جامعة أبو بكر بلقايد – تلمسان

Université Abou Bakr Belkaïd de Tlemcen

Faculté de Technologie
Département de Génie Biomédical

MEMOIRE DE PROJET DE FIN D'ETUDES

Pour l'obtention du Diplôme de

MASTER en GENIE BIOMEDICAL

Spécialité : **Informatique Biomédicale**

Présenté par : MISSALTI Djahida et TARCHID Nour El Houda.

Vers une pondération des caractéristiques de données médicales.

Soutenu le 23 septembre 2017 devant le Jury

Mme.	MEKKIOUI Nawel	*MCB*	Université de Tlemcen	Président
Mr.	CHIKH Mohamed Amine	*Prof*	Université de Tlemcen	Encadreur
Mme.	BENCHAIB Yassmine	*MAA*	Université de Tlemcen	Examinateur
Melle.	*BELAROUSSI Sarra*	Doctorante	Université de Tlemcen	Co-Encadreur

Année universitaire 2016-2017

Remerciements

Nous tenons premièrement à prosterner remerciant Allah le tout puissant de nous avoir donné le courage et la patience au Long de la préparation de cette mémoire.

Nous tenons d'abord à exprimer nos profonds remerciements à Mr CHIKH M.A, Professeur à l'université de Tlemcen, en tant que directeur de mémoire pour l'attention qu'il a porté sur ce mémoire et ces judicieux conseils, ses orientations et la qualité des commentaires et suggestions.

Nous tenons aussi à remercier Melle BELAROUSSI SARRA ,
En tant que Co-encadreur, De l'attention qu'elle a porté sur notre travail dès le début de ce projet, ses conseils, son aide, sa présence, et son suivi durant notre projet ; grâce à son intérêt et sa compétence scientifique, elle a largement contribuer à la diversité des travaux réalisés et de l'amélioration du manuscrite.
Au terme de ce travail,
Il nous a met agréable de présenter notre remerciements Aussi à Madame Mikiwi d'avoir accepté de présider ce jury
Nos remerciements à Madame Benchaib qui a participé à examiner ce travail.
Enfin nous adressons nos remerciements les plus sincères à tous ceux qui ont contribués de près ou de loin à la concrétisation de ce travail

M.djahida & T.NourElHouda

Dédicace

Je dédie ce modeste travail en premier lieu à

A mes chers parents aucune dédicace ne saurait exprimer l'amour, l'estime, le dévouement et le respect que j'ai toujours pour vous.

J'adresse tout mon affection pour votre confiance, tendresse, amour que me portent et me guident tous les jours. Merci pour avoir fait de moi ce que je suis aujourd'hui.

A mes sœurs et mes frères Les mots ne suffisent pas pour exprimer l'attachement, l'amour que je vous porte.

A Mon français personne le plus précieux de ma vie pour son soutien moral et sons gentillesse sans égal,

Tous mes vœux de bonheur, de santé de réussite et une belle vie.

A toute Mes neveux et nièces ;

A toute mes amies.

Messalti djahida

Dédicace

Je dédie ce travail à :

Mes très chers parents et mon très cher frère ;

A toute ma famille ;

A tous mes chères amies.

Tarchid NourElHouda

Résumé

Une sélection objective avant la phase de classification est essentielle pour l'obtention des résultats attendus. Nous nous intéressons dans ce travail à la hiérarchisation des variables des données médicales pour cela nous avons utilisé la technique de sélection d'attributs par l'approche filtre. Elle qui se compose de méthodes univariées et multivariées. Nous avons réalisé différentes expérimentations sur des différentes données. Nous avons discuté et comparé les résultats obtenus avec quelques travaux de la littérature. La sélection d'attributs est considérée comme une étape primordiale avant la classification. Elle permet de représenter un sous ensemble de variables réduit à partir d'un grand ensemble volumineux de variables et d'éliminer les variables redondantes, non pertinentes ou incomplètes. Elle facilite aussi la visualisation des données et elle nous permet une meilleure compréhension. Elle réduit la complexité de données d'apprentissage qui va nous économiser aussi le temps de la phase d'apprentissage tout en améliorant la précision du classifieur

Mots clés

Réduction de dimension, sélection de variables, approche filter univarie, approche filter multivarie, classification supervisé.

Abstract

We are interested in the hierarchy of variables of medical data for that we used the technique of feature selection by the filter approach which consists of univariate and multivariate methods. we realize experimentations on different data and we have discussed and compared the obtained results with a others in littirature to distinguish which are the best methods compared to others.So, the feature selection is a step that plays an important role in classification.It allows to represent a subset of variables from a large set of variables and eliminate redundant, irrelevant or noisy parameters. It also facilitates the visualization of data and helps us to orient towards better understanding. It reduces the complexity of learning phase that will reduce the time of the learning algorithm by improving the accuracy of the classifier.

Keywords

Dimension reduction, features selection, filter multivariate, filter univariate, supervised classification

ملخص

إن اختيار المتغيرات ، قبل مباشرة عملية التصنيف له أهمية بالغة في تحقيق النتائج المنتظرة من أجل ذلك نحن مهتمين بتحديد أولويات المتغيرات للمعطيات الطبية منأجل هذا قمنا باستخدام تقنية اختيار من سمات بواسطة منهاج التصفية الذي ينقسم إلى نوعين نوع يعالج كل متغير على حدى وأخر يعالج مجموعة من المتغيرات في نفس الوقت من أجل القيام بعدة تجارب على معطيات متنوعة من أجل مناقشة النتائج المتحصل عليها إضافة إلى مقارنة ما بين التقنيات لتمييز التي تقوم بإعطاء نتائج جيدة بالنسبة للتقنيات الأخرى

مع أن تقنية اختيار المتغيرات تلعب دور مهما في عملية التصنيف والتي تسمح بتمثيل مجموعة فرعية من المتغيرات انطلاقا من المجموعة الكبيرة وتقوم بحذف المتغيرات المكررة والغير ملائمة والإضافية كما أنها تسهل عملية توضيع المتغيرات لتوجيهنا نحو الفهم الجيد . تقلل من تعقيد البيانات التي تؤدي إلى توفير الوقت وتحسين دقة المصنف.

كلمات مفتاحية
اختيار المتغيرات ,نموذج التصفية ,المصنف

Table des matières

Chapitre II: Méthode utilisé pour la sélection et classification

Chapitre III: Résultats et discussion

Liste des tableaux

Liste des figures

Glossaries

FS	Feature Selection.
CFS	Correlation Based Feature Selection.
FCBF	Correlated based filter
mRMR	Min-Redundancy, Max-relevance
IG	Information Gain.
MI	Mutuel Information.
KNN	K-nearest neighbor
SVM	support Vector Machine
CART	Classification and Regression Tree.

Introduction générale :

L'apparition des grandes bases de données dans le domaine de l'apprentissage a exigé une réduction d'entrée avant d'entamer la tâche de classification des données.

Le problème majeur de notre travail concerne des données de grande dimension et aussi la présence des données incomplètes. Donc dans ce mémoire nous nous intéressons aux techniques de sélection de caractéristiques qui permettent une meilleure compréhension de la modélisation d'un problème, réduction de la dimensionnalité.

Toutefois, la sélection des attributs consiste à chercher dans l'ensemble des variables explicatives disponibles un ensemble optimal des caractéristiques les plus pertinentes, significatives adaptées à la résolution d'un problème particulier.

Une phase de sélection de variables constitue alors une étape très pertinente pour la classification. Par exemple, le processus de sélection ne réduit pas seulement le temps d'apprentissage mais il aide aussi à comprendre les résultats fournis par le classifieur et à améliorer la précision de la classification,

Dans ce travail, nous avons réalisé des expérimentations sur différentes données par l'approche filtre après une étude comparative entre les trois méthodes de sélection à savoir de l'approche filtre, l'approche wrapper, l'approche intégré.

Ce travail de master a pour but de réduire le nombre de variables parmi lesquelles certaines pouvant être peu significatives, corrélées ou non pertinentes

La présentation du plan de travail :

Chapitre 1 : dans ce chapitre nous introduisons et nous exposons le contexte principal de la sélection existante dans la littérature nous définissons les différentes mesures de pertinence rencontrées ainsi qu'une étude comparative entre les différentes approches filtre, wrappers et l'approche intégré.

Chapitre 2 : ce chapitre est consacré à la présentation des concepts des méthodes de sélection avec une revue de quelques techniques de classification existants dans l'état de l'art.

Chapitre 3 : dans ce chapitre nous discutons les résultats des expérimentations obtenus, de l'approche proposée, obtenues sur des bases de données avec une description détaillée.

Nous terminons ce travail par une synthèse de nos différentes contributions, et nous donnons quelques perspectives qui peuvent donner suite à ces travaux.

Chapitre I

Principes généraux des méthodes de sélection

I. Introduction

Les classifieur des données automatique font partie des modèles d'analyse et de fouille de données les plus utilisées. Malgré leur succès en analyse de données exploratoire, les techniques de classification doivent s'adapter à des grands volumes de données. En effet de l'évolution des technologies de stockage, le volume de données disponibles a progressivement éclaté en nombre d'individus mais aussi en nombre de descripteurs. Nous sommes donc très souvent en face de problème de la réduction de la dimensionnalité. On distingue généralement deux types d'approche qui peuvent être utilisée: Transformation de caractéristiques (consiste à construire de nouveaux attributs à partir de l'ensemble des variables originales) et la sélection des variables (permet de ne conserver qu'un sous ensemble pertinent de variables).

Nous présentons dans ce chapitre en détail les techniques de sélection de variables.
Une revue de quelques méthodes de sélection ainsi que une étude comparative entre ces approches.

I.1 Réduction de la dimensionnalité:

Les méthodes de réduction de la dimensionnalité sont généralement classées en deux catégories :

•**L'extraction de caractéristiques** elle permet de à remplacer l'ensemble initial des données par un nouvel ensemble réduit, construit à partir un ensemble de départ [1].

•**La sélection de caractéristiques** elle regroupe les algorithmes permettant de sélectionner un sous-ensemble de données parmi un ensemble de départ, en utilisant divers critères et différentes méthodes.

Notre travail est basé sur la réduction par l'approche par sélection permet de mieux comprendre la modélisation d'un problème et de limiter les mesures qui permettent la résolution du problème.

I.2 définition d'une donnée:

Une donnée est une information enregistrée au sens des bases de données, que l'on nomme aussi« individu» ou «instance», elle est caractérisée par un ensemble d'attributs.

I.2.1 Les différents types d'attributs:

Un attribut peut être de nature qualitative ou quantative ça dépend de l'ensemble des valeurs qu'il peut prendre. Un attribut est qualitatif, sa valeur est d'un type défini en extension (une couleur, une marque de voiture …) sinon, l'attribut est de nature quantative : (un entier, un réel..).

I.2.2 Les différentes valeurs d'attributs: [4]

• **Attributs à valeurs ordinales**: une notion d'ordres impose sur les ordinaux mais il n'est pas possible de calculer directement des distances entre les valeurs ordinales, les opérations d'addition et de soustraction ne sont pas possibles.

•**Attributs à valeurs nominales:** les valeurs sont des symboles (des noms) dont aucune relation (ordre ou distance) existe entre les nominaux.

•**Attributs de type intervalles:** les intervalles impliquent une notion d'ordre, et les valeurs sont mesurées dans des unités spécifiques et fixées. La somme, la différence et le produit de deux intervalles ne sont pas possibles (car le point zéro n'existe pas).

• **Attributs de type rapport (ratio):**toutes les opérations mathématiques sont autorisées sur les attributs de ce type.

I.2.3 Principes de pertinence, non pertinence et redondance :

Avant de définir le processus de la sélection des variables, il est nécessaire de connaitre la différence entre les types des attributs puisqu'il nous permet une meilleure compréhension de la performance d'une stratégie de la sélection d'attributs.

I.2.3.1 pertinence d'un attribut:

La performance d'un algorithme d'apprentissage dépend fortement d'une donnée utilisée dans la tâche d'apprentissage. La présence de caractéristiques redondantes ou non pertinentes peut influencer sur l'efficacité du processus étudiée.

Dans la littérature, il existe plusieurs définitions de la pertinence d'une caractéristique, la plus connue est celle de (John et al. [1994], John[1997]). Selon cette définition, une caractéristique est classée comme étant très pertinente, peu pertinente et non pertinente.

Très pertinente : Une caractéristique fi est dite très pertinente si son absence provoque une réduction remarquable de la performance de la classification.

Peu pertinente : Une caractéristique fi est dite peu pertinente si elle n'est pas "très pertinente" et un sous-ensemble V tel que la performance de V ∪ {fi}soit meilleure par rapport à la performance de V.

Non pertinente : Les caractéristiques qui ne sont ni "peu pertinentes" ni "très pertinentes" représentent les caractéristiques non pertinentes. Ces caractéristiques seront en général supprimées de l'ensemble de caractéristiques de départ [1].

I.2.3.2 Redondance d'un attribut:

Elle est exprimée en termes de corrélation entre les attributs. Deux attributs sont redondants (*Entre eux*) si leurs valeurs sont complètement corrélées [2]. D'autre façon les attributs redondantes exprime les variables qui sont doublons c'est -à-dire qui apportent le même type d'information.

I.3 Principe de la sélection :

La sélection de caractéristiques est généralement définie comme un processus de recherche permettant de trouver un sous-ensemble "pertinent" de caractéristiques parmi celles de l'ensemble de départ. La notion de pertinence d'un sous- ensemble de caractéristiques dépend toujours des objectifs et des critères du système.

I.3.1 Intérêt de la sélection:

La sélection des variables est un axe de recherche très prometteur depuis plusieurs années. Elle concerne divers applications permettant de choisir un sous-ensemble optimal de variables pertinentes, à partir d'un ensemble original des variables, selon un certain critère de performance. L'importance de la sélection de variables est résumée comme suit :

- Utiliser un sous-ensemble plus petit permet d'améliorer la classification si l'on élimine les attributs qui sont source de bruit. Cela permet aussi une meilleure compréhension des phénomènes étudiés.

-Des petits sous-ensembles d'attributs permettent une meilleure généralisation des données en évitant le sur-apprentissage.

-Une fois que les meilleurs attributs sont identifiés, le temps d'apprentissage et d'exécution sont réduits et en conséquence l'apprentissage est moins coûteux.

