**Projet : reconnaissance d’activité humaine via un smartphone.**

Il existe beaucoup d’applications permettant à un utilisateur de smartphone de connaître en direct (en temps réel) ses performances sportives (compter son nombre de pas, sa vitesse etc…). Pour cela, l’application a besoin d’identifier le type d’activité de l’utilisateur. L’idée du projet est de créer un programme qui permet de reconnaître l’activité de l’utilisateur (son mouvement : jogging - marche - monter les escaliers…) à partir de données captées sur son appareil mobile. Vous allez donc implémenter un algorithme "maison" de classification supervisée.

**En quoi consiste la classification supervisée ?**

***Principe :***

L'objectif de la classification consiste à attribuer une classe/une catégorie à chacune des observations d'un jeu de données.

On dispose au départ d'un ensemble de données d'apprentissage pour lesquelles le classement est connu (d'où le nom "supervisée"). Sur base de ces exemples, des règles vont pouvoir être établies pour permettre d'attribuer une catégorie à de nouvelles données du même type.

***Exemple de classification supervisée :***

De manière très schématique, supposons que l'objectif soit de déterminer un type de fleurs (iris Sétosa - iris Versicolor - iris Virginica) en fonction de la longueur du pétale, sa largeur, la longueur du sépale, sa largeur. On aurait ainsi un ensemble de données ressemblant à ceci :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Long pétale | Larg pétale | Long sépale | Larg sépal | Type |
| … | … | … | … | Sétosa |
| … | … | … | … | Sétosa |
| … | … | … | … | Sétosa |
| … | … | … | … | Sétosa |
| .. |  |  |  | Virginica |
| .. |  |  |  | Virginica |
| .. |  |  |  | Virginica |
| .. |  |  |  | Virginica |
| … |  |  |  |  |
| .. |  |  |  | Versicolor |
| .. |  |  |  | Versicolor |

A partir de ces exemples, l'algorithme va établir des règles, trouver des patterns qui lui permettront par la suite lorsqu'une nouvelle donnée arrive

|  |  |  |  |
| --- | --- | --- | --- |
| Long pétale | Larg pétale | Long sépale | Larg sépal |
| … | … | … | … |

de déterminer de quel type d'iris il s'agit.

***Quelles sont les différentes étapes d'une classification supervisée?***

* bien analyser et comprendre les données à disposition
* créer deux datasets (un pour l’apprentissage et un pour le test)
* créer le modèle
* évaluer et tester le modèle

***Revenons au projet : Phase 0***

Pour atteindre notre objectif de classification de mouvement, vous disposez de 15 fois 24 (+1) fichiers de données.



Un informaticien qui travaille sur des données ne peut évidemment pas faire fi de bien comprendre les données qu’il a à sa disposition et ce qu’il veut en faire. C’est une étape très importante pour éviter de travailler à l’aveugle et laisser passer des incohérences.

* Prenez donc d’abord le temps de découvrir à quoi correspondent les données qui sont à votre disposition en lisant attentivement les notes « context » et « content » du lien suivant : <https://www.kaggle.com/malekzadeh/motionsense-dataset> (Téléchargez ces données si vous ne les avez pas déjà)

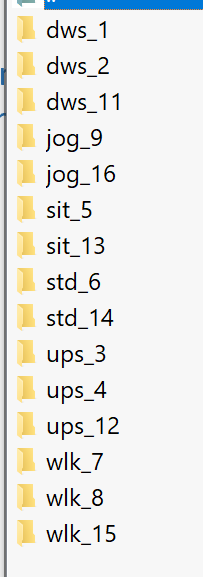
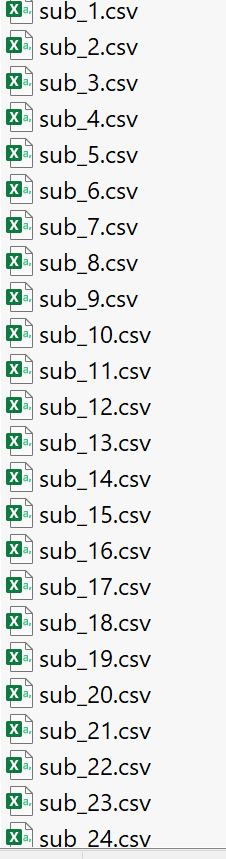
|  |  |
| --- | --- |
|  |  |
|  |  |

|  |  |
| --- | --- |
|  |  |

***Projet : Phase 1 – créer deux datasets***

# 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 :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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(x²+y²+z²)) de seconde pendant la première minute (si la donnée existe).

