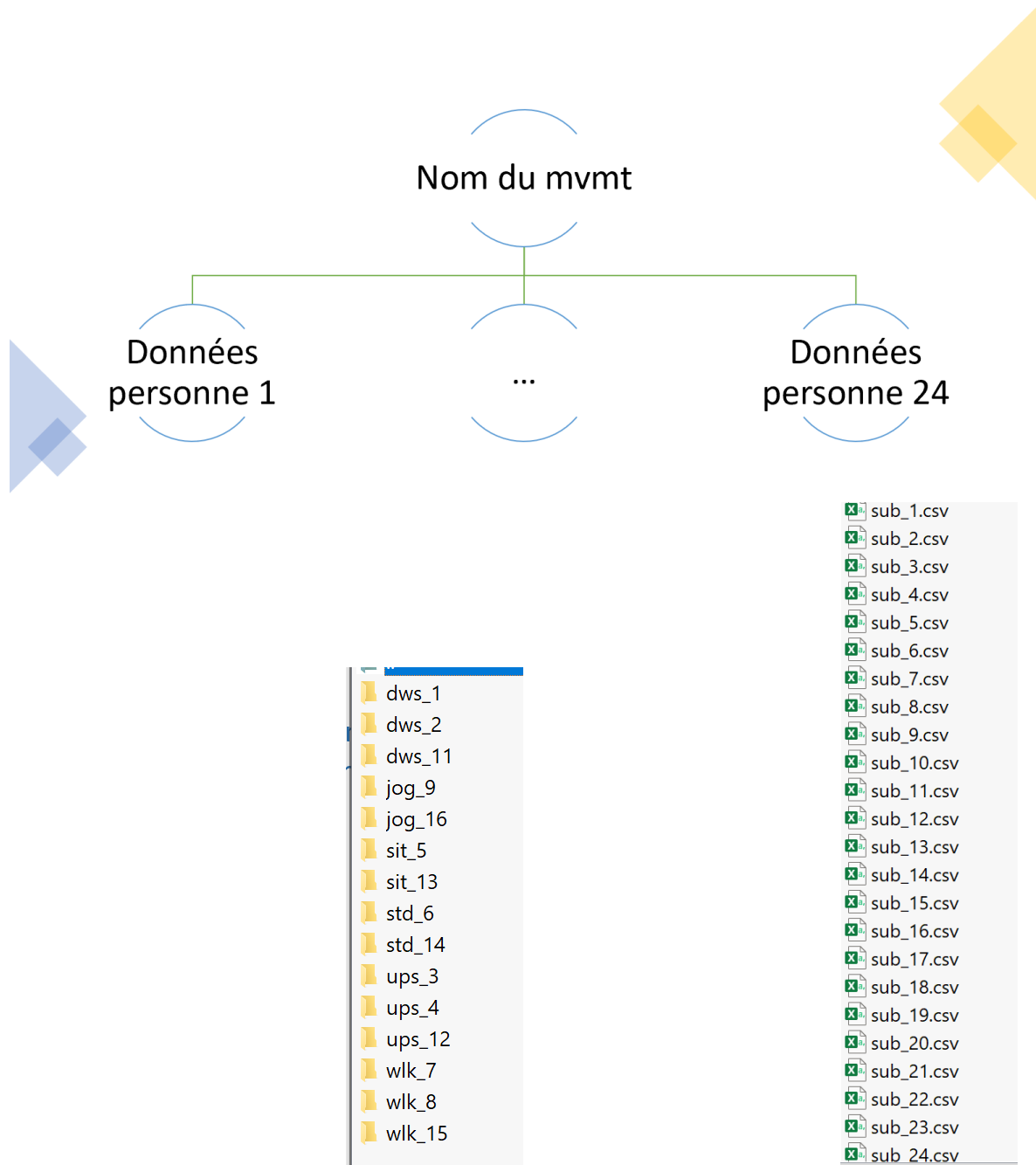


Projet : Phase 2 – Créez vos patterns

Rappel : STRUCTURE DES FICHIERS

360 fichiers (15*24) qui reprennent les données provenant du smartphone lorsqu'une des 24 personnes testées (fichier supplémentaire data_subjects_info.csv) fait un mouvement. Ces fichiers sont répertoriés comme suit :



et leur structure est construite comme suit (correspond à une expérimentation pour une personne)

Tps	Attitude	Gravity	Rotation	Acceleration
	X Y Z	X Y Z	X Y Z	X Y Z
0	1.27...	...		
1				
...				
1000???				

Le choix ici est de baser notre algorithme "maison" sur la norme du vecteur accélération pris à chaque dixième ($V_{acc} = \sqrt{x^2 + y^2 + z^2}$) de seconde pendant les 1000 premiers dixièmes de secondes si la donnée existe.

OBJECTIF PHASE INTERMEDIAIRE : Créer les datasets.

Deux fichiers, l'un appelé « trainSet.csv » et l'autre « testSet .csv » qui reprennent chacun une partie des données à notre disposition seront fournis sur Moodle. Le train set nous servira à établir nos patterns et le fichier test set à tester notre modèle sur de nouvelles données. La structure de chacun des deux fichiers sera la suivante :

Mvt	Vacc1	Vacc2	Vacc1000
1	0.39562
1							0	0	0
2									

Prenez le temps de visualiser ces deux fichiers et de comprendre les données qui y sont incluses. Faire l'exercice à la main dans excel.

