Réduction de dimensionnalité appliquée au Breast Cancer Wisconsin (Diagnostic) Data Set

AYOUB KOUSSY

MOHAMED AMINE TAIF

Jury:

Pr Mohamed LAZAAR

Sommaire:

- Jeu de données
- Réduction de dimensionnalité
- Résultats expérimentaux
- Conclusion

Jeu de données

jeu de données

Base des données :

Le jeu de données utilisé Breast Cancer Wisconsin (Diagnostic) Data Set , est un ensemble de données qui contient des informations sur des caractéristiques de cellules cancéreuses de sein.

	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

jeu de données

Visualisation des données :

En observant les quatre boîtes à moustaches représentant les colonnes radius mean, texture mean, perimeter mean et area mean on peut dire que les valeurs prise par les colons ne sont pas homogènes

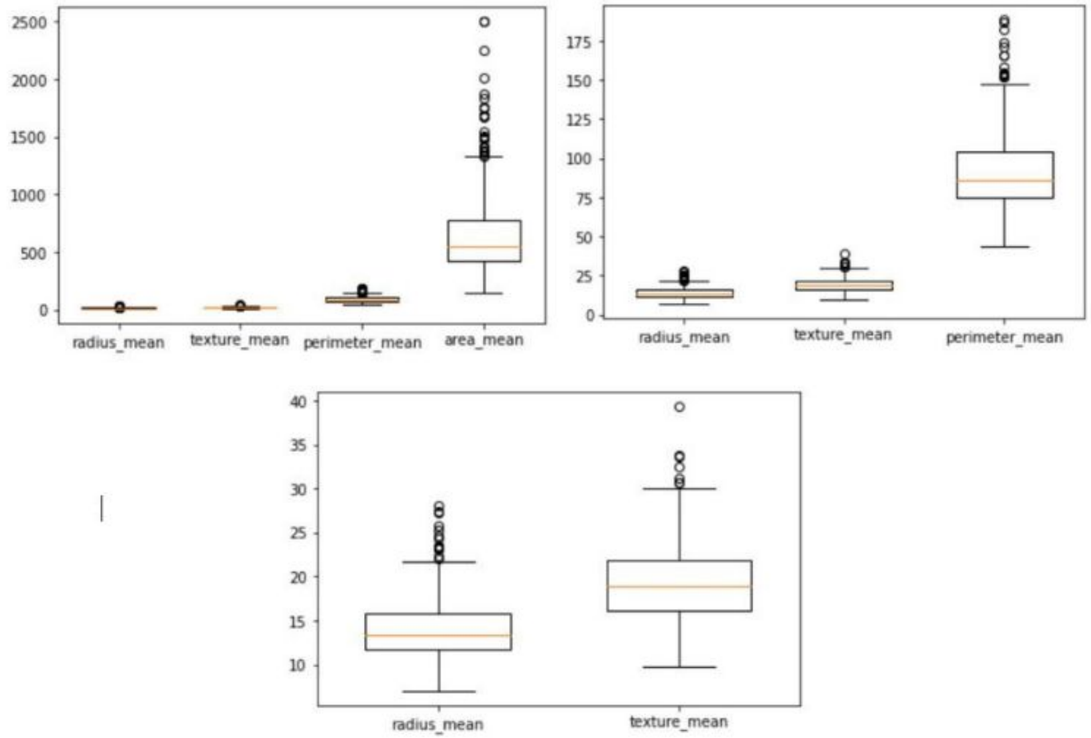


FIGURE 1.2 – Boîte à moustaches

jeu de données

Visualisation des données :

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.

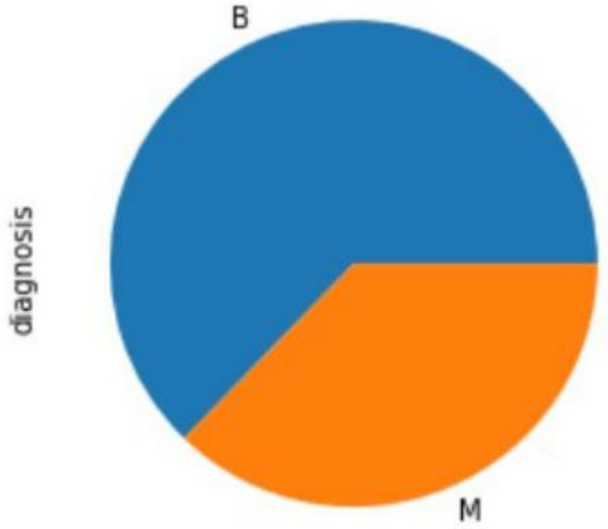


FIGURE 1.3 – Distrubition des class

Réduction de Dimensionnalité

PCA

PCA (Analyse de composantes principales)

