut ignorer potentiellement des données utiles, par contre le sur-échantillonnage augmente artificiellement la taille de l'ensemble des données et, par conséquent, augmente le temps de calcul de l'algorithme d'apprentissage.

#### **3.1- Oversampling (sur-échantillonnage) :**

La méthode la plus simple pour augmenter la taille de la classe minoritaire correspondante au sur-échantillonnage aléatoire, ce qui est un procédé non-heuristique qui équilibre la répartition de la classe à travers la réplique aléatoire des exemples positifs [22].

Néanmoins, puisque cette méthode réplique des exemples existants dans la classe minoritaire, plus de surajustement (overfitting ) est le plus susceptible de se produire.

Le chercheur Chawla a proposé la Synthetic Minority Over-sampling Technique (SMOTE) [7], une approche de sur-échantillonnage dans laquelle la classe minoritaire est sur-échantillonnée par la création des exemples synthétiques plutôt que par un sur-échantillonnage

avec remplacement.

La classe minoritaire est sur-échantillonnée en prenant chaque échantillon de classe minoritaire et l'introduction des exemples synthétique sur les segments de droite joignant toute / tous de la classe k minoritaire des voisins les plus proches. Selon la quantité de sur-échantillonnage requis, voisins du k plus proches voisins sont choisis au hasard.

De l'algorithme SMOTE d'origine, plusieurs modifications ont été proposées dans la littérature. Alors que l'approche SMOTE ne gère pas les ensembles de données avec toutes les caractéristiques nominales, il a été généralisé pour traiter des ensembles de données mixtes de fonctions continues et nominales.

Et si on parle des inconvénients du sur échantillonnage on trouve parmi eux qu'elle augmente le nombre d'exemples de formation, augmentant ainsi le temps d'apprentissage.

Compte tenu des inconvénients avec échantillonnage, encore l' échantillonnage est un moyen connue pour traiter des données déséquilibrées plutôt que d'un algorithme d'apprentissage sensible aux coûts. Il ya plusieurs raisons à cela. La raison la plus évidente est qu'il n'existe pas des implémentations rentable avec tous les algorithmes d'apprentissage et donc l'utilisation d'une approche wrapper-based (composé à base de) pour l'échantillonnage est la seule option selon les chercheurs [7].

### **3.2- Undersampling (sous-échantillonnage) :**

Le Sous-échantillonnage est une méthode efficace pour la classification des données non équilibrées. Cette méthode utilise un sous-ensemble de la classe majoritaire pour former le classificateur. Puisque de nombreux exemples de la classe majoritaire sont ignorés, l'ensemble de la formation devient plus équilibré et le processus de formation devient plus rapide [22].

La technique de prétraitement la plus commun est le sous-échantillonnage aléatoire de la majorité (RUS), EN RUS, les instances de la classe de majoritaire sont jetés au hasard dans l'ensemble de données. Cette technique a été utilisée entre autres par Kubat et Matwin [10] dans le but d'obtenir une base d'apprentissage plus équilibrée qui puisse diminuer les inconvénients évoqués des classifieurs.

Cependant, le principal inconvénient du sous-échantillonnage selon [7] qui est cette potentielle utile information contenue dans ces exemples ignorés qui sont négligées.

Pour certains problèmes comme la détection du fraude chevauchent problème de classification de données très déséquilibrées, où des échantillons non-fraude sont plus nombreux que le fraude d'échantillon. T. Maruthi Padmaja dans [28] a proposé une technique d'échantillonnage hybride, une combinaison de SMOTE pour sur-échantillonner les données minoritaires (échantillons de fraude) et sous-échantillonnage aléatoire pur sous-échantillonner les données de la majorité (échantillons non-fraude) si nous éliminons les valeurs extrêmes à partir des échantillons des minorités pour les ensembles de données déséquilibrées très asymétriques comme la détection de la fraude donc la précision de la classification peut être améliorée.

Les méthodes d'échantillonnage considèrent le décalage de classe et propriétés de l'ensemble de données dans son ensemble.

Il existe des inconvénients connus associés à l'utilisation d'échantillonnage à mettre en œuvre un apprentissage sensible aux coûts.

Le sous-échantillonnage peut ignorer les données potentiellement utiles. Par contre le principal inconvénient du suréchantillonnage, de notre point de vue, c'est que en faisant des copies exactes des exemples existants, il fait overfitting susceptible. En fait, avec sur échantillonnage il est assez fréquent pour un apprenant à générer une règle de classification pour couvrir un seul, répliqué, par exemple.

Bien que cela soit certainement moins vrai aujourd'hui que par le passé, de nombreux algorithmes d'apprentissage (par exemple, C4.5) ne gèrent pas encore directement les coûts dans le processus d'apprentissage. Une seconde raison de l'utilisation d'échantillonnage est que de nombreux ensembles de données très asymétriques sont énormes et la taille de l'ensemble de la formation doivent être réduites afin d'apprendre à être faisable.

Dans ce cas, undersampling (sous échantillonnage ) semble être un délai raisonnable, et une stratégie valable si on a besoin de se débarrasser de certaines données de formation, il pourrait encore être bénéfique pour rejeter une partie des exemple de classes majoritaire afin de réduire la taille de l'ensemble de la formation à la taille requise, puis emploient aussi un algorithme d'apprentissage sensible aux coûts, de sorte que le montant de données d'entraînement mis au rebut (disparu) est minimisé.

Une dernière raison qui pourrait avoir contribué à l'utilisation de l'échantillonnage plutôt que d'un algorithme d'apprentissage sensible aux coûts est que les coûts de mauvaise classification sont souvent inconnus.

Cependant, ce ne est pas une raison valable pour l'utilisation de l'échantillonnage sur l'algorithme d'apprentissage sensible aux coûts, puisque la question analogue qui se pose avec échantillonnage quelle devrait être la répartition des classes des données d'entraînement final être? Si cette information de coût n'est pas connu, une mesure telle que l'aire sous la courbe ROC pourrait être utilisé pour mesurer les performances du classificateur et les deux proches pourraient alors déterminer empiriquement le bon ratio de coûts / distribution de classe [27].

### **3.3- HYBRID Sampling Algorithme :**

Dans les données déséquilibrées, certaines fonctionnalités sont redondantes et même hors de propos, ces caractéristiques vont nuire aux performances de généralisation du l'apprentissage automatique (machine learning).

La sélection des fonctionnalités, un processus de choix d'un sous-ensemble de caractéristiques de ceux d'origine, est fréquemment utilisée comme une technique de prétraitement dans l'analyse des données. Il a été prouvé efficace pour réduire la dimensionnalité, amélioration de l'efficacité de l'exploitation minière, en augmentant la précision de l'exploitation minière et l'amélioration de résultat compréhensible.

Comme par exemple si on prend l'arbre de décision en présence de données déséquilibrée qui est une question de grande importance pratique, que de telles données est omniprésente dans une grande variété de domaines d'application. Nous proposons alors hybride échantillonnage

des données, qui utilise une combinaison de deux techniques d'échantillonnage tel que sur-échantillonnage aléatoire et sous-échantillonnage aléatoire, comme celle expliquée dans [16] pour créer un ensemble de données équilibrée pour une utilisation dans la construction de modèles de classification d'arbre de décision. Les résultats démontrent que notre méthodologie est souvent capable d'améliorer la performance d'un apprenant arbre de décision C4.5 dans le contexte de données déséquilibrées.

Cette nouvelle approche, la technique d'échantillonnage hybride est proposée pour améliorer les méthodes précédentes afin de résoudre les problèmes difficiles.

#### **4-Autres méthodes similaire :**

Chawla a proposer SMOTE-NC (Synthetic Minority Over-sampling Technique Nominal Continuous) et SMOTE-N (Synthetic Minority Over-sampling Technique Nominal), SMOTE peut également être prolongé pour les caractéristiques nominales [7].

Andrew Estabrooks et al. ont proposé une méthode de ré-échantillonnage multiple qui choisit le taux de ré-échantillonnage le plus appropriée adaptative [11].

Hongyu Guo et al. découvre exemples durs des classes majoritaires et minoritaires during le processus de renforcement, alors généré de nouveaux exemples de synthèse à partir d'exemples durs et les ajouter à des ensembles de données [12].

Han et Wen-Yuan Wang ont présenté deux méthodes de sur-échantillonnage nouvelle minoritaires, borderline-SMOTE1 et borderline-SMOTE2, dans laquelle seuls les exemples minoritaires près de la frontière sont sur-échantillonnés. Ces approches atteindre un meilleur taux de TP et F-valeur que frappa et aléatoires sur-échantillonnage méthodes [13].

David A. Cieslak, Nitesh V. Chawla a suggéré que la façon pour améliorer un classificateur d'échantillonnage de performance peut être traitée localement, au lieu d'appliquer des niveaux uniformes de l'échantillonnage à l'échelle mondiale. Ils ont proposé un cadre qui identifie d'abord les régions significatives de données et procède ensuite à trouver des niveaux d'échantillonnage optimaux dans chaque régions [14].

Show-Jane Yen et Yeu-Shi Lee [15], on proposée le 'cluster-based undersampling approaches' pour sélectionner les données représentatives en tant que données de formations pour améliorer la précision de la classification pour la classe minoritaire et étudier l'effet de la méthode sous-échantillonnage dans l'environnement de distribution de classe déséquilibrée.

