EFREI

Projet de Machine Learning

Classification du Diabète à l'aide du Machine Learning

Amine M'ZALI

26 Novembre 2024

Table des matières

Introduction **1**

1 Analyse des Données **2**
1.1 Présentation du Dataset 2
1.2 Analyse Exploratoire des Données (EDA) 2
1.2.1 Distribution de la Variable Cible 2
1.2.2 Analyse des Variables Indépendantes 3
1.3 Prétraitement des Données 4
1.3.1 Gestion des Valeurs Manquantes 4
1.3.2 Standardisation des Données 4
1.3.3 Vérification des Valeurs Aberrantes 4
1.3.4 Analyse des Corrélations 5

2 Méthodologie **7**
2.1 Séparation des Données 7
2.2 Modèles de Classification 7
2.2.1 Régression Logistique 7
2.2.2 Arbre de Décision 7
2.2.3 Réseau de Neurones 7
2.3 Optimisation des Hyperparamètres 7
2.3.1 Régression Logistique 8
2.3.2 Arbre de Décision 8
2.3.3 Réseau de Neurones 8

3 Résultats **9**
3.1 Performances des Modèles Avant Optimisation 9
3.2 Performances des Modèles Après Optimisation 9
3.3 Matrice de Confusion 10
3.4 Courbes ROC 10
3.5 Importance des Caractéristiques 11

4 Discussion **12**
4.1 Analyse des Résultats 12
4.2 Interprétation des Caractéristiques Importantes 12
4.3 Limites du Projet 12
4.4 Perspectives d'Amélioration 12

5 Conclusion **14**

Références **15**

A Annexes **16**
A.1 Code Source 16
A.2 Graphiques Supplémentaires 17

Table des figures

1.1 Distribution de la variable cible 3
1.2 Histogrammes des variables indépendantes 3
1.3 Boxplots des variables indépendantes 5
1.4 Matrice de corrélation 5

3.1 Matrice de confusion pour le Réseau de Neurones Optimisé 10
3.2 Courbes ROC pour les modèles optimisés 10
3.3 Importance des caractéristiques selon l'Arbre de Décision 11

A.1 Matrice de corrélation détaillée 17

Liste des tableaux

1.1 Description des variables du dataset 2
1.2 Nombre de valeurs manquantes par variable 4

3.1 Performances des modèles avant optimisation 9
3.2 Performances des modèles après optimisation 9

Résumé

Ce projet a pour objectif de prédire si un patient est diabétique en utilisant des techniques de machine learning sur le dataset Pima Indians Diabetes Database. Plusieurs modèles de classification ont été explorés, notamment la Régression Logistique, l'Arbre de Décision et le Réseau de Neurones. Une analyse approfondie des données a été réalisée, incluant le prétraitement, l'analyse exploratoire et l'optimisation des hyperparamètres. Les résultats montrent que le Réseau de Neurones optimisé offre les meilleures performances globales, avec un rappel élevé, ce qui est crucial pour la détection des patients diabétiques.

Introduction

Contexte

Le diabète est une maladie chronique qui affecte des millions de personnes dans le monde. La détection précoce du diabète est essentielle pour prévenir les complications graves et améliorer la qualité de vie des patients. Avec l'avènement du machine learning, il est possible de développer des modèles prédictifs efficaces pour identifier les patients à risque en se basant sur des mesures médicales.

Objectifs du Projet

Ce projet vise à :

- Analyser le dataset Pima Indians Diabetes Database pour comprendre les relations entre les variables.
- Prétraiter les données, notamment en gérant les valeurs manquantes et en standardisant les variables.
- Entraîner et comparer plusieurs modèles de classification : Régression Logistique, Arbre de Décision et Réseau de Neurones.
- Optimiser les hyperparamètres des modèles pour améliorer leurs performances.
- Évaluer les modèles en utilisant des métriques appropriées et sélectionner le meilleur modèle pour la prédiction du diabète.

Organisation du Rapport

Le rapport est structuré comme suit :

- **Chapitre 1** : Analyse des Données.
- **Chapitre 2** : Méthodologie.
- **Chapitre 3** : Résultats.
- **Chapitre 4** : Discussion.
- **Chapitre 5** : Conclusion.
- **Références**.

