PYTHON FOR DATA ANALYSIS

Youssef CHRAIBI - Sarah Cubuk
ESILV A4 – DIA3

Contents

Part 1 Data presentation

Part 2 Data visualization

Part 3 Data modelisation

Part 4 Conclusion

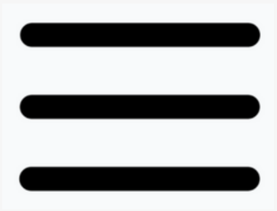
Data presentation

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:

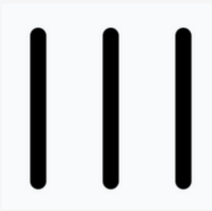
1. text(1)

2. horizontal line (2)



3. picture (3)

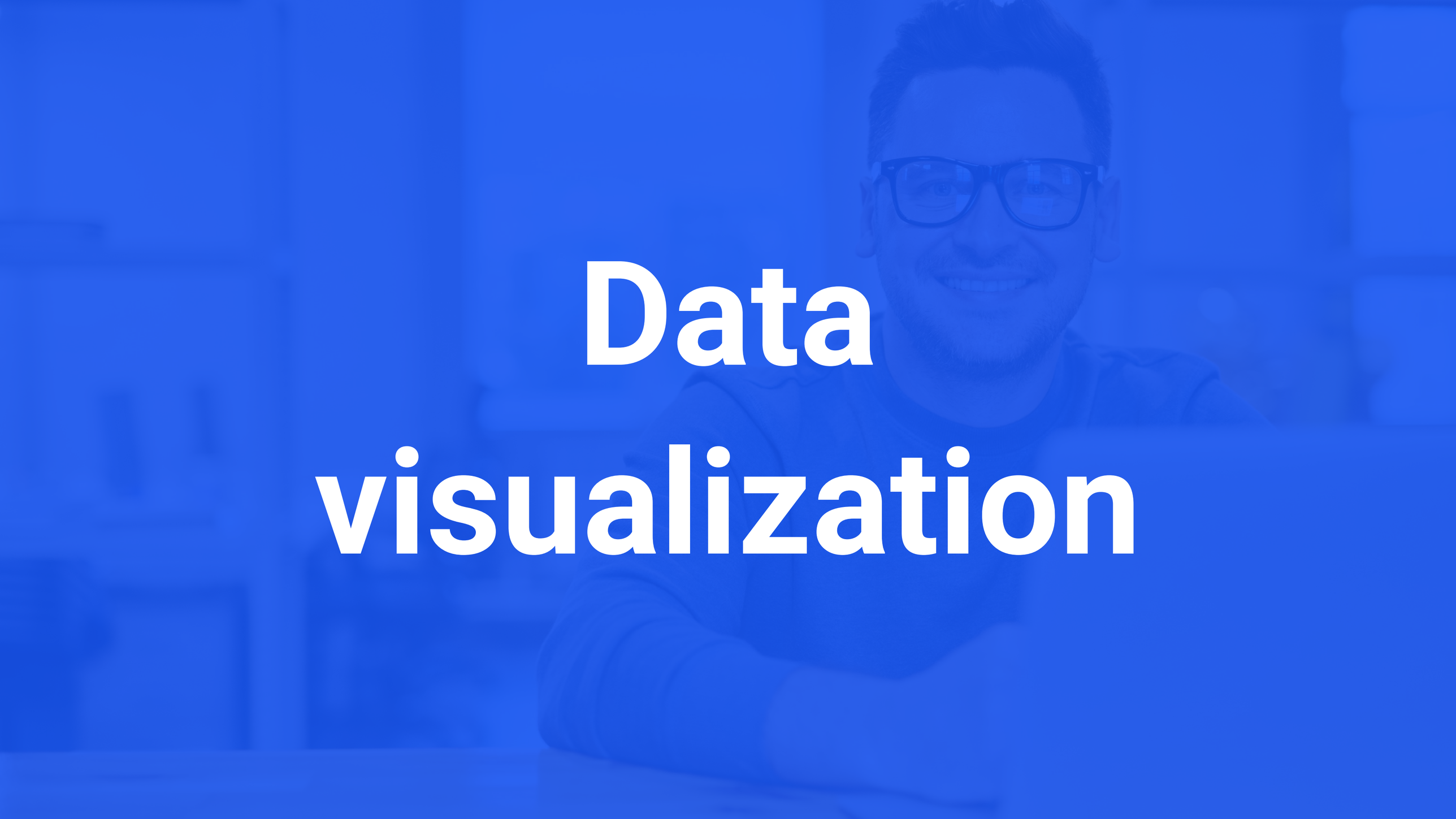
4. vertical line (4)



