Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes - RABAT

Rapport de Projet :
Ingénierie des données

Réalisé par :

Mohamed Amine TAIF
Ayoub KOUSSY

Encadré par :

Pr Mohamed LAZAAR

Année académique 2022/2023

Résumé

Ce rapport synthétise notre travail effectué dans le cadre d' un projet académique Dans ce contexte, nous nous intéressons à la prédiction des diagnostics de cancer du sein en utilisant la data set de Breast Cancer Wisconsin (Diagnostic). Le but de ce projet est de prédire les diagnostics malignant ou bien bénignes en utilisant la régression logistique. Cependant, pour améliorer les performances de notre modèle, nous avons décidé d'utiliser des techniques de réduction de dimensionnalité telles que PCA et la réduction de la densité de Fisher (LDA). Ces techniques nous permettent de minimiser la complexité des données et d'augmenter les performances de notre modèle. En somme, l'objectif de ce projet est de mettre en œuvre une méthode de prédiction efficace pour les diagnostics de cancer du sein en utilisant des techniques de réduction de dimensionnalité.

Abstract

This report summarizes our work carried out as part of an academic project. In this context, we are interested in predicting breast cancer diagnoses using the Breast Cancer Wisconsin (Diagnostic) dataset. The goal of this project is to predict malignant or benign diagnoses using logistic regression. However, to improve the performance of our model, we decided to use dimensionality reduction techniques such as PCA and Fisher density reduction (LDA). These techniques allow us to minimize data complexity and increase the performance of our model. In summary, the objective of this project is to implement an effective prediction method for breast cancer diagnoses using dimensionality reduction techniques.

Liste des abréviations

BDD	Base De Données
ACP	Analyse en composantes principales
LDA	Linear Discriminant Analysis

Table des figures

Table des matières

Introduction

Ce Projet a pour objectif mettre en œuvre les concepts clés de l'ingénierie des données et de décrire les principales techniques et outils utilisés dans ce domaine.

Le contexte de ce rapport est le rôle que joue l'ingénierie des données dans la collecte, la gestion et l'analyse de données pour en tirer des informations précieuses afin de prendre des décisions éclairées et maximiser la valeur extraite des données.

Nous présenterons quelques exemples de cas d'utilisation de l'ingénierie des données dans différents cas et examinerons les défis et opportunités liés à sa mise en œuvre.

Chapitre 1

jeu de données

Le jeu de données est un élément central de l'ingénierie des données, car il constitue la base sur laquelle repose toute analyse. Dans ce chapitre, nous allons examiner les différents aspects du jeu de données utilisé dans ce rapport, notamment sa provenance, sa structure et sa qualité. Nous présenterons également les étapes de visualisation et de nettoyage des données qui ont été mises en œuvre pour préparer le jeu de données à l'analyse.

1.1 Base des données

La description des bases de données est une étape cruciale de l'ingénierie des données qui consiste à comprendre les caractéristiques et les relations entre les différentes variables du jeu de données. Dans ce chapitre, nous allons décrire la base de données utilisée dans ce rapport, ainsi que son contenu et son structure.

Le jeu de données utilisé s'appel Breast Cancer Wisconsin (Diagnostic) Data Set , est un ensemble de données qui contient des informations sur des caractéristiques de cellules cancéreuses de sein. Les données ont été collectées en utilisant une technique d'analyse appelée Fine Needle Aspirate (FNA) qui consiste à aspirer des cellules à l'aide d'une aiguille fine. Les caractéristiques calculées à partir de ces images numérisées incluent des informations sur les noyaux cellulaires tels que le rayon, la texture, la zone, la symétrie et la dimension fractale. Les données sont organisées avec un numéro d'identification et un diagnostic (M = malin, B = bénin) . Il y a 30 caractéristiques différentes dans les données, qui ont été calculées en utilisant des moyennes, des erreurs standard et des valeurs maximales. Les données sont disponibles sur le serveur ftp de UW CS et sur le dépôt UCI Machine Learning Repository. Il n'y a pas de valeurs manquantes dans les données et la répartition des classes est de 357 tumeurs bénignes et 212 malignes.

La base de données sur le cancer du sein de Wisconsin (Diagnostic) contient des informations sur plus de 500 individus et comprend 32 colonnes. 30 de ces colonnes contiennent des valeurs numériques qui décrivent les caractéristiques des cellules cancéreuses de sein, tandis qu'une colonne contient des valeurs de chaîne pour indiquer le diagnostic (M = malin, B = bénin). Il y a également une colonne supplémentaire pour l'ID de chaque individu.