Chapitre 1

Analyse des Données

1.1 Présentation du Dataset

Le dataset utilisé est le *Pima Indians Diabetes Database*, disponible sur Kaggle. Il comprend 768 observations et 9 variables :

Variable	**Description**
Pregnancies	Nombre de grossesses
Glucose	Concentration de glucose plasmatique
BloodPressure	Pression artérielle diastolique (mm Hg)
SkinThickness	Épaisseur du pli cutané triceps (mm)
Insulin	Insuline sérique (mu U/ml)
BMI	Indice de masse corporelle (kg/m^2)
DiabetesPedigreeFunction	Fonction héréditaire du diabète
Age	Âge (années)
Outcome	Variable cible (0 : non diabétique, 1 : diabétique)

TABLE 1.1 – Description des variables du dataset

1.2 Analyse Exploratoire des Données (EDA)

Une analyse exploratoire a été effectuée pour comprendre la distribution des variables et leurs relations.

1.2.1 Distribution de la Variable Cible

La variable cible est déséquilibrée :
— 500 patients non diabétiques (65%).
— 268 patients diabétiques (35%).

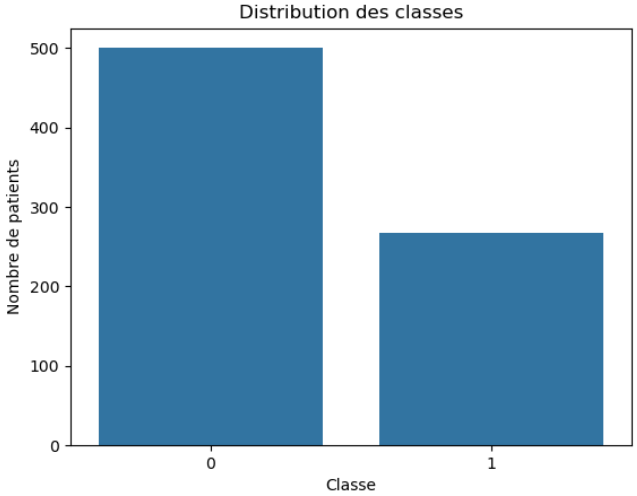

FIGURE 1.1 – Distribution de la variable cible

1.2.2 Analyse des Variables Indépendantes

Les histogrammes des variables indépendantes montrent que certaines variables, comme *Insulin* et *SkinThickness*, ont une distribution asymétrique avec des valeurs nulles improbables.

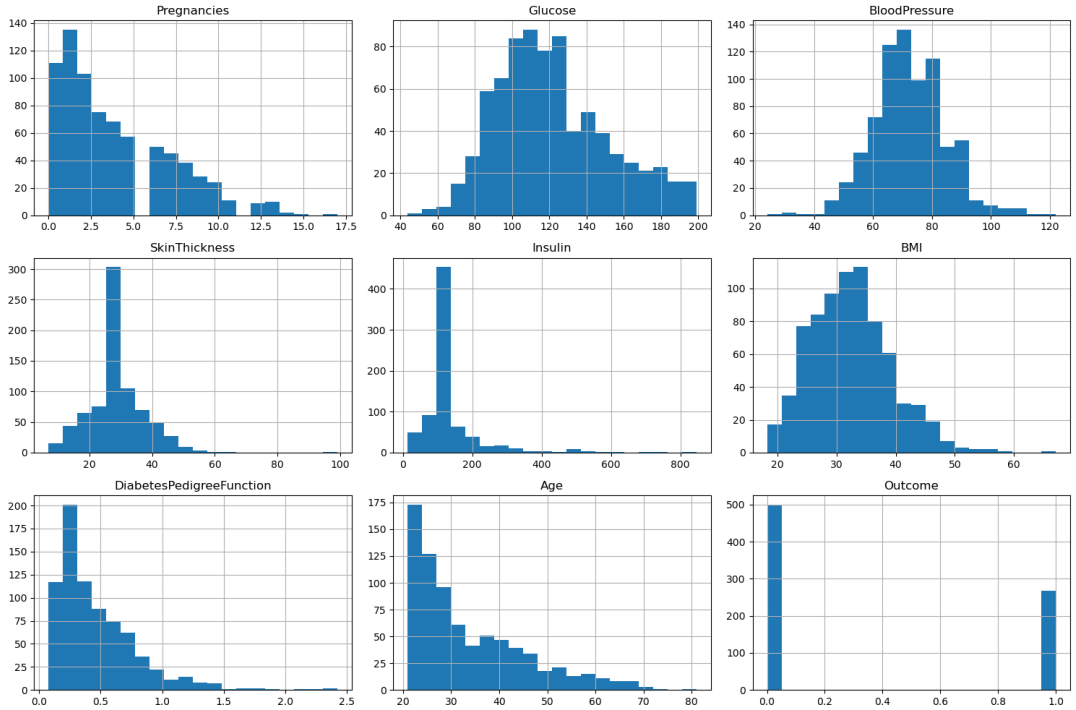

FIGURE 1.2 – Histogrammes des variables indépendantes

1.3 Prétraitement des Données

Pour améliorer la qualité des données, plusieurs étapes de prétraitement ont été réalisées.

