République Tunisienne
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Monastir
Institut Supérieur d'Informatique de Mahdia

Projet de fin d'études

RAPPORT

Présenté à

Institut Supérieur d'Informatique de Mahdia

Master Professionnel science de Données

Par

AICHA WAILI

PREDICTION DU DIABETE

Examinateur

Année universitaire 2023/2024

Table Des Matières

Liste Des Figures

1. Introduction

Le diabète est un problème de santé mondial croissant, affectant des millions de personnes et ayant des répercussions majeures sur les systèmes de santé. Ce rapport explore l'application des techniques de Data Mining pour la prédiction du diabète. Nous discuterons des méthodes utilisées, des résultats obtenus et de leurs implications.

1.1 Objectif principal

Ce projet vise à développer un modèle de prédiction du diabète utilisant des techniques de Data Mining. Le modèle sera développé et évalué à l'aide d'un ensemble de données de patients diabétiques et non diabétiques

1 .2 Connaissance de Domaine

Dans le contexte de l'ensemble de données sur le diabète, il est important de comprendre la signification des caractéristiques liées à la santé et le rôle qu'elles peuvent jouer dans la prédiction du début du diabète. Plongeons plus en profondeur dans ces caractéristiques :

Grossesses : Le nombre de grossesses précédentes peut être un indicateur important du risque de diabète chez les femmes.

Glucose : La concentration élevée de glucose dans le sang est un symptôme commun du diabète et peut indiquer un contrôle insuffisant de la glycémie.

BloodPressure : Une pression artérielle élevée est un facteur de risque de diabète et de complications associées.

SkinThickness : L'épaisseur du pli cutané est corrélée à l'obésité, qui est un facteur de risque majeur de diabète.

Insulin : Le niveau d'insuline sérique peut être perturbé chez les personnes atteintes de diabète de type 2.

BMI : Un indice de masse corporelle élevé est associé à un risque accru de diabète et de nombreuses autres maladies métaboliques.

DiabetesPedigreeFunction : Cette fonction évalue la probabilité de diabète en tenant compte de l'hérédité familiale.

Age : Le risque de diabète augmente avec l'âge, en partie en raison de changements métaboliques liés au vieillissement.

Outcome : Variable cible indiquant la présence ou l'absence de diabète, utilisée pour évaluer les performances des modèles de prédiction.

2. Méthode

Cette section présente une description des méthodes utilisées dans l'étude sur la prédiction du diabète, incluant la collecte des données, le prétraitement, la sélection des caractéristiques, l'utilisation d'algorithme de data mining k-NN et l'évaluation de modèle.

2.1 Collecte et prétraitement des données

2.1.1 Description de l'ensemble de données utilisé

L'ensemble de données utilisé dans cette étude provient de Kaggle, une plateforme en ligne populaire pour le partage de données et de projets liés à la science des données et à l'apprentissage automatique. Cet ensemble de données, comprenant 768 observations et 9 colonnes, inclut les caractéristiques suivantes : Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age et Outcome.

Source de données : Kaggle (https://www.kaggle.com/datasets/afrahalshrari/diabtes-data).

Dans la Figure 2.1, une vue sur l'ensemble de données est présentée, offrant un aperçu visuel des données utilisées dans cette étude.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 2.1 : Vue sur l'ensemble de données

2.1.2 Prétraitement des données

Gestion des valeurs manquantes : Les valeurs manquantes dans l'ensemble de données ont été traitées pour préserver l'intégrité des données et assurer la qualité des analyses à venir. Nous avons utilisé les approches suivantes pour gérer les valeurs manquantes spécifiques à certaines variables :

Glucose et BloodPressure : Les valeurs manquantes ont été remplacées par la moyenne des valeurs disponibles pour ces variables. L'imputation par la moyenne est une méthode simple qui permet de maintenir la distribution générale des données tout en comblant les lacunes.

SkinThickness, Insulin et BMI : Pour ces variables, les valeurs manquantes ont été remplacées par la médiane des valeurs disponibles. La médiane est une mesure de tendance centrale robuste aux valeurs extrêmes, ce qui en fait une option appropriée pour remplacer les valeurs manquantes dans ces cas.