5. graphic (5)

THE VARIABLES

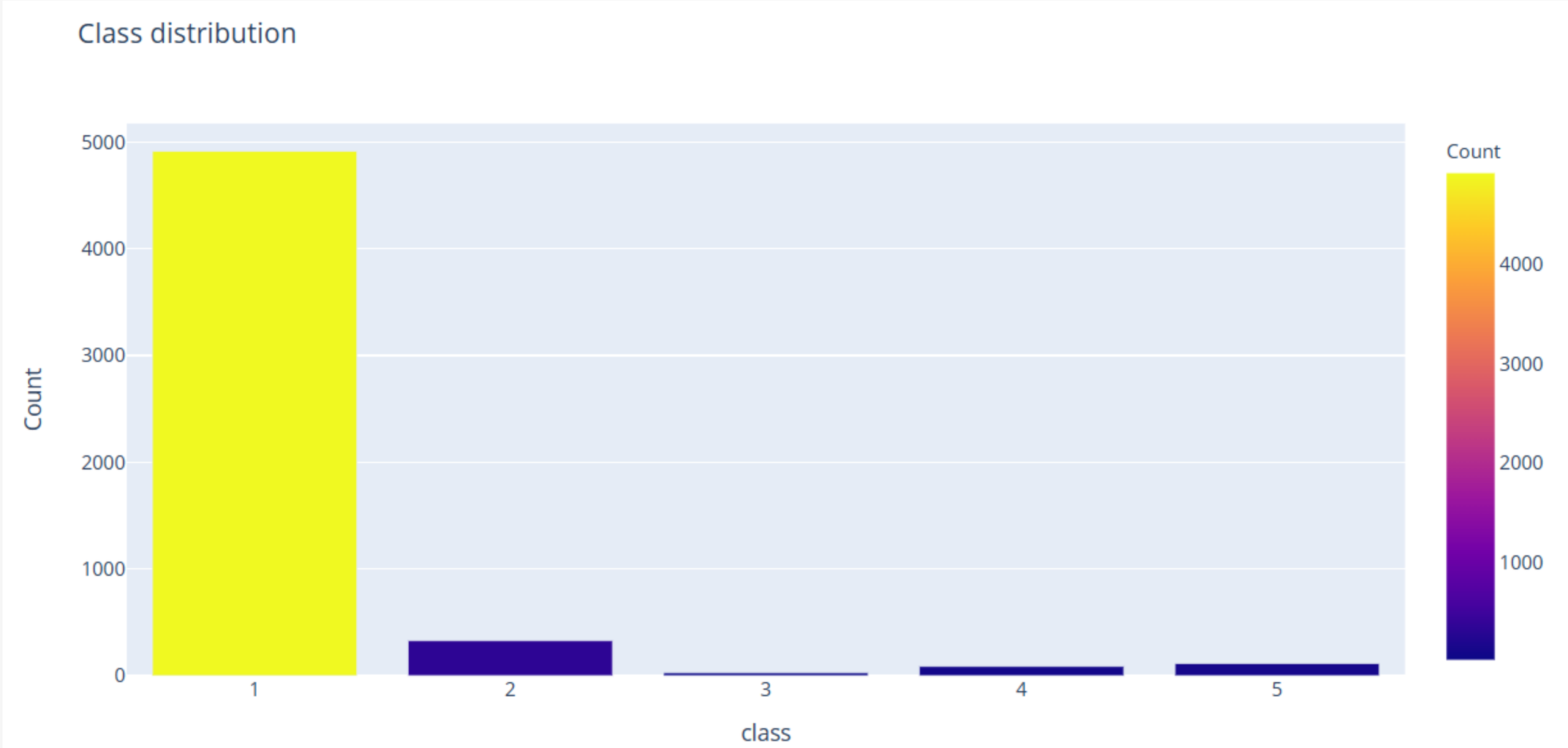
NAME	TYPE	DESCRIPTION
Height	Integer	Height of the block
Lenght	Integer	Length of the block
Area	Integer	Area of the block (height * lenght)
Eccen	Float	Eccentricity of the block (lenght / height)
p_black	Float	Percentage of black pixels within the block (blackpix / area)
p_and	Float	Percentage of black pixels after the application of the Run Length Smoothing Algorithm (RLSA) (blackand / area)
mean_tr	Float	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

