2022

BAGUIAN HAROUNA DIASSANA FATOUMATA dite SOUKOURA

17/06/2022



Rapport projet prétraitement et visualisation des données

Table des matières

**Aucune entrée de table des matières n'a été trouvée.**

Introduction

Le machine Learning est en plein essor, de nos jours les algorithmes sont devenu de plus en plus efficace et donne de bon resultats.les algorithme de machine Learning il y’a de toute sorte allant de la prédiction a la classification. Cet essor est favorisé par le boum de la masse de donnée. Ces données doivent être traité afin de les utilisé dans l’algorithme de machine Learning c’est le data prepocessing. Le traitement des données fait référence à la conversion de données brutes en informations significatives, et ces données sont également lisibles par machine.  Ainsi, le traitement des données implique la suppression des valeurs aberrantes, les oubliers et les normalisé les données. Après le traitement des données une visualisation est nécessaire afin de mieux comprendre nos données, c’est le data visualisation. La data visualisation (data viz ou représentation graphique de données ; elle s'écrit également data visualization) consiste à structurer visuellement des données recueillies et stockées. Ainsi, l'exploitation des données se fait plus facilement. Ensuite nous pouvons choisir le modèle adéquat pour entrainer avec nos données. Faire tout ceci est réservé a un public aguerri.

Mais si on pouvait faire du machine Learning sans coder ?

C’est au vu de répondre a cette problématique que nous avons créez data cleaner.

Mots clés : python, data prepocessing, data visualisation, streamlit, prediction, classification

Le machine Learning passe par plusieurs étapes : data collection and prepocessing, choix du modèle, training du modèle, évaluation et la prediction

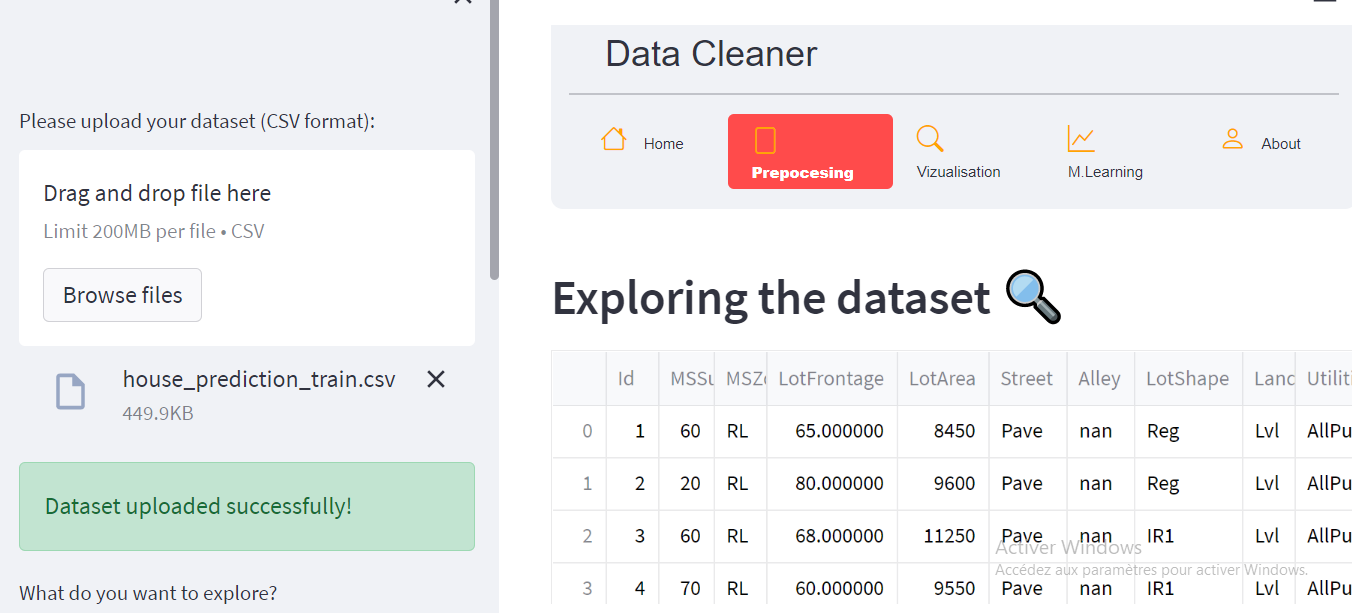
Data collection and prepoccessing

Le preprocessing est la deuxième étape du machine Learning juste après la collecte des données.