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919
std	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803
min	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000
25%	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310
50%	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500
75%	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000
max	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200

	symmetry_mean	texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	0.181162	25.677223	107.261213	880.583128	0.132369	0.254265	0.272188
std	0.027414	6.146258	33.602542	569.356993	0.022832	0.157336	0.208624
min	0.106000	12.020000	50.410000	185.200000	0.071170	0.027290	0.000000
25%	0.161900	21.080000	84.110000	515.300000	0.116600	0.147200	0.114500
50%	0.179200	25.410000	97.660000	686.500000	0.131300	0.211900	0.226700
75%	0.195700	29.720000	125.400000	1084.000000	0.146000	0.339100	0.382900
max	0.304000	49.540000	251.200000	4254.000000	0.222600	1.058000	1.252000

FIGURE 1.1 – Description des Colonnes de BDD

1.2 Visualisation des données

La visualisation des données est un outil puissant pour comprendre et communiquer les insights tirés des données. Dans ce chapitre, nous allons présenter les différentes techniques de visualisation utilisées dans ce rapport et comment elles ont aidé à mettre en évidence les tendances et les patterns dans les données. Nous décrirons également comment ces visualisations ont été utilisées pour communiquer les résultats de l'analyse de manière claire et concise.

En observant les quatre boîtes à moustaches représentant les colonnes radius mean, texture mean, perimeter mean et area mean, on peut remarquer que la boîte à moustaches correspondant à la colonne area mean est nettement plus grande que les autres boîtes à moustaches. Cela indique que les valeurs de la colonne area mean sont généralement plus élevées que les valeurs des autres colonnes. En comparant les colonnes radius mean et texture mean, on peut remarquer que la boîte à moustaches correspondant à la colonne radius mean est également plus grande que celle correspondant à la colonne texture mean. Il est également possible de remarquer que les boîtes à moustaches sont symétriques par rapport à leur median. Cela signifie que la distribution des données est plutôt régulière, sans grande asymétrie. Cependant, on peut observer plusieurs valeurs atypiques dans les boîtes à moustaches, indiquant la présence de quelques valeurs extrêmes qui peuvent influencer la moyenne et l'écart-type.

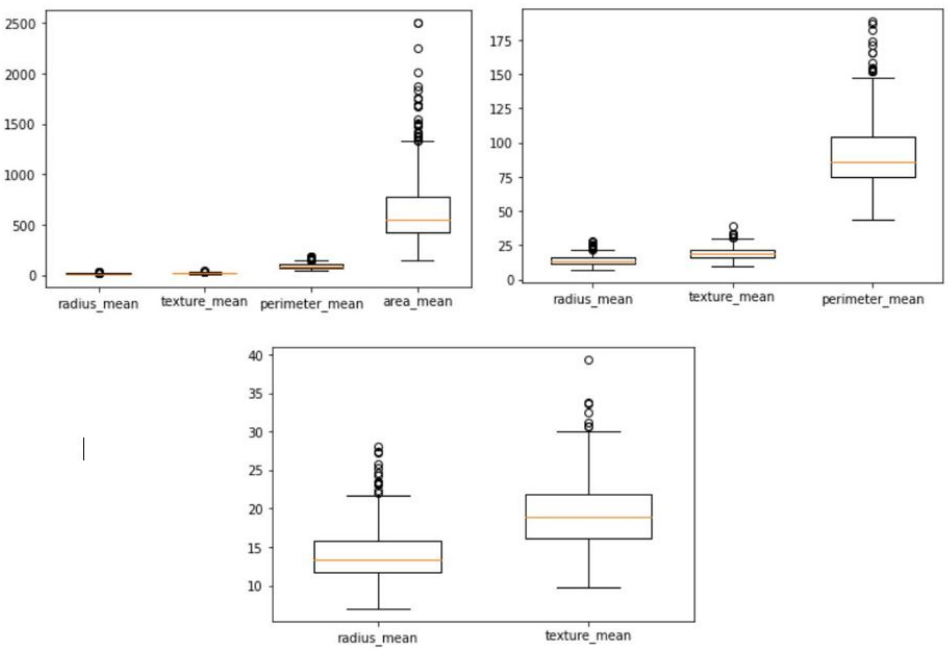

FIGURE 1.2 – Boîte à moustaches

En examinant les données, nous avons remarqué que le nombre d'observations dans les deux classes (malignes et bénignes) était presque équivalent. Cela est avantageux pour notre modèle de classification car cela signifie que les données ne sont pas déséquilibrées. Un déséquilibre des données peut entraîner un biais dans les résultats de la classification, car le modèle peut être plus enclin à prédire la classe majoritaire. En ayant un nombre équivalent d'observations dans les deux classes, nous pouvons être plus confiants que notre modèle sera capable de classer les données de manière équitable.

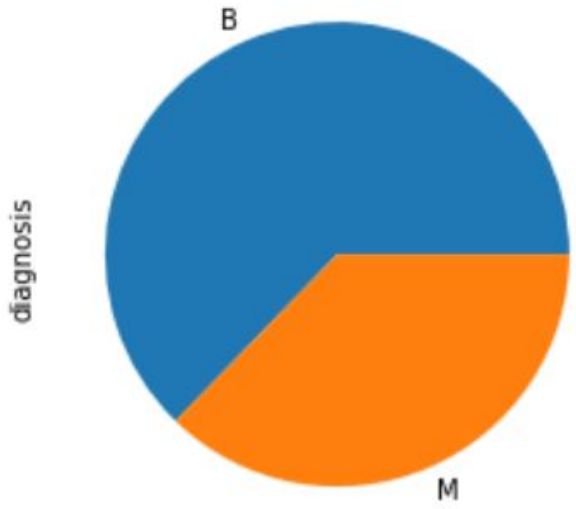

FIGURE 1.3 – Distrubition des class

1.3 Nettoyage des données

Le nettoyage des données est une étape cruciale de l'ingénierie des données qui consiste à préparer le jeu de données à l'analyse en corrigeant les erreurs et en supprimant les données redondantes ou obsolètes. Dans ce chapitre, nous allons décrire les différentes étapes de nettoyage des données mises en œuvre dans ce rapport et comment elles ont contribué à améliorer la qualité et l'exactitude des données utilisées dans l'analyse. Nous présenterons également les défis et les leçons apprises lors du processus de nettoyage des données.