#### **5-Etat de l'art et recherches précédentes :**

Les valeurs finales pour rappel, précision et valeur F pour la classe minoritaire de U2R lorsque les méthodes proposées sont appliquées sur ensemble de données d'intrusion KDDCup-99. On appliquant la technique SMOTE on aura les propriétés par défaut un rappel

de 80.15 % et une précision de 88.62 % et une F-mesure égale à 84.17 % testées établies par [1]. Elle ont trouvé aussi qu'avec la base mamography SMOTE a donné les résultats suivantes un rappel avec 58.04%, précision 64.96% et une f-mesure de 61.31%. Par contre la base de données blood transfusion a donné une précision de 76.07 et une F-mesure de 0.40 en utilisant la technique SMOTE avec l'algorithme KNN. Et en ce qui concerne le Brast cancer collection des résultats a donné une précision égale à 69.50 % et une F-mesure de 0.45. Selon [2]. La discussion aussi des meilleurs résultats de classification obtenus par les trois algorithmes de classification, SVM, Naive Bayes et J48 (C4.5), sur chaque ensemble de données sous chaque métrique de performance, la spécificité, G-moyens est sensible au temps de traitement. La comparaison également des résultats obtenus pour déterminer le meilleur classificateur pour chaque ensemble de données. Si on prend La BD Bank Marketing pour l'algorithme SVM on trouve la sensibilité 0.749 et une spécificité égale à 1 et le G-mean 0.87 Par contre Le Naive bayes 0.638, 0.848 et 0.74 en ordre et si on prend les arbres de décisions J48 ça donnera 0.904, 0.927 et pour le G-means 0.92, Les deux J48 et Naive Bayes montrent une meilleure performance dans ces deux données à petite échelle. Nous pouvons noter que la fécondité et la colonne vertébrale semblent les données les plus difficiles à traiter, car aucun des trois classificateurs a atteint plus de 90% pour le G-métrique signifie: moins de 50% pour la fertilité, et moins de 80% pour la colonne vertébrale. En revanche, les connaissances de l'utilisateur sont les données les plus faciles à être manipulées par les trois classificateurs avec au moins 95% pour les G-moyens. Nous pouvons noter que pour les grands ensembles de données complexes, SVM est beaucoup mieux que les deux autres classificateurs.[3].

Une mise en œuvre de trois algorithmes différents, à savoir, la régression logistique (LR), de réseau neuronal (NN) et Chi-squared Automatic Interaction Detection (CHAID) à un ensemble de données de marketing qui consistent en 2826 (17%) qui a acheté le produit (exemples positifs) et 14 130 (83%) qui ne achète pas le produit (des exemples négatifs). La performance des trois classificateurs a été basée sur la précision, taux de succès et de l'ASC et ont été comparés pour divers ensembles de données de déséquilibre générés par l'ensemble de données d'origine. Ils ont indiqué que le taux de succès (de précision) est une meilleure mesure de la performance de classificateur pour dataset déséquilibrée et CHAID peut être utilisé pour développer des modèles de marketing, une idée testée par [4].

En ce qui concerne la BD (Cardiac Surgery Data), Les Chercheurs [5] ont trouvé qu'avec le suréchantillonnage et sous-échantillonnage, la sensibilité de l'ensemble de test est passée à 69,4% et 68,7% respectivement. Le Suréchantillonnage a été signalé à être sujet de overfitting mais dans cette étude il n'y avait pas de problème de sur apprentissage. Les résultats de CART\_Bagg sont semblables à modèle CART pour l'ensemble de données d'origine et CHAID. Pendant ce temps, CART Boost améliore avec une sensibilité d'essai (27,9%) et de précision (42,2%). Prenant en considération le fait que le petit échantillon de classe minoritaire se traduira par beaucoup plus petit nombre de cas de minoritaires dans les échantillons de formation et d'essai, le CART, algorithmes CHAID et C5 ont été appliqués aux données originales sans partitionnement des données. Les deux CART et CHAID classés tous les 209 cas de minoritaires dans le groupe majoritaire (sensibilité = 0%), tandis que C5

correctement classé 28 (13,4%) des cas minoritaires.

Il en résulte que l'échantillonnage approche fonctionne mieux que l'ensachage et le renforcement de méthodes. La stimulation et l'ensachage n'ont pas amélioré la sensibilité des classificateurs d'arbres de décision.

En étudiant une approche hybride, CART classificateur avec la sélection des fonctionnalités et de la technique d'ensachage a été considéré pour évaluer la performance en termes de précision et de temps pour la classification des différents ensembles de données sur le cancer du sein. Nous avons utilisé l'algorithme d'arbre de décision, car elle produit des règles de classification lisibles par l'homme qui sont faciles à interpréter. Une méthode hybride est proposée d'améliorer la précision de la classification des ensembles de données sur le cancer du sein. Les données d'apprentissage est testée avec de 10 fois la validation croisée dans les paramètres donnant les résultats suivantes : avec une accuracy (précision) de 94.84 % avec l'algorithme CART et 97.85 % avec La Hybrida proche, donc on conclue que les résultats expérimentaux d'une approche hybride avec la combinaison de reprocessing, bagging avec CART qui a démontré l'amélioration de la précision (accuracy) de la classification des ensembles de données et ceci et selon [6].

## 6- Choix du Domaine D'application :

Comme la classification et le clustering, il existe aussi l'analyse factorielle discriminante ou analyse discriminante qui sont des techniques statistique qui vise à décrire, expliquer et prédire l'appartenance à des groupes prédéfinis (classes, modalités de la variable à prédire...) d'un ensemble d'observations (individus, exemples...) à partir d'une série de variables prédictives (descripteurs, variables exogènes...) stocker dans un fichier de base de données d'extension (.arff) qu'on vas définir dans le dernier chapitre.

Ces analyses sont utilisée dans de nombreux domaines sur les quelle ont vas appliquer des séries de tests en donnant des exemples dans ces domaines il existe les suivantes les plus populaire.

- [La médecine](#), par exemple pour détecter les groupes à hauts risques cardiaques à partir de caractéristiques telles que l'alimentation, le fait de fumer ou pas, les antécédents familiaux, etc.
- [Le domaine bancaire](#), lorsque l'on veut évaluer la fiabilité d'un demandeur de crédit à partir de ses revenus, du nombre de personnes à charge, des encours de crédits qu'il détient, etc.
- [La biologie](#), lorsque l'on veut affecter un objet à sa famille d'appartenance à partir de ses caractéristiques physiques. Les iris de Sir Ronald Fisher — qui est à l'origine de cette méthode — en est un exemple, il s'agit de reconnaître le type d'iris (setosa, virginica, et versicolor) à partir de la longueur/largeur de ses pétales et sépales.
- [L'informatique](#), pour la reconnaissance optique de caractères. L'analyse discriminante est utilisée pour reconnaître un caractère imprimé à partir d'informations simples, comme la présence ou non de symétrie, le nombre d'extrémités...

Pour moi j'ai choisi les deux plus sensibles et ou les méthodes d'analyse et les algorithmes ne sont pas appliquées précédemment sur les bases de données sur ces domaine les domaines se résume par qui suit :

### **6.1- Domaine Biologique :**

Un domaine qui parcourt tous ce qui concerne les plante et tous ce qui est végétale donc l'agriculture en générale, nous ans notre travail on a pris comme exemple les plantes, spécialement la plante Iris qui définit comme suite.

### **6.2- Le domaine de la médecine :**

De nos jours, la fouille de donnée est utilisée par de nombreuses applications, et dans divers domaines. En médecine par exemple, elle connaît de grands développements. Néanmoins, les applications médicales réalisant la fouille de donnée demeurent souvent au stade expérimental.

Comme on la mentionner précédemment c'est un domaine beaucoup sensible qui comprend les groupes à hauts risques cardiaques par exemple, ou bien la glycémie appelée aussi "taux de sucre" ou "taux de glucose" dans le sang en gérant l'étude du diabète quotidien chez les gens, nous nous sommes longuement interrogés sur les moyens concrets d'intégrer de tels systèmes à un environnement médical, en considérant les contraintes associées à ce domaine d'application.

## **7-Methodologie du travail :**

Sur les bases de données non équilibrées, les chercheurs ont procédé avec plusieurs méthodes pour les traitées pour ensuite les testées après êtres équilibrés.

Pour notre cas on procède avec la méthode d'échantillonnage (Sampling Technique ) qui se compose de deux méthode :

- 1- Le sur échantillonnages (OverSampling).
- 2- Le sous échantillonnages (UnderSampling).

Ensuite on procède avec une combinaison des deux méthodes précédentes

- 3- Hybride. ( OverSampling + UnderSampling).

### **7.1- Sur échantillonnages (SMOTE) :**

Pour notre cas ici on applique le sur échantillonnages qui est un filtre appelé SMOTE (synthetic Minority Over sampling) qui se trouve dans notre Logiciel de travail Weka [17] qui sera définis plus tard. SMOTE par son nom on comprend qu'il augmente la taille des données dans les classe minoritaires après être équilibré, on entamera avec l'algorithme de classification qui donneras des résultats d'une base équilibré, ces résultats seront commenté et comparées dans le dernier chapitre.

### **7.2- Sous échantillonnages ( Under Sampling) :**

Le sous échantillonnages et le contraire du précédent SMOTE, Son utilisation est de travailler avec la classe Majoritaire, j'essayerais avec mes base de données d'équilibrer leur classes en diminuant le nombre de la classe majoritaire de sorte que les deux classes se rapproche l'une de l'autre, et cette opération existe dans notre outil WEKA en forme de filtre supervisé appelé SpreadSubSample qui sera appliquer sur Les instances, qui nous donnera par suite une distribution nouvelle, donnant des classe équilibrées, ou peut les contrôlées bien sûr en jouant avec les paramètres des filtre . En dernier lieu On utilisera les 3 algorithmes choisis qu'on va voir dans le chapitre suivant pour classifier nos instance et en finale étudier les résultats.

### **7.3- La combinaison des deux méthodes (SMOTE and SpreadSubSample) :**

La technique consiste à appliquer les deux méthodes expliquer précédemment l'une après l'autre sur des mêmes instances. c.à.d. Appliquer SMOTE sur une base de données pour rapprocher la classe minoritaire de la classe majoritaire, ensuite appliquer sur cette même base de données le Filtre SpreadSubSample sur la classe majoritaire pour la diminuer pour plus d'efficacité sur l'équilibrage des classe , et finir en classifiant avec les algorithmes et commenter les résultats .

## **8-Conclusion :**

Nous nous sommes dans cette partie intéressés à la problématique des données déséquilibrées dans le contexte de certain classifieur, Un ensemble de données est déséquilibré si les catégories de classification ne sont pas à peu près également représentées.

Nous avons présenté et discuté les avantages et inconvénients des trois méthodes, tous concernant la manipulation des données, ce chapitre donne un aperçu de la classification des ensembles de données déséquilibrées. Au niveau des données, échantillonnage est l'approche la plus commune pour traiter des données déséquilibrées. De l'état de l'art et les recherches précédentes on peut conclure que le suréchantillonnage apparaît clairement comme mieux que sous-échantillonnage pour les classificateurs locaux, alors que certaines stratégies sous-échantillonnage surperforment l'échantillonnage lors de l'utilisation des classifieurs avec apprentissage global.