Normalisation des données : La normalisation des variables a été entreprise pour uniformiser l'échelle de toutes les caractéristiques, une pratique essentielle pour de nombreux algorithmes de Data Mining. Cela vise à prévenir toute prédominance d'une caractéristique sur les autres en raison de leurs échelles respectives, ce qui pourrait altérer les résultats de l'analyse.

2.2 Analyse Exploratoire des Données (AED)

L'Analyse Exploratoire des Données (AED) est une méthode statistique qui explore les caractéristiques principales d'un ensemble de données sans recourir à des tests statistiques formels. Elle utilise des techniques telles que les statistiques descriptives et les visualisations graphiques pour comprendre la structure des données, identifier les tendances, les modèles et les relations entre les variables.

2.3 Ingénierie des caractéristiques

Pour l'ingénierie des caractéristiques, deux approches principales ont été utilisées : l'analyse de corrélation et l'évaluation de l'importance des caractéristiques.

2.3 .1 Matrice de Corrélation

Pour l'ingénierie des caractéristiques, différentes approches ont été utilisées pour analyser les corrélations dans l'ensemble de données :

Matrice de Corrélation Globale : Une analyse de corrélation a été effectuée pour évaluer les relations entre chaque paire de variables, indépendamment de leur lien avec le diabète. Cette approche a permis d'identifier les corrélations potentielles entre les caractéristiques, facilitant ainsi la détection des redondances et la réduction de la dimensionnalité de l'ensemble de données.

Matrice de Corrélation avec le Diabète : Une analyse spécifique de corrélation a été réalisée pour évaluer la relation de chaque caractéristique avec le diagnostic de diabète. Cette analyse a permis d'identifier les caractéristiques les plus fortement associées au diabète, ce qui est crucial pour la prédiction précise de cette maladie. En se concentrant sur ces caractéristiques, il était possible de développer un modèle plus ciblé et efficace pour la prédiction du diabète.

2.4 Développement des modèles

Pour le développement du modèle, nous avons exploré deux approches différentes : le clustering non supervisé avec l'algorithme de clustering, et l'apprentissage supervisé avec l'algorithme des k plus proches voisins (k-NN).

Classification avec k-NN : Parallèlement, l'algorithme des k plus proches voisins (k-NN) a été utilisé pour l'apprentissage supervisé. La performance du classificateur k-NN a été évaluée en utilisant différentes valeurs de k, et k = 11 a été sélectionné comme paramètre optimal pour le modèle. L'algorithme k-NN utilise la similarité entre les caractéristiques des échantillons pour prédire la classe d'un nouvel échantillon en se basant sur les k voisins les plus proches dans l'ensemble d'entraînement. Comme représente à la figure 2.2.

