PYTHON FOR DATA ANALYSIS

SAMIA ZOBIRI – AMINATA SEYDI

2021– ESILV A4 – DIA5

ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

BLOCKS CLASSIFICATION

- The problem consists in classifying all the blocks of the page layout of a document that has been detected by a segmentation process. This is an essential step in document analysis in order to separate text from graphic areas.

The five classes are:

Text(1)

Horizontal line(2)

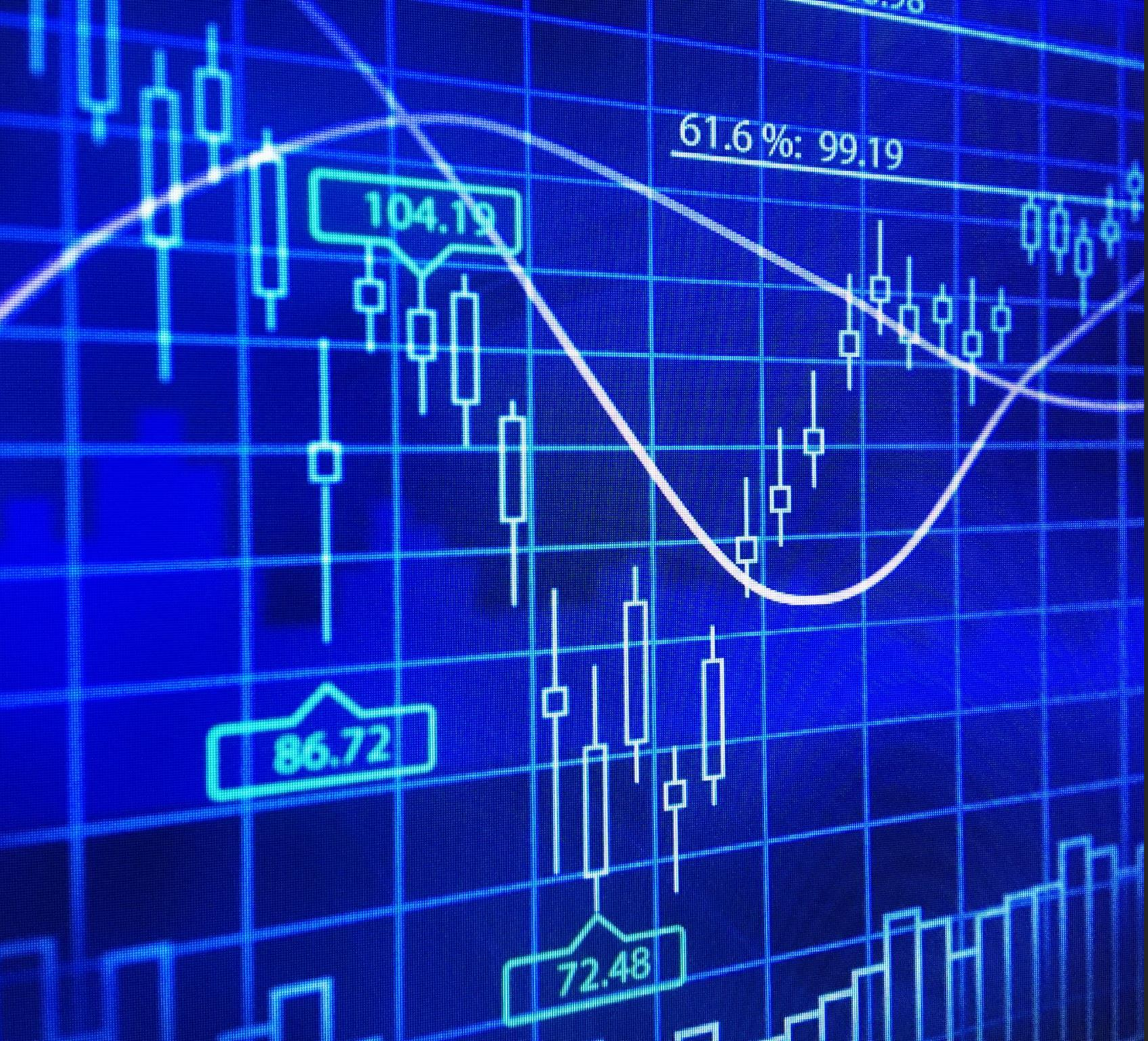
Picture(3)

Vertical line(4)

Graphic(5)