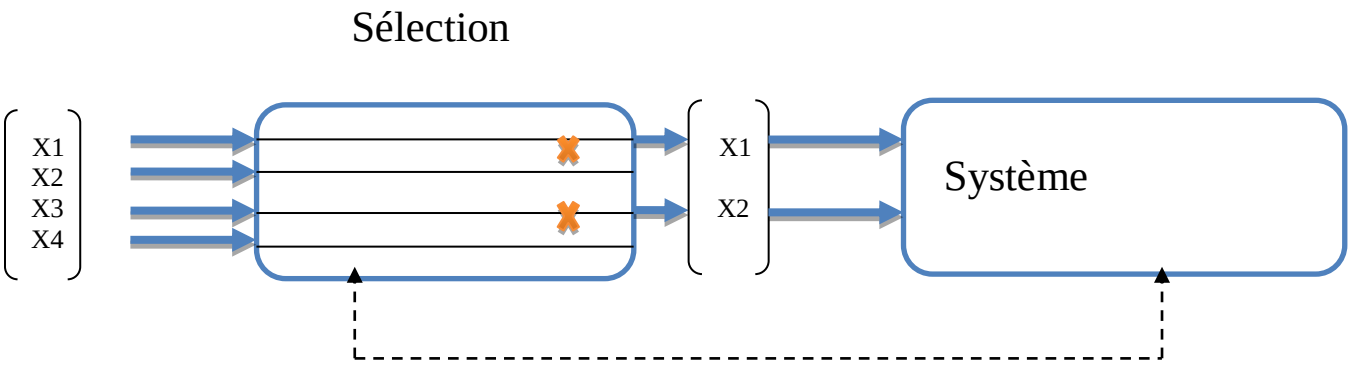

Figure I.1 .principe de sélection de variables.

Les bases de l'algorithme de sélection de variable est illustré dans ce qui suit:

I.1Algorithme générale de sélection

Input:
S-échantillon de données avec variable X,|X|=N J-
mesure d'évaluation à maximiser
GS -opérateur de successeur de génération
Output:
Solution-sous-ensemble des variables(pondérées)
L: = Point de départ(X);
Solution:={meilleur de L selon J};
Répéter
L:= Stratégie de recherche(L,GS(J),X);
X ': = {meilleur de L selon J};
Si J(X ') ≥J(Solution) ou (J(X') =J(Solution) et|X'| <|Solution|) puis Solution:=X';
Jusqu'à Stop (J, L)

I.3.2 Mesure de pertinence :

La mesure de pertinence des méthodes de sélection de variables sont basées sur des heuristiques calculant l'importance individuelle de chaque variable dans le modèle obtenu, ces heuristiques sont de différentes natures: statistique, probabiliste, information mutuelle ou celle qui mesure l'indépendance ou la vraisemblance entre les variables.

I.3.3 Processus global de la sélection de variables:

Une procédure de sélection d'attributs est généralement composée de quatre étapes illustrées dans la (Figure I.1) dans lequel nous trouvons les éléments clés suivants :

1) une procédure de génération de sous-ensembles de données.

2) Une fonction d'évaluation donnant la qualité de sous-ensembles de données.

3) Une condition d'arrêt.

4) Un processus de validation pour vérifier si l'objectif souhaité est atteint.

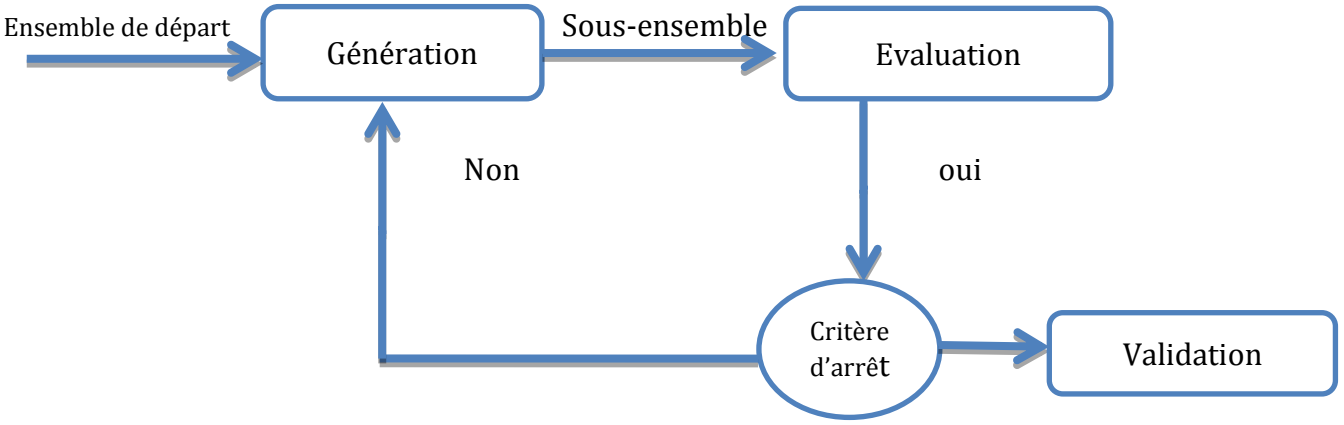

Figure I.2 La procédure de sélection.

I.3.3.1 La procédure de génération:

Elle permet à chaque itération de générer un sous-ensemble d'attributs qui sera évalué lors de la seconde étape de la procédure de sélection [3]. La procédure de génération est une recherche peut commencer par un ensemble vide des variables soit avec l'ensemble de tous les attributs c'est-à-dire les attributs sont itérativement ajoutés (Forward selection) ou retirés (Backward selection), soit avec un sous-ensemble d'attributs est créé de manière aléatoire à chaque itération (Random generation). Les méthodes de génération proposée dans l'été de l'art peuvent être classées en trois grandes approches de génération, la génération complète, la génération aléatoire et la génération séquentielle.

I.3.3.1.1 La génération complète:

Dans la procédure de génération complète, une recherche exhaustive est effectuée pour trouver l'ensemble optimal d'attributs sur tout l'espace des solutions possibles, plusieurs procédures de recherche heuristique sont proposées afin de réduire l'espace de recherche sans pour autant compromettre les chances de trouver le sous-ensemble optimal à évaluer.

I.3.3.1.2 La génération aléatoire :

Cette procédure de génération aléatoire n'évalue pas toutes les solutions possibles dans l'espace de recherche, contrairement aux procédures de génération complète. Un nombre maximal d'itérations est imposé afin de limiter le temps de calcul. L'avantage de cette procédure est qu'elle ne nécessite pas l'utilisation de fonction d'évaluation monotone. D'autre part, contrairement aux méthodes de génération complète dont la complexité est exponentielle vis-à-vis de la dimension initiale de l'espace d'attributs. Plusieurs méthodes sont basées sur les algorithmes génétiques (AG).

I.3.3.1.3 La génération séquentielle :

Le principe est l'ajout successif ou la suppression séquentielle (diminution / augmentation dans un espace de recherche) un ou plusieurs attributs pendant les itérations. On distingue alors deux approches de génération séquentielle:

- **L'approche de type Forward ou Ascendante:** cette approche commence par un sous ensemble vide d'attributs à chaque itération permet d'ajouter un ou plusieurs attributs.
- **L'approche de type Backward ou Descendante:** elle commence par un sous ensemble complet des attributs et à chaque itération permet de supprimer un ou plusieurs attributs.

Les algorithmes utilisant ces approches de génération sont caractérisés par leur simplicité et leur rapidité.

I.1.3.3.2 La fonction d'évaluation:

La fonction d'évaluation permet d'évaluer les attributs ou les sous-ensembles d'attributs générés à l'étape précédente. Elle est utilisée pour mesurer: La pertinence des attributs d'une manière individuelle, La pertinence des sous-ensembles d'attributs générés par l'une des différentes méthodes de génération présentées ci-dessous, Les méthodes de sélection sont classées en plusieurs approches(les filtres et les wrappers, intégrées), selon leur dépendance de l'algorithme d'apprentissage. Les méthodes filtres sont indépendantes de l'algorithme d'apprentissage, alors que les méthodes wrappers utilisent l'algorithme d'apprentissage comme une fonction d'évaluation selon différentes fonctions d'évaluation ont été proposée pour évaluer les attributs sont divisées en 5 catégories :

- **Les mesures d'erreur de classification** : L'attribut ou les sous-ensembles d'attributs sont évalués en fonction de la qualité de la performance obtenue.
- **Les mesures d'information**: permet de déterminer le gain d'information pour un attribut estimé à partir des probabilités.
- **Les mesures de consistance**: cherchent à évaluer si l'attribut (ou le sous-ensemble d'attributs) étudié contient les informations nécessaires à la discrimination des classes[2].
- **Les mesures de dépendance**: permet la mesure du degré de corrélation entre les attributs.
- **Les mesures de distance** : Les mesures de distance sont aussi nommées mesures de séparabilité, divergence ou de discrimination. Un attribut ou un sous ensemble d'attributs est sélectionnés il permet une meilleure séparabilité et cohérence des classes.

  En effet, le but est de :

  -maximiser la dispersion interclasses (séparabilité), afin que les points représentatifs des différentes classes forment dans l'espace d'attributs des nuages les plus séparés possibles les uns des autres.

  -minimiser la dispersion intra-classe(cohérence),afin que les nuages de points représentatifs de chaque classe soient les plus compacts possible [23].

1.3.3.3 Le critère d'arrêt :

C'est un choix souvent défini en fonction de la procédure de recherche et/ou du critère d'évaluation. Les critères d'arrêts les plus fréquents sont basées sur:

- **L'algorithme de génération:** on peut décider d'arrêter la recherche en fixant un seuil sur le nombre d'attributs à sélectionner ou sur le nombre d'itérations.
- **L'évaluation:** un seuil est fixé soit sur la fonction d'évaluation, soit entre deux itérations consécutives.
-

1.3.3.4. La procédure de validation:

La validation elle permet de tester la validité du sous-ensemble d'attributs sélectionnés en réalisant plusieurs tests il existe différentes approches de validation:

- **La méthode Holdout :** les données sont divisées en deux sous-ensembles : 2/3 pour l'ensemble d'apprentissage et1/3 pour l'ensemble de test.

- **La méthode validation croisée**: l'ensemble des données est divisées en N parties de tailles presque égales .Nous réalisons ainsi N fois la procédure de validation et chaque fois une des parties constitue l'ensemble test et les N-1 parties restantes pour former l'ensemble d'apprentissage.
- **La méthode de restitution :** l'ensemble d'apprentissage est utilisé comme ensemble de test.

En littérature, les méthodes de sélection des variables sont classées en trois catégories: filtre, wrapper et méthodes intégrées.

I.1.4. Les approches de sélection des variables:

I.1.4.1 Approche filtre:

L'approche filtre sélectionne un sous ensemble de variables en-prétraitement des données d'un modèle (l'étape de l'analyse des données) , le principe consiste à évaluer chaque attribut (cas univariées)pour lui assigner un score de pertinence ou bien évaluer un groupe d'attributs (cas multivariées) en lui assigner un score de pertinence. C'est-à-dire nous conduisent à une sélection des attributs les plus pertinents. Le processus de sélection est indépendant du processus de classification. Elle propose un sous ensemble de variables satisfaisant pour expliquer la structure des données qui se cachent et que le sous ensemble est indépendant de l'algorithme d'apprentissage choisi. De plus les procédures filtres sont moins couteuses en temps de calcul elles évitent les exécutions répétitives des algorithmes d'apprentissage sur différents sous ensemble de variables.

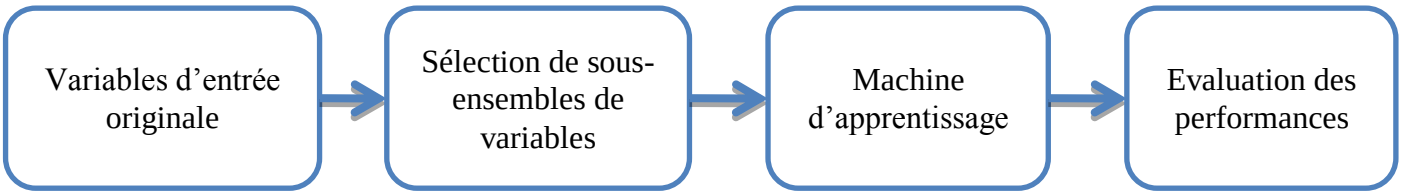

Figure I.3 Schéma générique des méthodes de filtre.

I.4.2 Approche wrappers :

Les wrappers ont été introduits par John et al. en 1994.Leur principe est de générer des sous- ensembles candidats et de les évaluer grâce à un algorithme de classification. Cette évaluation est faite par un calcul d'un score, par exemple un score d'un ensemble sera un compromis entre le nombre de variables éliminées et le taux de réussite de la classification sur un fichier de test. L'appel de l'algorithme de classification est fait plusieurs fois à chaque évaluation (c'est-à-dire à chaque sélection d'une variable, nous calculons le taux de classification pour juger la pertinence d'une caractéristique)car un mécanisme de validation croisé est fréquemment utilisé. Le principe de wrappers est de générer un sous ensemble bien adapté à l'algorithme de classification. Les taux de reconnaissance sont élevés car la sélection prend en compte le biais intrinsè que de l'algorithme de classification. Un autre avantage est sa simplicité conceptuelle.

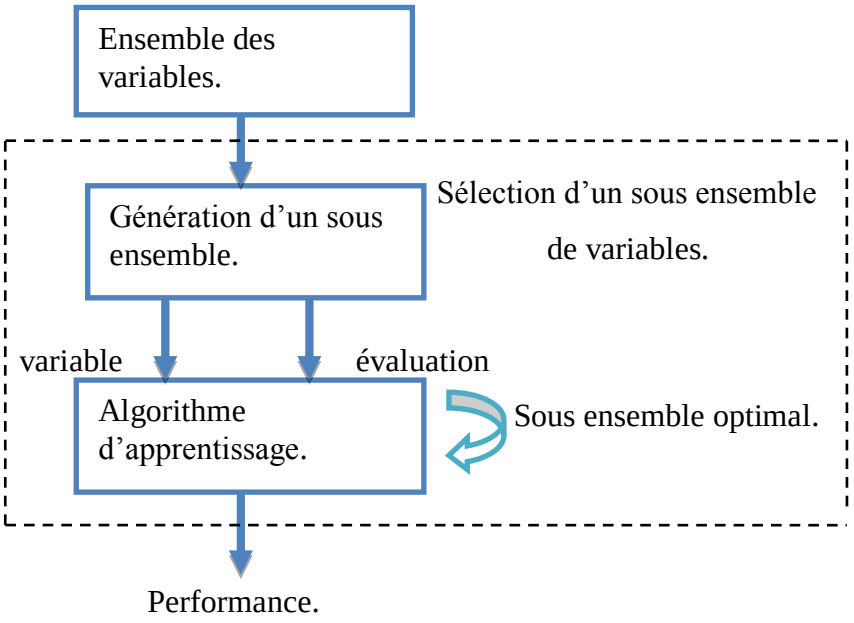

Figure I.3. Schéma générique des méthodes de wrappers.