Le prétraitement fait référence aux transformations appliquées à nos données avant de les alimenter à l’algorithme. Le prétraitement des données est une technique utilisée pour convertir les données brutes en un ensemble de données propres.

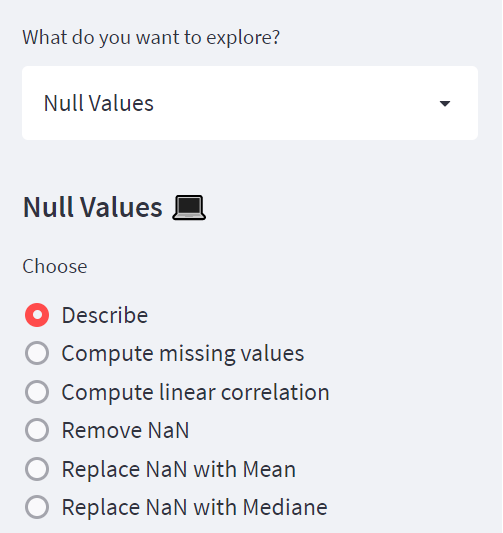
Pour le prétraitement de nos données on est passé par plusieurs étapes.

Uploader les données pour commencer le prétraitement.



1. **La suppression des missing values**

Le problème de la valeur manquante est assez courant dans de nombreux ensembles de données réels. Une valeur manquante peut fausser les résultats des modèles d'apprentissage automatique et/ou réduire la précision du modèle.



Apres avoir uploadé les données on fait une description générale des données. Ensuite on a la possibilité de supprimer les données aberrantes, de les remplacer par la moyenne ou par la médiane.

1. **Visualisation et suppression des outliers**

Plusieurs algorithmes de Machine Learning sont sensibles aux données d’entrainement ainsi qu’à leurs distributions. Avoir des O*utliers dans Training Set*d’un algorithme de Machine Learning peut rendre la phase d’entrainement plus longue. Sans mentionner que l’apprentissage sera biaisé. Par conséquent, le modèle prédictif produit ne sera pas performant, ou du moins, loin d’être optimal.

Pour la visualisation on a utilisé des boxplots.

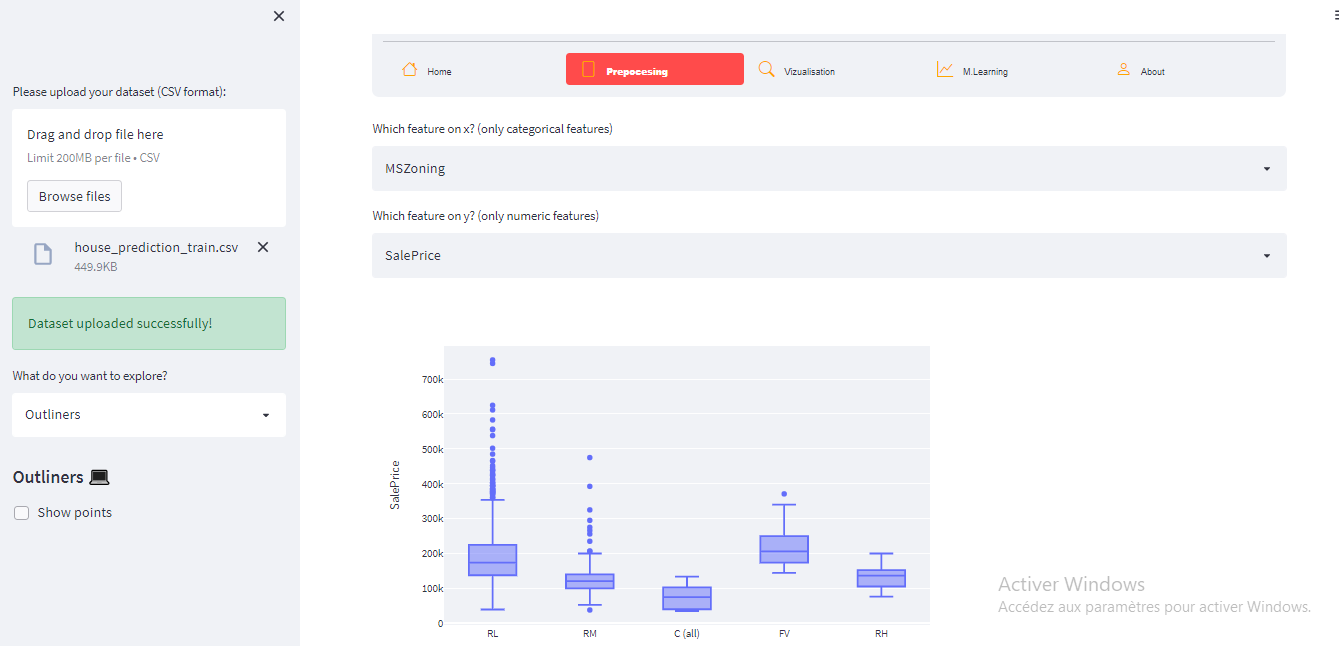
La suppression s’est faites par la méthodes de l’intervalle interquartile(IQR)

