

# Classification d'activités

## PPG-DaLiA

# Le dataset

- Données provenant de 15 sujets portant des capteurs physiologiques et de mouvement
- 27 GB de données, pour 11 attributs
- 8 activités différentes

# Réflexions sur le dataset

- Le dataset a été pensé pour un travail de régression, mais peut être utilisé pour un cas de classification.
- Il s'agit ici de prédire l'activité du sujet en fonction des mesures relevées par les capteurs
- Les capteurs du torse sont réglés pour faire des mesures selon une fréquence de 700Hz, alors que les capteurs du poignet les font pour des fréquences de 64Hz, 32Hz et 4Hz.
- Deux possibilités pour le preprocessing du dataset :
  - Augmenter le dataset pour synchroniser les mesures
  - Diminuer le dataset pour synchroniser les mesures
- Choix pris : Diminuer le dataset en raison d'un manque de puissance de calcul (ordinateur vieillissant) et le dataset est trop volumineux pour être utilisé sur Google Collab
- On a filtré toutes les données qui ne se calquent pas sur une fréquence de mesure de 4Hz
- Les données de chaque patient sont accessibles via un fichier pickle, données que l'on transfère dans un DataFrame pour faciliter leur manipulation

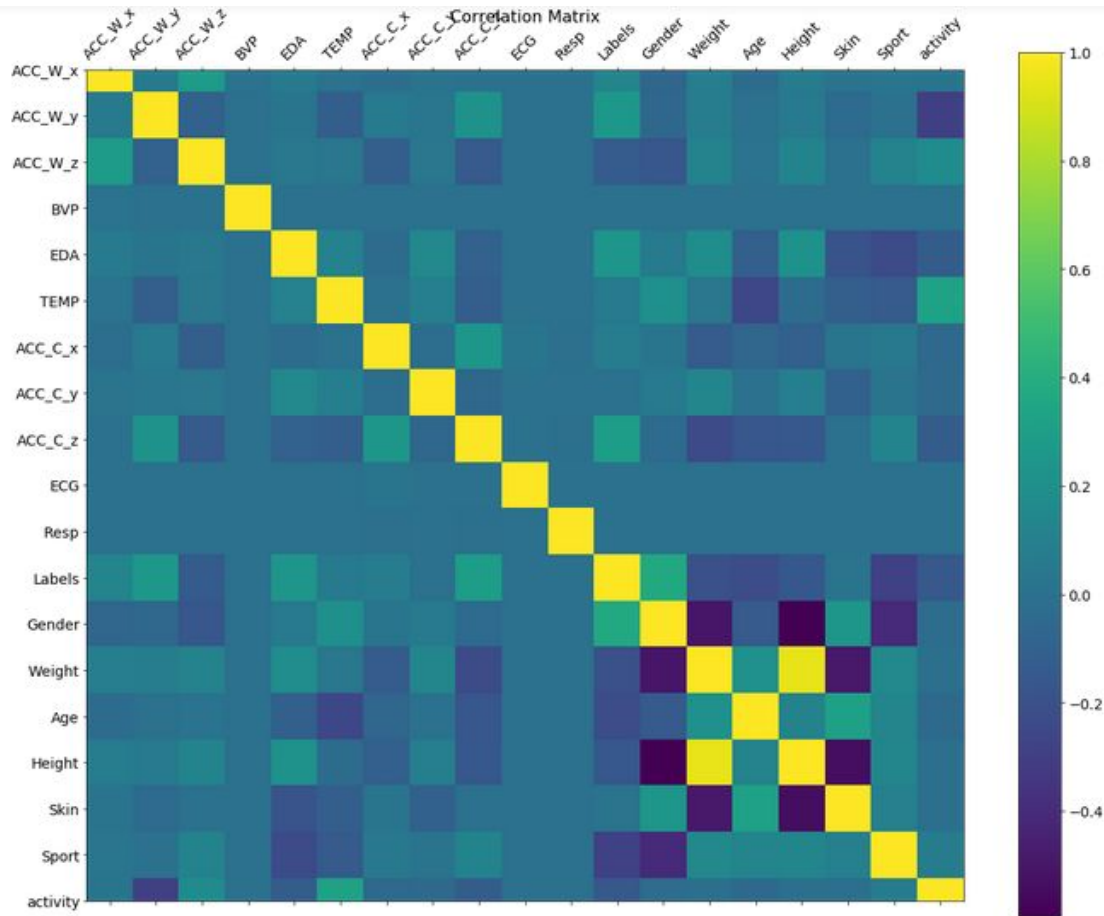
# Les variables

- Intuitivement, les valeurs de l'accélération, du rythme cardiaque, de la température devraient être explicatives de l'activité effectuée par le sujet.
  - L'attribut "Accélération" est un tableau avec les valeurs de l'accélération selon les axes x, y et z. On divise ainsi ce tableau en trois valeurs distinctes, nous avons ainsi 3 nouveaux attributs correspondant à l'accélération mesurée sur le torse, et 3 nouveaux attributs correspondant à l'accélération mesurée sur le poignet.
  - On intègre les informations relatives au patient (sa taille, son poids, son âge etc.) car ces informations pourraient être significatives.
  - Pour la liste exhaustive des variables, consulter le ReadMe joint au rendu.
- On calcule la matrice de corrélation du DataFrame pour déterminer quelles variables sont les plus explicatives.

# Les variables

Voici une matrice en gradient de couleur permettant de visualiser la corrélation des différentes variables.

