République Tunisienne

Ministère de l'Enseignement Supérieur

et de la Recherche Scientifique

Université de Tunis El Manar

Institut Supérieur d'Informatique d'El Manar

Rapport de Projet

Réalisé dans le cadre d'un module d'apprentissage

Spécialité : Ingénierie de développement logiciel

Par

Ghassen Mastouri (2IDL01)

Maha Houidi (2IDL01)

Mariem Ben Salah (2IDL01)

Mohamed Aziz Benboubakr (2IDL02)

Analyse et détection du diabète par fouille de données et apprentissage automatique sur des données cliniques

Encadrant académique : **Monsieur Sahbi Bahroun** Maître Assistant

Réalisé au sein de ISI

Table des matières

Table des figures

Liste des tableaux

Introduction générale

La révolution du numérique a incroyablement accru la quantité de données cliniques disponibles offrant ainsi des opportunités sans pareille pour améliorer la prévention et la gestion des maladies chroniques. Parmi celles-ci, le diabète représente un enjeu de santé publique majeur, en raison de sa progression rapide à l'échelle mondiale.

Face à ce défi, les techniques d'apprentissage automatique se sont imposées comme des outils puissants, capables d'extraire des connaissances utiles à partir de données complexes. Elles permettent aujourd'hui d'obtenir des résultats prometteurs en matière de diagnostic précoce. Dans un domaine aussi exigeant que la santé, où la précision et la fiabilité sont essentielles, le recours à des modèles prédictifs robustes constitue une avancée importante pour soutenir la prise de décision clinique.

C'est dans ce cadre que s'inscrit notre projet qui vise à analyser et exploiter un jeu de données réel et prédire la présence ou non du diabète grâce à des techniques d'apprentissage supervisé et non supervisé. Pour mettre en œuvre ce projet, nous avons recours à la méthodologie CRISP-DM spécifique aux projets de science des données. Cette méthodologie comprend six étapes essentiels : la compréhension du besoin métier, la compréhension des données, la préparation des données, la modélisation, l'évaluation e finalement le déploiement.

Le présent rapport documente chacune de ces étapes, en mettant l'accent sur la qualité des données, la rigueur analytique, la pertinence des modèles utilisés, ainsi que sur l'interprétation des résultats dans un contexte de détection de diabète.

Chapitre 1

Cadre du projet et compréhension métier

Plan

Introduction

Durant ce premier chapitre, nous allons introduire le cadre du projet et ses objectifs, suivie d'une compréhension du besoin métier.

1.1 Cadre du projet

Ce projet s'inscrit dans le cadre de notre formation en science des données au cours de la 2ème année de l'ingénierie de developpement logiciel au sein de l'institut supérieur d'informatique, avec pour objectif de mettre en pratique les différentes étapes d'un processus d'analyse de données réelles. Il s'agit d'un mini-projet complet mobilisant des techniques de fouille de données, d'apprentissage automatique supervisé et non supervisé, en s'appuyant sur un jeu de données médical lié à la prédiction de la présence de diabète chez des patients.

1.2 Objectifs du projet

Les principaux objectifs pédagogiques et techniques de ce projet sont les suivants :

- Explorer et prétraiter un dataset réel, fourni sous forme de fichier CSV, contenant des données cliniques relatives à différents patients.
- Appliquer des algorithmes d'apprentissage supervisé pour la tâche de classification.
- Mettre en œuvre des méthodes non supervisées afin de détecter des regroupements ou profils types de patients sans étiquette.
- Développer une interface graphique interactive permettant de présenter les résultats aux utilisateurs de façon intuitive et exploitable.

1.3 Compréhension métier

Dans le contexte de détection précoce des personnes à risque de diabète, l'objectif métier consiste à analyser un ensemble de données cliniques réelles afin de prédire la présence ou non de diabète chez des patientes, et d'explorer des structures de regroupement naturelles pour mieux comprendre les différents profils de santé. Ces résultats visent à illustrer comment les approches data-driven peuvent compléter l'expertise médicale et soutenir la prise de décision en milieu clinique.

Conclusion

Ce chapitre a présenté le cadre général du projet ainsi que l'aspect métier. Le chapitre suivant portera sur la phase de compréhension de données.

Chapitre 2

Phase de compréhension des données

Plan

Introduction

L'objectif de la phase de compréhension des données consiste à explorer et évaluer en profondeur les données disponibles afin d'en comprendre la structure, la qualité et la pertinence pour les objectifs du projet, ce qui servira à identifier d'éventuelles incohérences, valeurs manquantes ou anomalies qui pourraient affecter les analyses futures. Ce chapitre expose les résultats de cette étape, présentant les principales caractéristiques du jeu de données utilisé, les premiers constats réalisés lors de son exploration, ainsi que les visualisations clés et les recommandations initiales.

