**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)

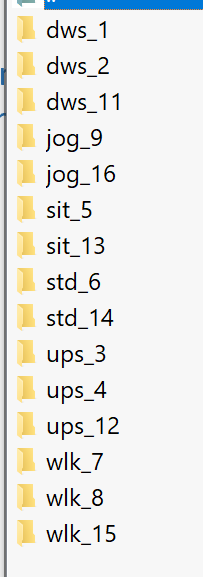
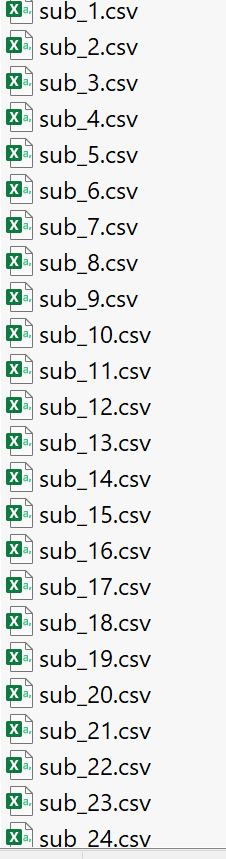
|  |  |
| --- | --- |
|  |  |
|  |  |

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

***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 : Vérifier les données

Par la suite, nous allons utiliser le vecteur accélération pour créer notre modèle (ce choix est totalement subjectif à des fins pédagogiques). La phase 1 va nous permettre de parcourir tous les fichiers et d’essayer de détecter d’éventuelles données manquantes ou aberrantes. Il a été établi que

* les données du vecteur accélération en x doivent suivre une loi normale de moyenne 0,00096087 et d’écart-type 0,38875666
* les données du vecteur accélération en y doivent suivre une loi normale de moyenne 0,05525659 et d’écart-type 0,61937128
* les données du vecteur accélération en z doivent suivre une loi normale de moyenne 0,0352192 et d’écart-type 0,4300345

# Enoncé phase 1 : validation des données.

Ecrivez le DA qui permet de parcourir les 360 fichiers csv à votre disposition. Pour chacun d’eux, vous allez imprimer dans un fichier csv son nom, son nombre de lignes, 0 ou 1 suivant qu’il manque ou non un des indices de temps, le nombre de données jugées aberrantes en x et pour chacune d’elles, l’indice de temps et la valeur en question. Idem pour les données en y et en z.

Implémentez-le en langage C.

|  |  |
| --- | --- |
| Générer les noms de fichiers en C    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). |  |