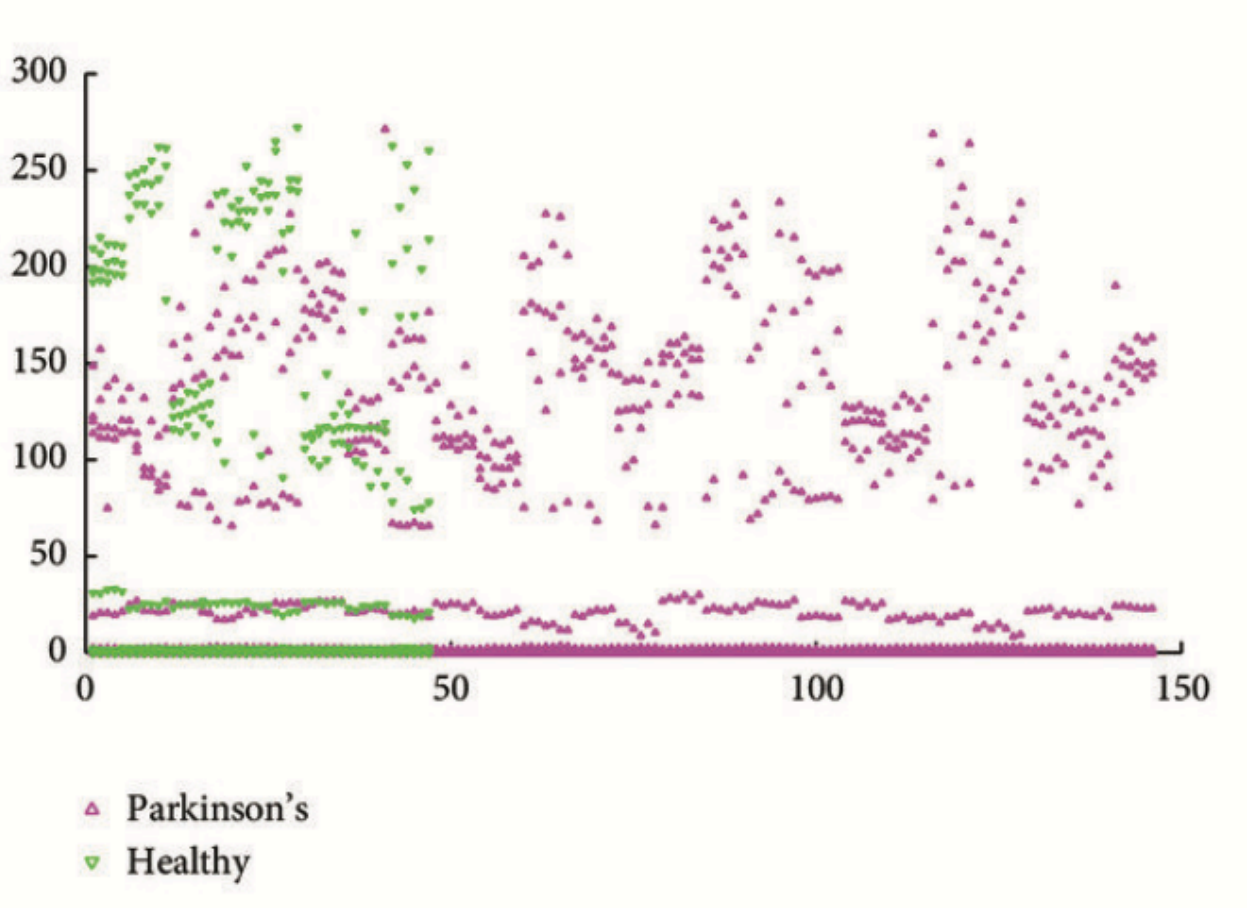
université
PARIS-SACLAY

RAPPORT PROJET INTRODUCTION A L'APPRENTISSAGE STATISTIQUE DE

SARAH BOUNDAOUI. & AHMED BAAROUN.



1.INTRODUCTION :

Ce projet est un projet à but ouvert qui consiste à choisir un dataset libre sur lequel nous allons devoir choisir un objectif, raffiné le data et accomplir l'objectif via un modele vu en cours.

