DATA ANALYSIS PROJECT

Joyce LAPILUS – Promo 2025 A4 DIA1
Nov 2023

TABLE OF CONTENTS

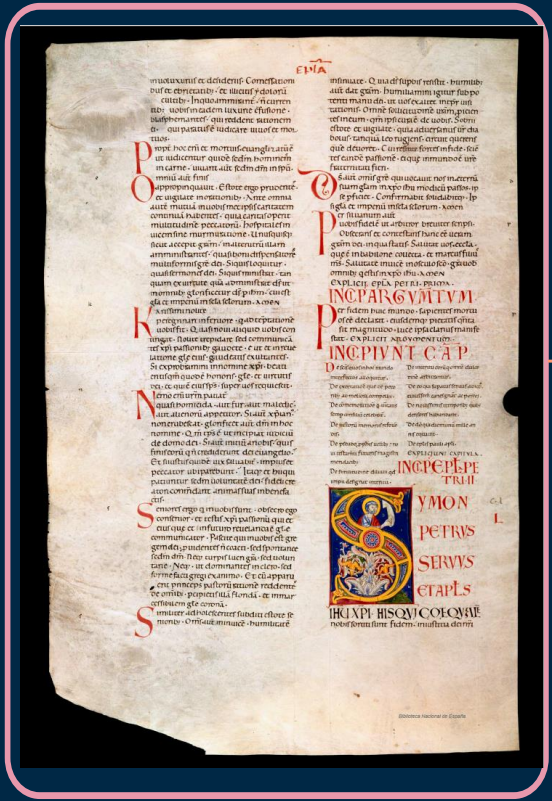
INTRODUCTION

Dataset overview

01

AVILA BIBLE

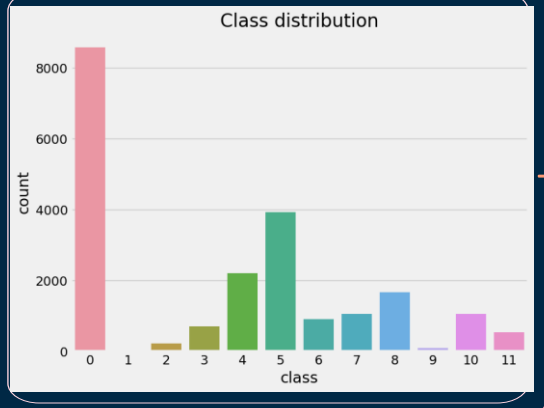
The dataset used is derived from the 12th-century Avila Bible.



OVERVIEW

Our dataset consists of:

- 20,867 samples with a split between training and test sets.
- The datasets have been splitted : 10,430 for the training & 10,437 for testing



Source of the dataset:
Stefano,Claudio, Fontanella,Francesco, Maniaci,Marilena, and Freca,Alessandra. (2018). Avila. UCI Machine Learning Repository. https://doi.org/10.24432/C5K02X.

My primary goal was to predict the copyist responsible for each manuscript among 12 distinct copyists labeled differently (A, B, C, etc.).

Problem statement

DATA

Preparation & Analysis

02

FEATURES

- 10 distinct features
  - Intercolumnar distance
  - Weight
  - Etc.
- Result of a paleographic analysis

```
ATTRIBUTE DESCRIPTION

ID      Name
F1       intercolumnar 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
```

study and academic discipline of the analysis of historical writing systems, the historicity of manuscripts and texts, subsuming deciphering and dating of historical manuscripts, including the analysis of historic handwriting, signification and printed media.

Definition of “paleography”

PREPROCESSING

```
train_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10430 entries, 0 to 10429
Data columns (total 11 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   intercolumnar distance                10430 non-null  float64
 1   upper margin                          10430 non-null  float64
 2   lower margin                          10430 non-null  float64
 3   exploitation                          10430 non-null  float64
 4   row number                            10430 non-null  float64
 5   modular ratio                         10430 non-null  float64
 6   interlinear spacing                   10430 non-null  float64
 7   weight                                10430 non-null  float64
 8   peak number                           10430 non-null  float64
 9   modular ratio / interlinear spacing   10430 non-null  float64
 10  class                                 10430 non-null  object
dtypes: float64(10), object(1)
memory usage: 896.5+ KB
```

