

Python for data analysis

IBO1

BOTRAN

Marvin

Jeu de données

PPG-DaLiA

Le dataset est tiré d'une expérimentation visant à obtenir des données sur l'activité d'une **population** de **15** personnes âgées de **20 à 40** ans.

Le groupe a réalisé simultanément un total de 7 activités sportives.

On suit leurs déplacements à l'aide de deux appareils :
une montre et un appareil se positionnant sur le torse.

Ces matériels utilisent la **PPG** permettant d'évaluer le **rhythme cardiaque** en éclairant les veines et d'un **accéléromètre** pour enregistrer les **déplacements**.





Problèmes



Objectif : Prédire l'activité d'un individu en fonction des observations

Pour l'atteindre, nous avons utilisé la technique du divide and conquer.

Diviser un problèmes complexe en le divisant en plusieurs sous problèmes.

1

Réunir l'ensemble des données dans un format exploitable

2

Trouver le modèle le plus adapté

3

Interpréter les résultats

Réunir l'ensemble des données
dans un format exploitable.

Une fois les fichiers au format pkl importés, je les ai visualisés avant de pouvoir les exploiter.

J'ai découvert que les objets chest et wrist étaient eux-mêmes des dict contenant des tableaux.

Pour pouvoir exploiter toutes ces données dans un dataframe contenant des données de même dimension, il a fallu recréer des listes pour chacune des sous-entrées de chest et de wrist.

Le deuxième problème majeur lié au dataset est la différence de longueur de colonnes due aux différentes fréquences d'enregistrement.

Le cas le plus complexe est celui du label car il faut ajouter manuellement les dernières valeurs.

Test des modèles

Le test des modèles à été fait une première fois sur un seul individu avant de prendre l'ensemble du dataset. On peut voir avec le classement des features que '**sport**' (régularité de la pratique du sport) influe assez sur les résultats. On pourrait tenter de grouper les individus par cette feature avant d'entrainer les modèles

La problématique portant sur de la classification, les modèles testés pour répondre à la problématique :

- DecisionTreeRegressor
- RandomForestRegressor

- MLPClassifier
- DecisionTreeClassifier
- RandomForestClassifier
- ExtraTreesClassifier

Après interpretation des resultats, on peut voir que le MLP n'etait pas adapté à l'étude.

La question qui se pose est l'objectif de déploiement de l'outil, deux possibilités s'offrent à nous :

- Etre capable de reconnaitre l'activite d'un grand nombre de personnes avec une acquite approximative
- Etre capable de reconnaitre l'activite d'un type de personne tres defini grace a un grand nombre de donnees sur une personne

Conclusion

Ce premier projet de machine learning et de python m'a permis d'expérimenter et mettre en œuvre les différents thèmes vus ce semestre en python for data science.