République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université de Mohamed El Bachir El Ibrahimi de Bordj Bou Arréridj
Faculté des Mathématiques et d'Informatique
Département d'informatique

MEMOIRE
Présenté en vue de l'obtention du diplôme
Master en informatique
Spécialité : Technologies de l'information et des communications

THEME

Impact des techniques de prétraitement sur la performance des modèles de classification du diabète.

Présenté par :

Dif marwa

Mahmoud Zineb Ghezlane

Soutenu publiquement le : 22/06/2025

Devant le jury composé de :

Président : Dr.Nouioua Mourad. MCA université de BBA

Examinateur : Dr.Saifi Linda. MCB université de BBA

Encadrante : Dr. Boutouhami Sara. MCB université de BBA

2024/2025

Dédicace

Au terme de ce travail, je remercie mon Dieu, le Tout-Puissant, de m'avoir donné le courage pour terminer ce projet. Je dédie :

À mon père : Tu laisses un grand vide dans ma vie, mais ta place est toujours dans mon cœur.

À ma chère mère : Pour son amour infini, ses sacrifices inestimables et son soutien au long de mon parcours.

À mes frères : Khalil et Marwen. Sources d'espoir et d'amour, pour leurs encouragements et leur présence constante.

À mon mari : Pour ses encouragements et son soutien qui m'ont permis de terminer ce travail.

À toute ma grande famille, à mes oncles et mes tantes et leurs enfants.

À mon binôme Marwa et à ma chère amie Yasmine.

À ma superviseure : Mme Boutouhami Sara Pour ses précieux conseils et son accompagnement.

Mahmoud Zineb Ghezlane

Dédicace

Au terme de ce travail, je rends grâce à Dieu, le Tout-Puissant, pour m'avoir accordé la force, la patience et la persévérance nécessaires à l'accomplissement de ce projet. Je dédie :

À mon père(Dif Amar) : Pour ses encouragements constants, ses conseils avisés et sa confiance qui m'a donné la force d'avancer.

À ma chère mère : Pour son amour inconditionnel, ses sacrifices inestimables et son soutien indéfectible tout au long de mon parcours

À mes frères : Aymen et Adem et à ma tendre sœur Lamis Sources d'espoir et d'amour, pour leurs encouragements et leur présence constante.

À toute ma grande famille Dif et Bahrata

À mon binôme Amira et À mes chères amies Yasmine et Abir.

À ma superviseure : Mme Boutouhami Sara Pour ses précieux conseils et son accompagnement.

Dif Marwa

Remerciement

C'est avec un immense plaisir que nous réservons ces quelques lignes en signe de gratitude et de reconnaissance à tous ceux qui ont contribué de près ou de loin à l'élaboration de ce travail. Nous rendons grand merci à Dieu Tout-Puissant qui nous a donné une grande volonté et le courage pour faire cet humble travail.

À notre encadrante Dr. Boutouhami Sara, pour sa compréhension, sa disponibilité, son aide et ses précieux conseils qui nous ont été très utiles pour l'achèvement de ce projet.

Nous exprimons également notre gratitude envers les membres du jury Dr.Nouioua Mourad et Dr.Saifi Linda pour l'intérêt qu'ils ont porté à notre projet de fin d'études et d'avoir accepté d'examiner notre travail.

Nos remerciements s'étendent à tous nos enseignants du département d'Informatique de l'Université BBA.

Résumé

Le diabète est une maladie chronique dont le diagnostic précoce est crucial pour prévenir les complications graves. Dans ce travail, nous étudions l'impact des techniques de prétraitement sur la performance des modèles de classification appliqués au diabète. Pour ce faire, nous utilisons deux bases de données médicales : la base Pima Indians et une base locale irakienne. Nous expérimentons trois algorithmes de classification : la régression logistique, les machines à vecteurs de support (SVM) et les arbre de décision. Nous appliquons deux méthodes de normalisation (MinMaxScaler et StandardScaler) ainsi que trois techniques de sélection de caractéristiques (SelectKBest, GenericUnivariateSelect, SelectFromModel). Les résultats, évalués par validation croisée, montrent que le choix judicieux des techniques de prétraitement permet d'améliorer significativement l'accuracy des modèles, avec des performances variables selon la nature des données et l'algorithme utilisé.

Mots-clés : Diabète, Classification, Prétraitement, Sélection de caractéristiques, Normalisation, Validation croisée.

Abstract

Diabetes is a chronic disease for which early diagnosis is crucial to prevent serious complications. In this work, we study the impact of preprocessing techniques on the performance of classification models applied to diabetes data. To this end, we use two medical datasets : the Pima Indians dataset and a local dataset from Iraq. We evaluate three classification algorithms : logistic regression, support vector machines (SVM), and decision trees. We apply two normalization techniques (MinMaxScaler and StandardScaler) and three feature selection methods (SelectKBest, GenericUnivariateSelect, SelectFromModel). The results, evaluated using cross-validation, show that a well-chosen preprocessing strategy significantly improves model accuracy, with varying performance depending on the nature of the data and the algorithm used.

Keywords : Diabetes, Classification, Preprocessing, Feature Selection, Normalization, Cross-Validation.

ملخص

داء السكري هو مرض مزمن، ويُعتبر تشخيصه المبكر ضرورياً لتفادي المضاعفات الخطيرة. في هذا العمل، ندرس تأثير تقنيات المعالجة المسبقة على أداء نماذج التصنيف المطبقة على بيانات مرضى السكري. لتحقيق ذلك، استخدمنا قاعدتي بيانات طبيتين: قاعدة بيانات Pima Indians وقاعدة بيانات محلية عراقية. قمنا باختبار ثلاثة خوارزميات تصنيف: الانحدار اللوجستي، وآلات الدعم الناقل (SVM)، وأشجار القرار. كما قمنا بتطبيق طريقتين للتطبيع (MinMaxScaler StandardScaler) وثلاث تقنيات لاختيار الخصائص (SelectKBest, GenericUnivariateSelect, SelectFromModel). أظهرت النتائج، التي تم تقييمها من خلال التحقق المتقاطع، أن الاختيار المناسب لتقنيات المعالجة المسبقة يُحسن بشكل كبير من دقة النماذج، مع اختلاف في الأداء حسب طبيعة البيانات والخوارزمية المستخدمة.

الكلمات المفتاحية: داء السكري، التصنيف، المعالجة المسبقة، اختيار الخصائص، التطبيع، التحقق المتقاطع.

Table des matières

Table des figures

Liste des tableaux

Introduction Générale

1. Contexte et problématique

Le diabète est une maladie métabolique chronique qui touche des millions de personnes à travers le monde. Son diagnostic précoce et précis est crucial pour prévenir ses complications graves, notamment cardiovasculaires, neurologiques et rénales. Avec l'évolution des technologies de traitement de données médicales, l'utilisation de techniques d'apprentissage automatique (machine learning) permet aujourd'hui de proposer des approches de classification efficaces pour assister les professionnels de santé. Grâce à la richesse des données médicales disponibles, il devient possible de développer des modèles intelligents capables d'identifier les patients à risque, de prédire la progression de la maladie, voire de simuler les effets potentiels de modifications de certains paramètres cliniques.

Cependant, la performance des modèles de classification dépend fortement de la qualité des données et des étapes de prétraitement appliquées en amont. En effet, des techniques telles que la normalisation ou la sélection des caractéristiques pertinentes peuvent significativement améliorer la précision des prédictions. Ainsi, une étude approfondie de l'impact de ces techniques de prétraitement sur différents jeux de données est essentielle pour optimiser les modèles d'aide au diagnostic.

2. Objectif du travail

Dans ce mémoire, nous nous intéressons à l'étude de la classification du diabète à partir de deux bases de données médicales distinctes. L'objectif principal est d'évaluer l'impact des techniques de prétraitement sur la performance des modèles de classification, en particulier sur l'amélioration de l'accuracy.

Les deux jeux de données utilisés sont :

- Une base locale provenant de l'Irak, collectée à partir de dossiers de patients dans deux hôpitaux spécialisés, contenant des données telles que le niveau de sucre sanguin, la créatinine, le cholestérol, le HBA1C, etc., avec une classification en trois catégories : Diabetic, Non-Diabetic, et Predict-Diabetic.
- La base de données Pima Indians du *National Institute of Diabetes and Digestive and Kidney Diseases*, qui contient des mesures diagnostiques de femmes d'origine amérindienne âgées de 21 ans ou plus, avec une classification binaire (diabétique ou non).

Le choix de ces deux bases de données, bien qu'hétérogènes, est pleinement justifié. D'un côté, la base Pima Indians est une référence standard dans les études sur le diabète. De l'autre, la base locale irakienne reflète des données plus réalistes et plus complexes, avec une classification à trois classes. Leur seule caractéristique commune étant la thématique du diabète, leur étude séparée permet de tester la robustesse et l'adaptabilité des techniques appliquées à des contextes très différents.

Dans notre démarche, nous avons appliqué :

- Deux techniques de **normalisation** : *MinMaxScaler* et *StandardScaler*,
- Trois méthodes de **sélection de caractéristiques** :
  - *SelectKBest*,
  - *GenericUnivariateSelect*,
  - *SelectFromModel* (sélection avec modèle), qui utilise l'importance des attributs estimée par un modèle d'apprentissage supervisé,
- Trois modèles de **classification supervisée** :
  - *Régression Logistique (Logistic Regression- LG)*,
  - *Machine à Vecteurs de Support (Support Vector Machine - SVM)*,
  - *Arbre de Décision (Decision Tree AD)*.

Pour garantir une évaluation rigoureuse des performances, nous avons utilisé la **validation croisée** (cross-validation), permettant une estimation fiable de la précision (accuracy) de chaque modèle selon les différents scénarios de prétraitement.

3. Organisation du mémoire

Ce mémoire est structuré comme suit :

- Le **chapitre 1** présente les concepts fondamentaux de la classification supervisée et des modèles utilisés.
- Le **chapitre 2** est dédié aux techniques de prétraitement des données, en particulier la normalisation et la sélection de caractéristiques.
- Le **chapitre 3** décrit notre méthodologie expérimentale, les bases de données utilisées, les résultats obtenus et une discussion critique de ces derniers.
- Enfin, une **conclusion générale** résume les principaux apports du travail et propose des pistes pour des travaux futurs.

Chapitre 1

Introduction à la fouille de données et à la classification supervisée

1.1 Introduction

Dans ce chapitre, nous abordons le processus de fouille de données, une discipline visant à extraire des connaissances utiles à partir de grandes quantités de données. Nous nous concentrons particulièrement sur la classification supervisée, une tâche clé qui consiste à développer des modèles capables de prédire la classe d'observations non étiquetées en se basant sur des exemples préalablement étiquetés. Cette tache revêt une grande importance dans de nombreux domaines, car elle permet de prendre des décisions basées sur les caractéristiques des données, en les classant dans des catégories prédéfinies. Parmi les techniques d'apprentissage supervisé, nous allons nous pencher sur trois techniques à savoir le Support Vector Machine (SVM), les Arbres de Décision (AD) et Régression logistique (LG)

1.2 Définition de la fouille de données

La fouille de données, ou data mining, est une discipline née de la croissance exponentielle des volumes de données et des avancées dans les technologies de stockage et de traitement. Elle vise à extraire des informations utiles et exploitables à partir de grands ensembles de données, permettant de mieux comprendre ces données ou de prévoir leur comportement futur. Elle s'inscrit dans le cadre du processus plus large de découverte de connaissances en bases de

données (Knowledge Discovery from Data ou KDD).[2]..Ce domaine mobilise des méthodes issues de la reconnaissance de formes, des statistiques et des mathématiques pour révéler des corrélations, des schémas et des tendances auparavant inconnus dans les données. Ces connaissances sont utilisées pour améliorer la prise de décision, optimiser les processus et renforcer les performances dans de nombreux secteurs comme la finance, la santé ou le commerce électronique.Elle utilise à la fois des outils manuels et automatiques pour extraire des connaissances utiles, contribuant ainsi à des applications variées telles que l'octroi de crédit, l'optimisation des ressources, le diagnostic médical, l'analyse génomique, et bien d'autres.Ce processus est d'une importance stratégique et économique, permettant d'améliorer la gestion des ressources humaines et matérielles, et de faciliter des applications complexes, comme l'analyse des pratiques commerciales, les moteurs de recherche, l'extraction de textes, et l'analyse des données temporelles.[3].

1.3 Techniques de classification supervisée

Les méthodes de la fouille de données peuvent être classées selon différents critères. Selon le traitement appliqué, elles sont généralement catégorisées en deux grandes classes : les méthodes supervisées et les méthodes non supervisées. Les méthodes supervisées sont utilisées lorsque les données sont étiquetées, ce qui signifie que chaque exemple est associé à une classe ou à une valeur de sortie. Ces méthodes visent à construire un modèle à partir de ces exemples étiquetés pour prédire la classe ou la valeur de sortie d'exemples non étiquetés. Les méthodes non supervisées sont employées lorsque les données ne sont pas étiquetées. Elles cherchent à découvrir des structures ou des régularités dans les données, sans préjuger des classes ou des valeurs de sortie [4].

Dans la démarche de classification supervisée, les classes ainsi que leur nombre sont préalablement définis. L'objectif principal est d'assigner des objets à des classes spécifiques en se basant sur leurs caractéristiques distinctives [2].

Le processus de classification supervisée s'articule autour des étapes suivantes :

1. **Etape de construction du modèle** : Durant étape, l'algorithme apprend à partir des données en créant un ensemble de règles de classification, représentant ainsi le modèle d'apprentissage.

2. **Etape de prédiction** : Pendant cette étape, les données de test sont utilisées pour évaluer l'exactitude des règles de classification établies lors de l'étape précedente. Si le modèle atteint un niveau de précision acceptable, les règles peuvent être appliquées à de nouvelles données.

La construction d'un modèle prédictif suit généralement trois phases distinctes :

- **Phase d'entraînement** : Un échantillon d'entraînement est utilisé pour développer le modèle.
- **Phase de validation** : Un échantillon de validation est employé pour évaluer la performance du modèle sur des données non utilisées dans l'entraînement, afin d'éviter le sur-apprentissage. La performance du modèle peut être évaluée selon différents critères.
- **Phase de test** : Un échantillon de test est utilisé pour évaluer la performance finale du modèle. Cette étape est essentielle pour obtenir une évaluation rigoureuse de l'efficacité du modèle.

Typiquement, les données sont aléatoirement réparties en trois échantillons :

- L'échantillon d'entraînement comprend généralement entre 50% et 80% des données.
- L'échantillon de validation représente entre 20% et 40% des données.
- L'échantillon de test utilise entre 5% et 10% des données (cette phase est parfois omise en pratique).

Cette approche garantit une évaluation précise de la performance du modèle tout en évitant le sur-apprentissage.

Il existe de nombreuses méthodes de classification supervisée ; nous citons quelques-unes et nous ne détaillons que les trois techniques que nous allons utiliser dans notre système :

- Régression logistique « RL »
- Arbres de décision « AD »
- Support victor machines « SVM »

1.3.1 La méthode de classification « Support Vector Machine (SVM) »

La méthode des machines à vecteurs de support (SVM) est une technique de classification supervisée introduite par Vapnik et Cortes en 1995 [2]. Contrairement à KNN, qui classe les points en fonction de leur proximité avec les voisins, SVM cherche à trouver un hyperplan opti-

mal séparant les classes en maximisant la marge entre elles. Cette approche est particulièrement efficace pour les problèmes de classification binaire et peut être étendue aux cas multiclasse à l'aide de techniques comme "one-vs-one" ou "one-vs-all" [5].

1.3.1.1 Principe de la construction

Le principe de l'algorithme SVM repose sur la recherche d'un hyperplan de séparation optimal qui maximise la distance entre les points les plus proches de chaque classe, appelés vecteurs de support [6]. L'algorithme nécessite :

- **Un ensemble de données d'apprentissage** D ;
- **Un noyau** K qui permet de transformer les données dans un espace de plus grande dimension si nécessaire ;
- **Un paramètre** C Un paramètre C qui contrôle le compromis entre maximisation de la marge et minimisation des erreurs

Dans l'exemple d'un modèle SVM (Support Vector Machine) illustré à l'aide d'une séparation de classes dans un espace bidimensionnel. Le SVM cherche à tracer un hyperplan optimal qui maximise la marge entre les points des deux classes : On a

- **deux classes** de données : **Classe 1** (cercles verts) et **Classe 2** (rectangles bleus).
- **L'hyperplan**(ligne noire) sépare les deux classes avec une marge maximale.
- **Les vecteurs de support**sont les points les plus proches de l'hyperplan dans chaque classe.

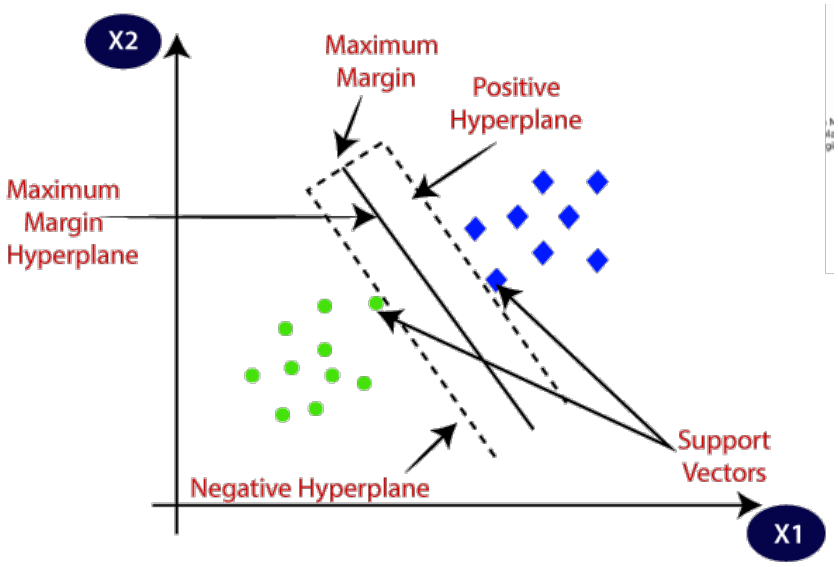

FIGURE 1.1 – exemle du principe de la classification par SVM

1.3.1.2 Hyperplan et vecteurs de support dans l'algorithme SVM

1. **Hyperplan** : il peut y avoir plusieurs lignes/limites de décision pour séparer les classes dans un espace à n dimensions, mais nous devons trouver la meilleure limite de décision qui aide à classer les points de données. Cette meilleure frontière est connue sous le nom d'hyperplan de SVM. Les dimensions de l'hyperplan dépendent des entités présentes dans le jeu de données, ce qui signifie que s'il y a 2 entités, alors l'hyperplan sera une ligne droite. Et s'il y a 3 caractéristiques, alors l'hyperplan sera un plan à 2 dimensions. Nous créons toujours un hyperplan qui a une marge maximale, c'est-à-dire la distance maximale entre les points de données [5]
2. **Vecteurs de soutien** :Les points de données ou vecteurs les plus proches de l'hyperplan et qui affectent la position de l'hyperplan sont appelés vecteurs de support. Puisque ces vecteurs supportent l'hyperplan, donc appelé vecteur de support [5]

1.3.2 Pseudo-Algorithme SVM[1]

Données d'entrée :

- $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ où :
  — $x_i \in R^d$: Vecteur de caractéristiques de l'observation i.
  — $y_i \in \{-1, +1\}$: Étiquette de classe associée.
- $K(x_i, x_j)$: Fonction noyau (kernel).
- $C > 0$: Paramètre de régularisation.

Résultat :

- Modèle de prédiction $f(x)$ pour de nouvelles observations.

Procédure :

1. **Choisir un noyau** $K(x_i, x_j)$ (linéaire, RBF, polynomial, etc.).
2. **Résoudre le problème dual** :

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{1.1}$$

   Sous les contraintes :

$$0 \leq \alpha_i \leq C \quad \text{et} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{1.2}$$

3. **Déterminer le biais** b à partir des vecteurs de support (via les conditions KKT).
4. **Fonction de décision** pour une nouvelle observation x :

$$f(x) = \sum_{i \in SV} \alpha_i y_i K(x_i, x) + b \tag{1.3}$$

   où SV est l'ensemble des indices des vecteurs de support.

5. **Règle de classification** :

$$\text{Si } f(x) > 0, \text{ alors } x \text{ est classé } +1, \text{ sinon } -1.$$

Terme	**Description**
x_i	Vecteur de caractéristiques (features) de dimension d.
y_i	Étiquette de classe (-1 ou $+1$ pour un problème binaire).
$K(x_i, x_j)$	Fonction noyau mesurant la similarité entre x_i et x_j.
C	Paramètre de régularisation contrôlant le compromis biais-variance.
α_i	Multiplicateur de Lagrange associé à chaque point d'entraînement.
SV	Ensemble des vecteurs de support ($\alpha_i > 0$).
b	Biais de l'hyperplan optimal.
$f(x)$	Fonction de décision pour la prédiction.

TABLE 1.1 – Explication des termes utilisés dans l'algorithme SVM.

1.3.2.1 Types de noyaux en SVM

Le choix du noyau est essentiel dans un SVM, car il permet de transformer les données afin de rendre la classification possible même lorsqu'elles ne sont pas séparables linéairement. Voici les noyaux les plus utilisés :

- **Noyau linéaire** : $K(x_i, x_j) = x_i \cdot x_j$ [7].
- **Noyau polynomial** : $K(x_i, x_j) = (x_i \cdot x_j + c)^d$ [8].
- **Noyau gaussien (RBF)** : $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$ [9].
- **Noyau sigmoïde** : $K(x_i, x_j) = \tanh(\alpha x_i \cdot x_j + c)$ [10].

1.3.2.2 Choix du paramètre C et de γ

Le paramètre C est un hyperparamètre crucial dans SVM :

- Un C faible permet une grande marge mais tolère des erreurs de classification.
- Un C élevé pénalise fortement les erreurs, ce qui peut entraîner un surapprentissage (overfitting) [11].

Le paramètre γ (dans le noyau RBF) contrôle l'influence d'un seul exemple d'apprentissage :

- Un γ faible donne un modèle plus général.
- Un γ élevé permet d'adapter le modèle aux données mais peut réduire la capacité de généralisation [12].

1.3.2.3 Avantages de SVM

- Efficace pour les problèmes en haute dimension .
- Bonne performance même avec un petit nombre de données .[6]
- Robuste face aux données bruitées (avec un choix adapté de C).[1]
- S'adapte aux problèmes non linéaires grâce aux noyaux .[10]

1.3.2.4 Inconvénients de SVM

- Sensible au choix des paramètres C et γ, ce qui nécessite un ajustement via validation croisée .

