The Avila Bible

PYTHON FOR DATA ANALYSIS

Augustin Vendroux
Tristan Solus

A little context

The Avila Bible is one of the largest and most spectacular codices in the Spanish National Library. The ornamentation features contrasting Italian and Spanish styles. The Italian decoration depicts the authors of the books and contains numerous capitals, either illuminated or colored in red, blue, yellow and dark green on lighter backgrounds in the same tone. The colors change in the Spanish decoration, which also has exceptional intertwined initials and whole-page figurative illustrations of subjects such as Noah's Ark, the symbols of the evangelists and scenes from the life of Christ.

Object: Manuscript
Dimensions: 435 pages, 58x39 cm
Remarks: Carolingian script
Date: XIth century

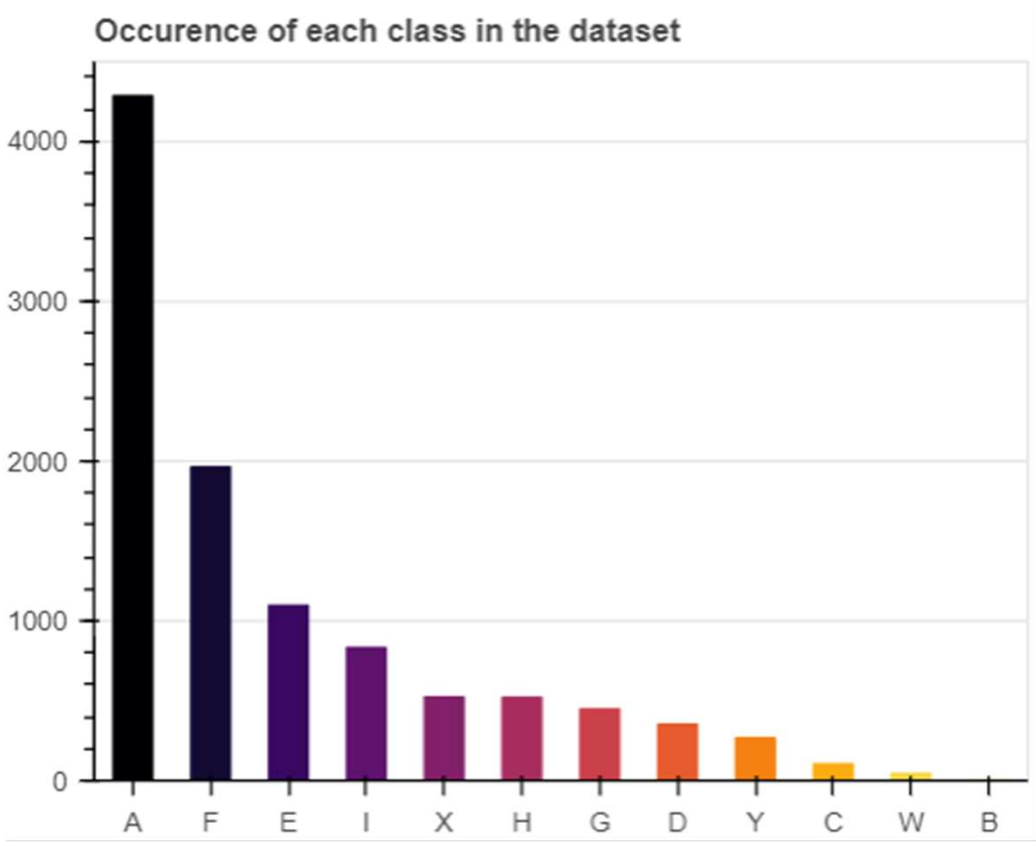
Dataset presentation

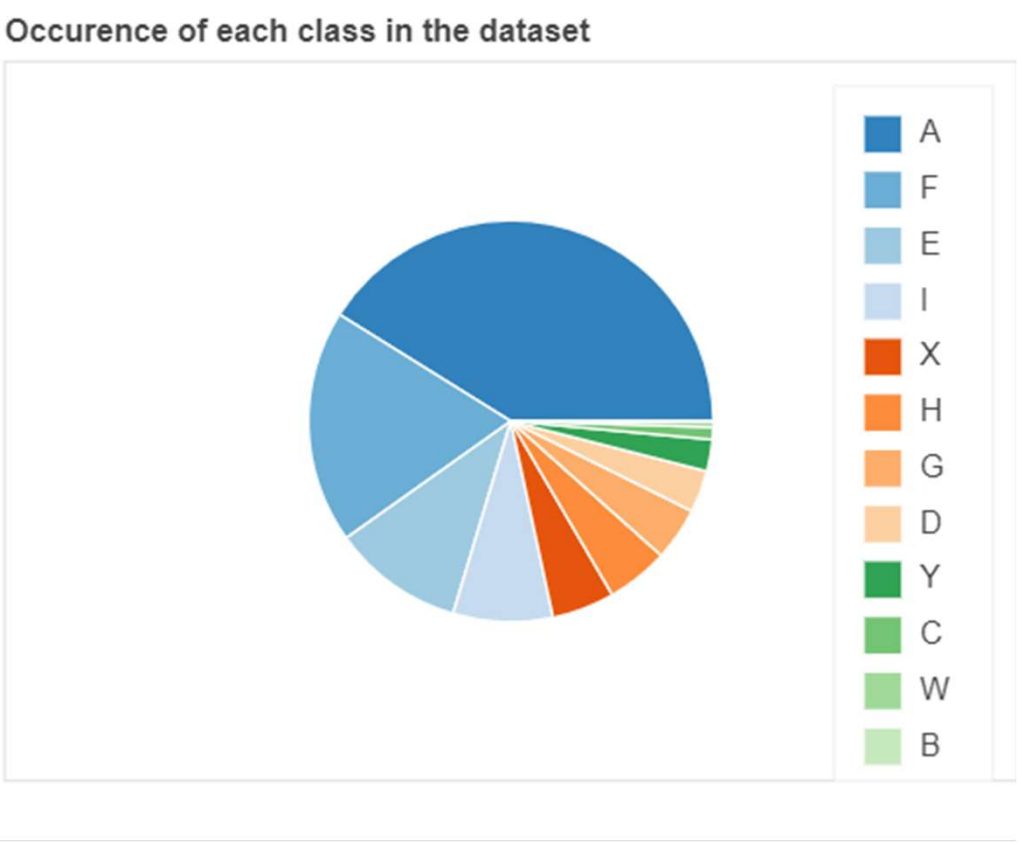
The dataset is composed of 20 867 instances of 10 attributes.
The attributes are the following:

- Intercolumnar distance: distance between two columns of a page
- Upper margin: distance between upper margin of the page and first line of text
- Lower margin: distance between lower margin of the page and last line of text
- Exploitation: fraction of the column filled with ink (ratio of black and white pixels in the same column)
- Row number: number of rows in the current column
- Modular ratio: estimation of the dimension of the handwritten character
- Interlinear spacing: distance between two rows, in pixels
- Weight: fraction of row filled with ink. Analogous to exploitation, but for a single row
- Peak number: estimation of the number of characters in a row
- Modular ratio/interlinear spacing: ratio of the two preceding attributes

The dataset is already split into one training set and one test set of about 10 430 instances each. In the dataset, the instances are given a class corresponding to the monk who wrote the extract. These classes are A, B, C, D, E, F, G, H, I, W, X and Y.

Our objective for this project will be to determine – predict – which monk wrote which extract.

