Lamoureux Léo 21602944
Pires Pinto Lola 21706287

ALGO : RANDOM FOREST

Etape 2

Datasets : choisis dans Scikit-learn

Dataset Classification :

- Lien : https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html

Dataset Régression :

- Lien : https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_diabetes.html

Nous avons choisi ces Datasets car parmi ceux proposés sur Scikit-learn, il s'agit de ceux que nous trouvons les plus clairs.

Pour une première approche nous préférions donc des datasets que nous comprenons bien afin d'apprendre à les manipuler correctement, ainsi que de bien comprendre les principes de base de nos algorithmes et comment les appliquer.

De plus, ils ont des applications que nous trouvons importantes : classification de tumeurs pour le cancer du sein, et prédiction de taux de glycémie pour le diabète.

Programme de test : Ouvrir le notebook fourni dans le mail.

RÉSULTATS OBTENUS :

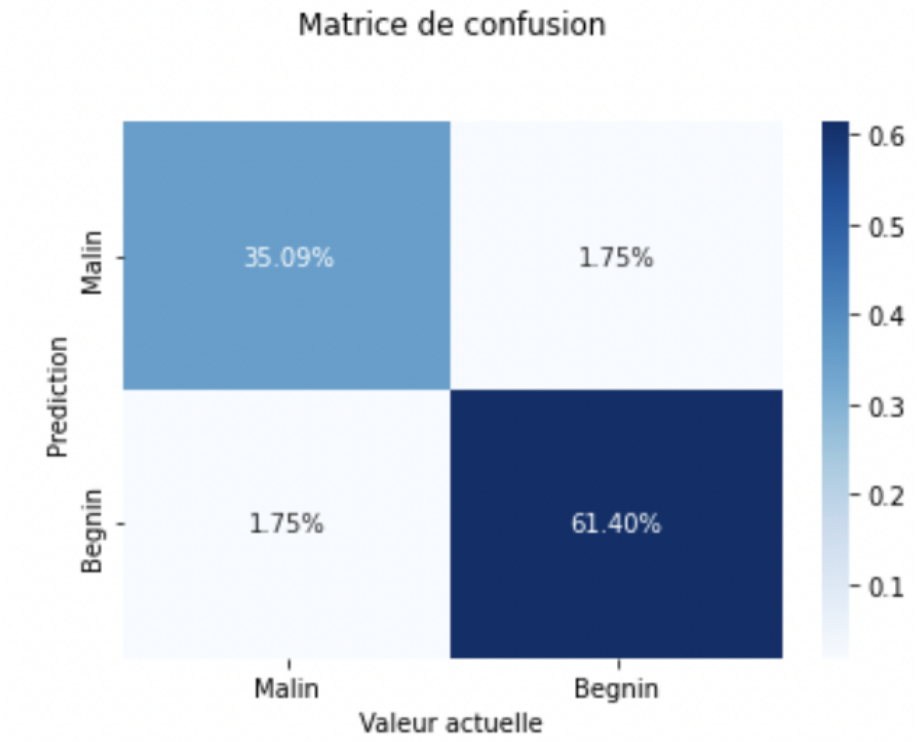
(PLUS D'ANALYSES SONT PRÉSENTÉES DANS LE CODE SOURCE)

Classification Breast cancer :

Pour évaluer notre modèle, nous avons choisi d'utiliser une matrice de confusion. Celle-ci compare, pour la variable cible, les données réelles à celles prédites par le modèle.

Elle nous permet de visualiser les rapports entre prédictions correctes et incorrectes.

```
[[40  2]
 [ 2 70]]
```



On observe 1,75% de faux positifs et 1,75% de faux négatifs, donc un total d'environ 3,5% de fausses prédictions contre environ 96% de bonnes prédictions au total.

A partir de cette matrice on peut calculer toutes les métriques nécessaires pour évaluer notre modèle.

La matrice graphique étant normalisée pour afficher des pourcentages, on calcule ces métriques à partir des valeurs exactes.

Nous avons choisi d'utiliser les métriques suivantes :

recall_score : le recall donne le pourcentage de positifs bien prédits par le modèle. Plus il est élevé, moins le modèle ne rate de cas positifs (tumeur maligne). C'est donc un indicateur important dans le cadre du cancer, ou minimiser les faux négatifs peut sauver des vies. Ici le recall à un score de : 40/(40+2) = 0,952%.