Distribution of the Output : As you can see, a very large part of the page blocks are text blocks (1) while image blocks (3) are the rarest

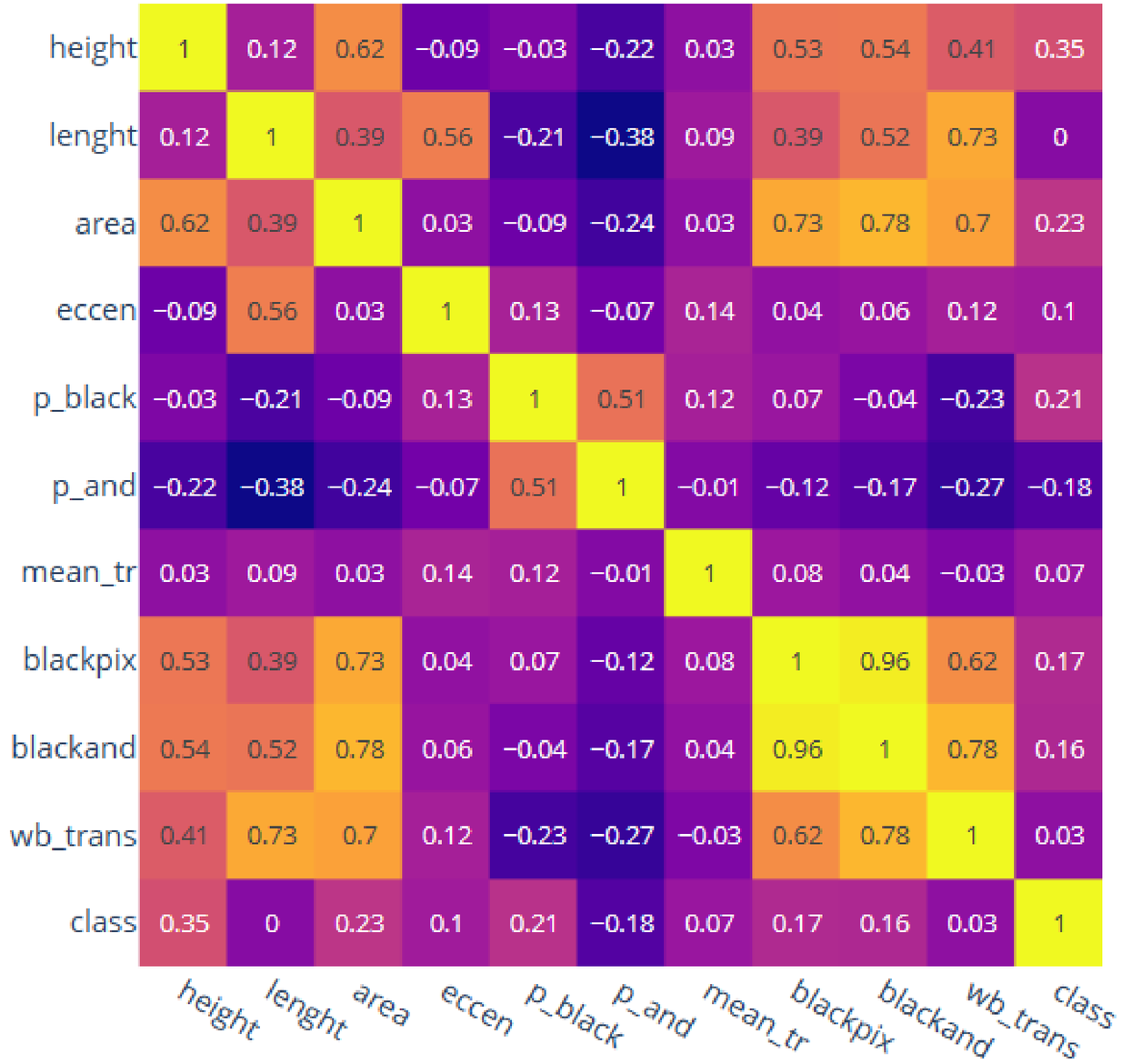
highly imbalanced dataset

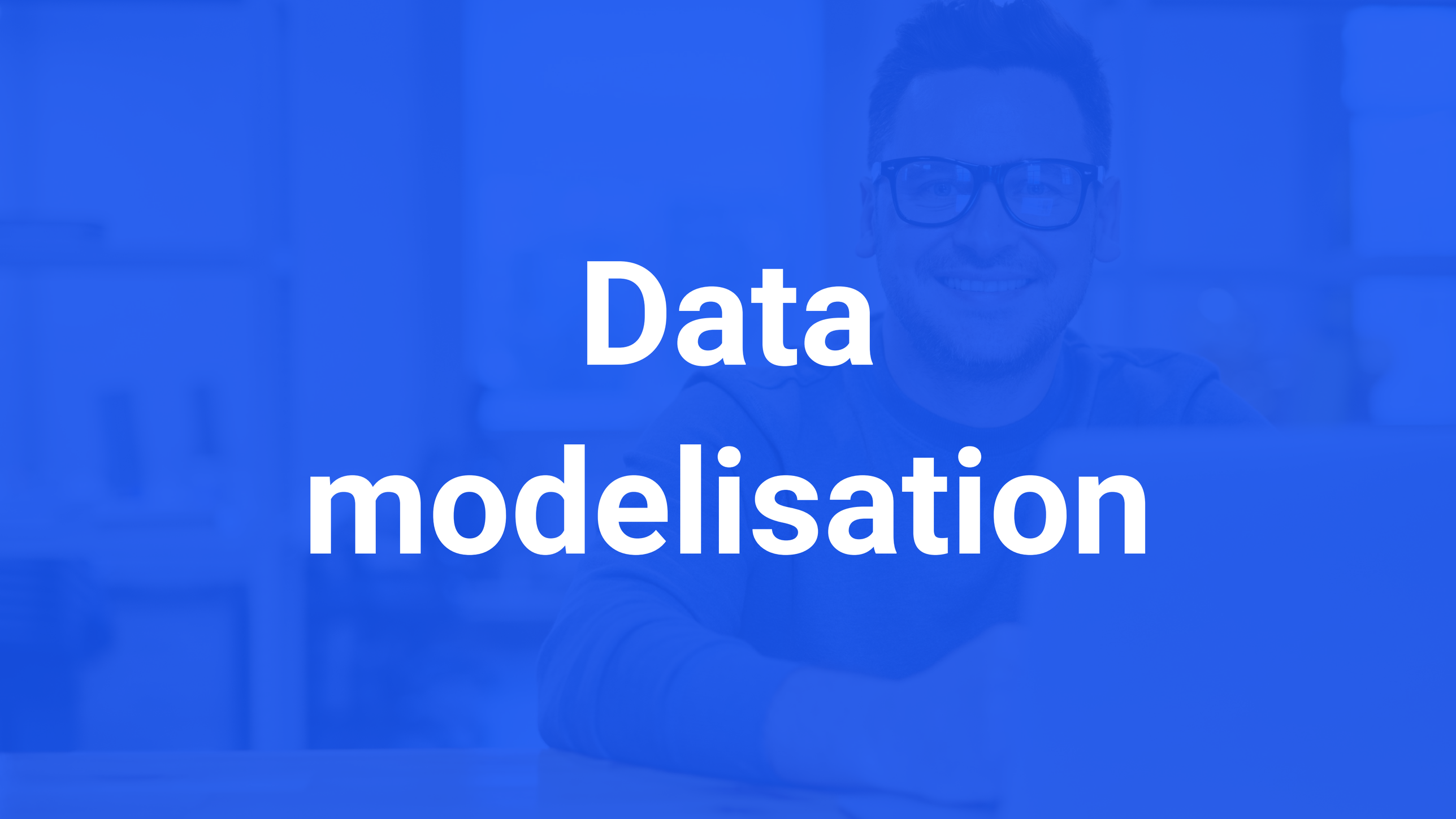


Correlation Matrix

Here, we notice that the 2 most correlated variables are blackand (Total number of black pixels in the bitmap of the block after the RLSA) and blackpix (Total number of black pixels in the original bitmap of the block) reaching a correlation rate of 96%.