Chaque ensemble de données peut être divisé en quartiles. Le premier quartile indique que 25 % des points de données sont inférieurs à cette valeur, tandis que le deuxième quartile est considéré comme le point médian de l'ensemble de données. La méthode interquartile trouve les valeurs aberrantes sur les ensembles de données numériques en suivant la procédure ci-dessous

* Trouvez le premier quartile, Q1.
* Trouvez le troisième quartile, Q3.
* Calculez l'IQR. IQR= Q3-Q1.
* Définissez la plage de données normale avec la limite inférieure comme Q1–1.5\*IQR et la limite supérieure comme Q3+1.5\*IQR.
* Tout point de données en dehors de cette plage est considéré comme aberrant et doit être supprimé pour une analyse plus approfondie.

Le concept de quartiles et d'IQR peut être mieux visualisé à partir de la boîte à moustaches. Il a le point minimum et maximum défini comme Q1–1.5\*IQR et Q3+1.5\*IQR respectivement. Tout point en dehors de cette plage est aberrant.

Choisir les données a représenté par des box plots pour visualiser les outliers afin de les supprimer.



1. **La normalisation des données**

Normalisation : La normalisation est une méthode de prétraitement des données qui permet de réduire la complexité des modèles.

Elle permet de donner la même importance à tous les features

Il existe plusieurs méthodes à l’instar de Standard scaling.

**Standard scaling**

La normalisation standardise la moyenne et l’écart-type de tout type de distribution de données, ce qui permet de simplifier le problème d’apprentissage en s’affranchissant de ces deux paramètres.

Pour effectuer cette transformation, on soustrait aux données leur moyenne empirique *m*et on les divise par leur écart-type *σ.*

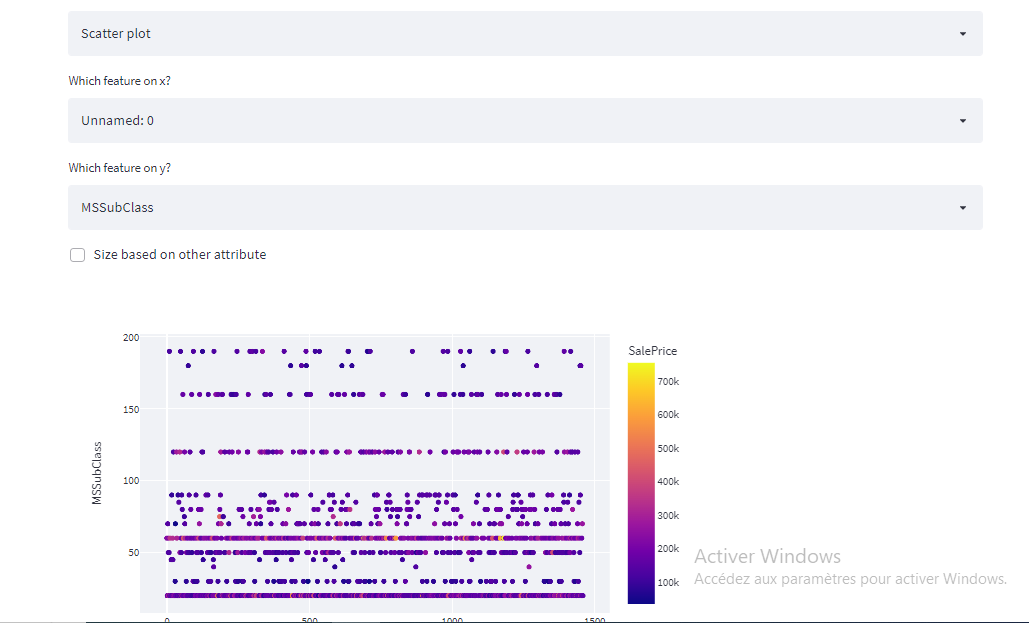
Data visualisation

La data visualisation est l’art et la manière de transformer la donnée en un formidable outil d’analyse. En montrant l’invisible, la data visualisation facilite et accélère la prise de décision.