- Temps de calcul élevé pour les grands jeux de données, notamment avec un noyau RBF .[7]
- Moins interprétable que d'autres modèles comme la régression logistique .[11]

1.3.3 La méthode de classification « arbre de décision (AD)»

Les arbres de décision sont une méthode efficace d'apprentissage supervisé utilisée pour classifier des données en fonction de leurs caractéristiques. Cette technique consiste à construire un arbre à partir d'un ensemble de données d'entraînement, où chaque nœud interne représente un test sur une caractéristique des données et chaque feuille représente une décision de classification. Ce modèle offre une grande facilité d'interprétation et d'explication grâce à sa représentation graphique : chaque chemin de la racine à une feuille indique les décisions prises. En outre, les arbres de décision peuvent être utilisés pour évaluer des actions potentielles en fonction de coûts, probabilités et bénéfices, permettant ainsi de déterminer des choix optimaux soit de manière visuelle, soit à travers des algorithmes formels [1].

1.3.3.1 Principe de la construction

Les arbres de décision sont des structures où chaque feuille représente une valeur de la variable-cible, et chaque embranchement correspond à une combinaison de variables d'entrée. Le processus de construction commence en plaçant tous les points de la base d'apprentissage dans le nœud racine, puis divise récursivement chaque nœud en fonction de la valeur d'un attribut testé à chaque étape. L'objectif est d'obtenir des sous-ensembles d'exemples contenant principalement des exemples appartenant à la même classe. Cela conduit à une construction top-down de l'arbre, de la racine vers les feuilles.[4] [2]

Algorithm 1: Pseudo-Algorithme de Construction d'un Arbre de Décision

1 **Début :**
- Initialiser l'arbre courant à l'arbre vide : la racine est le nœud courant;

Répéter :
- Décider si le nœud courant est terminal;
  - **Si** le nœud est terminal alors lui affecter une classe;
  - **Sinon** sélectionner un test créer autant de nouveaux nœuds fils qu'il y a de réponses possibles au test;
- Passer au nœud suivant non exploré s'il en existe jusqu'à un arbre de décision;

fin

Dans cet exemple (voir figure 1.2), le nœud racine est représenté par la condition "température > 37 °C". Les feuilles de l'arbre correspondent aux différentes classes ou décisions, telles que "malade" ou "sain". Les nœuds intermédiaires, tels que "température > 37 °C" et "toux", sont appelés attributs ou variables [13].

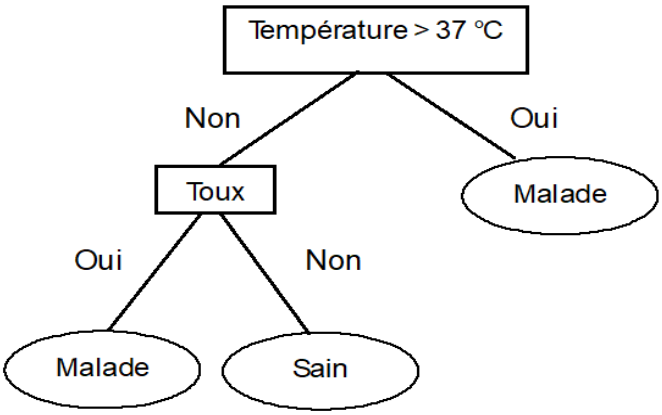

FIGURE 1.2 – Exemple d'un arbre de décision.

1.3.3.2 Les mesures de sélection d'attributs

Les mesures de sélection d'attributs sont cruciales pour choisir les attributs à chaque niveau de l'arbre. Différentes mesures sont utilisées, telles que l'indice d'impureté de Gini, le gain d'information et la réduction de la variance. Ces mesures aident à décider quelle est la meilleure façon de répartir les données à chaque niveau de l'arbre pour maximiser sa performance [4][7].

1. **L'indice de Gini** mesure la probabilité qu'un élément choisi au hasard dans un nœud soit mal classé s'il était classé aléatoirement en fonction de la distribution des classes dans le nœud.

$$Gini\ (P) = 1 - \sum_{k=1}^{c} (P(k/P))^2$$

2. **L'entropie** mesure le désordre dans un ensemble de données. Un ensemble parfaitement homogène a une entropie de zéro, tandis qu'un ensemble avec une répartition égale de toutes les classes à une entropie maximale.

$$Entropie\,(p) = -\sum_{k=1}^{c} P(k/p)\log(k/p)$$

3. **Le gain d'information** mesure la réduction de l'entropie obtenue en divisant les données selon une caractéristique particulière. Il représente la quantité d'information gagnée en divisant les données par rapport à cette caractéristique.

$$Gain(p,i) = i(p) - \sum_{j=1}^{n} pj.i(pj)$$

1.3.3.3 Choix de la bonne taille de l'arbre

En pratique, choisir la bonne taille pour un arbre de décision est crucial. Un arbre trop complexe, avec de nombreuses branches et feuilles correspondant à des sous-ensembles parfaitement homogènes, peut-être trop spécifique aux données d'entraînement. Cela peut entraîner une mauvaise généralisation du modèle à de nouvelles données, limitant ainsi sa capacité à rendre compte de la réalité que l'on cherche à modéliser. Il est donc important de trouver un équilibre entre la complexité de l'arbre et sa capacité à généraliser les prédictions à de nouvelles observations [4].

1.3.3.4 Avantages

- Décisions aisément interprétables.
- Classification très rapide.
- Facilité à manipuler des données catégoriques.
- Traitement facile des variables d'amplitudes très différentes.

1.3.3.5 Inconvénients

- La sensibilité au bruit et aux points aberrants.
- La sensibilité au nombre de classes (plus le nombre de classes est grand plus les performances diminuent).

1.3.4 La méthode de classification « Régression Logistique (RL) »

La régression logistique est une technique d'apprentissage supervisé largement utilisée pour résoudre des problèmes de classification. Contrairement à la régression linéaire qui prédit des valeurs continues, la régression logistique permet de modéliser la probabilité qu'une observa-

tion appartienne à une classe particulière. Elle est particulièrement adaptée aux cas de classification binaire, mais peut être généralisée aux situations à plusieurs classes. Son fonctionnement repose sur l'utilisation de la fonction logistique (ou sigmoïde), qui transforme une combinaison linéaire des variables explicatives en une probabilité [2, 5].

Ce modèle est apprécié pour sa simplicité d'interprétation, sa robustesse ainsi que sa capacité à être combiné avec des techniques de régularisation (comme L1 ou L2). Dans un contexte médical, par exemple, il permet de prédire la présence ou l'absence d'une maladie à partir de caractéristiques cliniques mesurées

1.3.4.1 Principe de fonctionnement

La régression logistique est un modèle statistique permettant de prédire la probabilité d'appartenance d'une observation à une classe donnée. Contrairement à la régression linéaire, elle est conçue pour des variables cibles qualitatives (catégoriques), notamment dans des contextes de classification binaire.[2, 6]

Le modèle utilise la fonction logistique (ou sigmoïde) pour transformer une combinaison linéaire des variables d'entrée en une probabilité. La formule de la régression logistique est :

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

où :

- $P(y = 1 \mid x)$ est la probabilité que l'événement $y = 1$ se produise,
- β_0 est le biais (constante),
- β_i sont les coefficients associés aux variables explicatives x_i.

Une fois la probabilité estimée, un seuil de décision (souvent 0.5) est appliqué : si $P(y = 1 \mid x) \geq 0.5$, alors l'observation est classée dans la classe 1, sinon dans la classe 0.

Les paramètres β sont estimés par la méthode de maximisation de la vraisemblance, qui permet d'optimiser l'ajustement du modèle aux données observées.

1.3.4.2 Étapes de modélisation

Algorithme 2 : Pseudo-algorithme de la Régression Logistique

- **Début :**
  - Initialiser les coefficients β (aléatoirement ou à zéro)
- **Répéter :**
  - Calculer la probabilité prédite $\hat{y}$ à l'aide de la fonction logistique
  - Calculer le coût (log-vraisemblance)
  - Mettre à jour les coefficients par descente de gradient ou autre méthode d'optimisation
- **Jusqu'à convergence des coefficients**
- **Fin :** Utiliser les $\hat{y}$ pour prédire la classe selon un seuil (souvent 0.5) [5, 1]

1.3.4.3 Avantages

- Permet une interprétation directe des coefficients sous forme de rapports de cotes (*odds ratio*).
- Modèle simple, rapide à entraîner et facile à implémenter.
- Convient aux données binaires et largement utilisé dans la pratique.
- Gère à la fois les variables explicatives continues et catégorielles.
- Offre une base solide pour l'analyse prédictive et l'inférence statistique.[3, 1]

1.3.4.4 Inconvénients

- Suppose une relation linéaire entre les variables explicatives et le logit de la probabilité.
- Moins performant pour les problèmes à forte non-linéarité ou à interactions complexes.
- Sensible aux valeurs aberrantes et aux multicolinéarités.
- La performance peut chuter si les classes sont déséquilibrées.[2, 5]

1.4 Conclusion

Dans ce chapitre, nous avons exploré le processus de fouille de données, en soulignant son rôle fondamental dans l'extraction de connaissances à partir de vastes ensembles de données. Nous nous sommes concentrés sur la classification supervisée, une approche essentielle qui

permet de construire des modèles prédictifs à partir de données annotées.

Nous avons analysé trois techniques majeures d'apprentissage supervisé : la machine à vecteurs de support (SVM), les arbres de décision (AD) et Régression Logistique (RL). Chacune de ces méthodes possède ses propres atouts et limites en termes d'interprétabilité, de précision, de performance et de sensibilité aux données.

Nous avons également abordé la méta-classification, une approche qui consiste à combiner les prédictions de plusieurs classificateurs afin d'améliorer la robustesse et la fiabilité des modèles. Cette méthode est particulièrement précieuse dans des contextes complexes où une seule technique de classification pourrait ne pas être suffisante.

Ces différentes approches constituent des outils puissants pour la construction de modèles prédictifs, notamment dans des domaines critiques comme le diagnostic médical. Le choix de la meilleure méthode dépend des caractéristiques spécifiques des données et des objectifs fixés. En alliant une compréhension approfondie de ces techniques à des applications concrètes, nous pouvons exploiter pleinement le potentiel de la fouille de données pour prendre des décisions éclairées et anticiper les tendances futures dans divers domaines.

Chapitre 2

Prétraitement des Données, Normalisation et Sélection des Caractéristiques

2.1 Introduction

Le prétraitement des données est une étape essentielle dans toute analyse de données et dans les projets d'apprentissage automatique (machine learning). Les données brutes recueillies sont souvent incomplètes, incohérentes ou mal formatées, ce qui peut nuire à la performance des modèles. Le prétraitement permet de préparer les données pour une analyse plus précise et efficace. La normalisation et la sélection des caractéristiques ou de *feature selection* en anglais sont des sous-étapes clés du prétraitement, permettant de réduire la complexité des données tout en maintenant leur pertinence pour les modèles d'apprentissage.

Dans ce chapitre, nous aborderons les étapes de prétraitement des données, en mettant l'accent sur la normalisation, la standardisation et la sélection des caractéristiques. Ces processus sont cruciaux pour assurer la qualité et la pertinence des données utilisées dans les modèles de machine learning.

2.2 Prétraitement des Données

2.2.1 Nettoyage des Données

Le nettoyage des données consiste à identifier et corriger les erreurs dans un jeu de données. Cela inclut la gestion des valeurs manquantes, des doublons, des erreurs de saisie et des anomalies. Des techniques comme l'imputation (remplir les valeurs manquantes avec des estimations basées sur les autres données) sont souvent utilisées pour traiter ces problèmes.

2.2.2 Transformation des Données

Les transformations visent à rendre les données plus appropriées pour l'analyse. Cela inclut la conversion des données catégorielles en variables numériques (par exemple, à l'aide de l'encodage one-hot) ou la transformation des variables non linéaires en variables linéaires via des méthodes telles que la prise du logarithme.

2.2.3 Gestion des Valeurs Manquantes

Les valeurs manquantes peuvent être imputées de diverses façons : en utilisant des méthodes simples comme la moyenne, la médiane ou des techniques plus complexes comme l'imputation par KNN (k-plus proches voisins). Le choix de la méthode dépend du type de données et de la quantité d'informations manquantes.

2.2.4 Gestion des Outliers

Les valeurs aberrantes (outliers) peuvent fausser les résultats de l'analyse. Il est donc crucial de les détecter et de décider de la manière de les traiter. Les méthodes courantes pour gérer les outliers incluent leur suppression ou leur transformation (par exemple, en utilisant la normalisation).

2.3 Normalisation des Données

2.3.1 Importance de la Normalisation

La normalisation des données est une étape de prétraitement indispensable dans de nombreux algorithmes d'apprentissage automatique. Elle consiste à transformer les valeurs des variables numériques afin qu'elles soient représentées sur une échelle comparable. En effet, lorsque les attributs d'un jeu de données présentent des unités ou des plages de valeurs très différentes (par exemple, des distances en kilomètres et des masses en grammes), les modèles peuvent accorder une importance disproportionnée à certaines variables, non pas à cause de leur pertinence, mais en raison de leur échelle. La normalisation permet de rendre les données comparables, en réduisant l'influence disproportionnée de certaines variables.

Cette transformation permet donc de réduire les biais induits par l'amplitude des variables et d'améliorer la convergence des algorithmes d'optimisation. Les modèles basés sur les distances (comme k-NN ou SVM) ou sur des gradients (comme les réseaux de neurones) sont particulièrement sensibles à cette problématique.

2.3.2 Méthodes de Normalisation

Il existe plusieurs techniques de normalisation qui peuvent être utilisées en fonction des caractéristiques des données et des exigences du modèle. Voici les trois principales méthodes utilisées dans ce travail :

- **Min-Max Scaling** : Cette méthode de normalisation transforme les données pour qu'elles soient dans une plage fixe, généralement entre 0 et 1, ou toute autre plage définie. Elle est définie par la formule suivante :

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

  Fonctionnement : Cette technique calcule la différence entre chaque valeur X et la valeur minimale X_{min}, puis la divise par l'écart entre la valeur maximale X_{max} et la valeur minimale. Ce processus rescale toutes les données dans une échelle commune sans altérer les relations entre les observations. Cependant, elle est sensible aux valeurs extrêmes (outliers), qui peuvent fortement influencer les résultats de la normalisation.

- **Standardisation (Z-score)** : La standardisation permet de transformer les données de manière à ce qu'elles aient une moyenne de 0 et un écart-type de 1. Elle est calculée avec la formule suivante :

$$X_{\text{std}} = \frac{X - \mu}{\sigma}$$

où μ est la moyenne et σ l'écart-type de la variable X.

Fonctionnement : La standardisation est particulièrement utile lorsque les données suivent une distribution normale ou lorsque l'on travaille avec des algorithmes qui sont sensibles à la variance, comme les k-plus proches voisins (k-NN) ou les réseaux de neurones. Elle permet de centrer les données autour de zéro et de les ajuster pour qu'elles aient une variance égale à 1. Ce processus réduit l'impact des valeurs extrêmes tout en maintenant les relations entre les variables.

- **Normalisation par la norme L2** : Cette méthode consiste à diviser chaque observation par sa norme L2, ce qui permet de rescaler les données pour que leur norme soit égale à 1. La formule est la suivante :

$$X_{\text{norm\_L2}} = \frac{X}{\|X\|_2}$$

où $\|X\|_2$ est la norme L2 de l'observation X, définie comme la racine carrée de la somme des carrés des valeurs de l'observation.

Fonctionnement : La normalisation L2 est couramment utilisée dans les modèles de type réseau de neurones, où chaque observation est "réajustée" de manière uniforme. Elle permet de donner à toutes les observations une échelle comparable, sans qu'aucune variable ne prenne le pas sur les autres en raison de sa valeur absolue. Cette méthode est particulièrement efficace lorsque les données sont très disparates ou de grande dimension.

2.3.3 Méthodes Utilisées dans notre Travail

Dans notre travail, nous avons utilisé deux méthodes de normalisation afin d'évaluer leur impact sur la performance des modèles de classification : la normalisation Min-Max et la standardisation Z-score.

- La **normalisation Min-Max** a été choisie pour sa capacité à maintenir l'échelle relative des variables tout en les ramenant dans un intervalle uniforme. Elle a notamment

été appliquée dans les expérimentations avec les modèles de classification utilisant des distances.

- La **standardisation Z-score**, quant à elle, a permis de traiter les cas où il était préférable de centrer les données et de les homogénéiser en termes de variance.

La normalisation doit être appliquée avant l'entraînement des modèles d'apprentissage automatique, notamment pour les algorithmes sensibles aux distances (comme les k-NN et SVM). Il est essentiel d'utiliser les mêmes paramètres de normalisation pour les données d'entraînement et de test afin d'éviter toute fuite de données (data leakage).

2.4 Sélection des Caractéristiques

La sélection des caractéristiques, également appelée sélection d'attributs ou *feature selection*, est une étape fondamentale du prétraitement des données. Elle consiste à identifier, parmi un ensemble initialement large de variables, celles qui sont les plus pertinentes pour une tâche d'apprentissage donnée. Cette étape permet de réduire la dimensionnalité des données, d'éliminer les variables redondantes, bruitées ou non informatives, tout en conservant les plus discriminantes. Ce processus est utilisé dans divers domaines, tels que l'apprentissage automatique, la bioinformatique, la vision par ordinateur ou encore la médecine. Dans ce dernier domaine, il joue un rôle crucial dans l'identification des symptômes ou caractéristiques les plus utiles pour le diagnostic, à partir de données patient comprenant des résultats cliniques, biologiques ou d'imagerie.

2.4.1 Objectifs de la sélection des caractéristiques

La sélection des caractéristiques poursuit plusieurs objectifs :

- **Réduction de la complexité du modèle** : Moins de variables signifie des modèles plus simples et plus rapides à entraîner.
- **Amélioration de la généralisation** : En éliminant le bruit, on réduit le sur-apprentissage (*overfitting*).
- **Facilité d'interprétation** : Un sous-ensemble de caractéristiques sélectionnées est souvent plus compréhensible.
- **Réduction du coût computationnel** : Moins de dimensions implique moins de res-

sources nécessaires.

2.4.2 Défis liés à la sélection des caractéristiques

Les principaux défis de la sélection des caractéristiques sont les suivants :

- **La malédiction de la dimensionnalité** : L'augmentation du nombre de caractéristiques rend les données éparses, accroît les temps de calcul et peut nuire aux performances des algorithmes d'apprentissage (*fléau de la dimension*) [14, 15].
- **Pertinence des attributs** : Il est difficile d'identifier les attributs véritablement informatifs pour la classification. Les attributs pertinents sont ceux dont les valeurs varient significativement selon les classes cibles [16, 17].
- **Redondance des attributs** : Des caractéristiques peuvent être fortement corrélées ou contenir la même information. Cela ajoute une complexité inutile et peut affecter l'interprétation du modèle. Des mesures telles que la corrélation ou l'information mutuelle permettent de détecter cette redondance [18, 19].

2.4.3 Avantages et Inconvénients de la sélection d'attributs

La sélection d'attributs offre plusieurs bénéfices, mais comporte aussi des limites [18, 19, 20].

2.4.3.1 Avantages

- **Simplification du modèle :** Moins d'attributs rendent le modèle plus simple et plus facile à interpréter.
- **Amélioration des performances :** Une meilleure qualité des données améliore la précision et réduit le temps de calcul.
- **Réduction du bruit et des redondances :** Elle permet d'éliminer les attributs inutiles ou corrélés.

2.4.3.2 Inconvénients

- **Risque de perte d'information :** Des attributs utiles peuvent être écartés.
- **Complexité ajoutée :** La sélection elle-même peut être coûteuse et sensible aux critères choisis.

- **Surajustement :** Une mauvaise sélection peut nuire à la généralisation du modèle.

2.4.4 Processus de Sélection des Caractéristiques

La sélection d'attributs suit généralement quatre étapes principales [20], comme illustré dans la figure suivante :

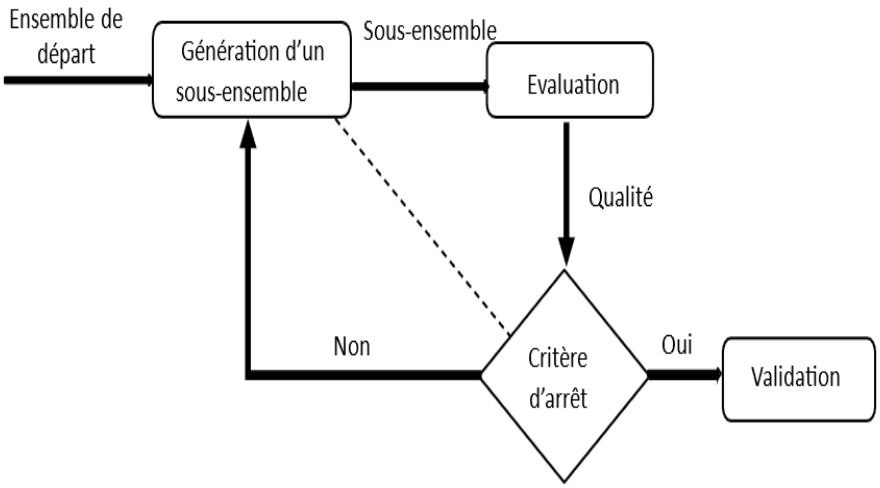

FIGURE 2.1 – Procédure générale pour la sélection d'attributs

- **Génération de sous-ensembles** : Cette étape consiste à générer différents sous-ensembles d'attributs pour être évalués. Cela peut être fait via des techniques de recherche exhaustive ou des méthodes heuristiques.
- **Évaluation des sous-ensembles** : Chaque sous-ensemble d'attributs est évalué en fonction de son impact sur la performance du modèle. La validation croisée est souvent utilisée pour estimer la performance.
- **Critère d'arrêt** : Le processus de sélection s'arrête lorsque l'ajout de nouvelles caractéristiques n'améliore plus la performance du modèle ou lorsque le nombre d'attributs atteint un seuil prédéfini.
- **Validation** : Une fois la sélection des caractéristiques terminée, il est important de valider que les caractéristiques sélectionnées ne conduisent pas à un surajustement et qu'elles génèrent bien des modèles performants sur des données nouvelles.

Cette approche en quatre étapes offre un cadre systématique pour mener efficacement la sélection d'attributs [21, 22].

2.4.5 Méthodes de Sélection des Caractéristiques

Les méthodes de sélection d'attributs peuvent être classées en trois grandes catégories : les méthodes filtrantes, les méthodes enveloppantes (*wrapper*) et les méthodes intégrées (*embedded*) Nous présentons ici une revue des principales approches utilisées dans chaque catégorie.

2.4.5.1 Méthodes filtrantes (*Filter methods*)

Ces méthodes évaluent chaque caractéristique indépendamment du modèle en utilisant des critères statistiques tels que la corrélation, le test du chi carré, l'information mutuelle, etc. Elles sont rapides et efficaces pour éliminer les caractéristiques non pertinentes, mais elles ne prennent pas en compte les interactions entre les caractéristiques. Les méthodes filtrantes évaluent chaque attribut indépendamment des modèles d'apprentissage utilisés. Elles se basent généralement sur des critères statistiques ou des mesures d'information. Ces techniques sont rapides et simples à mettre en œuvre, ce qui les rend particulièrement adaptées aux jeux de données de grande dimension.

