Réunion Scrum 4

Groupe de Carole

random over sample = bonne piste pour gérer les populations inégales

variables corréleées: supprimer des variables pour le ML ? attnete réponse client

ML: picaret (?) pour faux pos et faux neg

pour apercevoir prérésultat des modèles

recommandations de Souf:

encoder

scaling : en tester plusieurs

pour chaque scaler, faire tous les algos ⇒ comparaison /scaler et algo

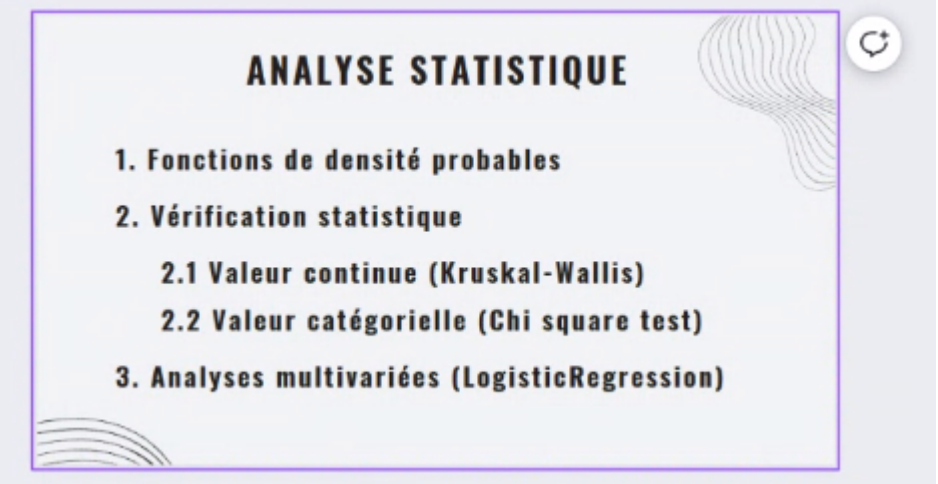
Groupe de Jonathan Inovie

ont remplacé Nan et 0 par médiane et moyenne selon (a)symétrie

Medic analytics

mise en commun avec les autres groupes remote en binomes/ trinomes

Groupe de Jesus Medi Mind



test ANOVA ?

kruskl un peu similaire à Anova

utiliser krukal wallis quand groupes pas égaux (quand pas la normalité pour les variables)

KW utilisable pour les valeurs ordinales

KW est une alternative non paramétrique à Anova

Anova: pour comparer les moyennes de 3 groupes ou plus. échantillons indépendants et échantillons doivent avoir variance égales et pop normalement ditribuées (distribution gaussienne)

uniquement pour variables num

Quand la distrib est normale, on devrait utiliser Anova

chi square test pour les variables catég

analyse multivariée avec regression log ??????????

Chronics analytics Lucas

acp analyse de comparaison des variables (pour distribution): Souf: à prendre avec beaucoup de pincettes mais intéressant !

Groupe de Mireille Data health

Groupe de Thibault Romao Génome

on attaqué le ML

bibliothèque lazy predict: réponse de Souf: lazy predict est super lazy et les résultats obtenus semblent très élevés. ici, on est dans une situation d'overfitting.
OK juste pour identifier des modèles mais il en manque.
Il faut partir sur une dizaine de classi par mie
Soufiane préfère picaret à Lazy

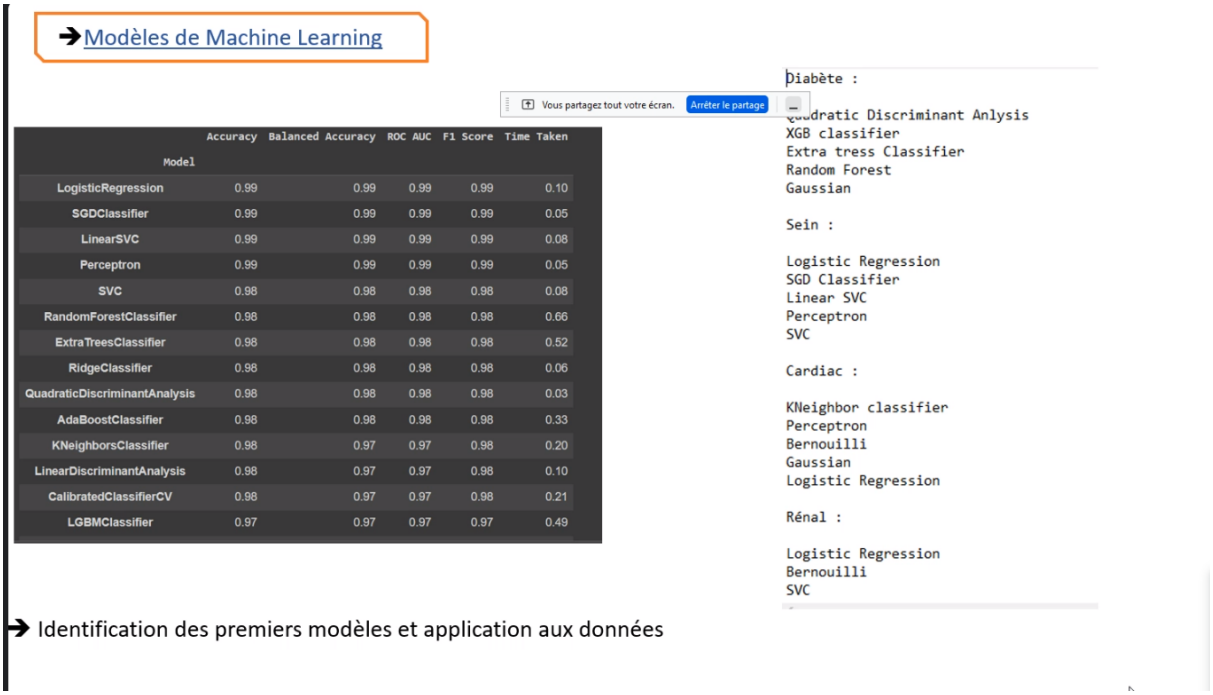
Boost

???

