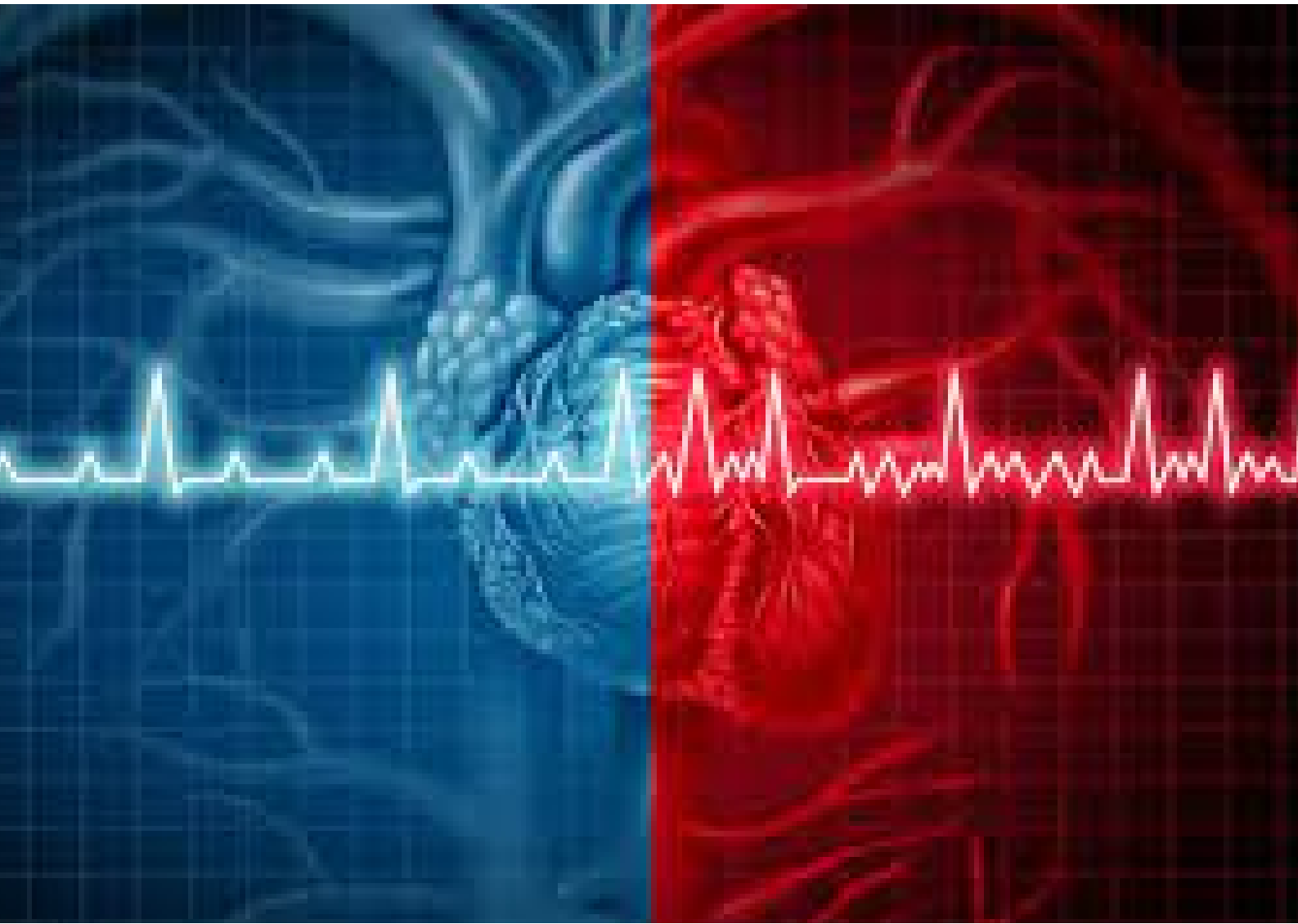
2023

RAPPORT OPTIMISATION D'UN MODÈLE

Encadrée par : Simplon , Sanofi

Lecarpentier Eden



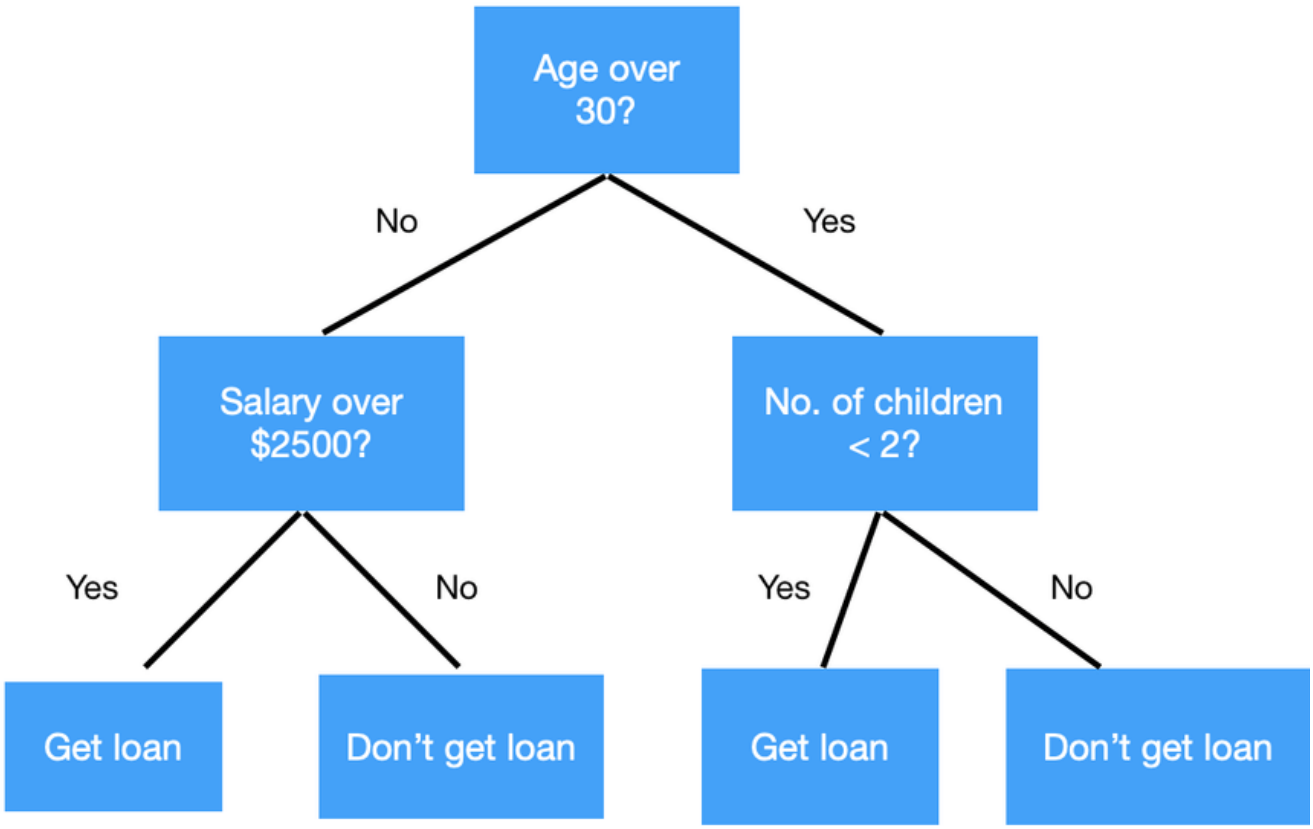
SOMMAIRE

MODÈLE DE RÉFÉRENCE

Le projet que je vais amélioré est sur la prédiction de probléme cardiaques .
On a une target contenant 2 classes Oui et Non .
C'est un modèle de classification en machine learning .
Mon modèle a était réalisée avec un RandomForest classifier .

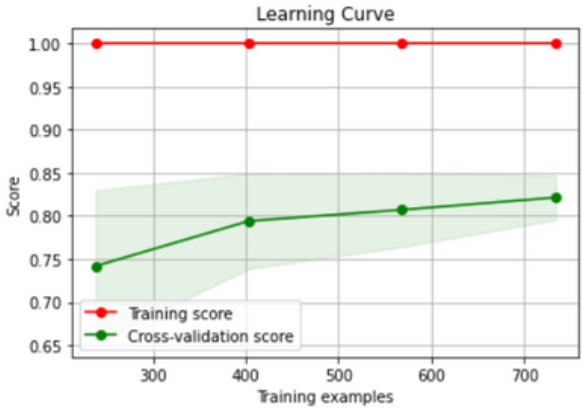
RandomForestClassifier

- Model de classification
- Le random forest fonctionne comme un arbres .
- On part du tron et on va dans les branches
- Cela permet de trouver toutes les corrélations entre les features



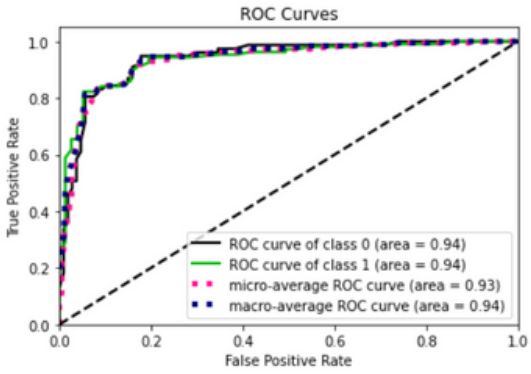
MODÈLE DE RÉFÉRENCE

LearningCurve



- La Learning Curve permet de voir les compétences du modèle sur le train(ligne rouge) et le test (ligne verte)
- Elle permet de trouver de l'overfitting ou Underfitting , le taux d'erreur possible
- L'espace des deux courbes montre le taux d'erreur possible.
- Les courbes sont probablement trop éloignées pour que le modèle sois efficace
- Les courbes risques également de se toucher