Les chercheurs ont prouvé dans [21] que les techniques d'échantillonnage hybrides peuvent faire mieux que simplement sur-échantillonnage et sous-échantillonnage.

## **CHAPITRE 4**

### **EXPERIMENTATIONS ET RESULTATS**

## **1-Introduction :**

L'objectif de ce chapitre est de mettre en œuvre notre méthodologie et tester les différentes approches pour équilibrer les données. Les algorithmes (Naïve bayésien, Le plus proche voisin, l'arbre de décision) servent à tester les échantillons et comparer les résultats avec les mesures de classification (F-mesure, Accuracy), et cela entre les différentes bases de données (Iris, Breast Cancer, Blood Transfusion). Pour enfin conclure à une meilleure décision et méthode utilisée pour le cas déséquilibré des données dans ce domaine de médecine.

## **2- Le Choix des algorithmes de classification :**

Les problèmes de classification constituent une famille de problèmes auxquels il est possible d'appliquer de nombreuses méthodes d'apprentissage supervisé. On a sélectionnées les 3 plus populaires et plus utilisées dans le domaine d'apprentissages de la fouille de données.

La première Méthode c'est la méthode naïve bayes qui est un type d'apprentissage supervisé qui repose sur une hypothèse simplificatrice forte : les descripteurs ( $X_j$ ) sont deux à deux indépendants conditionnellement aux valeurs de la variable à prédire ( $Y$ )<sup>1</sup>. Pourtant, malgré cela, il se révèle robuste et efficace. Ses performances sont comparables aux autres techniques d'apprentissage.

Pour la deuxième, Le KNN (K Nearest Neighbors) La méthode du plus proche voisin qui consiste à estimer la sortie associée à une nouvelle entrée  $x$ , la méthode des  $k$  plus proches voisins consiste à prendre en compte (de façon identique) les  $k$  échantillons d'apprentissage dont l'entrée est la plus proche de la nouvelle entrée  $x$ , selon une distance à définir.

Concernant le Dernier Algorithmes, Les Arbres de décision qui est un modèle prédictif permettant d'évaluer la valeur d'une caractéristique d'un système depuis l'observation d'autres caractéristiques du même système. Dans ces structures d'arbre, les feuilles représentent les valeurs de la variable-cible et les embranchements correspondent à des combinaisons de variables d'entrée qui mènent à ces valeurs.

Un arbre de décision décrit les données mais pas les décisions elles-mêmes, l'arbre serait utilisé comme point de départ au processus de décision. C'est une technique d'apprentissage supervisé: on utilise un ensemble de données pour lesquelles on connaît la valeur de la variable-cible afin de construire l'arbre (données dites étiquetées), puis on extrapole les résultats à l'ensemble des données de test.

## **3- Choix et description des bases de données utilisées :**

### **3.1- La Base de donnée 'Iris' :**

L'iris (nom masculin) est une plante vivace à rhizomes ou à bulbes de la famille des Iridacées (dont fait également partie le crocus).

Le genre Iris contient 210 espèces et d'innombrables variétés horticoles. On trouve souvent dans les jardins des Iris germaniques D'après le UCI Machine learning Repository.

On trouve des iris dans tout l'hémisphère nord, aussi bien en Europe qu'en Asie, en Afrique du

Nord et en Amérique du Nord.

C'est peut-être la plus connue des bases de données, qui se trouve dans la littérature de reconnaissance de formes.

L'article de Fisher est un classique dans le domaine et est fréquemment référencé à ce jour. (Voir Duda et Hart, par exemple.) L'ensemble de données contient trois classes de 50 cas chacun, où chaque classe se réfère à un type de plante de l'iris. Une classe est linéairement séparable de l'autre 2; les derniers ne sont pas linéairement séparables l'une de l'autre.

Attribut prédit : class de la plante Iris.

C'est un domaine extrêmement simple. Ces données diffèrent des données présentées dans l'article Fishers (identifié par Steve Chadwick, spchadwick '@' espeedaz.net). L'échantillon devrait être 35e: 4.9,3.1,1.5,0.2, "Iris-setosa" où l'erreur est dans la quatrième caractéristique. Le 38ème échantillon: 4.9,3.6,1.4,0.1, "Iris-setosa" où les erreurs sont dans les deuxième et troisième caractéristiques.

Renseignements sur les attributs :

1. longueur sépales cm
2. largeur des sépales en cm
3. La longueur des pétales en cm
4. largeur pétales en cm

Classe 5:

- Iris Setosa
- Iris versicolor
- Iris Virginica

**3.2– La Base de données Breast Cancer Wisconsin (Original) :**

Des Échantillons qui arrivent périodiquement, comme le rapporte le Dr Wolberg ses cas cliniques, La base de données reflète donc ce groupement chronologique des données. Cette information regroupant apparaît immédiatement au-dessous, après avoir été retiré de la donnée elle-même:

- Groupe 1: 367 cas (janvier 1989)
- Groupe 2: 70 cas (octobre 1989)
- Groupe 3: 31 cas (février 1990)
- Groupe 4: 17 cas (avril 1990)
- Groupe 5: 48 cas (août 1990)
- Groupe 6: 49 cas (Mise à jour Janvier 1991)
- Groupe 7: 31 cas (juin 1991)
- Groupe 8: 86 cas (novembre 1991)

total: 699 points (à partir de la base de données donné le 15 Juillet 1992)  
Notez que les résultats résumés ci-dessus dans l'utilisation passée référer à un ensemble de

données de la taille 369, tandis que le groupe ne dispose que de 1 367 cas. Parce qu'il contenait à l'origine 369 cas; 2 ont été retirés. Les énoncés suivants résument l'évolution du Groupe de 1 jeu initial de données:

- \* Groupe 1: 367 points: 200B 167m (janvier 1989)
- \* Révisée le 10 janvier 1991: Remplacé zéro noyaux nus dans 1080185 & 1187805
- \* Révisée novembre 22,1991: Suppression 765878,4,5,9,7,10,10,10,3,8,1 aucune trace
- \*: Suppression 484201,2,7,8,8,4,3,10,3,4,1 zéro épithéliale
- \*: Changé 0-1 dans la zone 6 de l'échantillon 1219406
- \*: Changé 0-1 dans la zone 8 de l'échantillon suivant:
- \*: 1182404,2,3,1,1,1,2,0,1,1,1

Les information sur les attribues :

2. Exemple de numéro de code: numéro d'identification
2. Clump Epaisseur: 1-10
3. Uniformité de Cell Taille: 1 - 10
4. Uniformité de Cell Forme: 1-10
5. Adhésion marginale: 1-10
6. unique des cellules épithéliales Taille: 1 - 10
7. Bare noyaux: 1-10
8. chromatine Bland: 1-10
9. normale Nucléoles: 1-10
10. Mitoses: 1-10
11. Classe: (2 pour bénigne, 4 pour maligne)(pas dangereux, dangereux)

La Source:

- \*créateur: Dr William H. Wolberg (médecin)  
Université de Wisconsin Hôpitaux  
Madison, Wisconsin, États-Unis
- \*donateur: Olvi Mangasarian (Mangasarian '@' cs.wisc.edu)  
Reçu par David W. Aha (aha '@' cs.jhu.edu)

**3.3-La BD du Centre Service de transfusion sanguine :**

Contient des données extraites du centre de service de transfusion sanguine dans la ville de Hsinchu Taïwan – c'est un problème de classification.

Ensemble de données Caractéristiques: multivariées.

Nombre d'instances: 748

Attribut Caractéristiques: réel

Nombre de Attributs: 5

Date de récoltés: 2008-10-03

Tâches associées: Classification

Valeurs manquantes? N / A

cette étude a adopté la base de données des donateurs de sang Centre Service de transfusion de Hsin-Chu Ville à Taiwan. Le centre transmet leur bus de service de transfusion sanguine à une université dans Hsin-Chu Ville pour recueillir le sang donné environ tous les trois mois. Pour construire un modèle de FRMTC, nous avons sélectionné 748 donateurs au hasard dans la base de données des donateurs.

Les information sur les attribues :

Étant donné le nom de la variable, type de variable, l'unité de mesure et une brève la description. Le "Blood Transfusion Service Center" est un problème de classification. L'ordre de cette liste correspond à l'ordre de chiffres le long des rangées de la base de données.

R (Récence - mois depuis la dernière donation),

F (Fréquence - nombre total de dons),

M (monétaire - le sang total des dons en C.C.),

T (Temps - mois depuis le premier don), et

une variable binaire représentant se il / elle a fait don de sang dans Mars 2007 (1 se présenter à un don de sang; 0 signifie pas don de sang).

Variable Type de données Description de mesure min max signifie std

Récence Mois quantitatives entrée 0,03 74,4 9,74 8,07

Fréquence quantitative Times, entrée 1 50 5,51 5,84

C.C. quantitative monétaire entrée de sang 250 12500 1378,68 1459,83

Temps Mois quantitatives entrée 2,27 98,3 34,42 24,32

Qu'il / elle a fait don de sang dans Mars 2007 binaire 1 = oui 0 = pas de sortie 0 1 1 (24%) 0 (76%)

La Source:

Propriétaire d'origine et donateurs Prof. I-Cheng Yeh, Département de la gestion de l'information Chung-Hua University, Hsin Chu, Taiwan 30067, R.O.C.

e-mail: icyeh '@' chu.edu.tw

TEL: 886-3-5186511

Date de Don: Octobre 3, 2008

#### **4- Les critères de mesure des performances des algorithmes :**

Lorsqu'une personne interroge une base de données (que ce soit un logiciel documentaire ou un moteur de recherche), elle attend un nombre de réponses (sous forme de documents) supérieur ou égal à un. À partir de l'ensemble de réponses obtenues mis en regard de l'attente de l'utilisateur, on peut mesurer les performances de l'algorithme de recherche mis en œuvre

pour retrouver un document. Les critères de mesure des performances sont le rappel et la précision et aussi la f-mesure qui est la combinaison des deux précédentes.