1.3.1 Gestion des Valeurs Manquantes

Les valeurs nulles (0) dans certaines variables médicalement impossibles ont été remplacées par *NaN* :

— Colonnes concernées : *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, *BMI*.

Le nombre de valeurs manquantes après remplacement :

Variable	**Valeurs manquantes**
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374
BMI	11

TABLE 1.2 – Nombre de valeurs manquantes par variable

Les valeurs manquantes ont été imputées avec la médiane de chaque variable pour éviter de biaiser les données.

1.3.2 Standardisation des Données

Les variables ont été standardisées en utilisant la méthode *StandardScaler* pour faciliter l'apprentissage des modèles, en particulier pour le Réseau de Neurones.

1.3.3 Vérification des Valeurs Aberrantes

Des boxplots ont été générés pour détecter les outliers.

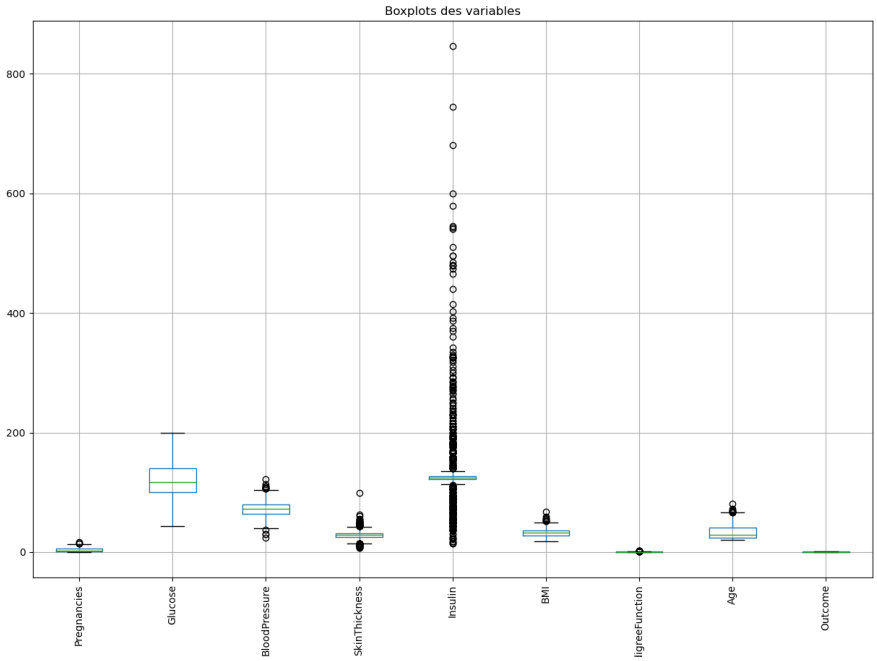

FIGURE 1.3 – Boxplots des variables indépendantes

1.3.4 Analyse des Corrélations

Une matrice de corrélation a été établie :