THE ATTRIBUTES

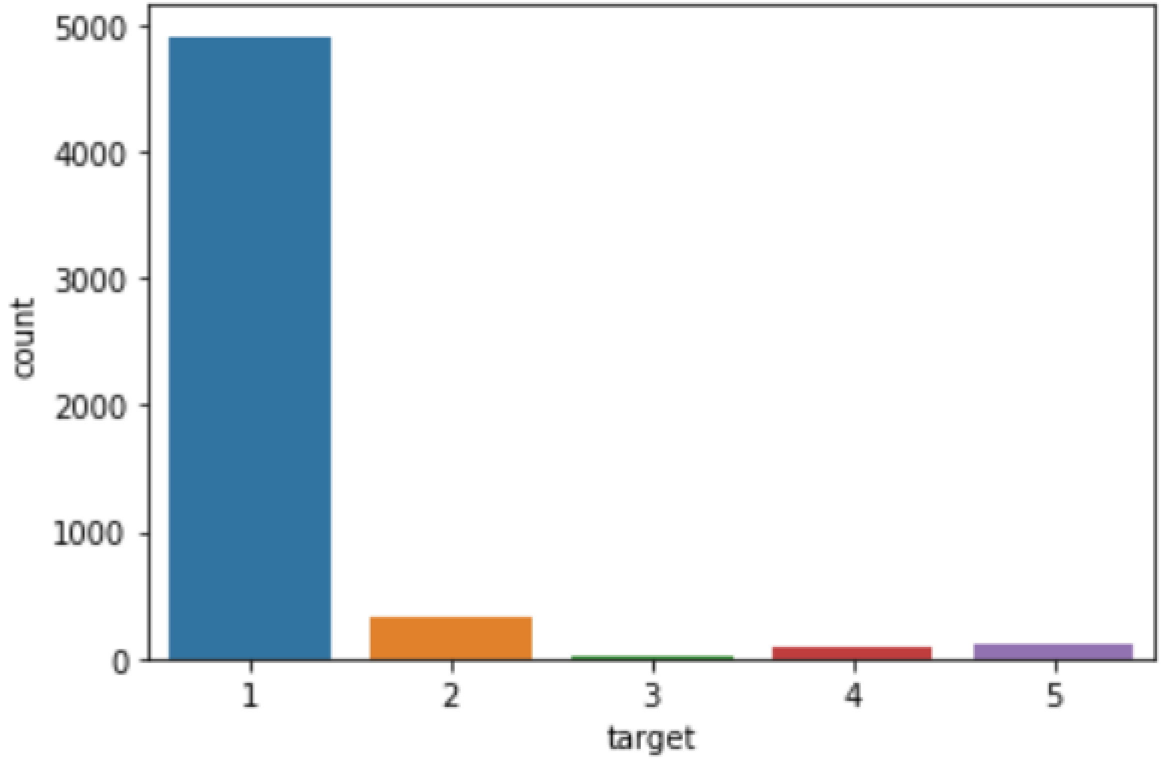
NAME	TYPE	DESCRIPTION
Height	Integer	Height of the block
Lenght	Integer	Length of the block.
Area	Integer	Area of the block (height * lenght)
Eccen	Continuous	Eccentricity of the block (lenght / height)
p_black	Continuous	Percentage of black pixels within the block (blackpix / area)
p_and	Continuous	Percentage of black pixels after the application of the Run Length Smoothing Algorithm (RLSA) (blackand / area)
mean_tr	Continuous	Mean number of white-black transitions (blackpix / wb_trans)
Blackpix	Integer	Total number of black pixels in the original bitmap of the block
Blackand	Integer	Total number of black pixels in the bitmap of the block after the RLSA
wb_trans	Integer	Number of white-black transitions in the original bitmap of the block