I.4.3 Approche intégrée :

Dans les méthodes de sélection de type "wrapper" la base d'apprentissage est divisée en deux parties : une base d'apprentissage et une base de validation pour valider le sous-ensemble de caractéristiques sélectionné. En revanche, les méthodes intégrées peuvent se servir de tous les exemples d'apprentissage pour établir le système. Cela constitue un avantage qui peut améliorer les résultats. Un autre avantage de ces méthodes est leur plus grande rapidité par rapport aux approches "Wrapper" parce qu'elles s'évitent que le classifieur recommence de zéro pour chaque sous-ensemble de caractéristiques. Les méthodes intégrées apprennent quelles caractéristiques contribuent le mieux à la précision du modèle pendant la création du modèle.

1.5 Etude comparatives entre les trois approches:

	Avantage	Inconvénient
Filtre	▪ Une faible complexité de calcul assurant la vitesse au modèle. ▪ Indépendant de l'algorithme de classification.	▪ ignorent l'interaction avec le classifieur. ▪ Risque de sur-Apprentissage
Wrapper	▪ peut être utilisée lorsqu'on travaille avec un très grand nombre d'attributs car elle est de complexité raisonnable.	▪ Les wrappers sont généralement les plus lents par rapport les trois méthodes. ▪ les wrappers ont tendance à avoir une plus grande complexité queles filtres pour le paramétrage dans les classifieur. ▪ Wrapper considèrent l'apprentissage automatique
intégrée	▪ Optimalité au sens du critère de performance ▪ moins d'intensité de calcul que les méthodes wrappers.	▪ ont tendance à avoir une plus grande complexité que les filtres pour le paramétrage dans les classificateurs. ▪ Spécifique à une machine à apprendre.

Tableau I.1étude comparative entre les approches.

Selon le critère d'évaluation utilisé dans le processus de sélection d'attributs, Les approches "filter" utilisent une fonction d'évaluation basée sur les caractéristiques de l'ensemble des données, indépendamment de tout algorithme de classification, afin de sélectionner certains attributs ou sous-ensemble d'attributs (mesures d'information, mesures de consistance, mesures de dépendance et mesures de distance) Ces méthodes sont rapides, plus générales et moins coûteuses en temps de calcul, ce qui leur permet d'opérer plus facilement avec des bases de données de très grandes dimensions. Cependant, comme elles sont indépendantes de l'étape de classification, elles ne permettent pas de garantir que le meilleur taux de classification soit obtenu dans l'espace retenu.

I.6 Sélection des variables dans la littérature: [1]

Auteurs	**Titres**	**Méthodes et expériences :**	**Résultat**
Hafa amel 2012	Sélection de Variables Biologiques par l'approche FILTER	Ce papier implémente la méthode de sélection ReliefF,fisherMI, mrMr pour sélectionner les gènes les plus pertinents de la base cancer de colon avec les classifieurs K-NN. Les bases de données sont: cancer de colon	Plusieurs tests ont été réalisés avec différents nombre de variables sélectionnées à partir : -5 variables le taux est 80%, -10 variables, un tauxde85%, -Après la sélection de20 variables on remarque une chute du taux de classification avec90%.

.

Auteurs	Titres	Méthodes et expériences :	Résultat
YuhangWing, FilliaMakedon, 2004 [WM04].	Application Of ReliefF to selecting informative genes for cancer classification using microarray data.	Ce papier implémente la méthode de sélection ReliefF pour sélectionner les gènes les plus pertinents des diffèrent base de données avec les classifieurs SVM et K-NN. Les bases de données sont: ALL leukemia, MLL leukemia	Apres la sélection de 150 gènes pour chaque base, les taux de classification sont : SVM : -ALL: 99% - MLL: 97%K-NN: --- -ALL: 100%-MLL: 98%
Shousken Li, RuiXia, Chingquing Zong, ChuiRan Hueing, 2009. [LXZH09]	A frame work of feature selection methods for text categorization.	Ce papier se focalise sur la classification des textes, il se base sur la sélection des termes et leurs classifications, Il compare six méthodes : DF (document frequency) ,MI (mutuel information ,IG(information gain), CHI-2 (X2- test, BNS (bi- normal separation) et WLLR (weighted log likelihood ratio), ces méthodes ont été implémenté pour mesurer le score entre les termes et leurs catégories. Les expériences ont été testé sur un corpus de R et ers-21578 dénomméR2 et 20NG est une collection d'environ 20000 termes de20 documents.	-DFscore = 0,004 -MIscore= 0,870 Cequi montre que MI score a exprimé une bonne information sur la catégorie.

Auteurs	Titres	Méthodes et Expériences	Résultat
YiZhang, Chris Ding,TaoLi, 2008. [ZDL08]	Gene selection Algoriyhme by combing relifF and MRMR	Ce papier combine deux méthodes de sélection RelifF et MRMR ou la première consiste à trouver un ensemble de gènes et la seconde est appliquée explicitement pour réduire la redondance ; an d'avoir un ensemble de gènes compacte et efficace. La classification a été réalisé avec SVM et Naive bayes. Les bases sont: ALL (Aculte, Lynphplastic ,Leukimia), ARR(Arrhythmia), GCM, HBC, MLL (lekemia).	Les taux de classification sont évalués après la sélection de30 gènes pour chaque base.
Pablo A.Estévez, Michel Tesmer, ClandioA. Perez, Jacek M.Zurada, 2009. [ATAM0]	Normalised mutuel information Feature Selection.	Ce papier, propose une normalisation de la méthode de sélection MIen NMIFS et GAMIFS. Le premier est normalisation de mutuelle information (of feature selection) pour la sélection des variables, le second est Une hybridation entre les algorithmes génétiques et l'information mutuelle. Les bases de données utilisées sont des bases de données artificielles : Sonar, Breiman, Spam base, Madelon, Arcene.	-NMIFS: Nombre De variables sélectionnées est 11avecun taux de classification de86,36%. -GAMIFS: Nombre de variables sélectionnées est 11avecun taux de classification de 90,96%

Tableau I.2 Quelques travaux sur la sélection de variables.

I.7 Conclusion :

Dans ce chapitre, nous avons présenté le principe de la sélection de variables. il inclus les différentes étapes de la sélection de caractéristique qui permet de réduire la durée de l'apprentissage et de simplifier le modèle apprise après avoir cité les différentes méthodes «wrapper» «filtre» «intégrée» avec une étude comparative entre ces approches de sélection.

Chapitre II

Méthodes utilisé pour la sélection et la classification

II.1 Introduction :

Pendant la préparation des données pour la classification on repose essentiellement sur la recherche d'une représentation pertinente des variables d'entrée, la recherche qui se base sur une sélection de variable ; comporte trois approches qui ont été proposées dans La littérature. L'approche filtre elle repose sur des caractéristiques générales des données pour évaluer et sélectionner un sous-ensemble sans impliquer l'algorithme d'apprentissage *(plus rapide),* elle devienne plus efficace du point de temps calcul comme on a cité déjà dans le chapitre précédent. Dans ce chapitre nous présentons une revue sur la méthode de sélection d'approche filtre qu'ils sont divisés en méthodes multivariées et univariées, nous introduisons aussi dans la section suivante les différents modèles de classification proposés comme K plus proches voisins(Kppv), Machines à vecteurs de support (SVM) ,et les arbres de décision pour la phase de classification nous citons aussi quelque algorithmes utilisé dans la littérature .

II.2 Méthode de sélection : approche filter

Les méthodes de filtrage effectuent le processus de sélection des attributs comme une étape de prétraitement sans l'algorithme de classification. L'évaluation se fait indépendamment d'un classificateur. La plupart des approches filtres classent les variables selon leur pouvoir individuel de prédiction de la classe qui peut être estimé de divers moyens **[5]**. Dans cette section, nous présentons les méthodes de filtrage univariées et multivariées qui ont été utilisé dans notre étude.

II.2.1 Approche filtre multivariée :

Elles sont des méthodes filtre capables de trouver des relations entre les caractéristiques elles permettent d'évaluer tous les attributs à la fois **[6]**. Elles donnent en sortie un groupe d'attributs le plus pertinents à partir l'ensemble de départ et éliminent les attributs les moins significatifs

II.2.2 Approche filtre univariée :

Elles permettent d'évaluer chaque attribut indépendamment et de sélectionner un groupe de variables qui portent des informations similaires*(ou les mêmes)* sur la sortie **[6]**, c'est-à-dire

qu'il suffit d'utiliser un seul (*ou quelques-uns)* de ces variables. Elles n'ont pas besoin de reconnaître qu'une caractéristique est importante *(en combinaison avec d'autres variables).*

II.2.3Etude comparatives entre Multivariée et Univariée :

	Avantage	Inconvénient
Univariées	▪ rapide. ▪ évolutives. ▪ Considèrent chaque entité séparément.	▪ Ignore la dépendance des variables ▪ Ignore l'interaction avec le classifieur
Multivariés	▪ Modèle indépendant de fonctionnalité. ▪ Evolution de tous les attributs à la fois. ▪ Plus grande complexité informatique.	▪ Un peu plus lent que l'univariées. ▪ Moins évolutive que l'univariées. ▪ Ignore l'interaction avec le classifieur.

Tableau I.1 : Etude comparative entre Multivariées et Univariées.

Les méthodes univariées sont simples et elles permettent de classer les variables d'entrée en fonction d'une certaine utilité pour la prédiction, elles considèrent chaque entité séparément c'est-à-dire elles évaluent *(et habituellement classent)* une seule caractéristique mais peut provoquer des erreurs fatales, tandis que les méthodes multivariées sont plus résistantes aux erreurs lors de la sélection, mais elles sont beaucoup plus exigeantes sur le plan informatique.

II.2.4 Présentation de quelques méthodes de sélection :

Dans cette partie nous présentons les algorithmes de filtre univariée et multivariée utilisé dans notre projet de fin d'études

II.2.4.1 Méthodes Univariées :

A- Algorithme ReliefF :

C'est une méthode de filtrage univariée de sélection des caractéristiques ; elle est introduite sous le nom de Relief dans **[7]** puis elle est améliorée et elle est adaptée au cas multi-classes par Kononenko sous le nom de ReliefF. Son principe est de calculer une mesure globale de la pertinence des caractéristiques en tenant compte la différence des distances entre des exemples d'apprentissage choisis aléatoirement et leurs plus proches voisins de la même classe et de l'autre classe **[1]**. La simplicité, la facilité de la mise en œuvre ainsi que la précision même sur des données bruitées, représentent les avantages de cette méthode. En revanche, sa technique aléatoire ne peut pas garantir la cohérence des résultats lorsque nous appliquons plusieurs fois la méthode sur les mêmes données **[2]**. Cette méthode ne prend pas en compte la corrélation éventuelle entre les caractéristiques.

A. Algorithme de Sélection de ReliefF

---

1. Initialiser les poids
2. Tirer aléatoirement une donnée Xi
3. Trouver les K Plus Proche voisin de Xi ayant les mêmes étiquettes (hits).
4. Trouver les K Plus Proche voisin de Xi ayant une étiquette différente de la classe de Xi (misses)
5. Pour chaque caractéristique mettre à jours les poids.

   $$Wd=wd-\sum_{j=1}^{k} diff\left(\frac{x_i,d_i,hits_j}{m*k}\right)$$

   $$+\sum_{j=1}^{k}\left(\frac{p(c)}{(1-p(class(xi))}\right)\sum_{j=1}^{k}\left(\frac{diffp(x_i,d_i,misscs_i)}{m*k}\right)$$

6. la distance utilisée est définie par :

   $$Diff\ ((x_i, d_i, x_j)=\frac{|x_i d-x_j d|}{max(d)-min(d)}$$

La variable K pour le calcul des plus proches voisins des hits et misses fixées manuellement. Les attributs sélectionnés sont ceux qui ont une pondération supérieure à un seuil donné, l'attribut subira une diminution de sa qualité s'il y'a une différence entre ses valeurs pour les individus sélectionnées et un plus proche voisin de la même classe, et obtiendra une

augmentation de sa qualité s'il y'a une différence entre ses valeurs pour l'instance sélectionnée et un plus proche voisin d'une autre classe. Le sous-ensemble sélectionné par ReliefF contient soit toutes les copies, soit aucune copie de l'attribut redondant elle est incapable de détecter les attributs redondants **[2]**.

B- Algorithme Fisher :

Le score de Fisher est une méthode pour déterminer les caractéristiques les plus pertinentes pour la sélection. Il utilise des méthodes discriminatives et des modèles statistiques génératifs pour y parvenir.

L'algorithme de Fisher est défini comme suit :$P=\frac{(\overline{x_1}-\overline{x_2})^2}{s_1^2-s_2^2}$

Ou x_k *et* s_k^2 la moyenne et l'écart type de l'attribut pour la classe k = 1; 2. Un score important indique donc que les moyennes des 2 classes sont significativement différentes **[8]**.