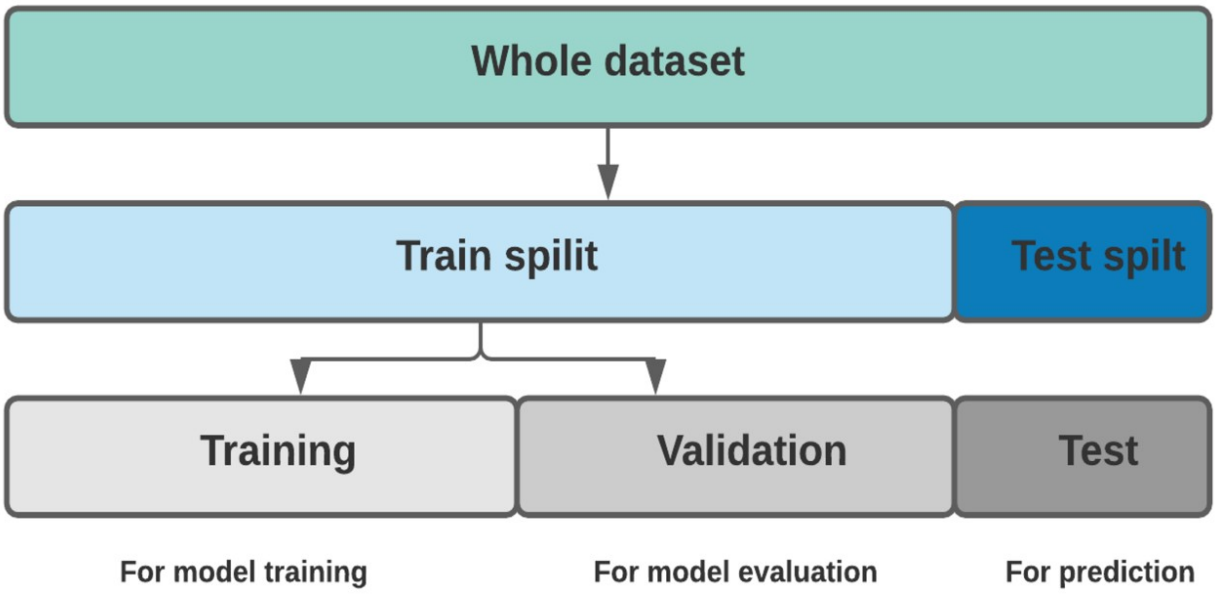

Figure 2.2 : K-NN

Dans la Figure 2.2 les variations des scores d'entraînement et de test en fonction de la valeur de k dans l'algorithme k-NN sont illustrées. Il est notable que le score de test atteint son niveau le plus élevé lorsque k est égal à 11.

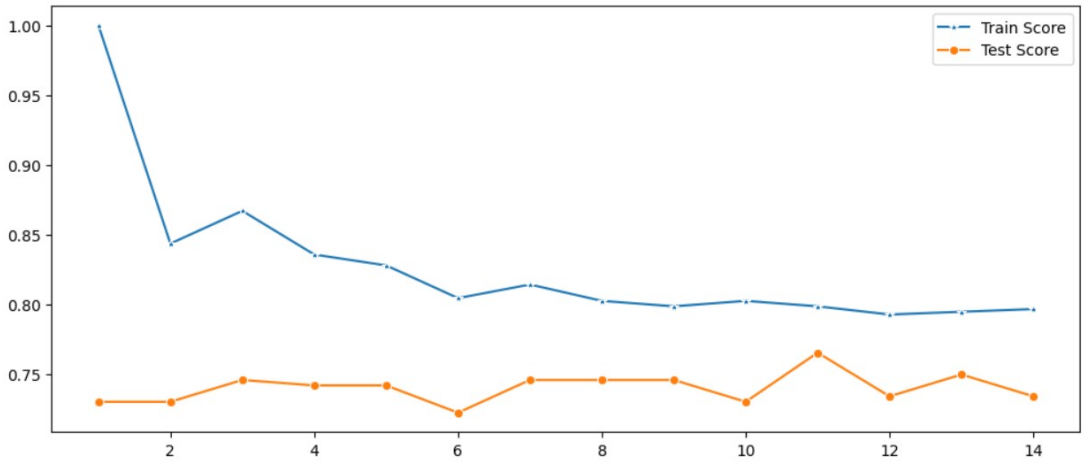

Figure 2.3 : Les variations des scores d'entraînement et de test

2.5 Évaluation des modèles

Pour évaluer les performances des modèles, différentes mesures ont été utilisées pour chaque approche :

2.5.1 Évaluation du k-NN :

Pour évaluer le modèle k-NN, plusieurs métriques ont été utilisées :

Matrice de confusion : Elle permet de visualiser les performances du modèle en comparant les prédictions du modèle avec les valeurs réelles. Cela fournit des informations sur les vrais positifs, les faux positifs, les vrais négatifs et les faux négatifs. Comme représente à la figure 2.4.