Lorsque le système retourne une réponse par rapport à une donnée et une classe, deux choix s'offrent à lui :

- Le message **appartient** selon lui à la classe
- Le message **n'appartient pas** selon lui à la classe

En face de ces deux possibilités de réponses du système, nous avons les deux cas où :

- Le message **appartient** à la classe
- Le message **n'appartient pas** à la classe

Cela donne alors 4 cas possibles différents :

Nom du cas	Abréviation	Description
Vrai positif	VP	Le système trouve <b>à raison</b> le message comme <b>appartenant</b> à la classe
Faux positif	FP	Le système trouve <b>à tort</b> le message comme <b>appartenant</b> à la classe
Vrai négatif	VN	Le système trouve <b>à raison</b> le message comme <b>n'appartenant pas</b> à la classe
Faux négatif	FN	Le système trouve <b>à tort</b> le message comme <b>n'appartenant pas</b> à la classe

Et donc nous avons la précision, le rappel et la F-mesure d'une classe i donnés par les formules :

Nom	Formule	Description
Précision	$P = \frac{vp}{vp + fp}$	Proportion de solutions trouvées qui sont pertinentes. Mesure la capacité du système à refuser les solutions non-pertinentes
Rappel	$R = \frac{vp}{vp + fn}$	Proportion des solutions pertinentes qui sont trouvées. Mesure la capacité du système à donner toutes les solutions pertinentes.
F-mesure	$F = \frac{2PR}{P + R}$	Moyenne harmonique de la précision et du rappel. Mesure la capacité du système à donner toutes les solutions pertinentes et à refuser les autres

Si on prend par exemple Les données comme documents attribuées à une classe i la définition des Caractéristique sera comme suite.

#### 4.1-Le rappel :

Le rappel est défini par le nombre de documents (données) pertinents retrouvés au regard du nombre de documents pertinents que possède la base de données. Cela signifie que lorsque

l'utilisateur interroge la base il souhaite voir apparaître tous les documents qui pourraient répondre à son besoin d'information. Si cette adéquation entre le questionnement de l'utilisateur et le nombre de documents présentés est importante alors le taux de rappel est élevé. À l'inverse si le système possède de nombreux documents intéressants mais que ceux-ci n'apparaissent pas dans la liste des réponses, on parle de silence. Le silence s'oppose au rappel.

$$Rappel_i = \frac{\text{documents correctement attribués à la classe } i}{\text{nombre de documents appartenant à la classe } i} \quad (6)$$

En statistique, le rappel est appelé sensibilité.

#### 4.2-La Précision :

La précision est le nombre de documents ou données pertinents retrouvés rapporté au nombre de documents total proposé par le moteur de recherche pour une requête donnée.

Le principe est le suivant : quand un utilisateur interroge une base de données, il souhaite que les documents proposées en réponse à son interrogation correspondent à son attente. Tous les documents retournés superflus ou non pertinents constituent du bruit. La précision s'oppose à ce bruit documentaire. Si elle est élevée, cela signifie que peu de documents inutiles sont proposés par le système et que ce dernier peut être considéré comme "précis". On calcule la précision avec la formule suivante :

$$Précision_i = \frac{\text{documents correctement attribués à la classe } i}{\text{nombre de documents attribués à la classe } i} \quad (7)$$

#### 4.3-La F-mesure :

C'est la moyenne harmonique de la précision et du rappel. Qui mesure la capacité du système à donner toutes les solutions pertinentes et à refuser les autres

Une mesure populaire qui combine la précision et le rappel est leur pondération, nommée F-mesure (soit *F-measure* en anglais) ou F-score :

$$F = \frac{2 \cdot (\text{précision} \cdot \text{rappel})}{(\text{précision} + \text{rappel})} \quad (8)$$

### 5- Méthode d'échantillonnage :

La méthode consiste à choisir comment notre testes va se dérouler c.a.d. Le résultat de la classification choisi sera tester selon des options de tests, Il ya quatre modes de test connue:

### **5.1- Utiliser l'ensemble d'apprentissage (using training set) :**

Le classificateur est évalué sur la façon dont il prédit la classe des cas où il a été formé sur.

### **5.2- Ensembles de testes fournis (supplied test set) :**

Le classificateur est évalué sur la façon dont il prédit la classe d'un ensemble d'instances chargées à partir d'un fichier.

### **5.3- La validation croisée ( cross validation) :**

La plus utilisée, le classificateur est évalué par validation croisée, à l'aide du nombre de plis qui sont entrés dans le champ de texte Folds par défaut on trouve 10 champs c.à.d. divisé la base de données en 10 champs ensuite tester.

### **5.4- Pourcentage scission (Pourcentage split) :**

Le classificateur est évalué de la façon dont il prévoit un certain pourcentage de données qui a lieu pour le test. La quantité de données détenues sur dépend de la valeur saisie dans le champ%, par défaut on trouve 66 % pour créer le modèle, le reste 34 % est pour le test.

## **6-Outil de travail :**

Il existe plusieurs Application et logiciel qui sont spécialisés dans l'analyse et l'extraction des connaissances à partir des données informatisées. Ce sont des logiciels qui aident l'analyste en exploration de données à trouver des motifs remarquables et intéressants.

On trouve *KNIME* (prononcer NAÏM), acronyme de *Konstanz Information Miner*.

Et aussi *R* qui est un langage et un environnement permettant d'effectuer des calculs statistiques et de créer leurs graphiques. *Tanagra* aussi qui est un logiciel libre d'exploration de données développé sous la direction de Ricco Rakotomalala du laboratoire ERIC de l'Université Lumière Lyon.

Nous ce qui nous intéressent c'est bien le logiciel Weka parce-qu'il est plus utilisé dans le domaine de la fouille de données, créé par l'université de Waikato (Nouvelle-Zélande). C'est une collection d'algorithmes d'apprentissage automatique mis en place pour effectuer des tâches d'exploration de données. Les algorithmes peuvent soit être appliqués directement à un ensemble de données soit être appelés directement par un code Java développé par une équipe informatique indépendante par exemple. *Weka* contient des outils pour les prétraitements des données, la classification, la régression, le *clustering*, les règles d'association et la visualisation. Il est également bien adapté au développement de nouveaux schémas pour l'apprentissage automatique. C'est un logiciel *open source* publié sous la LGPL [23].

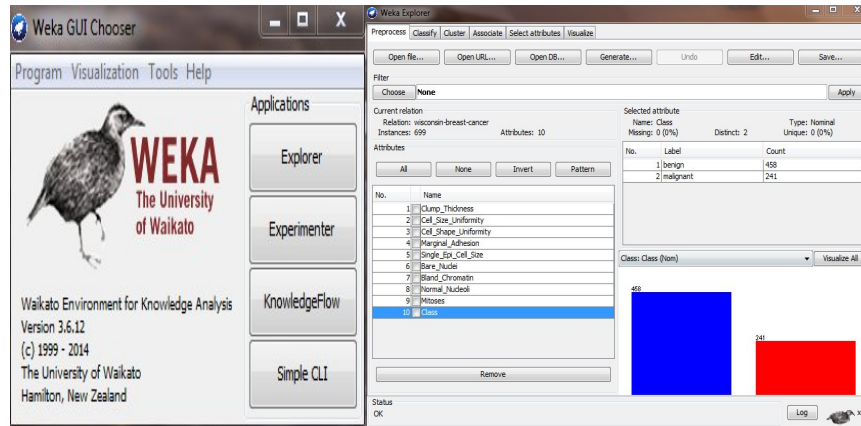


Figure 4.1-L'interface du logiciel Weka [23].

## 7-Expérimentations Sur La base de données IRIS:

### 7.1-IRIS Base référence :

#### 7.1.1- OverSampling (SMOTE) :

Voici ci-dessous le tableau d'une série de tests et résultats sur la base de données Iris qui donnera avec les 3 algorithmes (Naive Bayes, KNN, Arbre de décision) et l'application de la méthode SMOTE sur cette base avec les paramètres suivantes :

- class = 0 (la première classe) : la classe où on veut appliquer SMOTE, 0 = la première.
- Nearest Neighbors = 5 puis 10 puis 15 : Le nombre de voisins à utiliser
- Pourcentage = 100 : Le pourcentage des instances de SMOTE à créer.
- random seed = 1 : la semence utilisée pour l'échantillonnage aléatoire

Pour le nombre de voisins pour ce cas ça sera 100 pour la première classe et les autres ne changent pas. Note : neighbors = neighbors, AVG= moyenne des F-mesure, ACC= accuracy.

	F-mesure Setosa	F-mesure Versicolore	F-mesure virginica	AVG	ACC
Origine Naive Byes	1	0.941	0.939	0.960	<b>96 %</b>
Origine KNN	1	0.931	0.929	0.953	95.333 %
Origine Arbre Decision	0.990	0.940	0.950	0.960	<b>96 %</b>
NB SMOTE 5 neighbors	1	0.931	0.929	0.953	96.500 %
NB SMOTE 10 neighbors	1	0.941	0.939	0.960	97 %
NB SMOTE 15 neighbors	1	0.931	0.929	0.953	96.500 %
KNN SMOTE 5 neighbors	1	0.920	0.920	0.946	96%
KNN SMOTE 10 neighbors	1	0.920	0.920	0.946	96 %
KNN SMOTE 15 neighbors	1	0.931	0.929	0.953	96.500 %
DT SMOTE 5 neighbors	0.995	0.940	0.950	<b>0.961</b>	<b>97 %</b>
DT SMOTE 10 neighbors	0.995	0.940	0.950	<b>0.961</b>	<b>97 %</b>
DT SMOTE 15 neighbors	0.995	0.940	0.950	<b>0.961</b>	<b>97 %</b>

Table 4.1- résultats avec et sans SMOTE avec 3 classifieurs sur IRIS part 1.

On remarque ici dans le cas origine avant d'appliquer le filtre SMOTE que les algorithmes Naive Bayes et Les Arbres de décision donne des bon résultats, leur pourcentages sont presque égaux, mais on voit que dans la F-mesure de la classe Virginica est élevé de environ 0.02 avec les arbres de décision par rapport au Naive bayes qui dépasse de juste 0.001 dans les deux premières classes, donc dans ce cas les Arbres de décision sont plus performant que autres classifieur.

