ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

# Analyse des données d'Avila

Python for data analysis

# Introduction

La Bible d'Avila est un manuscrit du XIIe siècle dont les origines remontent à la région ombro-romaine d'Italie.

C’est un codex de grandes dimensions qui contient l'Ancien et le Nouveau Testament.

Elle est aujourd’hui entretenu dans le musée national d’Espagne, à Madrid.

# Description des données

L'ensemble des données d'Avila est une extraction de 800 images de la « Bible d'Avila ».

Elle se présente sous forme de tableau regroupant les caractéristiques lié à l'écriture des pages du manuscrit selon un copiste.

On peux distinguer les caractéristique suivant :

- « intercolumnar distance » (f1) : espacement inter-colonne
- « upper margin » (f2) : marge supérieur
- « lower margin » (f3) : marge inferieur
- « exploitation » (f4) :
- « row number » (f5) : nombre de lignes
- « modular ratio » (f6) : ratio entre la hauteur et la largeur des caractères
- « interlinear spacing (f7) » : espacement entre les lignes
- « weight » (f8) :
- « peak number » (f9) :
- « modular ratio / interlinear spacing » (f10) : rapport entre le ratio des caractères et l'espacement entre les lignes

Dans l’ensemble des données on y distingue 12 copistes représenté par des labels : A, B, C, D, E, F, G, H, I, W, X et Y

Et les valeurs inscrite dans l’ensemble des données sont normalisé par une Z-normalization

# Objectif de l'analyse

L'objectif de ce projet consiste à associer les caractéristique d'écriture d'une page du manuscrit d'Avila à un copiste et de pouvoir, par la suite, prédire la classification d'une nouvelle entrée donnée.

# Analyse des attributs

Dans un premier temps nous avons étudié les attributs afin d'étudier leur lien avec la classe cible.

Pour se faire nous avons comparé les variances des attributs par type de classe et visionner la corrélations de chaque attributs entre eux.

Nous avons ensuite sélectionné les attributs les plus pertinent selon des critères de corrélation faible et de variances haute

# Résolution du problème

Afin de répondre au problème du sujet nous avons opter pour une classification.

Nous avons étudié 6 modèles de classification afin de trouver le modèle correspondant au mieux à l'ensemble des données. Voici les modèles utilisés : k-NN, Logistic regression, Support Vector Machine (SVM), Random Forest, Decision Tree et Naive Bayes.

Pour chaque modèle nous avons effectué des phases d'entrainement et de comparaison.

- Phase d'entrainement :

Tout d'abord, nous récupérons le score de précision d'un modèle avec les paramètres par défaut pour avoir un avis préliminaire de son efficacité sur l'ensemble des données.

Ensuite, nous effectuons une étape de recherche pour trouver les meilleurs paramètres applicable au modèle.

Enfin, nous comparons le score de précision du modèle avec les paramètres par défaut et avec les nouveaux paramètres pour voir son potentiel.

- Phase de comparaison :

Lorsque l'on a trouvé les meilleurs paramètres associés à chaque modèle nous effectuons un dernier test de comparaison de score afin de récupérer le meilleur modèle.

Nous avons ainsi identifié « Random forest » comme étant le plus performant pour la classification de l'ensemble des données d'Avila.