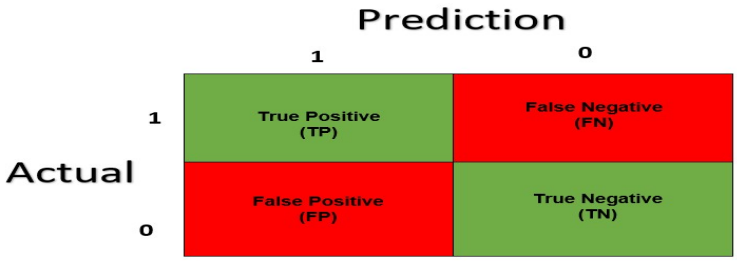

Figure 2.4 : Matrice de confusion

Rapport de classification : fournit une évaluation détaillée des performances du modèle pour chaque classe prédite. Voici une explication plus détaillée des mesures incluses dans ce rapport :

Précision (Precision) : La précision mesure la proportion d'instances prédites comme positives qui sont réellement positives. En d'autres termes, il s'agit du nombre de vrais positifs divisé par la somme des vrais positifs et des faux positifs. Une précision plus élevée indique un modèle qui produit moins de faux positifs.

Rappel (Recall) : Le rappel mesure la proportion d'instances réellement positives qui ont été correctement prédites par le modèle. Il est calculé en divisant le nombre de vrais positifs par la somme des vrais positifs et des faux négatifs. Un rappel plus élevé indique un modèle qui identifie un plus grand nombre de vrais positifs, réduisant ainsi les faux négatifs.

$$Recall = \frac{TP}{TP + FN}$$

Score F1 : Le score F1 est la moyenne harmonique de la précision et du rappel. Il combine ces deux mesures en un seul score qui tient compte à la fois des faux positifs et des faux négatifs. Le score F1 est calculé à l'aide de la formule :

$$F1\ score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

Un score F1 élevé indique à la fois une précision et un rappel élevés, ce qui est généralement souhaitable.

Exactitude (Accuracy) : L'exactitude mesure la proportion totale d'instances correctement classées par le modèle, qu'elles soient positives ou négatives. Elle est calculée en divisant le nombre total de prédictions correctes par le nombre total d'instances. Une exactitude élevée indique un modèle globalement performant, mais elle peut être trompeuse si les classes sont déséquilibrées.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Courbe ROC et AUC : La courbe ROC (Receiver Operating Characteristic) et l'aire sous la courbe (AUC) sont utilisées pour évaluer les performances du modèle en termes de sensibilité et de spécificité. Une valeur d'AUC proche de 1 indique de bonnes performances du modèle.

3. Résultats

La section des résultats expose les conclusions de l'analyse sur la prédiction du diabète à partir des méthodes de data mining. Elle présente les performances des modèles développés, incluant les résultats de classification avec k-NN, ainsi que les mesures d'évaluation telles que les matrices de confusion et la courbe ROC - AUC. Cette section offre un aperçu des conclusions tirées de l'étude.

3.1 Résultat de L'Analyse Exploratoire des Données (AED)

Dans cette phase d'Analyse Exploratoire des Données (AED), les résultats descriptifs ont été visualisés afin d'appréhender la distribution et les tendances des variables dans l'ensemble de données

3.1.1 Les statistiques descriptives pour les groupes diabète et non-diabète

Le graphique de la figure 3.1 montre que les personnes diabétiques ont des taux de glycémie moyens plus élevés que les personnes non diabétiques, avec un taux moyen de glycémie de 109,98 mg/dL pour les non diabétiques et de 141,26 mg/dL pour les diabétiques.