avila-tr.txt

avila-ts.txt

- Data was already
  - Splitted
  - Cleaned
  - Normalized

PREPROCESSING

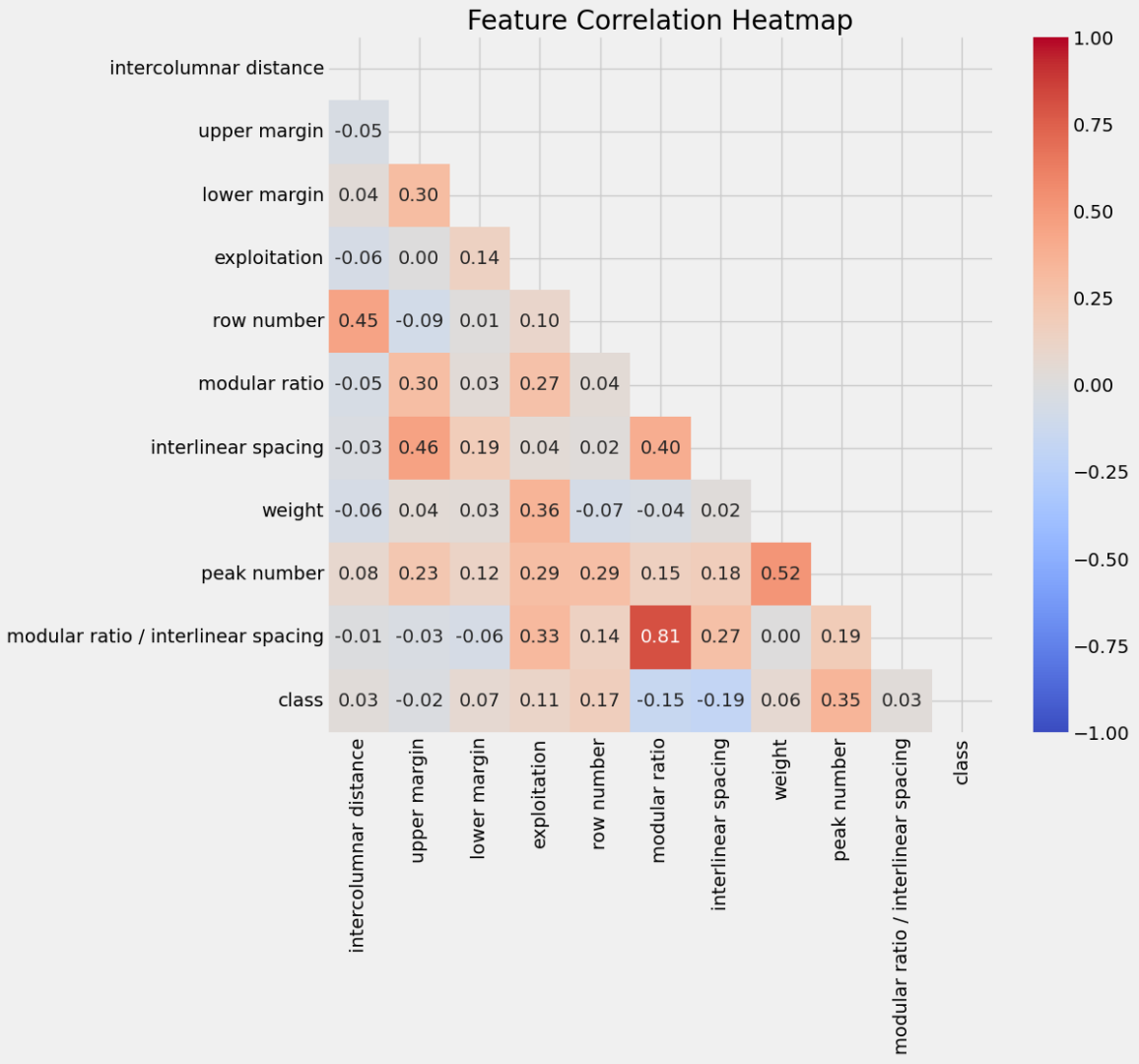
	intercolumnar distance	upper margin	lower margin	exploitation	row number	modular ratio	interlinear spacing	weight	peak number	modular ratio / interlinear spacing
count	10430.000000	10430.000000	10430.000000	10430.000000	10430.000000	10430.000000	10430.000000	10430.000000	10430.000000	10430.000000
mean	0.000852	0.033611	-0.000525	-0.002387	0.006370	0.013973	0.005605	0.010323	0.012914	0.000818
std	0.991431	3.920868	1.120202	1.008527	0.992053	1.126245	1.313754	1.003507	1.087665	1.007094
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	-0.541991	-0.372457	-0.516097
50%	0.043885	-0.055704	0.217845	0.095763	0.261718	-0.058835	0.220177	0.111803	0.064084	-0.034513
75%	0.204355	0.203385	0.352988	0.658210	0.261718	0.564038	0.446679	0.654944	0.500624	0.530855
max	11.819916	386.000000	50.000000	3.987152	1.066121	53.000000	83.000000	13.173081	44.000000	4.671232