La data visualisation simplifie la diffusion de l’information. Elle apporte des points de comparaison et d'analyse sur les tendances. Elle affine alors les prédictions sur les tendances à venir.

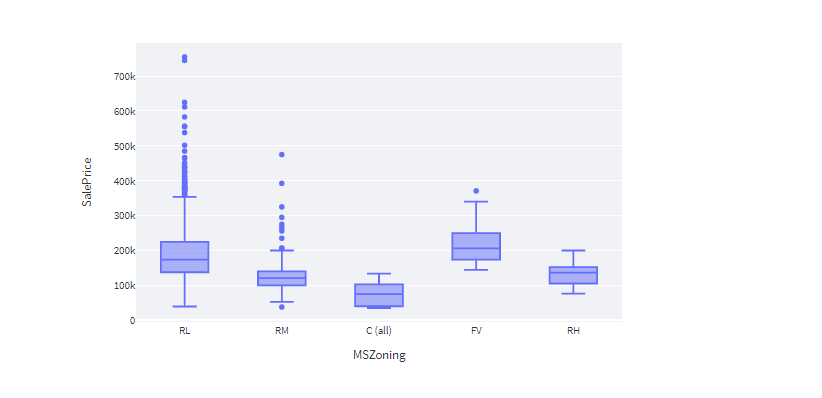
1. **Scatter plot**

Un nuage de points (également appelé nuage de points , nuage de points , nuage de points , nuage de points ou diagramme de dispersion )  est un type de tracé<https://en.wikipedia.org/wiki/Plot_(graphics)> ou de diagramme<https://en.wikipedia.org/wiki/Mathematical_diagram> utilisant des coordonnées cartésiennes<https://en.wikipedia.org/wiki/Cartesian_coordinate_system> pour afficher les valeurs de généralement deux variables<https://en.wikipedia.org/wiki/Variable_(mathematics)> pour un ensemble de données.



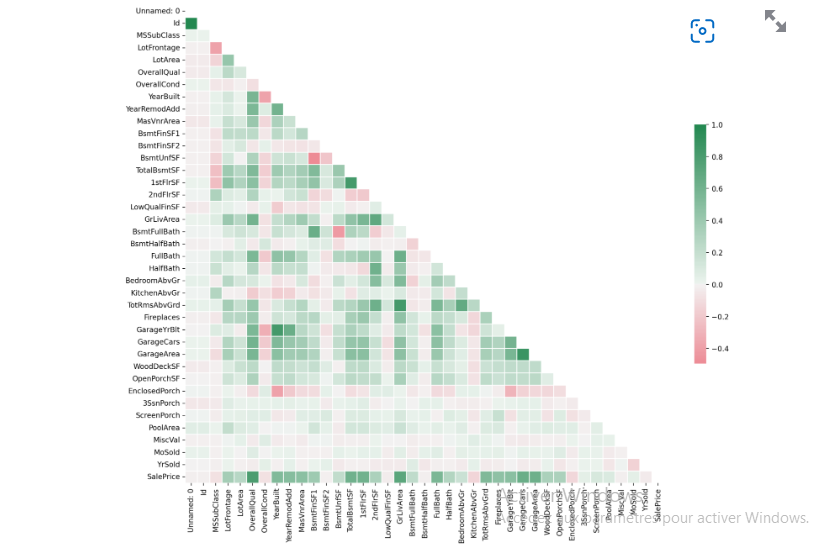
1. **Box plot**

Un box-plot est un graphique simple composé d'un rectangle duquel deux droites sortent afin de représenter certains éléments des données. La valeur centrale du graphique est la médiane (il existe autant de valeur supérieure qu'inférieures à cette valeur dans l'échantillon).