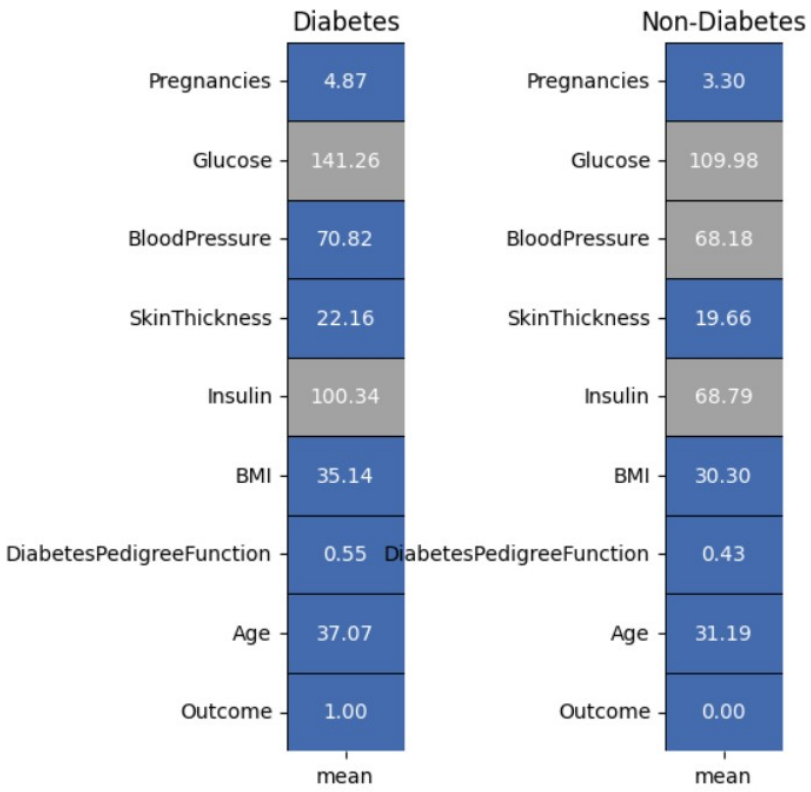

Figure 3.1 : les statistiques descriptives pour les groupes diabète et non-diabète

Les résultats de la variable "Outcome" montrent que parmi les personnes étudiées, 500 ne sont pas diabétiques (Outcome = 0) tandis que 268 sont diabétiques (Outcome = 1). Cela indique une proportion importante de personnes diabétiques dans l'échantillon analysé. Comme représente dans la figure 3.2.

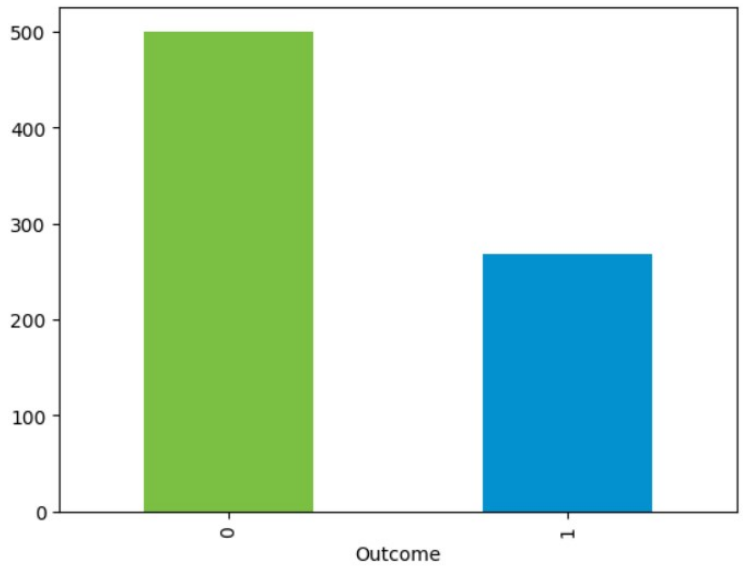

Figure 3.2 : Outcome

3.2 Résultat Ingénierie des caractéristiques

3.2.1 Matrice de Corrélation

En examinant les résultats de corrélation, différentes caractéristiques présentent des niveaux variés de corrélation avec le diabète. Comme représente dans La figure 3.3 et la figure 3.4

Matrice de Corrélation Globale :