Dans notre cas, nous avons dû supprimer une colonne qui ne contenait que des valeurs manquantes (NaN) pour préparer correctement nos données et les nettoyer. Les valeurs manquantes peuvent avoir un impact négatif sur les résultats de l'analyse et de la modélisation en raison de la perte d'informations. En supprimant cette colonne, nous avons pu éliminer tout bruit potentiel dans les données et améliorer la qualité des résultats de notre modèle. Cette étape de nettoyage des données est cruciale pour garantir que les résultats de notre analyse soient aussi précis que possible.

symmetry_worst	fractal_dimension_worst	Unnamed: 32
0.4601	0.11890	NaN
0.2750	0.08902	NaN
0.3613	0.08758	NaN
0.6638	0.17300	NaN
0.2364	0.07678	NaN

FIGURE 1.4 – NaN values

Chapitre 2

Réduction de dimensionnalité

2.1 Les méthodes de réduction de dimensionnalité

Dans ce chapitre, nous allons examiner les méthodes de réduction de dimensionnalité utilisées pour traiter les données de cancer du sein de Wisconsin. Les méthodes de réduction de dimensionnalité permettent de simplifier les données en réduisant le nombre de dimensions ou de caractéristiques sans perdre trop d'informations importantes. Certaines des méthodes couramment utilisées incluent l'analyse en composantes principales (PCA) et l'élimination a base de corrélation . Nous expliquerons les formules mathématiques, le pseudo code, les cas d'utilisation, les avantages et les inconvénients de chacune de ces méthodes.

2.1.1 l'analyse en composantes principales (PCA)

L'analyse en composantes principales (PCA) est une méthode de réduction de dimensionnalité qui consiste à trouver un sous-ensemble de caractéristiques qui explique le plus de variance dans les données. Elle est utilisée pour réduire la complexité des données tout en conservant les informations les plus importantes.

Les étapes de base pour effectuer une analyse en composantes principales sont les suivantes :

1. Centrer les données : Les données sont centrées en soustrayant la moyenne de chaque caractéristique à chaque observation. On peut le formaliser mathématiquement comme suit : $x'_{i,j} = x_{i,j} - \mu_j$, où $x_{i,j}$ est la valeur de la j-ème caractéristique pour l'i-ème observation, et μ_j est la moyenne de la j-ème caractéristique.
2. Calculer la matrice de covariance : La matrice de covariance est calculée à partir des données centrées pour mesurer les relations linéaires entre les caractéristiques. Elle est définie par la formule suivante : $C = \frac{1}{n}(X'^T X')$, où X' est la matrice des données centrées, et n est le nombre d'observations.
3. Trouver les vecteurs propres et les valeurs propres de la matrice de covariance : Les vecteurs propres sont les vecteurs qui définissent les directions dans lesquelles les données varient le plus. Les valeurs propres sont les variances dans ces directions. On peut les trouver en résolvant l'équation suivante : $Cv_i = \lambda_i v_i$, où v_i est le i-ème vecteur propre et λ_i est la i-ème valeur propre.

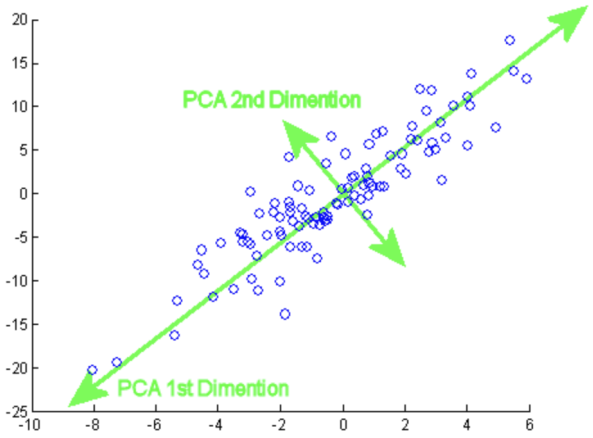

FIGURE 2.1 – PCA

4. Choisir les k premiers vecteurs propres : Les k premiers vecteurs propres qui ont les k plus

2.1.2 l'élimination à base de corrélation

La réduction de dimensionnalité de données à base de corrélation consiste à utiliser des techniques statistiques pour identifier les caractéristiques (variables) des données qui sont les plus corrélées les unes aux autres, puis éliminer, à partir de chaque deux caractéristiques corrélées, une d'eux en laissant l'autre, tout en comparant la corrélation à un seuil donné. Il s'agit d'une méthode naïve pour réduire la corrélation entre les caractéristiques et par la suite la dimension de nos données, mais reste une méthode à essayer dans notre projet. Une implémentation de cette méthode est comme la suivante :

```
temp = list(features) # la liste finale de features
temp2 = list(features)
for i in range(len(temp2)):
    for j in temp[i:]:
        if temp[i] != j : # pour eviter de comparer les memes colonnes
            # le seuil de comparaison de correlation est 0.7 ici
            if abs(df2[temp[i]].corr(df2[j])) > 0.7 :
                temp.remove(j)
non_corr_data = df2[temp] # dataframe final non correle
```