C- Algorithme Index de Gini (IG) :

Gini index est un algorithme de sélection des variables multivariées supervisé du modèle de filtre pour mesurer la capacité de la caractéristique de distinguer les classes. Compte tenu des classes C, l'index Gini d'une fonction f peut être calculé car l'indice Gini peut prendre la valeur maximale de 0.5 pour une classification binaire. Les caractéristiques les plus pertinentes ont des valeurs d'index Gini plus petites. L'index Gini de chaque caractéristique est calculé de manière indépendante et les fonctions k supérieures avec le plus petit indice Gini sont sélectionnées. Gagnez de l'information, cela n'élimine pas non plus les variables redondantes**[11]**.

$$\text{GiniIndex(f)} = 1 - \sum_{i=1}^{c} [p(i|f]^2.$$

D-Algorithme Gain d'Information (GI) :

Le gain d'information est un algorithme de sélection de variables univariées supervisé du modèle de filtre qui est une mesure de la dépendance entre la fonctionnalité et l'étiquette de classe. C'est l'une des techniques de sélection des attributs les plus puissantes et il est facile de

calculer et d'interpréter facilement. Gain d'information (IG) d'une caractéristique X et les étiquettes de classe Y sont calculées comme**[11]**

$$IG(x,y)=H(y)+H(x)-H(x|y).$$

Entropie (H) est une mesure de l'incertitude associée à une variable aléatoire. H (X) et H (X / Y) est l'entropie de X et l'entropie de X après observation de Y, respectivement.

$$H = \sum_i p(x_i) log_2(px_i)).$$

La valeur maximale du gain d'information est 1. Une caractéristique avec un gain d'information élevé est pertinente. Le gain d'information est évalué de manière indépendante pour chaque fonctionnalité et les fonctionnalités avec les valeurs top k sont sélectionnées comme caractéristiques pertinentes. Cet algorithme de sélection de fonctionnalité n'élimine pas les fonctionnalités redondantes**[11]**.

$$H(x|y)=-\sum_i p(y_j) \sum_i p(x_i|y_i) log_2 (p(x_i|y_j)).$$

E- Algorithme Chi Square: [9]

L'approche Chi Square évalue les variables individuellement en mesurant leur statistique chisquare. L'algorithme fournit un score qui suit une distribution de chisquare avec l'objectif de classer l'ensemble des caractéristiques d'entrée.

Cette approche est largement utilisée mais elle ne prend pas en compte l'interaction entre les variables. Si l'on suppose que la variable de classe est binaire, la valeur chisquare pour marquer l'appartenance de la variable v à la classe k est évaluée comme suit :

$$X^{2(D,k,v)}=\sum_{i=1}^{N}\left[\frac{(n_{i+}-\mu_{i+})^2}{\mu_{i+}}+\frac{(n_{i-}-\mu_{i-})^2}{\mu_{i-}}\right].$$

Où D est le jeu de données considéré, N est le nombre des variables d'entrée et le nombre des échantillons qui ont une classe positive pour la variable i et enfin n_{i+} représente la valeur attendue alors s'il existe une relation entre v et k.

En statistique, l'algorithme chisquare est utilisé pour vérifier si deux événements sont indépendants. Dans la sélection des caractéristiques, la statistique chisquare effectue un test d'hypothèse sur la distribution de la classe, se rapporte à la mesure de la variable considérée; L'hypothèse nulle représente une absence de corrélation.

F- Algorithme Kruskal-Wallis : [9]

Le test de Kruskal Wallis est une méthode non paramétrique *(Une méthode d'analyse, dans laquelle il n'y a aucune hypothèse quant à la distribution des données)* basée sur les rangs pour comparer les médianes de population parmi les groupes. La première étape consiste à classer ensemble tous les points de données entre tous les groupes. Et la mesure peut être formulée comme suit:

$$K(N\text{-}1)= \quad \frac{\sum_{i=1}^{g} ni(\bar{r}_i-\bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{ni} ni(\bar{r}_{ij}-\bar{r})^2} \cdot$$

Dans l'équation, nous avons: N est le nombre total d'observations dans tous les groupes Ni est le nombre d'observations dans le groupe 'i' Rij est le rang d'observation 'j' dans le groupe 'i'

$$\bar{r}_i \frac{\sum_{j=1}^{ni} r_{ij}}{n_i}$$

Ri est le rang moyen de toutes les observations, c'est-à-dire la somme de N nombres naturels / N

$$\frac{N+1}{2}$$

G-Algorithms - Spectral Feature Selection (spectrum): [9]

spectrum est une extension pour Laplacian score tenu compte de la matrice d'affinité K, du degré de matrice D et de la matrice La plachienne normalisée L, trois critères d'évaluation sont proposés pour mesurer la pertinence des caractéristiques de la manière suivante:

$$SC_{S,1}(f_i)=\widehat{f_i}\,\gamma(\mathcal{L})\widehat{f_i}=\sum_{j=1}^{n}\alpha_j^2\,\gamma(\lambda_i)$$

$$SC_{S,2}(f_i)=\frac{\widehat{f_i}\,\gamma(\mathcal{L})\widehat{f_i}}{1-(\widehat{f_i}\xi_1)^2}=\frac{\sum_{j=1}^{n}\alpha_j^2\gamma(\lambda_i)}{\sum_{j=1}^{n}\alpha_j^2}$$

$$SC_{S,3}(f_i)=\sum_{j=1}^{k}(\gamma(2)-\gamma(\lambda_i))^2\ \alpha_j^2$$

Dans les équations ci-dessus, $\widehat{f_i}=(D^{\frac{1}{2}}f_i).\left\|D^{\frac{1}{2}}f_i\right\|^{-1}$; $(\lambda_i;\xi_i)$ est le système électronique de $\mathcal{L}$; $\alpha_j=\cos\theta_j$ telque θ_j Est l'angle entre fi et ξj Et $\gamma(\cdot)$ est une fonction croissante qui est utilisée pour réévaluer les valeurs propres De L . Les principaux vecteurs propres de L sont les indicateurs de graphe pour indiquer optimum des données. Par En comparaison avec

ces vecteurs propres, spectrum sélectionne des fonctions qui attribuent des valeurs similaire aux instances qui sont similaires à K. il est démontré que laplacian score est un cas particulier du deuxième critère, $SC_{S,2}$, défini dans spectrum. Notez que spectrum évalue également les caractéristiques individuellement, donc elle ne peut pas gérer la Redondance des caractéristiques.

II.2.4.2 Méthodes multivariées :

a – Algorithme de sélection du variable basé sur la corrélation CFS:

La sélection de variables basée sur la corrélation (CFS) est un algorithme de filtrage simple qui se classe les sous-ensembles de caractéristiques selon une fonction d'évaluation heuristique basée sur la corrélation .le biais de la fonction d'évaluation concerne les sous-ensembles qui contiennent des fonctionnalités qui sont fortement corrélés avec la classe et non corrélés les uns avec les autres. Son importance est que les variables doivent être ignorées car elles auront une faible corrélation avec la classe **[13]**. Les variables redondantes doivent être éliminées car elles seront fortement corrélées avec une ou plusieurs des variables restantes. L'acceptation d'une fonction dépendra de la mesure dans laquelle il prédit les classes dans les zones de l'espace d'instance non déjà prédit par d'autres attributs. La fonction d'évaluation des sous-éléments caractéristiques est la suivantes $Ms = \frac{k\ \overline{\text{rcf}}}{\sqrt{\text{k+k(k}-1)\text{rff}}}$

Où M_s est le «mérite» heuristique d'un sous-ensemble de variables S contenant des variables Kr_{cf} est la corrélation moyenne de la classe de caractéristique (f ε S) et r_{ff} est l'intercorrélation caractéristique. Le numérateur de cette équation peut fournir une indication de la prédiction de la classe, un ensemble de variable ; et le dénominateur de la redondance qu'il existe parmi les caractéristiques **[9]**.

B- Algorithme filtrante basé sur la corrélation -FCBF: [9]

La méthode de filtrage rapide (FCBF) est basée sur une symétrie Incertitude (SU) qui est défini comme le rapport entre le gain d'information (IG) et l'entropie (H) de deux fonctions,

$$\text{x et y:SU(x,y)} = 2\,\frac{IG(x/y)}{\text{H(x)+H(y)}}.$$

Où le gain d'information est défini comme suit:

$$IG(x/y)=H(y)+H(x)-H(x,y),$$

Étant H (x) et H (x; y) l'entropie et l'entropie commune, respectivement. Cette méthode a été conçue pour des données de grande dimension et s'est révélé efficace en supprimant les variables non pertinentes et redondantes. Toutefois, il ne prend pas en compte l'interaction entre les variables.

C- Algorithme minimum Redondance Maximum Relevance (mRMR) :

" Min-Redundancy, Max-relevance" (mRMR) est une méthode de filtrage pour la sélection de caractéristiques proposée par Peng et al. En 2005 **[10]** . Cette méthode est basée sur des mesures statistiques classiques comme l'information mutuelle, la corrélation etc,.... L'idée de base est de profiter de ces mesures pour essayer de minimiser la redondance (mR) entre les variables et de maximiser la pertinence (MR). utilisent l'information mutuelle pour calculer les deux facteurs mR et MR. Le calcul de la redondance et de la pertinence d'une variable est donné par l'équation suivante **[1]**:

$$\text{Redondance (i)}=\frac{1}{|F|^2}\sum_{i,j\epsilon F} I(i,j).$$

$$\text{Pertinence (i)}=\frac{1}{|F|^2}\sum_{i,j\epsilon F} I(i,Y).$$

- $|F|$: représente la taille de l'ensemble de variables.
- I (i; j) : est l'information mutuelle entre la i eme et la j eme variable.
- I (i; y) : est l'information mutuelle entre la ieme variable et l'ensemble des étiquettes de la classe Y. Le score d'une variable est la combinaison de ces deux facteurs tel que :
- Score (i) =Pertinence (i) - Redondance (i).

L'information mutuelle (MI) est une mesure symétrique de l'information qui mesure la quantité d'informations pouvant être obtenues sur une variable aléatoire en observant une autre **[5]**. L'information mutuelle de la caractéristique f_i par rapport à la fonctionnalité $f_j$est donné par :

$$I(f_i; f_j)=\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}.$$

Où x est toutes les valeurs possibles de f_i et y est possible valeurs de f_j.

II.3 Phase de classification

La classification est une méthode de fouille des données, elle permet en particulier de scinder les données en plusieurs classes, il s'agit en effet d'extraire une règle générale à partir des données observées le terme de classification peut désigner trois approches distinctes : la classification supervisée , la classification non-supervisée *(automatic classification et clustering enanglais)*et la classification semi-supervisé. Les méthodes non supervisées ont pour but de constituer des groupes d'exemples (ou des groupes d'attributs) en fonction des données observées, sans connaissance apriori. En revanche les méthodes supervisées utilisent la connaissance a priori sur l'appartenance d'un exemple à une classe pour construire un système de reconnaissance de ces classes **[12]**.par contre la classification semi-supervisé à besoin d'un partie de données annotes pour pouvoir classe les autre classe non annotes. Nous décrivons dans cette section les méthodes de classification supervisée utilisé dans notre étude pour tester les performances des méthodes de sélection nous avons utilisé Kppv, SVM et l'arbre de décision.

II.3 .1 Classification supervisée

L'objectif de la classification supervisée est d'apprendre à l'aide d'un ensemble de données d'entraînement une procédure de classification qui permet de prédire l'appartenance d'un nouvel exemple à une classe **[4].** Les systèmes d'apprentissage permettant d'obtenir une telle procédure peuvent être basés sur des hypothèses probabilistes *(classifieur naïf de Bayes),* Sur des notions de proximité *(plus proches voisins)* ou sur des recherches dans des espaces d'hypothèses *(arbres de décisions, . . .).* Le fonctionnement de la classification supervisée se décompose en deux points le premier est la phase d'apprentissage, tout ce qui est appris par l'algorithme est représenté sous la forme des règles de classification que l'on appelle le modèle d'apprentissage. Le second point est la phase de la classification proprement dite, dans laquelle les données tests vont être utilisées pour estimer la précision des règles de classification générées pendant la première phase **[2].** Si la précision du modèle est considérée comme acceptable, la règle pourra être appliquée à des nouvelles données.

II.3 .2 Les méthodes de la classification utilisée :

II.3 .2.1 K-plus proche voisin Kppv:

Plus connus en anglais sous le nom K-nearest neighbor (K-NN), est l'un des algorithmes les plus populaires utilisés pour le classement dans les différents domaines de la reconnaissance des formes et de la fouille de données, c'est une méthode basée sur la notion de proximité (voisinage) entre exemples et sur le raisonnement à partir de cas similaire pour prendre une décision objet est classifié selon un vote majoritaire par ses voisins **[5]**, l'objet obtient la classe qui est la plus commune chez ses K plus proche voisins dans l'espace des caractéristiques**[15]**.

Le principe de cette méthode est de chercher pour chaque action à classer un ensemble d'actions de l'ensemble d'apprentissage parmi les plus proches possibles de l'action. L'action est alors affectée à la classe majoritaire parmi ces k plus proches voisins. La fixation du paramètre est délicate, une valeur très faible va engendrer une forte sensibilité au bruit d'échantillonnage. La méthode va devenir faiblement robuste. Un trop grand va engendrer un phénomène d'uniformisation des décisions. La plupart des actions vont être affectées à la classe la plus représentée. Pour remédier à ce problème, il faut tester plusieurs valeurs de et choisir le optimal qui minimise le taux d'erreurs de classification **[14]**.

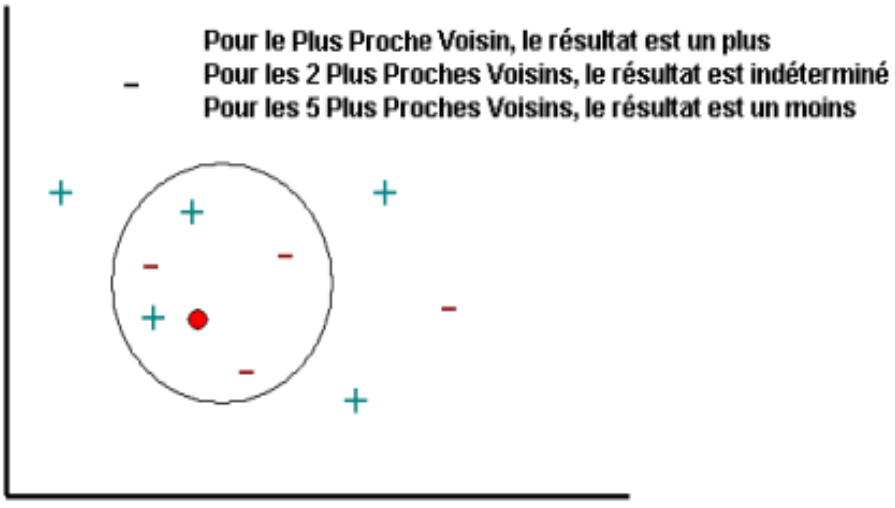

Figure II.1 : Classification avec KNN

II.3.2.1 Algorithme de KNN

Paramètre : le nombre k de voisins

Donnée : un échantillon de m exemples et leurs classes

-La classe d'un exemple X est c(X)

Entrée : un enregistrement Y

1. Déterminer les k plus proches exemples de Y en calculant les distances
2. Combiner les classes de ces k exemples en une classe c

Sortie : la classe de Y est c(Y)=c

Le choix de la distance est primordial au bon fonctionnement de la méthode.