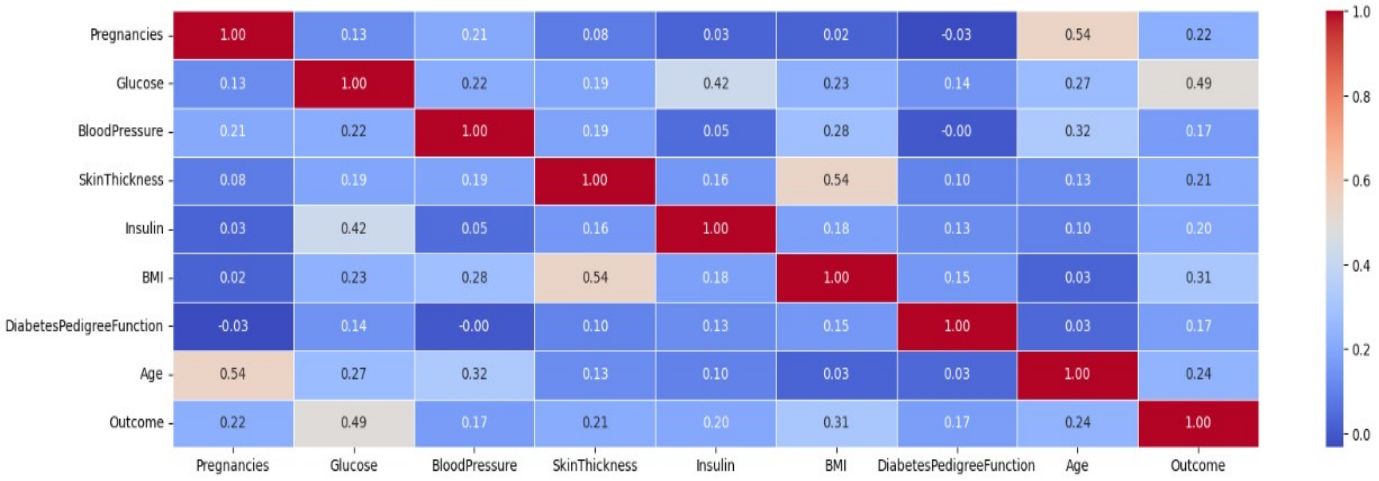

Figure 3.3 : Matrice de Corrélation Globale

Matrice de Corrélation avec le Diabète :

