Démarche de travail

* 1. Définir un objectif ****mesurable****
     1. **Objectif** du projet

Trouver le prix des laptops en fonction de leurs caractéristiques

* + 1. **Métrique**

**F1\_score : 90%**

* 1. EDA (Exploratory Data Analysis) Exploration des données
     1. **Checklist de base (non-exhaustive)**
        1. Analyse de la forme :
  + **Identification de la target** : « Price\_euros »
  + **Nombre de lignes et de colonnes** : 1303 lignes et 12 variables + 1 Target
  + **Types de variables** :

Une variable de type « Int », 2 variables dont la Target de type « Float » et 10 variables de type « Object »

* + **Identification des valeurs manquantes** : Aucune valeur manquante
    - 1. Analyse du fond :
  + **Visualisation de la target (Histogrammes / Boxplot)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Répartition des prix par tranche** | | | | | | |
| Prix | <= 1000 | >1000 & <= 1500 | >1500 & <= 2000 | >2000 & <= 3000 | >3000 & <= 4000 | >4000 |
| N | 683 | 316 | 167 | 118 | 19 | 4 |

La Target varie entre 174€ et 6099€

Le prix moyen est de 1124€

Les Quantiles se trouvent à 599€ et 1488€, La médiane à 977€

* + **Compréhension des différentes variables**

Une variable quantitative représentant le prix des laptops. C’est la Target

* + **Visualisation des relations features-target (Histogrammes / Boxplot)**
  + **Relation entre les variables**
  + **Identification des outliers**
  1. Preprocessing (pré-traitement des données)

**L’objectif est de** Transformer le data pour le mettre dans un format propice au Machine Learning

* + 1. **Checklist de base (non-exhaustive)**
* **Création du Train set/ Test set**
* **Élimination des NaN : dropna(), imputation, colonnes "vides"**
* **Encodage : OneHotEncoding, OrdinalEncoding ou fonction perso + replace() ou map()**

Pour la variable « Cpu », regrouper en I3, I5, I7, AMD et autre

Pour la variable « Gpu », regrouper en AMD, Nvidia, Intel et autre

* **Suppression des outliers néfastes au modèle (à faire apres la création d'un 1er modèle)**
* **Feature Selection : Éliminer les variables redondantes ou dont la variance est nulle**

Une variable ID quantitative que nous pourrons par la suite éliminer

On pourrait éventuellement éliminer la variable « Product ». Plusieurs infos de certains échantillons sont redondants avec d’autres variables comme la « Ram » et « Cpu »

* **Feature Engineering : Essayer de créer de nouvelles variables**

Séparer la hauteur et la largeur de la Résolution.

Plusieurs variables dont les échantillons sont à modifier (enlever « GB » pour « Ram », enlever « kg » pour « Weight »

Pour « Memory », creer colonnes HHD, SSD, Flash, Hybrid et indiquer la mémoire associée

Calculer le PPI (pixel per inch) 🡪 (np.sqrt((2560\*\*2) +(1600\*\*2)))/13,3

* **Feature Scaling : Normaliser les données**
  1. Modeling (Modélisation)

L’objectif est de Développer un modèle de Machine Learning répondant à l'objectif final

* + 1. **Checklist de base (non-exhaustive)**
* **Définir une fonction d'évaluation : Comparer la performance de modèle de ML**
* **Entrainement de différents modèles : Créer une liste de ML et les entrainer**
* **Optimisation avec GridSearchCV : Optimiser les hyperparamètres**
* **(Optionnel) (Analyse des erreurs des modèles et retour au Préprocessing/ EDA)**
* **Learning Cuve et prise de décision**
* **Matrice de confusion et rapport de classification**