Les distances les plus simples permettent d'obtenir des résultats satisfaisants

Les points positifs de cette méthode sont qu'elle ne pose aucune hypothèse sur la forme des classes à apprendre. La méthode est simple puisqu'il n'y a pas besoin d'apprentissage d'un modèle de classification et son pouvoir prédictif est souvent bon mais la performance de cette méthode diminue lorsque la dimension augmente, puisque pour chaque nouvelle classification, il est nécessaire de calculer toutes les distances de x à chacun des exemples d'apprentissage. De plus, la performance dépend fortement de k, le nombre de voisins choisi et il est nécessaire d'avoir un grand nombre d'observations pour obtenir une bonne précision des résultats.

II.3.2.2 Séparateurs à Vastes Marges (SVM) :

II.3.2.2.1 SVM pour deux classes :

Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais Support Vector Machine, SVM) ont été introduits dès 1992 [Cortes et Vapnik, 1995]. Les SVMs sont des classifieurs qui reposent sur deux idées clés : la notion de marge maximale et la notion de fonction noyau.

Le principe de base est de trouver l'hyperplan optimal qui sépare deux classes dans l'espace de description. Cet espace est celui qui maximise la distance entre les deux classes (la marge), le SVM utilise des fonctions de noyau qui, dans un espace augmenté, permettent

une séparation optimale des points en différentes catégories. Les données d'apprentissage sont utilisées pour découvrir l'hyperplan qui séparera au mieux les points(voir Figure II.2).

L'idée du SVM est d'utiliser sa fonction de noyau pour reconsidérer le même problème dans un espace de dimension plus élevée. Cet espace est caractérisé par la possibilité de trouver un séparateur linéaire qui permet de classer les points dans les deux groupes appropriés. Le séparateur linéaire peut ensuite être projeté dans l'espace d'origine où il devient habituellement non linéaire. Le critère d'optimisation est la largeur de marge entre les classes (l'espace vide de chaque coté des frontières de décision). La largeur de marge est caractérisée par la distance jusqu'aux échantillons d'entrainement le plus près. Ces échantillons s'appellent vecteurs de supports, ils définissent la fonction discriminante qui permet la classification. Le nombre de vecteurs de support est minimisé en maximisant la marge **[16]**.

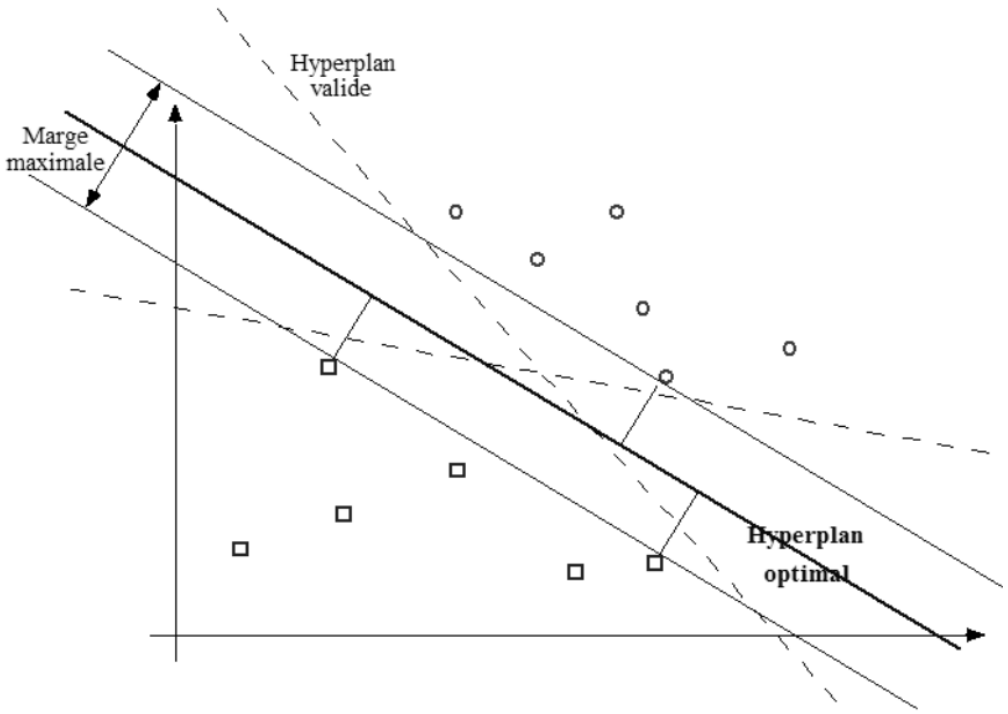

Figure II.2 : Classification avec des SVM

Les point positif du SVM est un modèle robuste et également peut fournir une bonne généralisation, mais son major inconvénient la sélection d'une mauvaise fonction noyau (ou de ces paramètres) peut conduire à produire un effet de sur-apprentissage **[17]**.

II.3.2.2.2 SVM Multi-classe :

Les méthodes des machines à vecteur support multiclasse, réduisent le problème multiclasse à une composition de plusieurs hyperplans biclasse permettant de tracer les frontières de décision entre les différentes classes **[18]**. Ces méthodes décomposent l'ensemble d'exemples en plusieurs sous ensembles représentant chacun un problème de classification binaire. Pour chaque problème un hyperplan de séparation est déterminé par la méthode SVM binaire.

On construit lors de la classification une hiérarchie des hyperplans binaires qui est parcourue de la racine jusqu'à une feuille pour décider de la classe d'un nouvel exemple. On trouve dans la littérature plusieurs méthodes de décomposition **[19]** :

a-Un contre tous

La méthode un-contre-tous (one against all) est la plus simple selon la formulation de Vapnik et elle produit autant de problèmes binaires que de classes **[20]**. Chaque problème binaire correspond à la discrimination des exemples d'une classe avec ceux des autres classes. Elle consiste à déterminer pour chaque classe k un hyperplan H_k la séparant de toutesles autres classes. Cette classe k est considéré comme étant la classe positive (+1) et les autres classes comme étant la classe négative **[19]**.

b-Un-contre-un

Est une méthode dite de un contre un. Au lieu d'apprendre N fonctions de décisions, ici chaque classe est discriminée d'une autre. Ainsi, N (N-1)/2 fonctions de décisions sont apprises et chacune d'entre elles effectue un vote pour l'affectation d'un nouveau point x. La classe de ce point x devient ensuite la classe majoritaire après le vote **[17]**.

II.3.2.3 Arbre de décision (CART) :

Les arbres de décision correspondent à un ensemble d'algorithme (CART, ID3, C4.5, CHAID ,etc.) **[22]**; **[21]** ; **[14]**. Et elles sont très utilisés depuis de nombreuses années dans le cadre de l'apprentissage supervisé .Ils peuvent traiter aussi bien des données représentées par des attributs quantitatifs, des attributs qualitatifs ou des représentations composites a pour objectif la classification et la prédiction , leur fonctionnement est basé sur un enchaînement hiérarchique ,le principe des arbres de décision est de réaliser la classification d'un exemple par une suite de tests sur les attributs qui le décrivent. Concrètement, dans la représentation graphique d'un arbre **[5]** :

1. Un nœud interne correspond à un test sur la valeur d'un attribut.
2. Une branche part d'un nœud et correspond à une ou plusieurs valeurs de ce test.
3. Une feuille est un nœud d'où ne part aucune branche et correspond à une classe.

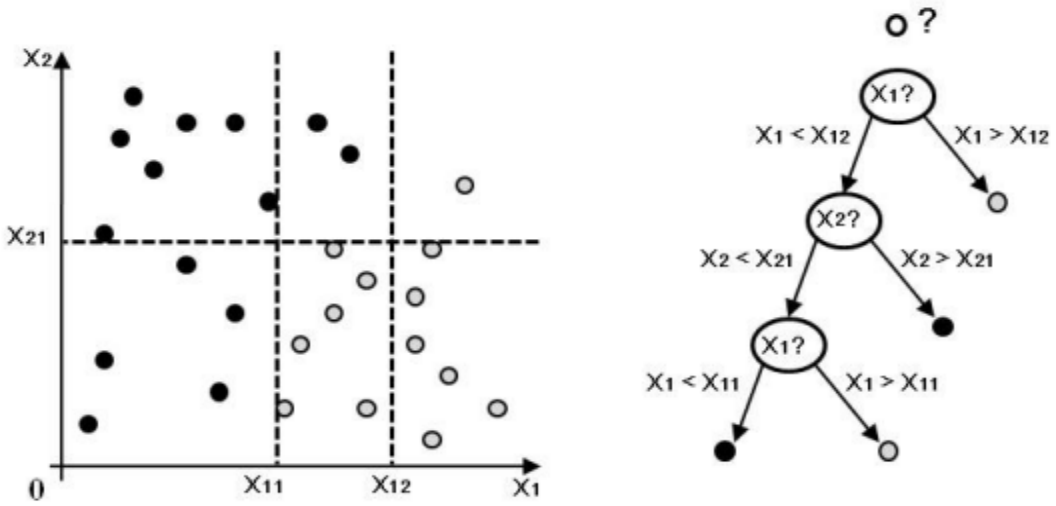

Figure II.3 : Classification avec des arbres de décision

a. L'algorithme d'apprentissage Cart :

Les principaux algorithmes de construction d'un arbre de décision sont : C4.5, et CART Dans ces algorithmes, à chaque étape de création d'un nœud, un critère de séparation entre les classes est utilisé pour décider que l'attribut est le plus pertinent pour la classification **[4]**.

II. 3.2.3 Algorithme d'apprentissage générique

Entrées : langage de description ; échantillon S

Début

Initialiser l'arbre à vide ;//la racine est le nœud courant

Répéter

Décider si le nœud est terminal alors affecter une classe ;

Sinon sélectionner un test et créer le sous-arbre ;

Passer au nœud suivant non exploré s'il existe ;

Jusqu'à obtenir un arbre de décision ;

Fin.

Les avantages procurés par les arbres de décision sont leur rapidité et, surtout, leur facilité quant à l'interprétation des règles de décision. Ce sont des méthodes non paramétriques qui ne font aucune hypothèse sur les données. Ils peuvent traiter des ensembles d'apprentissage avec des données manquantes. Cependant, les arbres de décision **[5]**.

II.3 .3 Validation croisée :

Les performances des méthodes de la classification dépendent généralement du nombre D'échantillons d'apprentissage qui permet d'avoir une estimation non optimiste de l'erreur de classification, il faut utiliser en test des échantillons qui ont servi à l'apprentissage *(utiliser une base d'apprentissage et une base de test)* plus ce nombre est élevé, plus les règles de classification seront fiable. En même temps, il est nécessaire de conserver un nombre significatif d'échantillons de test pour que l'évaluation de ces performances soit significative. La technique de la validation croisée est fréquemment utilisée pour répondre à ces deux besoins: elle consiste à diviser l'ensemble de départ en un certain nombre de sous ensembles de taille égale, chaque sous ensemble étant alors utilisé comme base de test, alors que l'union de tous les autres sous ensembles est utilisée comme base d'apprentissage **[4]**.
L'algorithme de validation croisée à k blocs (k-fold cross-validation) consiste à découper l'ensemble initial d'exemples D en k blocs. On répète alors k phases d'apprentissage-évaluation où une hypothèse h est obtenue par apprentissage sur k − 1 bloc de données
et testée sur le bloc restant.

II.3.3 Algorithme de cross-validation

1- Partitionner l'ensemble d'exemples en k sous-ensembles disjoints :
2- Pour tout i de 1 à k
   - Appliquer l'algorithme d'apprentissage sur le jeu d'apprentissage D_1-D_i pour obtenir une hypothèse.
   - Calculer l'erreur de R_i sur D_i
3- Retourner $R=\frac{1<i<k\ R_i}{k}$ comme estimation de l'erreur.

Avec k = n le nombre d'échantillons. Nous utiliserons la procédure de validation croisée pour évaluer les classifieur que nous mettrons en œuvre sur les jeux de données pour évaluer à la fois une méthode de sélection d'attributs et un classifieur.

II.4 Conclusion

Les sources de données peuvent être multiples et la fusion des données issues de chacune de ces sources conduit à la création d'un ensemble contenant des variables inutiles et redondantes. La solution proposé est la sélection d'un sous ensemble de variables pour permet d'extraire les variables plus pertinente parmi l'ensemble de données. L'étendue de ce chapitre montre la diversité des approches et des méthodes qui a été utilisées pour la sélection des données dans notre travail et les méthode SVM KNN CART pour la classification a pour but de tester les performance de ce dernier

Chapitre III

Résultats et Discussions

III.1 Introduction :

La qualité de système d'apprentissage repose sur la bonne représentation des données à traiter c'est exactement ce que fait la sélection d'attributs qui a pour but d'extraire l'information discriminante et pertinente est permet d'améliorer la performance de ce dernier et le rendre plus efficace .Dans cette section on va présenter l'expérimentation que nous avons réalisé .nous commençons par la description des bases de données que nous avons utilisé et Nous passons après à l'étape de sélection puis nous présentons les résultats que nous avons acquis par l'application de différent classifieur

III.2 Matériel et méthode :

Toutes les expérimentations réalisées ont été exécutées en utilisant le logiciel Matlab et aussi en utilisant la bibliothèque Feature sélection package pour la sélection des attributs

III.3 Description des ensembles de données

Nous avons utilisé 3 jeux de données; Nous présentons les caractéristiques de ces jeux de données :

III.3.1 hépatite : les données disponible dans la référence des données d'apprentissage par le cite UCI cette base de données contient 155 échantillons appartenant à deux classes cibles différentes. Il existe 19 attributs, 13 variable qualitative et 6 variable avec 6-8 valeurs discrètes. avec sortie indique si les patients atteints d'hépatite sont vivants ou morts la distribution de classe contient 32 cas pour la mort et 123 cas pour la vie **[24]**.