2.1 Présentation du jeu de données

Le jeu de données utilisé est un ensemble de données cliniques, contenant 768 observations et 8 variables explicatives quantitatives, ainsi qu'une variable cible binaire 'Outcome', qui indique la présence (1) ou l'absence (0) de diabète. Les données contiennent des facteurs de diabète variés relatifs à des données démographiques telles que l'âge et le nombre de grossesses, des marqueurs cliniques tels que le taux de glucose et d'insuline, ainsi que des indices de santé comme l'indice de masse corporelle. Nous pouvons aussi remarquer, selon la figure 2.1, qu'il y a aucune valeur manquante.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #  Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0  Pregnancies               768 non-null    int64
 1  Glucose                   768 non-null    int64
 2  BloodPressure             768 non-null    int64
 3  SkinThickness             768 non-null    int64
 4  Insulin                   768 non-null    int64
 5  BMI                       768 non-null    float64
 6  DiabetesPedigreeFunction  768 non-null    float64
 7  Age                       768 non-null    int64
 8  Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Figure 2.1 : Structure du dataset

La figure 2.2 présente les statistiques descriptives relatives à notre dataset. Une analyse du tableau permet d'identifier que les données révèlent des tendances et des anomalies notables. En effet, le nombre moyen de grossesses est de 3,85, avec une valeur maximale suspecte de 17, suggérant un

possible outlier. Une valeur médiane de 29 ans pour l'âge des patients, avec une moyenne de 33 ans, indique que la distribution est asymétrique et que la majorité sont des jeunes adultes. De plus, certaines variables cliniques, comme le glucose et la pression artérielle, présentent des valeurs minimales à 0, biologiquement impossibles, ce qui soulève des questions sur la qualité des données (possibilité de valeurs manquantes mal codées).

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768	768	768	768	768	768	768	768	768
mean	3.8451	120.8945	69.1055	20.5365	79.7995	31.9926	0.4719	33.2409	0.349
std	3.3696	31.9726	19.3558	15.9522	115.244	7.8842	0.3313	11.7602	0.477
min	0	0	0	0	0	0	0.078	21	0
25%	1	99	62	0	0	27.3	0.2438	24	0
50%	3	117	72	23	30.5	32	0.3725	29	0
75%	6	140.25	80	32	127.25	36.6	0.6263	41	1
max	17	199	122	99	846	67.1	2.42	81	1

Figure 2.2 : Aperçu des statistiques descriptives

2.2 Vérification des valeurs nulles ou aberrantes

Suite à l'analyse des statistiques descriptives, il est désormais essentiel de vérifier le nombre de valeurs nulles pour chaque variable afin d'identifier celles qui nécessitent un traitement.

Valeurs égales à zéro :

	0
Pregnancies	111
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374
BMI	11
DiabetesPedigre	0
Age	0
Outcome	500

Figure 2.3 : Aperçu des valeurs nulles

L'analyse exploratoire des données à l'aide de boxplots permet également de révèler plusieurs valeurs aberrantes et potentielles erreurs dans le dataset. Des variables comme Glucose, BloodPressure, SkinThickness, BMI et surtout Insulin présentent des valeurs nulles ou extrêmes peu réalistes. Par exemple, l'insuline montre une distribution très asymétrique avec de nombreux outliers au-delà de 300. De même, BMI et Diabetes Pedigree Function présentent une dispersion importante. Ces observations soulignent la nécessité d'un prétraitement rigoureux, incluant l'identification et la gestion des valeurs aberrantes, afin de garantir la fiabilité des modèles prédictifs appliqués par la suite.