2.2 Les classificateurs utilisés

Dans cette étude, nous avons utilisé des méthodes de réduction de dimensionnalité sur une base de données initiale comprenant 569 individus et 30 caractéristiques, classifiée en 2 classes (M = malignant, B = bénigne) pour simplifier la complexité des données. Nous allons ensuite présenter le classificateur utilisé pour diagnostiquer le cancer du sein à partir de ces données

simplifiées. Les classificateurs sont des algorithmes qui apprennent à classer les données en utilisant des données d'entraînement. Parmi les méthodes couramment utilisées, on peut citer la régression logistique :

La régression logistique est un modèle de classification utilisant la fonction logistique pour prédire la probabilité d'appartenance à une classe donnée. Elle peut être formalisée mathématiquement comme suit :

$$\hat{p} = \frac{1}{1 + e^{-w^T x}}$$

où $\hat{p}$ est la probabilité prédite d'appartenance à la classe positive, x est la variable indépendante, w est le poids à estimer.

La fonction coût (ou perte) utilisée pour optimiser les coefficients β est généralement la log-vraisemblance négative, elle est définie comme suit :

$$J(\beta) = -\frac{1}{n} \sum_{i=1}^{n} [y_i log(\hat{p}_i) + (1 - y_i) log(1 - \hat{p}_i)]$$

où n est le nombre d'observations, y_i est la variable dépendante pour la i-ème observation et $\hat{p}_i$ est la probabilité prédite pour la i-ème observation.

L'objectif est de minimiser cette fonction coût pour estimer les coefficients β. Cela peut être fait en utilisant des algorithmes d'optimisation tels que la descente de gradient ou la méthode des moindres carrés.

Pseudo-code de l'algorithme de régression logistique

1. Initialiser aléatoirement les coefficients β_i de w avec $w = (\beta_i)_{i=[1,n]}$
2. Pour chaque itération :
   a. Calculer la probabilité prédite $\hat{p}$ en utilisant la fonction logistique
   b. Calculer la fonction coût en utilisant la log-vraisemblance négative
   c. Calculer les dérivées par rapport aux coefficients β_i
   d. Mettre à jour les coefficients β_i en utilisant la méthode d'optimisation choisie (ex : descente de gradient)
3. Répéter les étapes (a-d) jusqu'à convergence ou atteinte d'un nombre maximum d'itérations
4. Utiliser les coefficients β_i optimisés pour faire des prédictions en utilisant la fonction logistique

La régression logistique est un modèle de classification couramment utilisé en raison de sa simplicité et de sa capacité à gérer des variables catégoriques et quantitatives.

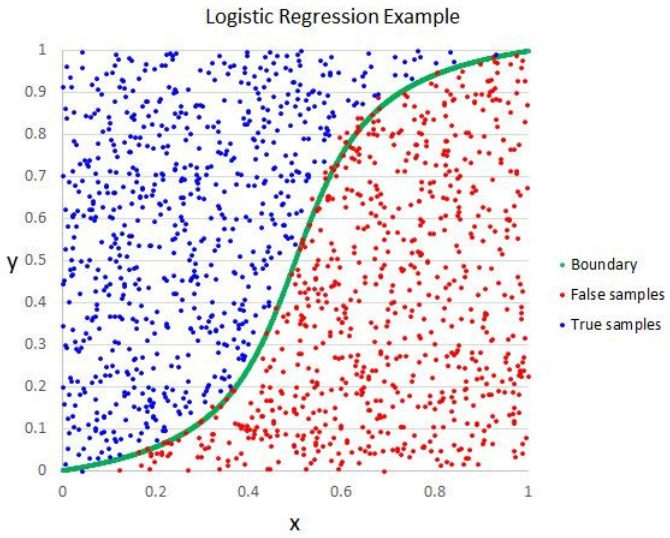

FIGURE 2.2 – PCA

Il existe plusieurs autres classificateurs dont les avantages différents, mais on optera dans notre projet uniquement à la régression logistique.

Chapitre 3

Résultats Expérimentaux

Dans ce chapitre, nous présenterons les différents outils utilisés dans la réalisation du projet, puis nous allons exposer les résultats des analyses réalisées.

