



Page Blocks Classification

Alexandre SMADJA – Brahim TALB – Ewen RONDEL

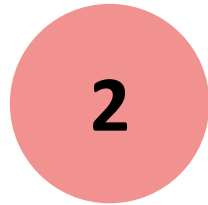
The background of the slide is a vibrant, multi-colored marbled pattern. It features swirling, organic shapes in shades of pink, red, orange, and yellow, with thin, delicate veins of blue and green interspersed throughout. In the center of the image is a large, white rectangular box with a thin black border. Inside this box, the word "Dataset" is written in a clean, black, sans-serif font. Below the text, there is a short, horizontal black line.

Dataset

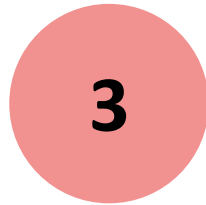
This dataset is made of blocks information of the page layout of different documents. Those blocks are labeled with 5 classes:



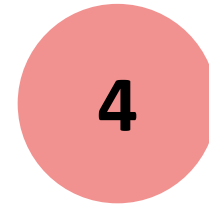
text



horizontal_line



graphic



vertical_line



picture

	height	length	area	eccen	p_black	p_and	mean_tr	blackpix	blackand	wb_trans	class
0	5	7	35	1.400	0.400	0.657	2.33	14	23	6	1
1	6	7	42	1.167	0.429	0.881	3.60	18	37	5	1
...
5472	7	41	287	5.857	0.213	0.801	1.36	61	230	45	1
5473	8	1	8	0.125	1.000	1.000	8.00	8	8	1	4

Class	Frequency	Percentage	Cumulative percentage
text	4913	89.8%	89.8%
horizontal_line	329	6.0%	95.8%
graphic	28	0.5%	96.3%
vertical_line	88	1.6%	97.9%
picture	115	2.1%	100%
Total	5473	100%	100%

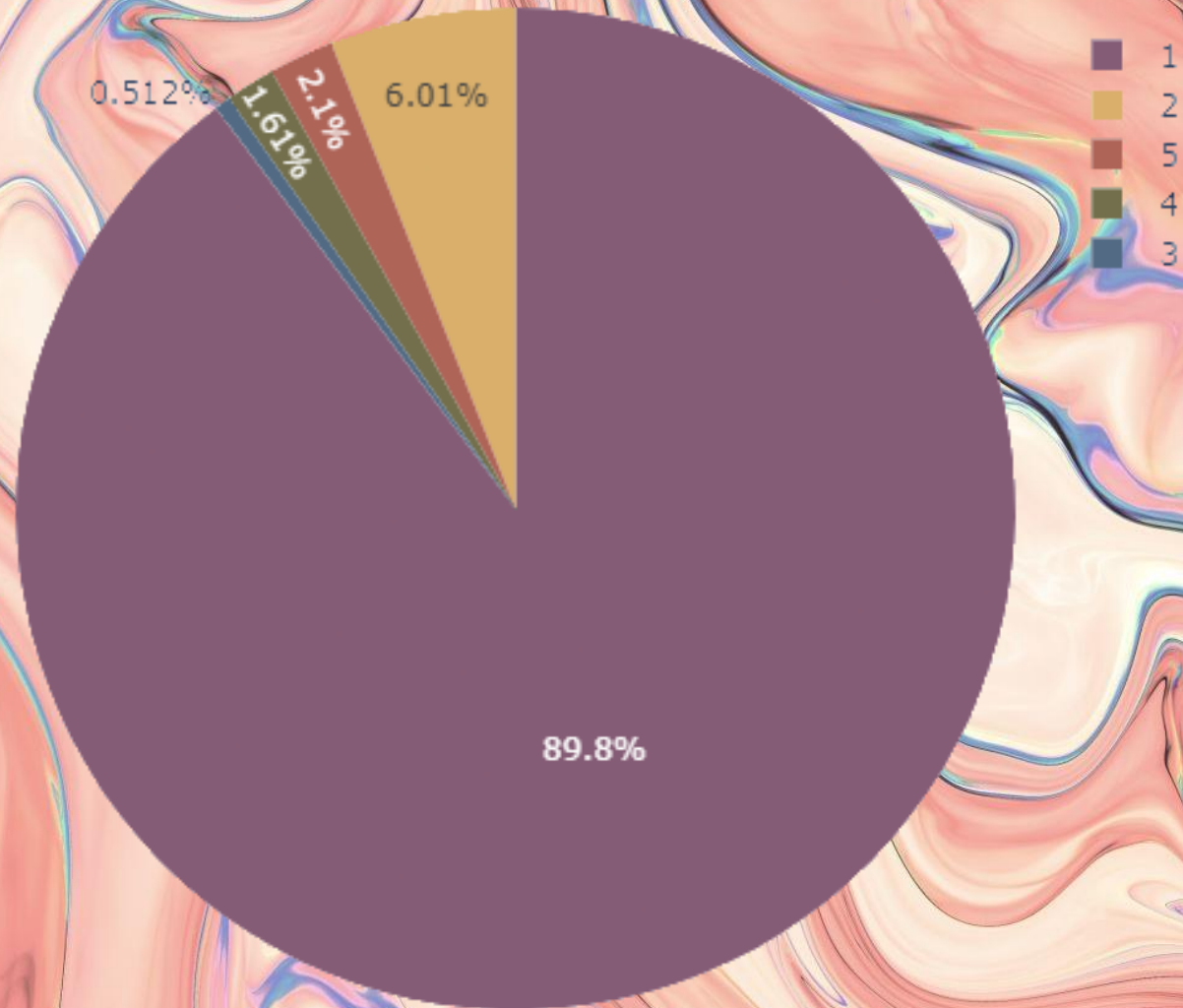
How to classify all the blocks of the page layout of a document that has been detected by a segmentation process ?