La PCA est une technique statistique qui permet de réduire la dimensionnalité des données en identifiant les directions de variation maximale dans les données.

PCA algorithm($\mathbf{X}$, k): top k eigenvalues/eigenvectors

% $\underline{\mathbf{X}}$ = N × m data matrix,
% … each **data point** $\mathbf{x}_i$ = column vector, i=1..m

- $\underline{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i$
- $\mathbf{X} \leftarrow$ subtract mean $\underline{\mathbf{x}}$ from each column vector $\mathbf{x}_i$ in $\underline{\mathbf{X}}$
- $\Sigma \leftarrow \mathbf{X}\mathbf{X}^T$ … covariance matrix of $\mathbf{X}$
- $\{ \lambda_i, \mathbf{u}_i \}_{i=1..N}$ = eigenvectors/eigenvalues of Σ
  … $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$
- Return $\{ \lambda_i, \mathbf{u}_i \}_{i=1..k}$
  % top k principal components

Corrélation

Méthode de corrélation

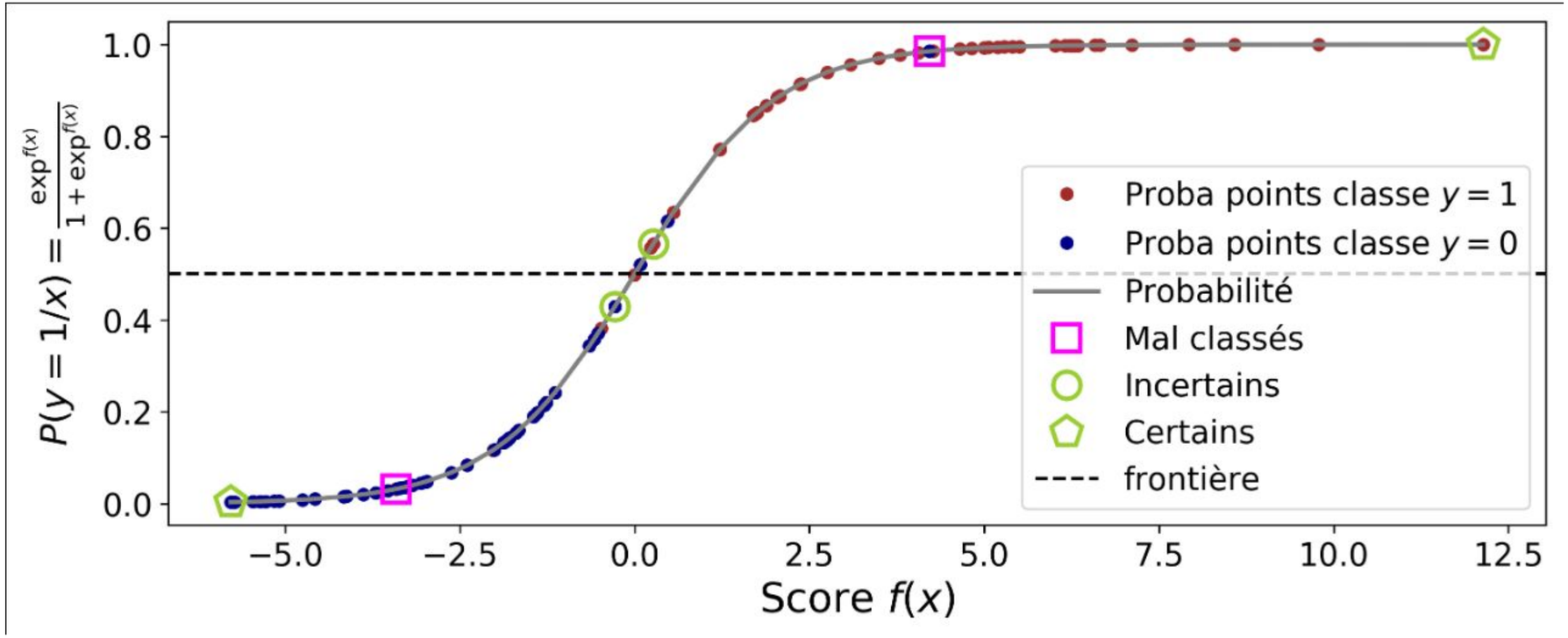
Après avoir identifié les caractéristiques qui sont les plus corrélées les unes aux autres, on élimine à partir de chaque deux caractéristiques corrélées, une d'eux en laissant l'autre, tout en comparant la corrélation à un seuil donné.

```
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
```