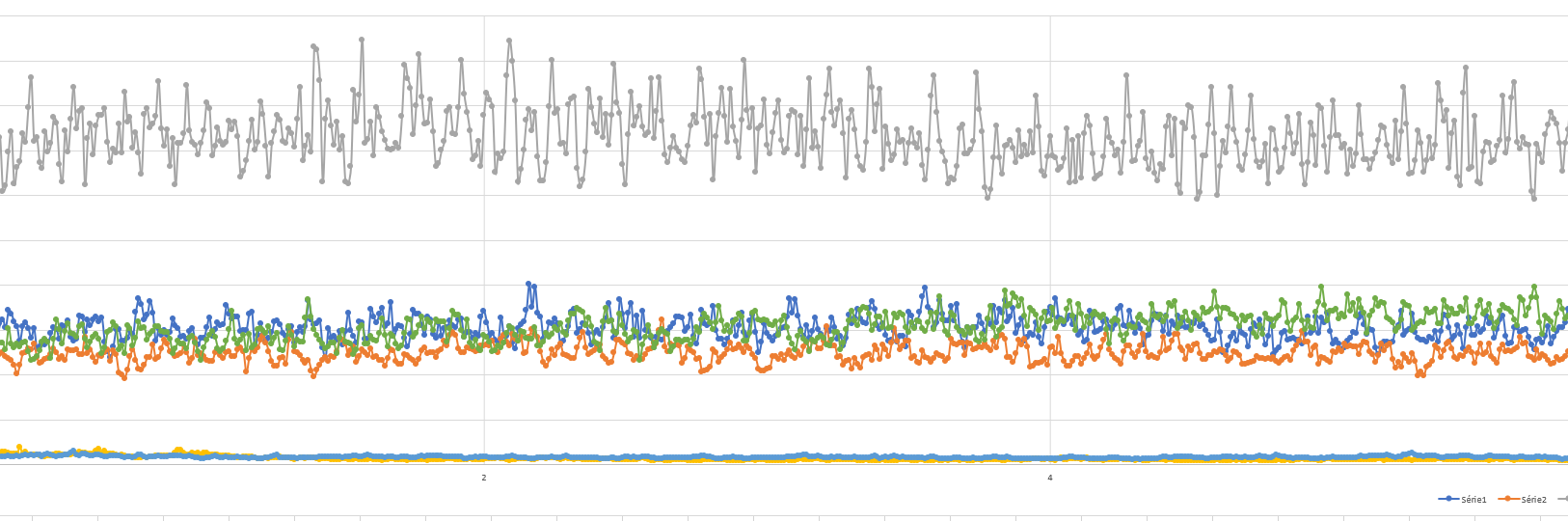
# OBJECTIF PHASE 1 : Créez vos datasets

## 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 l’apprentissage. Il permettra de trouver des patterns selon le mouvement.

  Bleu foncé: downstairs-Orange : upstairs-Gris : jogging-Jaune : sitting-Bleu clair : standing up -Vert : walking



## 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.

A partir des 360 fichiers csv, vous devez générer deux fichiers : trainSet.csv et testSet.csv. Le premier contiendra 90% des données, l’autre les 10% restant. La structure (à supposer que l’on puisse parler de structure pour des fichiers csv) de ces deux fichiers est la même :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mouvement | Genre | Index | Vacc | Vacc | … | Vacc |
| Mouvement | … | … | … | … | … | … |

## Vacc : vecteur accélération ( au y dixième de seconde. Il sera calculé chaque dixième de seconde ***pendant 1 minute*** si possible. NB : vous pouvez choisir d’autres données parmi celles qui sont proposées.

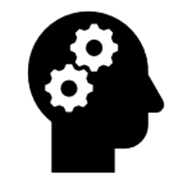
Index est une variable qui va de 1 à 360. Elle identifie l'expérience.

Vous devez donc parcourir tous les fichiers pour reconfigurer l’information sous la forme ci-dessus. Veillez à ce que le fichier testSet.csv

* contienne 10% de chaque type de mouvement
* contienne 1 ou 2 enregistrements correspondant à chaque sujet testé.

Dans un premier temps, vous réalisez le DA. Précisez aussi les structures de données que vous allez utiliser et remettez le devoir sur Moodle sous la forme nomEtudiant\_phase1.pdf pour la fin du cours.

NB : pour simplifier, vous donnerez un numéro à chaque mouvement (1=Downstairs / 2=jogging/ 3=upstairs/4=sit down/5=stand up/ 6=walking)



|  |  |
| --- | --- |
| Générer les noms de fichiers    Vous générerez les noms de fichiers au fur et à mesure comme suit :  Créez le tableau suivant :  char paths[15][LGPATH] = { "dws\_1/","dws\_2/","dws\_11/","jog\_9/","jog\_16/","sit\_5/","sit\_13/"…    Le chemin et nom du fichier sera : élément du tableau/sub\_numéro du fichier à traiter (de 1 à 24). |  |