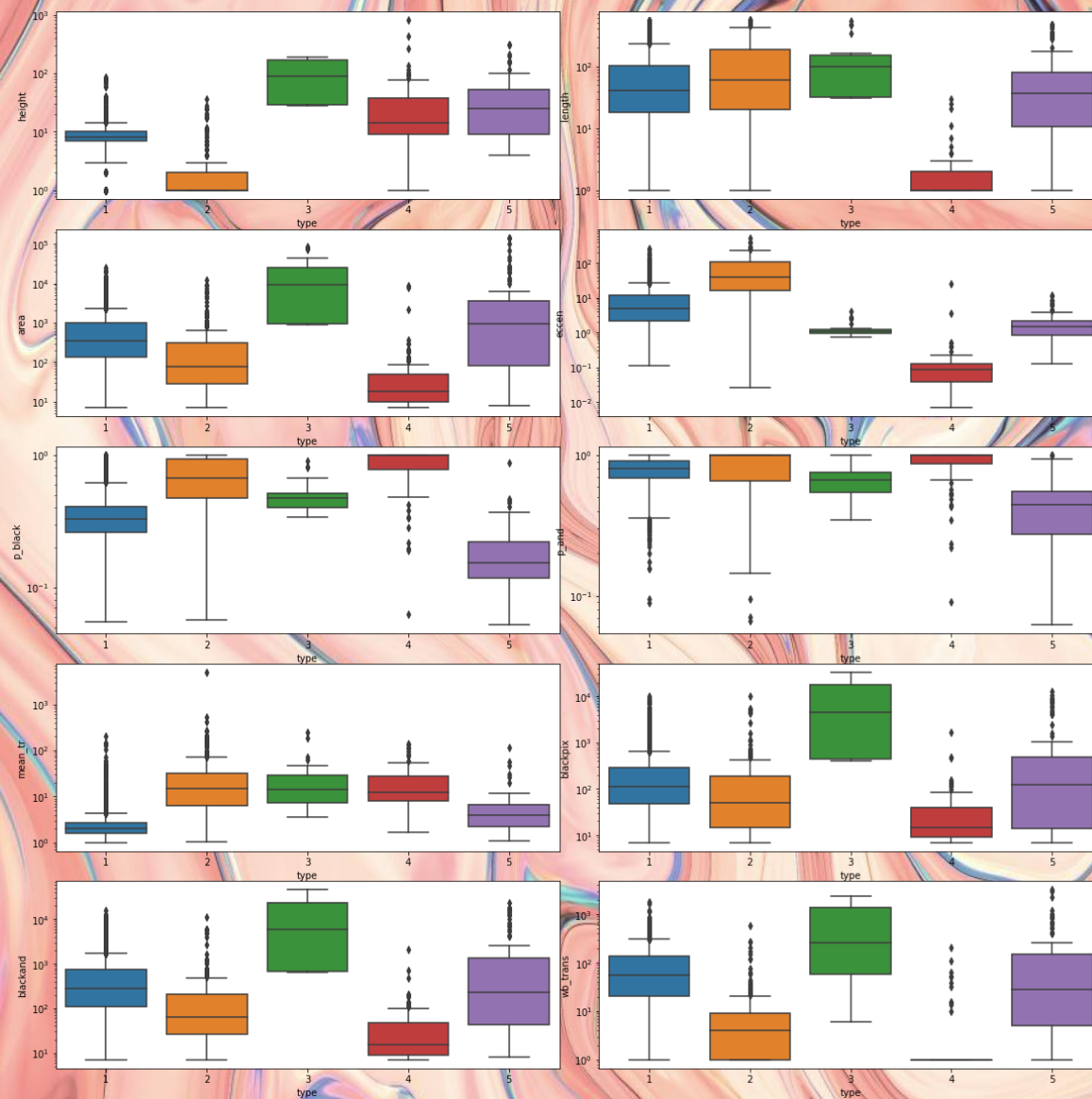
The background of the slide is a vibrant, multi-colored marbled pattern. It features swirling, organic shapes in shades of pink, red, orange, and yellow, with thin, delicate veins of blue and green interspersed throughout. In the center of the image is a large, white rectangular box with a thin black border. Inside this box, the text "Data Visualization" is written in a clean, black, sans-serif font. Below the text, there is a short, horizontal purple line.