3.1 Outils utilisés

3.1.1 Python

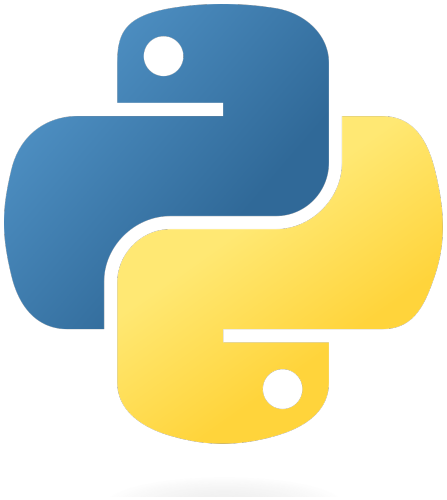

FIGURE 3.1 – Python

Python est un langage de programmation interprété, multi-paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions ; il est ainsi similaire à Perl, Ruby, Scheme, Smalltalk et Tcl.

3.1.2 Jupyter Notebook

Jupyter Notebook est un environnement de développement interactif (IDE) basé sur le web qui vous permet de créer et de partager des documents contenant du code en direct, des équations, des visualisations et du texte narratif. Il est basé sur le projet open-source IPython et est

couramment utilisé pour les tâches de science des données, de calcul scientifique et d'apprentissage automatique. Jupyter Notebook prend en charge de nombreux langages de programmation tels que Python, R, Julia, et bien plus encore. Il vous permet d'écrire, d'exécuter et de déboguer du code, de visualiser des données et de collaborer avec d'autres.

FIGURE 3.2 – jupyter notebook

L'interface notebook vous permet d'organiser votre code et votre texte narratif dans un seul document, facilitant le partage et la reproduction de votre travail. Il peut être exécuté localement sur votre ordinateur ou sur un serveur distant.

3.1.3 Numpy

FIGURE 3.3 – Numpy

La bibliothèque NumPy (http ://www.numpy.org/) permet d'effectuer des calculs numériques avec Python. Elle introduit une gestion facilitée des tableaux de nombres.

3.1.4 Pandas

Une librairie essentielle à l'analyse qui permet l'analyse la lecture et la manipulation des données (dans notre cas les fichiers CSV).

FIGURE 3.4 – Pandas

3.1.5 Scikit-Learn

Scikit-learn est une bibliothèque libre Python destinée à l'apprentissage automatique. Elle est développée par de nombreux contributeurs notamment dans le monde académique par des instituts français d'enseignement supérieur et de recherche comme Inria.

Elle propose dans son framework de nombreuses bibliothèques d'algorithmes à implémenter, clé en main. Ces bibliothèques sont à disposition notamment des data scientists.

FIGURE 3.5 – Scikit-Learn

Elle comprend notamment des fonctions pour estimer des forêts aléatoires, des régressions logistiques, des algorithmes de classification, et les machines à vecteurs de support. Elle est conçue pour s'harmoniser avec d'autres bibliothèques libres Python, notamment NumPy et SciPy.

3.1.6 Matplotlib

Matplotlib est une bibliothèque de tracé open-source pour Python. Elle fournit une interface de programmation orientée objet pour intégrer des graphiques dans des applications utilisant des kits d'interface graphique généraux tels que Tkinter, wxPython, Qt ou GTK. Matplotlib peut être utilisé pour créer une grande variété de visualisations statiques, animées et interactives en Python, notamment des tracés de lignes, des nuages de points, des barres, des histogrammes, des graphiques 3D, et plus encore. Il est largement utilisé en visualisation de données et en analyse de données dans les domaines de la science, de l'ingénierie, de la finance, etc.

FIGURE 3.6 – matplotlib

3.1.7 skfeature

skfeature est une bibliothèque open-source pour Python qui permet de sélectionner des caractéristiques pour les tâches d'apprentissage automatique. Il fournit une interface pour différents algorithmes de sélection de caractéristiques tels que la régression Lasso, la récursion de sélection de caractéristiques, la sélection de caractéristiques basée sur la corrélation, etc. Il est utilisé pour améliorer les performances des modèles d'apprentissage s.

3.2 Résultats sans réduction de dimensionnalité

La base de données initiale contenait beaucoup de colonnes qui étaient fortement corrélées entre elles, ce qui aurait pu affecter la précision de notre modèle. Les variables corrélées peuvent introduire des biais dans les résultats de la modélisation en donnant une importance excessive à certaines variables. Pour éviter cela, il est recommandé de sélectionner les caractéristiques les plus pertinentes pour notre modèle en utilisant des techniques de sélection des caractéristiques, telles que la régression linéaire ou l'analyse en composantes principales.

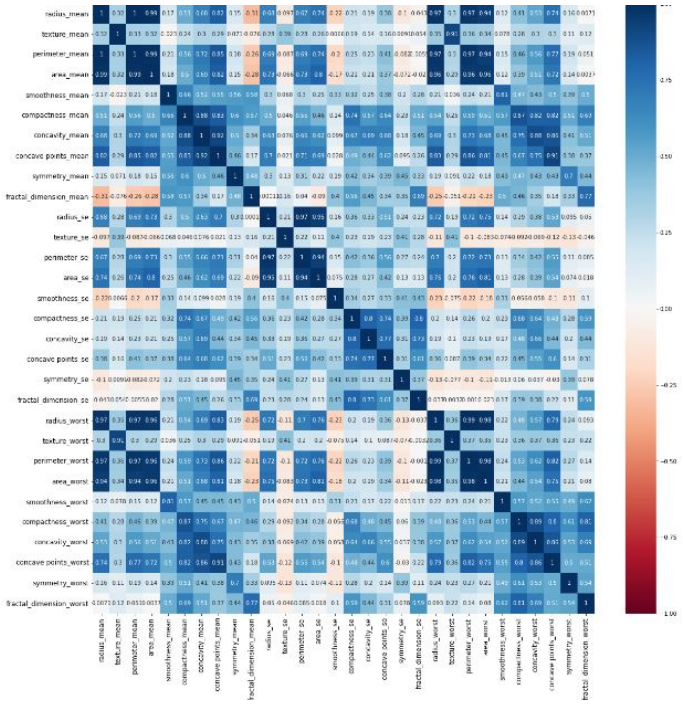

FIGURE 3.7 – des variables corrélées

mais on obtient en appliquant la régression logistique un score de 94% .