ROC CURVE



- La ROC Curve permet de vérifier l'entraînement du modèle, plus les lignes monte vite mieux seras l'entraînement.
- Les courbes représente 2 paramètres le taux de vrai et faux positif .
- Cette ROC Curve montre une bonne performance

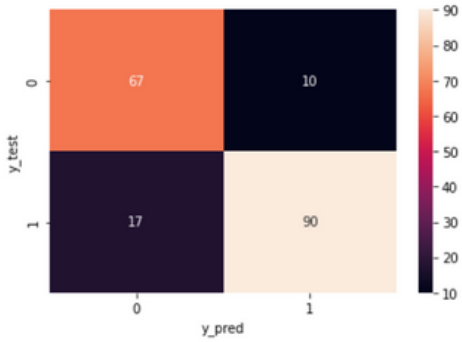
Classification Report

```
accuracy train : 1.000
accuracy test : 0.891
```

	precision	recall	f1-score	support
0	0.76	0.88	0.82	77
1	0.91	0.80	0.85	107
accuracy			0.84	184
macro avg	0.83	0.84	0.84	184
weighted avg	0.85	0.84	0.84	184

- La classification report nous permet de voir les performances sur les classes que l'on veut prédire grâce au metric .
- Nos classes : oui cette personne est à risque , non cette personne n'est pas à risque d'une crise cardiaque
- Le modèle de classification utilise les metrics : accuracy et recall
- L'accuracy n'est pas mauvaise par contre le recall la classe 0(Non) devrais se rapprocher le plus possible de 0 et l'inverse pour la classe 1(Oui)

Confusion Matrix



- La Confusion Matrix nous montre : les faux positifs , les faux négatifs , les vrais positifs , les vrais négatifs
- Vrai positif : devait avoir une crise cardiaque et là eux
- Faux positif : devait ne pas avoir de crise mai en a eux une
- Vrai négatif : devait ne pas avoir de crise et n'ont pas eux
- Faux négatif : devait ne pas avoir de crise est en ont eux une

OPTIMISATION DU MODÈLE

L'optimisation dun modèle se fait en plusieurs étapes pour se projet j'ai réalisée les choses suivantes: Trouvée de bon hyperparamètres , équilibrés les données .

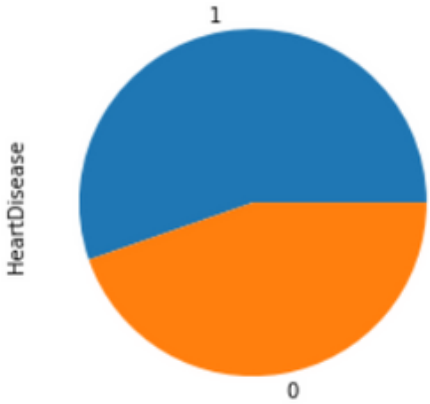
GridSearchCV

- GridSearchCV est utilisée pour trouver les meilleurs hyper paramètre du modèle a utilisée pour normalement améliorer les performances du modèle basé sur les données utilisée .
- Les hypers paramètres que je vais utiliser pour la logistic regression son les suivants : C , penalty ,solver .
- Les valeurs conseillé par le GridSearchCV sont : n_estimators , max_features , max_depth , criterion .
- Hyper Paramètre : ce sont des paramètres qui permettent de déterminer les valeurs nécessaires pour le paramètre du modèle .
- n_estimators : Indication sur le nombre d'arbres dans la forets j'en utilise 200
- max_features : Démontre le nombre maximal de features possible pour chaque arbres
- max_depth : Le nombre de split possible pour chaque arbre permet de lutté contre l'overfit , underfit
- criterion : Permet de vérifier la qualité d'un split en utilisant Gini Impurity qui permet de savoir a quel fréquences des éléments du set sont indiquer incorrectement .

SMOTE

Smote est utilisée lorsque la target n'est pas équilibrée , pour déterminer si on utilise le SMOTE ou non ou utilise un graphique :

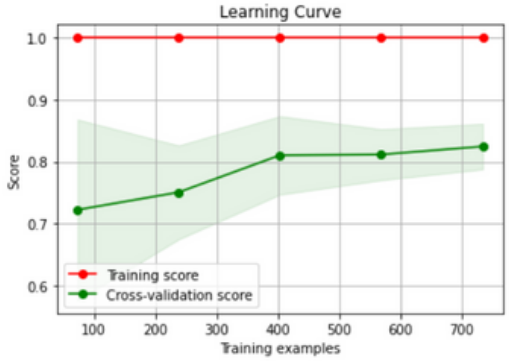
Le graphique suivant nous montre que les 2 class ne sont pas égale .
Les 2 classes étant si oui ou non on auras une crise cardiaque .