Si on utilise SMOTE et quand on compare avec les nombres d'instances utilisé pour crée une nouvelle instance on voit l'ors de l'utilisation de 10 voisins on a une meilleur précision de d'une 97 % pour le Naive Bayes. Pour le KNN ca augmente jusqu'à 96.5 % d'accuracy, et la F-mesure presque ne change pas, et si on utilise 15 voisin pour crée une nouvelle instance pour les arbres de décision ça donne toujours une meilleur performance, et le changement des nombres de voisins n'affecte pas les résultats alors dans notre cas ici les Arbres de décision sont les Plus performant avec et Sans SMOTE avec 97 % précision.

 Et maintenant on procède avec les paramètres SMOTE Suivantes on a :

- class = 0 (la première classe )
- Nearest Neighbors = 5
- Pourcentage = 100 puis 150 et ensuite 200
- random seed = 1

Dans ce cas le nombre pour 100 % de pourcentage est 100 instances de setosa et les autres ne changent pas et dans le cas où on applique SMOTE avec un paramètre de 150 ça donne 125 instances et les autres ne changent pas aussi parce qu'on applique que pour la première classe, en essayant de déséquilibrées les classes. Et enfin pour 200 % ça donne 150 instance de Setosa .

On rappelle que le fonctionnement de SMOTE et que si on choisit d'augmenter les instances de la classe minoritaire de 100 % dans ces paramètres, le nombre d'échantillons sera doublé. Avec 5 voisins, il prend chaque instance et explore les cinq voisins de cette dernière pour crée une nouvelle instance à partir de leurs façon d'étiqueter et distribuer les classes, Par contre SpreadSubSample qu'on va voir après, Supprime aléatoirement des données de la classe majoritaire, pour que les classes seront équilibrées.

Note :

prct = le pourcentage des instances créées par SMOTE,  
AVG = average la moyenne des F-mesure  
ACC = accuracy l'exactitude de la classification  
Origine = La base d'origine avant et sans modification.  
NB= Naive bayésien  
KNN= k plus proche voisin  
DT = arbre de décision

	F-mesure Setosa	F-mesure Versicolore	F-mesure virginica	AVG	ACC
Origine Naive Byes	1	0.941	0.939	0.960	<b>96 %</b>
Origine KNN	1	0.931	0.929	0.953	95.333 %
Origine Arbre Décision	0.990	0.940	0.950	0.960	<b>96 %</b>
NB SMOTE 100 prct	1	0.931	0.929	0.953	96.5 %
NB SMOTE 150 prct	1	0.920	0.920	0.946	96.44 %
NB SMOTE 200 prct	1	0.909	0.911	0.940	96.4 %
KNN SMOTE 100 prct	1	0.920	0.920	0.946	96%
KNN SMOTE 150 prct	1	0.920	0.920	0.946	96.444 %
KNN SMOTE 200 prct	1	0.931	0.929	0.953	97.200 %
DT SMOTE 100 prct	0.995	0.940	0.950	0.961	97 %
DT SMOTE 150 prct	0.996	0.922	0.929	0.949	96.444 %
DT SMOTE 200 prct	0.997	0.940	0.950	0.962	<b>97.600 %</b>

Table 4.2 - résultats avec et sans SMOTE avec 3 classifieur sur IRIS part 2.

Dans le cas de la classification de la base d'origine sans aucune modification, nous donnons que les arbres de décision sont les plus performants. Si on applique le filtre SMOTE avec différents pourcentages d'augmentation de la classe minoritaire, on aura que le Naive Bayes performe quand on utilise 100 % comme paramètre dans SMOTE avec 96.5 % Acc et 95.3 % Avg (moyenne des F-mesures). En ce qui concerne le K Nearest Neighbors, on remarque que à chaque fois quand on monte le pourcentage de SMOTE, on aura de meilleurs résultats, surtout pour 200 % qui donne 97.2 % Acc, même chose pour les Arbres de décision qui nous donnent des meilleurs résultats beaucoup plus dans le cas de 200 % (augmentation de la classe minoritaire) qui donne 97.6 % et 96.2 AVG. Dans ce cas, on conclut que les Arbres de décision sont les plus recommandés avec cette Base de données, mais il y a un seul inconvénient, c'est qu'il affecte les performances de la classe sur laquelle on a appliqué les méthodes.

### 7.1.2-UnderSampling ( SpreadSubSample) :

On ne peut pas appliquer cette méthode sur cette base parce que les classes de cette dernière sont égales entre elles, car le principe de cette technique est d'équilibrer les classes déséquilibrées en diminuant la classe majoritaire d'une façon à ce que ces nombres d'instance soient égaux au nombre des autres classes.

### 7.1.3-Hybride (SMOTE and SpreadSubSample):

La même chose pour cette méthode, on ne peut pas l'appliquer sur cette base car ses classes sont équilibrées.

## 7.2-IRIS base modifiée :

Pour rappel, on connaît que SMOTE crée des nouveaux exemples dans la classe minoritaire,

en générant synthétiquement plusieurs instances de la classe minoritaire. Précédemment on a eu des résultat de SMOTE appliquée sur la Base de Données IRIS de classes équilibrées tous de 50 instances, maintenant si on suppose artificiellement que IRIS a l'origine avait une classe minoritaire, supposant que c'est la classe Setosa, on supprime aléatoirement 25 instance de cette classe pour que le nombre de instance devient le demi que celui des autres classes afin d'appliquer SMOTE avec 100 % d'augmentation pour la classe devient équilibré de 50 instance égale aux autres Virginica et versicolor, on a après tests avec les classifieurs des nouveaux résultat, qui seront comparé aux précédents résultat quand la base était équilibré, cette technique est pour but de tester la performance de SMOTE et analysé les résultat procuré par ce dernier c.à.d. Es que les nouveaux échantillons crée par SMOTE se rapproche des instances origine supprimés dans notre cas ici ou non , le tableaux 4.3 éclairecis .

		F-mesure Setosa	F-mesure Versicolore	F-mesure virginica	AVG	ACC
<b>Base référence</b>	Origine Naive Byes	1	0.9410	0.9390	0.9600	96 %
	Origine KNN	1	0.9310	0.9290	0.9530	95.3330 %
	Origine Arbre Décision	0.9900	0.9400	0.9500	0.96	96 %
<b>Base modifiée</b>	modifiée Naive Byes	1	0.9500	0.9490	0.9663	96.6667 %
	Modifiée KNN	1	0.9500	0.9490	0.9663	96.6667 %
	Modifiée Arbre Décision	0.9900	0.9400	0.95	0.96	96 %

Table4.3 – Comparaison de Base Iris avant et après modification (SMOTE test).

On remarque ici qu'il n'ya pas un grand changement de résultats, il n ya pas de grandes différence sauf une légère écartassions avec l'algorithme KNN avec une moyenne de F-mesure de 0.953 dans le cas origine et 0.9663 dans le cas de IRIS modifiée, et aussi en voit que la F-mesure de la classe qu'on a modifié reste la même avec les trois algorithmes, donc on conclue que la Méthode SMOTE est efficace.et affecte positivement les données.

## 8- Expérimentations sur La Base de données breast cancer wisconsin:

### 8.1-Oversampling (SMOTE) :

Dans ce cas on applique la même série de tests précédents sur la base Breast Cancer qui donnera comme suite :

En premier lieu on utilise les paramètres suivants :

- class = 0 (la première classe)
- Nearest Neighbors = 5 puis 10 puis 15

-Pourcentage = 100  
-random seed = 1

Pour le nombre d'instance après l'application du filtre SMOTE sur les instances ça donnera que la classe Malignant aura 482 instance et la première benign 458 ne change pas par suite on aura :

	F-mesure Benign	F-mesure Malignant	AVG	Acc
Origine Naive Byes	0.9690	0.9440	0.9565	<b>95.9943 %</b>
Origine K NN	0.9630	0.9290	0.9460	95.1359 %
Origine Arbre Décision	0.9580	0.9210	0.9395	94.5637 %
NB SMOTE 5 neibrs	0.9620	0.9650	0.9635	96.3830 %
NB SMOTE 10 neibrs	0.9650	0.9670	0.9660	96.5957 %
NB SMOTE 15 neibrs	0.9650	0.9670	0.9660	96.5957 %
KNN SMOTE 5 neibrs	0.9780	0.9790	0.9785	97.8723 %
KNN SMOTE 10 neibrs	0.9780	0.9790	0.9785	97.8723 %
KNN SMOTE 15 neibrs	0.9800	<b>0.9820</b>	<b>0.9810</b>	<b>98.0851 %</b>
DT SMOTE 5 neibrs	0.9520	0.9560	0.9540	95.4255 %
DT SMOTE 10 neibrs	0.9540	0.9560	0.9550	95.5319 %
DT SMOTE 15 neibrs	0.9530	0.9570	0.9550	95.5319 %

Table 4.4- résultats de breast cancer part 1.

Dans le cas origine ici on remarque que les résultats de l'algorithme Naive Bayes sont supérieur aux autres résultats avec un pourcentage max de 95.994 % acc pas une grande différence vis-à-vis aux autres algorithmes

Si on applique SMOTE en augmentant à chaque fois le nombre de voisins a ce basé pour crée une nouvelle instance ça nous donnera des résultats plus rapprochées entre eux, Le Naive Bayes donne 96.59 % comme bonne classification (Accuracy), et le KNN donne 98.08% acc qui le meilleur résultat jusqu'à présent dans ce cas, et pour les Arbres de Décisions ça a données 95.53 % ACC. Et on voit ici aussi que la f-mesure de la classe minoritaire s'affecte légèrement avec la méthode SMOTE.

On conclut donc que Sans le filtre, Le Naive Bayes donne les meilleurs résultats, par contre si on applique le filtre SMOTE on trouve le KNN avec un pourcentage de classification correct 98.08 %, si on se base sur 15 voisins pour crée une instance qui est le plus performant.

