Lamoureux Léo 21602944
Pires Pinto Lola 21706287

ALGO : RANDOM FOREST

Etape 4

Classification : Type d'étoiles

Lien : https://www.kaggle.com/datasets/deepu1109/star-dataset

Nous avons choisi un dataset sur les types d'étoiles car nous le trouvons intéressant, facile à comprendre et à interpréter. Il est adapté pour notre algorithme dans un but de classification.

features

- La température de l'étoile.
- Sa luminosité.
- Son rayon.
- Sa magnitude absolue (une autre mesure de luminosité).
- Sa couleur.
- Sa classification spectrale.

Nous avons choisi comme cible le type de l'étoile. Notre modèle essaiera donc de classifier le type d'une étoile selon toutes ces caractéristiques.

Nettoyage des données

- Aucune donnée manquante dans le dataset.
- Nous remplaçons les données de type catégoriques (i.e. qualitatives, c.a.d. non quantifiables, par exemple un nom) par des valeurs numériques : Par exemple, les différentes classes spectrales seront représentées par un numéro, de 1 à n.
- Pas vraiment de notion de valeurs aberrantes sur les étoiles car il peut y avoir des écarts énormes en fonction des étoiles.

Régression : Prix des ordinateurs

Lien : https://www.kaggle.com/datasets/muhammetvarl/laptop-price

Nous avons choisi un dataset concernant le prix des ordinateurs selon leurs caractéristiques. Nous l'avons choisi car il semble être adapté à notre algorithme dans un but de régression.
De plus, étant étudiants en informatique, c'est une étude qui nous intéresse.

features

- L'entreprise qui l'ordinateur.
- Le modèle de l'ordinateur.
- Le type d'ordinateur (portable, fixe…).
- La taille de son écran.
- La résolution de son écran.
- Le CPU qui y est installé.
- Sa capacité de RAM.
- Sa mémoire.
- Le GPU qui y est installé.
- Son système d'exploitation.
- Son poids.

La cible que nous avons choisie est le prix en euros de l'ordinateur. Notre modèle tentera donc de prédire le prix d'un ordinateur selon toutes ses caractéristiques.

Nettoyage des données

- Aucune donnée manquante dans le dataset.
- Nous avons ici aussi remplacé les données catégoriques par des données numériques.
- Nous avons supprimé du dataset la colonne donnant l'ID de l'ordinateur sur la table car elle était inutile pour notre modèle.
- Nous retirons du dataset les tuples pour lesquels le prix de l'ordinateur est aberrant. Ces prix ne feraient que fausser le modèle car ils sont inhabituels et non représentatifs de l'échantillon.
  Nous les supprimons car nous ne voulons pas les remplacer par d'autres prix (exemple par le prix moyen, le prix qui revient

le plus souvent…) qui ne correspondent pas aux caractéristiques du tuple et pourraient fausser le modèle.