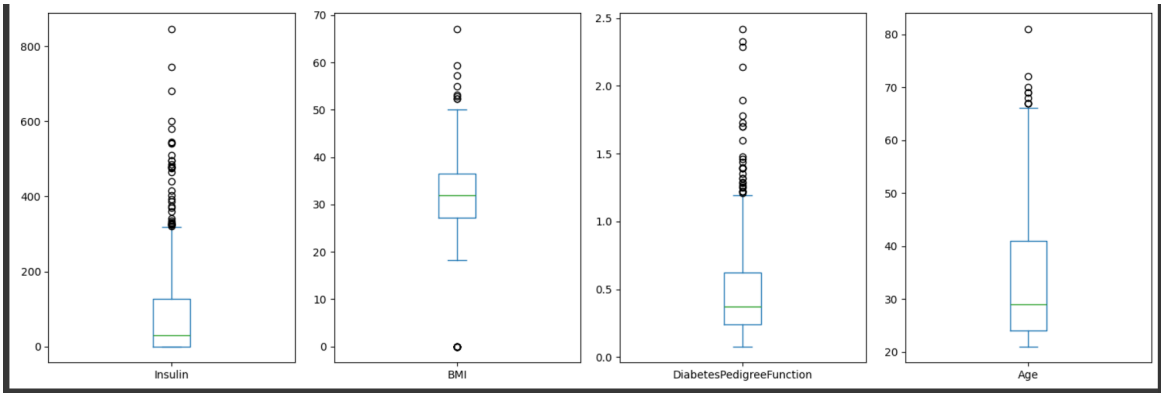

FIGURE 2.4 : Distribution par boxplots -1

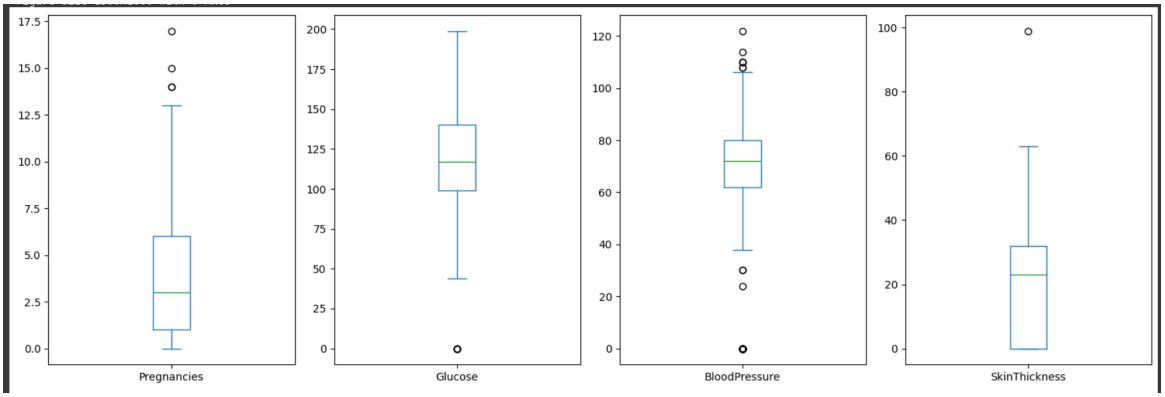

FIGURE 2.5 : Distribution par boxplots -2

2.3 Corrélations entre les variables

Pour étudier les corrélations entre les variables, nous avons recours à la matrice de corrélation, présentée dans la figure 2.6. Cette matrice permet d'identifier que l'âge et le nombre de grossesses (pregnancies) sont fortement corrélés et que le taux du glucose est la variable la plus corrélée avec la variable cible (la présence ou non du diabète).

Conclusion

Cette phase d'exploration nous a permis de bien comprendre le jeu de données et mettre en évidence des aspects essentiels tels que la présence de données incohérentes et des relations significatives entre certaines variables, ce qui est fondamental pour les phases à venir. Le chapitre suivant portera sur l'étape de nettoyage des données.