🚦 Maintenant on procède avec les paramètres SMOTE suivantes on a :

-class = 0 (la première classe )  
-Nearest Neighbors = 5  
-Pourcentage = 100 puis 150 et ensuite 200

-random seed = 1

Dans ce cas le nombre pour 100 % de pourcentage serais 458 instance de benign et l'autre Malignant aura 482 avant elle contenait que 241. Et dans le cas où on applique SMOTE avec un pourcentage de 150 ca donnera 602 pour la deuxième classe la première ne change pas. Pour un pourcentage de 200 dans les paramètres de SMOTE on aura 723 instance pour Malignant et la première ne change pas 458.

	F-mesure Benign	F-mesure Maligant	AVG	ACC
Origine Naive Byes	0.9690	0.9440	0.9565	95.9943 %
Origine KNN	0.9630	0.9290	0.9460	95.1359 %
Origine Arbre Décision	0.9580	0.9210	0.9395	94.5637 %
NB SMOTE 100 prct	0.9620	0.9650	0.9635	96.383 %
NB SMOTE 150 prct	0.9520	0.9680	0.9600	96.3208 %
NB SMOTE 200 prct	0.9560	0.9720	0.9640	96.613 %
KNN SMOTE 100 prct	0.9780	0.9790	0.9785	97.8723 %
KNN SMOTE 150 prct	0.980	0.9850	<b>0.9825</b>	98.3019 %
KNN SMOTE 200 prct	0.9790	0.9870	0.9830	<b>98.3912 %</b>
DT SMOTE 100 neibrs	0.9520	0.9560	0.9540	95.4255 %
DT SMOTE 150 neibrs	0.9620	0.9720	0.9670	96.7925 %
DT SMOTE 200 neibrs	0.9580	0.9740	0.9660	96.7824 %

Table 4.5- résultats de breast cancer part 2.

Si on modifie les paramètres de pourcentage de SMOTE et on test on aura que le Naive Bayes procure 96.61 % comme taux de classification correct acc, et pour le KNN a chaque fois quant augmente le pourcentage de SMOTE, les résultats et les taux augmentes aussi, avec 98.39 %pour 200% augmentation de SMOTE, pour les Arbres de Décision on aura au maximum d'accuracy 96.79 %. Dans ce cas on conclut que le plus proche voisin (KNN) performe mieux que les autres Algorithmes si on utilise SMOTE.

## 8.2-UnderSampling (SpreadSubSample) :

Si on utilise la méthode Undersamplig qui se résume dans notre Logiciel de traitement Weka a SpresSubSample qui est un filtre pour diminuer la classe majoritaire dans le cas de données déséquilibrées, on applique les 3 algorithme avec cette méthode sur notre base de données ca donne :

On procède avec les paramètres de filtres suivants :

Adjust Weights = False : (poids d'instance castrés seront ajustés afin de maintenir le poids total par classe)

DistributionSpread = (1.0 : l'écart maximal de distribution de classe)

Max count =0 : ( le nombre maximum pour une valeur de classe)

RandomSeed=1 : (définit le nombre aléatoire pour le sous-échantillonnage)

On trouve :

	F-mesure Benign	F-mesure Maligant	AVG	ACC
Origine Naive Byes	0.9690	0.9440	0.9565	95.9943 %
Origine KNN	0.9630	0.9290	0.946	95.1359 %
Origine Arbre Décision	0.9580	0.9210	0.9395	94.5637 %
NB SpreadSubsample	0.9620	0.9630	0.9625	<b>96.2656 %</b>
KNN SpreadSubSample	0.9590	0.9580	0.9585	95.8506 %
DT SpreadSubSample	0.9470	0.9490	0.9480	94.8133 %

Table 4.6-Résultat Avec Undesampling pour Breast Cancer.

On rappelant dans le cas origine qui nous donne le Naive Bayes comme Algorithme performant. Si on applique la méthode Undersamplig (SpreadSubSample), on voit dans la table 4.6 que le Naive Bayes reste performant avec 96.26 % ACC de classification correct.

### 8.3-Hypyde (SMOTE and SpreadSubSample) :

Ici cette méthode consiste à appliquer en premier lieu le Filtre SMOTE avec un pourcentage calculé (dans cette base il sera 45 %) sur la base déséquilibré en augmentant la classe Minoritaire d'une façon que le nombre d'instance de cette dernière augmente avec la moitié de la différence entre les classe en forme d'origine. Pour ensuite appliquer le deuxième filtre sur la classe majoritaire pour la diminuer a un niveau du nouveau nombre de la classe minoritaire pour enfin appliquer la classification avec les trois algorithmes voir les figures :

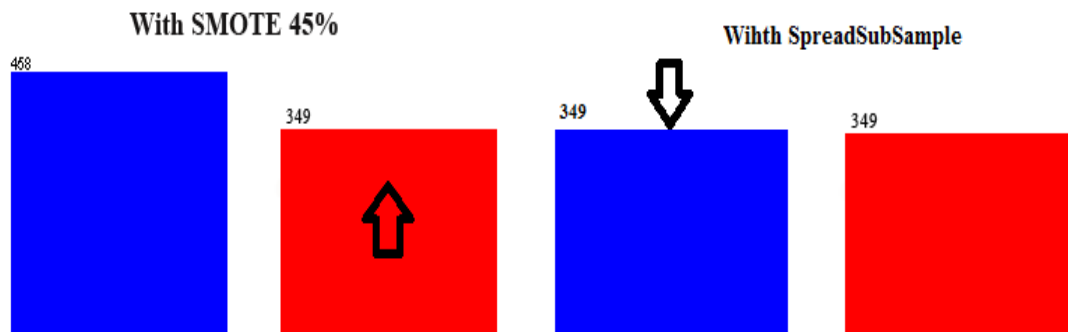


Figure 4.2 - le nombre d'instances après L'application de la hybride sur Breast Cancer.

Et dans ce cas on aura le tableau 4.7 des résultats suivants :

	F-mesure Benign	F-mesure Malignant	AVG	Acc
origine Naive Byes	0.969	0.944	0.9565	95.9943 %
Origine KNN	0.963	0.929	0.946	95.1359 %
Origine Arbre Décision	0.958	0.921	0.9395	94.5637 %
NB hybrid 45 %	0.964	0.964	0.964	96.4183 %
KNN Hybrid 45 %	0.977	0.977	<b>0.977</b>	<b>97.7077 %</b>
DT hybrid 45 %	0.957	0.957	0.957	95.702 %

Table 4.7- Résultat avec la méthode hybride pour breast cancer .

Dans ce cas avec la méthode hybride on remarque que le KNN donne les meilleur résultats avec 97.70 % Acc est une moyenne de F-mesure de 97.7 % AVG. On remarque ici aussi que la F-mesure de la classe minoritaire augmente ce qui est un bon avancement.

#### 8.4- La Meilleur Méthode pour Breast Cancer :

Pour aboutir à une meilleur méthode, on procède avec la comparaison en générale avec la F-mesure et l'accuracy c.a.d le taux de bonne classification des résultats précédente avec les trois méthodes, Puisque la f-mesure presque ne changes pas par rapport au cas origine, le taux élevé de bonne classification suffit à décider à partir de ces tableaux on trouve :

	Origine	SMOTE nbrs	SMOTE prct
<b>NB</b>	95.9943 %	96.5957 %	96.6130%
<b>KNN</b>	95.1359 %	98.0851 %	98.3912 %
<b>DT</b>	94.5637 %	95.5319 %	96.7925 %

Table 4.8-comparaison avec SMOTE

	Hybrid
<b>NB</b>	96.4180 %
<b>KNN</b>	97.7077 %
<b>DT</b>	96.7925 %

Table 4.9- comparaison avec Hybride

	SpreadSubSample
<b>NB</b>	96.2656 %
<b>KNN</b>	95.8506 %
<b>DT</b>	94.8133 %

Table 4.10- avec spreadsubsample

On remarque ici que les résultats augmentes par rapport au cas origine quand les bases de données étaient déséquilibrées, mais d'après les tableaux on remarque que la méthode SMOTE est la plus performante car elle augmente au maximum le taux de bonne classification.

#### 9- Travail Avec La Base de données Blood Transfusion :

### 9.1-Oversampling ( SMOTE) :


On applique les mêmes tests appliquées sur la base précédentes on a les tableaux suivant :  
Le nombre d'instance avec 100 % dans les paramètres de SMOTE devient 356 pour la classe minoritaire qui était 178. Pour 150 on a 445 et pour 200 ça donne 534 instances.

Note : neibrs = neighbors , AVG=\_moyenne des F-mesure, ACC= exactitude accuracy.

	F-mesure 1	F-mesure 2	AVG	Acc
origine Naive Byes	0.8500	0.2730	0.5615	75.1004 %
Origine K NN	0.8160	0.3290	0.5725	71.0843 %
Origine Arbre Décision	0.8600	0.4790	0.6695	<b>77.9116 %</b>
NB SMOTE 5 neibrs	0.7560	0.3660	0.5610	64.7568 %
NB SMOTE 10 neibrs	0.7610	0.3920	0.5765	65.7297 %
NB SMOTE 15 neibrs	0.7590	0.3710	0.5650	65.1892 %
KNN SMOTE 5 neibrs	0.7640	0.5830	0.6735	69.8378 %
KNN SMOTE 10 neibrs	0.7600	0.5820	0.6710	69.5135 %
KNN SMOTE 15 neibrs	0.7490	0.5500	0.6495	67.7838 %
DT SMOTE 5 neibrs	0.8010	0.6150	0.7080	73.7297 %
DT SMOTE 10 neibrs	0.7960	0.6390	0.7175	73.9459 %
DT SMOTE 15 neibrs	0.8030	0.6580	<b>0.7305</b>	<b>75.0270 %</b>

Table 4.11- Les résultats avec SMOTE sur Blood Transfusion part 1.

Dans le cas Origine (la base des résultats) sans aucun filtre on aura que l'Arbre de Décision est le meilleur Algorithme avec 77.91 % accuracy, Le Naive Bayes donne 65.72 % comme max accuracy si on se base sur 10 voisin pour créer une instance avec SMOTE, le KNN performe avec 69.83 % si on se base sur seulement 5 voisins, par contre les Arbres de Décision donnent un taux de 75.027 % acc avec 15 voisins. On voit ici que le f-mesure de la classe minoritaire augmente avec tous les algorithmes contrairement aux taux de bonne classification qui diminuent. De tout ça on conclut que les Arbres de Décision performe mieux que les autres classifier avec et sans utilisé le filtre SMOTE.

