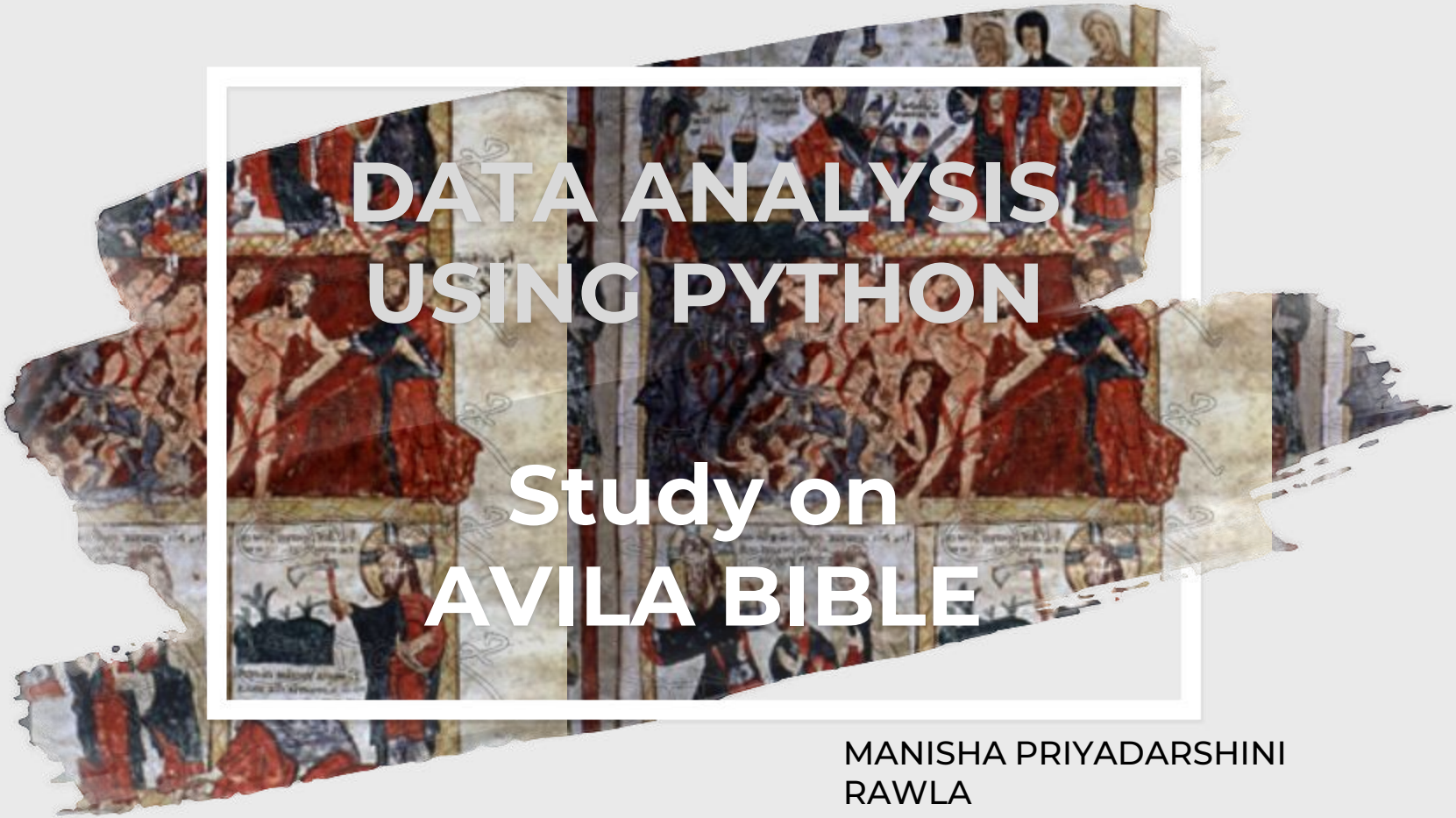


DATA ANALYSIS USING PYTHON

Study on AVILA BIBLE

MANISHA PRIYADARSHINI RAWLA

THE AVILA DATASET

- **The Avila** data set has been extracted from 800 images of the "Avila Bible", a giant Latin copy of the whole Bible produced during the XII century between Italy and Spain.
- The **paleographic analysis** of the manuscript has individuated the presence of 12 copyists. The pages written by each copyist are not equally numerous.
- Each pattern contains 10 features and corresponds to a group of 4 consecutive rows.
- The prediction task consists in associating each pattern to one of the 12 copyists (labeled as: A, B, C, D, E, F, G, H, I, W, X, Y).
- The data have has been normalized, by using the Z normalization method, and divided in two data sets: a training set containing 10430 samples, and a test set containing the 10437 samples.