→ Les variables les plus explicatives sont **la température** et **l'accélération** !



# La modélisation

- On teste plusieurs modèles pour la classification, avec leurs paramètres par défaut :

Gaussian Process		precision	recall	f1-score	support
	0.0	0.53	0.61	0.57	28077
	1.0	0.91	0.78	0.84	7279
	2.0	0.64	0.44	0.52	5199
	3.0	0.48	0.02	0.04	3682
	4.0	0.55	0.73	0.63	5472
	5.0	0.87	0.41	0.55	10854
	6.0	0.50	0.73	0.60	21919
	7.0	0.45	0.55	0.49	7446
	8.0	0.60	0.36	0.45	13588
accuracy				0.57	103516
macro avg		0.62	0.51	0.52	103516
weighted avg		0.60	0.57	0.55	103516

## Decision Tree

	precision	recall	f1-score	support
0.0	0.49	0.80	0.61	28077
1.0	0.93	0.87	0.90	7279
2.0	0.94	0.24	0.39	5199
3.0	0.00	0.00	0.00	3682
4.0	0.95	0.52	0.67	5472
5.0	0.78	0.65	0.71	10854
6.0	0.60	0.79	0.68	21919
7.0	1.00	0.00	0.00	7446
8.0	0.83	0.53	0.65	13588
accuracy			0.62	103516
macro avg	0.72	0.49	0.51	103516
weighted avg	0.69	0.62	0.59	103516

## Random Forest

	precision	recall	f1-score	support
0.0	0.70	0.72	0.71	28077
1.0	0.88	0.95	0.92	7279
2.0	0.79	0.64	0.71	5199
3.0	0.79	0.44	0.56	3682
4.0	0.86	0.87	0.86	5472
5.0	0.85	0.89	0.87	10854
6.0	0.80	0.87	0.83	21919
7.0	0.71	0.60	0.65	7446
8.0	0.82	0.81	0.82	13588
accuracy			0.78	103516
macro avg	0.80	0.75	0.77	103516
weighted avg	0.78	0.78	0.78	103516

## Neural Net

	precision	recall	f1-score	support
0.0	0.50	0.48	0.49	28077
1.0	0.53	0.58	0.56	7279
2.0	0.50	0.57	0.54	5199
3.0	0.41	0.44	0.42	3682
4.0	0.81	0.72	0.76	5472
5.0	0.68	0.57	0.62	10854
6.0	0.52	0.72	0.61	21919
7.0	0.38	0.36	0.37	7446
8.0	0.42	0.24	0.30	13588
accuracy			0.52	103516
macro avg	0.53	0.52	0.52	103516
weighted avg	0.52	0.52	0.52	103516

## AdaBoost

	precision	recall	f1-score	support
0.0	0.84	0.84	0.84	28077
1.0	0.97	0.99	0.98	7279
2.0	0.87	0.74	0.80	5199
3.0	0.85	0.86	0.86	3682
4.0	0.92	0.93	0.93	5472
5.0	0.94	0.95	0.95	10854
6.0	0.92	0.94	0.93	21919
7.0	0.75	0.76	0.76	7446
8.0	0.96	0.95	0.96	13588
accuracy			0.89	103516
macro avg	0.89	0.88	0.89	103516
weighted avg	0.89	0.89	0.89	103516



# La modélisation

- On peut observer que RandomForest et AdaBoost obtiennent les meilleurs résultats.