Run Length Smoothing Algorithm (RLSA)





Data modelisation

Modelisation

In this section we will be training various models using different classifiers. Out of them all, we will be choosing the best classifier to give us the most accurate prediction

Comparaison of models

Model	Train Accuracy	Test Accuracy	Rank
Logistic Regression	96,20%	95,71%	5
K Nearest Neighbor	97,21%	96,26%	4
Support Vector Machine (Linear Classifier)	96,30%	96,26%	4
Support Vector Machine (RBF Classifier)	96,96%	96,53%	2
Gaussian Naive Bayes	90,61%	90,68%	6
Decision Tree Classifier	99,79%	96,44%	3
Random Forest Classifier	99,58%	97,26%	1

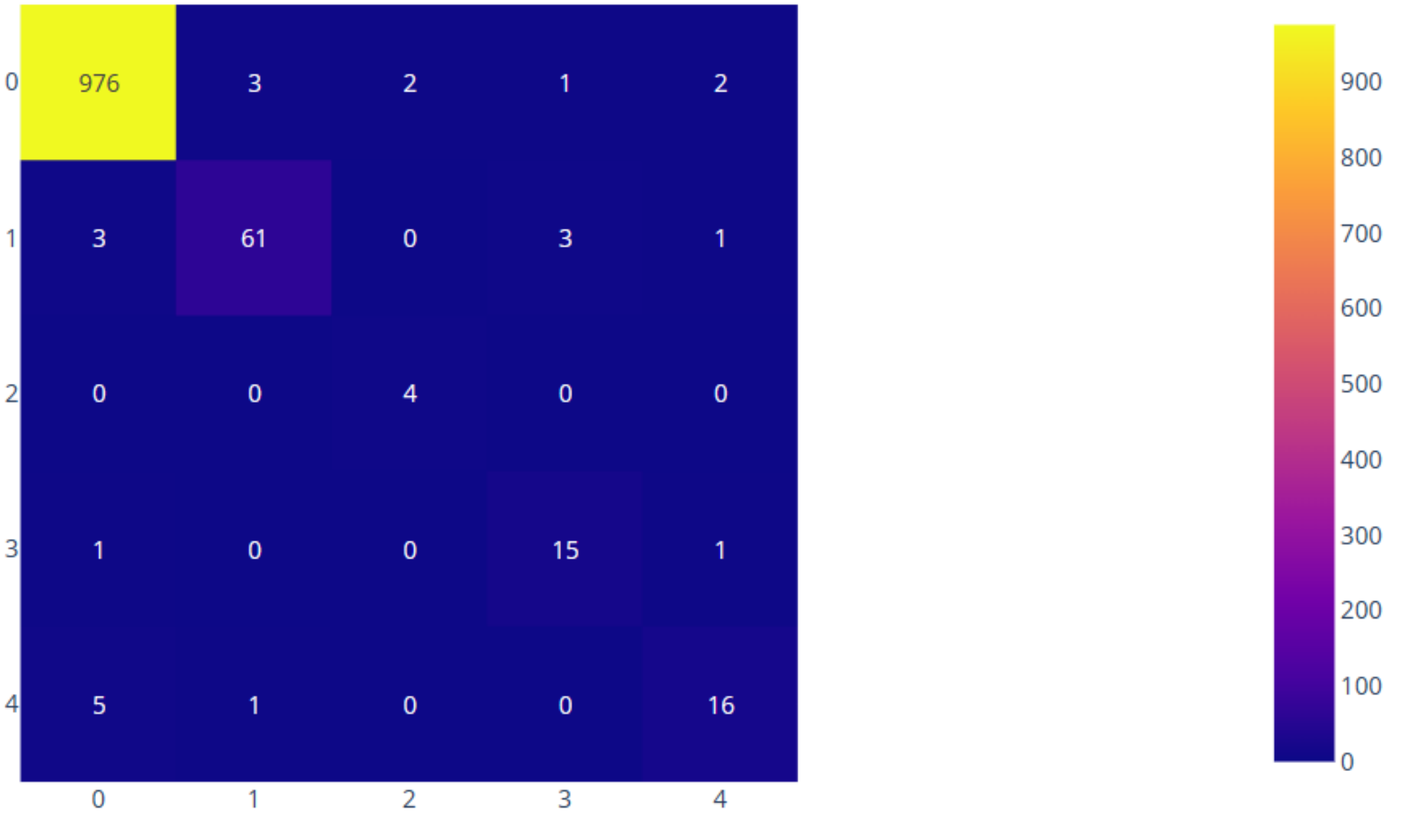
HyperTunning

what parameters should we choose ?

This is the most important section of this project. Here, the ultimate goal is to find an optimal combination of hyperparameters that minimizes a predefined loss function to give better results.