DATA VISUALISATION

As said earlier, this dataset is composed of 10 different features that each represent a particularity of a copyist.

The different features are :

F1 : Inter columnar distance

F2: upper margin

F3 : lower margin

F4: exploitation

F5: row number

F6 : modular ratio

F7: interlinear spacing

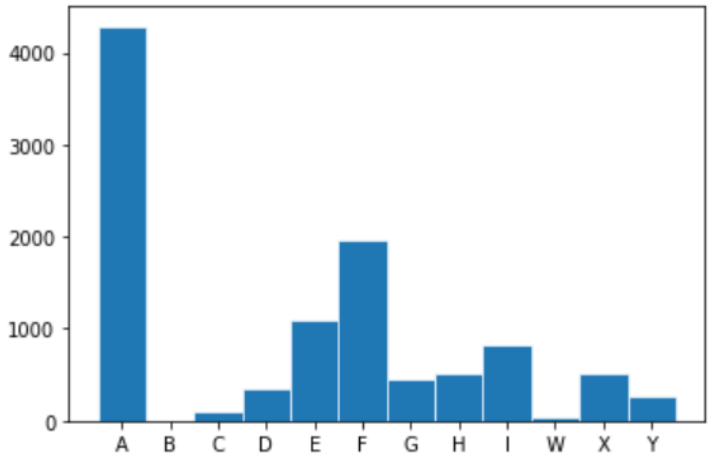
F8 : weight

F9: peak number

F10 : modular ratio/interlinear spacing

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Copyist
0	0.266074	-0.165620	0.320980	0.483299	0.172340	0.273364	0.371178	0.929823	0.251173	0.159345	A
1	0.130292	0.870736	-3.210528	0.062493	0.261718	1.436060	1.465940	0.636203	0.282354	0.515587	A
2	-0.116585	0.069915	0.068476	-0.783147	0.261718	0.439463	-0.081827	-0.888236	-0.123005	0.582939	A
3	0.031541	0.297600	-3.210528	-0.583590	-0.721442	-0.307984	0.710932	1.051693	0.594169	-0.533994	A
4	0.229043	0.807926	-0.052442	0.082634	0.261718	0.148790	0.635431	0.051062	0.032902	-0.086652	F
...	...	...	...	...	...	...	...	...	...	...	...
10425	0.080916	0.588093	0.015130	0.002250	0.261718	-0.557133	0.371178	0.932346	0.282354	-0.580141	F
10426	0.253730	-0.338346	0.352988	-1.154243	0.172340	-0.557133	0.257927	0.348428	0.032902	-0.527134	F
10427	0.229043	-0.000745	0.171611	-0.002793	0.261718	0.688613	0.295677	-1.088486	-0.590727	0.580142	A
10428	-0.301743	0.352558	0.288973	1.638181	0.261718	0.688613	0.069175	0.502761	0.625350	0.718969	E
10429	-0.104241	-1.037102	0.388552	-1.099311	0.172340	-0.307984	0.786433	-1.337547	0.999528	-0.551063	X