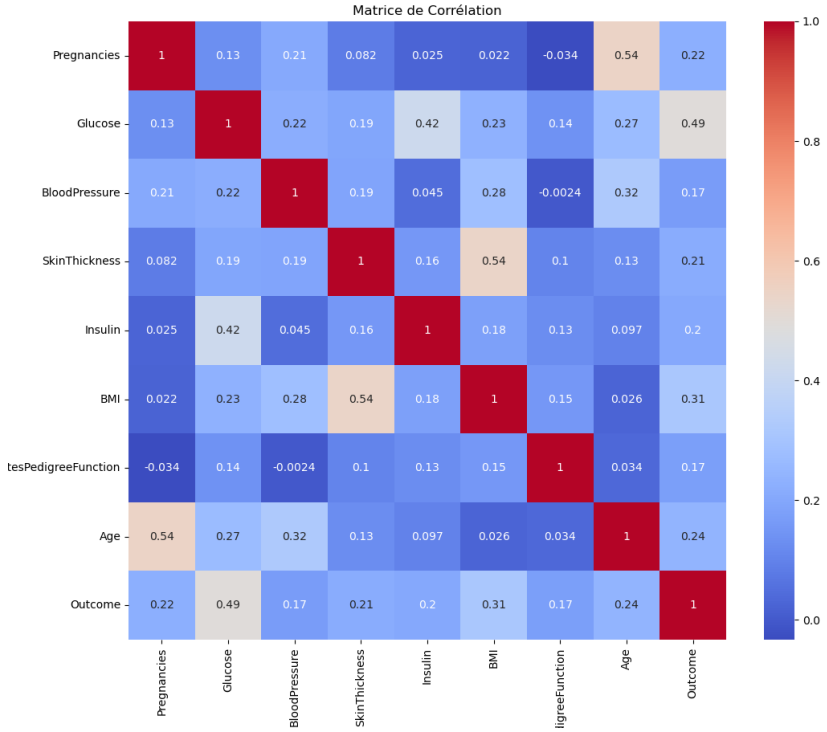

FIGURE 1.4 – Matrice de corrélation

Observations :

— *Glucose* a une forte corrélation positive avec l'issue du diabète.

— *BMI* et *Age* présentent également des corrélations significatives.

— Faible corrélation entre les variables indépendantes, réduisant le risque de multicolinéarité.

Chapitre 2

Méthodologie

2.1 Séparation des Données

Les données ont été divisées en ensembles d'entraînement (80%) et de test (20%) en utilisant une stratification pour maintenir la proportion des classes.

2.2 Modèles de Classification

Plusieurs modèles de machine learning ont été utilisés :

2.2.1 Régression Logistique

Un modèle de Régression Logistique a été choisi pour sa simplicité et son interprétabilité.

2.2.2 Arbre de Décision

L'Arbre de Décision est robuste aux valeurs aberrantes et permet d'identifier les caractéristiques les plus importantes.

2.2.3 Réseau de Neurones

Un Réseau de Neurones Multi-Couches (MLP) a été utilisé pour capturer les relations non linéaires entre les variables.

2.3 Optimisation des Hyperparamètres

Une recherche en grille (*GridSearchCV*) a été effectuée pour optimiser les hyperparamètres de chaque modèle.

2.3.1 Régression Logistique

Paramètres optimisés :
— *C* : [0.01, 0.1, 1, 10]
— *solver* : ['lbfgs', 'liblinear']

2.3.2 Arbre de Décision

Paramètres optimisés :
— *max_depth* : [None, 5, 10, 15]
— *min_samples_split* : [2, 5, 10]

2.3.3 Réseau de Neurones

Paramètres optimisés :
— *hidden_layer_sizes* : [(50,), (100,), (100,50)]
— *activation* : ['relu', 'tanh']
— *solver* : ['adam', 'sgd']

Chapitre 3

Résultats

3.1 Performances des Modèles Avant Optimisation

Modèle	Accuracy	Précision	Rappel	F1-Score
Régression Logistique	0.7013	0.5870	0.5000	0.5400
Arbre de Décision	0.6818	0.5532	0.4815	0.5149
Réseau de Neurones	**0.7468**	**0.6744**	**0.5370**	**0.5979**

TABLE 3.1 – Performances des modèles avant optimisation

Analyse :
Le Réseau de Neurones offre les meilleures performances globales avant optimisation.

3.2 Performances des Modèles Après Optimisation

Modèle Optimisé	Accuracy	Précision	Rappel	F1-Score
Régression Logistique	0.7013	0.5833	0.5185	0.5490
Arbre de Décision	0.6818	0.5532	0.4815	0.5149
Réseau de Neurones	**0.7403**	**0.6346**	**0.6111**	**0.6226**

TABLE 3.2 – Performances des modèles après optimisation

Analyse :
Après optimisation, le Réseau de Neurones montre une amélioration notable du rappel et du F1-score.