3.3 Résultats avec réduction de dimensionnalité

Dans ce chapitre, nous allons explorer les méthodes de réduction de dimension pour réduire la complexité des données.En utilisant des méthodes de réduction de dimension telles que l'analyse en composantes principales (ACP), nous allons créer des nouvelles variables qui sont un sous-ensemble des variables d'entrée et qui sont généralement plus pertinentes pour la tâche de modélisation. Nous allons utiliser ces méthodes pour réduire la complexité des données tout en conservant l'essentiel de l'information pour notre modèle de classification. Enfin, nous comparerons les résultats obtenus avec les méthodes de réduction de dimension avec les résultats obtenus avec l'ensemble des variables d'origine pour évaluer l'efficacité de ces méthodes.

3.3.1 corrélation

Pour la méthode de corrélation, on a essayé de tester le paramètre optimal pour obtenir les meilleurs résultats en utilisant cette méthode, on a expérimenté sur le seuil de comparaison et obtenue les plots suivants :

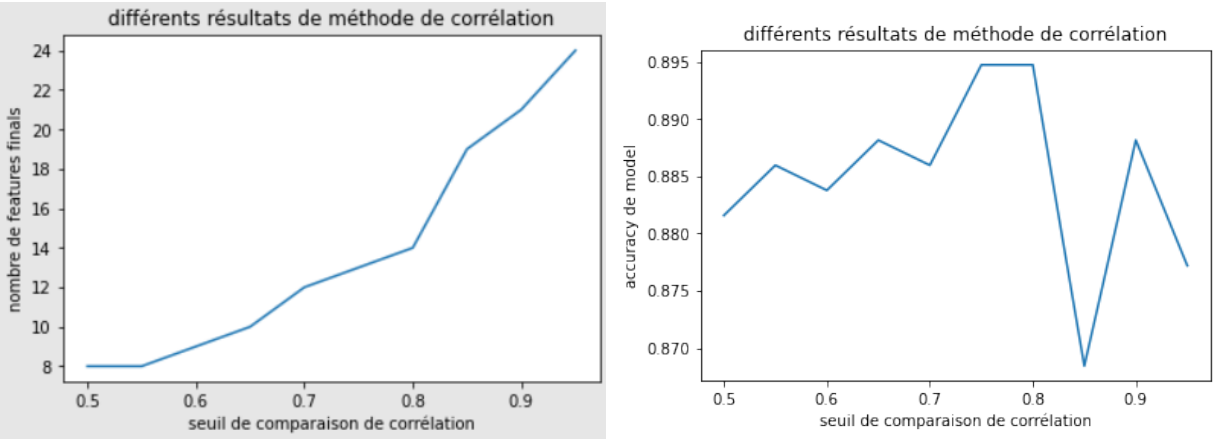

FIGURE 3.8 – Expérimentations avec la méthode de corrélation

D'après les résultats ci-dessus, un bon seuil de comparaison est modéré (environs de 0.8) pour obtenir une bonne précision ainsi qu'une réduction de caractéristiques importante.

On peut donc conclure que notre seuil doit obéir des conditions pour avoir des résultats acceptables. On optera pour un seuil de 0.8 pour tester notre méthode : On obtient une précision de 0.9488 dans notre essai.

3.3.2 Résultat de l'utilisation de PCA

La PCA nous a permet tout d'abord de créer un nouveau système de coordonnées pour représenter les données sans corrélation entres les dimensions. Cela permet de conserver les caractéristiques les plus importantes des données tout en suppriment les corrélations entre les variables en créant des nouvelles variables qui sont de combinaisons linéaires non corrélées des variables d'origine. Cela permet d'éviter les redondances dans les données et d'améliorer les performances des modèles de classification et de prévision.