 Maintenant en modifiant le paramètre pourcentage on a :

	F-mesure 1	F-mesure 2	AVG	Acc
origine Naive Byes	0.8500	0.2730	0.5615	75.1004 %
Origine KNN	0.8160	0.3290	0.5725	71.0843 %
Origine Arbre Décision	0.8600	0.4790	0.6695	<b>77.9116 %</b>
NB SMOTE 100 prct	0.7560	0.3660	0.5610	64.7568 %
NB SMOTE 150 prct	0.7240	0.4210	0.5725	62.6233 %
NB SMOTE 200 prct	0.7040	0.5090	0.6065	63.1006 %
KNN SMOTE 100 prct	0.7640	0.5830	0.6735	69.8378 %
KNN SMOTE 150 prct	0.7470	0.6370	0.6920	70.217 %
KNN SMOTE 200 prct	0.7170	0.6730	0.6950	69.6283 %
DT SMOTE 100 neibrs	0.8010	0.6150	0.7080	<b>73.7297 %</b>
DT SMOTE 150 neibrs	0.7450	0.6340	0.6895	69.9211 %
DT SMOTE 200 neibrs	0.7390	0.7060	<b>0.7225</b>	72.3481 %

Table 4.12- Les résultats avec SMOTE sur Blood Transfusion part 2.

Le NB donne 64.75 % avec seulement 100 % d'augmentation de pourcentage dans les paramètres de SMOTE, mais ici on voit que à chaque fois qu'on augmenta la classe minoritaire et on test les résultats diminues.

Le KNN donne 70.21 % avec 150 % comme meilleur taux, mais les Arbres de décision restent les plus performants avec un taux de bonnes classifications de 73.72 %.

## 9.2- Undersampling (SpreadSubSampling) :

On procèdent avec les paramètres du filtre suivantes :

Adjust Weights = False.                      Max count =0.  
DistribustionSpread = 1.0                      RandomSeed=1

	F-mesure 1	F-mesure 2	AVG	Acc
origine Naive Byes	0.85	0.273	0.5615	75.1004 %
Origine KNN	0.816	0.329	0.5725	71.0843 %
Origine Arbre Décision	0.86	0.479	<b>0.6695</b>	<b>77.9116 %</b>
NB SpreadSubsample	0.67	0.522	0.596	60.9551 %
KNN SpreadSubSample	0.615	0.574	0.5945	59.5506 %
DT SpreadSubSample	0.68	0.597	<b>0.6385</b>	<b>64.3258 %</b>

Table 4.13 –Undersampling avec la base Blood Transfusion.

Pour la méthode SpreadSubSample on a toujours que les arbres de décision sont les plus performants avec 64.32 %.

### 9.3-Hybrid (SMOTE and SpreadSubSample ) :

Le pourcentage calculé ici dans cette base de donnée est 110 % puisque la classe majoritaire contient 569 instances et la minoritaire a 178 instances, donc si on applique SMOTE avec un paramètre de pourcentage de 100 % a la classe minoritaire il double les instance a 356. Et nous avons que la classe majoritaire a 569 la différence sera  $569 - 178 = 391$  le résultat divisé sur deux donnera 195 en l'ajustant a nombre minoritaire 178 sonnera 373. Il faut qu'elle atteigne 373 instance pour ça il faut un pourcentage de 110 % puisque 100 % nous donne 356. La figure suivante montre le résultat :

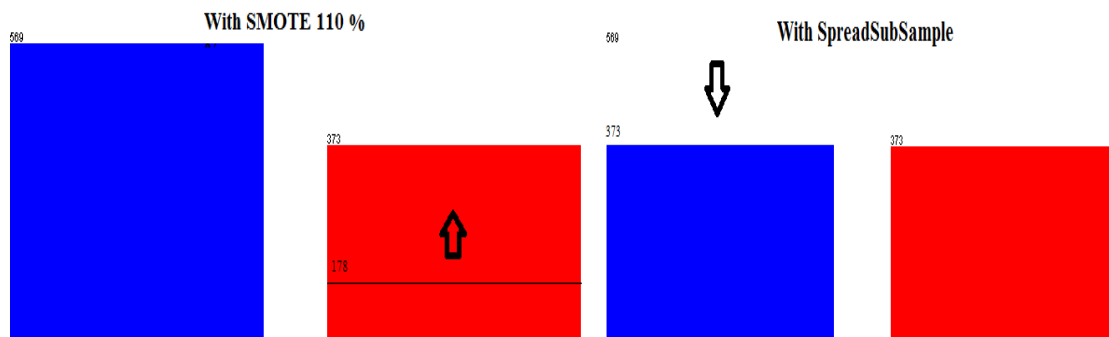


Figure 4.3- le nombre d'instances après L'application de l'hybride sur Blood Transfusion.

Et dans ce cas on a les tableaux 4.14 des résultats suivants :

	F-mesure 1	F-mesure 2	AVG	ACC
origine Naive Byes	0.85	0.273	0.5615	75.1004 %
Origine KNN	0.816	0.329	0.5725	71.0843 %
Origine Arbre Décision	0.86	0.479	0.6695	<b>77.9116 %</b>
NB hybrid 110 %	0.691	0.49	0.5905	61.5282 %
KNN Hybrid 110%	0.708	0.673	<b>0.6905</b>	<b>69.1689 %</b>
DT hybrid 110 %	0.684	0.689	0.6685	68.6327 %

Table 4.14 - Résultat avec la méthode hybride pour Blood Transfusion.

Pour la Hybride on remarque que le KNN à le grand taux de bonne classification ou classification correcte qui est de 69.1689 %, est à une grande f-mesure aussi des classes qui est de 69.05 %, on remarque aussi que la f-mesure augmente dans la classe minoritaire.

#### 9.4- Meilleur méthode pour Blood Transfusion :

On a les meilleurs résultats suivants :

	Origine		SMOTE Neibr		SMOTE Prct	
	ACC	AVG	ACC	AVG	ACC	AVG
<b>NB</b>	75.1004 %	0.5615	65.7297%	0.5765	64.7568%	0.561
<b>KNN</b>	71.0843 %	0.5725	69.5135%	0.671	70.217 %	0.692
<b>DT</b>	77.9116 %	0.6695	75.027 %	0.7305	73.7297%	0.708

Table 4.15- comparaison de méthode avec SMOTE.

	SpreadSubSample	
	ACC	AVG
<b>NB</b>	60.9551 %	0.596
<b>KNN</b>	59.5506 %	0.5945
<b>DT</b>	64.3258 %	0.6385

Table 4.16-undersamplig comparaison

	Hybride	
	ACC	AVG
<b>NB</b>	61.5282 %	0.5905
<b>KNN</b>	69.1689 %	0.6905
<b>DT</b>	68.6327 %	0.6685

Table 4.17-Hybrid comparaison

On remarque avec cette base de données le taux de bonne classification diminue si on applique les méthodes précédentes par rapport à l'origine des résultats, donc dans ce cas on compare avec la moyenne des F-mesure Average ou AVG.

Avec la moyenne AVG on trouve aussi que la méthode SMOTE performe et obtient des pourcentages plus élevées que ceux des autres méthodes et surtout avec les Arbres de décision donc on conclut que c'est la meilleure méthode conseillé pour travailler avec les bases de données dans le domaine de la médecine.

## 10- Conclusion :

Dans ce chapitre, on a étudié le cas de biais des données, Dans notre cas on a choisi les trois algorithmes les plus utilisées au paravent, qui sont le Naive Bayes, le plus proches voisin et les Arbres de Décisions. On a vu que les algorithmes de classification sont souvent biaisées en faveur de la classe majoritaire dans le cas déséquilibré, c'est pour ça qu'on a utilisés la méthode Sampling (échantillonnages) qui se devise en trois méthodes, le Sur-échantillonnage avec SMOTE qui s'est avéré la meilleur technique et la plus performantes d'après les résultat avec les trois algorithmes et surtout avec l'arbre de décision qui performe mieux que les

autres , car SMOTE donne des meilleurs résultats soit avec le taux de bonne classification ou la f-mesure, contrairement au sous-échantillonnage qui n'est pas conseiller parce que on risque de perdre des données utiles , et les résultat on montrer que le taux diminue avec cette méthode, même chose avec la Hybride qui est la combinaison des deux méthodes précédentes, ces taux et F-mesure sont faible par rapport à SMOTE. Dans ce cas et dans ce domaine de la médecine, l'algorithme Arbre de décision est la méthode oversampling avec SMOTE sont les plus conseillées pour classifier les données déséquilibrées dans le domaine de la médecine.

## **Conclusion Générale**

## Conclusion Générale

On s'est intéresser dans ce mémoire à la problématique des données déséquilibrées qui est un problème très connue dans la classification, cette dernier consiste à examiner les caractéristiques d'un objet et lui attribuer une classe, dans notre cas lorsque on a par exemple deux classe déséquilibrées , une classe majoritaire avec un large écart avec la deuxième qui sera minoritaire, En utilisant un algorithme, la classification sera biaisées en faveur de la classe majoritaire dans ce cas déséquilibré, beaucoup de méthode est solution sont proposées pour ce genre de problème, dans notre cas on a choisi la méthode d'échantillonnage (sampling), dans notre cas dans cet mémoire on a utilisé trois technique pour équilibrer les données, la première est le sur échantillonnages qui nous permet d'augmenter la classe minoritaire avec des données artificiel, crée par le filtre SMOTE, on peut modifier ces paramètres en augmentant le nombre d'instances artificiels qu' on veut créés ou la performance de ces dernières, la deuxième et le sous-échantillonnage ou bien le filtre SpreadSubSample qui permet de diminuer la classe majoritaire jusqu'au niveaux du nombre d'instances de la classe minoritaire c.à.d. de supprimer aléatoirement des données cette méthode n'est pas trop conseiller parce qu'on risque de perdre des données importantes, la troisième consiste à combiner les deux méthodes précédentes ensuite tester avec des classifieurs, on a choisi les algorithmes de classification les plus utilisées avec les données déséquilibrées est parce qu'il ne sont pas utilisé auparavant sur le domaines qu'on a choisis, la biologie et la médecine qui sont des domaine sensibles, et parce qu'ils ne sont pas tester auparavant sur les base de données choisis avec les méthodes mentionnées précédemment .