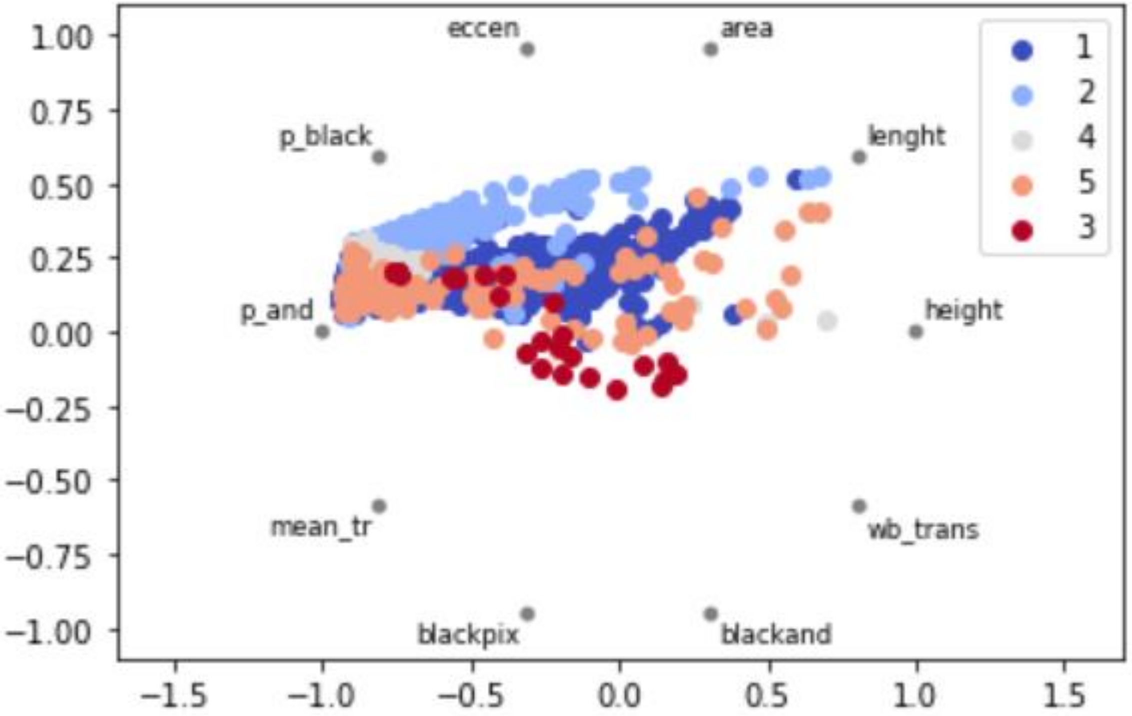
DATA VISUALIZATION

FIRST STEP

- We first change our “class” attribute to “target” on.
- As you can see it in the graphic, there is a large part of target values belonging to a text classification. The least represented one is 3 and corresponds to graphics.