Classificateur

Régression Logistique

Un classificateur qui permet de modéliser la probabilité d'appartenance à une classe binaire en utilisant une fonction de lien logistique (sigmoid) qui relie les variables explicatives à la probabilité cible. Cette méthode permet de construire des modèles de prédiction pour des données à deux catégories.



Résultats expérimentaux

Résultats expérimentaux

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

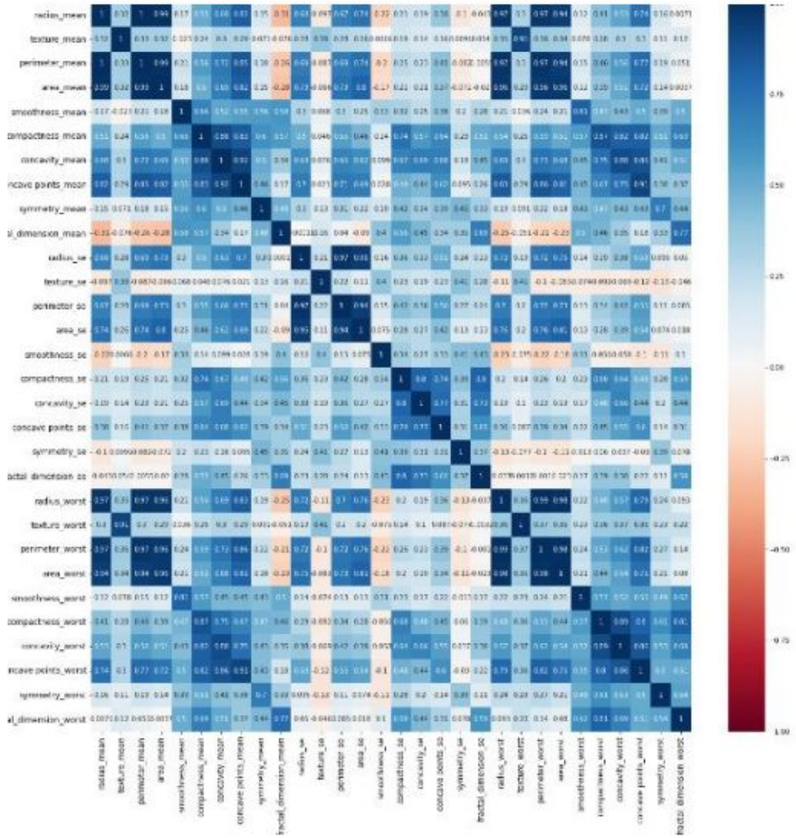
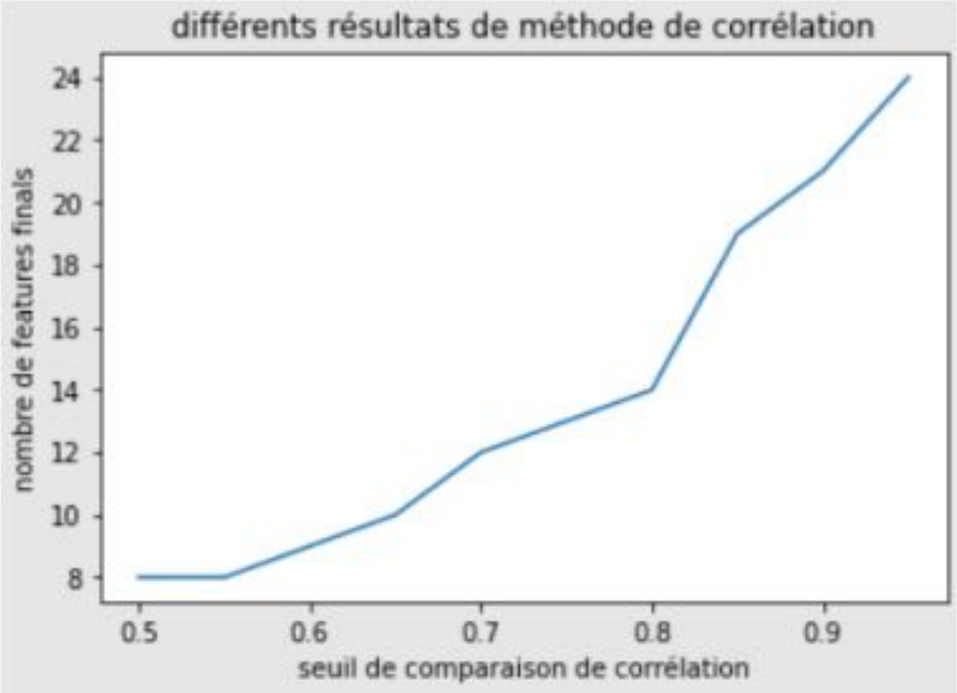