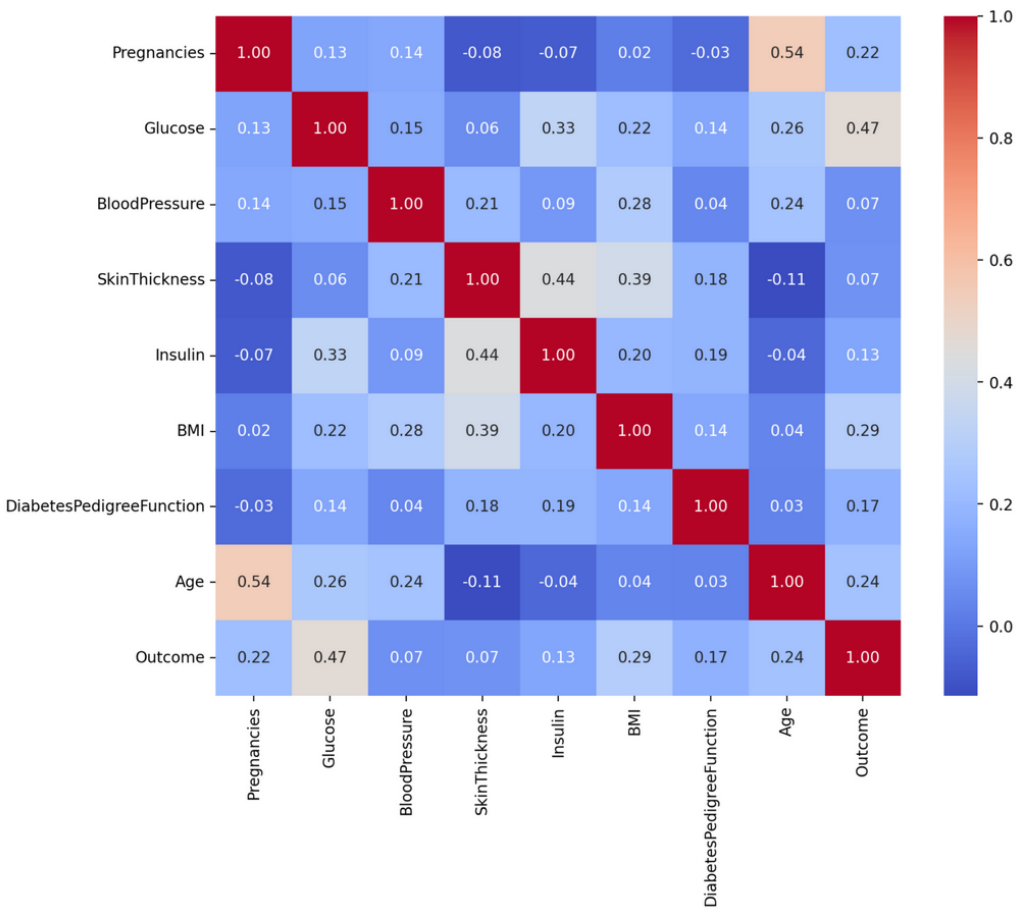

Figure 2.6 : Matrice de corrélation

Chapitre 3

Phase de péparation et nettoyage de données

Plan

Introduction

Le nettoyage des données est une étape essentielle du processus d'analyse. Dans ce chapitre, nous allons expliquer la méthode de gestion des données nulles ou aberrantes identifiées lors de l'analyse exploratoire.

3.1 Nettoyage de données

Afin de corriger les valeurs aberrantes détectées dans plusieurs variables (notamment Insulin, Glucose, BloodPressure, SkinThickness, etc.), nous avons adopté la méthode de remplacement par la médiane. Cette approche est adaptée dans ce contexte car elle est robuste aux valeurs extrêmes et permet de conserver la distribution globale des données sans introduire de biais important. Concrètement, pour chaque variable continue où des valeurs nulles ou irréalistes ont été identifiées (souvent égales à zéro, ce qui est médicalement incohérent), celles-ci ont été remplacées par la médiane non nulle de la variable correspondante.

3.2 Visualisation post-nettoyage

Après le remplacement des outliers par la médiane, les boxplots, dans la figure 3.1, montrent une nette amélioration de la distribution des variables. Les valeurs extrêmes ont été largement réduites, notamment pour Insulin, BMI, Glucose et SkinThickness, où les zéros incohérents ont été corrigés. Les distributions sont désormais plus centrées et homogènes, reflétant des données cliniquement plausibles. Quelques outliers persistants sont conservés lorsqu'ils sont justifiés, comme pour Pregnancies ou Diabetes Pedigree Function. Globalement, les variables présentent une répartition plus réaliste et exploitable pour l'analyse à venir.

Conclusion

Grâce à l'utilisation de la médiane pour le traitement des valeurs aberrantes, nous avons obtenu un jeu de données propre, cohérent et prêt pour les phases de modélisation et d'évaluation, explorées dans le prochain chapitre. Cette étape renforce la fiabilité des résultats futurs.