N° attribut	Description attribut	Valeur
1	Age	10-80 by step 10 years
2	sex	Male and female
3	stéroïde	No, yes
4	antivirale	No, yes
5	fatigue	No, yes
6	Malaise	No, yes
7	Anorexie	No, yes
8	Liver big	No, yes
9	Liver firm	No, yes
10	Plein palpable	No, yes
11	Naevus araignée	No, yes
12	Ascites	No, yes
13	varices	No, yes
14	Bilirubine	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
15	ALKanine.phosphatase	33, 80, 120, 160, 200, 250
16	Sgot	13, 100, 200, 300, 400, 500
17	Albumine	2.1, 3.0, 3.8, 4.5, 5.0, 6.0

18	Temps de prothrombine	10, 20, 30, 40, 50, 60, 70, 80, 90
19	Histologie	No ,yes
20	class	Die, alive

Tableau III.1 les attributs d'hépatite

III.3.2 Cardiotoco-graphy: Ce jeux de données est constitué de mesures de la fréquence cardiaque fœtale (FHR) et la contraction utérine (UC) sur les caractéristiques cardiotocographie classées par obstétriciens d'experts .contient 2126 cardiotocograms fœtales (CTG) sont automatiquement traitées et les caractéristiques de diagnostic respectives mesurées avec 1655 normale, 295 suspicieux et 176 pathologie. Les CTG ont également été classés par trois obstétriciens d'experts et une étiquette de classification de consensus attribué à chacun d'eux. Classification était à la fois par rapport à un modèle morphologique (A, B, C ...) contient 35 attribut et à un état fœtal (N, S, P). Par conséquent, l'ensemble de données peut être utilisé soit pour 10 classes ou expériences 3-classe **[24].** Dans ce tableau la suite des variables de cette base ne sont pas déclaré

N° attribut	Description attribut
1	LB - base FHR (battements par minute)
2	AC - Nombre d'accélérations par seconde
3	FM - Nombre de mouvements du fœtus par seconde
4	UC - Nombre de contractions utérines par seconde
5	DL - Nombre de décélérations lumière par seconde
6	DS - Nombre de décélérations graves par seconde
7	DP - Nombre de décélérations par seconde prolongée;
8	ASTV - pourcentage de temps à la variabilité à court terme anormale
9	MSTV - valeur moyenne de la variabilité à court terme
10	ALTV - pourcentage de temps à la variabilité à long terme anormale
11	MLTV - valeur moyenne de la variabilité à long terme
12	Largeur - largeur de l'histogramme FHR
13	Min - minimum de l'histogramme FHR
14	Max - Maximum de l'histogramme FHR
15	N max - Nombre de pics de l'histogramme
16	N zeros - Nombre de zéros d'histogramme
17	Mode - Mode d'histogramme
18	Moyenne - histogramme signifie
19	Médiane - médiane de l'histogramme
20	Écart - variance de l'histogramme
21	Tendency - tendance histogramme
22	CLASSE - motif FHR code de classe (1 à 10)
23	NSP - le code de classe d'état fœtal (N = normal; S = suspect; P = pathologique)

Tableau III.2 les attributs Cardiotography

III.3.3 leucimé: ce jeu de données concernait le cancer de sang contient 50 attributs avec 100 patients

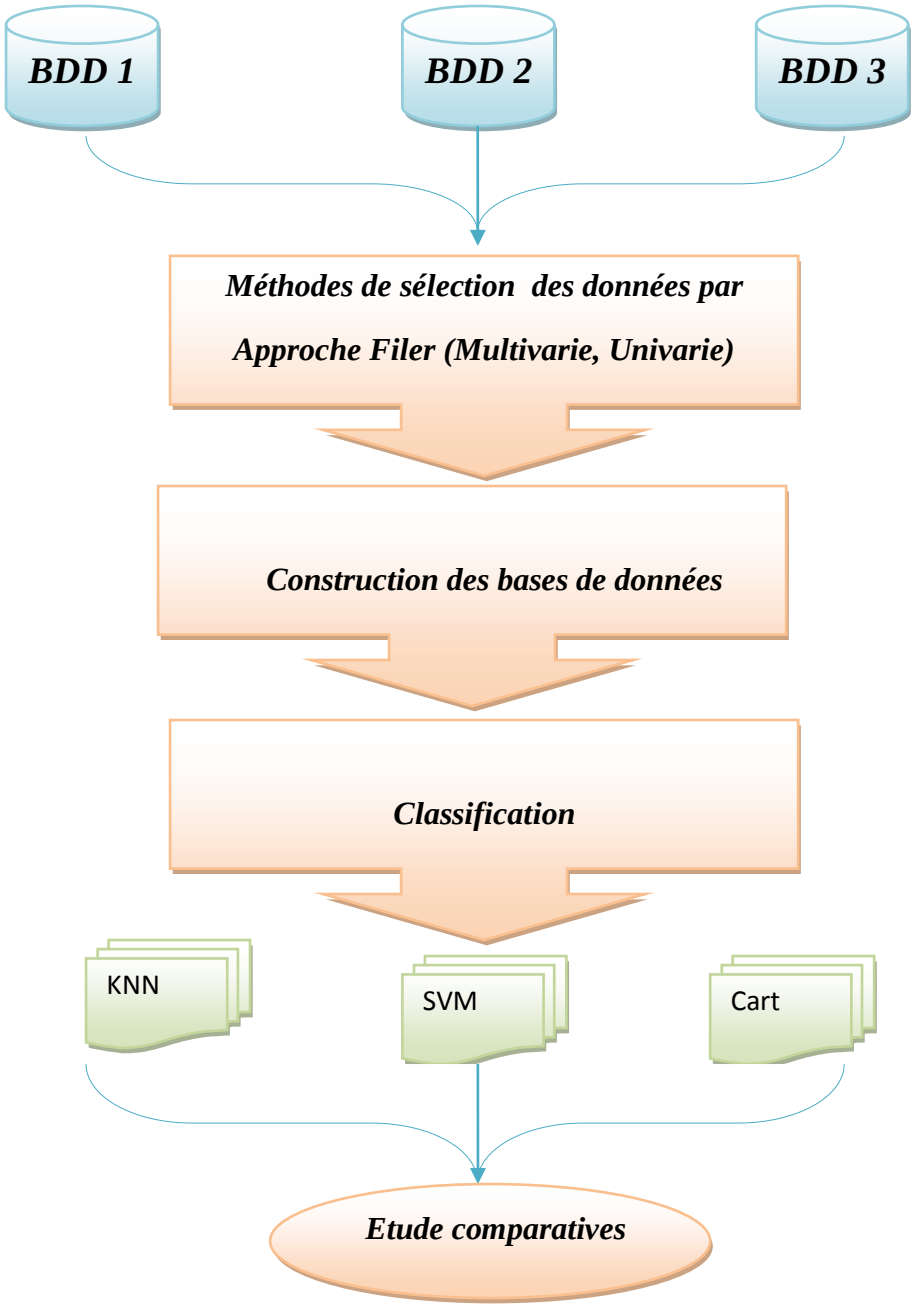

Figure III.1 Schéma représentatif de notre procédure

III.4 phase de sélection :

Dans cette étape nous présentons les variables sélectionnées par chaque méthode de sélection et on a construit par la suite notre base de données

III.4.1 pour la base hépatite :

méthode	
univarie	Fisher [12.17.14.13.11.19.6.18.5.10.1.2.15.3.4.7.9.16.8]
	ReliefF [6.13.11.12.7.14.9.15.10.19.1.16.3.18.17.8.2.5.4]
	Info gain[17.14.12.18.11.13.19.5.6.1.2.10.3.4.7.8.9.15.16]
	Gini Index [17.12.14.18.13.11.19.6.5.1.10.15.2.16.3.4.7.9.8]
	Chi-squar [17.12.14.18.13.11.19.6.5.10.1.2.3.4.7.8.9.15.16]
	Spectrum [2.4.12.13.10.7.11.6.9.19.3.16.14.5.8.15.1.18.17]
	Kruskal-Wallis [18.17.16.15.4.2.1.13.12.7.10.3.9.11.6.19.8.5.14]
Multivarie	CFS [17.12.18.13.1.2]
	FSCBF [1.2.11.12.13.14.17.18]
	MRmR [17.14.12.18.11.13.19.10.15.16]

Tableau III.3 les attributs d'hépatite sélectionné par les méthodes filter

III.4.2 pour la base Leucimé : pour la base Leucimé : pour les méthodes univarie on a choisi que les 25 premier attributs

méthode	
univarie	Fisher [31.18.50.42.38.22.30.37.43.2.21.28.7.26.34.9.11.17.5.49.19.4.15.8.46]
	ReliefF [40.35.47.16.24.40.34.31.49.44.22.4.8.5.39.41.37.32.50.20.42.23.21.15.19]
	Info gain [34.43.49.16.37.47.7.5.37.44.23.8.13.15.2.27.18.40.21.20.45.22.17.29.26]
	Gini Index [35.32.34.37.7.5.27.26.3.12.29.43.28.11.1.47.20.24.49.36.9.4.16.30.8]
	Chi-squar [48.40.31.33.19.39.50.44.41.10.23.13.14.17.46.8.30.38.22.2.18.15.42.45.6]
	Spectrum [18.30.10.48.46.40.50.38.31.14.28.37.20.41.34.17.36.47.9.24.49.44.35]
	Kruskal-Wallis [33.50.13.36.16.42.17.45.22.43.6.3.27.25.1.7.28.34.5.8.2.9.11.2.4.9.10.12]
Multivarie	CFS [2.3.5.6.8.10.11.12.13.14.15.16.17.18.19.20.21.22.23.24.25.28.31.36.39.40.41.44.45 47.48.49.50]
	FSCBF [48.40.50.8.19.14.13.2.12.28.23.42.36.49.16.6.21.20.3.29]
	MRmR [48.50.44.46.45.49.47.41.30.33]

Tableau III.4 les attributs leucimé sélectionné par les méthodes filter

III.4.3 pour la base cardiotography :

méthode		
univarie	Fisher	[26.27.28.29.30.31.32.33.34.35.13.15.6.11.22.24.16.17.10.23.21.9.18.19.8]
	ReliefF	[31.27.29.32.13.19.20.24.7.6.2.8.22.26.28.30.34.35.10.12.14.17.23.16.4.5]
	Info gain	[24.6.27.1.13.10.26.31.16.32.22.9.11.17.35.23.21.12.18.33.19.15.29.4.5.30]
	Gini Index	[1.2.3.4.5.6.8.9.10.11.12.13.14.15.16.17.19.20.21.22.23.24.25.26.27.28…]
	Chi-squar	[24.1.22.10.11.6.9.26.27.32.33.34.35.29.30.31.28.21.13.16.23.15.12.17.18]
	Spectrum	[14.34.28.7.29.30.33.35.15.32.31.26.20.25.11.27.13.24.6.19.12.8.2.1.10.16]
	Kruskal-Wallis	[35.34.33.32.31.30.29.28.27.25.23.22.21.20.18.17.16.15.14.12.9.6.5.4.2.1]
Multivarie	CFS	[1.6.10.11.12.13.18.22.24.26.27.28.29.30.31.32.33.35]
	FSCBF	[27.26.6.31.13.24.32.35.10.11.1.22.33.29.30.34.12.18.28]
	MRmR	[24.35.31.32.33.27.16.17.19.34]

Tableau III.5 les attributs cardiotography sélectionnés par les méthodes filter

III.5 phase de classification :

La sélection de caractéristiques est souvent un prétraitement pour passer à la classification dans le but de faire une validation , dans le cas de notre expérience nous avons utilisé la validation croisée qui assure la participation de tous les sous ensemble de variables par les performances de l'algorithme d'apprentissage utilisé (KNN , SVM, Cart) ensuite nous avons comparé les taux correcte de la classification, avant et après la sélection

III.5.1 les critères de performance d'un classifieur :

La performance de la classification est un critère d'évaluation important de la sélection de variables. Plusieurs paramètres de performance de classification se trouvent dans la littérature

- Vrai positif (VP): est le nombre de patient malade qui est classé malade
- Faux négatif (FN): est le nombre de malade qui sont classée non malade
- Faux positif (FP): est le nombre de patient non malade qui sont classée malade.
- Vrai négatif (VN): est le nombre de patient non malade qui sont classée non malade.

