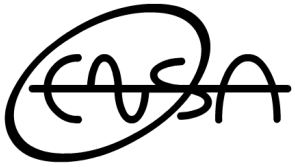

Rapport du projet ML

Prédiction des performances des élèves

Réalisé par:

- Elkhalifi Zahia.
- Ghammouri Asmae.
- Jabri Latifa.
- Kerroumi Hajar.

Année universitaire : 2022/2023

Objectif du projet :

l'objectif principal de ce projet est de prouver la possibilité de former et de modéliser un ensemble de données de petite taille et la faisabilité de créer un modèle de prédiction avec un taux de précision crédible. Cette recherche explore également la possibilité d'identifier les indicateurs clés dans le petit ensemble de données, qui seront utilisés pour créer le modèle de prédiction, en utilisant des algorithmes de visualisation et de regroupement. Les meilleurs indicateurs ont été introduits dans plusieurs algorithmes d'apprentissage automatique pour les évaluer pour le modèle le plus précis.

Un bon dataset :

On reconnait un bon dataset selon plusieurs caractéristiques :

- Il faut avoir assez d'observations complètes. Il n'existe pas de seuil à dépasser pour être capable de dire qu'on a « assez » de données. En revanche, on peut facilement voir si les données qu'on a sont complètes. Si une variable a plus de 1% de valeurs qui sont manquantes, on peut la supprimer car elle ne sera pas assez précise et peut fausser nos résultats.
- la qualité : il faut s'assurer que les données qu'ils renferment sont exactes. Lorsqu'on récupère les données via un API ou en faisant du scraping ou en communiquant avec des serveurs via SQL, il est facile de contrôler la qualité des données et vérifier si les données qu'on a sont correctes.
- Les données qu'on utilise doivent être également représentatives de la population. C'est particulièrement important lorsqu'on analyse des données de sondage, où

les réponses peuvent provenir majoritairement d'une infime partie de la population. Un autre exemple peut être celui des fraudes, où les datasets qui servent à entraîner des algorithmes de classification sont souvent déséquilibrés et contiennent peu de fraudes. Un algorithme de Machine Learning aura des difficultés à apprendre des données si elles ne couvrent pas tous les cas de figure.

- La dernière caractéristique importante est celle liée aux nombre de variables utiles variées. Les variables doivent d'abord provenir de sources variées (exemple : goûts musicaux, appareils utilisés, lieu, moment de la journée, etc.). En plus de cela, les données doivent être utiles, c'est-à-dire faire partie des types de données int, float et category, qui sont les 3 types utilisés par les modèles de Machine Learning. Les colonnes d'autres types ne seront pas utilisées donc rien ne sert à ce qu'elles soient en grand nombre. Par contre, on peut extraire de variables de type string des variables de type int, float ou category et donc augmenter le nombre de variables utiles.

Source des données:

- **Dataset utilisé :** Kaggle.

Analyse exploratoire des données:

L'analyse exploratoire des données est un moyen puissant d'explorer un jeu de données. Même lorsque l'objectif est d'effectuer des analyses planifiées, l'analyse exploratoire des données peut être utilisée pour le nettoyage de données, l'analyse des sous-groupes ou simplement pour

mieux comprendre nos données. Une étape initiale importante dans l'analyse des données consiste à représenter graphiquement les données.

Description de la phase de Pre-Processing des données :

- **Data Cleaning :**

La première étape consiste en un nettoyage des données incorrectes, incomplètes ou manquantes.
S'il manque des données dans le dataset, nous pouvons choisir de les ignorer dans le cas où la base de données est assez fournie et si de nombreuses données sont manquantes au sein de la même ligne. Nous pouvons également décider de remplir ces données manquantes de différentes manières : nous pouvons les remplacer par la valeur moyenne ou par la médiane
Pandas nous fournit des méthodes qui nous permettent d'effectuer ces traitements : fillna(), dropna().

- **Data Transformation.**

la conversion des étiquettes sous forme numérique afin de les convertir en une forme lisible par machine. Les algorithmes de machine learning peuvent alors décider la meilleure manière dont ces étiquettes doivent être utilisées : LabelEncoder, OrdinalEncoder, LabelBinarizer, et OneHotEncoder.

L’approche (l’algorithme) utilisée pour la résolution du problématique :

- Linear Regression() :

Les algorithmes de régression linéaire modélisent la relation entre des variables prédictives et une variable cible. La relation est modélisée par une fonction mathématique de prédiction. Le cas le plus simple est la régression linéaire univariée. Elle va trouver une fonction sous forme de droite pour estimer la relation. La régression linéaire multivariée intervient quand plusieurs variables explicatives interviennent dans la fonction de prédiction. Et finalement, la régression polynomiale permet de modéliser des relations complexes qui ne sont pas forcément linéaires.

- RidgeRegression() :

est l'une des méthodes de pénalisation les plus intuitives. Elle s'utilise pour limiter l'instabilité des prédictions liée à des variables explicatives trop corrélées entre elles.

La fonction de pénalisation se base sur la norme dite L2 qui correspond à la distance euclidienne. La régression ridge revient donc à minimiser la fonction de coût.

La pénalisation ridge va diminuer la distance entre les solutions possibles, sur la base de la mesure euclidienne.

- LassoRegression() :

c'est une forme de pénalisation qui permet de rendre nul certains coefficients de variables explicatives (contrairement à la régression ridge qui pourra aboutir à des coefficients proches de 0, mais jamais strictement nuls).

Le lasso est donc un algorithme qui permet également la simplification du modèle, en éliminant des variables.

- DecisionTreeRegression() :

L’arbre de décision est un algorithme qui se base sur un modèle de graphe (les arbres) pour définir la décision finale. Chaque nœud comporte une condition, et les branchements sont en fonction de cette condition (Vrai ou Faux). Plus on descend dans l’arbre, plus on cumule les conditions.