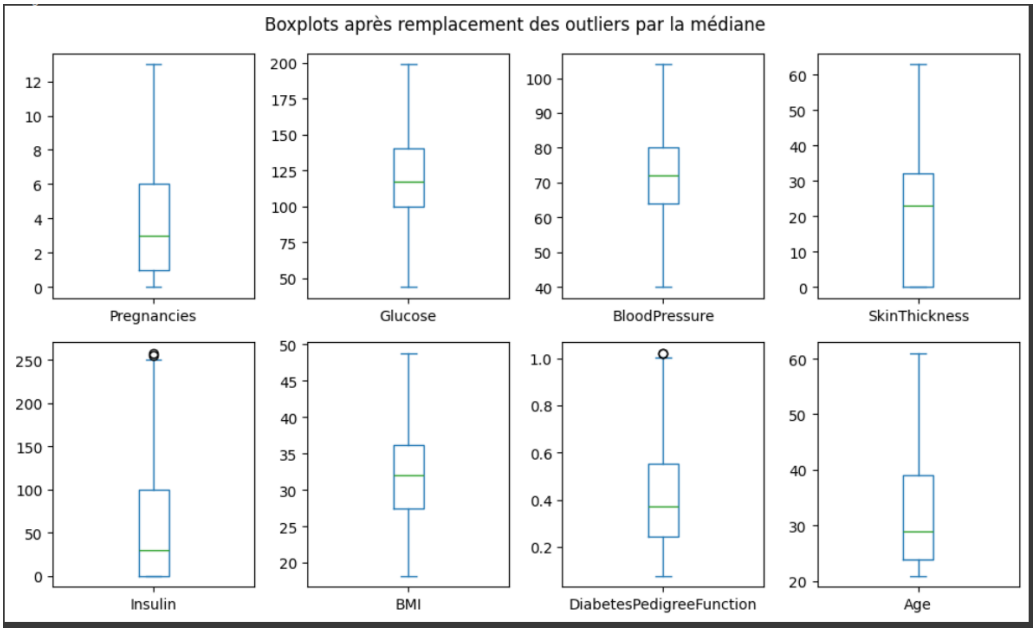

Figure 3.1 : Distribution par boxplots post-nettoyage

Chapitre 4

Phases de Modélisation et d'évaluation

Plan

Introduction

Après la phase de préparation des données, la modélisation consiste à appliquer différentes techniques d'apprentissage automatique afin de construire des modèles capables de prédire ou d'extraire des patterns à partir des données. Durant ce chapitre, nous allons justifier le choix des modèles d'apprentissage supervisé et non supervisé puis explorer le résultat de chacun.

4.1 Choix des modèles d'apprentissage

Le choix des modèles d'apprentissage utilisés dans ce projet repose sur les caractéristiques du dataset clinique étudié, notamment la nature numérique des variables explicatives et la nature qualitative de la variable cible binaire, ainsi que la complexité des interactions entre facteurs. La sélection intègre à la fois des approches supervisées, pour la classification du diabète, et non supervisées, pour le clustering.

Le tableau 4.1 résume les modèles retenus, en détaillant leurs avantages, leurs limites, ainsi que les raisons de choix.

Tableau 4.1 : Comparaison des modèles d'apprentissage

Modèle	**Type**	**Avantages**	**Limites**	**Justification du choix**
Régression logistique	Supervisé	Simple, interprétable	Linéaire, sensible aux outliers	Référence de base pour la comparaison
Random Forest	Supervisé	Robuste, gère non-linéarités	Moins interprétable	Bon compromis précision/interprétation
SVM (RBF)	Supervisé	Efficace sur marges étroites	Sensible aux paramètres	Performant sur données médicales
K-Means	Non supervisé	Rapide, intuitif	Nécessite k, sensible aux outliers	Exploration initiale des groupes
DBSCAN	Non supervisé	Détecte bruit et formes complexes	Sensible aux paramètres	Identifier anomalies et densités

4.2 Modèles d'apprentissage

Nous avons commencé tout d'abord par la séparation des données entre base d'apprentissage (80%) et base de test (20%).

4.2.1 Modèles de prédiction

4.2.1.1 Regression logistique

La regression logistique est utilisée pour modéliser la probabilité qu'un individu présente un diabète, à partir de caractéristiques cliniques fournies. Le modèle a été entraîné sur l'ensemble d'apprentissage et évalué sur l'ensemble de test. Pour étudier la performance du modèle, nous avons recours au calcul des métriques, présentées dans la figure 4.1, ainsi que la matrice de confusion.

	precision	recall	f1-score	support
0	0.7593	0.82	0.7885	100
1	0.6087	0.5185	0.56	54
accuracy	0.7143	0.7143	0.7143	0.7143
macro avg	0.684	0.6693	0.6742	154
weighted avg	0.7065	0.7143	0.7084	154

Figure 4.1 : Métriques pour regression logistique

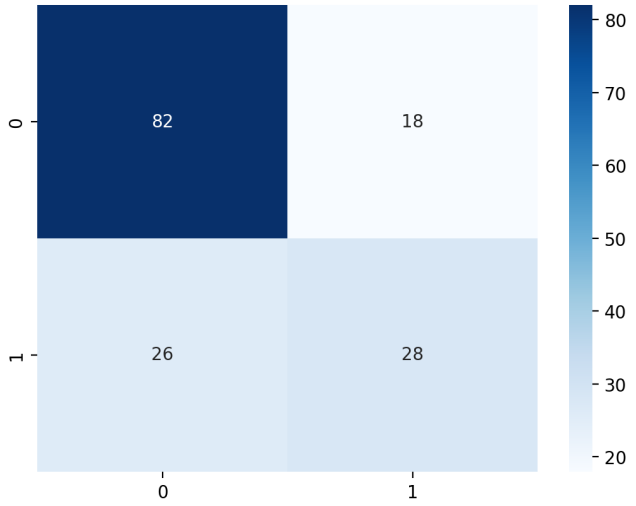

Figure 4.2 : Matrice de confusion de la regression logistique

- Le taux de précision global (accuracy) est de 71%.
- La précision indique la proportion de prédiction correcte. Nous pouvons remarquer que la classe 0 (absence de diabète) est mieux prédite que la classe 1 (présence de diabète)
- Le rappel (recall) qui est la capacité à retrouver les instances positives nous permet de conclure que le modèle permet de détecter 51,85% de vrais positifs.
- Le F1 score, étant le compromis entre la précision et le rappel, est meilleur pour la classe 0.
- La courbe ROC et AUC permet d'évaluer la capacité globale de discrimination du modèle. L'AUC est égale à 0.82 (>0.8) ce qui indique un bon modèle.

