

PREDICTION ET CLASSIFICATION DU CANCER DE SEINS

INTRODUCTION

Le cancer du sein demeure l'une des maladies les plus courantes et les plus dévastatrices touchant les femmes à travers le monde. Le dépistage précoce et l'identification précise des patients atteints de cette maladie sont d'une importance capitale pour un traitement réussi et une amélioration du taux de survie. Dans cette optique, les progrès de l'intelligence artificielle et de l'apprentissage automatique offrent de nouvelles possibilités pour la classification des patients atteints du cancer du sein. Ces technologies prometteuses peuvent faciliter la prise de décision médicale en identifiant rapidement et avec précision les patients atteints de cancer du sein à partir de leurs données cliniques.

PROBLEMATIQUE

Développer un modèle fiable et précis pour distinguer les patients atteints du cancer du sein de ceux qui ne le sont pas, en se basant uniquement sur les données cliniques disponibles, représentant un défi majeur. Il est essentiel de prendre en compte un large éventail de variables cliniques, tout en assurant la généralisation et la validité du modèle dans différents contextes hospitaliers. Ce projet vise à résoudre ces défis en développant un modèle avancé et fiable, permettant d'identifier rapidement les patients nécessitant une attention particulière et de fournir des traitements personnalisés.

Data Set/Variables

Nos données sont resumées dans un fichier csv nommé : **"breast-cancer.csv"**

Les variables contenues dans les données sont :

- Rayon moyen (mean_radius)
- Texture moyenne (mean_texture)
- Périmètre moyen (mean_perimeter)
- Surface moyenne (mean_area)
- Souplesse moyenne (mean_smoothness)

RESULTATS ET DISCUSSIONS

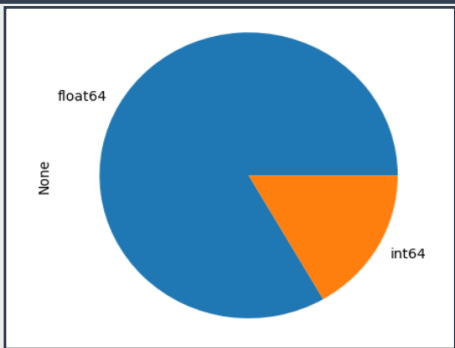
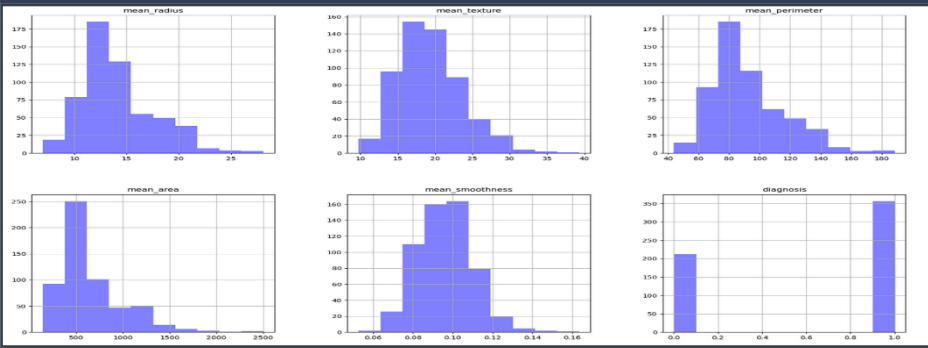
Modèle de prédiction (Régression logistique)

On a utilisé le modèle logistique pour procéder à la prédiction et classification du jeu de données.

Lecture du jeu de données

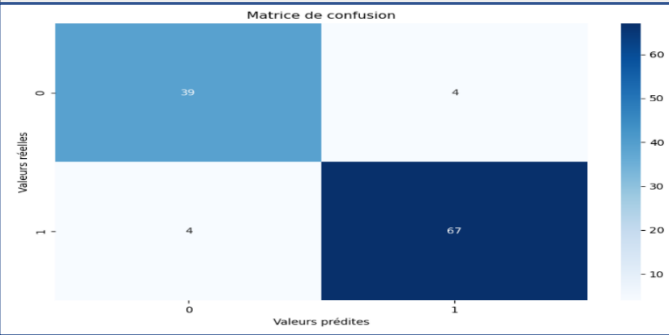
| | mean_radius | mean_texture | mean_perimeter | mean_area | mean_smoothness | diagnosis |
|---|-------------|--------------|----------------|-----------|-----------------|-----------|
| 0 | 17.99 | 10.38 | 122.80 | 1001.0 | 0.11840 | 0 |
| 1 | 20.57 | 17.77 | 132.90 | 1326.0 | 0.08474 | 0 |
| 2 | 19.69 | 21.25 | 130.00 | 1203.0 | 0.10960 | 0 |
| 3 | 11.42 | 20.38 | 77.58 | 386.1 | 0.14250 | 0 |
| 4 | 20.29 | 14.34 | 135.10 | 1297.0 | 0.10030 | 0 |

Visualisations



Evaluation du modèle

On a utilisé une matrice de confusion pour évaluer la performance du modèle



Rapport de classification du modèle

| Rapport de classification : | | | | |
|-----------------------------|-----------|--------|----------|---------|
| | precision | recall | f1-score | support |
| 0 | 0.93 | 0.91 | 0.92 | 43 |
| 1 | 0.94 | 0.96 | 0.95 | 71 |
| accuracy | | | 0.94 | 114 |
| macro avg | 0.94 | 0.93 | 0.93 | 114 |
| weighted avg | 0.94 | 0.94 | 0.94 | 114 |

CONCLUSION

En conclusion on admet que ce modèle de prédiction a su s'adapter avec le jeu de données et trouver une corrélation entre les données. Il nous a permis d'avoir résultats satisfaisant avec une précision d'environ 94%

Références

Documentations Scikit-learn