NB :

- chaque mouvement a été numéroté comme suit : 1- downstairs // 2 – upstairs // 3 – jogging// 4- sit down // 5 – stand up // 6 walking.
- le choix du fichier dans lequel l'enregistrement est stocké est aléatoire tout en respectant plus ou moins 90% des enregistrements dans le train set et 10% dans le test set.

Point théorie !!!



Notion d'overfitting et d'underfitting :

Il est très important de distinguer les données à partir desquelles les patterns sont générés des données sur lesquelles ces mêmes patterns sont testés. En effet, il est interdit de tester un modèle sur les valeurs qui ont servi à le créer.

De même, dans la phase de mise au point du modèle, on peut être amené à évaluer plusieurs fois le modèle. Ces tests ne peuvent pas non plus se faire tout le temps sur les mêmes données test sinon le risque est d'avoir un modèle qui « colle » au train set et au test set mais pas à de nouvelles données (**overfitting**). C'est pourquoi, dans la pratique, la **cross validation** est souvent utilisée. Cependant, pour faire simple dans notre projet, nous utiliserons uniquement un train set pour créer nos patterns et un test set pour évaluer notre modélisation.

Par ailleurs, il est bien sûr nécessaire de posséder suffisamment de données pour créer le modèle sinon risque **d'underfitting**.

OBJECTIF PHASE 2 : Créer les patterns.

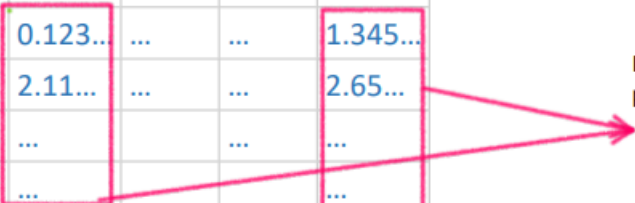
A disposition : le fichier « trainSet.csv ».

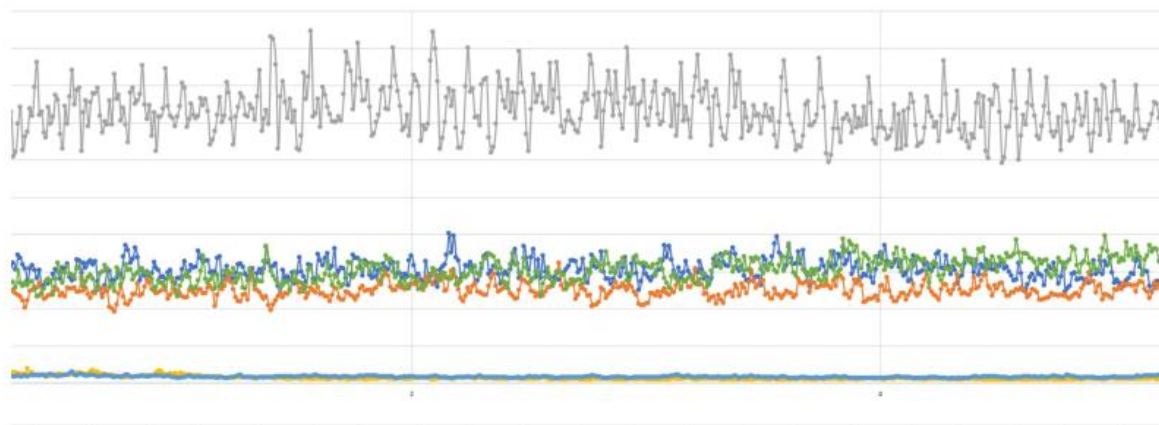
L'algorithme « maison » proposé se basera simplement sur les moyennes par mouvement et par dixième de seconde. Nous nous limiterons donc au « déplacement moyen » en termes de norme du vecteur accélération pendant les 1000 premiers dixièmes de secondes. Vous générerez ainsi 6 patterns.

Visualisation :

Mouvement	Vacc	Vacc	...	Vacc
1	0.123...	1.345...
1	2.11...	2.65...
1
1
3	0.123...	1.345...
3	2.11...	2.65...
3
...
6

Moyennes par mouvement et
par dixième de seconde





Le train set ressemble à ceci : Numéro du mouvement + 1000 normes

Enoncé phase 2 : créez vos patterns.

Ecrivez le DA qui permet, à partir du fichier « trainSet.csv » fourni, de générer les patterns correspondant à chaque type de mouvement. Ce nouveau fichier se nommera « patterns.csv » et aura la structure suivante : numéro du mouvement et 1000 valeurs moyennes

Mouvement	Vacc	Vacc	Vacc	Vacc	...	Vacc
1						
3						
4						
5						
2						
6						

Ce travail est à rendre en fin de séance sur le devoir Moodle prévu à cet effet.

Pour la fois prochaine, demandez à Chatgpt de vous générer ce programme et prenez-le pour le prochain labo.