1. **Corrélation matrix**

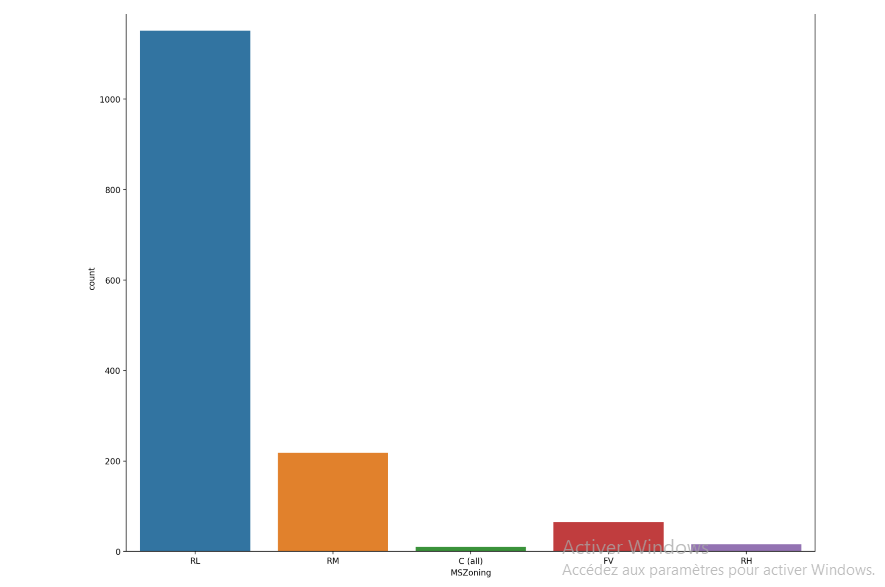
Une matrice de corrélation est simplement un tableau qui affiche les coefficients de corrélation pour différentes variables. La matrice représente la corrélation entre toutes les paires de valeurs possibles dans un tableau. C'est un outil puissant pour résumer un grand ensemble de données et pour identifier et visualiser des modèles dans les données données.



1. **Count plot**

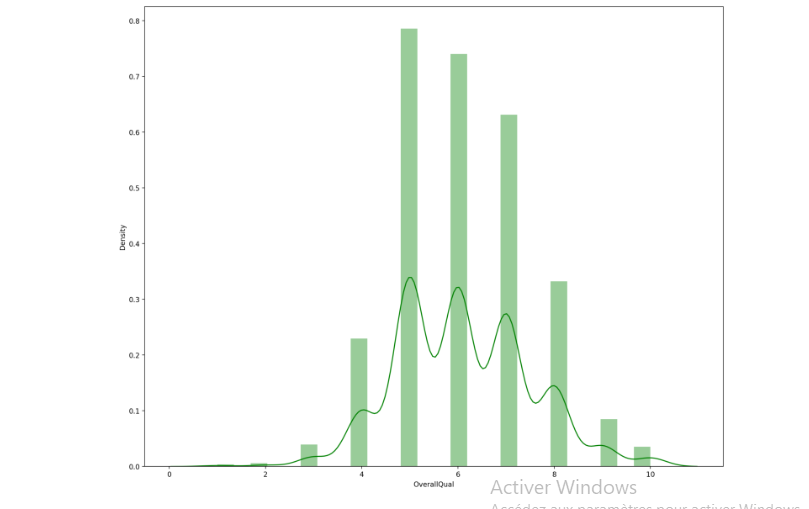
Le count plot est utilisé pour représenter l'occurrence (les comptes) de l'observation présente dans la variable catégorielle.

Il utilise le concept d'un graphique à barres pour la représentation visuelle.



1. **Distribution plot**

Les diagrammes de distribution sont d'une importance cruciale pour l'analyse exploratoire des données. Ils nous aident à détecter les valeurs aberrantes et l'asymétrie, ou à obtenir un aperçu des mesures de la tendance centrale (moyenne, médiane et mode).



Machine Learning

Apres avoir traiter nos données à travers le preprocesing (prétraitement des données), les visualisées dans la partie <visualisation>, on va procéder au machine Learning autrement appelée apprentissage automatique.

Le machine Learning est une technique de programmation informatique qui utilise des probabilités statistiques pour donner aux ordinateurs la capacité d’apprendre par eux-mêmes sans programmation explicite.

Cette partie se divisera principalement en deux volets à savoir :

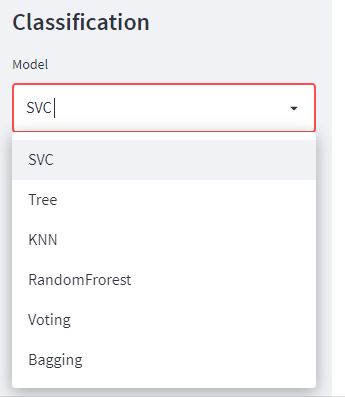
* La classification
* La régression

1. **La classification**

La classification est une partie du machine Learning où la variable de sortie est catégorie.

