Rapport  
 Le projet se focalise sur la catégorisation des objets célestes en trois groupes distincts : galaxy, star et quasars (QSO).

Pour atteindre cet objectif, nous avons exploité un ensemble de bibliothèques Python spécialisées dans la visualisation de données, le prétraitement, la détection d'anomalies et la modélisation des algorithmes d'apprentissage automatique. Ces bibliothèques ont été sélectionnées pour leur efficacité et leur polyvalence dans le traitement de données complexes, ainsi que leur capacité à faciliter l'analyse et l'interprétation des résultats. Les outils utilisés comprennent :

-Matplotlib.pyplot et Seaborn pour la représentation graphique des données, fournissant des outils visuels pour explorer les relations entre les différentes variables et présenter les résultats de manière claire et intuitive.

-Scikit-learn pour le prétraitement des données, la construction de modèles d'apprentissage automatique et l'évaluation des performances des modèles. Nous avons utilisé des fonctionnalités telles que le Local Outlier Factor pour la détection d'anomalies, le StandardScaler pour la mise à l'échelle des données, ainsi que divers algorithmes de classification tels que RandomForestClassifier, SVM et XGBoost.

-Yellowbrick pour la visualisation des performances des modèles, proposant une variété de graphiques tels que la matrice de confusion, le rapport de classification, les erreurs de prédiction de classe et la courbe ROC-AUC.

-Plotly pour la création de visualisations interactives et dynamiques, offrant une expérience immersive pour explorer les données et les résultats de la classification.

Grâce à l'utilisation de ces outils, nous avons pu mener une analyse approfondie des données célestes, construire des modèles de classification robustes et interpréter efficacement les résultats obtenus

Le fichier que nous avons utilisé pour classifier nos données était présenté de la sorte :

Les données comprennent les éléments suivants

-obj\_ID : Identifiant unique de l'objet dans le catalogue d'images utilisé par le CAS.

-alpha : Ascension droite (à l'époque J2000).

-delta : Déclinaison (à l'époque J2000).

-u, g, r, i, z : Filtres ultraviolet, vert, rouge, proche infrarouge et infrarouge respectivement -dans le système photométrique.

-run\_ID : Numéro d'exécution utilisé pour identifier le balayage spécifique.

-rereun\_ID : Numéro de réexécution pour spécifier comment l'image a été traitée.

-cam\_col : Colonne de la caméra pour identifier la ligne de balayage dans l'exécution.

-field\_ID : Numéro de champ pour identifier chaque champ.

-spec\_obj\_ID : ID unique utilisé pour les objets spectroscopiques optiques (ce qui -signifie que 2 observations différentes avec le même spec\_obj\_ID doivent partager la classe de sortie).

-class : Classe de l'objet (galaxie, étoile ou quasar).

redshift : Valeur de décalage vers le rouge basée sur l'augmentation de la longueur d'onde.

-plate : ID de la plaque, identifie chaque plaque dans le SDSS.

-MJD : Date julienne modifiée, utilisée pour indiquer quand une donnée SDSS donnée a été prise.

-fiber\_ID : ID de fibre qui identifie la fibre ayant pointé la lumière sur le plan focal dans chaque observation.

Après avoir analysé les données, la première chose qui a été observé est qu'il y a plus de galaxies que d'étoiles et de quasars.

Les corrélations fournies entre la variable cible "class" et les autres variables du DataFrame permettent de comprendre la relation linéaire entre ces variables. Une corrélation négative indique une relation où l'augmentation d'une variable est associée à la diminution de la probabilité de la classe "GALAXY", tandis qu'une corrélation positive suggère que l'augmentation d'une variable est associée à une augmentation de la probabilité des classes "STAR" ou "QSO" (quasar). Les valeurs proches de zéro indiquent une corrélation nulle, ce qui signifie qu'il n'y a pas de relation linéaire entre la variable et la cible. Les corrélations permettent d'identifier les variables les plus influentes dans la prédiction de la classe des objets, aidant ainsi à orienter les analyses et la modélisation des données.

Ensuite , pour pouvoir visualisé la classification nous avons utilisé les trois modèles suivants :

ConfusionMatrix

La matrice de confusion est un outil essentiel pour évaluer les performances d'un modèle de classification. Elle permet de visualiser le nombre de prédictions correctes et incorrectes faites par le modèle pour chaque classe.

Interprétation : La matrice de confusion présente les prédictions correctes sur la diagonale principale et les erreurs de classification hors de la diagonale. Cela permet d'identifier les classes mal prédites et les types d'erreurs que le modèle commet (faux positifs, faux négatifs).