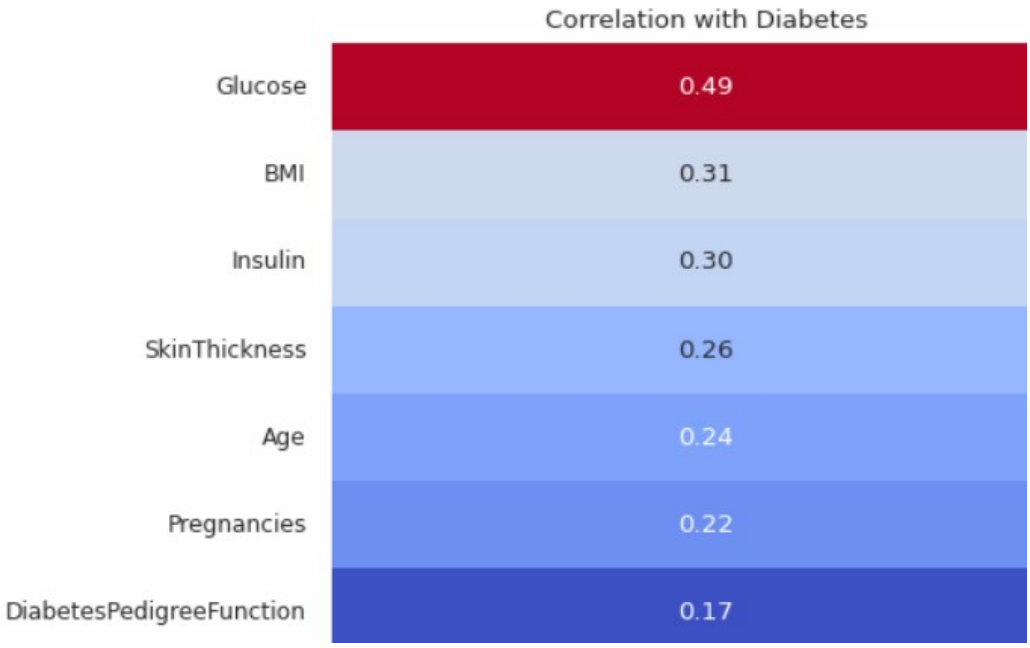

Figure 3.4 : Matrice de Corrélation avec diabète

3.3 Résultats des modèles développés

Dans la phase des résultats de modèle développé, les performances de modèle à été évaluée en termes de précision, de rappel et de score F1, avec l'analyse des matrices de confusion, des courbes ROC et des caractéristiques importantes pour la prédiction du diabète.

3.3.1 K-NN

Pour le K-NN, le score KNN obtenu est présenté dans la Figure 3.6.

```
[ ]  # Initialisation du classificateur KNN avec k=11
     knn = KNeighborsClassifier(11)
     # Entraînement du modèle sur les données d'entraînement
     knn.fit(X_train, y_train)
     # Évaluation de la précision du modèle sur les données de test
     knn.score(X_test,y_test)

     0.765625
```

Figure 3.6 : Score KNN

3.4 Résultats des performances de modelé

3.4.1 Résultats Évaluation du modèle k-NN :

Matrice de confusion :

La matrice de confusion fournit une évaluation détaillée des performances du modèle, indiquant le nombre de prédictions correctes et incorrectes pour chaque classe comme présente dans la

Figure 3.7.

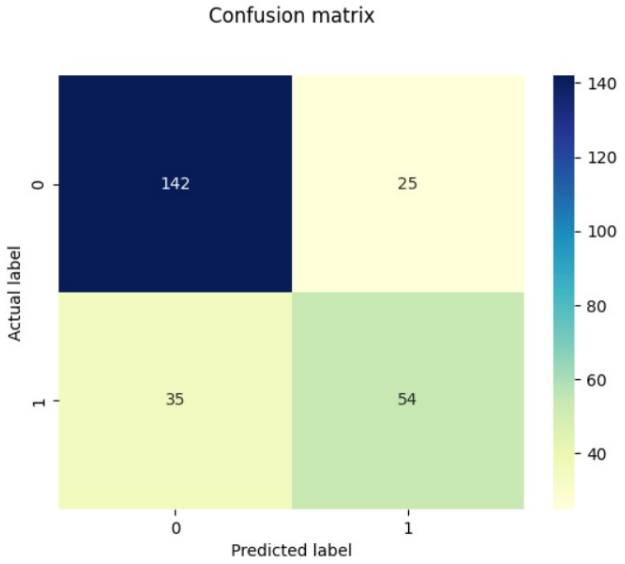

Figure 3.7 : Matrice de confusion

Rapport de classification :

Le rapport de classification fournit des mesures détaillées telles que la précision, le rappel et le score F1 pour chaque classe prédite par le modèle de classification. Ces résultats sont présentés dans la Figure 3.8.

	precision	recall	f1-score	support
0	0.80	0.85	0.83	167
1	0.68	0.61	0.64	89
accuracy			0.77	256
macro avg	0.74	0.73	0.73	256
weighted avg	0.76	0.77	0.76	256

Figure 3.8 : Rapport de classification.

Courbe ROC et AUC : La courbe ROC et l'AUC (Area Under the Curve) sont des outils d'évaluation essentiels pour mesurer les performances d'un modèle de classification. Ces résultats sont présentés dans la Figure 3.9.

