Avila dataset

Python Project

Akshay
Baba Yaya FALL
ESILV IBO A5

C. De Stefano, M. Maniaci, F. Fontanella, A. Scotto di Freca,
Reliable writer identification in medieval manuscripts through page layout features: The "Avila" Bible case,
Engineering Applications of Artificial Intelligence, Volume 72, 2018, pp. 99-110.

Description of the dataset

- Avila Dataset → extracted form 800 images of the Avila Bible
- 12 copyists determined by a palaeographic analysis of the manuscript labelled as A, B, C, D, E, F, G, H, I, W, X, Y
- Dataset normalized using Z-score normalization
  Divided in a training set of 10430 samples and a testing set of 10437 samples

Description of the dataset

ATTRIBUTE DESCRIPTION (features):

- ID Name
- F1 intercolumnar distance
- F2 upper margin
- F3 lower margin
- F4 exploitation
- F5 row number
- F6 modular ratio
- F7 interlinear spacing
- F8 weight
- F9 peak number
- F10 modular ratio/ interlinear spacing

Description of the dataset

CLASS DISTRIBUTION OF THE TRAINNG SET:

- A: 4286
- B: 5
- C: 103
- D: 352
- E: 1095
- F: 1961
- G: 446
- H: 519
- I: 831
- W: 44
- X: 522
- Y: 266

Goal

- The goal of the study is to identificate a sample to one of the copyist, using the features.

- The Avila dataset is splited in two:
  - 50% of the set will be the training set
  - 50% of the set will be the testing set

Reflexions

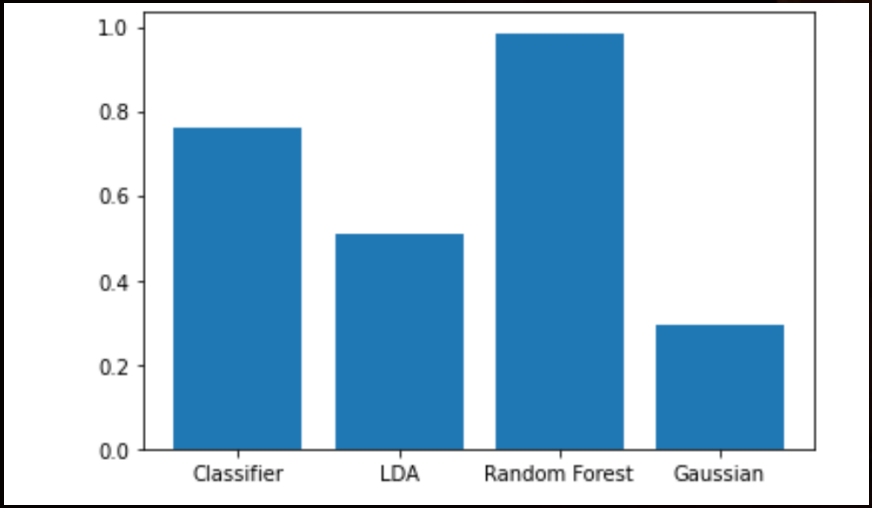
- We first decided to « play » with the data. Indeed, with 10000 hits, we can do a lot.
- We had the chance to have a complete dataset, without missing data.
- It made the data analysis work easier.

Reflexions

- How to predict who is the writer ?

- The path we choose is to compare the accuracy of several machine learning model :
  - The KNeighborsClassifier
  - The Linear Discriminant Analysis projection
  - The Random forest
  - The Gaussian

Model selection



- According to our tests, Random forest is the most efficient model for avila dataset.