- **Sélection par test du chi2** : Utilise le test du chi-carré pour évaluer la dépendance entre chaque caractéristique et la variable cible. Sélectionne les caractéristiques les plus importantes selon leur score. Adaptée aux caractéristiques catégorielles. Implémentée en Python sous le nom de **percentile** [23].
- **Sélection avec chi2 en mode k_best** : Utilise `GenericUnivariateSelect` pour sélectionner les k meilleures caractéristiques en fonction du score chi2. Permet de contrôler le nombre d'attributs sélectionnés. [24].
- **Sélection par ANOVA** : Utilise le test ANOVA pour évaluer la relation entre chaque caractéristique et la variable cible. Sélectionne les k meilleures caractéristiques en fonction de leur score d'ANOVA. Adaptée aux caractéristiques continues. Implémentée en Python sous **SelectKBest** [25].
- **Relief** : Mesure la pertinence des caractéristiques en comparant les distances entre des exemples et leurs voisins de même ou d'autre classe. Efficace sur des données bruitées mais peut manquer de cohérence et ignorer les corrélations [26, 22].
- **Best First Search** : Explore l'espace des caractéristiques en ajoutant progressivement celles les plus pertinentes, et peut revenir en arrière pour explorer des sous-ensembles prometteurs. [22].

- **FOCUS** : Effectue une recherche exhaustive sur l'ensemble des caractéristiques. Sensible au bruit et lent dans l'évaluation des sous-ensembles [22].
- **Branch & Bound** : Évalue chaque sous-ensemble de caractéristiques en les enlevant une à une. La variante ABB utilise une mesure monotone pour une sélection plus efficace [27].
- **Las Vegas Filter (LVF)** : Utilise un critère d'incohérence pour éliminer les sous-ensembles non pertinents, garantissant un sous-ensemble optimal, mais avec un coût computationnel plus élevé [28].

Avantages généraux : Les méthodes filtrantes sont rapides, indépendantes du modèle, et sont particulièrement utiles dans le cadre du prétraitement des données.

Inconvénients généraux : Elles ne tiennent pas compte des interactions complexes entre les variables et du comportement du modèle, ce qui peut limiter leur efficacité dans certains cas.

2.4.5.2 Méthodes enveloppantes (*Wrapper methods*)

Ces méthodes évaluent les sous-ensembles d'attributs en entraînant un modèle de machine learning sur chaque sous-ensemble. Ces méthodes offrent généralement de meilleures performances, mais elles sont plus coûteuses en termes de calcul. L'une des approches les plus courantes est la recherche en avant ou en arrière, où des sous-ensembles sont ajoutés ou supprimés en fonction de l'amélioration de la performance du modèle Les méthodes enveloppantes utilisent un algorithme d'apprentissage pour évaluer différentes combinaisons d'attributs. Elles tiennent compte des interactions entre variables, mais sont coûteuses en temps de calcul.

- **Sequential Forward Selection (SFS)** La méthode de Sélection Avant Séquentielle (SFS) est une approche heuristique de recherche. Elle commence avec un ensemble vide et ajoute successivement une caractéristique jusqu'à ce que le critère d'arrêt soit satisfait. Cette méthode était utilisée pour réduire la taille des données et améliorer les résultats de classification [20].
- **Sequential Backward Selection (SBS)** La méthode de Sélection Arrière Séquentielle (SBS) consiste à commencer avec l'ensemble complet de toutes les caractéristiques, puis à procéder à la suppression successive de ces caractéristiques. Bien que cette technique soit plus performante que la précédente SFS, son principal inconvénient réside

dans son temps de calcul plus important [20].

- **L'algorithme des essaims de lucioles (Firefly algorithm)** L'algorithme des essaims de lucioles, inspiré par le comportement lumineux des lucioles, est une technique d'optimisation globale basée sur la métaheuristique. Les lucioles, dans cet algorithme, représentent des solutions potentielles qui se déplacent dans l'espace de recherche en suivant certaines règles d'attraction et de luminosité [4].
- **Optimisation par essaims de particules (PSO)** L'optimisation par essaim de particules repose sur un ensemble d'individus, appelés particules, initialement répartis de manière aléatoire et homogène dans l'espace de recherche. Chaque particule est une solution potentielle et possède une mémoire contenant sa meilleure solution visitée jusqu'à présent. Elle a la capacité de communiquer avec les particules voisines. Avec ces informations, chaque particule ajuste sa position en suivant une tendance qui combine sa volonté de retourner vers sa meilleure solution locale et son mimétisme par rapport aux solutions de son voisinage. En combinant les optimums locaux et empiriques, l'ensemble des particules converge généralement vers la solution optimale globale du problème traité [13].

Avantages généraux Les méthodes enveloppantes prennent en compte l'interaction entre les caractéristiques et le modèle. Elles sont adaptées pour trouver le sous-ensemble optimal de caractéristiques pour un modèle spécifique et peuvent potentiellement améliorer les performances du modèle.

Inconvénients généraux Les méthodes enveloppantes sont coûteuses en temps de calcul, car elles nécessitent plusieurs évaluations du modèle. Elles présentent également un risque de sur-apprentissage si le critère d'arrêt n'est pas correctement ajusté.

2.4.5.3 Méthodes intégrées (*Embedded methods*)

Les méthodes intégrées réalisent la sélection des caractéristiques directement pendant le processus d'entraînement du modèle. Ces méthodes sont particulièrement intéressantes car elles combinent l'entraînement du modèle avec la sélection des caractéristiques, ce qui permet une sélection plus fine et adaptée au modèle. Des exemples de ces méthodes incluent l'importance des caractéristiques dans les arbres de décision comme les forêts aléatoires, ou encore l'utilisation de la régularisation, telle que Lasso pour la régression linéaire.

- **Régularisation Lasso (L1)** : Cette méthode de régularisation pénalise les coefficients de régression en leur appliquant une norme L1. Elle pousse certains coefficients à zéro, ce qui permet d'effectuer une sélection implicite des caractéristiques les plus pertinentes [29].
- **Régularisation Ridge (L2)** : Contrairement au Lasso, la régularisation Ridge applique une pénalité L2 sur les coefficients de régression, les réduisant sans les annuler complètement [30].
- **Elastic Net** : Cette méthode combine les régularisations L1 et L2, apportant ainsi une flexibilité accrue pour la sélection des caractéristiques [31].
- **Importance des caractéristiques dans les forêts aléatoires ou XGBoost** : Ces modèles attribuent une importance à chaque caractéristique en fonction de leur contribution à la réduction de l'impureté (comme l'indice de Gini ou l'entropie) dans les arbres de décision [32, 33].

Avantages généraux : Les méthodes intégrées permettent une sélection des caractéristiques plus efficace et directement adaptée au modèle. Elles sont souvent plus efficaces que les méthodes filtrantes et enveloppantes, car elles prennent en compte l'interaction entre les caractéristiques et le modèle.

Inconvénients généraux : Cependant, ces méthodes peuvent être plus complexes à mettre en œuvre et nécessitent un bon paramétrage du modèle. De plus, comme la sélection des caractéristiques est intégrée au processus d'entraînement, elles peuvent être coûteuses en temps de calcul, particulièrement pour des jeux de données volumineux.

Le choix de la méthode de sélection d'attributs dépend du contexte, du type de données, des ressources disponibles et du modèle utilisé. Une combinaison judicieuse de plusieurs approches peut souvent aboutir à de meilleures performances globales.

2.5 Conclusion

Le prétraitement des données constitue une étape essentielle dans tout projet d'apprentissage automatique. Parmi ses composantes fondamentales, la normalisation et la sélection des caractéristiques jouent un rôle central dans l'amélioration des performances des modèles. La normalisation permet de rendre les données comparables en les ramenant sur une échelle com-

mune, ce qui est crucial pour de nombreux algorithmes sensibles à la distance ou à la variance. La sélection des caractéristiques, quant à elle, vise à identifier les variables les plus pertinentes, réduisant ainsi la dimensionnalité, le bruit et la complexité computationnelle.

Ce chapitre a mis en lumière les défis liés à la sélection d'attributs, notamment la croissance de la dimension des données et la difficulté à évaluer la pertinence des variables. Plusieurs approches ont été présentées, notamment les méthodes filtre et wrapper, chacune ayant ses avantages et limites selon le contexte d'analyse. Une sélection rigoureuse des caractéristiques permet non seulement d'améliorer la qualité prédictive des modèles, mais aussi de faciliter leur interprétation.

En comprenant l'importance de choisir judicieusement les attributs et en explorant les diverses méthodes disponibles, les praticiens peuvent améliorer la qualité de leurs modèles et prendre des décisions plus éclairées dans leurs domaines respectifs.

Chapitre 3

Conception, Réalisation et Modélisation

3.1 Introduction

L'intégration de l'apprentissage automatique dans le domaine médical constitue aujourd'hui un levier majeur pour améliorer la qualité des diagnostics et favoriser l'émergence de traitements personnalisés. Dans le cadre de cette étude, nous avons exploité le potentiel de plusieurs techniques de classification supervisée pour prédire l'état de santé de patients en lien avec le diabète, à partir de deux jeux de données médicaux distincts : la base Pima Indians, largement utilisée dans la littérature scientifique, et une base locale issue d'hôpitaux en Irak, plus complexe et représentative de la réalité terrain.

Notre objectif principal est de construire des modèles d'apprentissage robustes capables d'apporter une aide fiable au diagnostic médical. Pour ce faire, nous avons mobilisé trois algorithmes d'apprentissage supervisé : la régression logistique (LG), la machine à vecteurs de support (SVM) et l'arbre de décision (AD), appliqués séparément sur chaque base de données.

Conscients de l'importance du prétraitement des données dans la qualité des résultats, nous avons intégré plusieurs étapes clés dans notre pipeline expérimental. Les modèles d'apprentissage étant particulièrement sensibles à la structure, à l'échelle et à la pertinence des variables d'entrée, un prétraitement adéquat permet non seulement de nettoyer les données, mais aussi de les transformer de manière à les rendre mieux adaptées à l'apprentissage.

Dans ce chapitre, nous étudions en détail l'impact de deux étapes fondamentales du prétraitement :

La normalisation des données, pour laquelle nous avons appliqué deux méthodes courantes : MinMaxScaler et StandardScaler.

La sélection d'attributs, que nous avons abordée à travers deux approches complémentaires : les méthodes filtrées, qui évaluent les caractéristiques indépendamment du modèle, et les méthodes basées sur des modèles, qui sélectionnent les attributs les plus informatifs en fonction de leur contribution aux performances.

Ces techniques jouent un rôle essentiel dans l'amélioration de l'efficacité des modèles, en particulier ceux sensibles à l'échelle ou à la dimensionnalité, comme les SVM et la régression logistique. Afin d'assurer la fiabilité de notre évaluation, nous avons appliqué une procédure rigoureuse de validation croisée à chaque expérimentation.

Nous avons mené cette étude sur deux bases de données très différentes mais complémentaires, toutes deux liées à la problématique du diabète. Cette double analyse nous permet d'évaluer la robustesse de nos conclusions et la capacité de généralisation des résultats dans des contextes variés. Notre objectif final est d'identifier la combinaison optimale entre techniques de normalisation et de sélection d'attributs, afin d'améliorer significativement les performances des modèles d'apprentissage automatique dans la prédiction du diabète.

L'ensemble de notre démarche méthodologique est synthétisé dans la figure suivante :

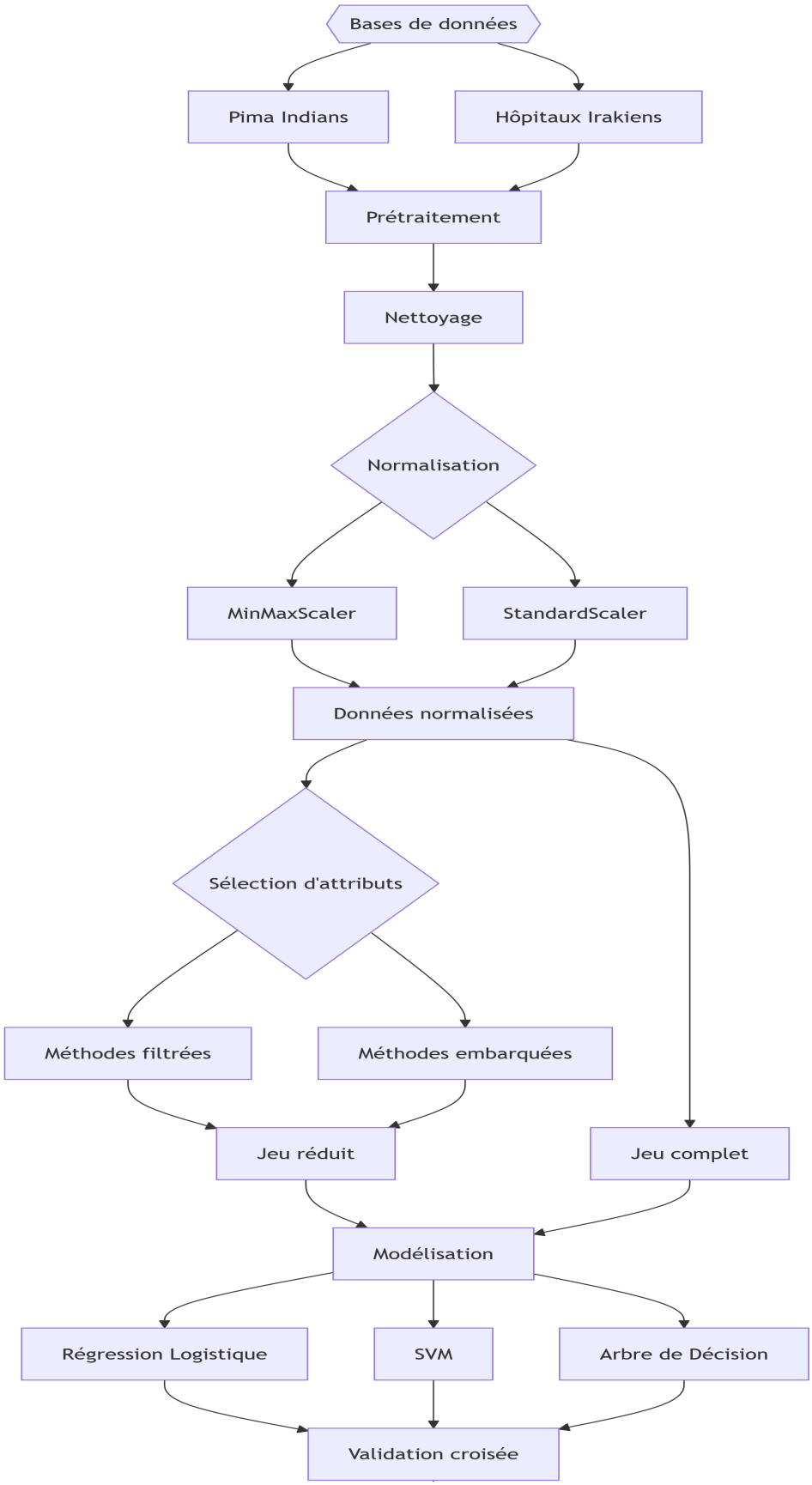

FIGURE 3.1 – Schéma Global de notre travail

3.2 Bases de données utilisées

Dans le cadre de notre étude, nous avons utilisé deux bases de données cliniques pour entraîner, tester et comparer les performances des algorithmes de classification appliqués à la détection du diabète. La première est une base de données locale, collectée dans des hôpitaux en Irak, qui propose une classification en trois catégories, tandis que la seconde est la célèbre base *Pima Indians Diabetes*, fréquemment utilisée dans la littérature pour la classification binaire du diabète.

3.2.1 La base de données médicale locale Irak sur le diabète

Cette base de données a été constituée en Irak à partir des dossiers médicaux de deux hôpitaux spécialisés. Elle regroupe des informations cliniques de patients diagnostiqués ou à risque de diabète. Contrairement aux jeux de données standards, cette base comporte une classification en trois catégories : **Non-Diabetic**, **Predict-Diabetic** et **Diabetic**, ce qui permet une détection plus fine des cas à surveiller, notamment ceux susceptibles d'évoluer vers la maladie.

Les attributs mesurés incluent notamment :

- **Age** : Âge du patient.
- **Gender** : Sexe.
- **Blood Sugar** : Niveau de glucose sanguin.
- **HBA1C** : Moyenne du taux de glucose sur les trois derniers mois.
- **Creatinine** : Indicateur de la fonction rénale.
- **Cholesterol** : Taux de cholestérol.
- **Triglycerides** : Taux de triglycérides.
- **Blood Pressure** : Tension artérielle.
- **Diagnosis** : Classe cible (Non-Diabetic, Predict-Diabetic, Diabetic).

Cette base présente un intérêt particulier en raison de sa granularité à trois classes et de son origine locale, ce qui permet une évaluation des modèles sur des données réelles dans un contexte médical régional.

3.2.2 La base de données Pima Indians Diabetes

La base de données *Pima Indians Diabetes* est fournie par le *National Institute of Diabetes and Digestive and Kidney Diseases*. Elle a été construite à partir d'un échantillon de femmes d'origine amérindienne âgées d'au moins 21 ans. L'objectif est de prédire la présence ou non du diabète à partir de variables cliniques mesurées lors de consultations médicales.

Les attributs mesurés sont les suivants :

- **Pregnancies** : Nombre de grossesses.
- **Glucose** : Concentration plasmatique en glucose.
- **Blood Pressure** : Pression artérielle diastolique.
- **Skin Thickness** : Épaisseur du pli cutané tricipital.
- **Insulin** : Taux d'insuline sérique.
- **BMI** : Indice de masse corporelle.
- **Diabetes Pedigree Function** : Antécédents familiaux du diabète.
- **Age** : Âge de la patiente.
- **Outcome** : Classe cible (0 = non diabétique, 1 = diabétique).

Cette base, largement utilisée dans les travaux d'apprentissage automatique, constitue une référence pour la classification binaire du diabète. Elle permet d'évaluer et de comparer les performances des modèles dans un cadre standardisé.

3.3 Techniques d'évaluation des résultats

Dans cette section, nous présentons les méthodes utilisées pour évaluer les performances des différents classificateurs appliqués dans notre étude, à savoir l'Arbre de Décision (AD), la Régression Logistique (LG) et la Machine à Vecteurs de Support (SVM). L'objectif est de mesurer la capacité de généralisation de chaque modèle à partir des bases de données sélectionnées.

3.3.1 Validation croisée

La validation croisée est une méthode largement utilisée pour estimer la performance d'un modèle d'apprentissage automatique sur des données indépendantes. Plutôt que d'utiliser une unique séparation des données en un ensemble d'entraînement et un ensemble de test, cette

approche consiste à diviser les données en k sous-ensembles appelés *folds*. Le processus est itératif : à chaque itération, un fold est utilisé comme ensemble de validation, tandis que les $k-1$ autres servent à l'entraînement. Ce mécanisme est répété k fois, de sorte que chaque sous-ensemble est utilisé une fois pour la validation.

Dans notre étude, nous avons appliqué une validation croisée à $k = 5$ plis pour chaque modèle testé. Cette technique permet d'obtenir une évaluation plus robuste des performances des modèles, tout en limitant les risques de surapprentissage (*overfitting*) ou de sous-apprentissage (*underfitting*).

Application aux modèles testés

- Pour le **classificateur Arbre de Décision (AD)**, nous avons testé différentes profondeurs maximales afin d'optimiser le compromis entre biais et variance. Les résultats correspondants sont détaillés en **Annexe B**.
- Pour la **Régression Logistique (LG)** et le **classificateur SVM**, des expérimentations ont été menées en utilisant les hyperparamètres standards, avec normalisation préalable des données. Des ajustements manuels ont été effectués pour améliorer la convergence et la stabilité des modèles. Voir **Annexe A**.

Les performances de chaque modèle ont été évaluées à l'aide des métriques classiques telles que l'*accuracy* (exactitude), la *précision*, le *rappel*, et la *F-mesure*. Ces indicateurs sont présentés et comparés dans la section suivante.

3.3.2 Critères et mesures d'évaluation

Pour mesurer la performance des modèles, il existe des indices ou critères qui permettent de quantifier l'écart entre les prédictions du modèle et les valeurs réelles. Ces critères servent à évaluer la précision, la sensibilité, la spécificité ou d'autres aspects de la performance prédictive du modèle. Dans cette section, nous examinerons ces différents indices et critères afin d'évaluer la performance de nos modèles d'apprentissage automatique dans la prédiction du diabète.

- **Matrice de confusion :** Dans les problématiques de classification, la plupart des indices de performance sont calculés à partir d'une matrice de confusion. Cette matrice affiche le nombre de succès et d'échecs de prédiction pour chaque catégorie de la variable

(attribut) à prédire. La matrice de confusion est une table qui montre chaque classe dans les données d'évaluation, ainsi que le nombre ou le pourcentage de prédictions correctes et incorrectes.

Dans le cas d'une tâche de classification supervisée binaire, où la modalité de la variable à prédire correspond à la classe « positive » et l'autre à la classe « négative », on nomme les coefficients de la matrice de confusion de la manière suivante :

— VN : Nombre de vrais négatifs (True Negative TN)
— FN : Nombre de faux négatifs (False Negative FN)
— FP : Nombre de faux positifs (False Positive FP)
— VP : Nombre de vrais positifs (True Positive TP)

		Y prédit par le modèle	
		Y=1	Y=0
Y réel(Y')	Y'=1	Nombre de 1 prédits correctement Vrai Positifs (VP) True Positif (TP)	Nombre de 1 prédits en 0 Faux Négatif (FN) False Negatif (FN)
	Y'=0	Nombre de 0 prédits en 1 Faux Positifs (FP) False Positif (FP)	Nombre de 0 prédits correctement Vrai Négatif (VN) True Negatif (TN)

TABLE 3.1 – Matrice de Confusion

- **Accuracy (Exactitude, justesse)** : (Proportion de prédictions correctes).il s'agit d'une description d'erreurs systématiques, d'une mesure du biais statistique ; faible précision provoque une différence entre un résultat et une valeur "vraie".

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%$$

- **Précision (Precision) :** Proportion de solutions trouvées qui sont pertinentes. A quel point les prédictions positives sont précises.

$$\text{Précision} = \frac{TP}{TP+FP} \times 100\%$$

- **Rappel (Sensitivity, Recall) :** Proportion des solutions pertinentes qui sont trouvées.

Mesure la capacité du système à donner toutes les solutions pertinentes. Couverture des observations vraiment positives.

$$\text{R}appel(\text{sensitivity, recall}) = \frac{TP}{TP+FN} \times 100\%$$

- **F-mesure (F-score) :** La F-mesure correspond à un compromis de la précision et du rappel donnant la performance du modèle. Moyenne harmonique de la précision et du rappel. Mesure la capacité du modèle à donner toutes les solutions pertinentes et à refuser les autres.

$$F1 \text{ score} = \frac{2 \times (Rappel \times Prcision)}{Rappel + Prcision} \times 100\%$$

- **Matrice de confusion pour la prédiction :** Nous avons réduit l'analyse aux classes $\{1,2,3\}$.