Cette comparaison est faite en utilisant une matrice de confusion où les lignes et les colonnes sont respectivement vraies et pré-dictées. VP et VN représentent les décisions correctes prises par le classificateur, alors que FN et Les FP sont des erreurs de classification. De cette matrice, plusieurs caractéristiques utiles de la classification peuvent être utilisées **[25]**

Précision: Le critère de performance le plus utilisé est le taux de classification correct, Connu sous le nom de précision
(TC %) : [TC = 100 * (VP + VN) / (VN + VP + FN+FP)]
Sensibilité: ou rappel est la proportion de VP sur le nombre total de cas positifs.
(Se%) : [Se = 100 * VP / (VP + FN)]
Spécificité: est la proportion de VN sur le nombre total de cas négatifs
(Sp %) : [Sp = 100 * VN / (VN + FP)] **[26]**

III.6. résultats obtenus :

Nous avons lancé diverses expérimentations qui prennent en compte deux aspects le nombre des attributs sélectionnés par chaque méthode et leur influence sur le taux de la classification, Pour réaliser nos expérimentations, nous utilisons un CV- K=10 fois sur les ensembles de données leucimé et hépatite et K=15 pour la base cardiotography

III.6.1 la base Leucimé :

a- Méthode KNN : Nous avons appliqué sur la base de leucémie les méthodes univarie et multivarie en utilisant le classifieur KNN pour calculer le taux de classification .Après plusieurs expérimentations nous avons retenu la valeur de K=3 pour le classifieur KNN .Nous notons aussi que le taux de classification obtenu sans la sélection des attributs c'est-à-dire en gardent les 50 attributs étant de 80%

❖ Résultat des méthodes multivarie :

	Sans sélection	FSCBF	CFS	MRMR
N.V.S	50 attributs	20 attributs	34 attributs	10 attributs
TC%	80	78	80	**81**

Tableau III.6 Résultats obtenus utilisant méthode multivarie appliquent KNN sur la base leucimé

❖ Résultat des méthodes univarie :

N.V.S	Fisher	ReliefF	Info gain	KrusKal Wallis	Gini index	Chisquare	Spectrum
10	77	76	73	77	70	78	73
20	78	80	75	79	73	84	77
25	77	80	73	79	70	**85**	74
30	**85**	78	77	80	75	80	80
35	83	79	79	80	73	81	81
40	84	77	79	83	80	80	80
Sans sélection	80						

Tableau III.7 Résultats obtenus utilisant méthode multivarie appliquent KNN pour la leucimé

Discussion :

Tableaux (III.6) et (III.7) présent les différent taux de classification en adaptant les deux approches de sélection (univariée et multivariée) .Nous remarquons clairement que la méthode MRmR (approche multivariée) à donné un taux presque similaire (TC=81%) au cas de sans sélection avec seulement dix attributs. Par contre pour l'approche univarie nous citons les deux méthodes ficher et chisquare respectivement
(30et 25) attributs qui ont donne un score meilleur de TC= 85%

b- Méthode SVM: nous avons testé les mêmes méthodes de sélection citées ci-dessus avec la classifieur SVM .Après plusieurs expérimentations nous avons retenu la fonction Linear il a obtenu un taux de classification TC=81% sans sélection d'attributs

❖ Résultat des méthodes multivarie :

	Sans sélection	FSCBF	CFS	MRMR
N.V.S	50 attributs	20 attributs	34 attributs	10 attributs
TC%	81	**83**	78	79

Tableau III.8 Résultats obtenus pour les méthodes multivarie appliquons SVM

❖ Résultat des méthodes univarie :

N.V.S	Fisher	ReliefF	Info gain	KrusKal Wallis	Gini index	Chisquare	Spectrum
10	80	**85**	77	80	79	83	82
20	76	83	82	77	82	77	**85**
25	83	**87**	85	80	79	74	82
30	79	85	83	80	83	71	80
35	75	83	85	77	84	78	82
40	77	82	84	80	84	73	84
Sans sélection	81						

Tableau III.9 Résultats obtenus pour les méthodes univarie appliquons SVM

Discussion :

Les tableaux(III.8) et (III.9) montre que dans le cas de l'approche multivarie c'est la méthode FCBF qui a donnée le meilleur résultat (TC=83%) avec seulement 20 attribut Par contre par l'approche univarie nous notons que le meilleur résultat et de taux de classification = 87% avec les 25 attributs suivi par TC=85 avec 10 attributs données par la méthode ReliefF et après moins la méthode Gain d'Information et la méthode spectrum de TC=85% avec respectivement 25 , 35 et 20 attributs

c-Arbre de décision (C4.5) : le dernier classifieur que nous avons testé est l'arbre de décision (C4.5) il a obtenu un taux de classification de 74% sans sélection

❖ Résultat des méthodes multivarie :

	Sans sélection	FSCBF	CFS	MRMR
N.V.S	50 attributs	20 attributs	34 attributs	10 attributs
TC%	74	77	**79**	75

Tableau III.10 Résultats obtenus par les méthodes multivarie appliquons Cart

❖ Résultat des méthodes univarie :

N.V.S	Fisher	ReliefF	Info gain	KrusKal Wallis	Gini index	Chisquare	Spectrum
10	75	74	69	70	63	68	72
20	**83**	76	73	73	61	69	71
25	78	80	78	70	65	71	68
30	73	71	70	72	66	73	68
35	79	73	70	72	68	73	69
40	75	71	72	75	70	78	77
Sans sélection	74						

Tableau III.11 Résultats obtenus par les méthodes univarie applique Cart

Discussion :

D'après les (tableaux III.10 , III.11) qui présent les différent taux de classification en adaptant les deux approches de sélection (univariée et multivariée) .Nous remarquons clairement que le méthodes univarie Fisher donne un meilleur taux de classification de TC=83% avec seulement 20 attributs , et suivi par la méthodes ReliefF TC=80% avec 25 attributs . Par contre pour l'approche multivarie nous citons la méthode CFS d'un taux de classification égale à 79 %.

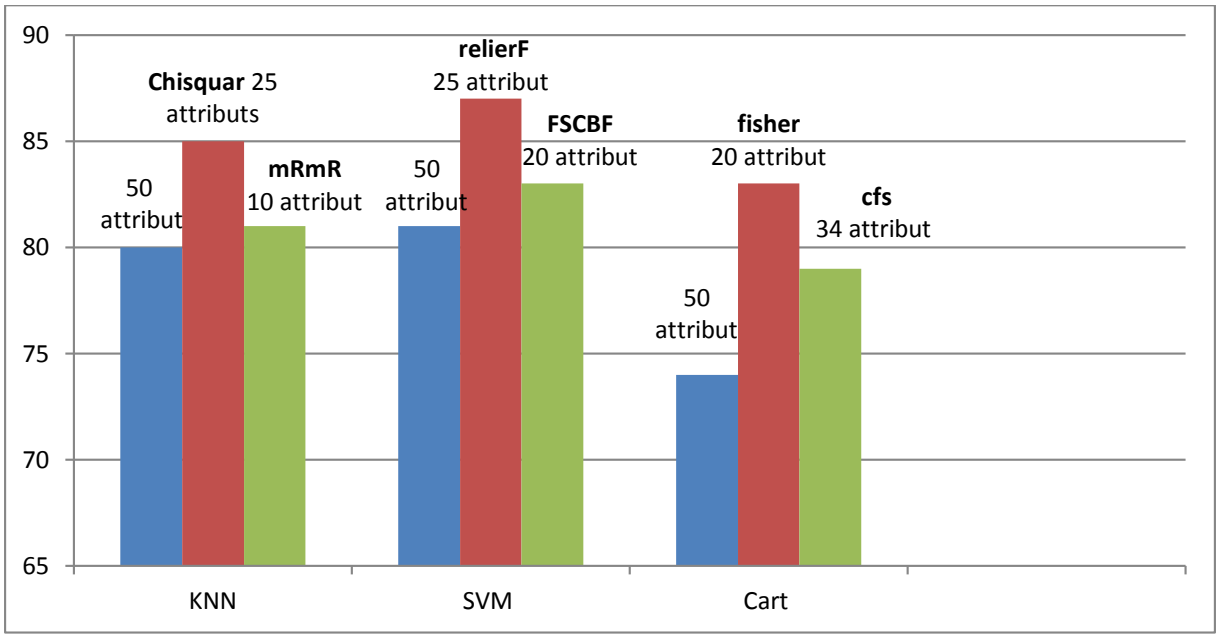

Figure III.1 les meilleurs résultats obtenus par chaque classifieur avec leucimé

Conclusion

L'analyse des tableaux montre que la sélection améliore la performance de classifieur et réduire le temps de classification en comparaisons avec la classification sans sélection de attributs, on note que les méthode univarie donnes les meilleures résultat par rapport aux méthodes multivarie car les méthode univarie sélectionne les attributs les plus pertinents sans prend en considération les interaction entre les attributs tendis que la meilleur résultat est la méthode ReliefF avec la classifieur SVM de TC=87% avec un sensibilité de 89% et spécificité 81.92 pour la base leucimé ,Qui dû a la simplicité , la facilité de la mise en œuvre ainsi que la précision même sur des données bruitées .on conclure que l'efficacité des méthodes de sélection dépend aussi sur la qualité des données , les nombre des attributs sélectionné et le classifieur utilisé .

III.6.2 la base hépatite :

a- KNN : Nous avons appliqué sur la base d'hépatite les méthodes univarie et multivarie en utilisant le classifieur KNN pour calculer le taux de classification .Après plusieurs expérimentations nous avons retenu la valeur de K=7 pour cette base de données .Nous notons aussi que le taux de classification obtenu sans la sélection des attributs c'est-à-dire en gardent les 19 attributs étant de 75.48%

❖ Résultat des méthodes multivarie

	Sans sélection	FSCBF	CFS	MRMR
N.V.S	19 attributs	6 attributs	8 attributs	8 attributs
TC%	75.48	81.29	**82.58**	76.77

Tableau III.12 Résultats obtenus par les méthodes multivarie appliquons KNN sur l'hépatite

❖ Résultat des méthodes univarie

N.V.S	Fisher	ReliefF	Info gain	KrusKal Wallis	Gini index	Chisquare	Spectrum
5	81.29	80.64	81.29	77.41	81.29	82.58	80.64
9	**84.51**	80	81.93	75.48	84.30	**84.51**	83.87
14	81.29	75.48	80.64	74.83	76.77	81.29	78.70
Sans sélection	75.48						

Tableau III.13 Résultats obtenus par les méthodes univarie appliquons KNN sur l'hépatite

Discussion :

Tableaux (III.12) et (III.13) présent les différent taux de classification en adaptant les deux approches de sélection (univariée et multivariée) .Nous remarquons clairement que la méthode MRmR (approche multivariée) à donné un taux presque similaire (TC=76.77%) au cas de sans sélection avec seulement 8 attributs. Par contre pour l'approche univarie nous citons les deux méthodes Ficher et chisquare avec 9 attributs qui représente un agréable résultat 84.51%

b- SVM: nous avons testé les mêmes méthodes de sélection citées ci-dessus avec la classifieur SVM .Après plusieurs expérimentations nous avons retenu la fonction Linear il a obtenu un taux de classification TC=79.35% sans sélection d'attributs

❖ Résultat des méthodes Multivarie **:**

	Sans sélection	FSCBF	CFS	MRMR
N.V.S	19 attributs	6 attributs	8 attributs	8 attributs
TC%	79.35	**81.93**	80.64	76.12

Tableau III.14 Résultats obtenus par les méthodes multivarie appliquant SVM sur l'hépatite

❖ Résultat des méthodes univarie :

N.V.S	Fisher	ReliefF	Info gain	KrusKal Wallis	Gini index	Chisquare	Spectrum
5	73.54	80	76.77	76.12	77.40	80.64	78.70
9	77.41	80	76.12	77.41	80	77.41	80
14	81.93	80.64	**83.22**	78.06	**83.22**	81.29	81.29
Sans sélection	79.35						

Tableau III.15 Résultats obtenus par les méthodes multivarie appliquant SVM sur hépatite

Discussion :

Les tableaux (III.14) et (III.15) montre que dans le cas de l'approche multivarie c'est la méthode FCBF qui a donnée le meilleur résultat TC=81.93% avec seulement 6 attributs Par contre par l'approche univarie nous notons que les méthodes Gini index et Info-gain réalise une identique amélioration aux niveaux de taux de sélection de 83.22 % avec 14 attributs sélectionné Notons que la sensibilité était plus importante avec la méthode Gain d'information qu'avec la méthode de Index de géni où la qualité de l'apprentissage dépend énormément de la sensibilité de prédiction de la maladie.

c-Arbre de décision (C4.5) : nous avons testé est l'arbre de décision (C4.5) il a obtenu un taux de classification de 76.13 % sans sélection

❖ Les résultats des méthodes Multivarie **:**

	Sans sélection	FSCBF	CFS	MRMR
N.V.S	19 attributs	6 attributs	8 attributs	8 attributs
TC%	76.13	**78.06**	77.30	**78.06**

Tableau III.16 Résultats obtenus par les méthodes multivarie appliquons Cart sur l'hépatite

❖ Les résultats des méthodes univarie **:**

N.V.S	Fisher	ReliefF	Info gain	KrusKal Wallis	Gini index	Chisquare	Spectrum
5	79.93	78.06	76.77	80.65	78.06	78.71	80.00
9	78.71	71.61	78.71	76.77	77.42	77.42	80.65
14	77.42	72.26	79.35	76.77	79.03	**82.58**	74.19
Sans sélection	76.13						

Tableau III.17 Résultats obtenus par les méthodes univarie appliquons Cart sur l'hépatite

Discussion :

D'après les (tableaux III.16 , III.17) qui présent les différent taux de classification en adaptant les deux approches de sélection (univariée et multivariée) avec le classifieur Cart Nous remarquons que le méthodes univarie la méthode chi-squar donne un meilleur taux de classification de TC=82.58% avec 14 attributs . Par contre pour l'approche multivarie nous citons la méthode FSCBF et mRmR qui donne un taux de classification identique de TC= 78.06%.

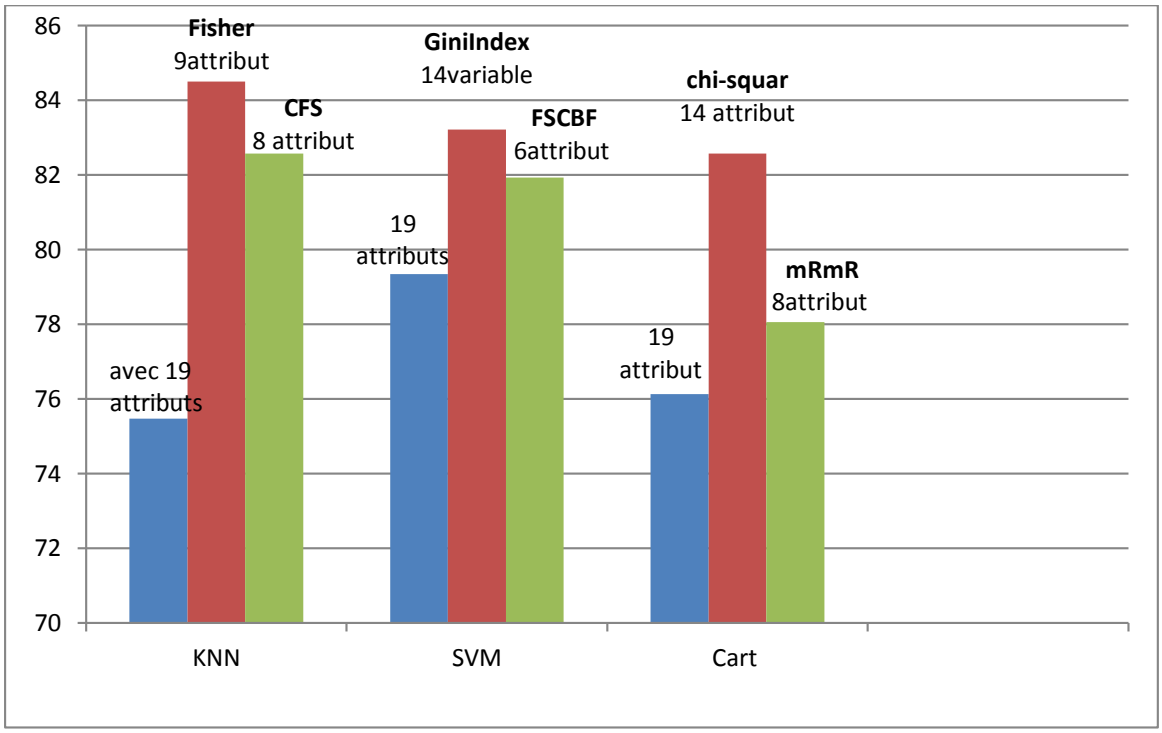

Figure III.2 les meilleurs résultats obtenus par chaque classifieur avec hépatite