Dans cette partie on mit en place des modèles, qui une fois bien entrainés, essayeront de développer une fonction qui classifiera avec précision la sortie à partir des variables d’entrées.

* Choix du modèle



Dans cette partie on offre la possibilité à l’utilisateur de choisir un modèle parmi les six ci-dessus.

Apres le choix du modèle l’utilisateur devra procéder au paramétrage des hyperparamètres ainsi que l’entrainement.

* Paramétrage du modèle

On offre la possibilité de choisir les meilleurs hyperparamètres afin que le modèle ait de très bonnes performances.

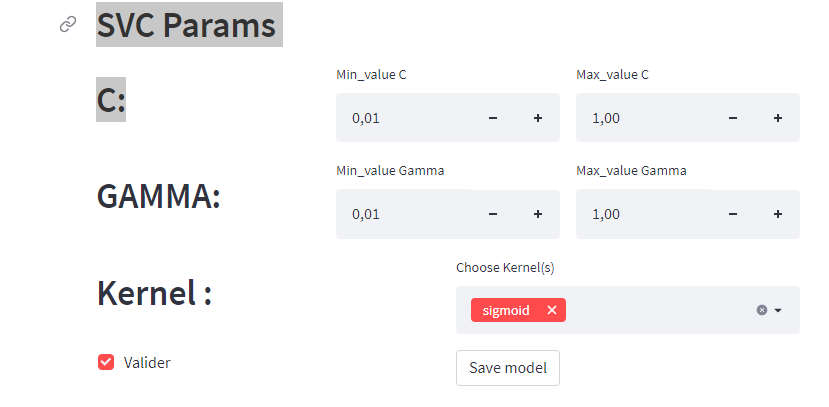
Deux possibilités existent en utilisant le gridsearch ou le randomsearch.



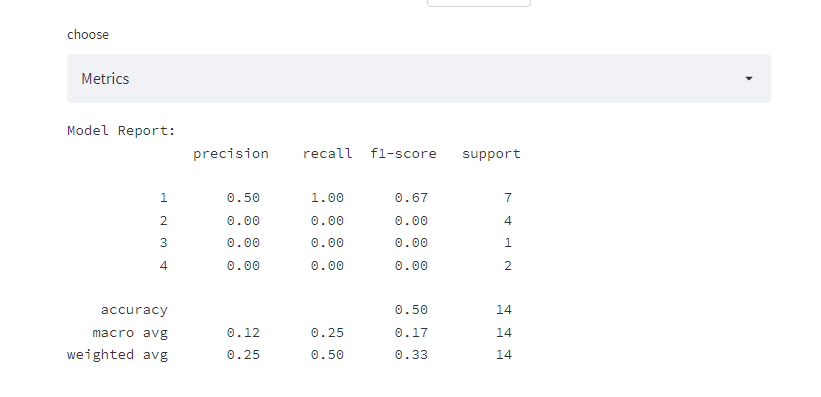
La différence principalement entre le gridsearch et le randomsearch étant que le gridsearch fait la recherche avec un pas assez petit contrairement au randomsearch.

Par conséquence le premier prend plus de temps mais obtient des hyperparamètres meilleurs que le second.

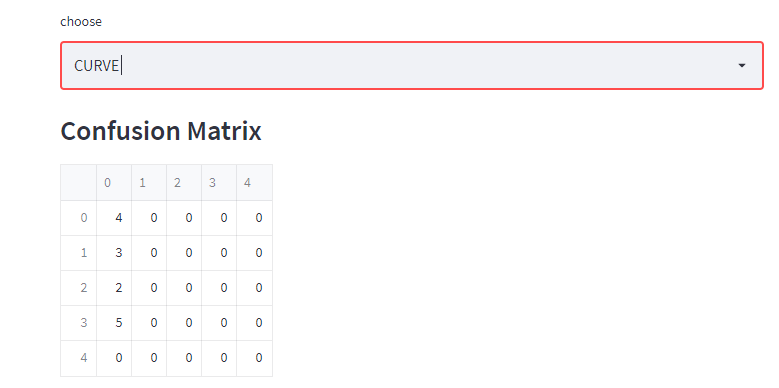
* Visualisation des performances de notre modèle après entrainement



Après avoir cliquer sur valider on obtient les performances suivantes :

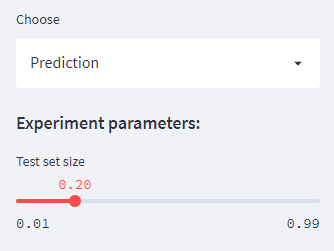


Également la matrice de confusion pour nous éclairer d’avantages sur les classes que notre modèle pourra parfaitement identifier et celles qui le sont moins.