Utilisation : La matrice de confusion est utile pour évaluer la sensibilité (recall), la spécificité et la précision du modèle pour chaque classe.

ClassificationReport

Le rapport de classification fournit une analyse plus détaillée des performances du modèle en affichant des mesures de précision, de rappel, de F1-score et de support pour chaque classe.

Interprétation : Le rapport de classification fournit une vue globale des performances du modèle en termes de précision (capacité à ne pas classer à tort en positif), de rappel (capacité à identifier tous les vrais positifs), de F1-score (moyenne harmonique de la précision et du rappel) et de support (nombre d'occurrences réelles de chaque classe dans l'ensemble de données).

Utilisation : Il permet de comparer les performances du modèle entre les différentes classes et de déterminer si le modèle a des difficultés avec une classe particulière.

ClassPredictionError

L'erreur de prédiction de classe permet de visualiser les instances mal classées et de comprendre les motifs d'erreur du modèle.

Interprétation : Cette visualisation met en évidence les classes qui sont souvent confondues par le modèle. Elle permet d'identifier les classes pour lesquelles le modèle a des difficultés à distinguer.

Utilisation : Elle aide à comprendre les erreurs spécifiques du modèle et à identifier les classes pour lesquelles des améliorations sont nécessaires.

ROCAUC

La courbe ROC (Receiver Operating Characteristic) et l'aire sous la courbe ROC (ROC-AUC) sont des mesures de la performance d'un modèle de classification binaire.

Interprétation : La courbe ROC représente le taux de vrais positifs en fonction du taux de faux positifs pour différents seuils de classification. L'aire sous la courbe ROC (ROC-AUC) mesure la capacité du modèle à discriminer entre les classes.

Utilisation : La courbe ROC et l'ROC-AUC sont utiles pour évaluer la performance d'un modèle de classification binaire et comparer différents modèles entre eux.

En utilisant ces différentes visualisations ensemble, les praticiens peuvent obtenir une compréhension approfondie des performances du modèle, identifier les classes problématiques, comprendre les types d'erreurs effectuées par le modèle et évaluer sa capacité à discriminer entre les classes. Cela permet une analyse plus complète et une meilleure prise de décision pour l'amélioration du modèle.

Un rapport de classification offre une vue complète des performances d'un modèle de classification. Il permet de comprendre comment le modèle se comporte pour chaque classe individuelle, en tenant compte à la fois des erreurs de type I et de type II, ce qui est essentiel pour évaluer la qualité globale du modèle.

Voici pourquoi chaque élément est inclus dans un rapport de classification :

Précision : La précision mesure la proportion de prédictions positives du modèle qui étaient réellement correctes. C'est important car elle nous donne une idée de la fiabilité des prédictions positives du modèle. Une haute précision est cruciale lorsque les faux positifs sont coûteux ou indésirables.

Rappel: Le rappel mesure la proportion de vrais positifs qui ont été correctement identifiés par le modèle. C'est essentiel car il nous indique la capacité du modèle à détecter correctement les instances positives réelles. Un rappel élevé est crucial lorsque les faux négatifs sont coûteux ou indésirables.

F1-score : Le F1-score est une mesure de la précision et du rappel, calculée comme la moyenne harmonique entre ces deux mesures. Il offre une évaluation globale de la performance du modèle, en tenant compte à la fois de la précision et du rappel. C'est une métrique utile pour évaluer l'équilibre entre la précision et le rappel.

Support : Le support représente le nombre total d'instances de chaque classe dans les données de test. Cela permet de mettre en contexte les autres mesures en fournissant une indication de la distribution des classes dans l'ensemble de données. Cela est important pour comprendre la représentativité des prédictions du modèle.

En conclusion, notre analyse a couvert plusieurs aspects essentiels de l'apprentissage automatique appliqué à la classification des objets stellaires. Nous avons exploré en détail les données, utilisant des bibliothèques spécialisées telles que Matplotlib, Seaborn, Scikit-learn, Yellowbrick et Plotly pour visualiser, prétraiter et modéliser les données. Nous avons également utilisé différents modèles de classification, tels que le SVC, le RandomForestClassifier et le XGBoostClassifier, pour construire des solutions robustes.

En évaluant les performances de nos modèles, nous avons utilisé une gamme d'outils, y compris la matrice de confusion, le rapport de classification, l'erreur de prédiction de classe et la courbe ROC-AUC, pour obtenir une compréhension approfondie de la capacité prédictive de nos modèles. Chaque outil nous a fourni des informations précieuses sur les forces et les faiblesses de nos modèles, nous permettant d'identifier les domaines où des améliorations sont nécessaires.

Reste dans le fichier jupyter