L'utilisation du SMOTE permettras de crée des données synthétiques c'est a dire des données qui différé légèrement des données originel . Cela permet d'avoir plus de données d'une des class qui en manque donc équilibré les données pour essayer d'améliorée le modèle .

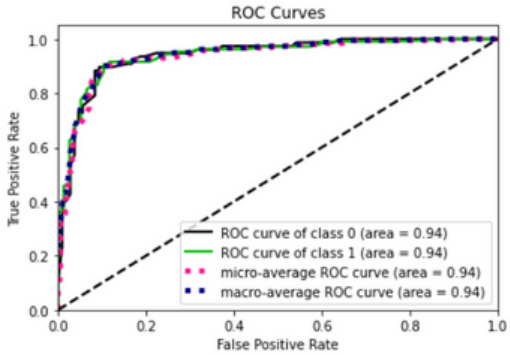
MODÈLE OPTIMISÉ PERFORMANCE

LearningCurve



- La Learning Curve est toujours en overfitting
- La courbe test s'entraîne légèrement mieux

ROC CURVE



- La Roc Curve est resté relativement similaire a la précédente .
- On peut s'apercevoir que les lignes montes de façon régulière contrairement à la précédente Roc Curve .

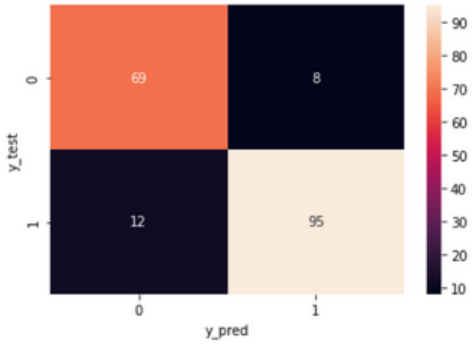
Classification Report

```
accuracy train : 0.857
accuracy test : 0.837
```

	precision	recall	f1-score	support
0	0.80	0.87	0.83	77
1	0.90	0.84	0.87	107
accuracy			0.85	184
macro avg	0.85	0.86	0.85	184
weighted avg	0.86	0.85	0.85	184

La class 1 a augmenter de 0.04 passant de 0.80 à 0.84.
La classe 0 a baisser de 0.01 passant de 0.88 à 0.87.

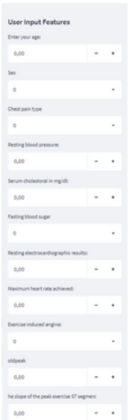
Confusion Matrix

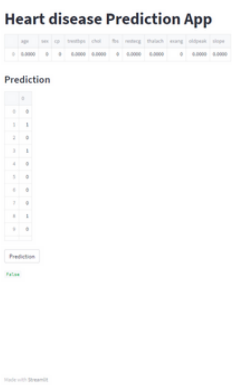


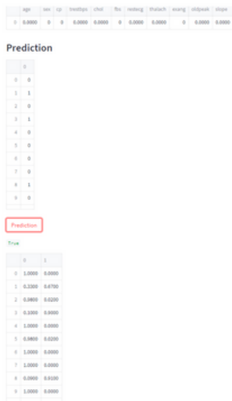
La confusion matrix a plus de True positive est le reste a diminué .

TEST UNITAIRE SUR L'APPLICATION

Application







- Ajout d'un boutton
- Le bouton permet de montré les prédiction une fois le bouton cliqué est non avant

Test unitaire

1. Vérification que mon dataset est le bon nombres de colonnes .
2. Vérification que les prédictions de l'application reste entre 0 et 1 si mon modèle aller au dessus ou en dessous de ces nombres cela voudrais dire que mon modèle est défectueux .
3. Vérification que le nombre de prédiction sois égale au nombre d'échantillions donner

```
test_main.py ...                                                    [100%]
======================== 3 passed in 1.56s ========================
```

CONCLUSION

Le projet à eux du mal à être optimisée probablement du à un grand manque de donnée car la base de données est petite.

Le modèle pourrais être encore améliorée en ajoutant de nouvelle données et de la diversité dans les données.

Cette diversité des données est nécessaire pour que le modèle sois le plus performants possible sur des données qu'il n'a jamais vue .

Le problème le plus voyant est l'overfitting même après l'avoir optimisée les conclusion amenée précédemment aiderait le modèle a s'amélioré d'avantages .