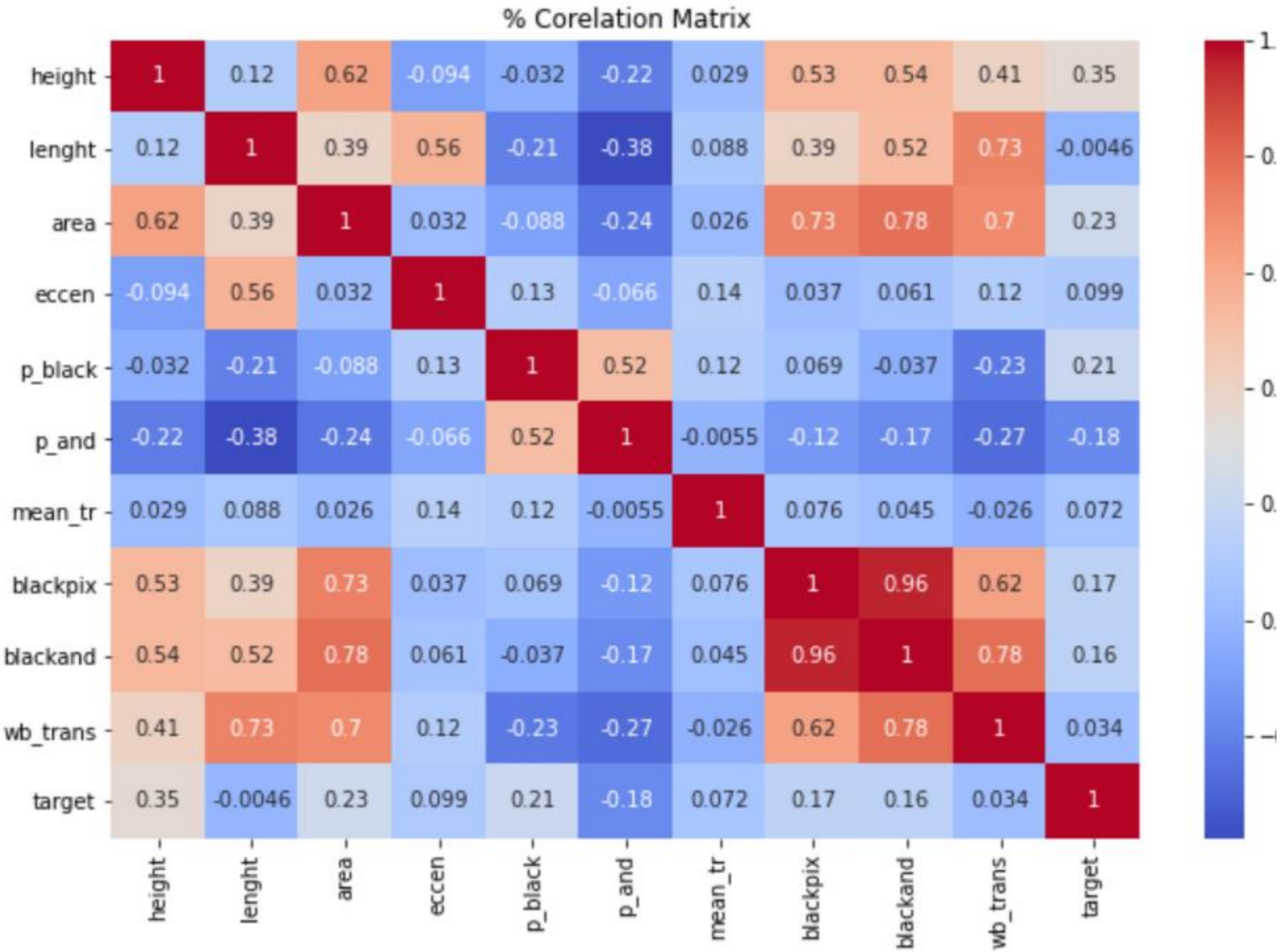


- This graph allows to project our 10-dimensional data set into a 2D space where the influence of each dimension can be interpreted as a balance between the influence of all dimensions.



THE CORELATION MATRIX

- Here we can see thanks to the correlation matrix that the most correlated features are blackpix (Total number of black pixels in the original bitmap of the block) and black_and(Total number of black pixels in the bitmap of the block after the RLSA.) with 96% correlation.



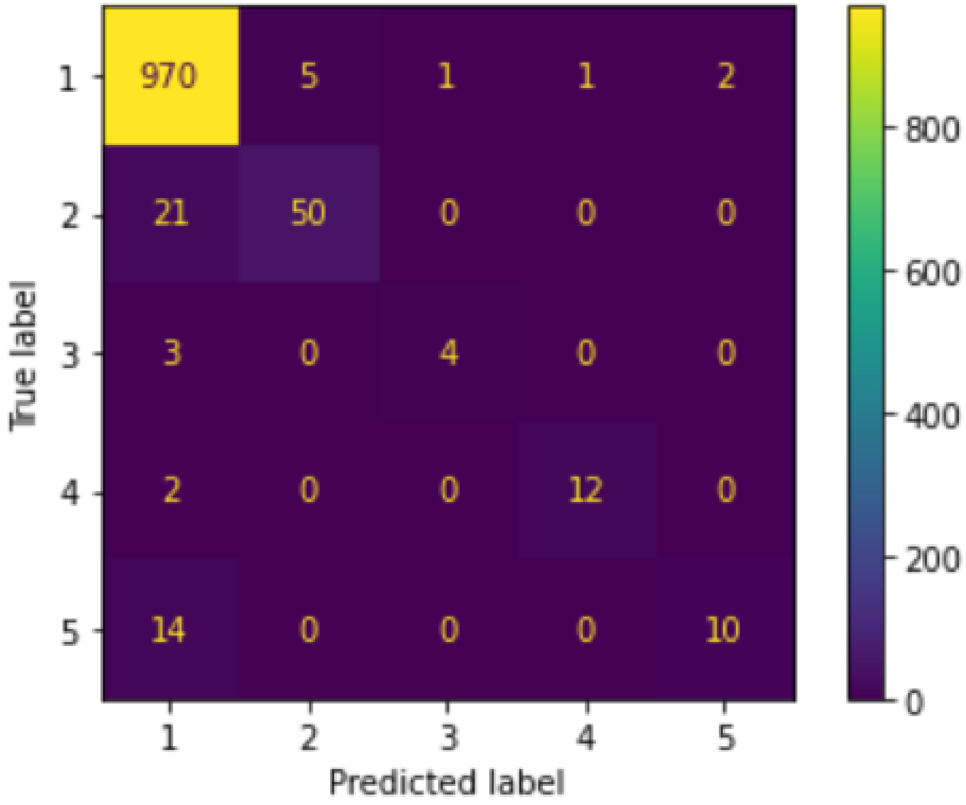
% Corelation Matrix

	height	lenght	area	eccen	p_black	p_and	mean_tr	blackpix	blackand	wb_trans	target
height	1	0.12	0.62	-0.094	-0.032	-0.22	0.029	0.53	0.54	0.41	0.35
lenght	0.12	1	0.39	0.56	-0.21	-0.38	0.088	0.39	0.52	0.73	-0.0046
area	0.62	0.39	1	0.032	-0.088	-0.24	0.026	0.73	0.78	0.7	0.23
eccen	-0.094	0.56	0.032	1	0.13	-0.066	0.14	0.037	0.061	0.12	0.099
p_black	-0.032	-0.21	-0.088	0.13	1	0.52	0.12	0.069	-0.037	-0.23	0.21
p_and	-0.22	-0.38	-0.24	-0.066	0.52	1	-0.0055	-0.12	-0.17	-0.27	-0.18
mean_tr	0.029	0.088	0.026	0.14	0.12	-0.0055	1	0.076	0.045	-0.026	0.072
blackpix	0.53	0.39	0.73	0.037	0.069	-0.12	0.076	1	0.96	0.62	0.17
blackand	0.54	0.52	0.78	0.061	-0.037	-0.17	0.045	0.96	1	0.78	0.16
wb_trans	0.41	0.73	0.7	0.12	-0.23	-0.27	-0.026	0.62	0.78	1	0.034
target	0.35	-0.0046	0.23	0.099	0.21	-0.18	0.072	0.17	0.16	0.034	1

