PROJET IAS
BINOME : Merimi Fatima-Zahra et Joanna Radola

9 Prise en main d'un problème, mise en route du projet

9.1 choix d'un problème

Choisir the data set à étudier

Depression DataSet : https://open.canada.ca/data/en/dataset/8d9bbce2-7bea-4f1b-95b8-4a1c866c30de

https://data.world/vizzup/mental-health-depression-disorder-data/workspace/file?filename=Mental+health+Depression+disorder+Data.xlsx

https://www.kaggle.com/arashnic/the-depression-dataset

Student Performance DataSet : https://archive.ics.uci.edu/ml/datasets/Student+Performance#

Car evaluation Safety : https://archive.ics.uci.edu/ml/datasets/Car+Evaluation

Fashion MINIST : https://www.kaggle.com/zalando-research/fashionmnist

9.2 exploration préliminaire des données

data, metadata

9.3 définition de la tâche, et choix d'une métrique de performance

9.4 Rapport d'étape

(a) Quel est l'objectif, globalement ? Car evaluation safety, déterminer si la voiture sont sûres avec des degré (safety: low, medium, high.) à partir des features données dans le data set. On souhaite que l'algorithme fonctionne au mieux sur des voitures qui sont dans le test/validation set.

(b) Quel type de tâche devrions-nous probablement accomplir (supervisée ou non, et dans chaque cas, quelle sous-catégorie) ? C'est une tâche de classification multi-classe (SVC et perceptron) avec trois classes, safety: low, medium, high.

(c) Quelle est la structure des données ? (Vous pouvez et devriez souvent faire certaines suppositions, selon les besoins). Faut-il faire attention à certains aspects du formatage ? Les features dans ce data set sont catégoriels, on peut les encoder en one-hot vector ; et pour le nombre de prote et personnes les transformer en int. On n'a pas de données manquantes et les valeurs sont notées avec les même mots clés.

Data Set Characteristics:	Multivariate	**Number of Instances:**	1728	**Area:**	N/A
Attribute Characteristics:	Categorical	**Number of Attributes:**	6	**Date Donated**	1997-06-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	1528219

Attributes:
buying: vhigh, high, med, low.
maint: vhigh, high, med, low.
doors: 2, 3, 4, 5more.
persons: 2, 4, more.
lug_boot: small, med, big.
safety: low, med, high.

(d) Quel algorithme semble bien adapté ? On peut utiliser la plupart des algorithmes de classification mais on choisit de tester particulièrement l'algorithme K-means clustering et random forest.

(e) Y a-t-il des points de vigilance particuliers ? (Choix des hyperparamètres, mesure des performances, ensemble de données non équilibrées, etc). Les voitures classées high dans safety sont rares. On devrait faire attention à ne pas hyperclasser les medium. Et dans notre cas d'étude, un faux high est plus cher qu'un faux low/med.