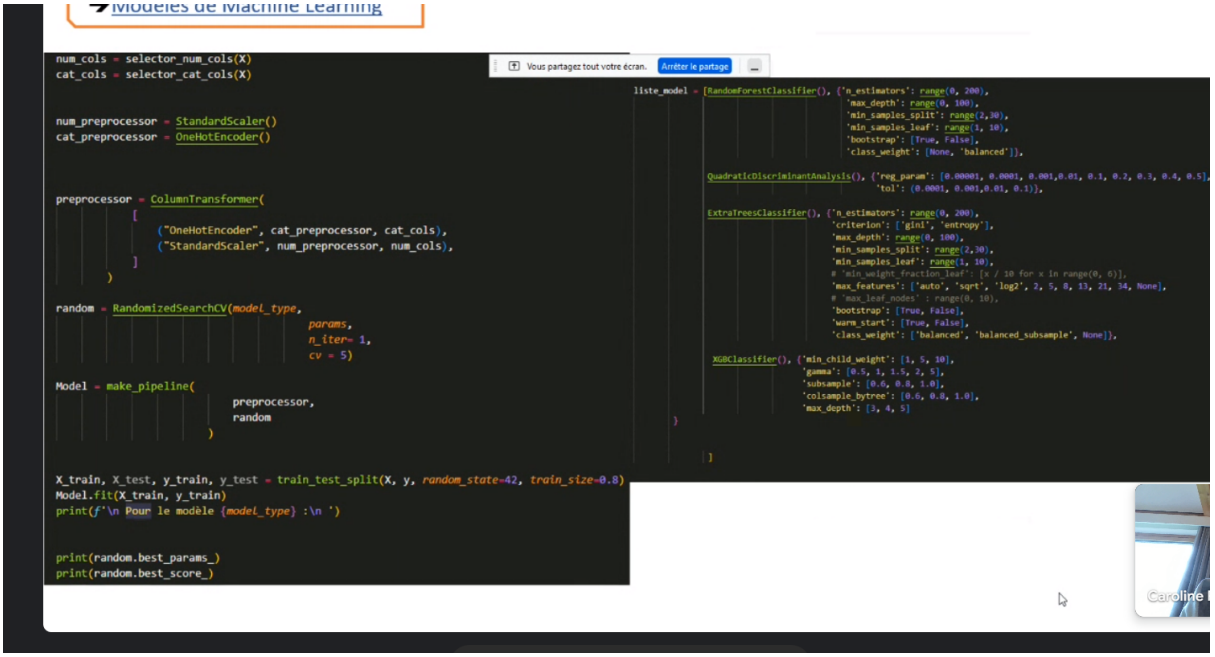
réseau de neurones

modèles d'ensemble

random forest, tree



Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
LogisticRegression	0.99	0.99	0.99	0.99	0.10
SGDClassifier	0.99	0.99	0.99	0.99	0.05
LinearSVC	0.99	0.99	0.99	0.99	0.08
Perceptron	0.99	0.99	0.99	0.99	0.05
SVC	0.98	0.98	0.98	0.98	0.08
RandomForestClassifier	0.98	0.98	0.98	0.98	0.66
ExtraTreesClassifier	0.98	0.98	0.98	0.98	0.52
RidgeClassifier	0.98	0.98	0.98	0.98	0.06
QuadraticDiscriminantAnalysis	0.98	0.98	0.98	0.98	0.03
AdaBoostClassifier	0.98	0.98	0.98	0.98	0.33
KNeighborsClassifier	0.98	0.97	0.97	0.98	0.20
LinearDiscriminantAnalysis	0.98	0.97	0.97	0.98	0.10
CalibratedClassifierCV	0.98	0.97	0.97	0.98	0.21
LGBMClassifier	0.97	0.97	0.97	0.97	0.49



Souf: un encodeur peut suffire mais il faut tester pls scaler avec pls algos

Q&A

Q: Peut on valider le fait qu'on supprime toutes les variables dont p value>0.05 ?

A: NON, il faudrait avoir des médecins avec nous pour décider cela

Q: foie: déséquilibre des classes: utilisation random over sampler ?

A: oui, c'est une bonne méthode

il faiut aussi rééquilibrer pour Breast et diabète

Q: Breast: à partir de qd corrélation assez forte pour supprimer variable ?

A: ne travailler que avec les mean
radius / perimetre / aire : on en garde un seul

0.8: fortement corrélé

.99: ça fait presque doublon

Q: rein hemo créat corrélés 87%

A: client veut garder les 2

Q Diabete skin thinkness à garder ?

A: oui, tester les ML avec ou sans ?