- Data was already
  - Splitted
  - Cleaned
  - Normalized

- Normalization
  - Standardizes the Scales
  - Improves Model Performance
  - Enhances Interpretability
  - Reduces Numerical Instabilities

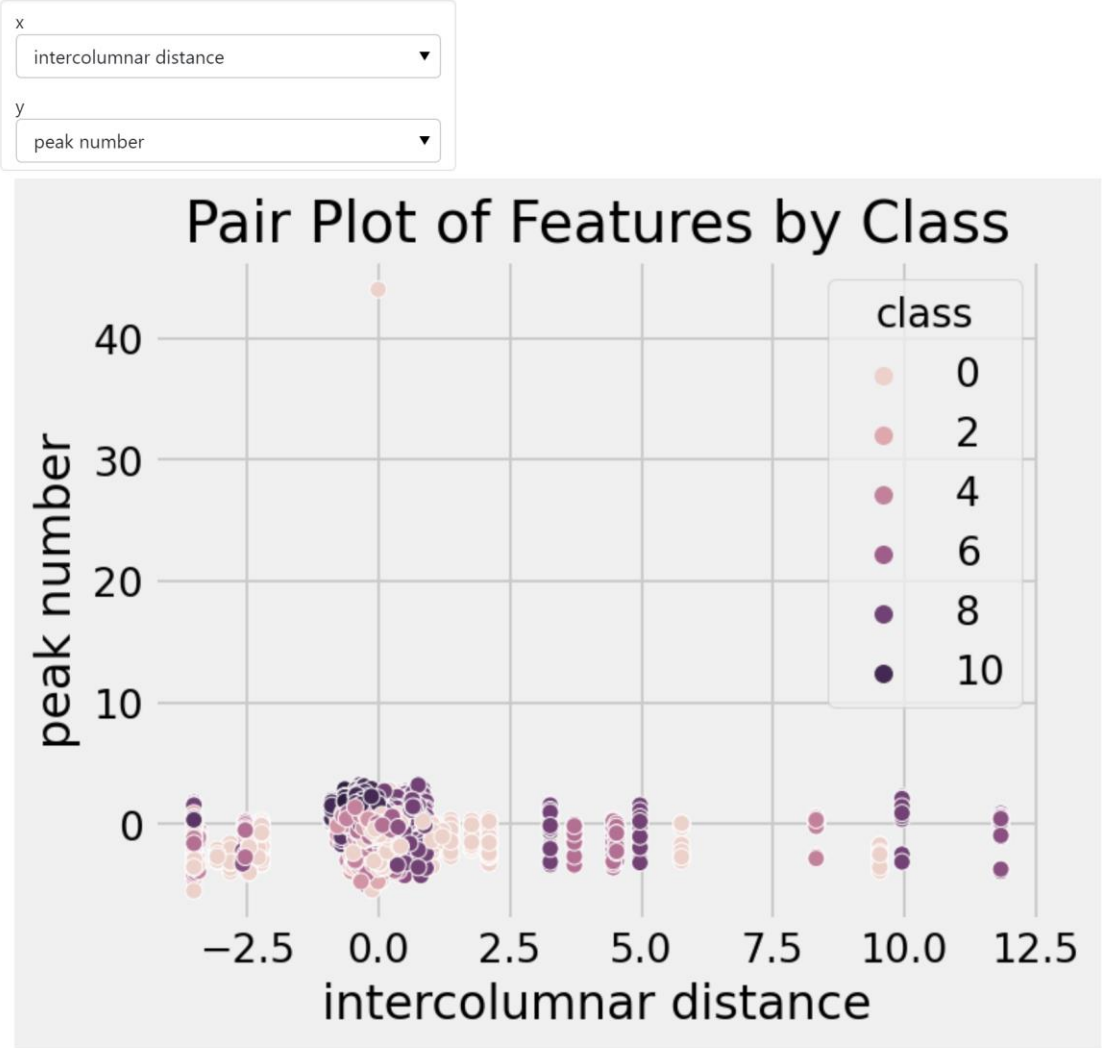
