Rapport Analysis of massive data

Epitech project – 5/12/21 12/12/21

arthur.perno@epitech.eu

eliott.m-barali@epitech.eu

Pour ce projet nous avons choisi de nous organiser de la manière suivante :

- Eliott s'est chargé de :
  - La génération du dataset personnalisé (pour le 1er exercice)
  - La visualisation des outliers
  - Le développement de visualization1 .py
- Arthur s'est chargé de :
  - L'analyse du dataset incluant la matrice de corrélation et les différents histogrammes
  - Le développement de visualization2.py

Dans le cadre de ce projet, nous avons choisi d'utiliser le dataset open exoplanet catalog.

Nous sommes tous deux passionnés par la découverte d'autres planètes d'où notre choix de ce dataset.

Ce dataset possède des données très intéressantes comme la taille des planètes, leur rayon, leur température ou encore des informations sur les étoiles autour desquelles elles orbitent.

Tout d'abord, la 1ère analyse de ce dataset s'est faite en plusieurs temps :

- La compréhension des différentes colonnes et leur importance
- La création d'histogrammes pour repérer la tendance des variables
- Le calcul de l'écart-type pour vérifier la distribution
- La création et l'analyse d'une matrice de corrélation

Dans l'analyse de ce dataset, plusieurs variables importantes sont ressorties : la masse des planètes, le rayon, la période de révolution (nombre de jours mis par la planète pour faire un tour de son étoile), le demi grand axe (la moitié du grand axe qui correspond au plus long diamètre de l'ellipse effectuée par la planète), l'excentricité (écart de forme entre l'orbite d'une planète et un cercle parfait, plus il est élevé, plus l'orbite aura une forme elliptique voire parabolique ou hyperbolique), l'inclinaison (correspond au degré d'inclinaison de la planète) et enfin la température de la planète.

En analysant les histogrammes de ces différentes variables, on se rend compte que certains sont assez illisibles notamment en ce qui concerne la masse et la période de révolution. En effet, certaines

planètes du dataset sont significativement plus massives et/ou larges que les autres : on en déduit qu'il s'agit d'outliers et pour permettre une analyse visuelle plus simple on écarte ces dernières des histogrammes en ne prenant que les planètes en dessous de 10 masses jupitériennes et en dessous de 10000 jours de révolution.

Il est important de noter que le calcul du nombre de bins des histogrammes se fait via la règle de Freedman-Draconis selon le calcul suivant : h = 2*iqr*n^(-1/3) où iqr correspond à l'écart interquartile et n au nombre d'observations totales.

Le calcul de l'écart-type est effectué sur chaque variables et l'on peut voir que deux ressortent notamment :

- La période de révolution qui a un écart-type plus grand que sa moyenne, cela signifie que l'échantillon de données n'est peut-être pas encore suffisamment précis et que certaines données sont trop éloignées de la majeure partie des autres.
- La température qui a un écart-type élevé qui correspond à peu près à la moitié de la moyenne, ce qui démontre une distribution assez balancée des valeurs

Enfin, la matrice de corrélation permet de faire ressortir certaine corrélation :

- On observe une corrélation positive faible (0,42) entre la masse et le rayon de la planète : cela signifie que la taille et la masse d'une planète ne corrèlent pas forcément
- On peut voir également une corrélation positive moyenne (0,66) entre la température de la planète et la température de son étoile ce qui intuitivement semble assez logique, plus une étoile est chaude et plus la planète qui l'orbite le sera aussi
- On observe ensuite une autre corrélation positive moyenne (0,69 et 0,61) entre le rayon et la température, cela peut correspondre au fait que plus un corps céleste est chaud et plus il va perdre de la matière de ses couches externes et donc sa taille va irrémédiablement se réduire
- Enfin on peut voir une très forte corrélation (0,97) entre la période de révolution et le demi grand axe ce qui est tout à fait logique étant donnée que le demi grand axe correspond à une distance de l'ellipse de l'étoile et la période est le temps que met l'étoile à circuler sur cette ellipse