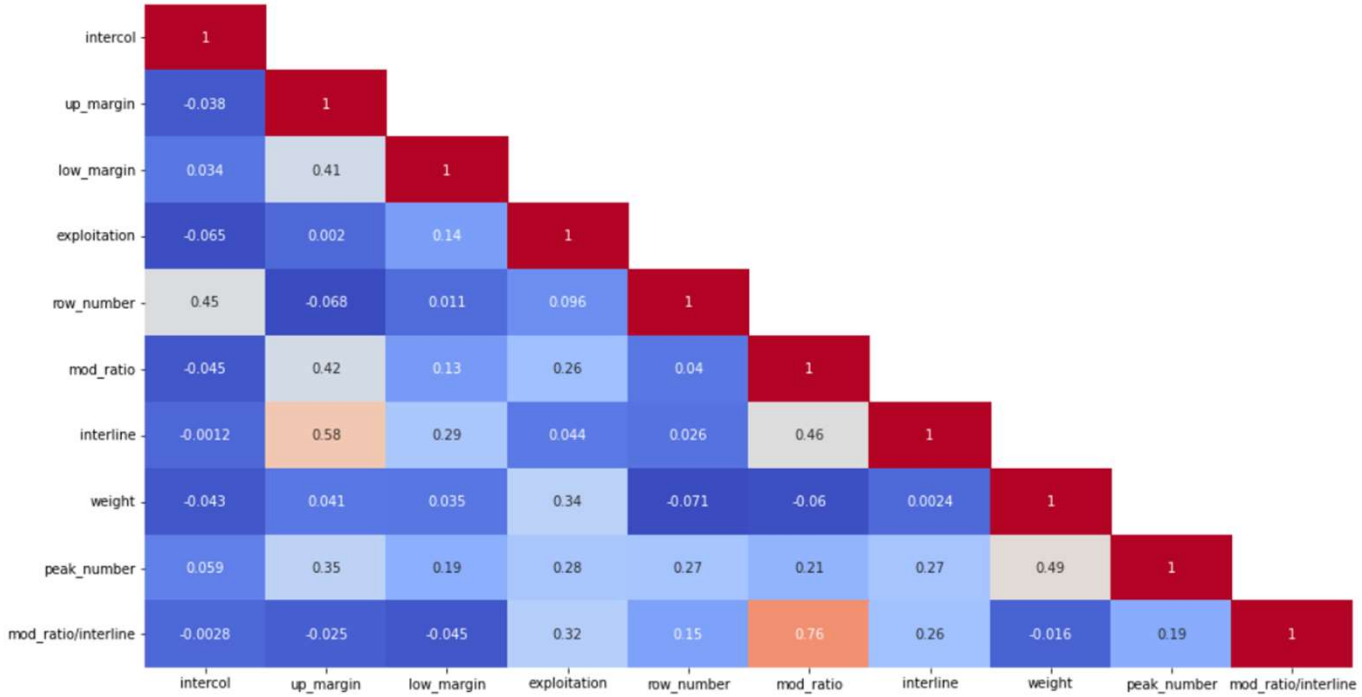




Data Visualization & Transformations

Correlation Matrix

Without any surprises, mod_ratio/ interline and mod_ratio are highly correlated. Moreover mod_ratio/ interline get a very low correlation with all the other data so we can conclude that this value is not very useful, thus, we will get rid of this line. We can also see that peak_number/weight and interline/ up_margin are highly correlated.