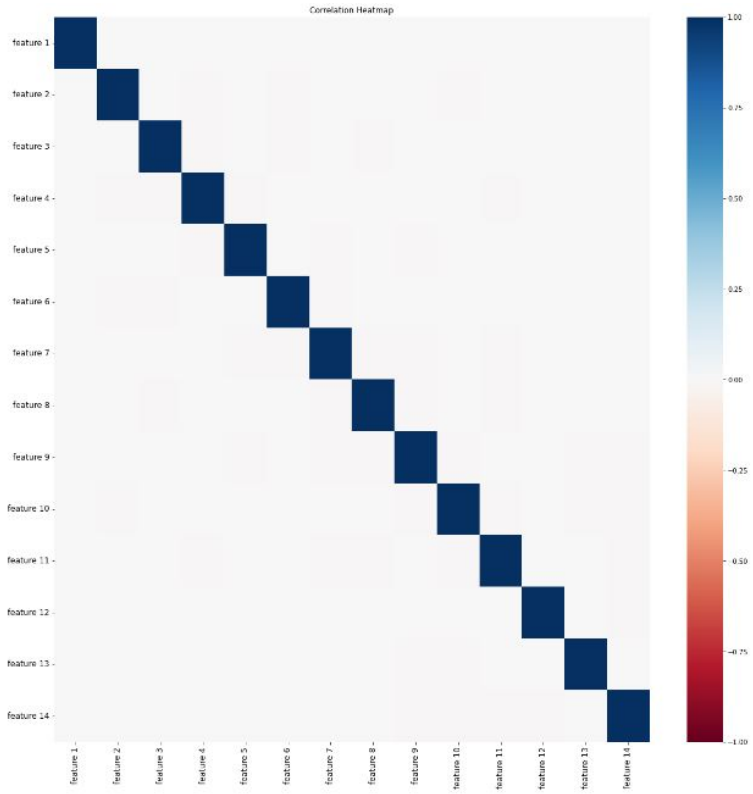

FIGURE 3.9 – corrélation nulle entre les vaiables

Pour réduire les dimensions d'une base de données à l'aide de la PCA, il est nécessaire de choisir les composantes principales les plus représentatives. Pour ce faire, on utilise généralement le graphe de la variance cumulée en fonction du nombre de composantes. Ce graphe montre la proportion de variance totale expliquée par chaque composante principale ajoutée. En choisissant un nombre de composantes qui représente une proportion suffisante de la variance totale, on peut conserver les caractéristiques les plus importantes des données tout en réduisant le nombre de dimensions. Il est généralement recommandé de choisir un nombre de composantes qui permet de conserver une proportion de la variance totale de 80% à 90%.

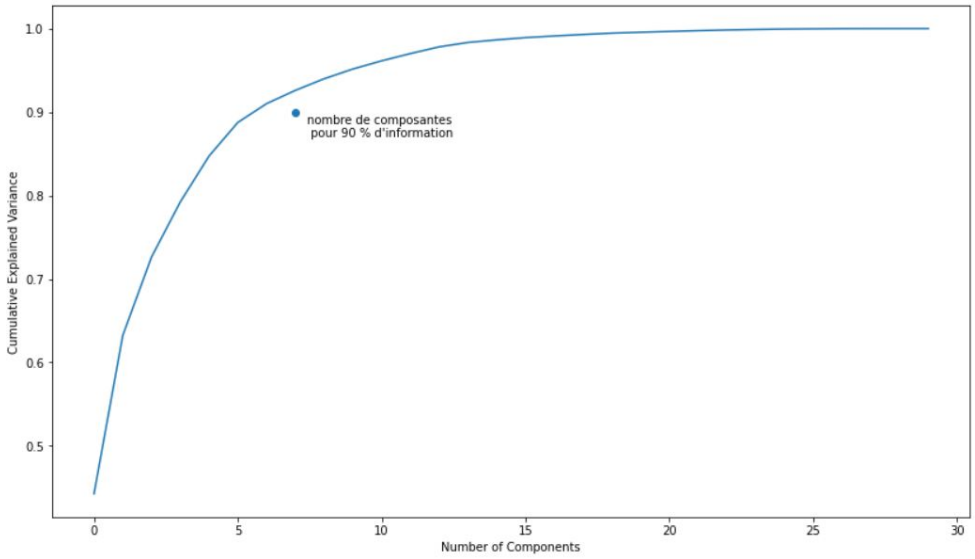

FIGURE 3.10 – Cumulative explained Variance

D'après le graphe de la variance cumulée en fonction du nombre de composantes, nous avons choisi les 7 premières composantes principales pour conserver près de 91% de l'information,voici la distribution de variance explique entre ces composantes.