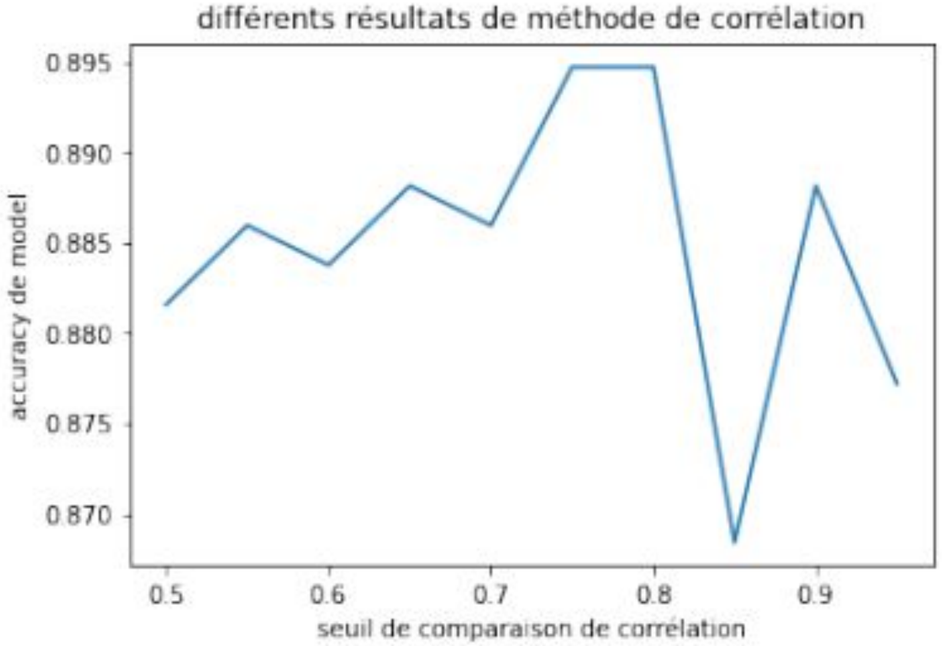
FIGURE 3.7 – des variables corrélées

Résultats expérimentaux

Résultats avec réduction de dimensionnalité:

corrélation





Résultats expérimentaux

Résultats avec réduction de dimensionnalité:

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.