		Classe prédite par le modèle		
		Y=1	Y=2	Y=3
Classe réelle	Y'=1	m_{11}	m_{12}	m_{13}
	Y'=2	m_{21}	m_{22}	m_{23}
	Y'=3	m_{31}	m_{32}	m_{33}

TABLE 3.2 – Matrice de confusion restreinte aux classes 1, 2 et 3.

m_{ij} : nombre de patients de **classe réelle** i prédits comme **classe** j par le modèle.
m_{cc} : nombre de patients de la **classe** c correctement prédits (vrais positifs de la classe c).

— **Accuracy** : Correspond à la proportion d'observations bien classées.

$$\text{Accuracy} = \frac{\sum_i m_{ii}}{\sum_{i,j} m_{ij}}$$

— **Taux d'erreur global** : Le taux d'erreur global, correspond à la proportion d'observations mal classées, qui dépend du ratio entre la trace de la matrice de confusion (c'est-à-dire la somme des coefficients diagonaux, donc le nombre de bonnes prédictions), et la somme de tous les coefficients (autrement dit le nombre total de

prédictions) :

$$E = 1 - \frac{\sum_i m_{ii}}{\sum_{i,j} m_{ij}}$$

— **Précision par rapport à une classe** : La précision d'un classifieur par rapport à une certaine classe (autrement dit, par rapport à une certaine modalité de la variable à prédire), se mesure comme la proportion d'individus, parmi tous ceux pour lesquels le classifieur a prédit cette classe, qui appartiennent réellement à celle-ci.

$$Prcision_{\text{classe } c}(P_c) = \frac{m_{cc}}{\sum_i m_{ci}}$$

— **Rappel par rapport à une classe** :Le rappel d'un classifieur par rapport à une certaine classe se mesure, quant à lui, comme la proportion d'individus, parmis tous ceux qui appartiennent réellement à cette classe, pour lesquels le classifieur a prédit cette classe c.

$$Rappel_{\text{classe } c}(R_c) = \frac{m_{cc}}{\sum_j m_{cj}}$$

— **F-mesure par rapport à une classe** : On peut résumer les mesures de précision de rappel par rapport à une classe c en un seul indicateur, en calculant la moyenne harmonique :

$$F_{\text{classe } c} = \frac{P_c \times R_c}{P_c + R_c}$$

3.4 Construction des Modèles avec l'ensemble global des données

Dans cette première phase de notre étude, nous avons construit des modèles de classification en utilisant **l'intégralité des données disponibles** pour chacune des deux bases : la base **Pima Indians Diabetes** (classification binaire) et la base **locale irakienne** (classification multiclasse).

L'objectif est de comparer les performances des modèles selon trois axes principaux :

- **La méthode de normalisation appliquée** :
  — **MinMaxScaler** (mise à l'échelle entre 0 et 1)

— **StandardScaler** (centrée-réduite, moyenne = 0, écart-type = 1)

- **L'algorithme de classification utilisé** :
  — **Régression Logistique (LR)**
  — **Support Vector Machine (SVM)**
  — **Arbre de Décision (AD)**
- **Le jeu de données utilisé** :
  — **Pima Indians Diabetes**
  — **Base locale irakienne**

Chaque combinaison a été évaluée à l'aide d'une **validation croisée à 5 plis**, ce qui permet d'estimer la capacité de généralisation des modèles tout en limitant les risques de surapprentissage.

3.4.1 Méthodologie

Pour chaque base de données :

1. Séparation des données en `X` (variables explicatives) et `y` (variable cible).
2. Application de chaque méthode de normalisation indépendamment.
3. Entraînement de chaque modèle sur le jeu normalisé.
4. Évaluation des modèles par validation croisée (5 folds).

Ainsi, un total de 2 normalisations $\times$ 3 modèles $\times$ 2 bases $=$ 12 expériences ont été réalisées.

3.4.2 Résultats

Accuracy moyenne par validation croisée (base Pima Indians Diabetes)

Normalisation	LR	SVM	AD
MinMaxScaler	0.79	0.81	0.77
StandardScaler	0.81	0.74	0.77

TABLE 3.3 – Performances (accuracy moyenne) sur la base Pima Indians

Accuracy moyenne par validation croisée (base locale irakienne)

Normalisation	LR	SVM	AD
MinMaxScaler	0.925	0.96	0.995
StandardScaler	0.93	0.89	0.995

TABLE 3.4 – Performances (accuracy moyenne) sur la base locale irakienne

Discussion des résultats La comparaison des performances sur les deux bases de données révèle des tendances significatives concernant l'impact des méthodes de normalisation sur les différents algorithmes.

Base Pima Indians Diabetes

- **Régression Logistique (LR)** : Performe légèrement mieux avec StandardScaler (**0.81** contre 0.79 avec MinMaxScaler)
- **SVM** : Montre une préférence marquée pour MinMaxScaler (**0.81** contre 0.74 avec StandardScaler)
- **Arbre de Décision (AD)** : Reste stable (**0.77**) quelle que soit la méthode de normalisation

Base locale irakienne

- **Performances globales** : Accuracy systématiquement plus élevée que sur Pima Indians
- **Arbre de Décision** :
  - Performance exceptionnelle avec MinMaxScaler (**0.995**)
  - Chute importante avec StandardScaler (0.68, soit une différence de **0.315**)
- **SVM** : Maintient une meilleure performance avec MinMaxScaler (**0.96** contre 0.89)
- **Régression Logistique** : Stabilité remarquable (0.925 contre 0.93, différence de **0.005** seulement)

Conclusions

- L'interaction entre méthode de normalisation et algorithme varie considérablement selon le jeu de données
- Pour la base irakienne, **MinMaxScaler** est crucial pour SVM et AD
- Pour Pima Indians, **StandardScaler** est préférable pour LR

- L'AD montre le comportement le plus variable selon la base de données

Cette analyse démontre l'importance cruciale d'adapter la stratégie de prétraitement aux caractéristiques spécifiques des données et au choix de l'algorithme.

3.5 Application des techniques de Sélection d'attributs (caractéristiques)

Dans cette seconde étape, nous avons appliqué des techniques de sélection de variables de type **filtre** et **wrapper** afin d'évaluer leur impact sur la performance des modèles. Cette démarche a été conduite sur les deux jeux de données (Pima Indians Diabetes et base locale irakienne), en tenant compte également des deux méthodes de normalisation (*MinMaxScaler* et *StandardScaler*).

3.5.1 Méthodologie

Pour chaque jeu de données et pour chaque méthode de normalisation, les étapes suivantes ont été réalisées :

1. **Application d'une méthode de filtrage pour la sélection de variables** : Les méthodes de filtrage ont été appliquées pour évaluer la pertinence des variables en fonction de leur relation avec la variable cible. Les techniques de sélection testées comprennent :
   - **Méthode de sélection basée sur le test du chi2 (SelectPercentile)** : Cette méthode utilise le test du chi2 pour évaluer l'indépendance entre chaque caractéristique et la variable cible, et sélectionne un pourcentage fixe des meilleures caractéristiques.
   - **Méthode de sélection d'attributs avec chi2 en mode kbest (Generic Univariate_Select)** : Cette méthode applique le test du chi2 pour sélectionner les k meilleures caractéristiques, où k est défini empiriquement. Cette approche est spécifiquement adaptée aux variables **catégoriques**.
   - **Méthode de sélection des meilleures caractéristiques avec ANOVA (SelectKBest)** : Utilisée pour les variables **numériques**, la méthode ANOVA analyse la variance entre les classes et sélectionne les caractéristiques ayant la plus grande différence entre les groupes.

2. **Application des techniques de sélection d'attributs de type wrapper** : Les méthodes wrapper évaluent les sous-ensembles de variables en entraînant un modèle de classification et en mesurant sa performance. Les techniques wrapper testées sont :
   - **Sélection par Élimination Séquentielle (SBS)** : Cette technique commence par un ensemble vide de caractéristiques et ajoute progressivement les caractéristiques qui améliorent le plus les performances du modèle. L'objectif est de trouver le sous-ensemble de variables qui maximise la performance du modèle.
   - **Sélection par Ajout Séquentiel (SFS)** : Contrairement à SBS, cette méthode commence avec toutes les caractéristiques et élimine successivement celles qui apportent le moins d'information pour la classification. Elle permet de réduire la dimensionnalité tout en conservant les caractéristiques les plus pertinentes.
   - **Sélection de caractéristiques avec la technique SelectFromModel** : Cette méthode repose sur un modèle d'apprentissage préalable (par exemple, un arbre de décision ou un modèle linéaire) pour déterminer l'importance des caractéristiques. Les variables les plus importantes sont ensuite sélectionnées en fonction des scores d'importance attribués par le modèle.
3. **Sélection des k meilleures variables** : Après l'application des méthodes de sélection, les k meilleures caractéristiques sont retenues, où k est défini empiriquement en fonction des résultats de chaque test. Ce processus permet de réduire la dimensionnalité des jeux de données tout en préservant les variables les plus influentes pour le modèle.
4. **Entraînement et validation des modèles** : Les modèles de classification suivants ont été utilisés pour tester l'impact de la sélection des variables :
   - **Régression Logistique (LR)**
   - **Machines à Vecteurs de Support (SVM)**
   - **Arbre de Décision (AD)**

   Les performances des modèles ont été évaluées sur les sous-ensembles de variables sélectionnées pour comprendre leur impact sur la précision du modèle.
5. **Évaluation des performances par validation croisée à 5 plis** : Afin d'obtenir des évaluations robustes et éviter le surapprentissage (overfitting), nous avons utilisé une validation croisée à 5 plis pour tester les modèles sur des sous-ensembles de données différents à chaque itération. Les métriques utilisées pour évaluer la performance incluent l'accuracy, la précision, le rappel et la F1-score.

3.5.2 Résultats de la Base de données Pima Indians Diabetes

3.5.2.1 Technique de sélection : Chi2 ((SelectPercentile))

a. Min-Max Normalization

% des attributs	Attributs sélectionnés	AD	LG	SVM
10%	Glucos	0.681	0.75	0.75
20%	Glucos,BMI	0.688	0.77	0.77
30%	Glucos,BMI,Age	0.70	0.76	0.77
40%	Glucos,BMI,Age	0.70	0.76	0.77
50%	Pregnancies,Glucos,BMI,Age	0.72	0.76	0.79
60%	Pregnancies,Glucos,BMI,Diabetes Pedigree Function ,Age	0.74	0.76	0.81
70%	Pregnancies,Glucos,BMI,Diabetes Pedigree Function ,Age	0.74	0.76	0.81
80%	Pregnancies,Glucos,Insulin,BMI,Diabetes Pedigree Function ,Age	0.75	0.77	0.82
90%	Pregnancies,Glucos,BloodPressure,Insulin,BMI,Diabetes PedigreeFunction ,Age	0.75	0.77	0.82
100%	Pregnancies,Glucos,BloodPressure,Skin Thickness,Insulin,BMI,Diabetes Pedigree Function ,Age	0.77	0.76	0.81

TABLE 3.5 – Résultat de la Technique de sélection : Chi2 avec Min-Max Normalization

b. Standard Normalization

% des attributs	Attributs sélectionnés	AD	LG	SVM
10%	Glucos	0.681	0.75	0.75
20%	Glucos,BMI	0.688	0.77	0.75
30%	Glucos,BMI,Age	0.70	0.78	0.72
40%	Glucos,BMI,Age	0.70	0.78	0.72
50%	Pregnancies,Glucos,BMI,Age	0.72	0.76	0.70
60%	Pregnancies,Glucos,BMI,Diabetes Pedigree Function ,Age	0.74	0.77	0.68
70%	Pregnancies,Glucos,BMI,DiabetesPedigreeFunction ,Age	0.74	0.77	0.68
80%	Pregnancies,Glucos,Insulin,BMI, DiabetesPedigree-Function ,Age	0.75	0.78	0.68
90%	Pregnancies,Glucos,BloodPressure,Insulin,BMI, Diabetes PedigreeFunction,Age	0.75	0.77	0.67
100%	Pregnancies,Glucos,BloodPressure, Skin Thick-ness,Insulin,BMI,DiabetesPedigreeFunction,Age	0.77	0.77	0.70

TABLE 3.6 – Résultat de la Technique de sélection : Chi2 avec Standard Normalization

Discussion des résultats L'utilisation de Chi2 (SelectPercentile) pour sélectionner un certain pourcentage des meilleurs attributs a été explorée dans notre étude. Nous avons varié le pourcentage de sélection de 10% à 100% et observé son impact sur la performance des modèles AD, LG et SVM avec deux méthodes de normalisation.

Dans le cas du SVM avec normalisation Min-Max, nous avons constaté que l'ajout d'attributs a généralement amélioré sa performance. Avec seulement l'attribut Glucose, nous avons obtenu une accuracy de 0.75%, qui est passée à 0.77% avec l'ajout de BMI. L'accuracy a continué d'augmenter jusqu'à 0.82% avec 80-90% des attributs, avant de diminuer légèrement à 0.81% avec tous les attributs.Mais avec normalisation Standard, nous observons une tendance inverse. L'accuracy est passée de 0.75% avec un seul attribut à 0.70% avec tous les attributs, montrant une dégradation progressive avec l'ajout d'attributs. Concernant l'arbre de décision, l'ajout d'attributs a généralement amélioré la performance du modèle. L'accuracy est passée de

0.681% avec Glucose seul à 0.77% avec tous les attributs, indépendamment de la méthode de normalisation. Pour la régression logistique, l'ajout d'attributs a eu un impact modéré. L'accuracy est restée stable entre 0.75% et 0.78%, avec une légère préférence pour la normalisation Standard. Ces résultats confirment l'importance des attributs Glucose, BMI et Age. Cette étude démontre également l'intérêt de la sélection d'attributs pour optimiser les performances des modèles prédictifs tout en réduisant leur complexité..

3.5.2.2 Technique de sélection : ANOVA (SelectKBest)

a. Min-Max Normalization

Nbr attributs	Attributs sélectionnés	AD	LG	SVM
1	Glucos	0.681	0.75	0.75
2	Glucos,BMI	0.688	0.77	0.77
3	Glucos,BMI,Age	0.70	0.76	0.77
4	Pregnancies,Glucos,BMI,Age	0.72	0.76	0.79
5	Pregnancies,Glucos,BMI, DiabetesPedigreeFunction,Age	0.74	0.76	0.81
6	Pregnancies,Glucos,Insulin,BMI, DiabetesPedigreeFunction ,Age	0.75	0.77	0.82
7	Pregnancies,Glucos,BloodPressure,Insulin,BMI, DiabetesPedigree Function ,Age	0.75	0.77	0.82
8	Pregnancies,Glucos,BloodPressure, SkinThickness,Insulin,BMI, DiabetesPedigreeFunction ,Age	0.77	0.76	0.81

TABLE 3.7 – Résultat de la Technique de sélection : ANOVA (SelectKBest) avec Min-Max Normalization

b. Standard Normalization

Nbr attributs	Attributs sélectionnés	AD	LG	SVM
1	Glucos	0.681	0.75	0.75
2	Glucos,BMI	0.688	0.77	0.75
3	Glucos,BMI,Age	0.70	0.78	0.72
4	Pregnancies,Glucos,BMI,Age	0.72	0.76	0.70
5	Pregnancies,Glucos,BMI, Diabetes Pedigree Function ,Age	0.74	0.77	0.68
6	Pregnancies,Glucos,Insulin,BMI, Diabetes Pedigree Function ,Age	0.75	0.78	0.68
7	Pregnancies,Glucos,BloodPressure,,Insulin, BMI,DiabetesPedigree Function,Age	0.75	0.77	0.67
8	Pregnancies,Glucos,Blood Pressure, SkinThickness,Insulin,BMI, DiabetesPedigreeFunction ,Age	0.77	0.77	0.70

TABLE 3.8 – Résultat de la Technique de sélection : ANOVA (SelectKBest) avec Standard Normalization

Discussion des résultats L'analyse des résultats de la sélection d'attributs par ANOVA (SelectKBest) révèle plusieurs tendances intéressantes concernant l'impact du nombre d'attributs et de la méthode de normalisation sur les performances des modèles.

Pour le SVM avec normalisation Min-Max, nous observons une amélioration constante des performances avec l'ajout d'attributs. L'accuracy passe de 0.75% avec seulement Glucose à 0.82% avec 6-7 attributs, avant de diminuer légèrement à 0.81% avec tous les attributs. Cette tendance suggère que le SVM bénéficie de l'information additionnelle jusqu'à un certain point, au-delà duquel des attributs moins pertinents peuvent introduire du bruit.En revanche, le SVM avec normalisation Standard montre une tendance opposée. L'accuracy diminue progressivement de 0.75% avec un seul attribut à 0.67% avec 7 attributs, avant de remonter légèrement à 0.70% avec tous les attributs. Cette différence marquée souligne l'importance cruciale du choix de la méthode de normalisation pour les SVM. L'arbre de décision (AD) présente une amélioration constante avec l'ajout d'attributs, indépendamment de la méthode de normalisation. L'accuracy passe de 0.681 avec Glucose seul à 0.77% avec tous les attributs, démontrant

la capacité de l'AD à exploiter efficacement l'information additionnelle. Pour la régression logistique (LG), l'ajout d'attributs a un impact modéré. L'accuracy reste relativement stable entre 0.75% et 0.78% pour les deux méthodes de normalisation, avec une légère préférence pour la normalisation Standard. Ces résultats confirment l'importance primordiale de l'attribut Glucose, suivi par BMI et Age. Ils mettent également en évidence l'efficacité de la sélection d'attributs pour optimiser les performances des modèles tout en réduisant leur complexité.

3.5.2.3 Technique de sélection : GenericUnivariate- Select

a. Min-Max Normalization

Nbr attributs	Attributs sélectionnés	AD	LG	SVM
1	Glucos	0.681	0.75	0.75
2	Glucos,BMI	0.688	0.77	0.77
3	Glucos,BMI,Age	0.70	0.76	0.77
4	Pregnancies,Glucos,BMI,Age	0.72	0.76	0.79
5	Pregnancies,Glucos,BMI, DiabetesPedigreeFunction,Age	0.74	0.76	0.81
6	Pregnancies,Glucos,Insulin,BMI, DiabetesPedigreeFunction,Age	0.75	0.77	0.82
7	Pregnancies,Glucos,BloodPressure,,Insulin ,BMI,DiabetesPedigreeFunction ,Age	0.75	0.77	0.82
8	Pregnancies,Glucos,BloodPressure,Skin Thickness,Insulin,BMI,DiabetesPedigreeFunction,Age	0.77	0.76	0.81

TABLE 3.9 – Résultat de la Technique de sélection : GenericUnivariate- Select avec Min-Max Normalization

b. Standard Normalization

Nbr attributs	Attributs sélectionnés	AD	LG	SVM
1	Glucos	0.681	0.75	0.75
2	Glucos,BMI	0.688	0.77	0.75
3	Glucos,BMI,Age	0.70	0.78	0.72
4	Pregnancies,Glucos,BMI,Age	0.72	0.76	0.70
5	Pregnancies,Glucos,BMI, DiabetesPedigreeFunction ,Age	0.74	0.77	0.68
6	Pregnancies,Glucos,Insulin,BMI, DiabetesPedigree-Function ,Age	0.75	0.78	0.68
7	Pregnancies,Glucos,Blood Pressure, Insulin,BMI,DiabetesPedigree Function,Age	0.75	0.77	0.67
8	Pregnancies,Glucos,Blood Pressure,Skin Thickness,Insulin,BMI,DiabetesPedigreeFunction,Age	0.77	0.77	0.70

TABLE 3.10 – Résultat de la Technique de sélection : GenericUnivariate- Select avec Standard Normalization

Discussion des résultats L'analyse des résultats de la sélection d'attributs par GenericUnivariateSelect montre des tendances similaires à celles observées avec ANOVA, ce qui est cohérent puisque ces deux méthodes reposent sur des tests statistiques pour évaluer la pertinence des attributs.

Pour le SVM avec normalisation Min-Max, nous observons une amélioration progressive des performances avec l'ajout d'attributs. L'accuracy augmente de 0.75% avec Glucose seul à 0.82% avec 6-7 attributs, avant de diminuer légèrement à 0.81% avec tous les attributs. Le SVM avec normalisation Standard présente à nouveau une tendance inverse préoccupante. L'accuracy diminue de 0.75%avec un seul attribut à 0.67% avec 7 attributs, avant de remonter légèrement à 0.70% avec tous les attributs. Cette observation renforce l'importance critique du choix de la méthode de normalisation pour les SVM dans ce contexte. L'arbre de décision montre une amélioration constante avec l'ajout d'attributs, passant de 0.681% avec Glucose seul à 0.77% avec tous les attributs, indépendamment de la méthode de normalisation. Cette robustesse face à la méthode de normalisation est une caractéristique avantageuse des arbres de décision. La

régression logistique maintient des performances stables entre 0.75% et 0.78% quelle que soit la normalisation utilisée, avec une légère préférence pour la normalisation Standard lorsque 3 ou 6 attributs sont sélectionnés (0.78%).

3.5.2.4 Technique de sélection : Sélection par Élimination Séquentielle (SBS)

La méthode de Sélection Par Elimination Séquentielle (SBS) a été appliquée au jeu de données pour sélectionner les attributs (caractéristiques) les plus pertinents en supprimant progressivement les attributs les moins importants. Le tableau 3.11 ci-dessous résume les résultats de l'application de la méthode SBS à notre jeu de données avec le modèle AD. Les itérations montrent la performance du modèle après la suppression séquentielle des caractéristiques.

Max Features	**Attributs sélectionnés**	**Accuracy du test**
7	[Pregnancies,Glucose,BloodPressure,SkinThickness, Insulin, BMI, Age]	0.753
6	[Pregnancies,Glucose,SkinThickness,Insulin, BMI, Age]	0.753
5	[Pregnancies,Glucose,SkinThickness,BMI, Age]	0.727
4	[Pregnancies,Glucose,SkinThickness,BMI, Age]	0.727
3	[Pregnancies,Glucose,SkinThickness,BMI,Age]	0.727
2	[Pregnancies,Glucose,BloodPressure, SkinThickness,BMI,Age]	0.746
1	[Pregnancies,Glucose,SkinThickness,BMI,Age]	0.727

TABLE 3.11 – Résultat de La Sélection Par Elimination Séquentiel(SBS)combinée avec AD

Discussion des résultats de l'application de SBS combinée avec AD En utilisant la méthode de Sélection Par Élimination Séquentielle (SBS), nous avons progressivement réduit le nombre d'attributs (caractéristiques) tout en évaluant l'impact sur la performance du modèle. Même en réduisant le nombre d'attributs jusqu'à un minimum de 1, le modèle maintient une certaine capacité de prédiction, bien que l'accuracy diminue par rapport à la configuration optimale. Le modèle atteint sa meilleure accuracy de 0.753% lorsqu'il est entraîné avec un sous-ensemble de 6 ou 7 attributs. Au fur et à mesure que le nombre d'attributs diminue à 5 et

moins, des variations mineures de l'accuracy sont observées, mais dans l'ensemble, le modèle conserve une certaine stabilité dans sa performance, maintenant une accuracy de 0.727% pour plusieurs configurations. La sélection d'un nombre optimal d'attributs est essentielle pour équilibrer la complexité du modèle avec sa capacité prédictive. Dans ce cas, un sous-ensemble de 6 attributs (Pregnancies, Glucose, SkinThickness, Insulin, BMI, Age) semble offrir le meilleur compromis entre performance et simplicité, avec une accuracy identique à celle obtenue avec 7 attributs. Ces résultats confirment l'efficacité de la méthode SBS pour simplifier les modèles tout en maintenant des performances acceptables. La suppression de BloodPressure n'a pas affecté la performance, tandis que la suppression d'Insulin a entraîné une légère baisse d'accuracy, suggérant l'importance relative de ces attributs dans la prédiction du diabète.