Data Visualization

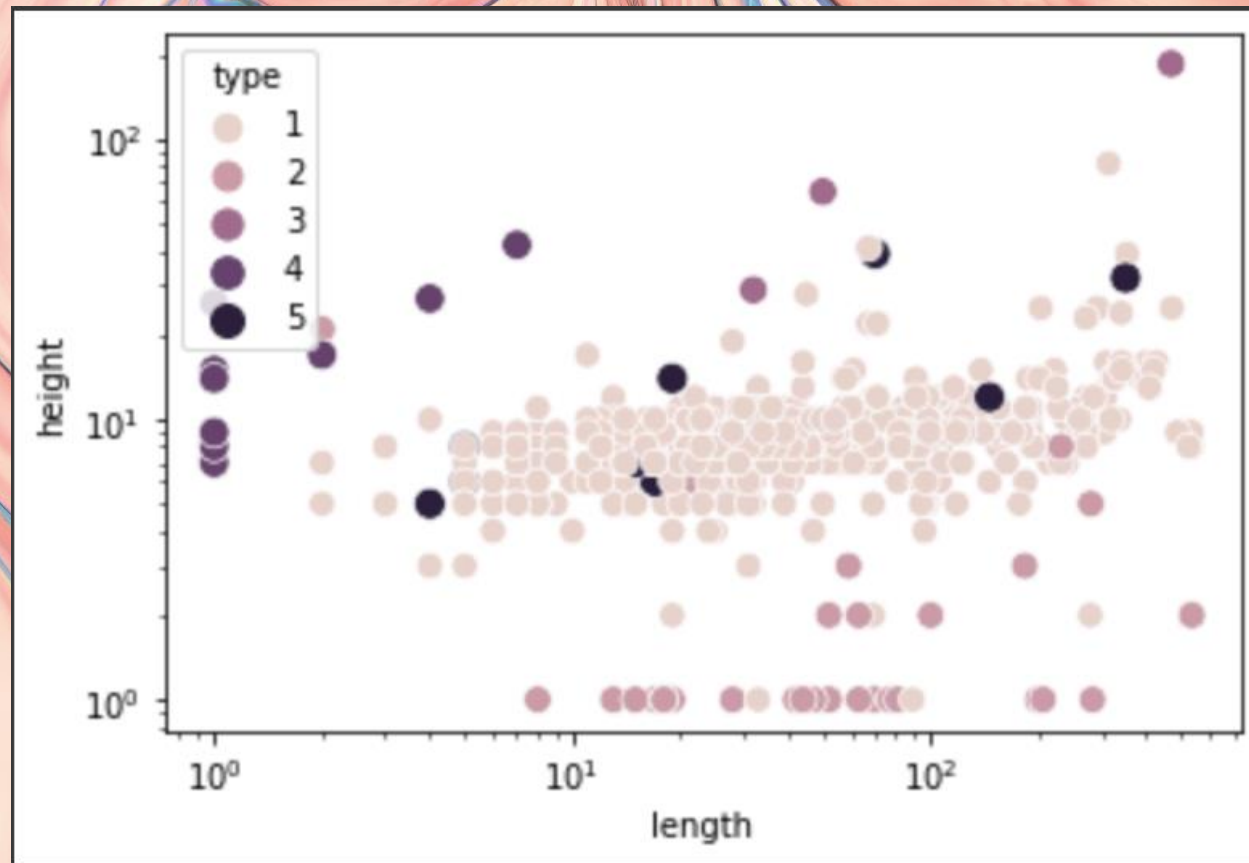
Piechart of all 5 different classes



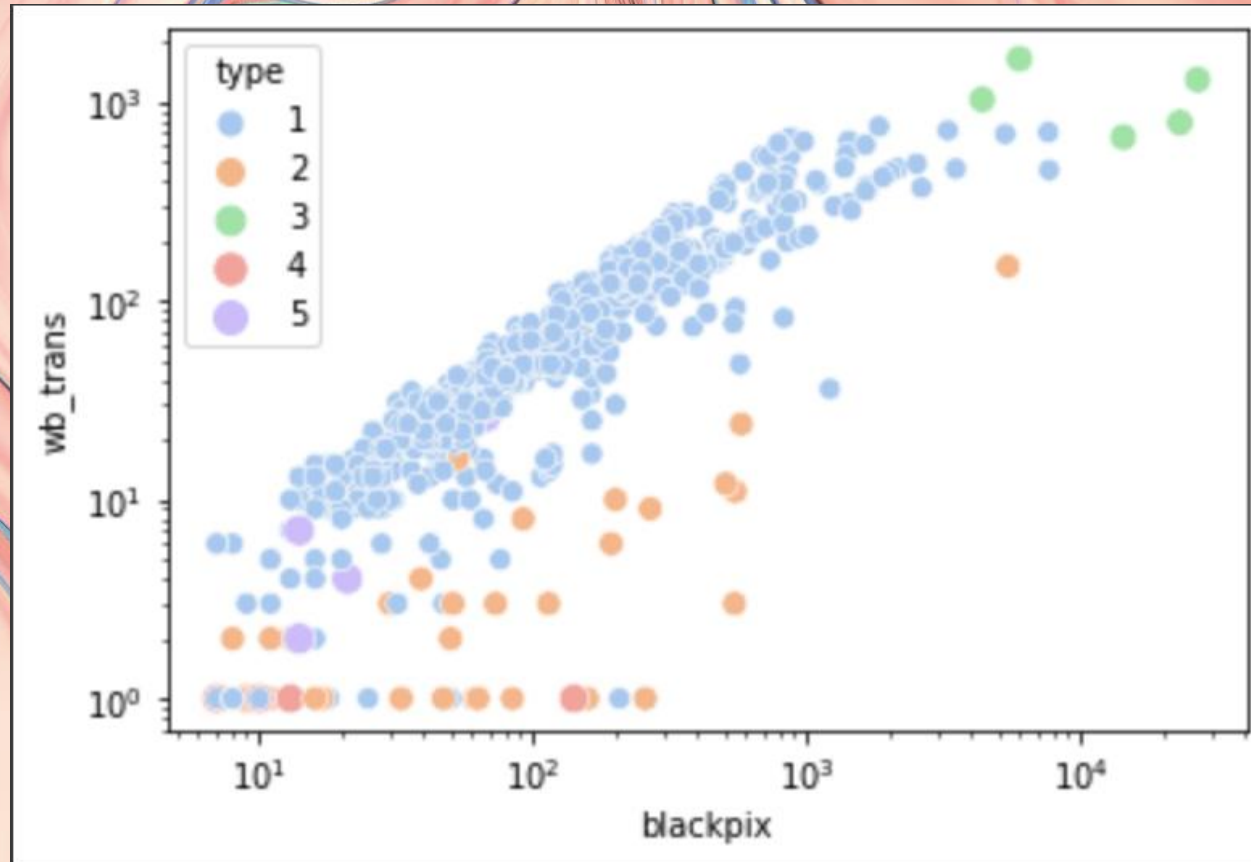
Boxplots of all variables by class using a logarithmic scale



Scatterplot of the height according to the length by class



Scatterplot of the number of black pixels according to the number of white/black transitions



The background of the slide is a vibrant, multi-colored marbled pattern. It features swirling, organic shapes in shades of pink, red, orange, and yellow, with thin, delicate veins of blue and green interspersed throughout. In the center of the image is a large, white rectangular box with a thin black border. Inside this box, the words "Machine Learning" are written in a clean, black, sans-serif font. Below the text, there is a short, horizontal purple line.

