Prédiction/Classification du Cancer des Seins

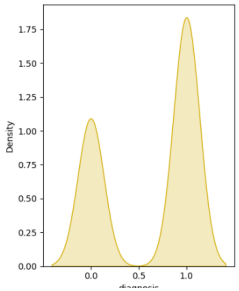
Un cancer du sein résulte d'un dérèglement de certaines cellules qui se multiplient et forment le plus souvent une masse appelée tumeur. Il en existe différents types qui n'évoluent pas de la même manière. Certains sont « agressifs » et évoluent très rapidement, d'autres plus lentement. Les cellules cancéreuses peuvent rester dans le sein. Ce projet vise à résoudre un problème concernant une classification des patients ayant le cancer des seins et ceux n'ayant pas le cancer.

Résumé

En tant que développeur Data, nous créons un modèle permettant de classifier une personne ayant un cancer des seins ou non. Cela est réalisable que par le Machine Learning. Les données sont téléchargées sur une plateforme dont le lien est mentionné dans les outils. Notre dataset ne comporte pas de valeur abbérantes. Ainsi, grâce à la visualisation , nous avons préparé différents algorithmes pour entrainer notre modèle. Le meilleur score est obtenu par la régression Logistique et le Random Forest avec accurancy = 93% et une précision = 94%.

Exploration de données :

Le dataset comporte six(06) colonnes dont 5 features et un target. Le taget est la colonne diagnosis. Le target est composé de valeur 0 et 1, soit Maligne et Bénigne. Alors, nous avons fait une réprésentation graphique.



Modèle utilisé :

1. Regression Linéaire.
Score = 0.6316916969686668
Erreur = 0.0859127881783686

2. Regression Logistique.
Score =0.9300699300699301
Erreur =0.06993006993006994

3. KNeighbors Classifier
Score = 0.8881118881118881
Erreur = 0.11188811188811189

4. SVC en utilisant Vector Machine Algorithm
Score = 0.9230769230769231
Erreur =0.07692307692307693

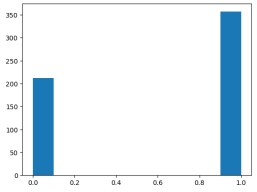
5. SVC (Support Vector Class) en utilisant Kernel Support Vector Machine Algorithm
Score =0.9090909090909091
Erreur =0.09090909090909091

6. Naïve Bayes Algorithm (GaussianNB)
Score =0.9230769230769231
Erreur =0.07692307692307693

7. Decision Tree Algorithm
Score =0.9090909090909091
Erreur = 0.11888111888111888

8. Random Forest Classification algorithm
Score =0.9300699300699301
Erreur =0.06993006993006994

RESULTAT



L'histograme ci-contre est lea visulation de notre target. Nous avons donc les resultat ci-dessous .

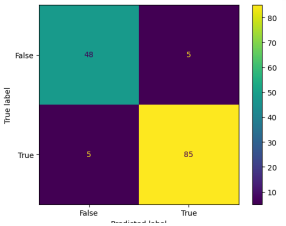
Ces modèles sont ceux qui ont plus de précision

L'erreur s'obtient grâce à cet algorithme

```
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error

model = RandomForestClassifier(n_estimators=100, min_samples_leaf=10, random_state=1)
model.fit(X_train, Y_train)
predictions = model.predict(X_test)
print("L'erreur  est:")
mean_squared_error(predictions, Y_test)
```

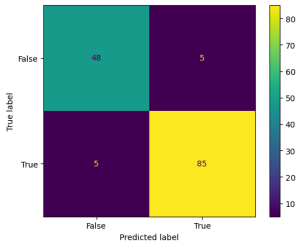
```
L'erreur  est:

0.06293706293706294
```

Cas 1 : Régression Logistique



Matrice de confusion de l'algorithme Régression Linéaire
Accuracy = 93%
Précision = 94%.
Vrai Positive (VP): 83
Vrai Negative (VN): 44
Faux Positive (FP): 9
L'erreur est: : 0.0699

Cas 2: Random Forest Classification



Matrice de confusion de Random Forest Classification algorithm
Accuracy = 93%
Précision = 94%.
Vrai Positive (VP): 85
Vrai Negative (VN): 48
Faux Positive (FP): 5
Faux Negative (FN): 5
L'erreur est: : 0.0699

Conclusion :

Notre travail a porté sur la prédiction/Classification du cancer de sein. Le resultat obtenu sur les différents modèles et la matrice de confusion, nous permettent de dire que l'algorithme Random Forest et Régression Logistique sont éfficaces. Ils donnent un accurancy = 93% et une précision de 94%.

Github Link :

dataset Link: click here