Le tableau 3.12 ci-dessous résume les résultats de l'application de la méthode SBS à notre jeu de données avec le modèle svm.

Max Features	**Attributs sélectionnés**	**Accuracy du test**
7	[Pregnancies,Glucose,SkinThickness,Insulin, BMI,DiabetesPedigreeFunction, Age]	0.792
6	[Pregnancies, Glucose,SkinThickness,Insulin, BMI,DiabetesPedigreeFunction, Age]	0.792
5	[Pregnancies,Glucose,SkinThickness,Insulin, BMI,DiabetesPedigreeFunction,Age]	0.792
4	[Pregnancies,Glucose,SkinThickness,Insulin, BMI,DiabetesPedigreeFunction,Age]	0.792
3	[Pregnancies,Glucose,SkinThickness,Insulin, BMI,DiabetesPedigreeFunction, Age]	0.792
2	[Pregnancies,Glucose,SkinThickness,Insulin, BMI,DiabetesPedigreeFunction, Age]	0.792
1	[Pregnancies,Glucose,SkinThickness,Insulin, BMI,DiabetesPedigreeFunction,Age]	0.792

TABLE 3.12 – Résultat de La Sélection Par Elimination Séquentiel(SBS)combinée avec SVM

Discussion des résultats de l'application de SBS combinée avec SVM En utilisant la méthode de Sélection Par Élimination Séquentielle (SBS), nous avons progressivement réduit le nombre d'attributs tout en évaluant l'impact sur la performance du modèle SVM. Les résultats présentés dans le tableau 3.12 révèlent un phénomène particulier : l'accuracy du test reste constante à 0.792% pour toutes les configurations, de 7 attributs jusqu'à 1 attribut. Cette stabilité remarquable de la performance suggère que le modèle SVM possède une grande robustesse face à la réduction du nombre d'attributs. Contrairement à d'autres algorithmes qui peuvent montrer une dégradation de performance lors de la suppression de caractéristiques importantes, le SVM maintient ici une performance élevée et constante. Cette constance dans les résultats pourrait également indiquer que le modèle SVM est capable d'extraire efficacement l'information pertinente même lorsque le nombre d'attributs est théoriquement réduit. La performance de 0.792207 est significativement supérieure à celle obtenue avec l'algorithme AD (0.753%), ce qui confirme l'efficacité du SVM pour ce jeu de données particulier. Le tableau 3.13 ci-dessous résume les résultats de l'application de la méthode SBS à notre jeu de données avec le modèle LG.

Max Features	**Attributs sélectionnés**	**Accuracy du test**
7	[Pregnancies,Glucose,SkinThickness,Insulin, BMI,DiabetesPedigreeFunction, Age]	0.798
6	[Pregnancies,Glucose,Insulin,BMI, DiabetesPedigreeFunction, Age]	0.798
5	[Pregnancies, Glucose,Insulin,BMI, DiabetesPedigreeFunction]	0.798
4	[Pregnancies,Glucose,BMI, DiabetesPedigreeFunction]	0.798
3	[Pregnancies,Glucose,BMI, DiabetesPedigreeFunction]	0.805
2	[Pregnancies,Glucose,BMI, DiabetesPedigreeFunction]	0.777
1	[Pregnancies,Glucose,BMI, DiabetesPedigreeFunction]	0.777

TABLE 3.13 – Résultat de La Sélection Par Elimination Séquentiel(SBS)combinée avec LG

Discussion des résultats de l'application de SBS combinée avec LG En utilisant la méthode de Sélection Par Élimination Séquentielle (SBS), nous avons progressivement réduit le nombre d'attributs tout en évaluant l'impact sur la performance du modèle de Régression Logistique (LG). Les résultats présentés dans le tableau 3.13 révèlent plusieurs observations intéressantes.

Initialement, avec 7 attributs, le modèle atteint une accuracy de 0.798%. En réduisant à 6 attributs par l'élimination de SkinThickness, la performance reste identique, suggérant que cet attribut n'apporte pas d'information significative au modèle LG. De même, la suppression d'Age à l'étape suivante (5 attributs) et d'Insulin ensuite (4 attributs) n'affecte pas la performance, qui se maintient à 0.798%. De façon remarquable, lorsque le nombre d'attributs est réduit à 3, la performance augmente légèrement pour atteindre 0.805%, ce qui représente la meilleure accuracy observée. Cette amélioration suggère que certains attributs pourraient introduire du bruit ou de la redondance, et que leur élimination permet au modèle de mieux généraliser. Ces observations indiquent que pour le modèle LG, un sous-ensemble optimal de 3 attributs (Pregnancies, Glucose, BMI, DiabetesPedigreeFunction) offre la meilleure performance, avec une accuracy de 0.805%. Cette configuration représente un excellent compromis entre simplicité du modèle et capacité prédictive, démontrant l'efficacité de la méthode SBS pour identifier les attributs les plus pertinents.

Discussion des résultats de la technique SBS L'application de la méthode de Sélection Par Élimination Séquentielle (SBS) aux trois modèles (AD, SVM et LG) nous permet de tirer plusieurs conclusions importantes sur l'efficacité de cette technique et sur l'importance relative des attributs pour la prédiction du diabète. Premièrement, nous observons des comportements distincts selon les algorithmes utilisés. Le modèle AD montre une sensibilité modérée à la réduction d'attributs, avec une performance optimale à 0.753% pour 6 attributs. Le SVM présente une robustesse remarquable avec une accuracy constante de 0.792% quelle que soit la configuration. La LG offre les résultats les plus intéressants, atteignant sa meilleure performance (0.805%) avec seulement 3 attributs. Cette comparaison met en évidence l'importance de l'interaction entre la méthode de sélection d'attributs et l'algorithme d'apprentissage. Certains modèles, comme le SVM, semblent moins sensibles à la sélection d'attributs, tandis que d'autres, comme la LG, peuvent bénéficier significativement d'une réduction judicieuse du nombre de

caractéristiques. En termes d'attributs, Glucose et BMI apparaissent comme particulièrement importants, étant conservés dans les configurations optimales de tous les modèles. Pregnancies figure également dans les meilleures configurations pour AD et LG. Ces observations concordent avec les connaissances médicales sur les facteurs de risque du diabète. La technique SBS s'est révélée particulièrement efficace pour la LG, permettant d'améliorer la performance tout en réduisant considérablement la complexité du modèle. Pour l'AD, elle a permis d'identifier un sous-ensemble d'attributs offrant un bon compromis performance-simplicité. Pour le SVM, les résultats suggèrent une grande robustesse du modèle. En conclusion, la méthode SBS constitue un outil précieux pour optimiser les modèles prédictifs en identifiant les sous-ensembles d'attributs les plus pertinents, contribuant ainsi à améliorer à la fois la performance et l'interprétabilité des

3.5.2.5 Technique de sélection : Sélection par Ajout Séquentiel (SFS)

3.5.2.6 La Sélection Par Ajout Séquentiel (SFS)

La méthode (SFS) a été appliquée au jeu de données pour sélectionner les attributs en ajoutant progressivement les attributs les plus informatifs. Le tableau 3.14 ci-dessous résume les résultats de l'application de la méthode SFS à notre jeu de données avec le modèle SVM

Max Features	Attributs sélectionnés	Accuracy du test
1	[Glucose]	0.7661
2	[Glucose,BMI]	0.7857
3	[Glucose,BMI,Insulin]	0.7597
4	[Glucose,BMI,,Insulin,AGE]	0.7662
5	[Glucose,BMI,Insulin,AGE]	0.76622
6	[Glucose,BMI,Insulin,AGE]	0.7662
7	[Glucose,BMI,Insulin,AGE]	0.7662

TABLE 3.14 – Résultat de La Sélection Par ajout Séquentiel(SFS)combinée avec SVM

Discussion des résultats de l'application de SFS combinée avec SVM : En utilisant la méthode de Sélection Par Ajout Séquentiel (SFS), nous avons progressivement augmenté le nombre d'attributs tout en évaluant l'impact sur la performance du modèle SVM. Les résultats

présentés dans le tableau 3.14 révèlent plusieurs observations intéressantes.

Avec un seul attribut, Glucose, le modèle atteint déjà une accuracy remarquable de 0.7661%, soulignant l'importance primordiale de cette caractéristique dans la prédiction du diabète. L'ajout de BMI comme deuxième attribut améliore significativement la performance, portant l'accuracy à 0.7857%, ce qui représente le pic de performance observé dans cette analyse. De façon surprenante, l'ajout d'un troisième attribut, Insulin, entraîne une diminution de la performance à 0.7597%. Cette baisse suggère que l'information apportée par Insulin pourrait introduire du bruit ou interférer avec les patterns déjà identifiés par le modèle à partir de Glucose et BMI. L'ajout d'Age comme quatrième attribut permet de récupérer partiellement la performance, avec une accuracy de 0.7662%. Cependant, les ajouts subséquents d'attributs (jusqu'à 7) ne modifient plus la performance, qui reste stable à 0.7662%. suggérant une certaine robustesse du modèle face aux variations du nombre d'attributs au-delà d'un certain seuil. En conclusion, cette analyse SFS confirme que la combinaison optimale pour le modèle SVM est constituée de seulement deux attributs : Glucose et BMI, permettant d'atteindre une accuracy de 0.7857%. Cette configuration représente un excellent compromis entre simplicité du modèle et capacité prédictive. Le tableau 3.15 ci-dessous résume les résultats de l'application de la méthode SFS à notre jeu de données avec le modèle LG.

Max Features	**Attributs sélectionnés**	**Accuracy du test**
1	[Glucose]	0.7597
2	[Glucose,BMI]	0.7727
3	[Glucose,BMI,DiabetesPedigreeFunction]	0.7662
4	[Glucose,BMI,DiabetesPedigreeFunction, Pregnancies]	0.7987
5	[Glucose,BMI,DiabetesPedigreeFunction, Pregnancies,BloodPressure]	0.7012
6	[Glucose,BMI,DiabetesPedigreeFunction, Pregnancies,BloodPressure]	0.7012
7	[Glucose,BMI,DiabetesPedigreeFunction, Pregnancies,BloodPressure]	0.7012

TABLE 3.15 – Résultat de La Sélection Par ajout Séquentiel(SFS)combinée avec LG

Discussion des résultats de l'application de SFS combinée avec LG : En utilisant la méthode de Sélection Par Ajout Séquentiel (SFS), nous avons progressivement augmenté le nombre d'attributs tout en évaluant l'impact sur la performance du modèle de Régression Logistique (LG). Avec un seul attribut, Glucose, le modèle atteint une accuracy de 0.7597%, confirmant l'importance fondamentale de cette caractéristique dans la prédiction du diabète. L'ajout de BMI comme deuxième attribut améliore la performance à 0.7727%, renforçant l'idée que ces deux attributs sont particulièrement informatifs. L'ajout de DiabetesPedigreeFunction comme troisième attribut entraîne une légère diminution de la performance à 0.7662%, suggérant une possible interaction complexe entre ces trois variables. Cependant, l'ajout de Pregnancies comme quatrième attribut conduit à une amélioration significative, portant l'accuracy à 0.7987%, ce qui représente la meilleure performance observée dans cette analyse. De façon surprenante, l'ajout de BloodPressure comme cinquième attribut provoque une chute importante de la performance à 0.7012%. Cette diminution drastique pourrait indiquer que BloodPressure introduit du bruit ou des patterns contradictoires qui perturbent le modèle de régression logistique. Les ajouts subséquents d'attributs (jusqu'à 7) ne modifient plus la performance, qui reste stable à 0.7012%. Il est également notable que le tableau indique que les attributs sélectionnés restent les mêmes pour les configurations de 5 à 7 attributs, ce qui pourrait suggérer une limitation dans l'implémentation de la méthode SFS ou que les attributs restants n'apportent pas d'amélioration selon le critère de sélection utilisé. En conclusion, cette analyse SFS démontre que la combinaison optimale pour le modèle LG est constituée de quatre attributs : Glucose, BMI, DiabetesPedigreeFunction et Pregnancies, permettant d'atteindre une accuracy remarquable de 0.7987%. Cette configuration représente un excellent compromis entre complexité du modèle et capacité prédictive, surpassant même les performances obtenues avec le modèle SVM (0.7857% avec deux attributs).

Le tableau 3.16 ci-dessous résume les résultats de l'application de la méthode SFS à notre jeu de données avec le modèle AD.

Max Features	**Attributs sélectionnés**	**Accuracy du test**
1	[Glucose]	0.7142
2	[Glucose,BMI]	0.7012
3	[Glucose,BMI]	0.7012
4	[Glucose,BMI]	0.7012
5	[Glucose,BMI]	0.7012
6	[Glucose,BMI]	0.7012
7	[Glucose,BMI]	0.7012

TABLE 3.16 – Résultat de La Sélection Par ajout Séquentiel(SFS)combinée avec AD

Discussion des résultats de l'application de SFS combinée avec AD : En analysant les résultats de l'application de la méthode SFS combinée avec le modèle d'Arbre de Décision (AD), nous observons un comportement spécifique qui mérite d'être détaillé. Lorsque le paramètre "Max Features" est fixé à 1, seul l'attribut Glucose est sélectionné, permettant d'atteindre une accuracy de test de 0.7142%. Cette performance relativement élevée avec un seul attribut souligne l'importance fondamentale de la glycémie dans la prédiction du diabète. En augmentant "Max Features" à 2, l'attribut BMI est ajouté à la sélection, mais contrairement aux attentes, l'accuracy diminue légèrement à 0.7012%. Cette baisse suggère que l'ajout de cet attribut pourrait introduire une certaine complexité que le modèle AD ne parvient pas à exploiter efficacement dans ce contexte spécifique. De façon plus surprenante, l'augmentation du nombre maximal d'attributs de 3 à 7 ne modifie ni les attributs sélectionnés (qui restent limités à Glucose et BMI), ni l'accuracy du test qui demeure stable à 0.7012%. Cette stagnation indique que le modèle AD, dans cette configuration particulière, ne bénéficie pas de l'ajout d'attributs supplémentaires. En examinant ces résultats, il apparaît que le modèle AD obtient ses meilleures performances lorsqu'un seul attribut (Glucose) est sélectionné. L'ajout d'attributs supplémentaires n'améliore pas l'accuracy du modèle et entraîne même une légère baisse de performance, ce qui pourrait être interprété comme un signe de surapprentissage .

Discussion des résultats de l'application de SFS : L'application de la méthode de Sélection par Ajout Séquentiel (SFS) à nos trois modèles (LG, SVM et AD) révèle des insights importants sur la prédiction du diabète. Tous les modèles identifient Glucose comme l'attribut le plus déterminant, confirmant son rôle crucial dans le diagnostic. BMI apparaît systématiquement comme le second attribut le plus informatif. Le modèle LG démontre la meilleure capacité d'exploitation des attributs multiples, atteignant une accuracy maximale de 0.7987% avec quatre attributs (Glucose, BMI, DiabetesPedigreeFunction et Pregnancies). Le SVM atteint son pic de performance (0.7857%) avec seulement deux attributs (Glucose et BMI), tandis que l'AD performe mieux avec un seul attribut (Glucose, 0.7142%) qu'avec des combinaisons plus complexes. Ces résultats soulignent l'importance d'adapter la sélection d'attributs au modèle utilisé. La LG bénéficie d'une combinaison plus riche d'attributs, alors que le SVM et l'AD favorisent la parcimonie. L'ajout d'attributs au-delà d'un certain seuil entraîne généralement une baisse de performance, illustrant le compromis classique entre complexité du modèle et capacité de généralisation. La méthode SFS s'avère particulièrement efficace pour identifier les configurations optimales spécifiques à chaque modèle, permettant d'améliorer simultanément la performance prédictive et l'interprétabilité des résultats.

3.5.2.7 Technique de sélection : Sélection de caractéristiques avec la technique SelectFromModel

Le tableau 3.17ci-dessous résume les résultats de l'application de la méthode SelectFromModel à notre jeu de données avec le modèle AD

Nbr d'attributs	Attributs sélectionnés (AD)	Accuracy
1	[Glucose]	0.7142
2	[Glucose, BMI]	0.7012
3	[Glucose, BMI, Age]	0.7337
4	[Glucose, BMI, Age]	0.7337
5	[Glucose, BMI, Age]	0.7337
6	[Glucose, BMI, Age]	0.7337
7	[Glucose, BMI, Age]	0.7337
8	[Glucose, BMI, Age]	0.7337

TABLE 3.17 – Résultat de la sélection de caractéristiques avec la technique SelectFromModel combinée avec AD

Discussion des résultats L'application de SelectFromModel avec l'Arbre de Décision révèle des insights significatifs pour la prédiction du diabète. Avec un seul attribut, Glucose est identifié comme la caractéristique la plus déterminante (accuracy : 0.7142%). L'ajout de BMI entraîne une légère baisse de performance (0.7012%), suggérant une interaction complexe entre ces variables que l'AD peine à exploiter efficacement. La configuration optimale est atteinte avec trois attributs (Glucose, BMI, Age), permettant une accuracy de 0.7337%. Remarquablement, l'augmentation du nombre maximal d'attributs de 4 à 8 ne modifie pas cette sélection et maintient l'accuracy stable, indiquant que les autres variables ne sont pas jugées suffisamment significatives par la méthode. Ces résultats démontrent l'efficacité de SelectFromModel pour identifier un sous-ensemble optimal et parcimonieux de caractéristiques. Cette analyse confirme également la cohérence avec d'autres méthodes de sélection qui ont identifié Glucose et BMI comme attributs cruciaux, tout en soulignant l'importance spécifique de l'âge dans le contexte de l'Arbre de Décision pour la prédiction du diabète.

Le tableau 3.18 ci-dessous résume les résultats de l'application de la méthode SelectFromModel à notre jeu de données avec le modèle LG avec la normalisation MinMax

a. Min-Max Normalization .

Nbr d'attributs)	Attributs sélectionnés (LG)	Accuracy
1	[Glucose]	0.7597
2	[Glucose,BMI]	0.7727
3	[Glucose,BMI,DiabetesPedigreeFunction]	0.7662
4	[Glucose,BMI,DiabetesPedigreeFunction]	0.7662
5	[Glucose,BMI,DiabetesPedigreeFunction]	0.7662
6	[Glucose,BMI,DiabetesPedigreeFunction]	0.7662
7	[Glucose,BMI,DiabetesPedigreeFunction]	0.7662
8	[Glucose,BMI,DiabetesPedigreeFunction]	0.7662

TABLE 3.18 – Résultat de La Sélection de caractéristiques avec la technique SelectFromModel combinée avec LG

Le tableau 3.19 ci-dessous résume les résultats de l'application de la méthode SelectFrom-Model à notre jeu de données avec le modèle LG avec la normalisation Standard

b. Standard Normalization .

Nbr d'attributs)	Attributs sélectionnés (LG)	Accuracy
1	[Glucose]	0.7597
2	[Glucose,BMI]	0.7727
3	[Pregnancies,Glucose,BMI]	0.7922
4	[Pregnancies,Glucose,BMI]	0.7922
5	[Pregnancies,Glucose,BMI]	0.7922
6	[Pregnancies,Glucose,BMI]	0.7922
7	[Pregnancies,Glucose,BMI]	0.7922
8	[Pregnancies,Glucose,BMI]	0.7922

TABLE 3.19 – Résultat de La Sélection de caractéristiques avec la technique SelectFromModel combinée avec LG

Discussion des résultats Cette différence souligne l'influence considérable de la méthode de normalisation sur l'importance relative attribuée aux attributs par le modèle LG. La normalisation Standard semble mieux préserver ou mettre en évidence l'information prédictive

contenue dans l'attribut Pregnancies, permettant au modèle d'atteindre une performance supérieure. Dans les deux cas, l'augmentation du nombre maximal d'attributs au-delà de trois n'entraîne aucun changement dans la sélection ni dans la performance, indiquant que les attributs supplémentaires ne sont pas considérés comme suffisamment informatifs par la méthode SelectFromModel. Ces résultats démontrent que la normalisation Standard est plus adaptée à notre jeu de données pour la prédiction du diabète avec la Régression Logistique, permettant d'atteindre une accuracy de 0.7922% avec seulement trois attributs (Pregnancies, Glucose et BMI). Cette configuration optimale offre un excellent compromis entre parcimonie du modèle et performance prédictive, tout en soulignant l'importance de choisir une méthode de normalisation appropriée lors de l'application de techniques de sélection d'attributs.Cette différence souligne l'influence considérable de la méthode de normalisation sur l'importance relative attribuée aux attributs par le modèle LG. La normalisation Standard semble mieux préserver ou mettre en évidence l'information prédictive contenue dans l'attribut Pregnancies, permettant au modèle d'atteindre une performance supérieure. Ces résultats démontrent que la normalisation Standard est plus adaptée à notre jeu de données pour la prédiction du diabète avec la Régression Logistique, permettant d'atteindre une accuracy de 0.7922% avec seulement trois attributs (Pregnancies, Glucose et BMI). Cette configuration optimale offre un excellent compromis entre parcimonie du modèle et performance prédictive, tout en soulignant l'importance de choisir une méthode de normalisation appropriée lors de l'application de techniques de sélection d'attributs.

Le tableau 3.20 ci-dessous résume les résultats de l'application de la méthode SelectFromModel à notre jeu de données avec le modèle SVM avec la normalisation MinMax.

a. Min-Max Normalization .

Nbr d'attributs)	Attributs sélectionnés (SVM)	Accuracy
1	[Glucose]	0.7662
2	[Glucose,BMI]	0.7857
3	[Glucose,BMI]	0.7857
4	[Glucose,BMI]	0.7857
5	[Glucose,BMI]	0.7857
6	[Glucose,BMI]	0.7857
7	[Glucose,BMI]	0.7857
8	[Glucose,BMI]	0.7857

TABLE 3.20 – Résultat de La Sélection de caractéristiques avec la technique SelectFromModel combinée avec SVM

Le tableau 3.21 ci-dessous résume les résultats de l'application de la méthode SelectFrom-Model à notre jeu de données avec le modèle SVM avec la normalisation Standard

b. Standard Normalization .