- Composition of the training dataset

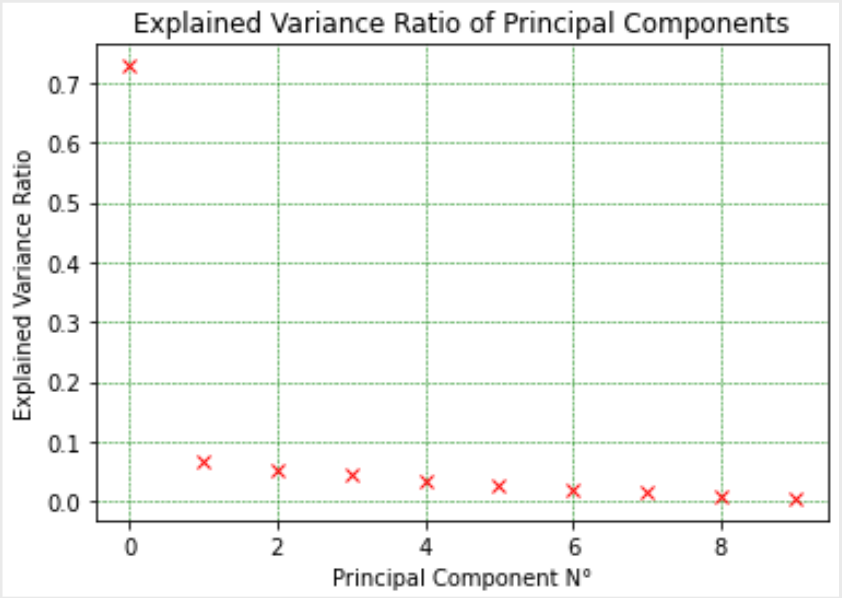


Distribution of samples by copyists

PCA AND LINK BETWEEN THE FEATURES

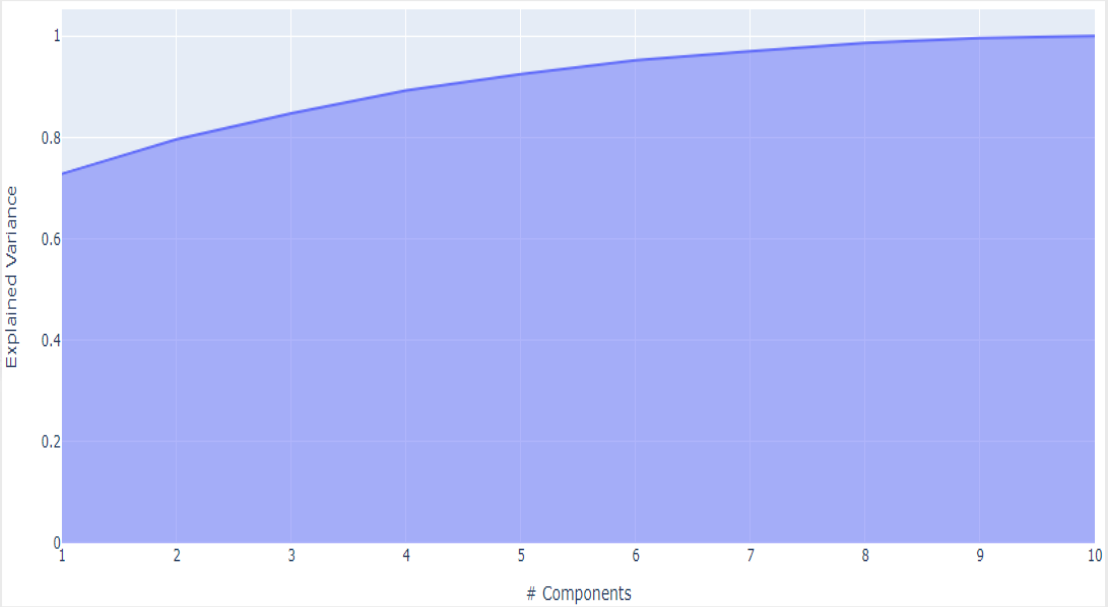
PCA makes maximum variability in the dataset more visible by rotating the axes.

PCA identifies a list of the principal axes to describe the underlying dataset before ranking them according to the amount of variance captured by each.



Explained variance ratio of the first 10 components

On this dataset, nearly 80% of the total variance is explained only by using the 2nd components of the PCA.