Machine Learning

Linear Regression



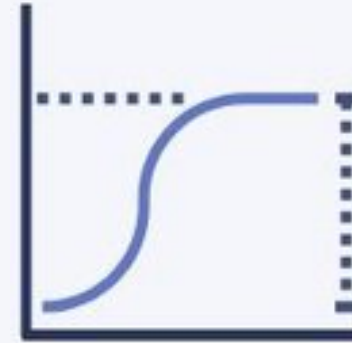
0.218s

30.9%

Training time

Testing score

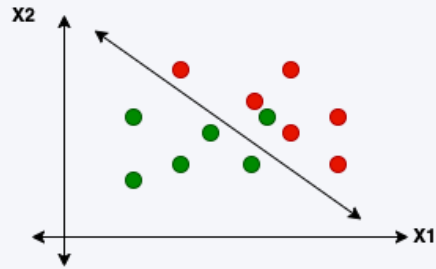
Logistic Regression



0.955s

94.8%

Linear Discriminant Analysis

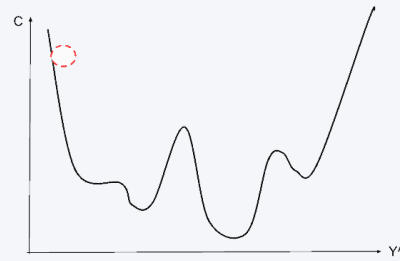


0.527s

94.3%

95.3%

Linear Classifiers with SGD

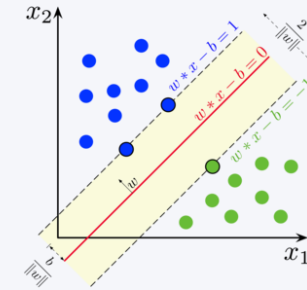


14.37s

93.6%

93.5%

Linear Support Vector Classification



8.12s

95.6%

95.9%

Gaussian Naive Bayes



Training time

22.45s

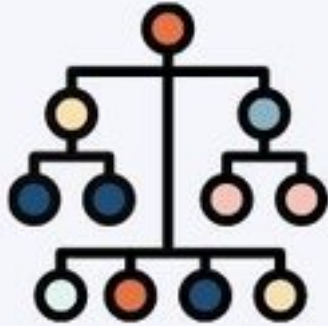
Training score

95.8%

Testing score

96.2%

Decision Tree Classifier



1.562s

95.3%

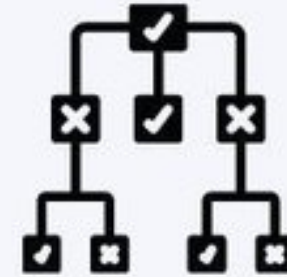
96.0%

Training time

Training score

Testing score

Random Forest Classifier



264.8s

96.5%

97.0%

K-Nearest Neighbors



Training time

22.45s

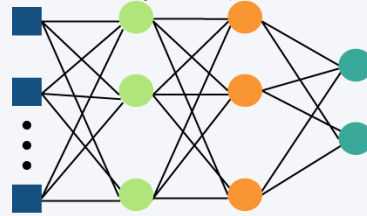
Training score

95.8%

Testing score

96.2%

Multi-Layer Perceptron Classifier



Training time

666.8s

Training score

96.5%

Testing score

96.9%

Model	Training time	Training score	Test score
Linear Regression	0.218s	/	30.9%
Logistic Regression	0.955s	/	94.8%
Linear Discriminant Analysis	0.527s	94.3%	95.3%
Linear Classifiers with SGD	14.37s	93.6%	93.5%
Linear Support Vector Classification	8.12s	95.6%	95.9%
Gaussian Naive Bayes Classifier	0.22s	/	91.0%
Decision Tree Classifier	1.562s	95.3%	96.0%
Random Forest Classifier	264.758s	96.5%	97.0%
K-Nearest Neighbors	22.45s	95.8%	96.2%
Multi-Layer Perceptron Classifier	666.818s	96.5%	96.9%

The image features a vibrant, multi-colored marbled background with swirling patterns of pink, orange, blue, and yellow. In the center, there is a white rectangular box with a thin black border. Inside this box, the letters 'API' are written in a large, black, sans-serif font. Below the text, a thin horizontal line is drawn across the width of the box.

API

← → ↻ 127.0.0.1:8000/api/ Importer les marque-pages...

Number of neighbors: 7

Weights: distance ▾

Algorithm: ball_tree ▾

Leaf size: 31

P: 2

Metric: minkowski ▾

Submit

← → ↻ 127.0.0.1:8000/api/45/model_results Importer les marque-pages...

[Home](#)

Number of neighbors 7

Weights distance

Algorithm ball_tree

Leaf size 31

p 2

Metric minkowski

Train score 1.00

Test score 0.97

Confusion matrix of the test set:

	1	2	3	4	5
1	983	5	1	0	3
2	6	56	0	0	0
3	3	0	2	0	0
4	7	1	0	11	1
5	9	0	0	0	7