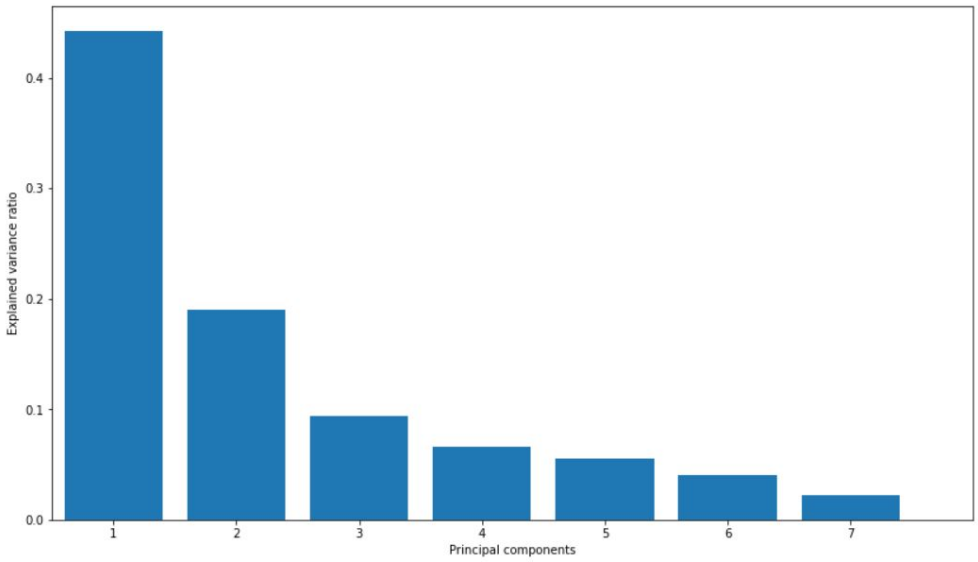

FIGURE 3.11 – distribution de variance expliqué

En utilisant ces 7 composantes principales, nous avons réussi à réduire le nombre de dimensions de la base de données tout en conservant une grande proportion de la variance totale. Cela permet de conserver les caractéristiques les plus importantes des données tout en éliminant les redondances et les corrélations inutiles entre les variables. Cette étape de réduction de dimension permet d'améliorer les performances des modèles de classification et de prévision en utilisant un sous-ensemble pertinent des variables d'entrée.à la fin on obtenir un score de plus de 96% en utilisent le régression logistique.

3.3.3 comparions globale

Dans ce rapport, nous avons comparé l'accuracy obtenue en utilisant la régression logistique avec différentes approches de réduction de la complexité des données. Plus précisément, nous avons utilisé l'approche de base, qui ne comporte aucune réduction, l'Analyse en Composantes Principales (PCA), et enfin l'élimination des colonnes corrélée. Les résultats obtenus montrent que l'accuracy la plus élevée a été obtenue en utilisant PCA, avec 96,25%, suivi de l'approche de base avec 94,95% et enfin l'élimination d'une colonne corrélée avec 94,75%. Il est important de noter que l'utilisation de PCA et l'élimination d'une colonne corrélée ont permis de réduire la complexité des données, tout en maintenant un niveau d'accuracy élevé.

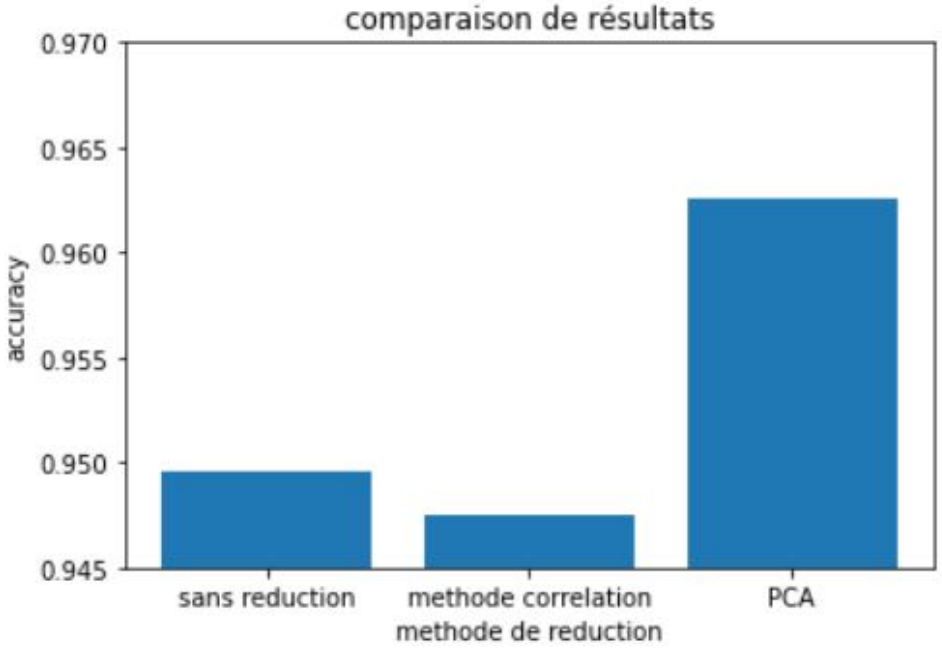

FIGURE 3.12 – comparaison des accuracy

Chapitre 4

Conclusion

Étant un étudiant en filière Ingénierie d'Intelligence Artificielle (2IA), ce projet a été une bonne opportunité pour s'initier au domaine de l'ingénierie des données : un domaine possible comme voie de carrière pour notre filière.

Généralement, ce projet a été une très bonne occasion pour consolider nos connaissances en Python et Analyse de données et nous a notamment permis de comprendre les étapes de classification et de pré-traitement des données volumineuse avec des dimension très grands .

Depuis l'obtention de la base de données initiale, au traitement manuel pour éliminer les détails passant au pré-traitement automatique et tester quelques algorithmes de réduction de dimensionnalité. nous a pu découvrir les étapes que suit la création d'un modèle et plus précisément un modèle basée sur des données avec des dimensions très grand.

Bibliographie

[1] <https://scikit-learn.org/stable/supervised_learning.html#supervised-learning>, 2023.

[2] <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>, 2023.

[3] <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>, 2023.

[4] <https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch12/5214889-fra.htm>, 2023.