DATA MODELISATION

LOGISTIC REGRESSION

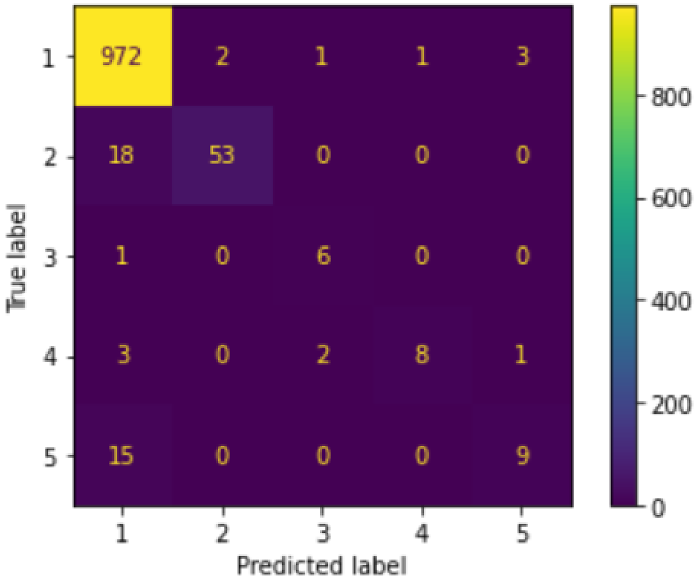
- We can see that a large ppart of predictions corresponds to text classification. Also with logistic regression we can see that 15 predicted values corresponfing to 1 were wrongly predicted (they belong to 2 (horizontal line))
- Logistic regression accuracy : 0.9553



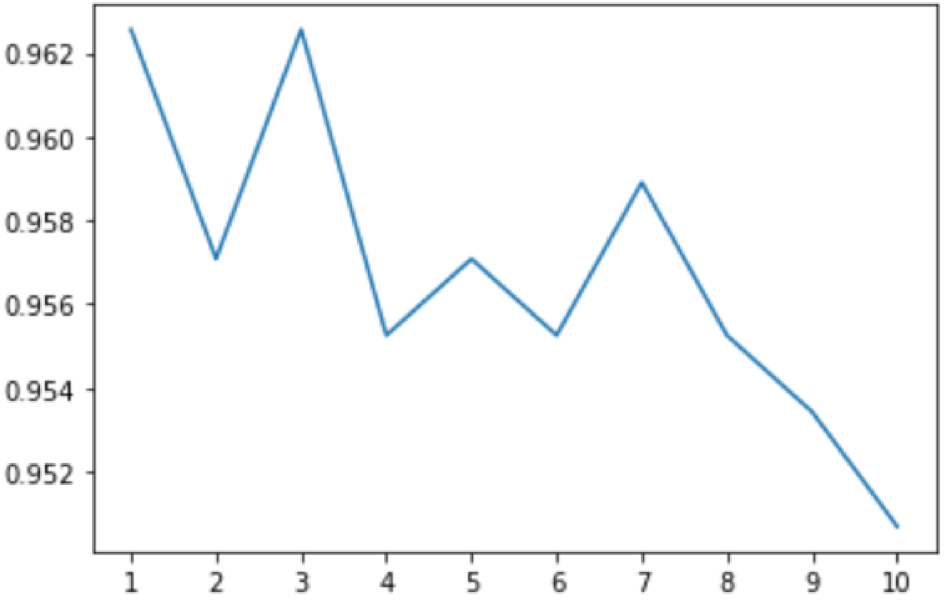
	1	2	3	4	5
1	970	5	1	1	2
2	21	50	0	0	0
3	3	0	4	0	0
4	2	0	0	12	0
5	14	0	0	0	10