Nbr d'attributs)	Attributs sélectionnés (SVM)	Accuracy
1	[Glucose]	0.7662
2	[Glucose,BMI]	0.7857
3	[Pregnancies,Glucose,BMI]	0.7727
4	[Pregnancies,Glucose,BMI]	0.7727
5	[Pregnancies,Glucose,BMI]	0.7727
6	[Pregnancies,Glucose,BMI]	0.7727
7	[Pregnancies,Glucose,BMI]	0.7727
8	[Pregnancies,Glucose,BMI]	0.7727

TABLE 3.21 – Résultat de La Sélection de caractéristiques avec la technique SelectFromModel combinée avec SVM :

Discussion des résultats L'application de SelectFromModel avec SVM montre des résultats révélateurs concernant l'impact des techniques de normalisation. Pour les deux méthodes, Glucose est identifié comme l'attribut principal (accuracy : 0.7662%), et l'ajout de BMI améliore

la performance à 0.7857% dans les deux cas. La différence clé apparaît avec trois attributs : la normalisation Min-Max ne sélectionne aucun attribut supplémentaire, maintenant l'accuracy à 0.7857%, tandis que la normalisation Standard ajoute Pregnancies mais avec une légère baisse de performance à 0.7727%. Ce résultat contraste avec ceux de la Régression Logistique, où l'ajout de Pregnancies avec normalisation Standard améliorait la performance. Dans les deux cas, aucun changement n'est observé au-delà de ces sélections initiales, indiquant que les autres attributs ne sont pas jugés suffisamment informatifs. Ces résultats suggèrent que la normalisation Min-Max est plus adaptée au SVM pour notre jeu de données, permettant d'atteindre une accuracy optimale de 0.7857% avec seulement deux attributs (Glucose et BMI).

3.5.3 Optimisation des Combinaisons de Caractéristiques : La base de données Pima Indians Diabetes

3.5.3.1 Évaluation des performances pour différentes combinaisons de caractéristiques

Après cette étape, nous avons examiné toutes les combinaisons de caractéristiques qui ont donné une bonne accuracy. Notre objectif était d'identifier les ensembles de caractéristiques les plus performants pour chaque modèle.

Modèle	Accuracy	Attributs sélectionnés
AD	0.7077	[0,1,5]
LG	0.7922	[0,1,5]
SVM	0.7987	[0,1,5]
AD	0.7142	[1,4,5]
LG	0.7727	[1,4,5]
SVM	0.7857	[1,4,5]
AD	0.7337	[1,5,7]
LG	0.7727	[1,5,7]
SVM	0.7922	[1,5,7]
AD	0.7012	[1,6,7]
LG	0.7402	[1,6,7]
SVM	0.7922	[1,6,7]
AD	0.7272	[0,1,4,5]
LG	0.7857	[0,1,4,5]
SVM	0.7987	[0,1,4,5]
AD	0.7077	[0,1,4,6]
LG	0.7922	[0,1,4,6]
SVM	0.7922	[0,1,4,6]
AD	0.7467	[0,1,5,6]
LG	0.7987	[0,1,5,6]
SVM	0.8051	[0,1,5,6]
AD	0.7272	[1,4,5,6]
LG	0.7662	[1,4,5,6]
SVM	0.8051	[1,4,5,6]
AD	0.7662	[1,5,6,7]
LG	0.7662	[1,5,6,7]
SVM	0.8116	[1,5,6,7]
AD	0.7597	[0,1,4,5,6]
LG	0.8051	[0,1,4,5,6]
SVM	0.8376	[0,1,4,5,6]
AD	0.7597	[0,1,5,6,7]
LG	0.7662	[0,1,5,6,7]
SVM	0.8246	[0,1,5,6,7]

TABLE 3.22 – Évaluation des performances pour différentes combinaisons de caractéristiques

Nous avons évalué plusieurs combinaisons de caractéristiques pour améliorer les performances des modèles de classification. Parmi celles testées, les meilleures performances ont été obtenues avec les colonnes : **0 : "Pregnancies" , 1 : "Glucose", 4 : "Insulin" , 5 : "BMI" et 6 : "DiabetesPedigreeFunction"**.

3.5.3.2 Résultats Finaux d'Amélioration des Performances

Après application de cette combinaison optimale, nous avons reconstruit et évalué les modèles avec validation croisée, montrant une nette amélioration des performances :

a. Classifieur : La régression logistique (LG)

- **Accuracy initiale :** 79%
- **Accuracy améliorée :** 80%

Cette amélioration à 80% suggère que la sélection de caractéristiques a permis au LG de mieux capturer les différences entre les classes.

b. Classifieur :la machine à vecteurs de support (SVM)

- **Accuracy initiale :** 74%
- **Accuracy améliorée :** 83%

Le SVM a atteint 83%, illustrant une meilleure partition de l'espace des données avec des caractéristiques pertinentes..

c. Classifieur : Arbre de décision (AD)

- **Accuracy initiale :** 77%
- **Accuracy améliorée :** 77%
- L'AD ne consiste aucune amélioration

Ces résultats soulignent l'importance cruciale de la sélection des caractéristiques dans la modélisation, réduisant le bruit et améliorant la généralisation des modèles.

3.5.4 Résultats de la Base de données Irakienne Diabetes Dataset

3.5.4.1 Technique de sélection : Chi2 ((SelectPercentile))

Basée sur l'utilisation des scores Chi2 (f_classif) pour sélectionner un certain pourcentage des meilleures caractéristiques. Dans notre étude, nous avons varié le pourcentage de sélection

de 10% à 100% par paliers de 10%.

a. Min-Max Normalization

% des attributs	Attributs sélectionnés	AD	LG	SVM
10%	BMI	0.92	0.865	0.895
20%	HbA1c,BMI	0.965	0.91	0.935
30%	AGE,HbA1c,BMI	0.99	0.91	0.935
40%	AGE,HbA1c,TG,BMI	0.975	0.915	0.92
50%	AGE,HbA1c,Chol,TG,BMI	0.99	92.5	0.94
60%	Gender,AGE,HbA1c,Chol,TG,BMI	0.99	0.92	0.945
70%	Gender,AGE,HbA1c,Chol,TG,VLDL,BMI	0.995	0.92	0.945
80%	Gender,AGE,Urea,HbA1c,Chol,TG,VLDL,BMI	0.995	0.92	0.945
90%	Gender,AGE,Urea,Cr,HbA1c,Chol,TG,VLDL,BMI	0.995	0.92	0.94
100%	Gender, AGE, Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, BMI	0.995	0.92	0.95

TABLE 3.23 – Résultat de La technique de sélection : Chi2 avec Min-Max Normalization

b. Standard Normalization

% des attributs	Attributs sélectionnés	AD	LG	SVM
10%	BMI	0.92	0.89	0.915
20%	HbA1c,BMI	0.97	0.93	0.975
30%	AGE,HbA1c,BMI	0.99	0.93	0.97
40%	AGE,HbA1c,TG,BMI	0.975	0.93	0.97
50%	AGE,HbA1c,Chol,TG,BMI	0.99	94	0.97
60%	Gender,AGE,HbA1c,Chol,TG,BMI	0.99	0.935	0.965
70%	Gender,AGE,HbA1c,Chol,TG,VLDL,BMI	0.995	0.94	0.96
80%	Gender,AGE,Urea,HbA1c,Chol,TG,VLDL,BMI	0.995	0.935	0.945
90%	Gender,AGE,Urea,Cr,HbA1c,Chol,TG,VLDL,BMI	0.995	0.935	0.935
100%	Gender, AGE, Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, BMI	0.995	0.935	0.91

TABLE 3.24 – Résultat de La technique de sélection : Chi2 avec Standard Normalization

Discussion des résultats L'analyse des résultats obtenus avec la technique de sélection Chi2 (SelectPercentile) révèle des tendances intéressantes sur les performances des modèles.

Pour l'Arbre de Décision (AD), nous observons une amélioration rapide des performances dès l'utilisation de 20% des attributs (HbA1c, BMI), atteignant 0.965 avec Min-Max. L'ajout de l'attribut Age porte cette performance à 0.99. Au-delà, les performances se stabilisent autour de 0.99-0.995, suggérant que l'AD capture efficacement les relations avec peu d'attributs. La Régression Logistique (LG) montre une progression plus graduelle jusqu'à un plateau de 0.92-0.94 avec 50-60% des attributs. La normalisation Standard offre généralement de meilleurs résultats pour ce modèle. Quant au SVM, ses performances suivent une courbe similaire à celle de la LG, mais avec une sensibilité plus marquée à la méthode de normalisation. Avec la normalisation Standard, le SVM atteint son pic de performance (0.975) avec seulement 20% des attributs, puis décline légèrement avec l'ajout d'attributs supplémentaires. Ce comportement contraste avec la normalisation Min-Max, où les performances augmentent progressivement jusqu'à 0.95 avec 100% des attributs.

Ces résultats confirment l'importance de la sélection d'attributs. Le choix de 5 attributs

(Age, HbA1c, Chol, TG, BMI) représente un excellent compromis entre complexité et performance. La supériorité de l'AD (0.995 d'accuracy) suggère que les relations entre les attributs et le diabète sont non-linéaires et bien capturées par ce modèle.

3.5.4.2 Technique de sélection : ANOVA (SelectKBest)

Sélection des k meilleures caractéristiques basées sur des tests univariés (ANOVA ou f_classif). Nous avons exploré différentes valeurs de k pour identifier le nombre optimal de caractéristiques à retenir

a. Min-Max Normalization

Nbr des atts	Attributs sélectionnés	AD	LG	SVM
1	BMI	0.92	0.865	0.895
2	HbA1c,BMI	0.965	0.91	0.935
3	AGE,HbA1c,BMI	0.99	0.91	0.935
4	AGE,HbA1c,TG,BMI	0.975	0.915	0.92
5	AGE,HbA1c,Chol,TG,BMI	0.99	92.5	0.94
6	Gender,AGE,HbA1c,Chol,TG,BMI	0.99	0.92	0.945
7	Gender,AGE,HbA1c,Chol,TG,VLDL,BMI	0.995	0.92	0.945
8	Gender,AGE,Urea,HbA1c,Chol,TG,VLDL,BMI	0.995	0.92	0.945
9	Gender,AGE,Urea,Cr,HbA1c,Chol,TG,VLDL,BMI	0.995	0.92	0.94
10	Gender, AGE, Urea, Cr, HbA1c, Chol, TG, HDL, VLDL, BMI	0.99	0.92	0.945
11	Gender, AGE, Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, BMI	0.995	0.92	0.95

TABLE 3.25 – Résultat de La technique de sélection : ANOVA (SelectKBest) avec Min-Max Normalization

b. Standard Normalization

Nbr des atts	Attributs sélectionnés	AD	LG	SVM
1	BMI	0.92	0.89	0.915
2	HbA1c,BMI	0.97	0.93	0.975
3	AGE,HbA1c,BMI	0.99	0.93	0.97
4	AGE,HbA1c,TG,BMI	0.975	0.93	0.97
5	AGE,HbA1c,Chol,TG,BMI	0.99	94	0.97
6	Gender,AGE,HbA1c,Chol,TG,BMI	0.99	0.935	0.965
7	Gender,AGE,HbA1c,Chol,TG,VLDL,BMI	0.995	0.94	0.96
8	Gender,AGE,Urea,HbA1c,Chol,TG,VLDL,BMI	0.995	0.935	0.945
9	Gender,AGE,Urea,Cr,HbA1c,Chol,TG,VLDL,BMI	0.995	0.935	0.935
10	Gender, AGE, Urea, Cr, HbA1c, Chol, TG, HDL, VLDL, BMI	0.99	0.93	0.92
11	Gender, AGE, Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, BMI	0.995	0.935	0.91

TABLE 3.26 – Résultat de La technique de sélection : ANOVA (SelectKBest) avec Standard Normalization

Discussion des résultats L'utilisation de SelectKBest pour la sélection des k meilleures caractéristiques basées sur des tests univariés tels que Chi2 ou f_classif a été explorée dans notre étude.

Avec un seul attribut (BMI), les trois modèles atteignent déjà des performances respectables (0.92% pour AD, 0.89-0.915% pour SVM), soulignant l'importance cruciale de cet indicateur. L'ajout d'HbA1c comme second attribut améliore considérablement les performances, particulièrement pour le SVM qui atteint 0.975% avec la normalisation Standard. L'Arbre de Décision montre une excellente capacité d'adaptation, atteignant 0.99% avec seulement trois attributs (Age, HbA1c, BMI) et maintenant cette performance élevée indépendamment du nombre d'attributs ajoutés par la suite. Cette stabilité suggère que l'AD capture efficacement les relations complexes dans les données avec un minimum d'information. La Régression Logistique bénéficie davantage de la normalisation Standard, avec une performance optimale de 0.94% pour 5 attributs, contre 0.925% avec Min-Max. Cette différence souligne l'importance de la préparation des données pour ce modèle. Le SVM présente un comportement intéressant : ses performances culminent avec peu d'attributs (2-5) en normalisation Standard, puis diminuent progressivement avec l'ajout d'attributs supplémentaires. Ce phénomène illustre la "malédiction de la dimensionnalité" où l'ajout de variables moins pertinentes peut dégrader les performances du modèle.

Ces observations confirment qu'un sous-ensemble optimal de 5 attributs (Age, HbA1c, Chol, TG, BMI) permet d'atteindre des performances quasi-maximales tout en simplifiant considérablement le modèle, facilitant ainsi son implémentation clinique et son interprétation par les professionnels de santé.

3.5.4.3 Technique de sélection : GenericUnivariate-Select

Sélection univariée basée sur les scores Chi2, avec différentes stratégies de sélection comme k_best. Nous avons testé la sélection en variant le nombre de caractéristiques de 1 à 11.

a. Min-Max Normalization

Nbr des atts	Attributs sélectionnés	AD	LG	SVM
1	BMI	0.92	0.865	0.895
2	HbA1c,BMI	0.965	0.91	0.935
3	AGE,HbA1c,BMI	0.99	0.91	0.935
4	AGE,HbA1c,TG,BMI	0.975	0.915	0.92
5	AGE,HbA1c,Chol,TG,BMI	0.99	0.925	0.94
6	Gender,AGE,HbA1c,Chol,TG,BMI	0.99	0.92	0.945
7	Gender,AGE,HbA1c,Chol,TG,VLDL,BMI	0.995	0.92	0.945
8	Gender,AGE,Urea,HbA1c,Chol,TG,VLDL,BMI	0.995	0.92	0.945
9	Gender,AGE,Urea,Cr,HbA1c,Chol,TG,VLDL,BMI	0.995	0.92	0.94
10	Gender,AGE,Urea,Cr,HbA1c,Chol,TG,HDL,VLDL,BMI	0.99	0.92	0.945
11	Gender,AGE,Urea,Cr,HbA1c,Chol,TG,HDL,LDL, VLDL,BMI]	0.995	0.92	0.95

TABLE 3.27 – Résultat de La technique de sélection : GenericUnivariate-Select avec Min-Max Normalization

b.Standard Normalization

Nbr des atts	Attributs sélectionnés	AD	LG	SVM
1	BMI	0.92	0.89	0.915
2	HbA1c,BMI	0.97	0.93	0.975
3	AGE,HbA1c,BMI	0.99	0.93	0.97
4	AGE,HbA1c,TG,BMI	0.975	0.93	0.97
5	AGE,HbA1c,Chol,TG,BMI	0.99	0.94	0.97
6	Gender,AGE,HbA1c,Chol,TG,BMI	0.99	0.935	0.965
7	Gender,AGE,HbA1c,Chol,TG,VLDL,BMI	0.995	0.94	0.96
8	Gender,AGE,Urea,HbA1c,Chol,TG,VLDL,BMI	0.995	0.935	0.945
9	Gender,AGE,Urea,Cr,HbA1c,Chol,TG,VLDL,BMI	0.995	0.935	0.935
10	Gender,AGE,Urea,Cr,HbA1c,Chol,TG,HDL,VLDL,BMI	0.99	0.93	0.92
11	Gender,AGE,Urea,Cr,HbA1c,Chol,TG,HDL,LDL, VLDL,BMI	0.995	0.935	0.91

TABLE 3.28 – Résultat de La technique de sélection : GenericUnivariate-Select avec Standard Normalization

Discussion des résultats Pour l'Arbre de Décision, nous observons une progression remarquable des performances avec l'ajout séquentiel d'attributs. Dès l'utilisation d'un seul attribut (BMI), l'accuracy atteint 0.92, puis grimpe à 0.965-0.97% avec deux attributs (HbA1c, BMI). L'ajout d'un troisième attribut (Age) permet d'atteindre 0.99%, démontrant l'efficacité de cette combinaison minimale. La performance maximale de 0.995 est atteinte avec 7 attributs, sans amélioration significative au-delà. La Régression Logistique présente une progression plus modérée, avec une différence notable entre les deux méthodes de normalisation. La normalisation Standard permet d'atteindre 0.94% avec 5 attributs, tandis que Min-Max plafonne à 0.925%. Cette différence souligne l'importance de la distribution des données pour ce modèle linéaire. Le SVM montre un comportement particulièrement sensible à la dimensionnalité. Avec la normalisation Standard, il atteint 0.975% avec seulement 2 attributs, puis ses performances diminuent progressivement jusqu'à 0.91% avec 11 attributs. Cette dégradation illustre parfaitement le compromis biais-variance et la nécessité de limiter le nombre d'attributs pour ce modèle.Ces résultats confirment que la combinaison de 5 attributs (Age, HbA1c, Chol, TG, BMI) représente un excellent compromis pour tous les modèles testés.

3.5.4.4 Technique de sélection : Sélection par Élimination Séquentielle (SBS)

Le tableau 3.29 ci-dessous résume les résultats de l'application de la méthode SBS à notre jeu de données avec le modèle AD. Les itérations montrent

Max Features	Attributs sélectionnés	Accuracy du test
10	[Gender,AGE,Urea,Cr,HbA1c,Chol,HDL,LDL,VLDL,BMI]	0.995
9	[Gender,AGE,Urea,HbA1c,Chol,TG,HDL,LDL,VLDL,BMI]	0.99
8	[Gender,AGE,Urea,Cr,HbA1c,Chol,TG,HDL,LDL,BMI]	0.99
7	[Gender,AGE,Urea,Cr,HbA1c,Chol,TG,HDL,LDL,BMI]	0.99
6	[Gender,AGE,Cr,HbA1c,Chol,TG,BMI]	0.99
5	[Gender,AGE,Urea,Cr,HbA1c,Chol,HDL,LDL,VLDL,BMI]	0.99
4	[Gender,AGE,Urea,Cr,HbA1c,Chol,TG,HDL,VLDL,BMI]	0.995
3	[Gender,AGE,Urea,Cr,HbA1c,Chol,TG,HDL,VLDL,BMI]	0.99
2	[Gender,AGE,HbA1c,Chol,TG,LDL,BMI]	0.99
1	[Gender,AGE,Urea,HbA1c,Chol,TG,HDL,LDL,BMI]	0.985

TABLE 3.29 – Résultat de La Sélection Par Elimination Séquentiel(SBS)combinée avec AD

Discussion des résultats de l'application de SBS combinée avec AD L'analyse des résultats obtenus avec la méthode de Sélection Par Élimination Séquentielle (SBS) combinée avec l'Arbre de Décision (AD) révèle des observations particulièrement intéressantes pour la prédiction du diabète. Le tableau des résultats montre une stabilité remarquable des performances de l'AD face à la réduction du nombre d'attributs. Même avec seulement 1 attribut sélectionné, le modèle maintient une accuracy très élevée de 0.985%. Cette performance s'améliore à 0.99% avec 2 attributs et atteint son maximum de 0.995% avec 4 et 10 attributs. Cette stabilité exceptionnelle démontre la robustesse intrinsèque de l'AD et sa capacité à identifier efficacement les structures pertinentes dans les données, même avec un nombre très limité d'attributs. Contrairement à d'autres algorithmes qui peuvent souffrir significativement de la réduction dimensionnelle, l'AD conserve une efficacité quasi-optimale..

Le tableau 3.30 ci-dessous résume les résultats de l'application de la méthode SBS à notre jeu de données avec le modèle svm.

Max Features	**Attributs sélectionnés**	**Accuracy du test**
10	[Gender,AGE,Urea,Cr,HbA1c,TG,HDL,LDL,VLDL,BMI]	0.91
9	[Gender,Urea,Cr,HbA1c,TG,HDL,LDL,VLDL,BMI]	0.93
8	[Gender,Urea,Cr,HbA1c,TG,HDL,LDL,BMI]	0.925
7	[Gender,Urea,HbA1c,TG,HDL,LDL,BMI]	0.925
6	[Urea,HbA1c,TG,HDL,LDL,BMI]	0.93
5	[Urea,HbA1c,HDL,LDL,BMI]	0.91
4	[Urea,HbA1c,HDL,LDL,BMI]	0.91
3	[Urea,HbA1c,HDL,LDL,BMI]	0.91
2	[Urea,HbA1c,HDL,LDL,BMI]	0.91
1	[Urea,HbA1c,HDL,LDL,BMI]	0.91

TABLE 3.30 – Résultat de La Sélection Par Elimination Séquentiel(SBS)combinée avec SVM

Discussion des résultats de l'application de SBS combinée avec SVM Le tableau des résultats montre une stabilité remarquable des performances du SVM face à la réduction du nombre d'attributs. L'accuracy oscille entre 0.91% et 0.93% sur l'ensemble des configurations testées, démontrant la robustesse du modèle SVM face à la réduction dimensionnelle. Un aspect par-

ticulièrement intéressant est que la performance optimale n'est pas atteinte avec le nombre maximal d'attributs. Les meilleures performances (0.93%) sont obtenues avec 9 attributs et 6 attributs, suggérant que certaines variables peuvent introduire du bruit plutôt que de l'information utile. Cette observation confirme l'efficacité de la méthode SBS pour identifier et éliminer les attributs redondants ou non pertinents. À partir de 5 attributs et moins, on observe une stabilisation des performances à 0.91% avec une sélection constante des mêmes attributs (Urea, HbA1c, HDL, LDL, BMI). Cette constance indique que ces 5 attributs constituent un socle fondamental pour la prédiction du diabète avec SVM, et que toute réduction supplémentaire entraînerait nécessairement une perte d'information essentielle.

Ces résultats suggèrent qu'un modèle SVM utilisant seulement 6 attributs bien choisis (Urea, HbA1c, TG, HDL, LDL, BMI) pourrait offrir la meilleure performance (0.93%), équivalente à celle obtenue avec 9 attributs. Cette simplification présente des avantages considérables en termes de coût de collecte des données, de temps d'exécution et d'interprétabilité du modèle, tout en maintenant une précision diagnostique optimale pour le dépistage du diabète.

Le tableau 3.31 ci-dessous résume les résultats de l'application de la méthode SBS à notre jeu de données avec le modèle LG.