VISUALIZATIONS

- Correlation matrix
- After encoding the labels



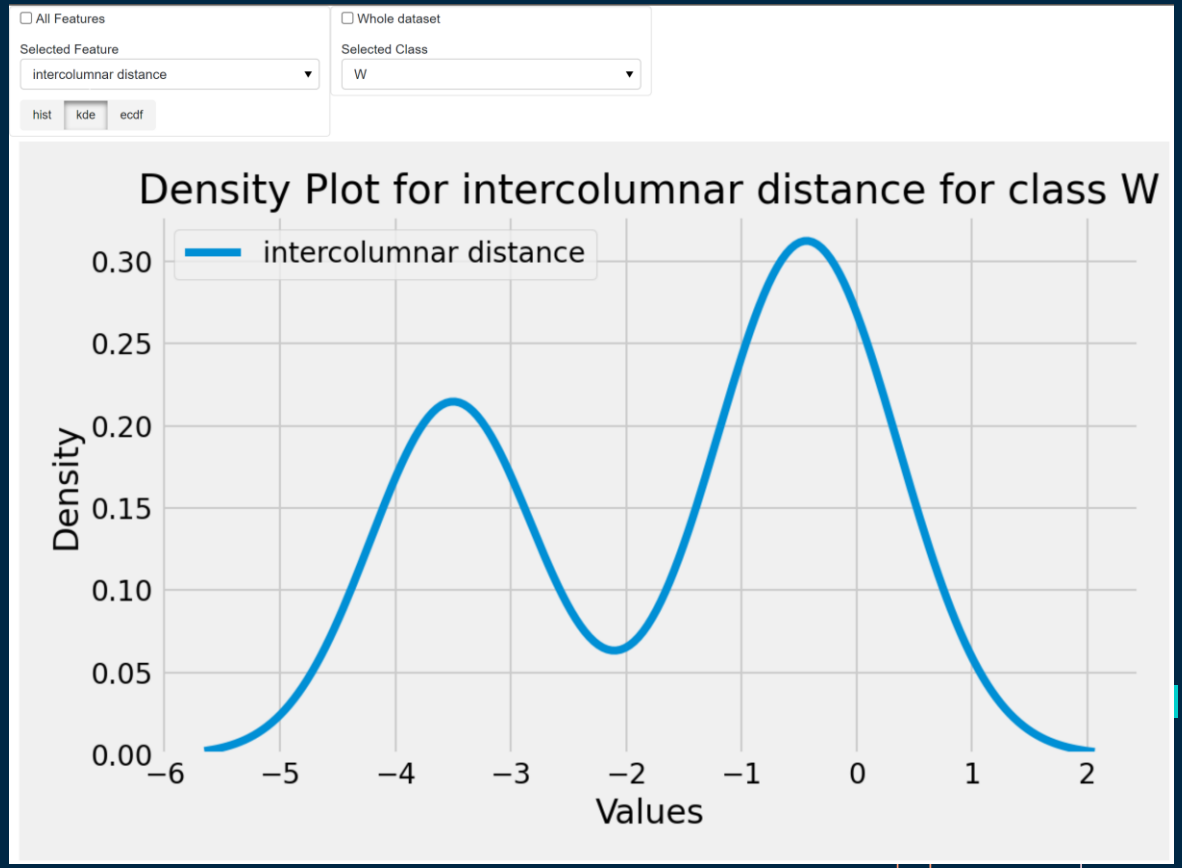
VISUALIZATIONS

- Interactive pair plots
- With *Panel* library
- To identify
  - Correlations
  - Patterns
  - And more.