KNN ALGORITHM

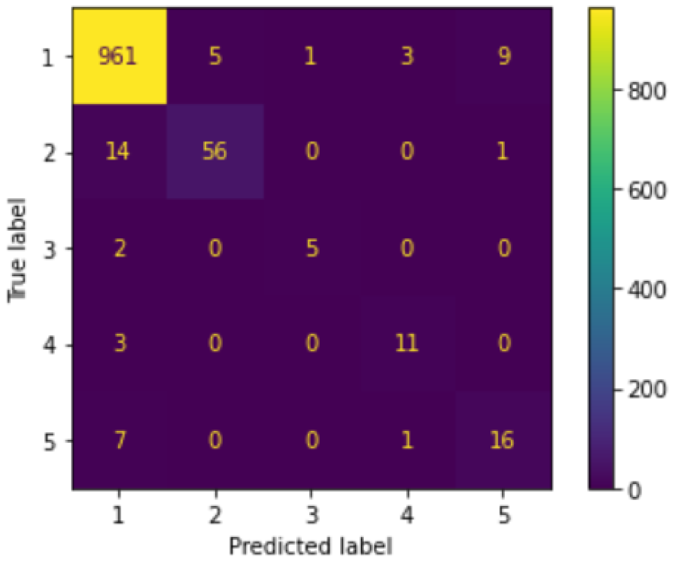
- We must find the numbers of neighbors N which allows the best accuracy
- Best accuracy for K-nearest neighbors is for N=5
- K-nearest Neighbors accuracy : 0.9571

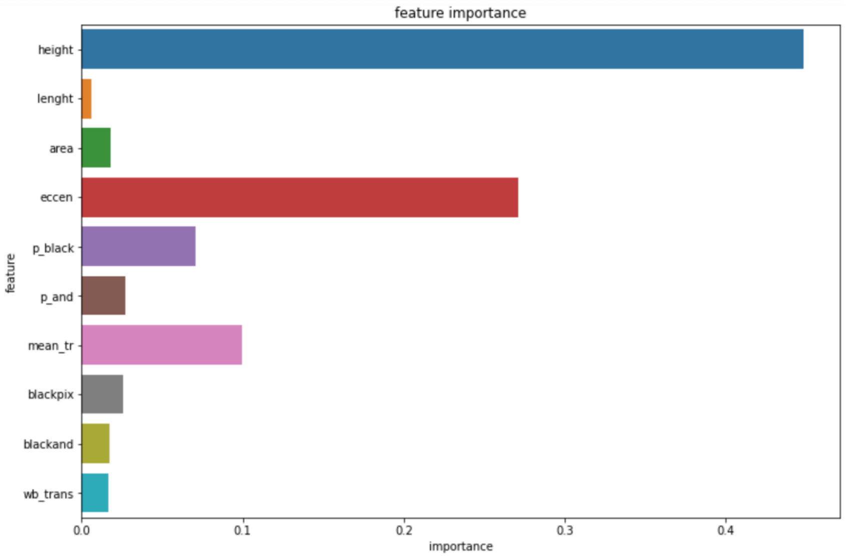


True label \ Predicted label	1	2	3	4	5
1	972	2	1	1	3
2	18	53	0	0	0
3	1	0	6	0	0
4	3	0	2	8	1
5	15	0	0	0	9