Les résultats obtenues par une série de tests ont montrés que la majorité des cas si on prend la premier base de donnée IRIS qui est la plus équilibrées de tous les bases, que nous avons pris comme base a se référencé, affirment que les arbres de décision donnent toujours une meilleur performance avec SMOTE soit en modifiant son paramètre pourcentage ou le nombre de voisin a se basés pour créer une instance. Concernant la base Breast canser qui est une base déséquilibrée, ses résultat ont montré que le classifieur Naive Bayes performe dans le cas origine avant d'appliquer un filtre, et le KNN si on applique SMOTE soit en modifiant ces nombres de voisins ou le pourcentage d'augmentation de la classe minoritaire et on a vu que le Naive Bayes performe avec SpreadSubSample et la hybride donne KNN comme meilleur classifieur pour cette base, on a conclu que SMOTE est la meilleur méthode pour breast cancer wisconsin. Pour la base de donnée Blood Transfusion on a eu que les arbres de décision les plus performants qu'aux autres algorithmes pour les deux méthodes SMOTE et SpreadSubSample, et pour la Hybride on a trouvé le KNN, et pour meilleur méthode on a conclu avec SMOTE, car les résultats par rapport à la base avec cette méthode reste robuste, surtout avec la classe minoritaire sur laquelle on a ajouté de nouveaux instance pour l'équilibrée.

En raison du temps manquant, j'ai pas pu approfondir mes recherches avec les données déséquilibrées, en essayant des nouveaux bases de données dans ce domaine médicale avec autre méthodes, ou autre algorithmes comme le SVM ou les réseaux neurones, ça reste à tester au future, si j'aurais l'occasion de poursuivre mon doctorat avec d'autre domaines et de nouvelle bases réels utilisées à l'échelle quotidien, ou utilisé l'apprentissage non supervisé.

## Bibliographie

- [1] Nitesh V. Chawla, Aleksandar Lazarevic, Lawrence O. Hall, Kevin Bowyer. "SMOTEBoost: Improving Prediction of the Minority Class in Boosting", usa.
- [2] Katarzyna Borowska , Magdalena Topczewska. "DATA PREPROCESSING IN THE CLASSIFICATION OF THE IMBALANCED DATA " Journal : Advances in Computer Science Research, vol. 11, pp. 31-46, 2014.
- [3] Shu Zhang, Samira Sadaoui & Malek Mouhoub "An Empirical Analysis of Imbalanced Data Classification" , Department of Computer Science, University of Regina, SK, Canada, janvier 2015.
- [4] Duman, E., Ekinci, Y., Tanriverdi, A . "comparaison des classificateurs alternatives pour la base de données Marketing: Le cas des ensembles de données déséquilibrées." Journal : Expert Systems with Applications, 39 (1), 48-53 (2012).
- [5] Bee Wah Yap, Khatijahhusna Abd Rani, Hezlin Aryani Abd Rahman, Simon Fong, Zuraida Khairudin, Nik Nairan Abdullah. "An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets". Faculty of Science and Technology, University of Macau, China
- [6] D.Lavanya and Dr.K.Usha Rani "ENSEMBLE DECISION TREE CLASSIFIER FOR BREAST CANCER DATA". Publié au International Journal of Information Technology Convergence and Services (IJITCS) Vol.2, No.1, a February 2012
- [7] Nitesh V. Chawla<sup>1</sup>, Kevin W. Bowyer<sup>2</sup>, Lawrence O. Hall<sup>1</sup>, W. Philip Kegelmeyer "SMOTE: Synthetic Minority Over-sampling Technique". Journal of Artificial Intelligence Research 16 (2002), 321 -- 357. Submitted 09/01; published 06/02.
- [8] Y. Sun, A.K.C. Wong, M.S. Kamel, "Classification of imbalanced data: a review", International Journal of Pattern Recognition and Artificial Intelligence 23 (4) (2009).
- [9] Dr. Abdlhamid DJEFFAL . "Cour d'introduction a la fouille de donnée". Université Mohamed khider Biskra.
- [10] Kubat & Matwin " Addressing The Curse Of Imbalanced Training Set: One-Sided Selection", Proc. 14th International Conference on Machine Learning, p179-186. 1997
- [11] Andrew Estabrooks, Taeho Jo and Nathalie Japkowicz: A Multiple Resampling Method for Learning from Imbalanced Data Sets. Computational Intelligence 20 (1) (2004) 18-36
- [12] Hongyu Guo, Herna L Viktor: "Learning from Imbalanced Data Sets with Boosting and

Data Generation: The DataBoost-IM Approach” . Sigkdd Explorations 6 (1) (2004) 30-39

[13] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao ” Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning”, (2005)

[14] David A. Cieslak, Nitesh V. Chawla “ Start Globally, Optimize Locally, Predict Globally: *Improving Performance on Imbalanced Data*”

[15] Show-Jane Yen et Yeu- Shi Lee “Cluster-based under-sampling approaches for imbalanced data distributions”, Journal : Expert Systems with Application : An international journal Volume 36 Issue 3, April, 2009.

## Liens Internet

[16] [http://ieeexplore.ieee.org/xpl/login.jsp?reload=true&tp=&arnumber=4583030&url=http%3A%2F%2Fieeexplore.ieee.org%2Fexpls%2Fabs\\_all.jsp%3Farnumber%3D4583030](http://ieeexplore.ieee.org/xpl/login.jsp?reload=true&tp=&arnumber=4583030&url=http%3A%2F%2Fieeexplore.ieee.org%2Fexpls%2Fabs_all.jsp%3Farnumber%3D4583030)

[17] <http://sourceforge.net/projects/weka/files/weka-3-6/3.6.12/>

[18] <https://scaron.info/doc/intro-arbres-decision/>

[19] [http://en.wikipedia.org/wiki/Naive\\_Bayes\\_classifier](http://en.wikipedia.org/wiki/Naive_Bayes_classifier)

[20] <http://fr.wikipedia.org/wiki/K-moyennes>

[21] [http://fr.wikipedia.org/wiki/R%C3%A8gle\\_d%27association](http://fr.wikipedia.org/wiki/R%C3%A8gle_d%27association) Consulté en mai 2015.

[22] [http://www.grappa.univ-lille3.fr/~torre/Recherche/Indiana/Documents/rapport\\_APitti.html](http://www.grappa.univ-lille3.fr/~torre/Recherche/Indiana/Documents/rapport_APitti.html)

[23] [http://en.wikipedia.org/wiki/Weka\\_\(machine\\_learning\)](http://en.wikipedia.org/wiki/Weka_(machine_learning)) Consulté en mai 2015.

[24] <http://www.grappa.univ-lille3.fr/polys/fouille/sortie005.html#toc15>

[25] <http://www.support-vector.net/>

[26] [http://fr.wikipedia.org/wiki/Apprentissage\\_automatique](http://fr.wikipedia.org/wiki/Apprentissage_automatique)

[27] <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.467.3090>

[28] [http://link.springer.com/chapter/10.1007%2F978-3-540-77046-6\\_43](http://link.springer.com/chapter/10.1007%2F978-3-540-77046-6_43)

[29] <http://www.simafore.com/blog/bid/111124/Decision-tree-accuracy-effect-of-unbalanced-data>

[30] [http://www.ieeefinalyearprojects.org/Data\\_mining\\_projects\\_in\\_java\\_dotnet.html](http://www.ieeefinalyearprojects.org/Data_mining_projects_in_java_dotnet.html)

## تلخيص

في هذه المذكرة، درسنا مشكلة تصنيف البيانات الغير متوازنة . في الواقع، هذا النوع من البيانات، يؤثر ويحط من أداء المصنفات. وقد تم اختيار قواعد البيانات من المجال الطبي. ثم تم اختبار ثلاث طرق لتحقيق التوازن بين البيانات (تضخيم البيانات (oversampling)، تصغير حجم البيانات (Undersampling)، تضخيم ثم تصغير (Hybrid) ). للقيام بعملية التصنيف، تم تنفيذ ثلاثة خوارزميات وتقييمها ( The naive bayes, Nearest neighbors, Decision trees ). أظهرت النتائج أن أفضل خوارزمية التصنيف مع هذه قواعد البيانات في المجال الطبي، هو شجرة القرار Decision tree ، الذي أدائه بشكل أفضل من الآخرين، وأفضل طريقة هي SMOTE OverSampling .

**الكلمات المفتاحية:** تصنيف موجه. البيانات الغير متوازنة ، تضخيم البيانات. SMOTE, تصغير حجم البيانات ، استخراج البيانات.

## Abstract

In this theses, we have studied the problem of unbalanced data classification. Indeed, this kind of data, the imbalanced data affects and degrades the performance of the classifiers. Databases from medical field were chosen. Three methods were tested to balance the data (oversampling, Undersampling, Hybrid). For the classification task, three algorithms were implemented and evaluated (the naive Bayes, nearest neighbors, and decision trees). The results show that the best classifier algorithm with this databases in the medical field, is the decision tree, which performs better than the others, and the best method is the oversampling technique with SMOTE.

**Keywords:** supervised classification, Imbalanced Data, SMOTE, oversampling, undersampling, Data mining.

## Résumé

Dans ce mémoire, on a étudié le problème de la classification des données non équilibrées. En effet, le biais de données affecte et dégrade les performances des classifieurs. On a choisis des bases de données du domaine médicales. Trois méthodes ont été testées pour équilibrer les données (suréchantillonnage, souséchantillonnage , hybride ). Pour la tache de classification, trois algorithmes ont été appliqués et évaluer (le naive bayésien, le plus proche voisin, les arbres de décisions). Les résultats montrent que le meilleur algorithme de classification de ces bases de données dans le domaine médicale est l'arbre de décision qui performe mieux que les autres, et la meilleur méthode d'échantillonnage est le sur échantillonnage avec SMOTE .

**Mots clés :** Classification supervisé, Imbalanced Data, SMOTE, oversampling, undersampling, Data Mining.