→ En modifiant les paramètres de RandomForest, on arrive au résultat suivant

→ `n_estimator = 50` :

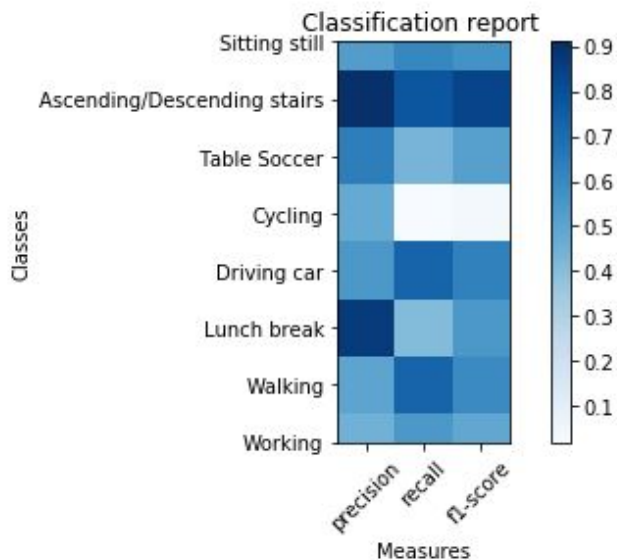
Random Forest

	precision	recall	f1-score	support
0.0	0.97	0.99	0.98	28077
1.0	1.00	1.00	1.00	7279
2.0	1.00	0.98	0.99	5199
3.0	0.99	0.96	0.98	3682
4.0	0.99	0.99	0.99	5472
5.0	0.99	1.00	0.99	10854
6.0	0.99	0.99	0.99	21919
7.0	0.98	0.98	0.98	7446
8.0	1.00	0.99	0.99	13588
accuracy			0.99	103516
macro avg	0.99	0.98	0.99	103516
weighted avg	0.99	0.99	0.99	103516

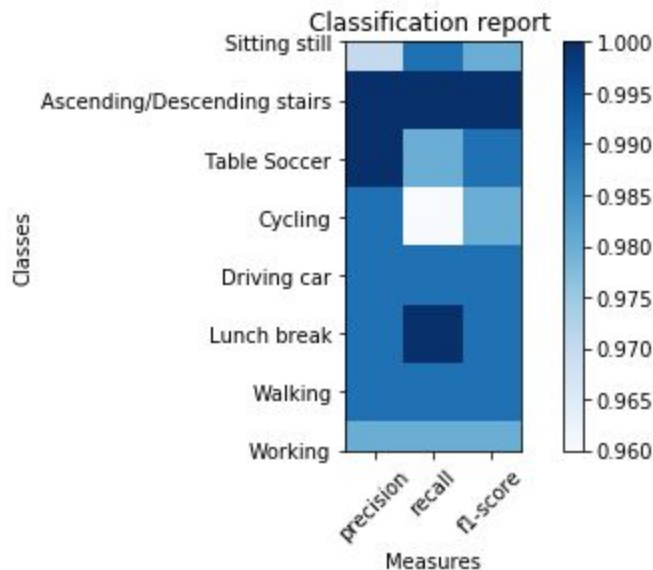
# La modélisation

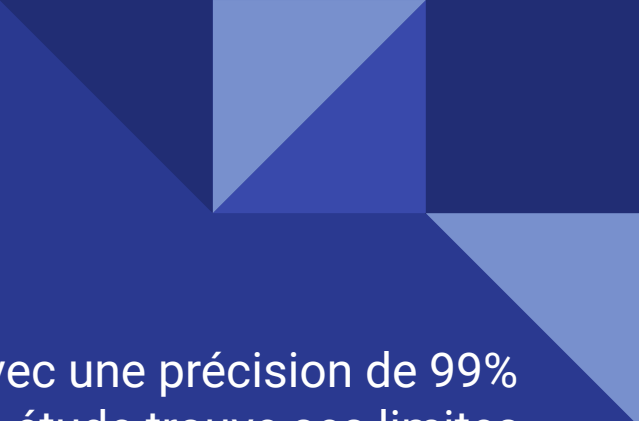
Afin de mieux visualiser les résultats de prédiction, on peut se référer aux gradients de couleurs suivants :

Gaussian Process



Random Forest





En conclusion, l'activité des sujets peut être prédite avec une précision de 99% à l'aide du classifieur RandomForest. Néanmoins, cette étude trouve ses limites dans le fait que le dataset originel a dû être diminué pour pouvoir être exploité avec les ressources disponibles.

Il vaudrait la peine de tester ce même modèle sur le dataset originel à l'aide d'une plus grande puissance de calcul.