DECISION TREE



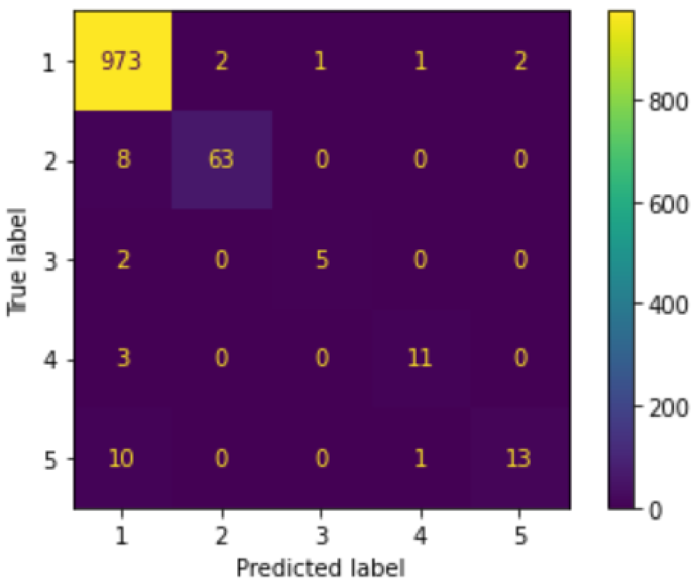


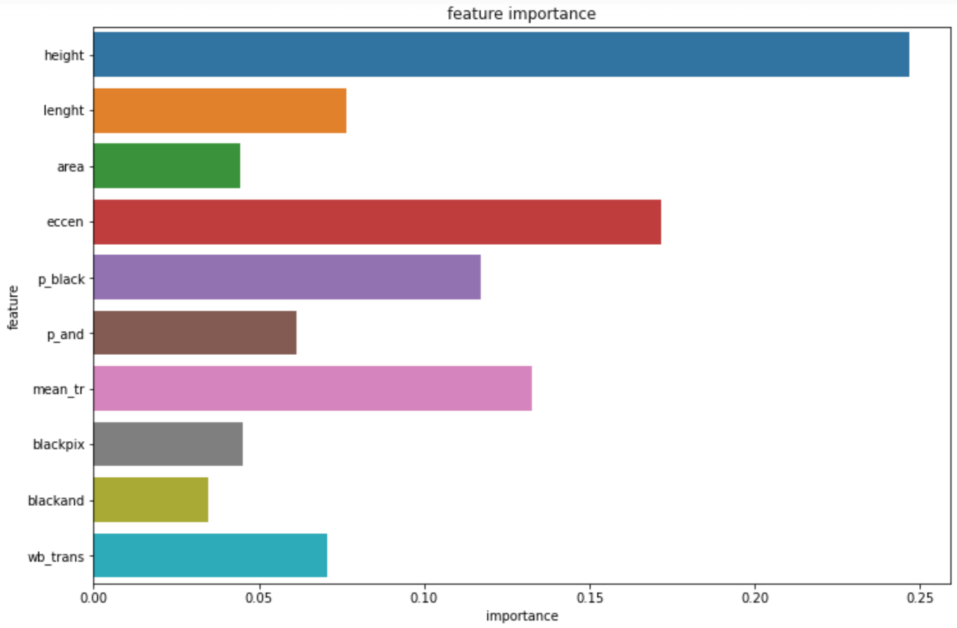
- Here we can see that the most important features are height eccen and mean-tr surprisingly even though they're not the best correlated features to target
- Decision tree accuracy : 0.958

RANDOM FOREST

- We can see the more we add trees to the forest, more the model get accurate.

Number of trees	Accuracy
10	0.9699
20	0.9726
100	0.9726

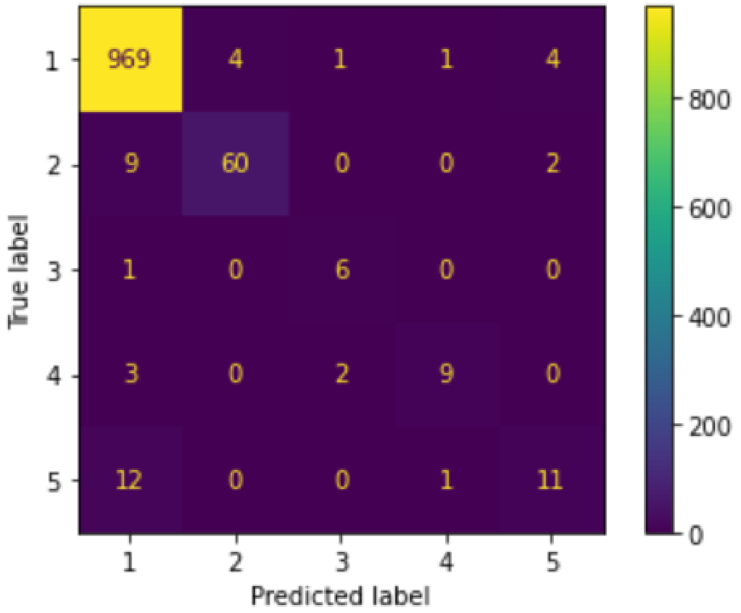




- Here the most important features to the modelling is also height, eccen and mean_tr

BOOSTING

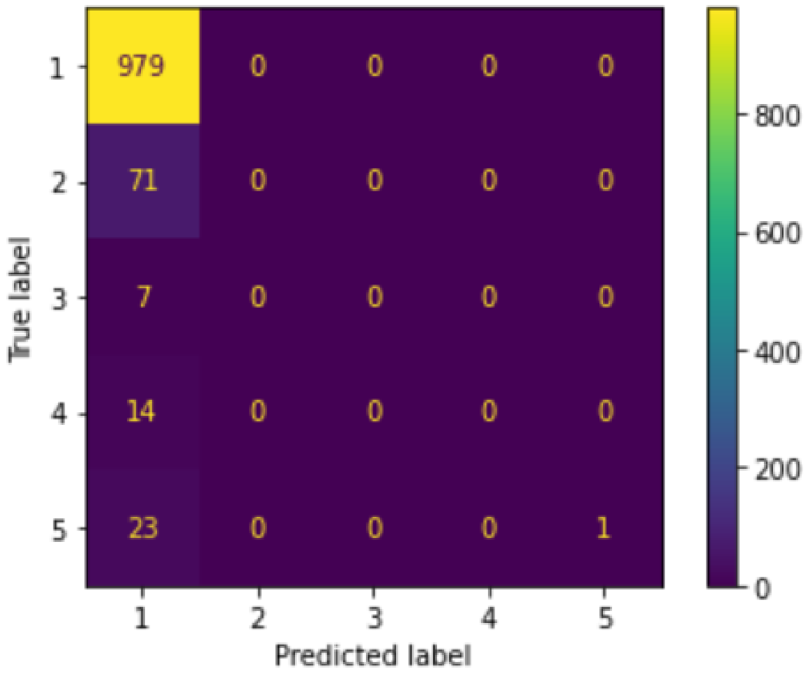
- the best boosting model is for 20 trees. For 100 trees, the model may overfit.



