

# Analysis of the PPG- DaLiA Data Set

---

GUILLAUME GERARD



# But de l'étude

Prédire l'activité de la personne

Ce dataset a été réalisé dans le but de fournir à la communauté un dataset assez grand à analyser pour pouvoir améliorer la détection de via Photoplethysmography.

Nous utiliserons ici ce Dataset non pas pour prédire l'activité cardiaque de la personne en fonction de son activité mais l'inverse.

# Description du Dataset

- Le but de l'étude est de déterminer l'activité d'une personne.
- Le dataset utilisé afin de répondre à cette problématique est le dataset PPG Field Study Dataset. Il a été réalisé dans le but de fournir un Dataset de taille suffisantes pour améliorer la Photoplethysmography
- Ce dataset permet de répertorier les particularités cardiaques de différents individus dans différentes conditions. Ces individus sont tous différents, certains sont sportifs, d'autres sédentaires.
- Dans ce dataset, nous avons différentes informations telles que les particularités cardiaques en fonction d'activités réalisées comme monter des escaliers, conduire, faire du vélo, etc.
- Nous avons aussi des informations sur les caractéristiques personnelles des individus de l'expérience telles que leur Age, leur taille, leurs poids, etc.
- Deux Dispositifs ont été utilisés pour récolter les données :
  - RespiBAN sur le torse
  - Empatica E4 au poignet
- Ces deux dispositifs nous génère entre autre des données relatives à l'activité cardiaque tel que l'ecg ou la différence de couleur entre le sang oxygéné et non oxygéné.
- Les données des appareils n'étant pas récupérées à la même fréquence, il faudra adapter le traitement des données

# Réflexions sur la question posée

Plusieurs problèmes étaient à résoudre avant même d'analyser les données :

- Quelles données sont utiles ? J'ai fait le choix d'importer toutes les données qui n'étaient pas indiquées comme inutiles dans la description du Dataset (comme TEMP/Rpeak/etc)
- Les données étant enregistrées sur des fréquences différentes, comment les échelonner ? J'ai échelonné par rapport au nombre de données présentes dans la catégorie activité

## Choix de réalisation

- Pour traiter les données, j'ai choisi d'échelonner toutes les données par rapport aux activités. Les activités nous étant fournis par sujet dans la colonne « activity ». Nous savons toutes les 15 secondes l'activité réalisée
- Il a donc fallu trouver les rapports entre les données provenant entre autre des différents appareils et les données indiquées dans activity puis les échelonner (Pour la valeur de l'ecg récupéré par la machine RespiBan, on prend 1 valeur sur 175 par exemple)
- Ufois mon Dataset formé, je l'ai exporté en csv pour ensuite pouvoir l'utiliser si besoin dans l'API par la suite. Il en est de même pour le modèle de RandomForestClassifier.

# Conclusion

Avec en moyenne 90% de confiance dans les données prédites, nous pouvons dire que le dataset permet bien d'estimer l'activité par rapport aux particularités cardiaques et physique des personnes.

Les étapes suivantes sont:

- Jouer avec les paramètres pour augmenter la confiance
- Transformer le modèle en API django Rest

Finalement, ce projet m'a permis d'accroître les compétences déjà acquises au cours de l'année en python.