Conclusion :

pour ce jeu de données (hépatite) on trouve que la sélection ne donne pas une bon amélioration au niveaux de performance particulièrement a chaque fois on a réduisons le nombre d'attribut ce qui est observé dans les tableaux qui est dû a la taille base de données ; on vois que les meilleur résultat a été obtenu par la méthode Fisher (9) par la classifieur KNN avec sensibilité 92.68% avec un bon prédiction de la maladie , suivi par la méthode Gini Index(14attribut) avec la classifieur SVM .et la méthode chisquar avec (14 attributs) Après une discussion on trouve que les(9 attributs) sélectionné par la méthode Fisher son les plus pertinent pour la détection de l'hépatite comme Sgot (sont des enzymes libérées dans le sang par le foie) ,Ascites,Temps de prothrombine, ictère

III.6.2 la base Cardiotography :

a- KNN : Nous avons appliqué sur la base de cardiotography les méthodes univarie et multivarie en utilisant le classifieur KNN .Après plusieurs expérimentations nous avons retenu la valeur de K=3 pour le classifieur KNN .Nous notons aussi que le taux de classification obtenu sans la sélection des attributs c'est-à-dire en gardent les 36 attributs étant de 42.23%

❖ Résultats des méthodes multivarie**:**

	Sans sélection	FSCBF	CFS	MRMR
N.V.S	36 attributs	19 attributs	19 attributs	10 attributs
TC%	42.23	74.41	**74.78**	60.90

Tableau III.18 Résultats obtenus par les méthodes multivarie appliquons KNN sur cardiotography

❖ Résultats des méthodes univarie**:**

N.V.S	Fisher	ReliefF	Info gain	KrusKal Wallis	Gini index	Chisquare	Spectrum
5	70.60	66.79	70.13	72.24	24.60	66.84	25.54
10	81.44	61.61	68.39	81.03	28.45	76.71	54.75
15	**82.78**	48.07	72.76	80.66	31.56	76.52	54.47
25	72.01	49.15	74.5	51.59	42.38	72.77	49.62
Sans sélection	42.23						

Tableau III.19 Résultats obtenus par les méthodes univarie appliquons KNN sur cardiotography

Discussion :

Tableaux (III.18) et (III.19) présent les différent taux de classification en adaptant les deux approches de sélection (univariée et multivariée) .Nous remarquons clairement que la méthode CFS (approche multivariée) à donné un taux moyenne (TC=74.78%) avec seulement 19 attributs. Par contre pour l'approche univarie nous citons la méthode Ficher avec 15 attributs qui ont donne un score meilleur de TC= 82.78%

b- SVM : nous avons testé les mêmes méthodes de sélection citées ci-dessus avec la classifieur SVM .Après plusieurs expérimentations nous avons retenu la fonction Linear il a obtenu un taux de classification TC=90.51% sans sélection d'attributs

❖ Les résultats des méthodes multivarie**:**

	Sans sélection	FSCBF	CFS	MRMR
N.V.S	35 attributs	19 attributs	19 attributs	10 attributs
TC%	90.51	**95.14**	94.30	91.25

Tableau III.20 Résultats obtenus par les méthodes multivarie appliquons SVM sur cardiotography

❖ Les résultats des méthodes univarie **:**

N.V.S	Fisher	ReliefF	Info gain	KrusKal Wallis	Gini index	Chisquare	Spectrum
5	69.99	81.57	55.28	63.19	74.23	39.88	28.51
10	73.02	79.82	84.68	73.02	75.23	78.29	59.94
15	93.15	86.31	96.28	92.76	75.15	**99.66**	60.98
25	93.60	**99.92**	98.82	91.52	89.93	98.60	73.56
Sans sélection	90.51						

Tableau III.21 Résultats obtenus par les méthodes univarie appliquons SVM sur cardiotography

Discussion :

Les tableaux (III.20) et (III.21) montre que dans le cas de l'approche multivarie c'est la méthode FCBF qui a donnée le meilleur résultat (TC=95.14%) avec seulement 19 attribut Par contre par l'approche univarie nous notons que le meilleur résultat et de taux de classification TC= 99.92% avec la méthode ReliefF (25 attributs)

c- Cart (C4.5) : nous avons testé est l'arbre de décision (C4.5) il a obtenu un taux de classification de 96.92 % sans sélection

❖ Les résultats des méthodes multivarie**:**

	Sans sélection	FSCBF	CFS	MRMR
N.V.S	35 attributs	19 attributs	19 attributs	10 attributs
TC%	96.92	99.86	**99.95**	92.74

Tableau III.22 Résultats obtenus par les méthodes multivarie appliquons Cart sur cardiotography

❖ Les résultats des méthodes univarie **:**

N.V.S	Fisher	ReliefF	Info gain	KrusKal Wallis	Gini Index	Chisquare	Spectrum
5	70.60	81.60	79.68	72.24	54.99	66.36	37.91
10	86.58	82.40	93.65	89.63	71.82	95.34	69.29
15	99.67	93.32	88.82	98.47	83.53	99.67	88.69
25	99.36	**99.98**	99.25	99.52	85.18	**99.95**	90.45
Sans sélection	96.92						

III.23 Résultats obtenus par les méthodes univarie appliquons Cart sur cardiotography

Discussion :

D'après les (tableaux III.22 , III.23) qui présent les différent taux de classification en adaptant les deux approches de sélection (univariée et multivariée) avec le classifieur Cart Nous remarquons que le méthodes univarie la méthode ReliefF donne un meilleur taux de classification de TC=99.98 % avec 25 attributs . Par contre pour l'approche multivarie nous citons la méthode CFS avec (19 attributs) qui donne un taux de classification identique de TC= 99.95%.

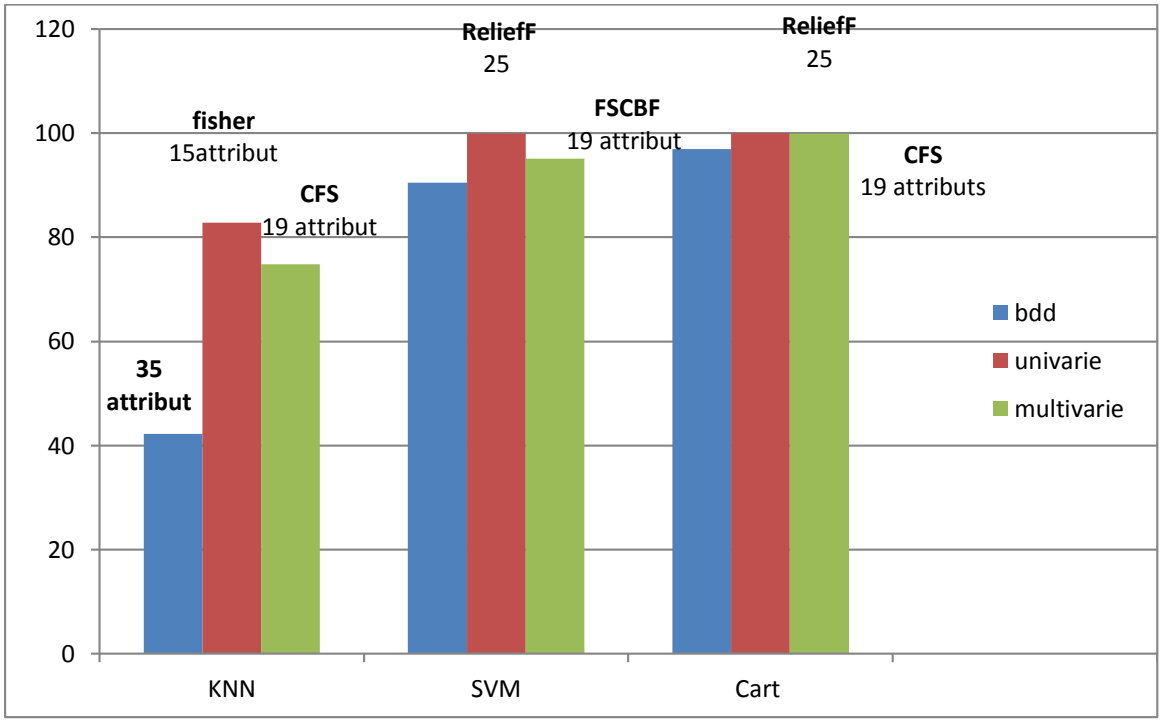

Figure III.3 les meilleurs résultats obtenus par chaque classifieur avec cardiotography

Conclusion

La sélection a été très bienfaisante avec la classification par KNN particulièrement la proche Filter univarie d'après l'état de l'art la méthode Fisher on trouve son capacité de sélection Est beaucoup plus appliquée qu'avec les bases d'une sortie binaires. La méthode ReliefF avec ca précision sur les données bruité permet d'amélioré le taux de classification allant de 90% à 99.92%

Conclusion générale et perspectives

Dans de nombreux domaines de classification, la dimensionnalité des données entraînent des problèmes de superposition et d'autres problèmes graves aux algorithmes d'apprentissage. La restructuration de la dimension est une solution pour ce genre de problème. Notre principale contribution de faire une étude comparative entre les méthodes de sélection d'approche filter. Les résultats obtenus montrent que la majorité des filtres testés ont amélioré les performances des classifieur avec un nombre réduit de attributs .Notez que les meilleures méthodes de sélection c'est les méthode univarie avec les trois bases de données Cela peut être interprété par l'avantage des méthodes univarie qui sont faite la sélection sans tenu compte de la interaction entre les attributs et la précision sur les données bruité c'est à dire considère chaque entité séparément . Cependant, la précision de la classification dépend de la phase de prétraitement des données et aussi avec le choix du classifieur, ses paramètres, et la qualité des données utilisé.

- Nous espérons réalisés en perspectives à ce travail d'autres expérimentations afin d'approfondir nos résultat d'aide au diagnostic médicale.
- Nous proposons de combiner différentes classifieur afin d'améliorer les performances.
- Nous proposons un changement d'environnement d'exécution.
- Nous proposons d'appliquer les méthodes de sélection sur des données avec un nombre de descripteurs beaucoup plus large.
- Nous proposons une hybridation entre les points forts de ces méthodes de sélection.
- Nous proposons aussi une interface graphique compressible par les médecins.

Références Bibliographies

[1] CHOUAIB Hassan *''Sélection de caractéristiques'', Université Paris Descartes,* 2011

[2] HAFA Amel, *''Sélection de Variables Biologiques par l'approche FILTER'', université Tlemcen, 2012*

[3] M. Kalakech, "Sélection semi-supervisée d'attributs: Application à la classification de textures couleur," Université de Lille 1**,** 2011

[4] ALAOUI Abdia, *Application des techniques des métas heuristiques pour l'optimisation de la tâche de la classificationde la fouille de données, université d'ORAN* **,***2012*

[5] El Akadi Ali, *" contribution a la sélection des variable pertinentes en classification supervisé ", université Mohamed V-AGDAL ,2012.*

[6] Petr Posik *Feature selection and extraction, TECHNICAL UNIVERSITY IN PRAGUE, 2015*

[7] *K Kira and L. Rendell*. *,''A practical approach to feature selection'', 1992.*

[8] José Crispin Hernandez Hernandez2008 *Algorithmes métaheuristiques hybrides pour la sélection de gènes et la classification de données de biopuces,Université d'Angers,2008*

[9] Zhao et AL ''Advancing *Feature Selection Research ASU Feature Selection Repository'*

[10] Peng, F Long, and C Ding.*'' Feature selection based on mutual information : criteria of max-dependency, max-relevance, and min-redundancy''. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005.*

[11]*International Journal of Innovative Research in Technology & Science(IJIRTS) comparison of filer based feature selection algorithme an overview ISSN:2321-1156 p108*

[12] Edmundo Bonilla Huerta , ''*Logique floue et algorithmes génétiques pour le prétraitement de données de biopuces et la sélection de génes'', Université d'Angers, 2008.*

[13] BEN BRAHIM , *Afef BEN BRAHIM, Stable and Efficient Feature Selection Methods for High Dimensional Data '',Université de Tunis,2015*

[14] Henriet, L *Système d'évaluation et de classification multicritères pour l'aide à la décision, construction de modèles et procédures d'affectation. Thèse de doctorat en science. Université Paris Dauphine.2000*

[15] Shakhnarovich, Gregory, T. Darrell et P. Indyk: *Nearest-Neighbor Methods in Learning and Vision. The MIT Press , 2006*

[16] Cortes et V. Vapnik: *Support vector network. Learning machine, 1995*

[17] BELAROUCI Sara,*''Traitement et classification des données médicales non Equilibrées'' université Tlmcen,2016*

[18] L. Hamel. *Knowledge discovery with support vector machines. Wiley Edition, 2009.*

[19] Abdelhamid DJEFFAL *,Utilisation des méthodes SupportVector Machine (SVM) dans l'analysedes bases de données,Université Mohamed Khider– Biskra,2011*

[20] V.N. Vapnik. *Statistical Learning Theory. Edition Wiley, 1998*

[21] Breiman, L. et al. *Classification and Regression Trees. Chapman and Hall, New York, 1984 .*

[22] Quinlan, J. R. (1993). *C4.5 : programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.*

[23] BOUBLENZA Amina , *Coopération entre classifieur hétérogènes pour la reconnaissance des données médicales, une thèse de doctorat université aboubekr belkaid,2016.*

[24] *http://archive.ics.uci.edu/ml/*

[25]Mathieu Feuilloy *,étude d'algorithmes d'apprentissage artificiel pour la prédiction de la syncope chez l'homme. Université d'Angers,2009*

[26] Amine chikh, Nessma Settouti. *Renforcement de l'apprentissage structurel pour la reconnaissance du diabète. Rapport de Magister, .2011*