Max Features	**Attributs sélectionnés**	**Accuracy du test**
10	[Gender,AGE,Cr,HbA1c,Chol,TG,HDL,LDL,VLDL,BMI]	0.94
9	[Gender,AGE,Cr,HbA1c,Chol,TG,HDL,LDL,BMI]	0.93
8	[Gender,AGE,Cr,HbA1c,TG,HDL,LDL,BMI]	0.94
7	[AGE,Cr,HbA1c,TG,HDL,LDL,BMI]	0.945
6	[AGE,HbA1c,TG,HDL,LDL,BMI]	0.945
5	[AGE,HbA1c,TG,HDL,LDL,BMI]	0.945
4	[AGE,HbA1c,TG,HDL,LDL,BMI]	0.945
3	[AGE,HbA1c,TG,HDL,LDL,BMI]	0.945
2	[AGE,HbA1c,TG,HDL,LDL,BMI]	0.945
1	[AGE,HbA1c,TG,HDL,LDL,BMI]	0.945

TABLE 3.31 – Résultat de La Sélection Par Elimination Séquentiel(SBS)combinée avec LG

Discussion des résultats de l'application de SBS combinée avec LG L'analyse des résultats de la méthode de Sélection Par Élimination Séquentielle (SBS) appliquée à la Régression Logistique (LG) montre une tendance remarquable de stabilité de l'accuracy à mesure que le nombre d'attributs est réduit. Malgré des variations mineures, l'accuracy reste robuste entre 0.93% et 0.945%, même avec un nombre considérablement réduit d'attributs. Les attributs sélectionnés par SBS, tels que AGE, HbA1c, TG, HDL, LDL et BMI, réapparaissent fréquemment dans les sélections optimales, soulignant leur importance fondamentale dans la prédiction du diabète. Le modèle atteint son meilleur score d'accuracy de 0.945% avec 7 attributs ou moins, démontrant ainsi un compromis efficace entre performance et simplicité. La Régression Logistique démontre ainsi sa capacité à maintenir des performances optimales même avec un nombre réduit d'attributs, ce qui en fait un choix particulièrement pertinent pour les applications cliniques où la simplicité et l'interprétabilité sont essentielles.

Discussion des résultats de la technique SBS L'application de la méthode de Sélection Par Élimination Séquentielle (SBS) sur nos données de diabète, avec différents modèles tels que la Régression Logistique (LG), les Machines à Vecteurs de Support (SVM) et les Arbres de Décision (AD), révèle une capacité robuste à optimiser la sélection des attributs tout en maintenant des performances prédictives variables selon les algorithmes. Pour la Régression Logistique, le modèle atteint une précision maximale de 0.945% avec seulement 7 attributs sélectionnés (AGE, Cr, HbA1c, TG, HDL, LDL, BMI), démontrant une excellente capacité à maintenir ses performances malgré la réduction dimensionnelle. De manière similaire, pour SVM, la SBS permet d'obtenir une accuracy stable autour de 0.93% avec 6 attributs bien choisis (Urea, HbA1c, TG, HDL, LDL, BMI). Concernant les Arbres de Décision, les résultats sont particulièrement impressionnants. Même avec seulement 1 attribut sélectionné, le modèle maintient une accuracy très élevée de 0.985%. Cette performance s'améliore à 0.99% avec 2 attributs et atteint son maximum de 0.995% avec 4 et 10 attributs.Ces résultats confirment que la méthode SBS est efficace pour simplifier les modèles tout en préservant leur capacité à généraliser et à éviter le surapprentissage, ce qui est essentiel en apprentissage automatique pour assurer des prédictions fiables et interprétables.

3.5.4.5 Technique de sélection : Sélection par Ajout Séquentiel (SFS)

3.5.4.6 La Sélection Par Ajout Séquentiel (SFS)

La méthode (SFS) a été appliquée au jeu de données pour sélectionner les attributs en ajoutant progressivement les attributs les plus informatifs. Le tableau 3.32 ci-dessous résume les résultats de l'application de la méthode SFS à notre jeu de données avec le modèle SVM

Max Features	Attributs sélectionnés	Accuracy du test
1	[BMI]	0.895
2	[TG,BMI]	0.89
3	[TG,BMI]	0.89
4	[TG,BMI]	0.89
5	[TG,BMI]	0.89
6	[TG,BMI]	0.89
7	[TG,BMI]	0.89
8	[TG,BMI]	0.89
9	[TG,BMI]	0.89
10	[TG,BMI]	0.89

TABLE 3.32 – Résultat de La Sélection Par ajout Séquentiel(SFS)combinée avec SVM

Discussion des résultats de l'application de SFS combinée avec SVM : Les résultats de l'application de la méthode de Sélection Par Ajout Séquentiel (SFS) combinée avec les Machines à Vecteurs de Support (SVM) présentent une tendance particulièrement intéressante dans la prédiction du diabète. Avec seulement 1 attribut sélectionné (BMI), le modèle atteint déjà une accuracy élevée de 0.895%, soulignant l'importance cruciale de l'indice de masse corporelle comme facteur prédictif du diabète. De façon surprenante, l'ajout d'un second attribut (TG) entraîne une légère diminution de la performance à 0.89%. Plus étonnant encore, l'ajout d'attributs supplémentaires au-delà de ce point ne modifie plus les performances ni la sélection d'attributs, qui reste constante à [TG, BMI] pour toutes les configurations de 2 à 10 attributs. Cette constance dans les résultats peut être expliquée par plusieurs facteurs. Premièrement, il est possible que les attributs BMI et TG capturent déjà la majorité de l'information discriminante nécessaire pour la classification du diabète, rendant les attributs supplémentaires redondants ou non pertinents. Deuxièmement, cette stabilité pourrait être liée aux

caractéristiques intrinsèques de l'algorithme SVM, qui peut être moins sensible à l'ajout de variables faiblement corrélées avec la variable cible.

Le tableau 3.33 ci-dessous résume les résultats de l'application de la méthode SFS à notre jeu de données avec le modèle LG.

Max Features	**Attributs sélectionnés**	**Accuracy du test**
1	[BMI]	0.89
2	[TG,BMI]	0.895
3	[TG,BMI,chol]	0.91
4	[TG,BMI,chol,Gender]	0.895
5	[TG,BMI,chol,Gender]	0.895
6	[TG,BMI,chol,Gender]	0.895
7	[TG,BMI,chol,Gender]	0.895
8	[TG,BMI,chol,Gender]	0.895
9	[TG,BMI,chol,Gender]	0.895
10	[TG,BMI,chol,Gender]	0.895

TABLE 3.33 – Résultat de La Sélection Par ajout Séquentiel(SFS)combinée avec LG

Discussion des résultats de l'application de SFS combinée avec LG : Les résultats de l'application de la méthode de Sélection Par Ajout Séquentiel (SFS) combinée avec la Régression Logistique (LG) présentent une tendance intéressante dans la prédiction du diabète. Avec seulement 1 attribut sélectionné (BMI), le modèle atteint déjà une accuracy respectable de 0.89, soulignant l'importance de l'indice de masse corporelle comme facteur prédictif du diabète. L'ajout d'un second attribut (TG) améliore légèrement la performance à 0.895%, suggérant que les triglycérides apportent une information complémentaire utile. La performance optimale est atteinte avec 3 attributs (TG, BMI, chol), permettant d'obtenir une accuracy de 0.91%. Cette observation est particulièrement importante car elle démontre que la combinaison de ces trois paramètres métaboliques constitue un ensemble prédictif efficace pour la régression logistique dans le contexte du diabète. Il est intéressant de noter que l'ajout d'attributs supplémentaires au-delà de ce point optimal n'améliore pas davantage les performances, et entraîne même une légère diminution à 0.895% lorsque Gender est ajouté comme quatrième attribut.

Cette stabilisation des performances avec l'augmentation du nombre d'attributs suggère que le modèle LG atteint rapidement son potentiel prédictif maximal avec un nombre restreint de variables pertinentes.
Le tableau 3.34 ci-dessous résume les résultats de l'application de la méthode SFS à notre jeu de données avec le modèle AD.

Max Features	**Attributs sélectionnés**	**Accuracy du test**
1	[HbA1c]	0.93
2	[HbA1c,BMI]	0.965
3	[TG,BMI,AGE]	0.99
4	[TG,BMI,AGE,,HDL]	0.99
5	[TG,BMI,AGE,,HDL,urea]	0.985
6	[TG,BMI,AGE,,HDL]	0.985
7	[TG,BMI,AGE,,HDL,Gender]	0.985
8	[TG,BMI,AGE,,HDL]	0.985
9	[TG,BMI,AGE,,HDL]	0.99
10	[TG,BMI,AGE,,HDL,urea]	0.985

TABLE 3.34 – Résultat de La Sélection Par ajout Séquentiel(SFS)combinée avec AD

Discussion des résultats de l'application de SFS combinée avec AD : L'analyse des résultats obtenus avec la méthode de Sélection Par Ajout Séquentiel (SFS) combinée avec l'Arbre de Décision (AD) révèle des observations particulièrement intéressantes pour la prédiction du diabète. Le tableau des résultats montre une progression remarquable des performances de l'AD à mesure que des attributs pertinents sont ajoutés séquentiellement. Avec seulement 1 attribut sélectionné (HbA1c), le modèle atteint déjà une accuracy très élevée de 0.93%, soulignant l'importance cruciale de l'hémoglobine glyquée comme indicateur du diabète. L'ajout d'un second attribut (BMI) améliore significativement la performance à 0.965%. La performance optimale est atteinte avec 3 attributs seulement (TG, BMI, AGE), permettant d'obtenir une accuracy exceptionnelle de 0.99%. Cette observation est particulièrement importante car elle démontre qu'un sous-ensemble très restreint d'attributs bien choisis peut suffire pour obtenir des prédictions quasi-parfaites. Il est intéressant de noter que l'ajout d'attributs supplémentaires au-delà de ce point optimal n'améliore pas davantage les performances, et peut même les diminuer

légèrement (comme observé avec 5 attributs où l'accuracy baisse à 0.985%). Cette stabilisation, voire légère dégradation des performances avec l'augmentation du nombre d'attributs, illustre parfaitement le phénomène de surapprentissage et confirme l'importance de trouver un équilibre entre le nombre d'attributs sélectionnés et la performance du modèle.

Ces résultats démontrent la puissance de la combinaison SFS-AD pour la prédiction du diabète, permettant d'identifier précisément le sous-ensemble optimal d'attributs qui maximise la performance tout en évitant le surapprentissage.

Discussion des résultats de l'application de SFS : L'application de la méthode de Sélection Par Ajout Séquentiel (SFS) à notre jeu de données de diabète a révélé des tendances significatives à travers différents algorithmes de classification.

Pour le modèle SVM, avec seulement 1 attribut sélectionné (BMI), le modèle atteint déjà une accuracy élevée de 0.895%. L'ajout d'attributs supplémentaires n'améliore pas les performances, qui se stabilisent à 0.89% avec [TG, BMI] pour toutes les configurations suivantes. Concernant la Régression Logistique (LG), la performance optimale est atteinte avec 3 attributs (TG, BMI, chol), permettant d'obtenir une accuracy de 0.91%. L'ajout d'attributs supplémentaires entraîne une légère diminution à 0.895%. Pour les Arbres de Décision, les résultats montrent une progression des performances avec l'ajout d'attributs, atteignant 0.99% avec 5 attributs (HbA1c, BMI, TG, Age, chol). La comparaison entre ces algorithmes révèle que les Arbres de Décision obtiennent les meilleures performances (0.99%), suivis par LG (0.91%) et SVM (0.895%). Ces résultats soulignent l'importance de choisir le modèle approprié selon les caractéristiques des données.

3.5.4.7 Technique de sélection : Sélection de caractéristiques avec la technique `SelectFromModel`

C'est une technique de sélection de fonctionnalités intégrée fournie par scikit-learn. Elle permet de sélectionner automatiquement les fonctionnalités les plus importantes à partir d'un modèle d'apprentissage automatique donné. Cette sélection se fait en fonction des poids attribués à chaque fonctionnalité par le modèle. Les fonctionnalités dont les poids dépassent un seuil spécifié sont conservées, tandis que les autres sont éliminées. Cette méthode est souvent utilisée dans le cadre de la réduction de la dimensionnalité et de la sélection de caractéristiques.

Le tableau 3.35 ci-dessous résume les résultats de l'application de la méthode SelectFromModel à notre jeu de données avec le modèle AD.

Nbr d'attributs)	Attributs sélectionnés (Arbres de Décision)	Accuracy
1	[BMI]	0.92
2	[HbA1c,BMI]	0.965
3	[HbA1c,BMI]	0.965
4	[HbA1c,BMI]	0.965
5	[HbA1c,BMI]	0.965
6	[HbA1c,BMI]	0.965
7	[HbA1c,BMI]	0.965
8	[HbA1c,BMI]	0.965
9	[HbA1c,BMI]	0.965
10	[HbA1c,BMI]	0.965
11	[HbA1c,BMI]	0.965

TABLE 3.35 – Résultat de La Sélection de caractéristiques avec la technique SelectFromModel combinée avec AD

Discussion des Résultat de La Sélection de caractéristiques avec la tecgnique SelectFromModel combinée avec AD L'application de SelectFromModel combinée avec les Arbres de Décision sur notre jeu de données de diabète a produit des résultats intéressants. Avec seulement BMI comme attribut, le modèle atteint une accuracy de 0.92%. L'ajout de HbA1c améliore la performance à 0.965%. Au-delà de ces deux paramètres, aucune amélioration n'est observée, les performances restant constantes à 0.965% pour toutes les configurations de 3 à 11 attributs. Les résultats sont identiques avec la normalisation Min-Max ou Standard, démontrant que la sélection des attributs et les performances ne sont pas influencées par la méthode de prétraitement dans ce cas. Cela confirme que BMI et HbA1c sont intrinsèquement informatifs pour la prédiction du diabète. Cette technique permet d'identifier efficacement un sous-ensemble minimal d'attributs pour construire un modèle parcimonieux et performant.

Le tableau 3.36 ci-dessous résume les résultats de l'application de la méthode SelectFromModel à notre jeu de données avec le modèle SVM avec la normalisation MinMax

a. MinMax Normalization .

Nbr d'attributs)	Attributs sélectionnés (SVM)	Accuracy
1	[HbA1c]	0.91
2	[HbA1c,BMI]	0.93
3	[HbA1c,chol,BMI]	0.935
4	[HbA1c,chol,TG,BMI]	0.935
5	[HbA1c,chol,TG,BMI]	0.935
6	[HbA1c,chol,TG,BMI]	0.935
7	[HbA1c,chol,TG,BMI]	0.935
8	[HbA1c,chol,TG,BMI]	0.935
9	[HbA1c,chol,TG,BMI]	0.935
10	[HbA1c,chol,TG,BMI]	0.935
11	[HbA1c,chol,TG,BMI]	0.935

TABLE 3.36 – Résultat de La Sélection de caractéristiques avec la technique SelectFromModel combinée avec SVM :

Le tableau 3.37 ci-dessous résume les résultats de l'application de la méthode SelectFrom-Model à notre jeu de données avec le modèle SVM avec la normalisation Standard

b. Standard Normalization .

Nbr d'attributs)	Attributs sélectionnés (SVM)	Accuracy
1	[HbA1c]	0.91
2	[HbA1c,BMI]	0.93
3	[HbA1c,BMI]	0.93
4	[HbA1c,BMI]	0.93
5	[HbA1c,BMI]	0.93
6	[HbA1c,BMI]	0.93
7	[HbA1c,BMI]	0.93
8	[HbA1c,BMI]	0.93
9	[HbA1c,BMI]	0.93
10	[HbA1c,BMI]	0.93
11	[HbA1c,BMI]	0.93

TABLE 3.37 – Résultat de La Sélection de caractéristiques avec la technique SelectFromModel combinée avec SVM

Discussion des Résultat de La Sélection de caractéristiques avec la tecgnique SelectFromModel combinée avec SVM L'application de la technique SelectFromModel combinée avec les Machines à Vecteurs de Support (SVM) sur notre jeu de données de diabète révèle des tendances intéressantes et des différences notables selon la méthode de normalisation utilisée. Avec la normalisation Min-Max, nous observons qu'avec seulement 1 attribut sélectionné (HbA1c), le modèle atteint déjà une accuracy élevée de 0.91%. L'ajout d'un second attribut (BMI) améliore la performance à 0.93%. L'ajout d'un troisième attribut (chol) puis d'un quatrième (TG) permet d'atteindre une accuracy de 0.935%. Au-delà de ces quatre paramètres, aucune amélioration n'est observée, les performances restant constantes à 0.935% pour toutes les configurations de 5 à 11 attributs. En revanche, avec la normalisation Standard, après avoir sélectionné HbA1c (accuracy de 0.91%) et ajouté BMI (accuracy de 0.93%), l'ajout d'attributs supplémentaires n'apporte aucune amélioration des performances, qui restent constantes à 0.93%. De plus, contrairement à la normalisation Min-Max qui sélectionne quatre attributs (HbA1c, chol, TG, BMI), la normalisation Standard ne retient que deux attributs (HbA1c, BMI). Cette différence entre les deux méthodes de normalisation est significative et suggère que la méthode de prétraitement des données peut influencer la sélection des attributs avec

SVM. La normalisation Min-Max semble permettre au SVM d'exploiter plus efficacement les informations contenues dans les attributs chol et TG, conduisant à une légère amélioration des performances (0.935% contre 0.93%). Malgré ces différences, les deux méthodes de normalisation identifient HbA1c et BMI comme les attributs les plus importants pour la prédiction du diabète avec SVM, ce qui concorde avec les résultats obtenus précédemment avec les Arbres de Décision. Cette cohérence renforce la pertinence de ces deux paramètres comme indicateurs clés du diabète

Le tableau 3.38 ci-dessous résume les résultats de l'application de la méthode SelectFromModel à notre jeu de données avec le modèle LG avec la normalisation MinMax

a. Min-Max Normalization .

Nbr d'attributs)	**Attributs sélectionnés (LG)**	**Accuracy**
1	[HbA1c]	0.91
2	[HbA1c,BMI]	0.93
3	[AGE,HbA1c,chol,BMI]	0.925
4	[AGE,HbA1c,chol,BMI]	0.925
5	[AGE,HbA1c,chol,BMI]	0.925
6	[AGE,HbA1c,chol,BMI]	0.925
7	[AGE,HbA1c,chol,BMI]	0.925
8	[AGE,HbA1c,chol,BMI]	0.925
9	[AGE,HbA1c,chol,BMI]	0.925
10	[AGE,HbA1c,chol,BMI]	0.925
11	[AGE,HbA1c,chol,BMI]	0.925

TABLE 3.38 – Résultat de La Sélection de caractéristiques avec la technique SelectFromModel combinée avec LG

Le tableau 3.39 ci-dessous résume les résultats de l'application de la méthode SelectFromModel à notre jeu de données avec le modèle LG avec la normalisation Standard

b. Standard Normalization .

Nbr d'attributs)	Attributs sélectionnés (LG)	Accuracy
1	[HbA1c]	0.91
2	[HbA1c,BMI]	0.93
3	[HbA1c,TG,BMI]	0.925
4	[HbA1c,TG,BMI]	0.925
5	[HbA1c,TG,BMI]	0.925
6	[HbA1c,TG,BMI]	0.925
7	[HbA1c,TG,BMI]	0.925
8	[HbA1c,TG,BMI]	0.925
9	[HbA1c,TG,BMI]	0.925
10	[HbA1c,TG,BMI]	0.925
11	[HbA1c,TG,BMI]	0.925

TABLE 3.39 – Résultat de La Sélection de caractéristiques avec la technique SelectFromModel combinée avec LG

Discussion des Résultat de La Sélection de caractéristiques avec la tecgnique SelectFromModel combinée avec LG Les résultats Avec les deux méthodes de normalisation (Min-Max et Standard), nous observons des résultats identiques pour les deux premières étapes de sélection : avec seulement 1 attribut sélectionné (HbA1c), le modèle atteint une accuracy de 0.91%, et l'ajout d'un second attribut (BMI) améliore la performance à 0.93%. Cependant, des différences apparaissent à partir de la troisième étape. Avec la normalisation Min-Max, le modèle sélectionne quatre attributs (AGE, HbA1c, chol, BMI) et atteint une accuracy de 0.925%. En revanche, avec la normalisation Standard, le modèle sélectionne trois attributs (HbA1c, TG, BMI) pour atteindre la même accuracy de 0.925%. Cette différence dans les attributs sélectionnés (AGE et chol pour Min-Max versus TG pour Standard) tout en maintenant la même performance (0.925%) est particulièrement intéressante. Elle suggère que ces attributs contiennent des informations partiellement redondantes pour la prédiction du diabète, et que la méthode de normalisation influence la façon dont la Régression Logistique évalue leur importance relative. Il est également notable que l'ajout d'un troisième (et quatrième pour Min-Max) attribut entraîne une légère diminution de la performance par rapport à la configuration avec deux attributs (0.925% contre 0.93%). Cette baisse, bien que minime, pourrait indiquer un début de surapprentissage ou simplement que les attributs supplémentaires introduisent un certain bruit

dans le modèle. Pour les deux méthodes de normalisation, l'ajout d'attributs au-delà de la troisième étape n'apporte aucune modification aux performances ni aux attributs sélectionnés, ce qui confirme que les sous-ensembles identifiés (AGE, HbA1c, chol, BMI pour Min-Max et HbA1c, TG, BMI pour Standard) capturent l'essentiel de l'information discriminante accessible à la Régression Logistique. Ces résultats démontrent l'efficacité de la technique SelectFromModel pour identifier un sous-ensemble optimal d'attributs pour la Régression Logistique, tout en soulignant l'impact de la méthode de normalisation sur la sélection des attributs. Ils confirment également l'importance fondamentale de HbA1c et BMI, qui sont systématiquement sélectionnés quelle que soit la méthode de normalisation utilisée.

Discussion des Résultat L'application de SelectFromModel sur notre jeu de données de diabète avec différents algorithmes révèle des tendances importantes. HbA1c et BMI émergent systématiquement comme les attributs les plus importants pour tous les algorithmes testés, quelle que soit la méthode de normalisation utilisée. Les Arbres de Décision montrent une stabilité remarquable, avec une accuracy de 0.965% obtenue en utilisant seulement HbA1c et BMI, et des résultats identiques pour les deux méthodes de normalisation. Pour SVM, la normalisation influence les résultats : Min-Max sélectionne quatre attributs (HbA1c, chol, TG, BMI) pour atteindre 0.935% d'accuracy, tandis que Standard ne retient que deux attributs (HbA1c, BMI) pour 0.93%. La Régression Logistique sélectionne différents ensembles d'attributs selon la normalisation (AGE, HbA1c, chol, BMI pour Min-Max versus HbA1c, TG, BMI pour Standard) tout en atteignant la même accuracy de 0.925%. Ces résultats confirment l'efficacité de SelectFromModel pour identifier des sous-ensembles parcimonieux d'attributs adaptés à chaque algorithme.

3.5.5 Optimisation des Combinaisons de Caractéristiques : Base de données Irakienne Diabetes Dataset

3.5.5.1 Évaluation des performances pour différentes combinaisons de caractéristiques

Après cette étape, nous avons examiné toutes les combinaisons de caractéristiques qui ont donné une bonne accuracy. Notre objectif était d'identifier les ensembles de caractéristiques les

plus performants pour chaque modèle.