VISUALIZATIONS

- Interactive dashboard
- With *Panel* library
- Distribution of features
  - One feature or All features
  - Per class or the whole dataset



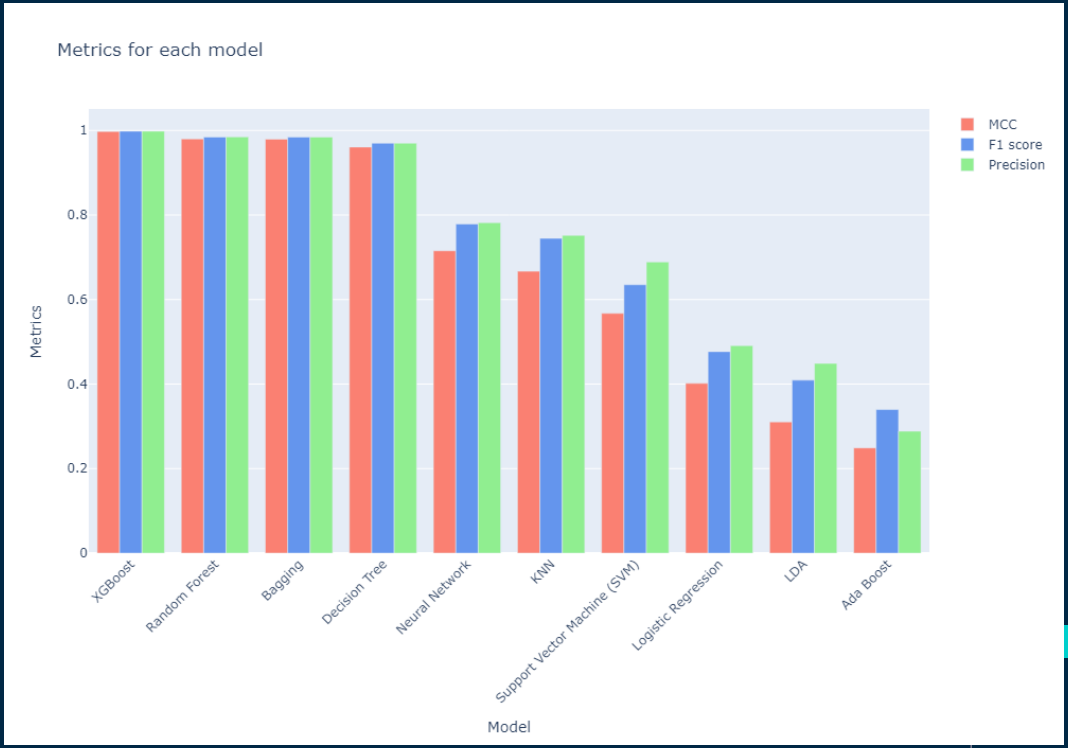
MODELING

Evaluation

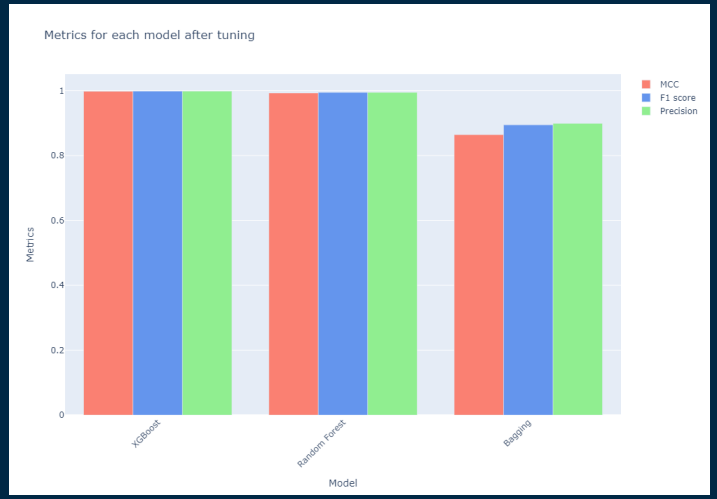
03

CLASSIFIERS

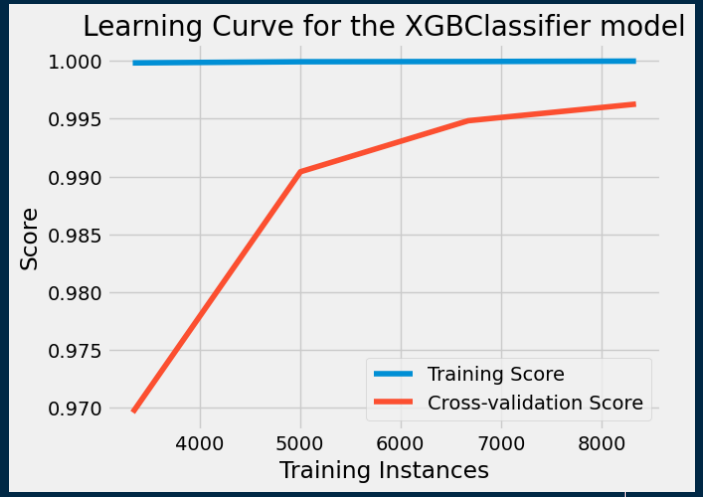
- Matthews' Correlation Coefficient (MCC)
  - Provides a balanced measure
- Precision
  - Focuses on the accuracy of positive predictions
- F1 score
  - Balances precision and recall



PARAMETERS TUNING