4.2.1.2 Forêt aléatoire

La forêt aléatoire est un algorithme basé sur la combinaison de plusieurs arbres de décision construits sur des sous-échantillons aléatoires d'observations et des variables. Le modèle a été entraîné avec 100 arbres. Les résultats sur la base de test sont :

- Le taux de précision global (accuracy) est de 75%.
- La précision est de 72%. Nous pouvons remarquer que la classe 0 est mieux prédite que la classe 1. (82 FN vs 28 TP).
- Le rappel (recall) : 69%.
- Le F1 score : 70%. Il est aussi meilleur pour la classe 0 (82%)
- L'AUC est égale à 0.81 (>0.8) ce qui indique un bon modèle.

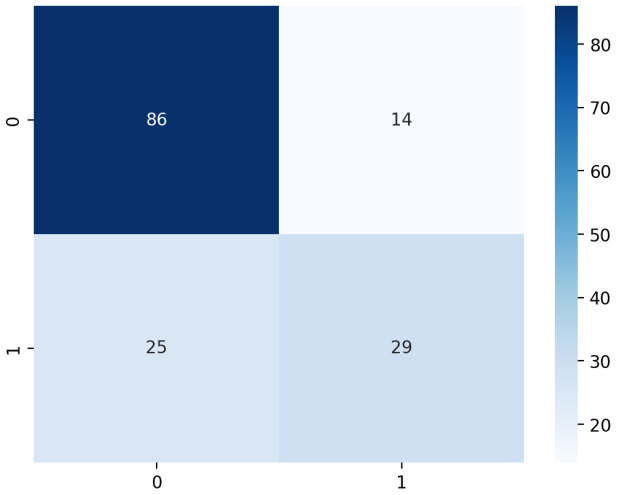

Figure 4.3 : Matrice de confusion de la forêt aléatoire

Contrairement à la régression logistique, la forêt aléatoire ne suppose aucune forme fonctionnelle entre les variables et la variable cible, ce qui lui permet de mieux capter des effets combinés.

4.2.1.3 Machine à Vecteurs de Support (SVM)

Le Support Vector Machine est un modèle efficace pour la classification binaire, surtout lorsqu'il existe une frontière complexe entre les classes. Le SVM est utilisé pour maximiser la marge entre les individus diabétiques et non diabétiques, sur la base de leurs caractéristiques cliniques. Le modèle donne les résultats présentés dans les figures 4.4 et 4.5, ainsi qu'un AUC égal à 0.79.

	precision	recall	f1-score	support
0	0.7961	0.82	0.8079	100
1	0.6471	0.6111	0.6286	54
accuracy	0.7468	0.7468	0.7468	0.7468
macro avg	0.7216	0.7156	0.7182	154
weighted avg	0.7438	0.7468	0.745	154

Figure 4.4 : Métriques pour SVM

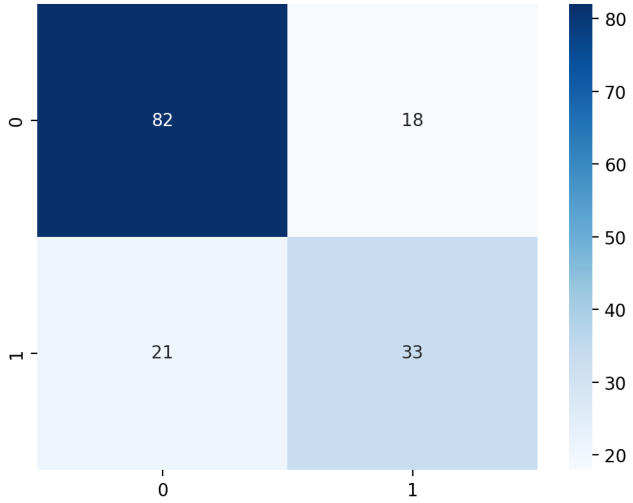

Figure 4.5 : Matrice de confusion de SVM

4.2.2 Modèles non supervisés

4.2.2.1 Clustering K-means

Le K-Means est un algorithme de partitionnement qui divise les données en k clusters, en minimisant la variance intra-cluster. Le modèle a été entrainé avec K=3 après une réduction de dimensions en 2 dimensions avec PCA. Le résultat est présenté dans la figure 4.6

4.2.2.2 DBSCAN

L'algorithme DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est basé sur la densité locale de points. Contrairement à K-Means, il ne nécessite pas de spécifier le nombre de clusters et peut détecter des points aberrants (outliers), ce qui est utile pour identifier des profils atypiques de patients. Les paramètres eps et min-samples ont été ajustés empiriquement après réduction en 2D via PCA. Le résultat est présenté dans la figure 4.7

DBSCAN a permis d'identifier un noyau dense de patients ayant des caractéristiques similaires (groupe homogène), mais aussi plusieurs points considérés comme bruit. Ces points pourraient correspondre à des cas limites ou à des individus présentant un profil clinique atypique.

