BAYET Hugo
MARIE Valentin
MAXANT Alexandre

Rendu Atelier composant Metier

dépot git :https://github.com/anderbro/EPSI_AtelierComposantMetier

Intro :

L'objectif est d'analyser le jeu de données fourni, et de réussir à trouver quoi prédire et à prédire un champ choisi. Nous allons donc essayer de voir si le lieu, le type d'établissement influe la note d'évaluation de l'établissement.

Hypothèse :

On a pensé qu'il serait intéressant de prédire le résultat de Synthese_eval_sanit d'après les autres données dont nous disposons.

Préparation des données :

Pour cet exercice nous n'avons pas trouvé pertinent d'avoir certaines colonnes qui référencent des informations que l'on ne trouvait pas pertinentes. On à notamment enlevé les colonnes suivantes :

- APP_Libelle_etablissement
- Adresse_2_UA
- Numero_inspection
- Agrement

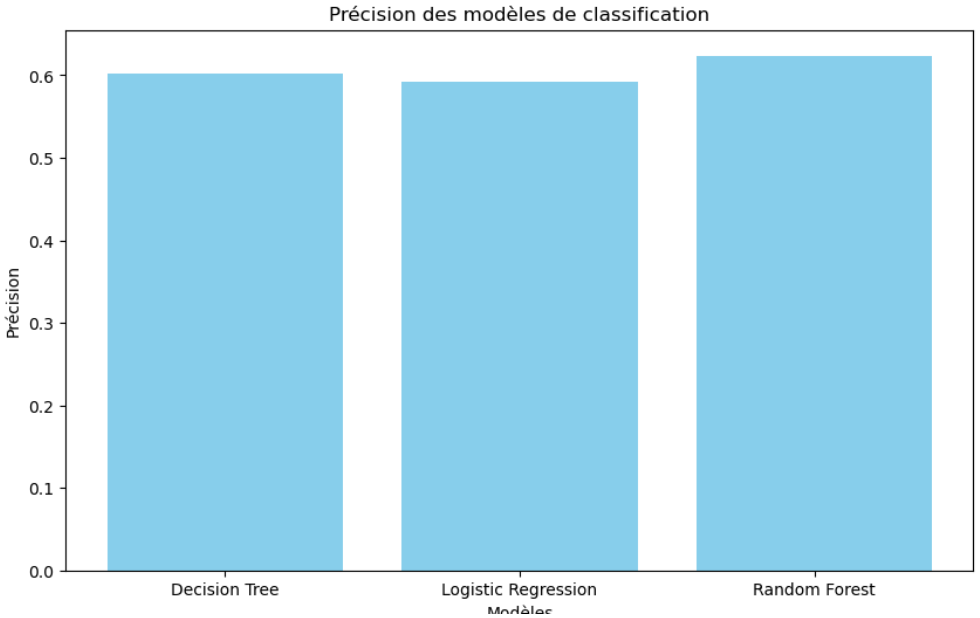
Pour la suite de l'exercice il nous à aussi fallu trier les données. On à pris le parti de supprimer les données qui ne correspondaient pas au format attendu. Par exemple, certains formats de département ne correspondaient à rien. De plus, pour certaines lignes il y avait plusieurs champs vides, nous avons aussi pris le parti de les supprimer. Nous avons fait la plupart de ces changements à la main ca très minimes, mais il aurait fallu à terme utiliser python ou par exemple un ETL comme Talend ou FME.

Pratique :

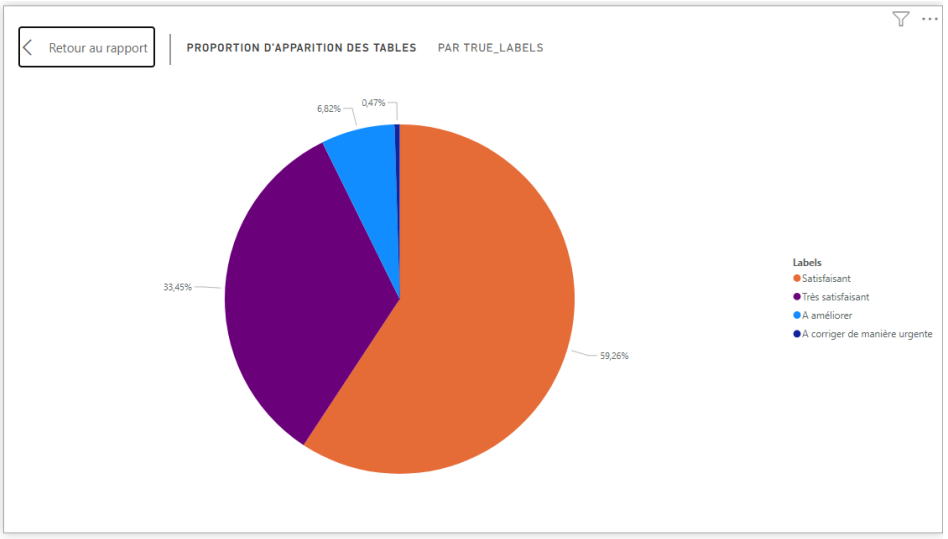
On a testé 3 différents algo de prédictions :

- Decision tree
- Logistic regression model
- Random forest

On peut voir ci-dessous les différents degrés de précisions des 3 algos testés.
On peut donc voir que la prédiction à l'aide d'un algorithme Random Forest semble avoir le meilleur résultat avec les paramètres énoncés ci dessous.




```
Decision Tree Accuracy: 0.6024224394390141
Decision Tree Classification Report:
                              precision    recall  f1-score   support

                   A améliorer       0.15      0.12      0.14       321
A corriger de manière urgente       0.00      0.00      0.00        22
                  Satisfaisant       0.67      0.77      0.72      2789
             Très satisfaisant       0.54      0.41      0.46      1574

                      accuracy                           0.60      4706
                     macro avg       0.34      0.33      0.33      4706
                  weighted avg       0.58      0.60      0.59      4706

Logistic Regression Accuracy: 0.5926476838079048
Logistic Regression Classification Report:
                              precision    recall  f1-score   support

                   A améliorer       0.00      0.00      0.00       321
A corriger de manière urgente       0.00      0.00      0.00        22
                  Satisfaisant       0.59      1.00      0.74      2789
             Très satisfaisant       0.00      0.00      0.00      1574

                      accuracy                           0.59      4706
                     macro avg       0.15      0.25      0.19      4706
                  weighted avg       0.35      0.59      0.44      4706

Random Forest Accuracy: 0.6236719082022949
Random Forest Classification Report:
                              precision    recall  f1-score   support

                   A améliorer       0.15      0.06      0.08       321
A corriger de manière urgente       0.00      0.00      0.00        22
                  Satisfaisant       0.67      0.80      0.73      2789
             Très satisfaisant       0.55      0.43      0.48      1574

                      accuracy                           0.62      4706
                     macro avg       0.34      0.32      0.32      4706
                  weighted avg       0.59      0.62      0.60      4706
```



Data-vis :

Nous avons utilisé l'outil power bi pour une meilleure visualisation des résultats de la prédiction réalisée par notre modèle.
On peut voir que la prédiction semble plutôt proche des valeurs de départ.