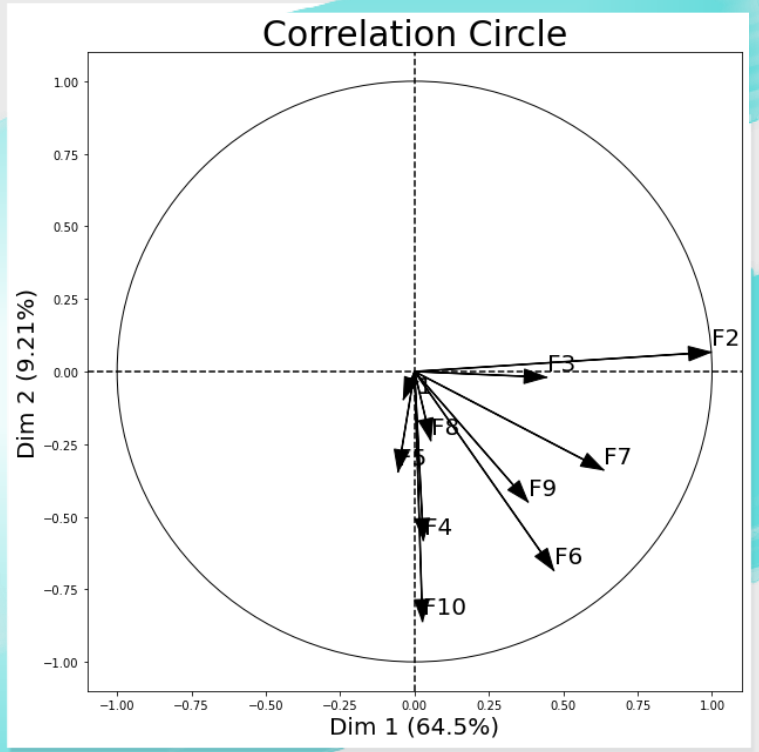


- Cumulative Explained variance ratio of the first 10 components

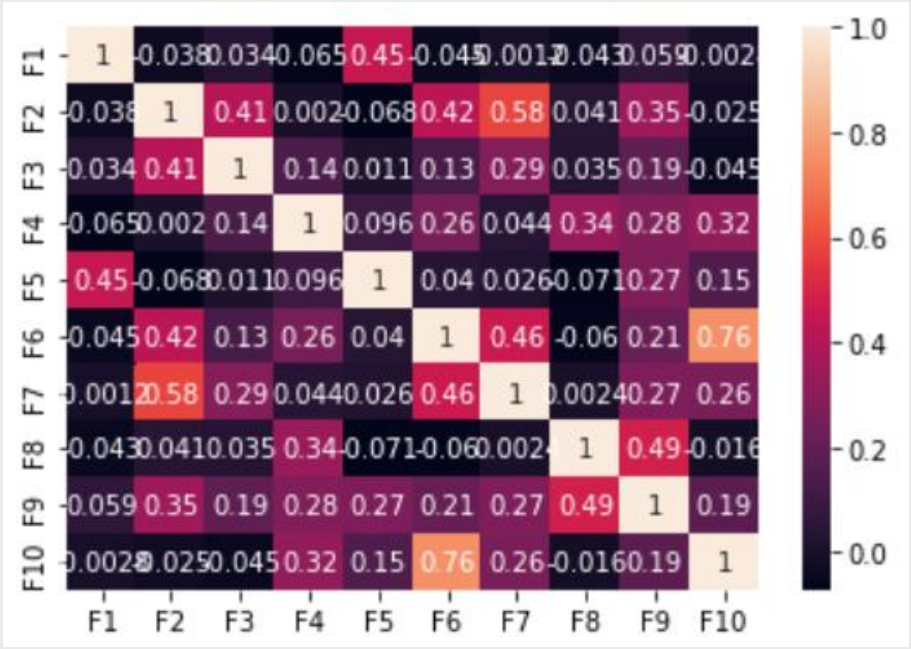
- Correlation circle help us to visualize the correlations between the original dataset features. The principal components are shown via coordinates.

- And this was predictable, because, for instance, **F6** and **F7** are representing the modular ratio and inter linear spacing of the copyist whereas **F10** is a feature composed of the modular ratio over the inter linear spacing, so **F10** is closely related to both features because it is directly composed by them.

- The correlation circle shows that some features are very correlated like **F10** and **F6** or **F7** and **F6**, but some aren't correlated at all like **F10** and **F2** or **F4** and **F3**.



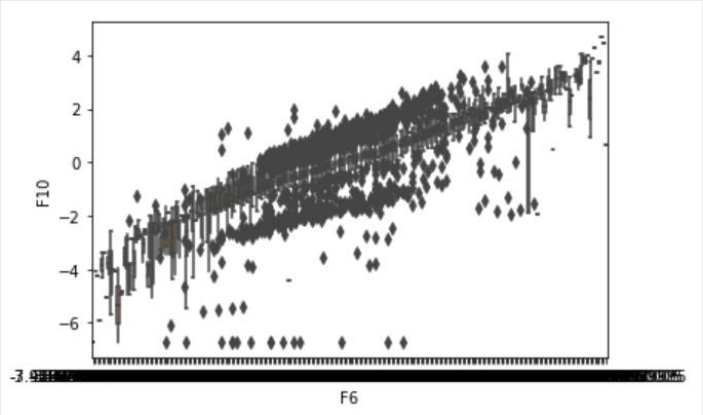
With the correlation matrix of the two datasets, earlier assumptions are now settled.