Number of trees	Accuracy
10	0.9607
20	0.9635
100	0.9717

BAGGING

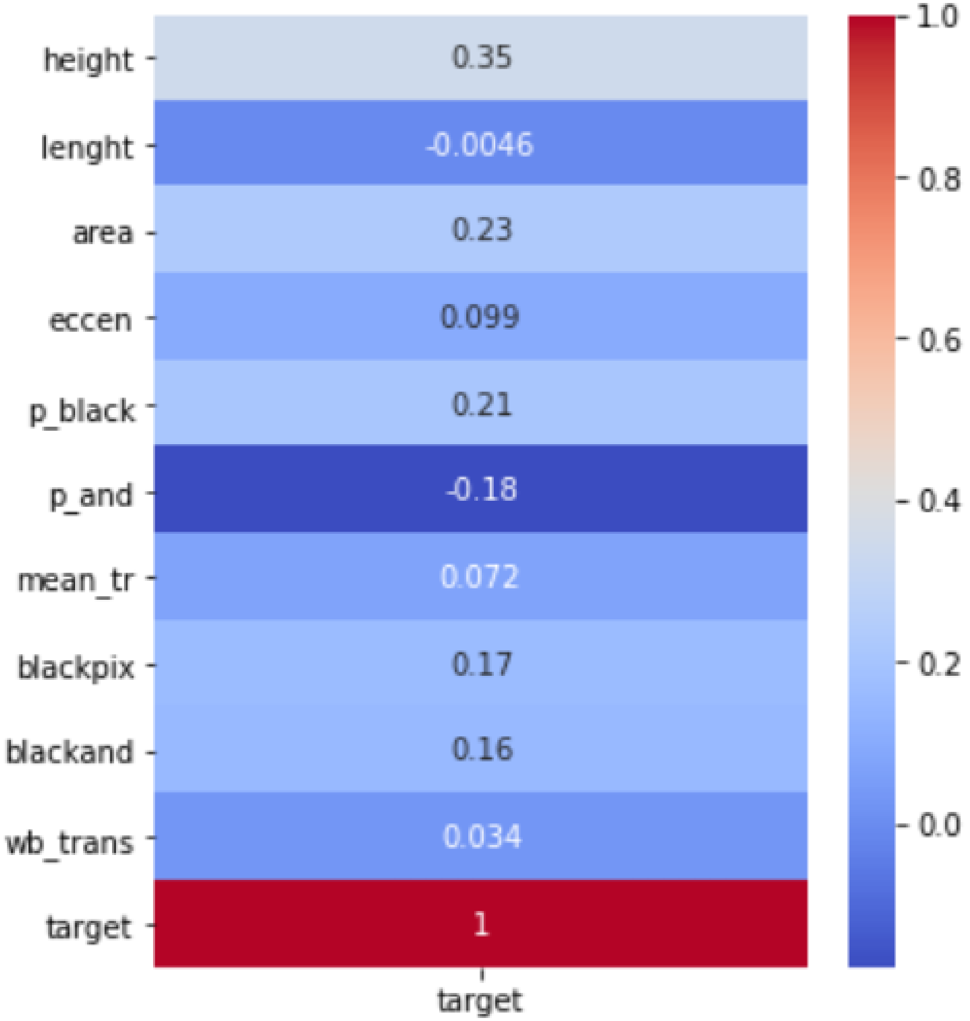
- the best bagging model is with 10 trees



Number of trees	Accuracy
10	0.895
20	0.8959
100	0.9014

DATA MODELIZATION DROPPING LEAST CORRELATED FEATURES

- We can see that wb_trans, mean_tr and lenght are the least correlated features to target, hence we are going to drop them. Also, we are going to change the dtype of target as category.



- The accuracy enhanced since we dropped the least correlated features. Also the best classifiers to fit the model are Decision Tree, Random Forest and Bagging. Hence we are going to keep those three for the rest of the project.

Model	Accuracy
Logistic Regression	0.9653
KNN Algorithm	0.9699
Decision Tree	0.9662
Random Forrest with 100 trees	0.9726
Boosting	0.9726
Bagging	0.9306