1.1 CHOIX D'UN DATASET :

Pour se faire, nous avons choisi le dataset sur la maladie du PARKINSON .

Lien vers le dataset : https://www.kaggle.com/datasets/sagarbapodara/parkinson-csv

1.2 OBJECTIF DU PROJET :

L'objectif de notre projet est d'approfondir nos connaissances acquises au cours d'introduction à l'apprentissage statistique et aussi de découvrir de nouveaux algorithmes.
Avec ce type de donnée on peut prédire si une personne est malade ou pas et ça en utilisant plusieurs modèles de classifications, et une comparaison entre ces modèles pour savoir lequel est le plus performant.Et pour arriver à cet objectif on est passé par plusieurs étapes de traitement de nos données : récupération des données , exploration du dataset, nettoyage des données, corrélation, PCA...

2.ETAPES DU PROJET

2.1 RECUPERATION DES DONNÉES

	name	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ	Jitter:DDP	MDVP:Shimmer	...	Shir
0	phon_R01_S01_1	119.992	157.302	74.997	0.00784	0.00007	0.00370	0.00554	0.01109	0.04374	...	
1	phon_R01_S01_2	122.400	148.650	113.819	0.00968	0.00008	0.00465	0.00696	0.01394	0.06134	...	
2	phon_R01_S01_3	116.682	131.111	111.555	0.01050	0.00009	0.00544	0.00781	0.01633	0.05233	...	
3	phon_R01_S01_4	116.676	137.871	111.366	0.00997	0.00009	0.00502	0.00698	0.01505	0.05492	...	

Nous avons commencé par récupérer le dataset pour l'afficher, nous avons remarqué qu'il y'avait 195 lignes et 24 colonnes traitant les fréquences en Hz.
On a regardé le type de valeur de chaque colonnes, ensuite nous avons vérifié si le dataset proposé est sans valeur nulle et sans valeurs aberrantes. Nous avons aussi affiché des données générales comme la moyenne, les quartiles pour ensuite trouver des corrélations plus facilement.

	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ	Jitter:DDP	MDVP:Shimmer	MDVP:Shimmer(dB)
count	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000
mean	154.228641	197.104918	116.324631	0.006220	0.000044	0.003306	0.003446	0.009920	0.029709	0.282251
std	41.390065	91.491548	43.521413	0.004848	0.000035	0.002968	0.002759	0.008903	0.018857	0.194877
min	88.333000	102.145000	65.476000	0.001680	0.000007	0.000680	0.000920	0.002040	0.009540	0.085000
25%	117.572000	134.862500	84.291000	0.003460	0.000020	0.001660	0.001860	0.004985	0.016505	0.148500
50%	148.790000	175.829000	104.315000	0.004940	0.000030	0.002500	0.002690	0.007490	0.022970	0.221000
75%	182.769000	224.205500	140.018500	0.007365	0.000060	0.003835	0.003955	0.011505	0.037885	0.350000
max	260.105000	592.030000	239.170000	0.033160	0.000260	0.021440	0.019580	0.064330	0.119080	1.302000

2.2 EXPLORATION DU DATASET :

Nous avons commencé par vérifier le nombre de malades et non malades et nous avons remarqué que la personne malade a le statut 1 et la personne non malade a le statut 0, à l'aide d'un diagramme.

Ensuite nous avons visualisé la distribution de chaque donnée selon différents diagrammes, pour pouvoir vérifier l'utilisation du modele sur cette population de donnée, en effet on remarque plus au moins que toutes les colonnes suivent donc la courbe gaussienne qui nous montre que le modele s'appliquera de manière efficaces à cette plage de données.

Figure 1 : Densité PPE.