As we can see, after testing several parameters, the following parameterization gives us the best randomForest score :

RandomForestClassifier(
criterion='entropy',
max_depth=11,
min_samples_split=7)

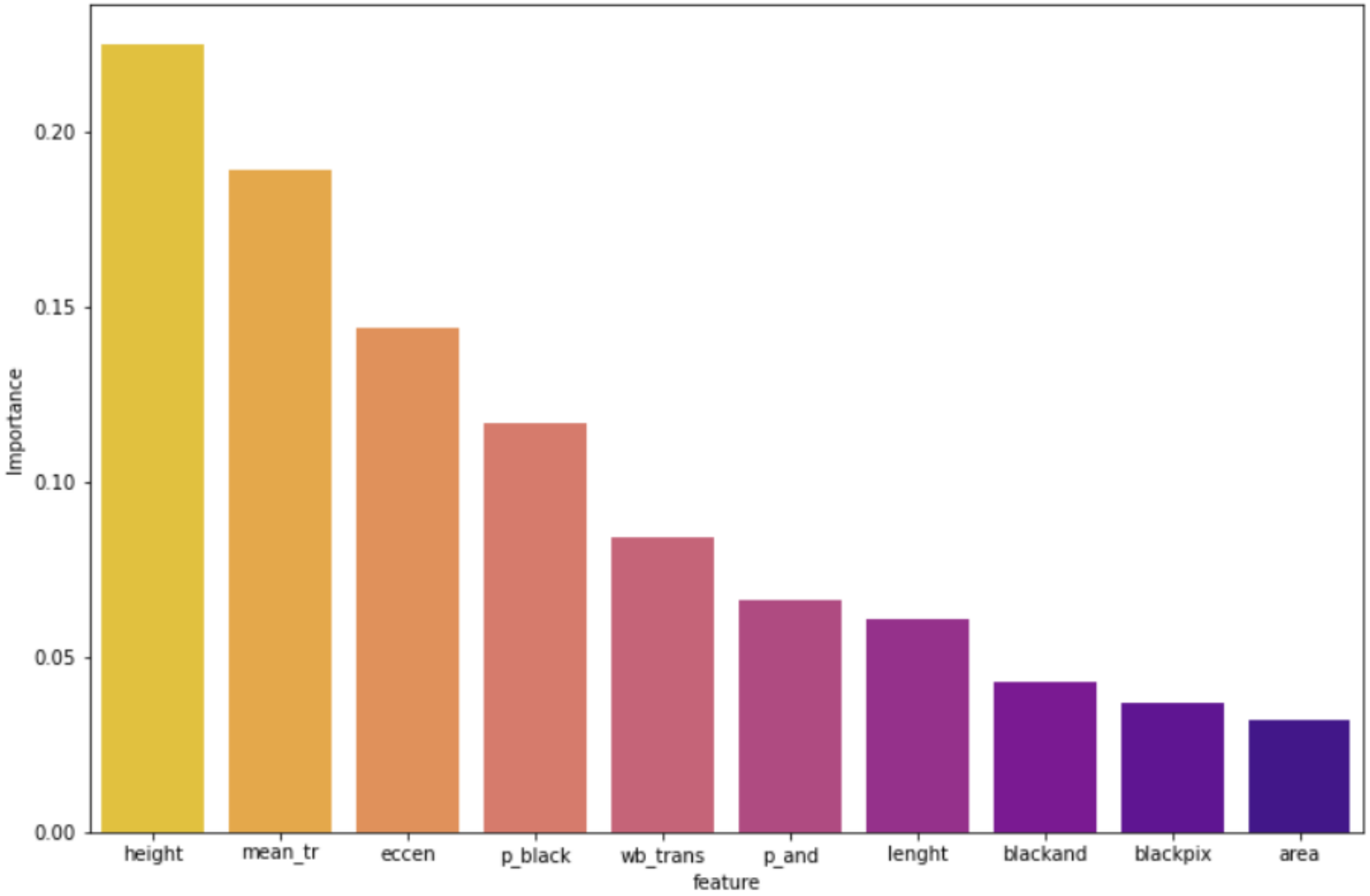


Indeed, we get a Train Accuracy of 99.36% and a Test Accuracy of 97.89%

Features importance

The most important features are : height, mean_tr, eccen for the modelisation

	feature	importance
0	height	0.225000
1	mean_tr	0.189000
2	eccen	0.144000
3	p_black	0.117000
4	wb_trans	0.084000
5	p_and	0.066000
6	lenght	0.061000
7	blackand	0.043000
8	blackpix	0.037000
9	area	0.032000



Conclusion

We applied multiple machine learning models on a dataset consisting of 5473 example coming from 54 distinct documents. We found that the Random Forest Classifier was the most efficient to classify page blocks and the quality of classification is 0.9789.


Merci !