3.3 Matrice de Confusion

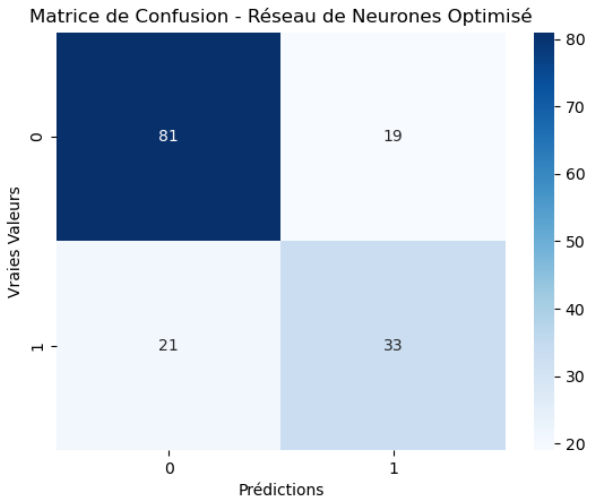

FIGURE 3.1 – Matrice de confusion pour le Réseau de Neurones Optimisé

Observation :
Le modèle détecte correctement 61% des patients diabétiques.

3.4 Courbes ROC

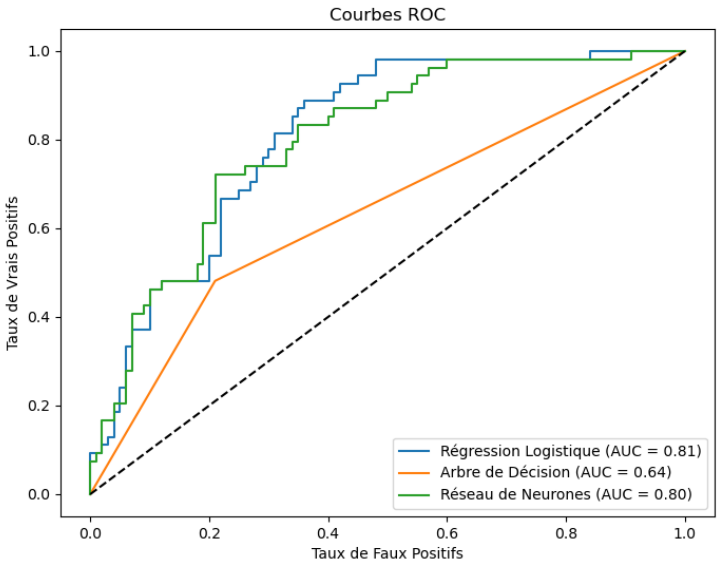

FIGURE 3.2 – Courbes ROC pour les modèles optimisés

Analyse :
La Régression Logistique a la plus grande aire sous la courbe (AUC), suivie de près par le Réseau de Neurones.

3.5 Importance des Caractéristiques

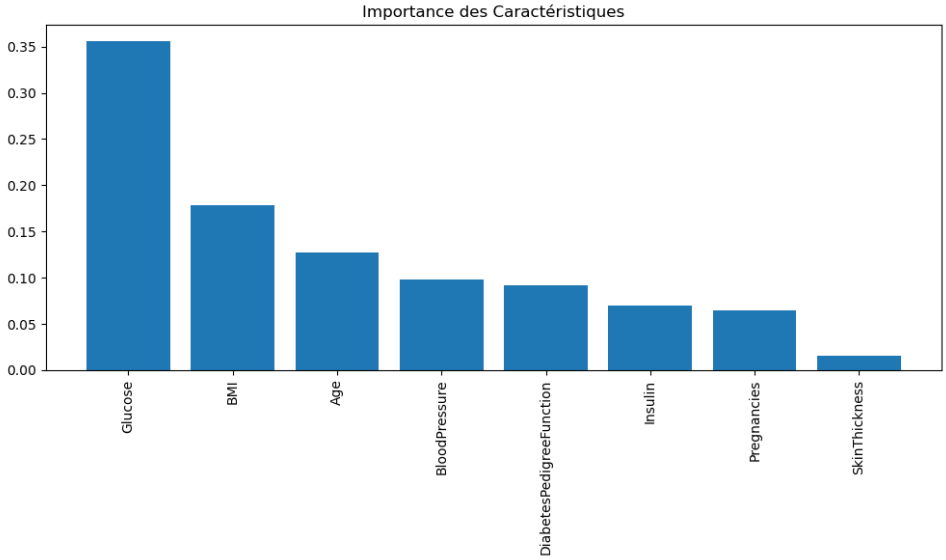

FIGURE 3.3 – Importance des caractéristiques selon l'Arbre de Décision

Observation :

Le *Glucose* est la caractéristique la plus importante, suivi du *BMI* et de l'*Age*.

Chapitre 4

Discussion

4.1 Analyse des Résultats

— Le Réseau de Neurones optimisé offre les meilleures performances globales, avec un rappel de 61%, ce qui est crucial pour la détection des patients diabétiques.
— La Régression Logistique a la meilleure AUC, indiquant une bonne capacité à distinguer les classes, mais son rappel est inférieur.
— L'Arbre de Décision n'a pas montré d'amélioration significative après optimisation.