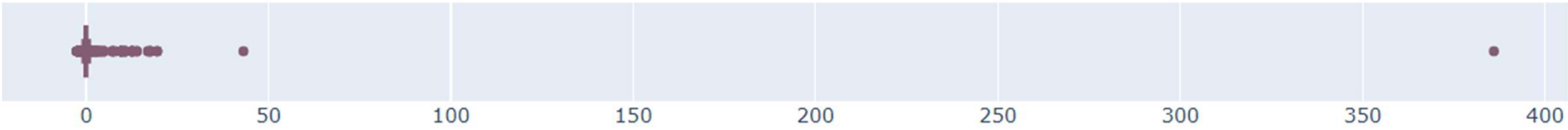


Describe function

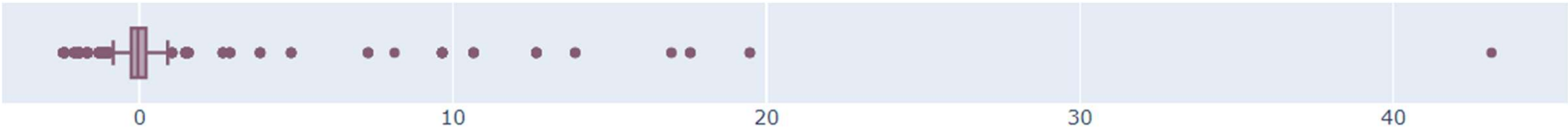
	intercol	up_margin	low_margin	exploitation	row_number	mod_ratio	interline	weight	peak_number	mod_ratio/interline
count	10429.000000	10429.000000	10429.000000	10429.000000	10429.000000	10429.000000	10429.000000	10429.000000	10429.000000	10429.000000
mean	0.000827	0.033630	-0.000556	-0.002433	0.006354	0.013948	0.005570	0.010234	0.012891	0.000803
std	0.991475	3.921056	1.120251	1.008564	0.992100	1.126296	1.313812	1.003515	1.087715	1.007141
min	-3.498799	-2.426761	-3.210528	-5.440122	-4.922215	-7.450257	-11.935457	-4.247781	-5.486218	-6.719324
25%	-0.128929	-0.259834	0.064919	-0.528002	0.172340	-0.598658	-0.044076	-0.542001	-0.372457	-0.516103
50%	0.043885	-0.055704	0.217845	0.095763	0.261718	-0.058835	0.220177	0.111754	0.064084	-0.034621
75%	0.204355	0.203385	0.352988	0.658210	0.261718	0.564038	0.446679	0.654900	0.500624	0.530885
max	11.819916	386.000000	50.000000	3.987152	1.066121	53.000000	83.000000	13.173081	44.000000	4.671232

As our goal is to predict which monk is writing, we evaluate the quality of the data set using the function .describe(). The dataset is supposed to be standardized, but we can see that the standard deviation of the upper margin is way above 1 and the max is 386. It probably means there is a value we need to get rid of.

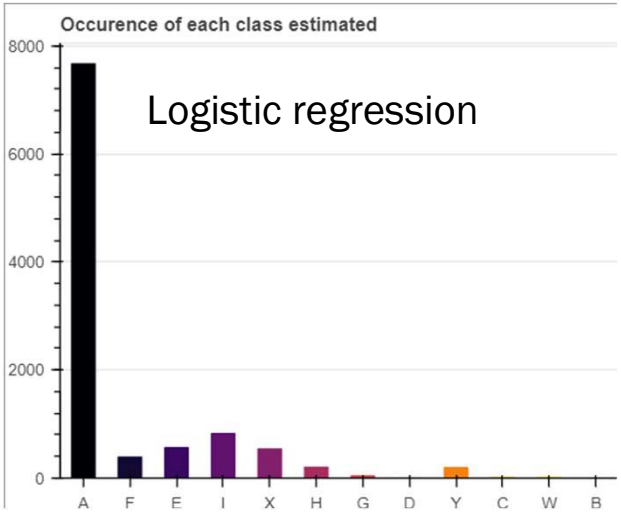
Cleaning the dataset

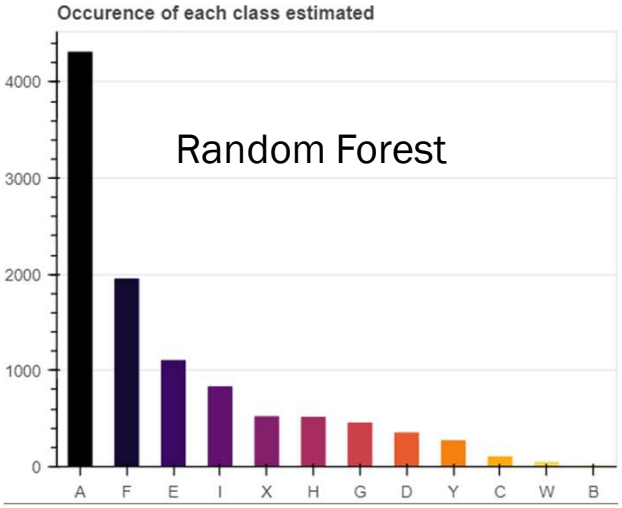


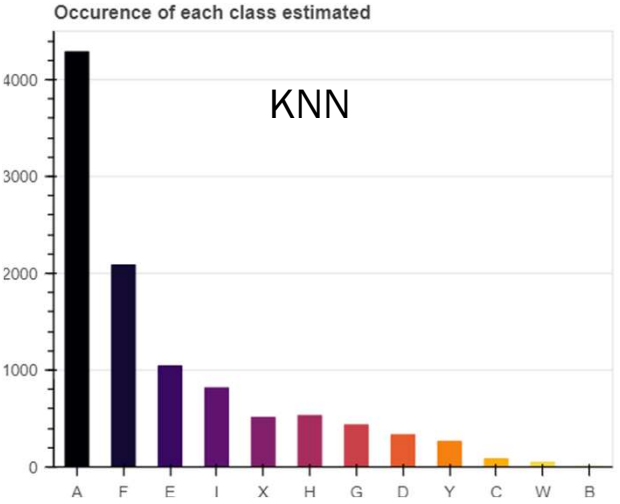
In order to detect any abnormal value, we show a boxplot. It is clear now that we have a singularity we need to delete prior to any other manipulation.