Metrics comparison after tuning the top 3 models



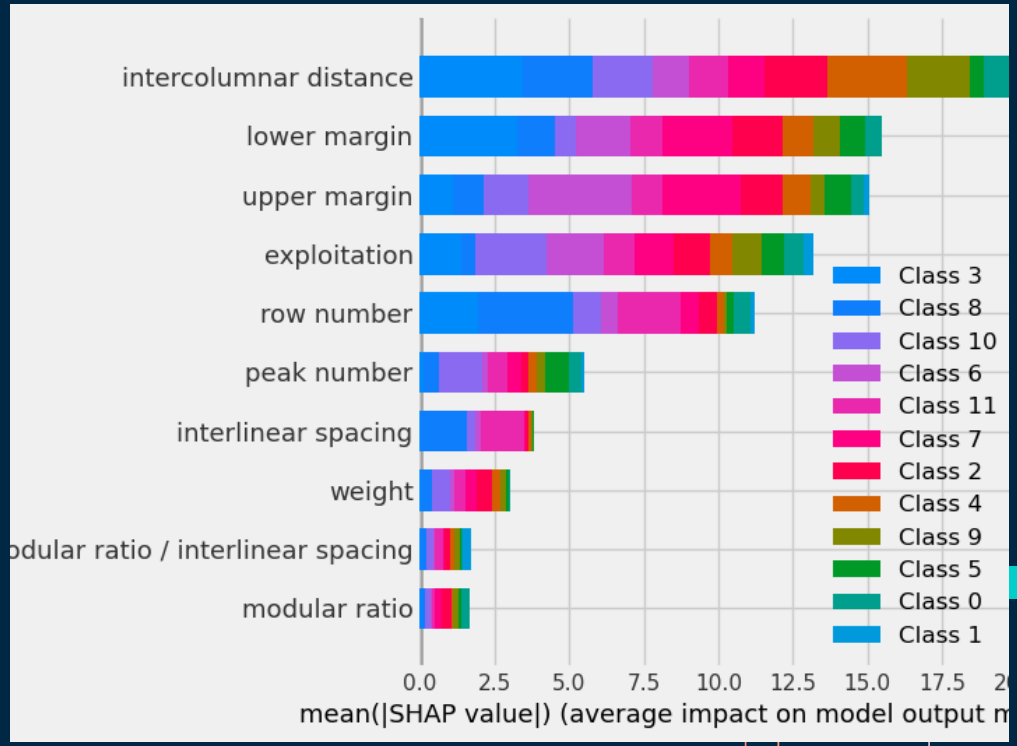
Learning Curve of **XGBoost** model

FEATURE IMPORTANCE

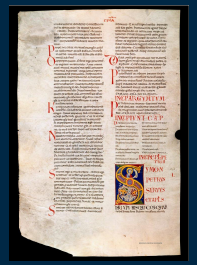
Here is how the **XGBoost** model explains its predictions.

Metrics:

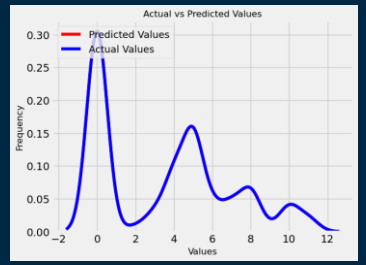
- MCC: 99.8%
- Precision: 99.8%
- F1 score: 99.8%



CONCLUSION



Discovered a new domain



XGBoost is the best model

Created a prediction API with Flask

Do you have any questions?

Joyce LAPILUS – ESILV A4 DIA1
Promo 2025

THANKS