4.2 Interprétation des Caractéristiques Importantes

Les caractéristiques les plus influentes sont :
— *Glucose* : Forte corrélation avec le diabète.
— *BMI* : Un indice de masse corporelle élevé est associé à un risque accru de diabète.
— *Age* : Le risque de diabète augmente avec l'âge.

4.3 Limites du Projet

— **Déséquilibre des Classes** : Le dataset est déséquilibré, ce qui peut affecter les performances des modèles.
— **Taille du Dataset** : Un dataset de 768 observations peut limiter la capacité du modèle à généraliser.
— **Valité Externe** : Les résultats peuvent ne pas être généralisables à d'autres populations.

4.4 Perspectives d'Amélioration

— **Équilibrage des Classes** : Utiliser des techniques comme SMOTE pour équilibrer le dataset.

— **Ensembles de Modèles** : Explorer des modèles comme Random Forest ou XGBoost.
— **Collecte de Données Supplémentaires** : Obtenir plus de données pour améliorer la robustesse du modèle.

Chapitre 5

Conclusion

Ce projet a permis de comparer plusieurs modèles de classification pour la prédiction du diabète. Le Réseau de Neurones optimisé s'est avéré être le plus performant, offrant un bon équilibre entre précision et rappel. Les caractéristiques les plus influentes identifiées sont le *Glucose*, le *BMI* et l'*Age*. Des améliorations peuvent être apportées en traitant le déséquilibre des classes et en explorant d'autres modèles.

Références

— Scikit-learn Documentation : `https://scikit-learn.org/`
— Dataset Pima Indians Diabetes Database : `https://www.kaggle.com/uciml/pima-indians-dia`

Annexe A

Annexes

A.1 Code Source

Listing A.1 – Prétraitement des données

```
# Remplacer les z ros par NaN dans les colonnes sp cifiques
colonnes_a_corriger = ['Glucose', 'BloodPressure', 'SkinThickness', '
    Insulin', 'BMI']
df[colonnes_a_corriger] = df[colonnes_a_corriger].replace(0, np.nan)

# Imputation des valeurs manquantes avec la m diane
df[colonnes_a_corriger] = df[colonnes_a_corriger].fillna(df[
    colonnes_a_corriger].median())
```

Listing A.2 – Entraînement du Réseau de Neurones

```
# Instanciation du mod le avec les meilleurs hyperparam tres
mlp = MLPClassifier(hidden_layer_sizes=(100, 50), activation='tanh',
    solver='adam', max_iter=1000, random_state=42)

# Entra nement du mod le
mlp.fit(X_train, y_train)

# Pr dictions sur l'ensemble de test
y_pred_mlp = mlp.predict(X_test)
```

A.2 Graphiques Supplémentaires

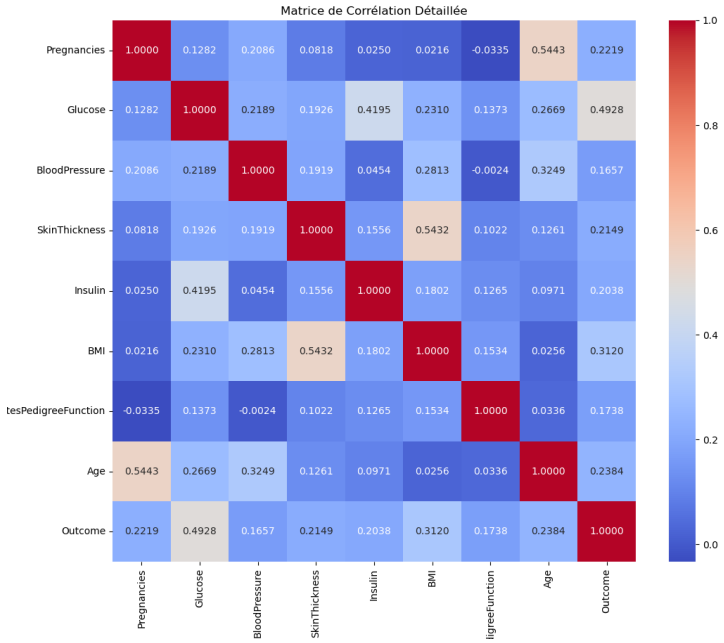

FIGURE A.1 – Matrice de corrélation détaillée