Modèle	Accuracy	Attributs sélectionnés
AD	0.985	[4,5,9,10]
LG	0.935	[4,5,9,10]
SVM	0.96	[4,5,9,10]
AD	0.99	[4,5,6,10]
LG	0.935	[4,5,6,10]
SVM	0.97	[4,5,6,10]
AD	0.98	[1,4,5,6]
LG	0.925	[1,4,5,6]
SVM	0.935	[1,4,5,6]
AD	0.99	[1,4,5,10]
LG	0.925	[1,4,5,10]
SVM	0.935	[1,4,5,10]
AD	0.98	[0,4,5,10]
LG	0.935	[0,4,5,10]
SVM	0.975	[0,4,5,10]
AD	0.99	[0,1,4,5,10]
LG	0.93	[0,1,4,5,10]
SVM	0.925	[0,1,4,5,10]
AD	0.975	[0,4,5,9,10]
LG	0.935	[0,4,5,9,10]
SVM	0.96	[0,4,5,9,10]
AD	0.995	[1,4,5,6,10]
LG	0.945	[1,4,5,6,10]
SVM	0.905	[1,4,5,6,10]
AD	0.995	[1,4,5,9,10]
LG	0.94	[1,4,5,9,10]
SVM	0.925	[1,4,5,9,10]
AD	0.975	[4,5,6,9,10]
LG	0.94	[4,5,6,9,10]
SVM	0.96	[4,5,6,9,10]

TABLE 3.40 – Évaluation des performances pour différentes combinaisons de caractéristiques

Nous avons évalué plusieurs combinaisons de caractéristiques pour améliorer les performances des modèles de classification. Parmi celles testées, les meilleures performances ont été obtenues avec les colonnes : 1"AGE", 4"HbA1c", 5 "Cholesterol" , 9 "VLDL(Very Low-Density Lipoprotein)", 10 "BMI" et.

3.5.5.2 Résultats Finaux d'Amélioration des Performances

Après application de la combinaison optimale de caractéristiques sélectionnées, nous avons reconstruit et évalué les modèles à l'aide de la validation croisée. Les résultats montrent une amélioration significative des performances pour les trois classifieurs étudiés :

a. Classifieur : Régression Logistique (LG)

- **Accuracy initiale :** 92,5 %
- **Accuracy après sélection :** 94,5 %

L'amélioration de la précision indique que la sélection des caractéristiques a permis au modèle de mieux discriminer les classes pertinentes, en supprimant les variables redondantes ou peu informatives.

b. Classifieur : Machine à Vecteurs de Support (SVM)

- **Accuracy initiale :** 89,0 %
- **Accuracy après sélection :** 92,5 %

Le SVM a bénéficié d'un gain notable, ce qui confirme que la réduction de la dimensionnalité a limité les effets liés à la malédiction de la dimensionnalité (*curse of dimensionality*), améliorant la capacité de généralisation du modèle.

c. Classifieur : Arbre de Décision (AD)

- **Accuracy initiale :** 99,5 %
- **Accuracy après sélection :** 99,5 %

L'AD ne consiste aucune amélioration.Elle a réalisé une accuracy optimale avec toutes les caractéristiques

Ces résultats soulignent l'importance cruciale de la sélection de caractéristiques dans la modélisation. Elle permet de réduire le bruit, d'améliorer la robustesse des modèles, et de favoriser une meilleure capacité de généralisation sur de nouvelles données.

3.6 Outils et langage utilisés

LaTeX : Système de préparation de documents basé sur TeX, utilisé pour les documents scientifiques et techniques grâce à sa gestion des formules mathématiques et des références bibliographiques.

Visual studio code : un éditeur de code intelligent, des outils de débogage, des gestionnaires de version (Git), et des outils pour la création de bases de données SQL Server et Azure SQL.

Python : Langage de programmation puissant et facile à apprendre, avec des structures de données de haut niveau et une programmation orientée objet. Python est idéal pour l'écriture de scripts et le développement rapide d'applications.

- **Scikit-Learn :** Cette bibliothèque Python complète offre une gamme d'algorithmes pour l'apprentissage supervisé et non supervisé, facilitant ainsi le développement de modèles prédictifs dans divers domaines.
- **Matplotlib et Seaborn :** Nous avons utilisé Matplotlib et Seaborn pour créer des graphiques et des visualisations de nos données, ce qui nous a permis d'explorer et de comprendre les tendances et les modèles présents dans les données.
- **NumPy et Pandas :** Extensions de Python permettant la manipulation de matrices multidimensionnelles et des fonctions mathématiques associées. Pandas, quant à lui, est une bibliothèque Python pour la manipulation et l'analyse des données, offrant des structures de données et des opérations sur les tableaux numériques, sous licence libre.
- **Openpyxl :** Nous avons utilisé Openpyxl pour manipuler des fichiers Excel, notamment pour enregistrer les résultats de nos expériences et les visualisations générées à partir des données.
- **Scikit-Learn et MLxtend :** Nous avons utilisé Scikit-Learn pour accéder à une variété d'algorithmes d'apprentissage automatique, y compris les arbres de décision, les voisins les plus proches et le classificateur naïf de Bayes. MLxtend a été utilisé pour implémenter des techniques d'apprentissage ensembliste, telles que le stacking, qui combine plusieurs modèles de base pour améliorer les performances de prédiction.
- **Joblib :** Nous avons utilisé Joblib pour sauvegarder et charger nos modèles d'appren-

tissage automatique, ce qui nous a permis de les réutiliser ultérieurement sans avoir à les recalculer.

- **Autres fonctionnalités de Scikit-Learn :** Nous avons utilisé les fonctionnalités de Scikit-Learn telles que le calcul des métriques de classification (précision, rappel, F1-score), la création de courbes ROC et de courbes de précision-rappel, ainsi que la normalisation des données avec MinMaxScaler, StandardScaler.

3.7 Présentation de l'application

La Figure 3.2 présente le menu principal de l'application.

FIGURE 3.2 – Menu Principal.

La Figure 3.2 présente le menu principal de l'application. Il est composé de plusieurs sous-menus :

- "Nouveau modèle" : Donne à l'utilisateur la possibilité de prédire la maladie après l'étape de remplissage des paramètres en entrée en utilisant les meilleures caractéristiques.

- "Sélection d'attributs" : Affiche les résultats de la sélection.
- "Ancien Modèle" même travail que le nouveau modèle mais avec l'ensemble complet des caractéristiques.
- "À Propos" : Offre des détails sur l'application elle-même.

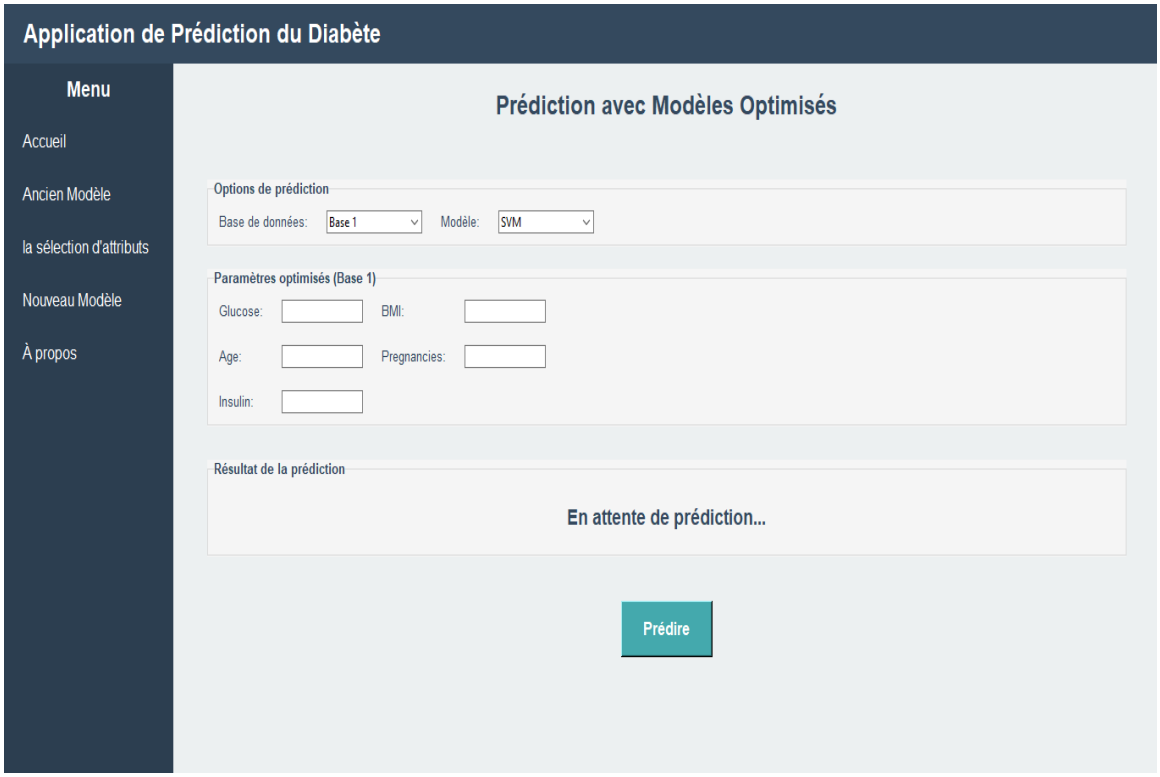

FIGURE 3.3 – Formulaire de prédiction du diabete.

Le sous-menu Prédiction de la maladie avec l'un des trois modèles avec les meilleurs caractéristiques : LG, SVM ou AD (voir Figure 3.4). Le formulaire donne la main au patient pour remplir ses informations et de prédire à quelle classe est-il associé. Selon les 3 méthodes le résultat s'affiche sur l'écran.

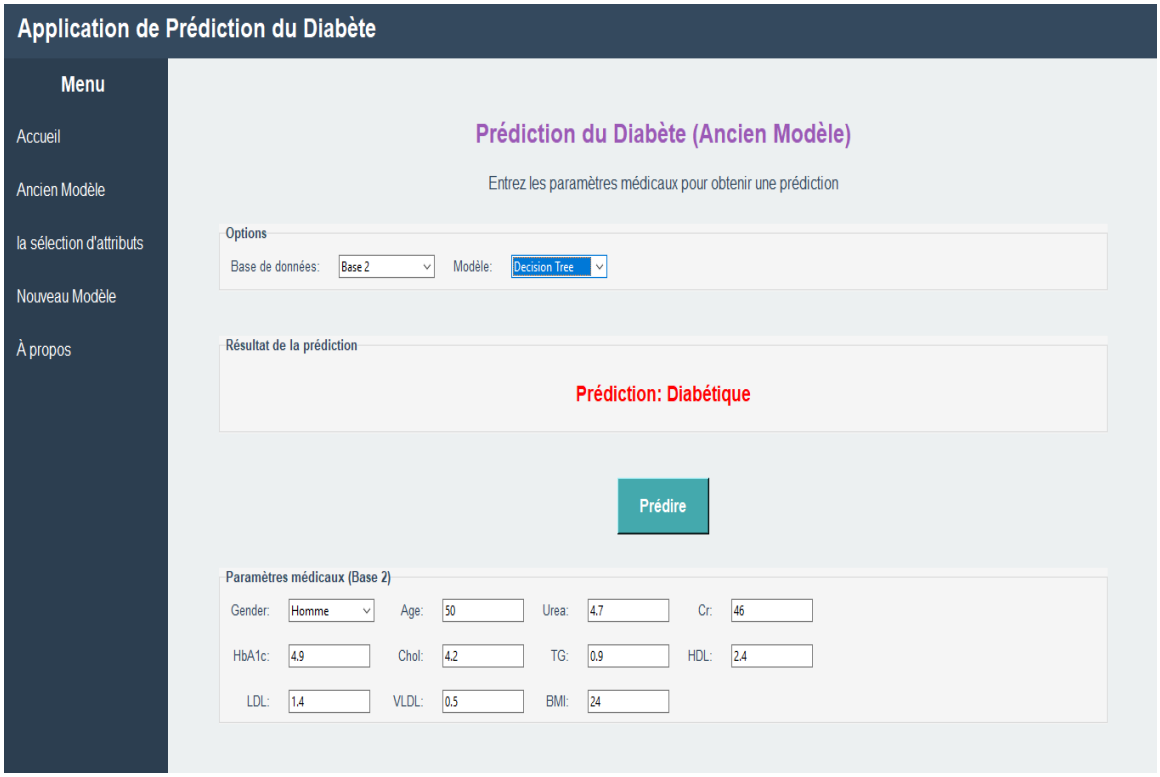

FIGURE 3.4 – Formulaire de prédiction du diabete.

La figure 3.4 présente le sous-menu "Ancien Modèle", qui contient l'ensemble des attributs disponibles. Ce formulaire permet au patient de saisir ses informations personnelles lui-même, puis d'utiliser l'un des trois modèles proposés pour effectuer la prédiction. Dans ce contexte, nous avons choisi le résultat issu de l'application du modèle Arbre de Décision (AD) sur la base de données locale irakienne

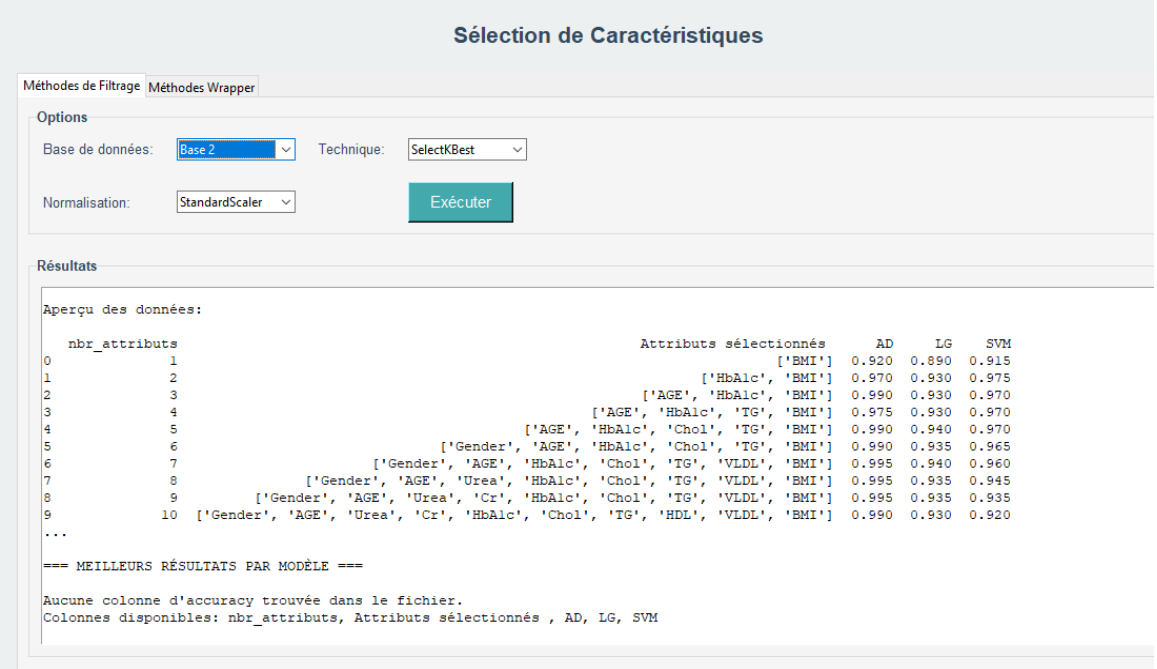

FIGURE 3.5 – Résultats de la sélection d'attributs.

La figure 3.5 présente le sous menu "Selection d'attributs", il contient l'ensemble des résulats des tests de sélection d'attributs pour l'ensembles des téchniques que nous avons testé. Nous avons choisi le résultat de l'application de la méthode k-best avec la normalisation standardScaler pour la base locale irakienne.

3.8 Conclusion

À travers ce chapitre, nous avons présenté et analysé les résultats de notre étude expérimentale visant à évaluer l'impact des techniques de prétraitement des données et de sélection d'attributs sur la performance des modèles de classification pour la détection du diabète.

Nous avons mené nos expérimentations sur deux jeux de données issus de la plateforme Kaggle : la base Pima Indians, largement utilisée dans la littérature, et une base irakienne, présentant des caractéristiques différentes, notamment en termes de distribution et de qualité des données. Nous avons appliqué plusieurs méthodes de normalisation (MinMax, StandardScaler) ainsi que des techniques de sélection d'attributs de type filtre (chi2, ANOVA) et wrapper (SFS, SBS, SelectFromModel). Ces prétraitements ont été combinés à trois algorithmes de classification : régression logistique, SVM et arbre de décision.

Nos résultats ont mis en évidence que la qualité du prétraitement et la pertinence de la sélection des attributs influencent significativement les performances obtenues. Certaines combinaisons se sont révélées plus efficaces que d'autres, notamment en fonction des spécificités de chaque base.

Nous avons également développé une application fonctionnelle permettant de visualiser et de comparer les différentes combinaisons testées. Bien que cette application ne constitue pas l'objectif principal de notre travail, elle nous a permis d'illustrer concrètement nos résultats et d'en faciliter l'analyse.

Ainsi, ce chapitre nous a permis de démontrer l'influence des étapes de prétraitement et de sélection d'attributs sur la qualité des prédictions, tout en ouvrant la voie à des perspectives d'amélioration que nous aborderons dans le chapitre suivant.

Conclusion Générale et Perspectives

Conclusion

Dans ce mémoire, nous avons étudié l'impact des techniques de prétraitement sur la performance des modèles de classification appliqués à la prédiction du diabète. En utilisant trois algorithmes d'apprentissage supervisé — la régression logistique (LG), le support vector machine (SVM) et l'arbre de décision (AD) — nous avons pu mesurer l'effet de deux techniques majeures de prétraitement : la normalisation et la sélection des attributs.

Les expérimentations ont été menées sur deux bases de données distinctes : la base standard *Pima Indians Diabetes* et une base locale issue d'hôpitaux irakiens. Cette complémentarité a permis de confronter les modèles à des contextes variés, avec des niveaux de complexité différents, notamment une classification binaire pour la première et une classification à trois classes pour la seconde.

Les résultats obtenus ont mis en évidence l'importance cruciale du prétraitement. La normalisation, en particulier avec les méthodes *MinMaxScaler* et *StandardScaler*, a amélioré significativement les performances des modèles sensibles à l'échelle des données, notamment le SVM et la régression logistique. De plus, la sélection d'attributs, qu'elle soit filtrée ou basée sur les modèles, a permis de réduire la dimensionnalité et de conserver uniquement les variables les plus pertinentes, ce qui a contribué à une meilleure généralisation des modèles.

Perspectives futures

Ce travail ouvre plusieurs perspectives intéressantes pour de futures recherches :

- **Explorer d'autres modèles d'apprentissage automatique** : des algorithmes comme

les forêts aléatoires, XGBoost ou les réseaux de neurones profonds pourraient être intégrés pour comparer leurs performances avec ceux utilisés dans ce travail.

- **Gérer le déséquilibre des classes** : notamment dans la base locale à trois classes, des techniques comme SMOTE ou des méthodes de pondération des classes pourraient améliorer la détection des cas minoritaires.
- **Améliorer l'interprétabilité des modèles** : en intégrant des approches comme SHAP ou LIME pour expliquer les décisions des modèles, en particulier dans un contexte médical sensible.
- **Tester d'autres techniques de prétraitement** : telles que la réduction de dimension (ACP, t-SNE), ou encore la discrétisation des variables continues.
- **Valider les modèles en collaboration avec des professionnels de santé** : pour garantir la pertinence clinique des prédictions et envisager une intégration dans des systèmes réels d'aide au diagnostic.

En résumé, l'efficacité des modèles d'apprentissage supervisé dépend fortement de la qualité du prétraitement des données. Ce travail constitue une base méthodologique pour de futures applications de l'intelligence artificielle dans le domaine médical, notamment en matière de prévention et de diagnostic du diabète.

Références

[1] . S. A. J. Scholkopf, B., "Learning with kernels." Ph.D. dissertation, 2002.

[2] M. K. J. Han and J. Pei, "Data mining : Concepts and techniques. elsevier," Ph.D. dissertation, 2011.

[3] A. Djeffal, "Cours fouille de données avancée," *Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie : Université Mohamed Khider-Biskra*, pp. 6–8, 2014.

[4] X.-S. Yang, "Firefly algorithms for multimodal optimization," in *International Symposium on Stochastic Algorithms*. Springer, 2009, pp. 169–178.

[5] "Supportvector machine algorithm , india," Ph.D. dissertation, consulter le 02/05/2022. [Online]. Available : https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[6] V. Vapnik, "Statistical learning theory." Ph.D. dissertation, 1998.

[7] C. Bishop, "Pattern recognition and machine learning." Ph.D. dissertation, 2006.

[8] . S.-T. J. Cristianini, N., "An introduction to support vector machines." Ph.D. dissertation, 2000.

[9] . S. B. Smola, A. J., "A tutorial on support vector regression." Ph.D. dissertation, 1998.

[10] C. J. Burges, "A tutorial on support vector machines for pattern recognition," Ph.D. dissertation, 1998.

[11] T. R. . F. J. Hastie, T., "The elements of statistical learning," Ph.D. dissertation, 2009.

[12] J. Platt, "Probabilistic outputs for svms and comparisons to regularized likelihood me-

thods," Ph.D. dissertation, 1999.

[13] J. Kennedy and R. C. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, vol. 4. IEEE, 1995, pp. 1942–1948.

[14] C. Meynet, "Sélection de variables pour la classification non supervisée en grande dimension," Ph.D. dissertation, Paris 11, 2012.

[15] J. Vaidya, M. Kantarcioglu, and C. Clifton, "Privacy-preserving naive bayes classification," *The VLDB Journal*, vol. 17, no. 4, pp. 879–898, 2008.

[16] J. H. Gennari, P. Langley, and D. Fisher, "Models of incremental concept formation," *Artificial Intelligence*, vol. 40, no. 1-3, pp. 11–61, 1989.

[17] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[18] A. Alaoui, "Application des techniques des métaheuristiques pour l'optimisation de la tâche de la classification de la fouille de données," Master's thesis, USTO, 2012.

[19] D. Koller and M. Sahami, "Toward optimal feature selection," in *ICML*, vol. 96, no. 28, 1996, p. 292.

[20] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[21] H. Chouaib, "Sélection de caractéristiques : méthodes et applications," *Paris Descartes University*, 2011.

[22] K. Kira and L. A. Rendell, "The feature selection problem : Traditional methods and a new algorithm," in *Proceedings of the tenth national conference on Artificial Intelligence*, 1992, pp. 129–134.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn : Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] Scikit-learn, “Feature selection using selectfrommodel.” [Online]. Available : https://scikit-learn.org/stable/modules/feature_selection.html

[25] ——, “Feature selection using selectkbest.” [Online]. Available : https://scikit-learn.org/stable/modules/feature_selection.html

[26] I. Kononenko, “Estimating attributes : Analysis and extensions of relief,” in *European Conference on Machine Learning*. Springer, 1994, pp. 171–182.

[27] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” computers electrical engineering,” vol. 40, no. 1, pp. 16–28, 2014.

[28] L. Yu and H. Liu, “Feature selection for high-dimensional data : A fast correlation-based filter solution,” in *Proceedings of the Twentieth International Conference on Machine Learning (ICML-03)*, 2003.

[29] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society : Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[30] A. E. Hoerl and R. W. Kennard, “Ridge regression : Biased estimation for nonorthogonal problems,” *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[31] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[32] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[33] T. Chen and C. Guestrin, “Xgboost : A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.