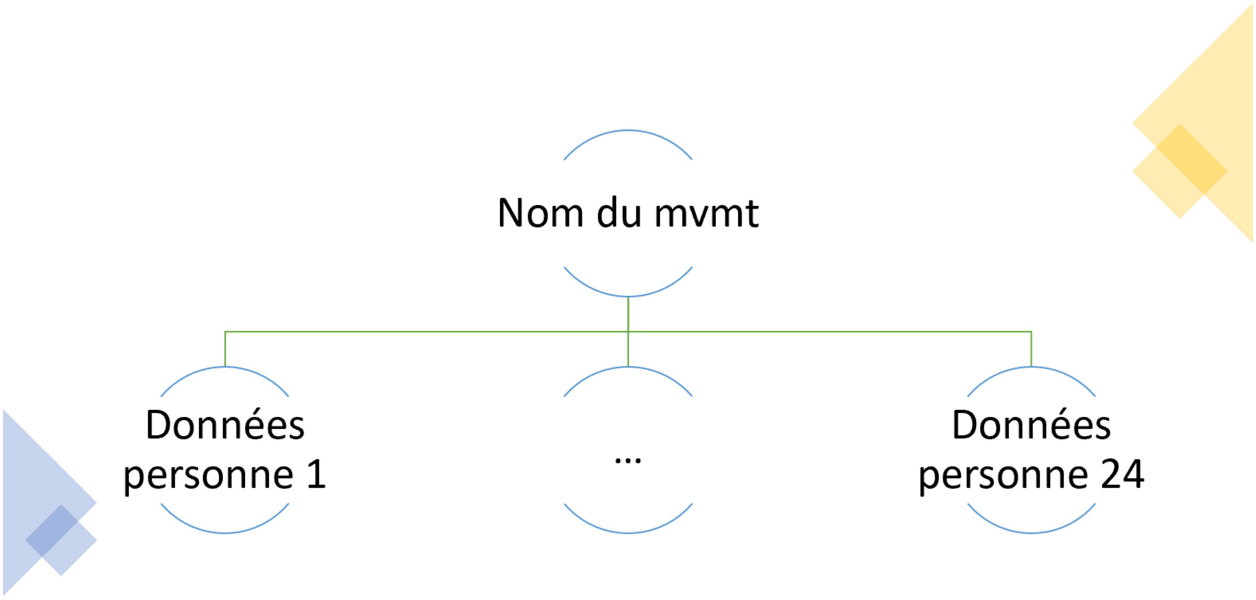
Phase 1 : Créez vos datasets

lundi 5 juillet 2021 14:35

Rappel : STRUCTURE DES FICHIERS

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



et leur structure est construite comme suit :

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

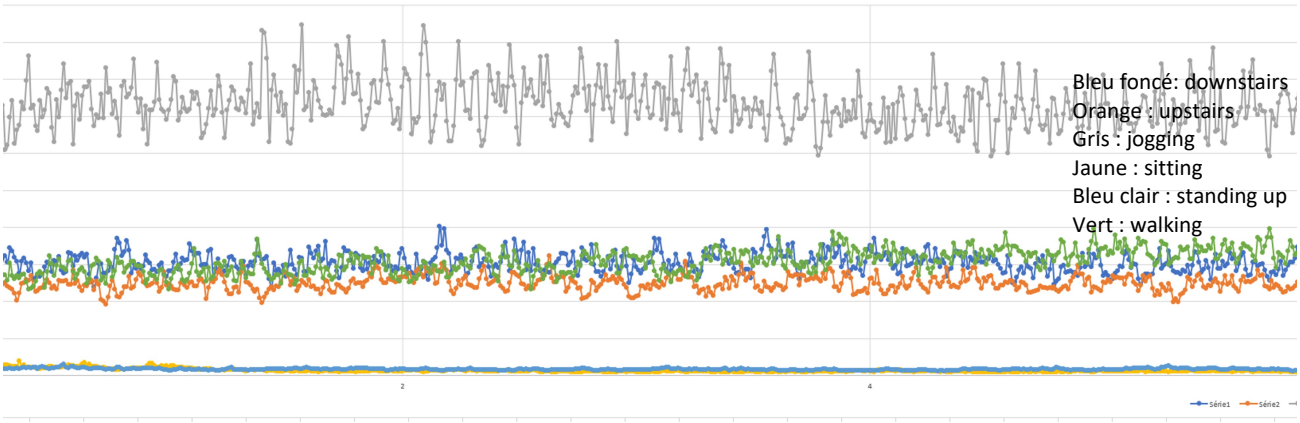
Le choix ici est de baser notre algorithme "maison" sur le vecteur accélération pris à chaque dixième (Vacc = Racine($x^2+y^2+z^2$)) de seconde pendant la première minute (si la donnée existe).

OBJECTIF PHASE 1 : Créez vos datasets

2 fichiers qui reprennent chacun une partie des données comme suit :

Mouvement	Genre	Index	Vacc	Vacc	...	Vacc
Mouvement	...	...	...	...	...	...

Un des fichiers va servir de training set pour trouver des patterns selon le mouvement.



L'autre fichier sera le test set pour tester les patterns sur base de données jamais utilisées.

Point théorie !!!

Il est très important de distinguer le fichier à partir duquel les patterns (et/ou les règles) sont générés du fichier sur lequel 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 amenés à évaluer plusieurs fois le modèle. Ces tests ne peuvent pas non plus se faire tout le temps sur le même fichier test sinon le risque est d'avoir un modèle qui « colle » au train set et au test set mais pas à des données nouvelles (risque d'overfitting). C'est pourquoi, dans la pratique, la cross validation est souvent utilisée. Cependant, pour faire simple dans ce projet, nous utiliserons uniquement un train set pour créer nos patterns et un test set pour évaluer notre modélisation.

Enoncé phase 1 : créez votre train set et votre test set.

Le modèle sera créé à partir du vecteur accélération calculé à chaque dixième de seconde ***pendant 1 minute*** si possible. Vous devez donc obtenir deux fichiers trainSet.csv - testSet.csv "structurés" comme suit :

Mouvement	Genre	Index	Vacc	Vacc	…	Vacc
Mouvement	…	…	…	…	…	…

Vacc : vecteur accélération au y dixième de seconde
Index est une variable qui va de 1 à 360, cela identifie l'expérience.

Vous devez donc parcourir tous les fichiers pour reconfigurer l'information sous la forme ci-dessus. Le fichier testSet.csv devra contenir (environ) 10% des enregistrements (1 enregistrement sur 10).

Avant de réaliser votre DA, voyez comment traiter des fichiers csv.
Réalisez le DA correspondant à l'énoncé ci-dessus – précisez aussi les structures de données que vous allez utiliser et remettez sur le devoir Moodle un fichier pdf NomEtudiant_Phase1.pdf.