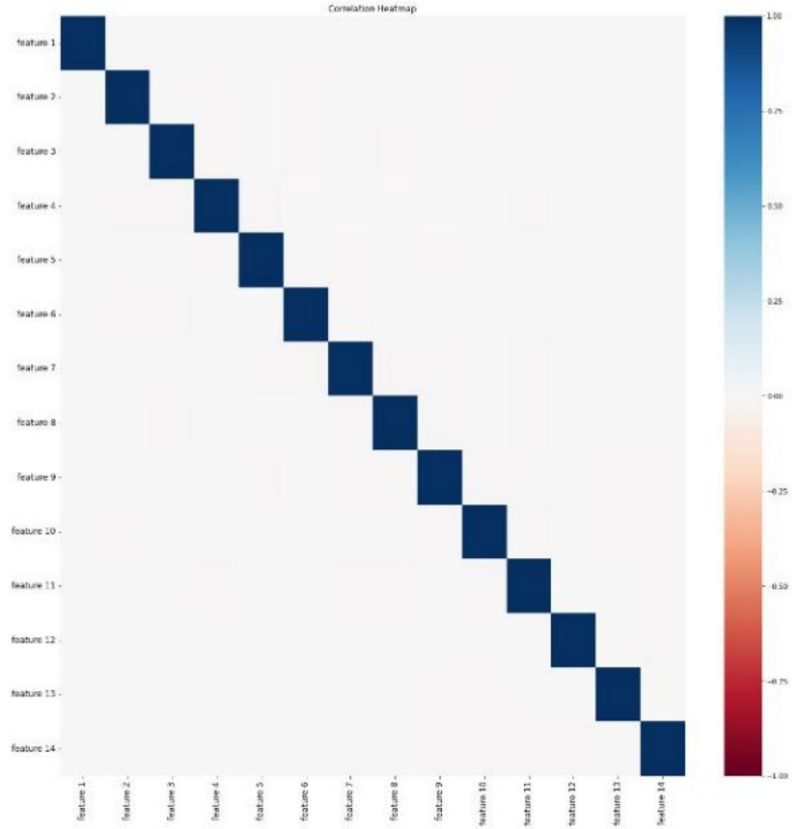


FIGURE 3.9 – corrélation nulle entre les vaiables

Résultats expérimentaux

Résultats avec réduction de dimensionnalité:

