

Réunion Scrum 4

Groupe de Carole

random over sample = bonne piste pour gérer les populations inégales

variables corrélées: supprimer des variables pour le ML ? attendre 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 \Rightarrow comparaison /scaler et algo

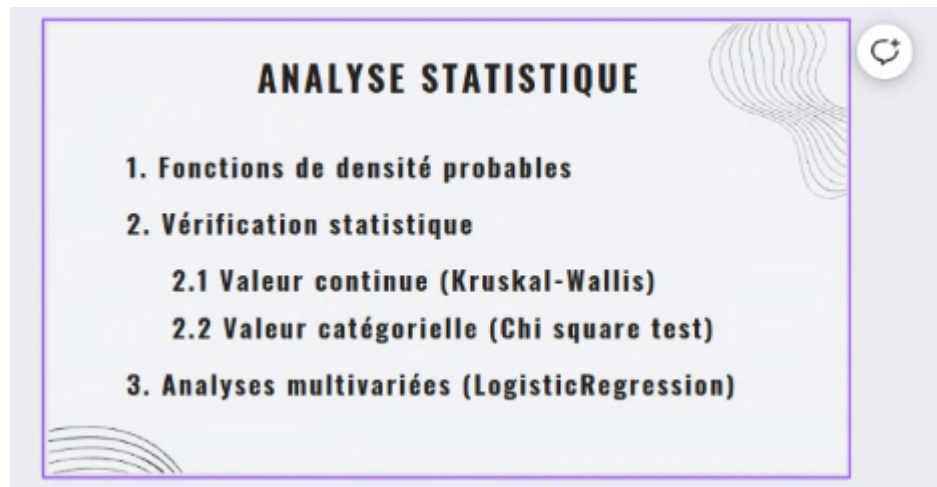
Groupe de Jonathan Inovie

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 ?

kruskal un peu similaire à Anova

utiliser kruskal 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 distribuées (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

→ [Modèles de Machine Learning](#)

Vous partagez tout votre écran. [Arrêter le partage](#)

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

Diabète :

Quadratic Discriminant Anlysis
XGB classifier
Extra tress Classifier
Random Forest
Gaussian

Sein :

Logistic Regression
SGD Classifier
Linear SVC
Perceptron
SVC

Cardiac :

KNeighbor classifier
Perceptron
Bernouilli
Gaussian
Logistic Regression

Rénal :

Logistic Regression
Bernouilli
SVC

→ Identification des premiers modèles et application aux données

Réunion Scrum 4

3

7 MOODELES DE MACHINE LEARNING

```

num_cols = selector_num_cols(X)
cat_cols = selector_cat_cols(X)

num_preprocessor = StandardScaler()
cat_preprocessor = OneHotEncoder()

preprocessor = ColumnTransformer(
    [
        ("OneHotEncoder", cat_preprocessor, cat_cols),
        ("StandardScaler", num_preprocessor, num_cols),
    ]
)

random = RandomizedSearchCV(model_type,
                             params,
                             n_iter=1,
                             cv=5)

Model = make_pipeline(
    preprocessor,
    random
)

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42, train_size=0.8)
Model.fit(X_train, y_train)
print(f'\n Pour le modele (model_type) :\n ')

print(random.best_params_)
print(random.best_score_)

liste_model = [
    RandomForestClassifier(),
    QuadraticDiscriminantAnalysis(),
    ExtraTreesClassifier(),
    NBClassifier()
]

```

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

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 medecins avec nous pour decider cela

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

A: oui, c'est une bonne methode

il faut aussi reequilibrer pour Breast et diabete

Q: Breast: a partir de qd correlation 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 ?