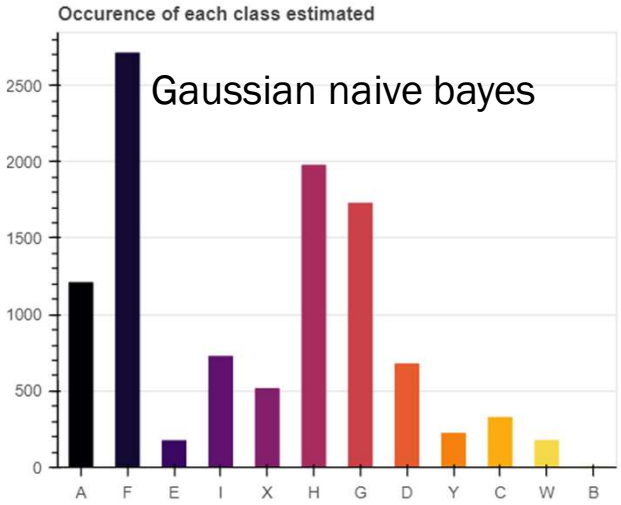


After deleting the value, we plot another boxplot to make sure our modification is efficient.

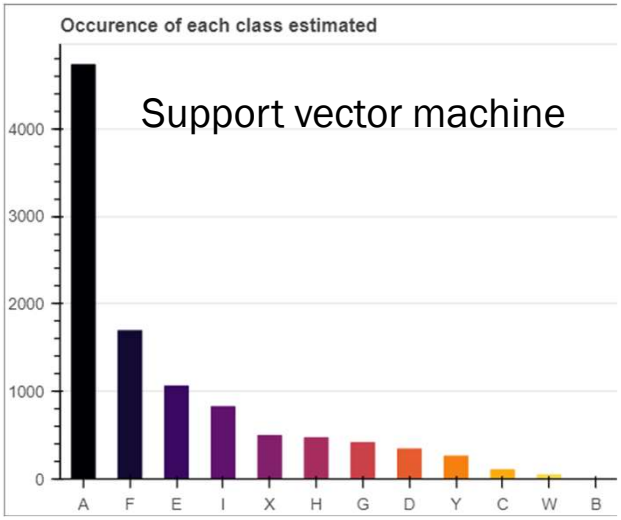






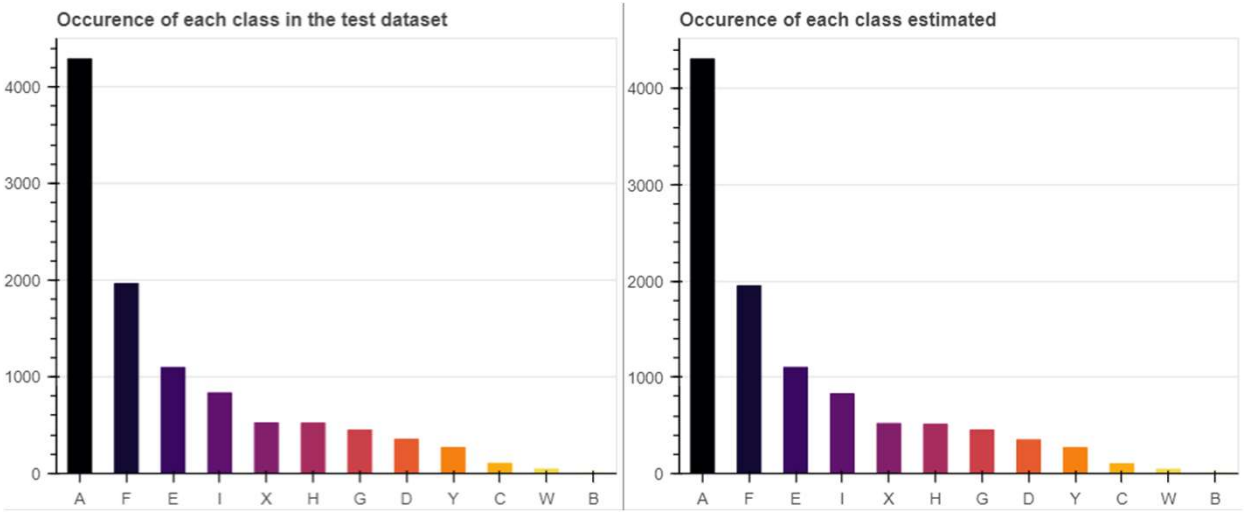


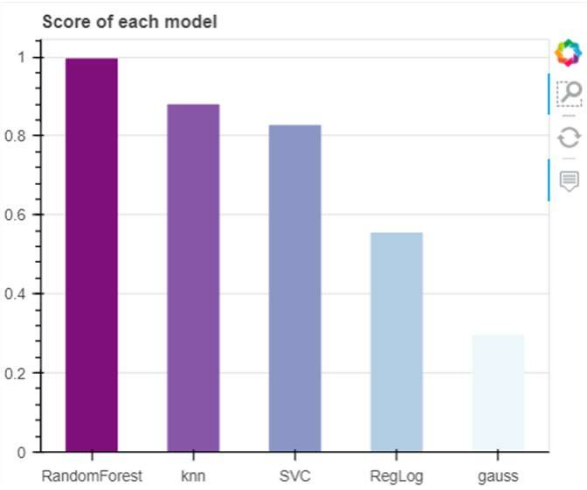
Models predictions



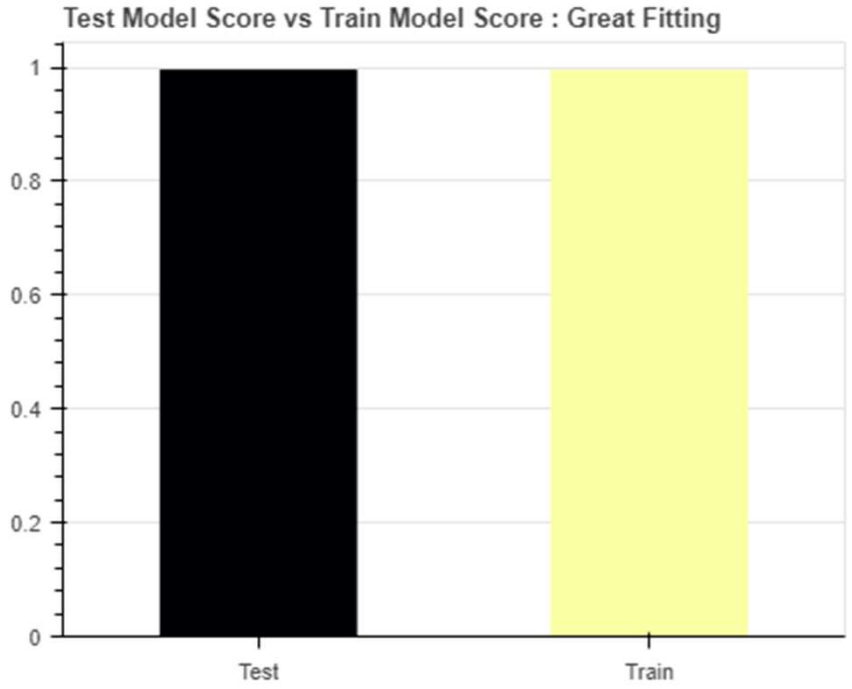
Models accuracy comparison

It appears that the best fitting model is the Random Forest, with an accuracy close to 100%.





Overfitting test



As the accuracy is really high, we suspect an overfitting, which we try to detect using the test_fitting() function.

As we can see, the result is really satisfying, we managed to avoid overfitting using cross validation.

Sources:

Avila Bible description: Avila bible, Spanish National Library, Madrid at Spain is culture.
Dataset description: Jupyter Notebook Viewer (nbviewer.org)