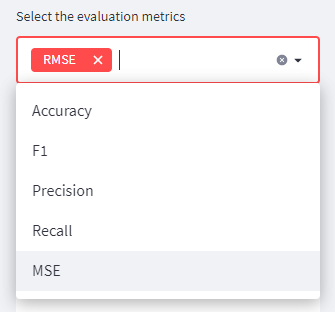


1. **Prédiction**

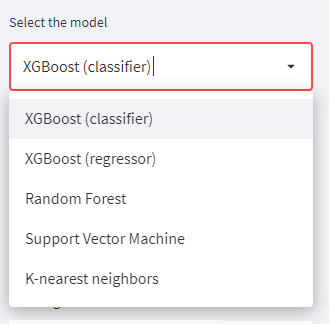
Dans cette partie on offre à l’utilisateur la possibilité diviser ses données : un pourcentage pour l’entrainement et un autre pour le test.



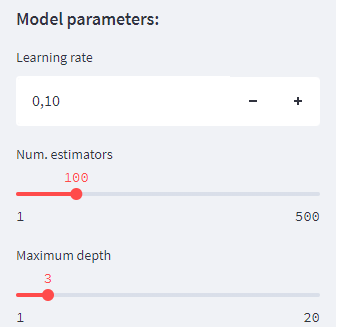
A coté de cela on offre également la possibilité de choisir la métrique pour évaluer la performance du modèle de prédiction.



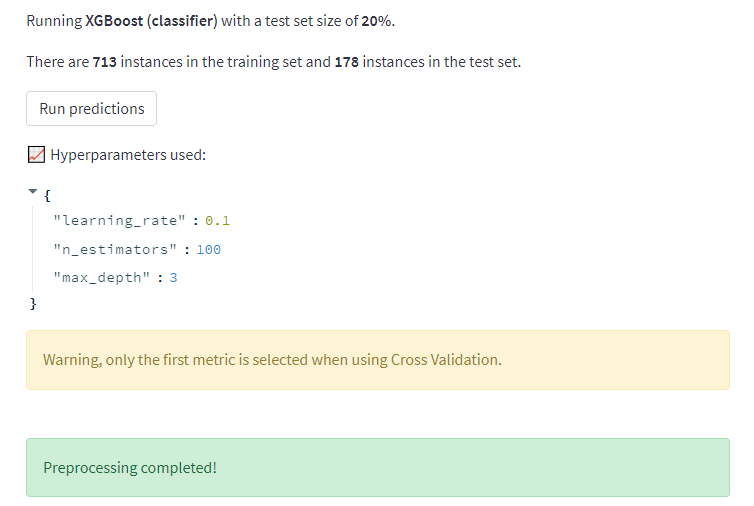
Choisir le modèle



Par ailleurs on peut également paramétrer le modèle comme dans la partie classification.