Nous avons donc environ 95% de cas positifs qui sont bien prédits par le modèle.

accuracy_score : il donne le pourcentage de prédictions correctes (positives comme négatives). Il est donc aussi important à prendre en compte, et à associer avec le recall pour avoir des informations précises. Il nous apprend ici que nous avons environ 97% de prédictions correctes (négatives comme positives). Valeur exactes = (40+70)/114 = 110/114 = 0,965% de précision.

Le modèle obtenu est donc très précis sur ce dataset.

Regression Diabete :

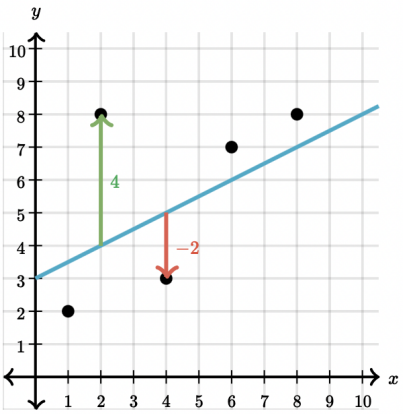
Les métriques de régression usuellement utilisées sont MAE (Erreur absolue moyenne), MSE (erreur quadratique moyenne), RMSE (erreur quadratique moyenne) et R^2.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

MAE = moyenne effectuée sur la différence absolue entre les valeurs réelles et celles prédites. Il mesure la moyenne des résidus dans l'ensemble des données.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

MSE = somme des différences entre valeur réelle et prédiction au carré le tout divisé par le nombre de valeurs. C'est donc la moyenne des différences au carré. Il mesure la variance des résidus.



Résidus = à quel point les valeurs réelles sont écartées de la droite de régression. Ce sont donc les erreurs observées.

(Exemple : voir schéma à gauche).

Plus la somme des différences est proche de 0 (plus les valeurs sont ajustées à la droite de régression), plus le modèle est précis. Donc plus la valeur de MSE est proche de 0, plus le modèle est précis.

lien du schéma

Ces deux métriques apportent donc des informations sur la précision du modèle.

MSE pénalise beaucoup plus les grandes erreurs que MAE. En effet, une grande différence entre valeur prédite et réelle sera élevée au carré dans MSE et aura donc plus d'impact dans la valeur finale. C'est pourquoi dans notre cas (prédiction d'un taux de glycémie pour

détecter le diabète) on utilisera la MSE car une grande erreur de prédiction peut être grave. Les grandes erreurs de prédictions doivent être détectées (pénalisées).

RMSE = $\sqrt{MSE}$, ce qui correspond à une mesure d'écart type, qui ramène la valeur à la même échelle que celle de l'erreur de prédiction, c'est pourquoi nous préférons choisir RMSE. Tout comme MSE, plus il est élevé, moins le modèle est précis.

Nous avons aussi utilisé le R^2(ou coefficient de détermination).

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Numérateur = calcul l'erreur entre valeurs réelles et prédites.
Dénominateur = calcul la variance des valeurs du dataset.

R^2 calcule donc les erreurs par rapport au niveau de variation présent dans les données. Plus il est proche de 1, plus la fraction est proche de 0, donc plus les erreurs du modèle sont petites par rapport à la variance des données : le modèle est précis. À l'inverse, si il est proche de 0 le modèle n'est pas précis.

Si les erreurs sont plus grandes que la variance des données on peut obtenir un R^2négatif.

C'est un indicateur très utilisé en statistiques et assez visuel pour définir la qualité d'une régression (associable à un pourcentage pour une valeur entre 0 et 1). Ici le score est très bas, donc le modèle est imprécis.

Nos résultats :

```
RMSE =  62.10979059407985
r2_score =  0.34239707898572114
```

RMSE est élevé et R^2 est petit, notre modèle en régression n'est donc pas précis du tout.

CONCLUSION :

Les modèles obtenus avec les algorithmes de type Random Forest sont efficaces sur notre premier dataset dans un but de classification, mais ne le sont pas sur notre deuxième dataset en régression.

Après cette première approche, on peut donc penser que ce type d'algorithme est plus efficace pour des problèmes de classification.