Avila analysis

Python for data analysis - Project

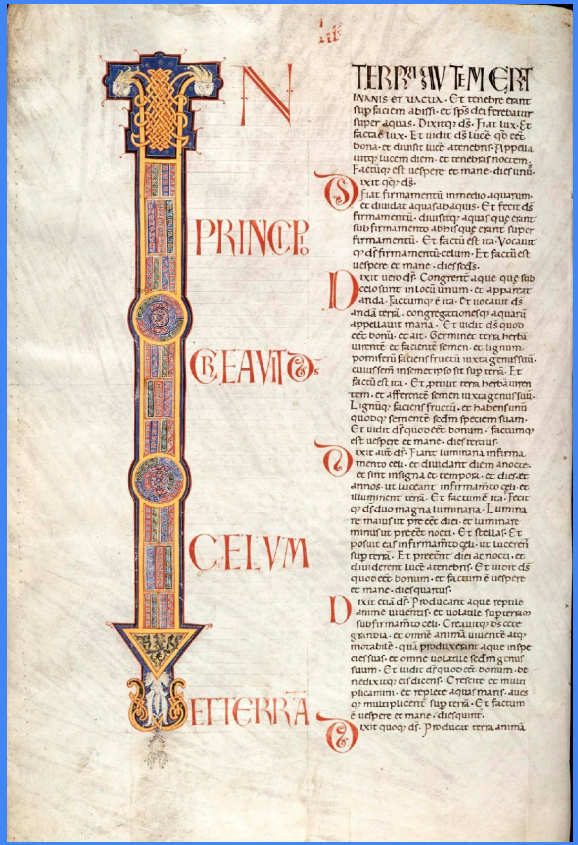
Marc-Etienne Dartus | Caillieux Nicolas

Le dataset

Le dataset représente des caractéristiques sur les **pages**, **colonnes** et **lignes** de la bible Avila provenant d'images.

Les pages sont écrites en 2 colonnes de plusieurs lignes.

Chaque ligne du dataset représente les caractéristiques d'une ligne écrite, de la colonne qui la contient, et de la page qui la contient.



Le problème

Le but de la **prédiction** sur ces données, est de pouvoir prédire *quel copiste a écrit une ligne donnée, en fonction des différents patterns de ce texte.*

La Bible ayant été écrite par 12 copistes, assimilés à des lettres, le dataset dispose de 12 **classes** de **cible**.

Il s'agit bien ici d'appliquer un algorithme de **classification**, pour associer des patterns d'écritures à un des copistes ayant écrit la bible.

Les données
La cible

La colonne cible de notre classification sera donc la colonne **label,** qui correspondra à une lettre représentant 1 des 12 copistes:

A, B, C, D, E, F, G, H, I, W, X, Y

Les données
Les valeurs

Le dataset dispose de 10 valeurs *Z-normalisées*, que nous avons apprit à comprendre en détail :

- **intercolumnar distance** : Distance entre les 2 colonnes de la page
- **upper margin** : Marge en haut de page
- **lower margin** : Marge en bas de page
- **exploitation** : Remplissage d'encre de la colonne
- **row number** : Nombre de lignes de la colonne
- **modular ratio** : Ratio de la hauteur/largeur des caractères
- **interlinear spacing** : Espace entres les lignes
- **weight :** Remplissage d'encre d'une ligne
- **peak number** : Nombre de pics d'encre d'une ligne si on projette ses pixels sur l'axe vertical
- **modular ratio / interlinear spacing:** Simple ratio de 2 valeurs précédentes

Les données
Création de données

Après avoir étudié les données pour chaque valeur de plus près, nous avons cherché à en créer de nouvelles, qui seraient révélatrice d'information utile au model.

Ces réflexions ont donné naissance à 2 nouvelles colonnes :

- **mrg_ratio** : ratio de *upper margin / lower margin*.
  → *Créée dans une logique de rassembler l'info sur les marges*

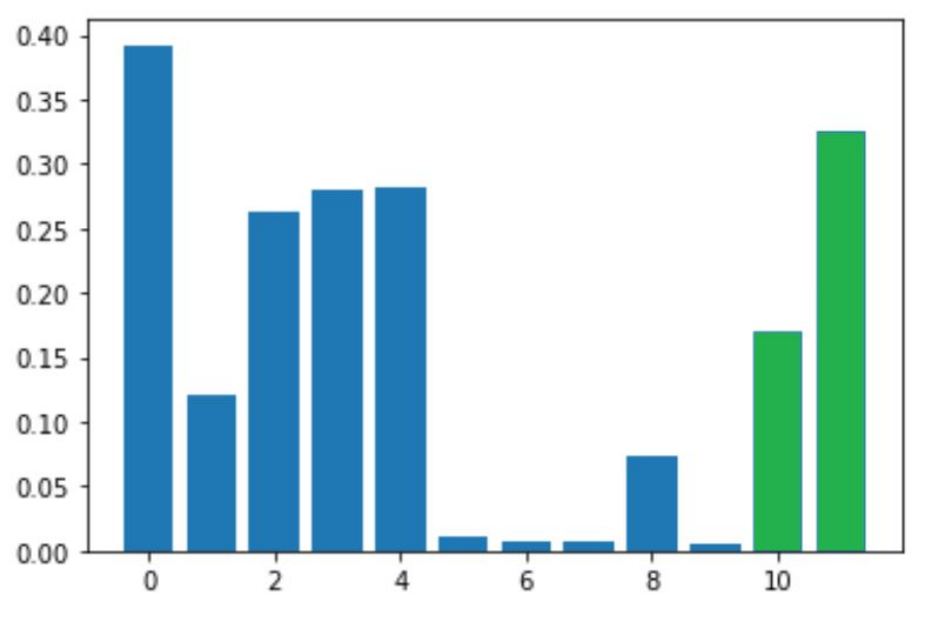
- **spacing_ration** : ratio de *intercolumnar distance / (upper margin + lower margin).*
  → *Créée dans une logique de rassembler l'info sur l'espacement sur la page de façon générale.*