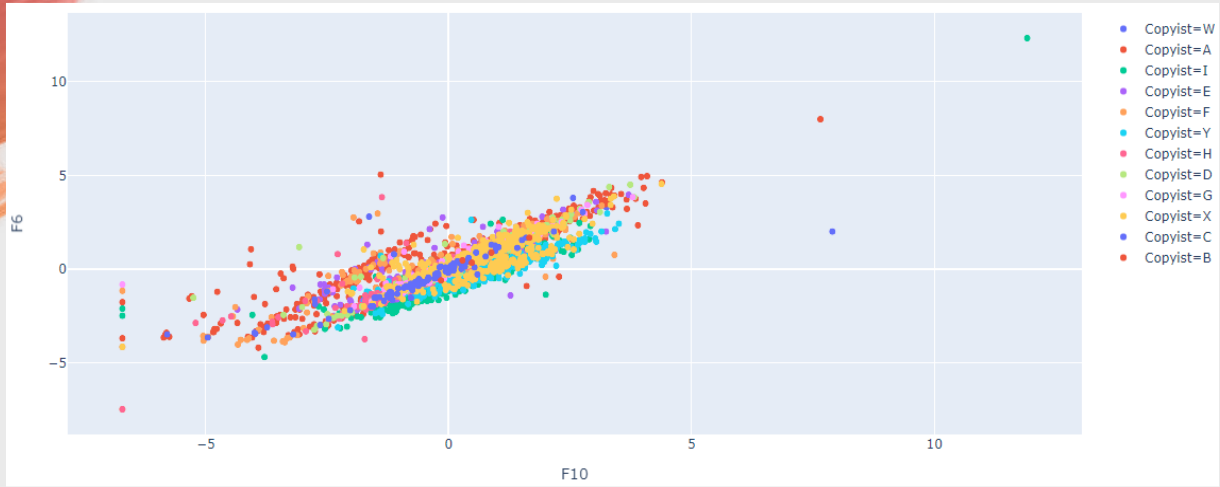
Correlation Matrix

F10, **F6** and **F7** are strongly correlated. But we discover other surprising correlations like between **F5** and **F1** which correspond to row number and Inter columnar distance Or between **F2** and **F7** which correspond to upper margin and interlinear spacing.

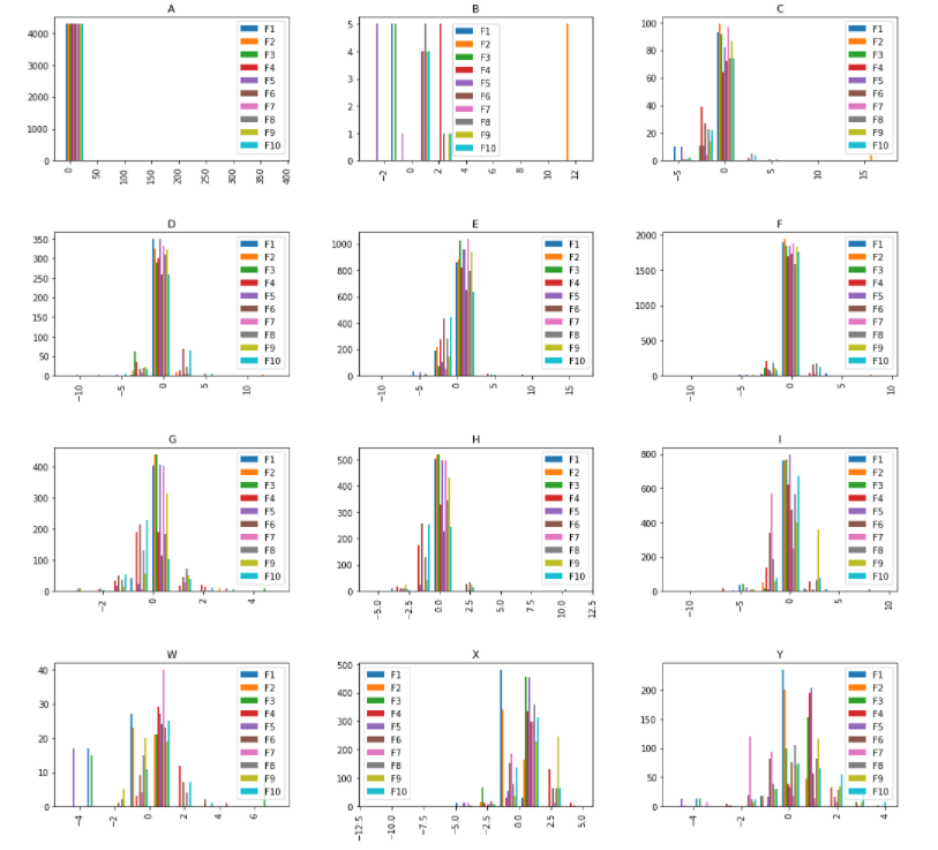
To illustrate the correlation between F6 and F10 we can use the following plots.