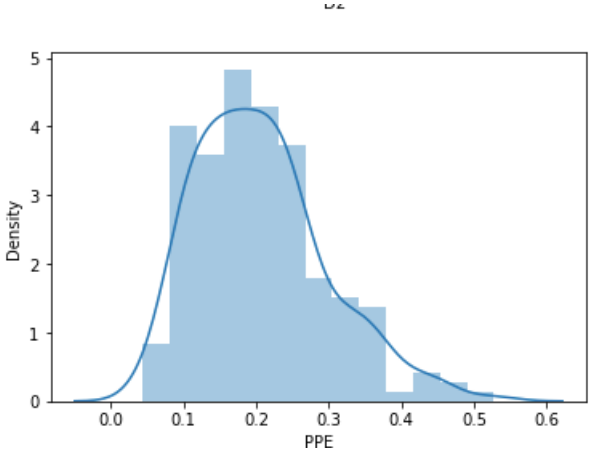
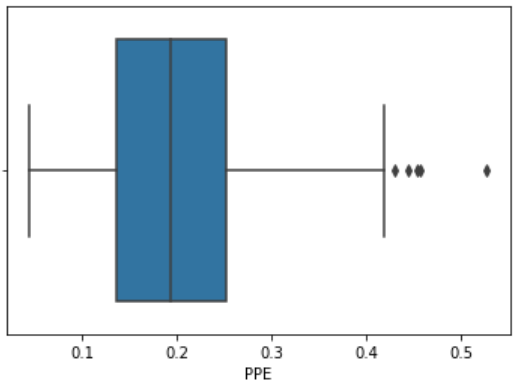


Figure 2 : Box plot PPE



Ensuite on a fait la moyenne de chaque donnée par rapport au statut et on a affiché la corrélation de chaque colonne, nous avons aussi utilisé la heatmap qui nous permet de voir les différentes corrélations ainsi que les features inutiles.Donc par extension la création d'un modele efficace.

2.3 CREATION DU MODELE:

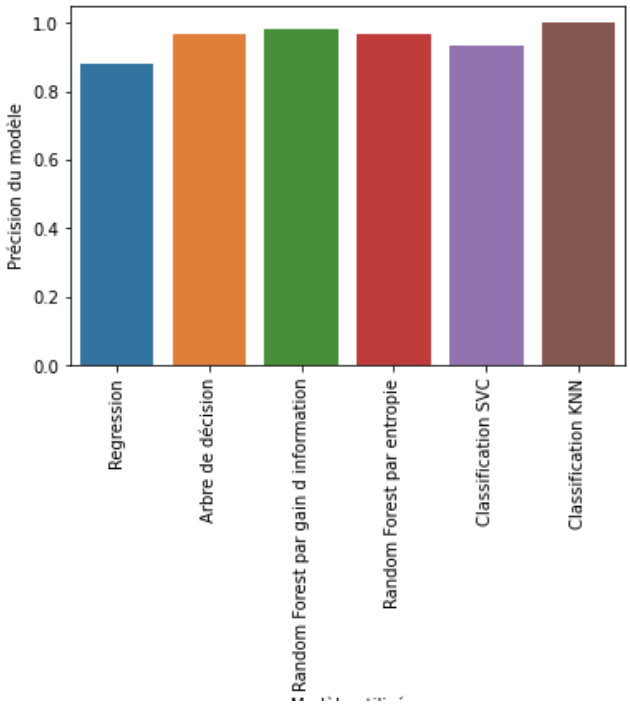
On commence par utiliser l'OverSample et MinMaxScaler, l'un permet de créer des échantillons aléatoires et l'autre met à l'échelle les données entre -1 et 1 pour une utilisation simplifié de l'algorithme k-near neighbour ainsi qu'un PCA avec une variance de 95% pour avoir encore une meilleure précision lors de l'utilisation du modele.

Ensuite on sélectionne beaucoup de classifieur pour avoir des différents scores pour ensuite choisir le modele le plus performant selon son score.

Nous avons aussi réalisé des courbes de ROC qui permettent de montrer les performances du modele, Le modele le plus performant est k nearest neighbors car il a le meilleur score

	Modèle utilisé	Précision du modèle
0	Regression	0.881356
1	Arbre de décision	0.966102
2	Random Forest par gain d information	0.983051
3	Random Forest par entropie	0.966102
4	Classification SVC	0.932203
5	Classification KNN	1.000000

AxesSubplot(0.125,0.125;0.775x0.755)



CONCLUSION:

Le projet que nous avons réalisé nous a permis d'approfondir nos connaissances en machine learning, et de bien comprendre la classification, et ses différents modèles, pendant la réalisation de ce projet nous avons rencontré des difficultés comme l'obligation d'installer certaine librairies.Globalement ce projet était trés passionnant et enrichissant.