Les données
Création de données

Nous avons procédé à l’évaluation de la pertinence de ces nouvelles colonnes via une Feature Importance permutée des différents modèles.

Le résultat est plus ou moins satisfaisant selon les modèles, mais ces colonnes ressortent pertinentes d’une manière générale.

```
--- Permuted Feature importance ---
Feature: intercolumnar_dist, Score: 0.39229
Feature: upper_mrg, Score: 0.12086
Feature: lower_mrg, Score: 0.26270
Feature: exploit, Score: 0.28090
Feature: row_num, Score: 0.28117
Feature: modular_ratio, Score: 0.01116
Feature: spacing, Score: 0.00642
Feature: weight, Score: 0.00675
Feature: peak_num, Score: 0.07415
Feature: modular/spacing, Score: 0.00539
Feature: spacing_ratio, Score: 0.17032
Feature: mrg_ratio, Score: 0.32462
```



Les modèles appliqués

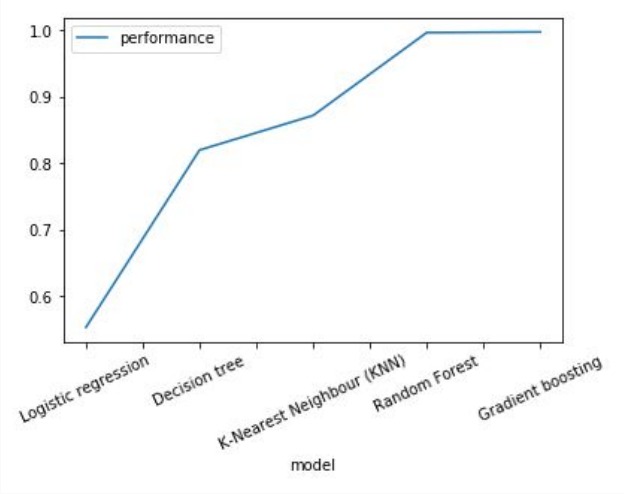
Une fois passé les méthodes de découverte et de feature engineering, vient la résolution du problème, que nous avons abordée via différents algorithmes de classification:

- KNN
- Random Forest
- Decision Tree
- Logistic Regression
- Gradient Boosting

Nous avons également cherché à affiner les résultats avec un paramétrage automatique des hyper-paramètres de certains modèles.

Résultats

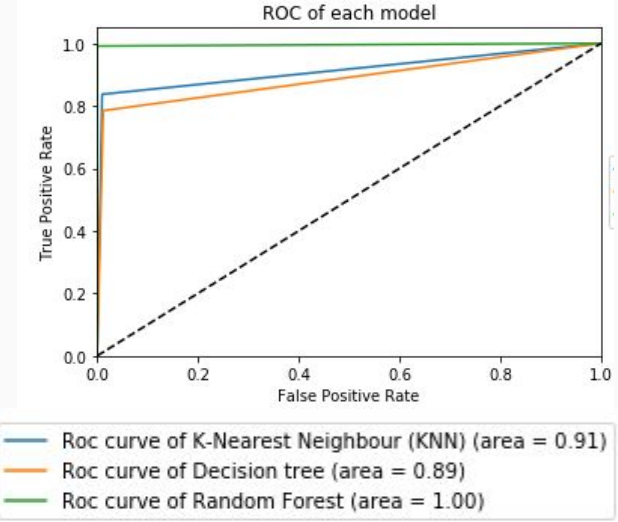
Après paramétrage et exécution des différents modèle, nous arrivon à répondre efficacement au problème avec des prédictions précises:



Accuracy:

Random Forest : 0.99

Gradient Boosting : 0.99



Conclusion

Expérience

Ce projet est une application intéressante de machine learning. Car bien qu'il soit difficile à l'oeil humain de différencier les auteurs d'une certaine habitude d'écriture et calligraphie, le machine learning y arrive particulièrement bien, du moins avec le dataset Avila.

Apprentissage

L'exploration des données, leur visualisation et leur compréhension à été une étape importante dans l'élaboration du modèle. Cela nous a permis de comprendre en profondeur chaque variable de façon concrète, et par la suite de créer de l'information utile pour le modèle.

Résultat

En plus de ce feature engineering, il nous a suffit de tester plusieurs modèles, d'effectuer des recherches d'hyper-paramètres optimaux, et le résultat retenu est d'une précision très satisfaisante.