PCA

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

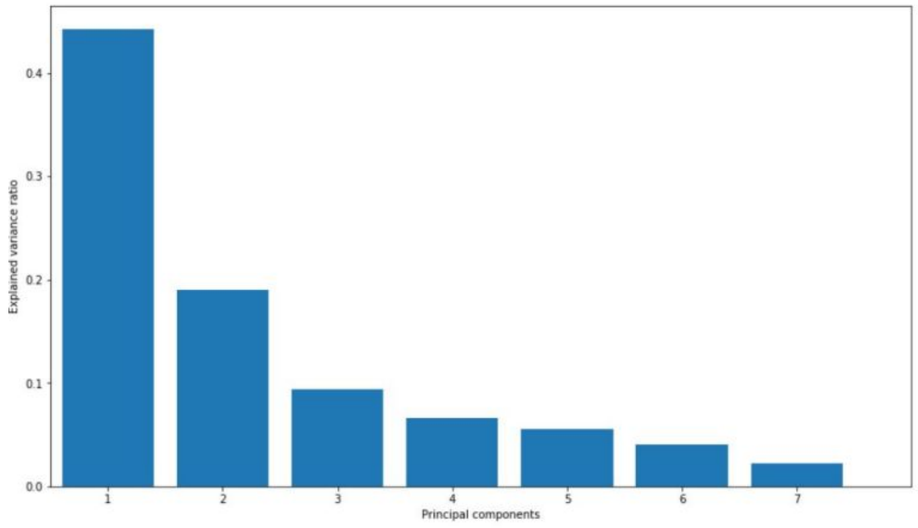


FIGURE 3.11 – distribution de variance expliqué

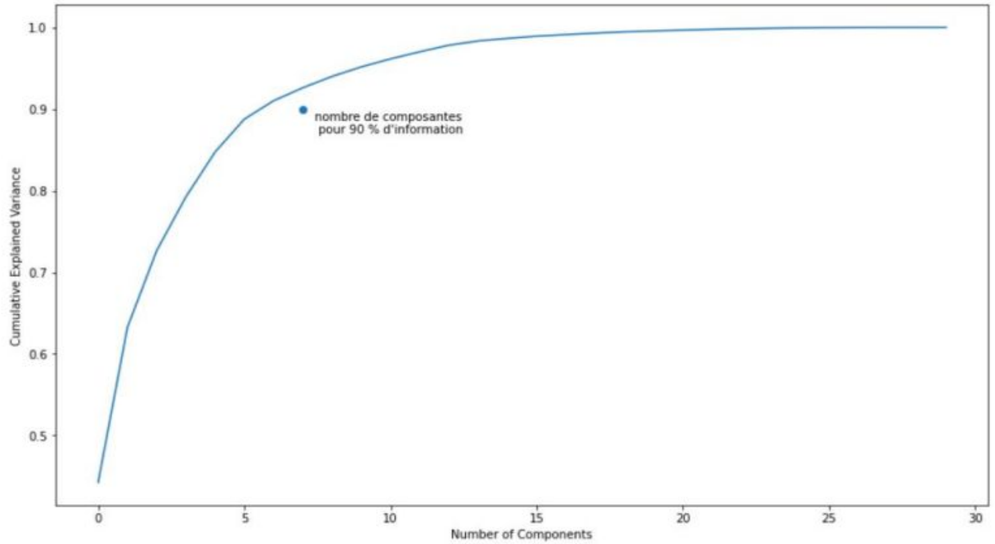


FIGURE 3.10 – Cumulative explained Variance

Résultats expérimentaux

comparions globale :

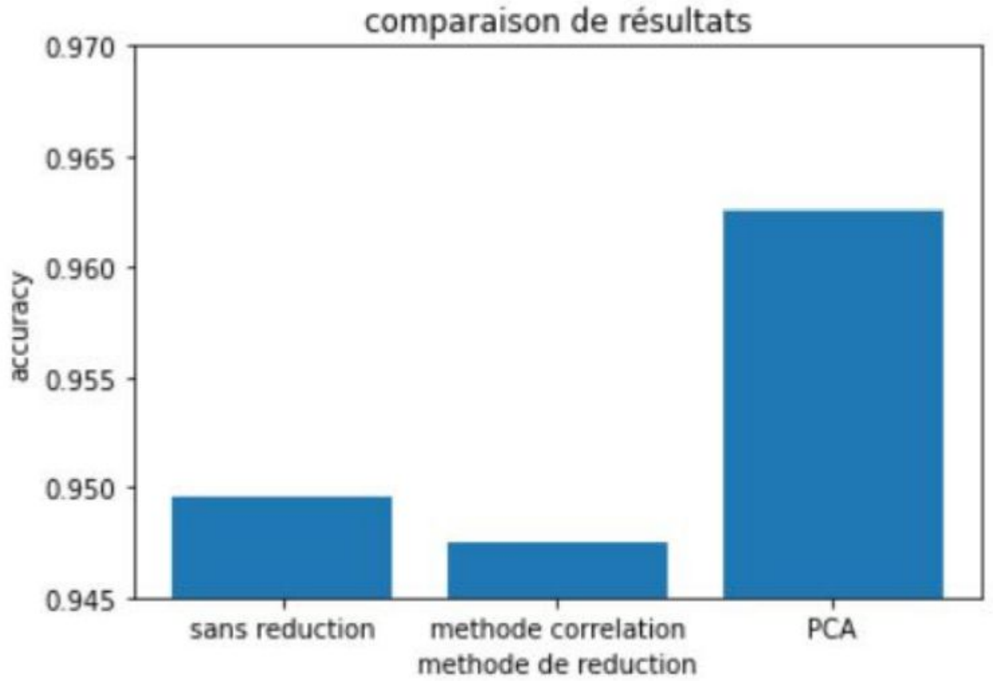


FIGURE 3.12 – comparaison des accuracy

Conclusion

Conclusion

En conclusion, les résultats obtenus montrent que l'utilisation de PCA a permis d'obtenir la meilleure accuracy de 96,25%, suivi de l'approche de base avec 94,95% et enfin l'élimination d'une colonne corrélée avec 94,75%. Ces résultats indiquent qu'il est possible de réduire la complexité des données tout en maintenant un niveau d'accuracy élevé en utilisant des techniques comme PCA ou l'élimination de colonnes corrélées.

Sources

- https://scikit-learn.org/stable/supervised_learning.html# supervised-learning, 2023.
- https://scikit-learn.org/stable/modules/generated/sklearn.decomposition, 2023. .
- https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data, 2023
- https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch12/5214889-fra.htm, 2023.

Merci pour votre attention