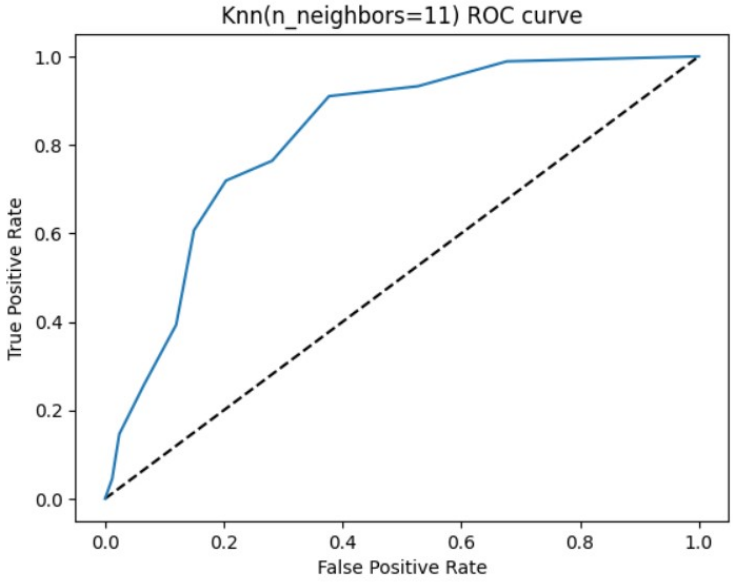

Figure 3.9 : Courbe ROC et AUC

4. Discussion

Dans la phase de discussion, le résultat de modèle et des analyses sont interprétés, en mettant en évidence les observations importantes, les limitations et les implications pratiques recherches.

4.1 Importance de caractéristiques

Ces données mettent en évidence l'importance cruciale de certaines caractéristiques dans la prédiction du diabète. Par exemple, les niveaux moyens de **glucose**, **d'insuline** et **d'IMC (BMI)** sont nettement plus élevés chez les individus atteints de diabète, soulignant leur rôle majeur en tant que facteurs de risque clés pour cette maladie.

De plus, l'âge moyen des individus diabétiques est légèrement supérieur à celui des non-diabétiques, suggérant que l'âge peut également influencer le risque de diabète. Cette observation souligne l'importance de surveiller attentivement ces caractéristiques pour identifier les personnes susceptibles de développer le diabète.

En comprenant pleinement l'impact de ces caractéristiques sur le risque de diabète, les professionnels de la santé peuvent mieux cibler leurs efforts préventifs et fournir des interventions adaptées pour réduire la prévalence de cette maladie chronique.

4.2 k-NN

Les résultats de la matrice de confusion, du rapport de classification et de la courbe ROC fournissent une évaluation détaillée des performances de votre modèle KNN dans la prédiction du diabète.

Matrice de confusion :

Le modèle a correctement prédit 142 des 167 cas de non-diabète (classe 0), mais a mal classé 25 comme diabétiques.

Pour les cas de diabète (classe 1), le modèle a correctement prédit 54 des 89 cas, mais a manqué 35 cas en les classant à tort comme non-diabétiques.

Rapport de classification :

La précision pour la classe 0 est de 80%, ce qui signifie que 80% des cas prédits comme non-diabète l'étaient effectivement.

Le rappel pour la classe 0 est de 85%, ce qui indique que le modèle a correctement identifié 85% des cas réels de non-diabète.

Pour la classe 1, la précision est de 68%, ce qui signifie que 68% des cas prédits comme diabétiques le sont réellement.

Le rappel pour la classe 1 est de 61%, ce qui montre que le modèle a identifié correctement 61% des cas réels de diabète.

Le score F1, qui combine précision et rappel, est de 0.83 pour la classe 0 et de 0.64 pour la classe 1.

Courbe ROC :

Le taux de faux positifs est proche de zéro pour toutes les valeurs du taux de vrais positifs, ce qui indique une capacité du modèle à bien discriminer entre les deux classes.

En conclusion, le modèle KNN présente des performances globalement acceptables, avec une précision et un rappel raisonnable pour les deux classes. Cependant, il existe une marge d'amélioration, notamment en réduisant le nombre de faux négatifs pour la classe 1, c'est-à-dire en identifiant plus efficacement les cas de diabète.

5. Conclusion

En conclusion, cette étude a exploré deux méthodes, à savoir l'algorithme KNN pour la classification dans le contexte de la prédiction du diabète. Les résultats ont révélé des performances acceptables pour le KNN en termes de précision et de rappel. Cette approche offre des possibilités prometteuses pour la prédiction du diabète.