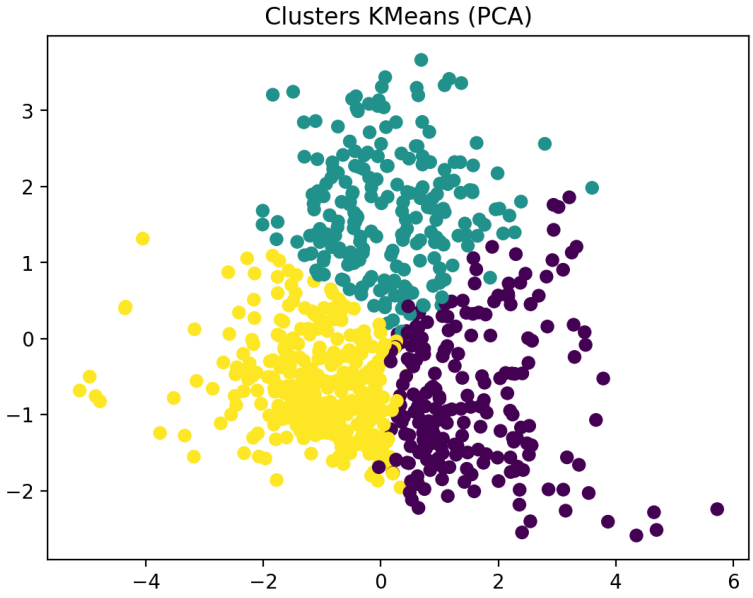

Figure 4.6 : Kmeans avec PCA

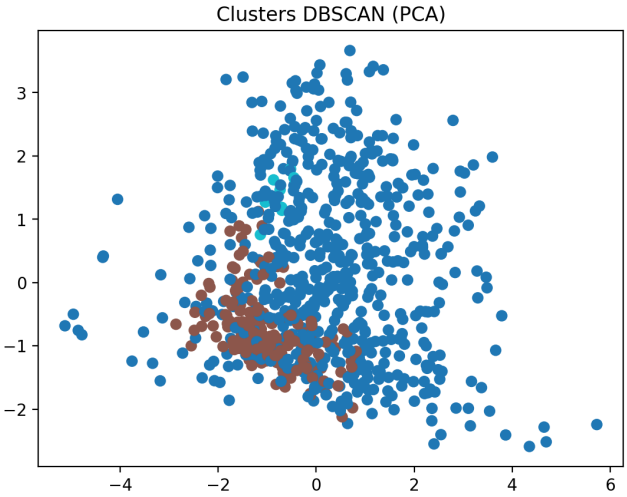

Figure 4.7 : DBSCAN avec PCA

4.3 Evaluation et comparaison des modèles

L'évaluation des modèles appliqués sur les données cliniques a permis de comparer leurs performances selon plusieurs critères : précision, rappel, F1-score et interprétabilité. Parmi les modèles supervisés, la forêt aléatoire s'est distinguée par sa robustesse et ses excellents scores globaux, faisant d'elle un candidat pertinent pour une application réelle (ayant le taux d'exactitude le plus élevé). Le SVM a également affiché de bonnes performances, en particulier pour séparer les cas limites, tandis que la régression logistique, bien que plus simple, a le taux de précision le plus faible parmi les trois. Du côté non supervisé, K-Means a permis de dégager des profils relativement cohérents avec les classes réelles, tandis que DBSCAN a révélé des sous-groupes denses et des anomalies potentiellement significatives. Cette comparaison met en lumière la complémentarité des approches testées, et oriente

vers une combinaison judicieuse entre performance prédictive et capacité explicative dans une optique de soutien au diagnostic médical.

Conclusion

Ce chapitre a présenté les différents modèles d'apprentissages utilisés lors de la phase de modélisation, avec justification de choix et évaluation de performances. Ces résultats sont suivis de l'étape d'évaluation. Le chapitre suivant exploitera la phase du déploiement avec la présentation de l'interface graphique.

Chapitre 5

Phase de déploiement

Plan

Introduction

Dans le cadre de la phase de déploiement, une interface utilisateur a été conçue pour regrouper l'ensemble des fonctionnalités développées au cours du projet. Cette interface vise à rendre l'analyse accessible, interactive et exploitable par un utilisateur non expert.