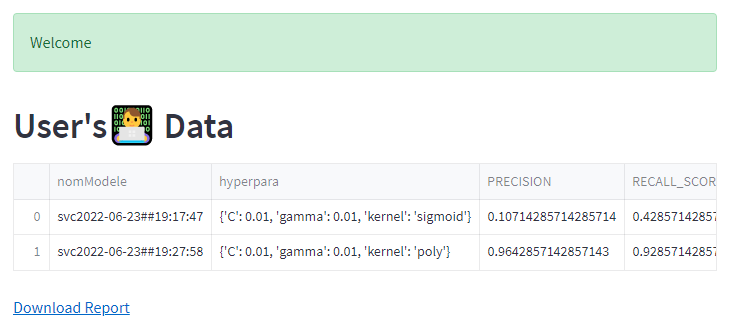


Enfin on entraine le modèle

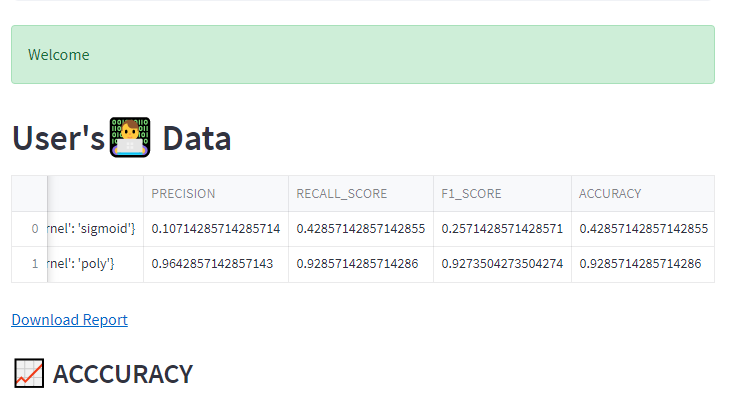


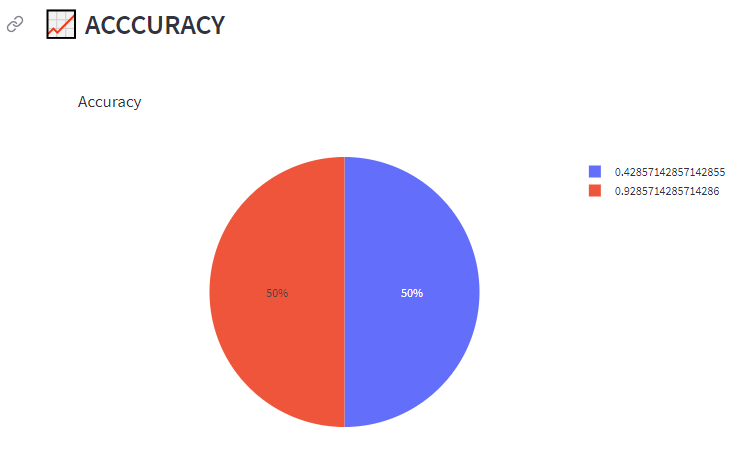
# Le rapport

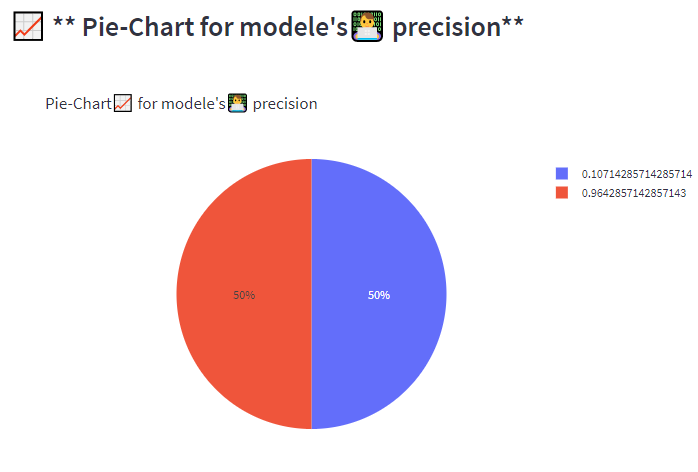
Cette dernière partie comportera un aperçue claire des modèles entrainés avec leur précision, les meilleurs hyperparamètres, l’accuracy ainsi que des graphiques pour la visualisation de ces données.

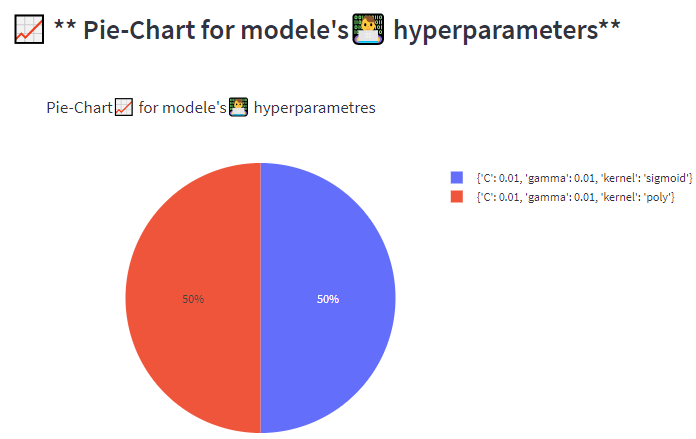


On a la possibilité de télécharger ces données pour les réutiliser après.









# Conclusion

Ce projet nous ont permis d’avoir une maitrise en termes de prétraitement des données et visualisation.

Pouvoir télécharger des données, les visualiser par des graphiques concis, détecter et supprimer les valeurs manquantes, le pourcentage de valeurs manquantes, entrainer des modèles d’apprentissage sur des données bien traiter et enfin sauvegarder sous forme de base de données, telles sont les principaux objectifs réalisés de ce projet.

Références

<https://github.com/antonin-lfv/Online_preprocessing_for_ML/blob/master/main.py>

Le Framework Streamlit