Seaborn plot of F10 correlation with F6



Plotly Express Scatter of the linear correlation between F6 and F10 features

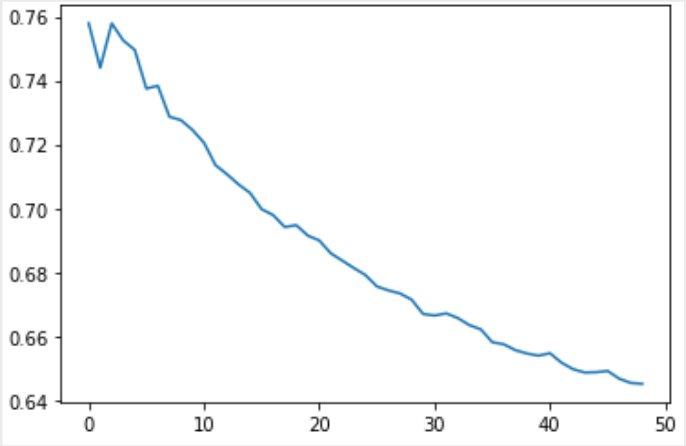


- Here are the features distribution for each copyist.
- We can see that some copyists have featured more and more normally scattered.
- But some behave very unpredictably accordingly, these are the copyist with the least train samples (n.b. the histogram of 2 2 section).

PREDICTION

Data have already been Z normalized, learning algorithms can also benefit from scaling the data.
13 KNeighbors accuracy by number of neighbors·

Thus, began prediction using
multiple models.



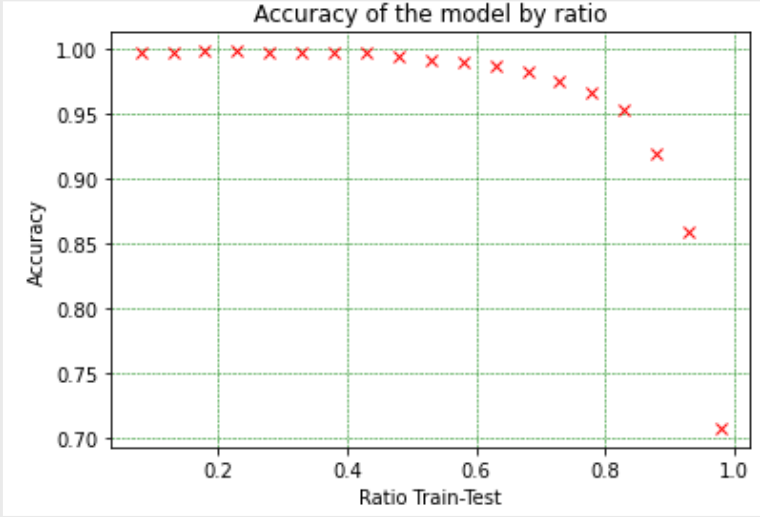
Kneighbors accuracy by number of neighbors

KNeighbors gives roughly a 74% accuracy.

The most efficient hyperparameters with GridSearch method where (n_neighbors = 4, weights = "distance", metric="euclidian")

The most efficient hyperparameters with GridSearch method where (criterion='entropy', n_estimators=500, random_state=42)

- RandomForest gives roughly a 98,02% accuracy.



Accuracy of RandomForest model by test/train ratio

- We can see that the test/train ratio is stable even approaching the 70%. But we can also conclude that a ratio of 33% test/train gives us a pretty good accuracy, which is optimal for our case.

THANK YOU!