5.1 Interface graphique

Pour une facilité de manipulation et d'exploitation pour tous types d'utilisateurs, nous avons implémenté une interface graphique interactive en utilisant la bibliothèque Streamlit de Python. Cette interface, telle que présentée dans la figure 5.1 offre la possibilité d'importer un fichier de données sous format CSV, sinon d'utiliser le dataset par défaut sur lequel nous avons basé notre étude. Elle fournit également une navigation fluide en plusieurs sections.

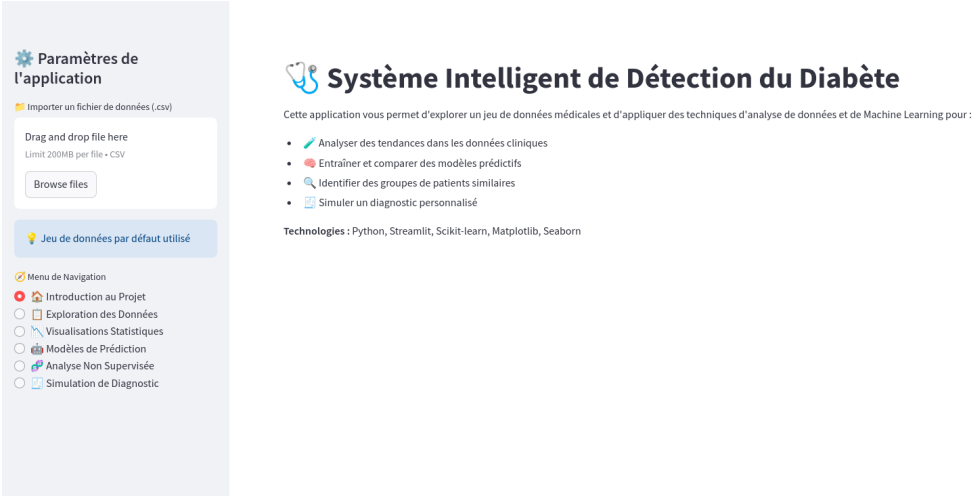

Figure 5.1 : Aperçu général de l'interface utilisateur

- **Exploration de données** : Cette section est dédiée à la compréhension du dataset et des différents variables avec des statistiques descriptives.
- **Visualisation de données** : Cette section est dédiée à l'exploration visuelle des données. Elle propose des graphiques interactifs tels que des histogrammes, des corrélations par heatmap, et des boxplots, permettant à l'utilisateur de comprendre rapidement la distribution des variables et la présence d'éventuelles anomalies.
- **Modèles de prédiction et d'analyse non supervisée** : Ces deux sections permettent de consulter les performances des modèles supervisés ainsi que les résultats des modèles non supervisés. L'interface est enrichie avec des graphiques et courbes utiles.

5.2 Simulation de Diagnostic

Une partie centrale de l'application est dédiée à la simulation de diagnostic. L'utilisateur peut saisir les valeurs des variables médicales via des sliders intuitifs. Ensuite, il choisit le modèle prédictif à utiliser (parmi ceux entraînés), et lance la prédiction.

Le résultat s'affiche instantanément sous forme d'un diagnostic estimé (présence ou non de diabète) accompagné d'une probabilité d'estimation et celle d'être diabète, ce qui renforce l'interprétabilité du résultat et la confiance de l'utilisateur.

Figure 5.2 : Aperçu de l'interface de simulation de diagnostic

Conclusion

Ce chapitre a présenté l'interface graphique implémentée, qui offre une solution complète, allant de l'analyse exploratoire jusqu'à la prédiction personnalisée. Elle constitue une passerelle concrète entre les résultats de la fouille de données et leur application réelle dans un cadre décisionnel ou médical.

Conclusion générale

Ce projet d'analyse et de fouille de données appliqué au diagnostic du diabète s'est appuyé sur la méthodologie CRISP-DM pour structurer les différentes phases de l'étude. À travers une analyse exploratoire rigoureuse, nous avons identifié les variables les plus influentes, traité les valeurs aberrantes, et appliqué des techniques de modélisation supervisée et non supervisée.

Les modèles prédictifs développés ont permis d'obtenir des résultats satisfaisants en termes de performance, mettant en évidence la pertinence des données utilisées. Le déploiement d'une interface interactive a permis de concrétiser l'application des résultats, en offrant à l'utilisateur une plateforme de visualisation, d'analyse et de simulation diagnostique.

Ainsi, ce travail démontre l'importance de la science des données dans l'aide à la décision médicale et ouvre la voie à des améliorations futures, notamment par l'intégration de données supplémentaires ou l'optimisation continue des modèles.