MUNI
FI

Wine quality classification

Current progress of implementation

Miroslav Mažgut (525136), Filip Gregora (525265)

Faculty of Informatics, Masaryk University

April 3, 2025

What was implemented

- Exploration analyzes
- Data preprocessing
- Models:
  - Naive Bayes
  - Regression Tree
  - Linear Regression
  - Support vector machines (SVM)
  - Neural Network
- Model evaluation

Exploration analyzes

Number of samples: 6496 from, Red wine: 1599 (25%),
White wine: 4898 (75%)

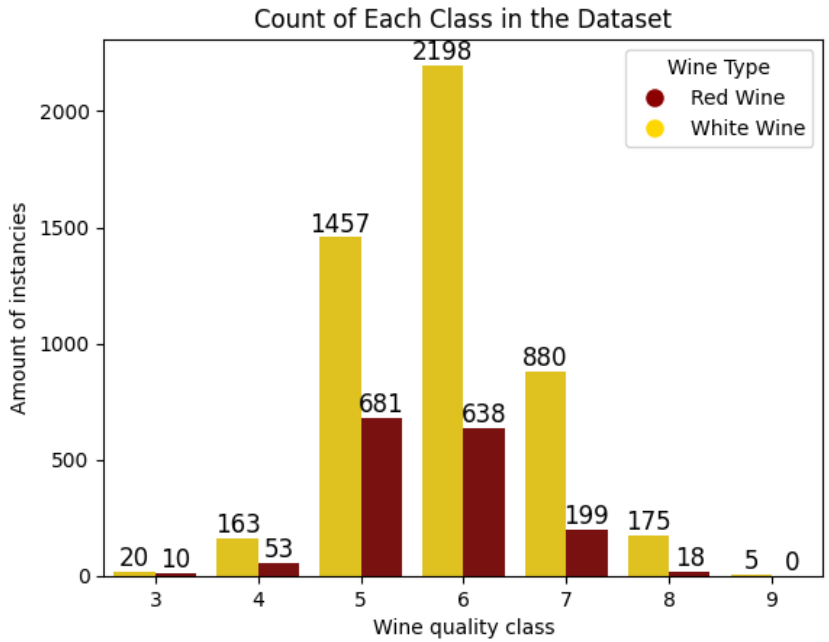
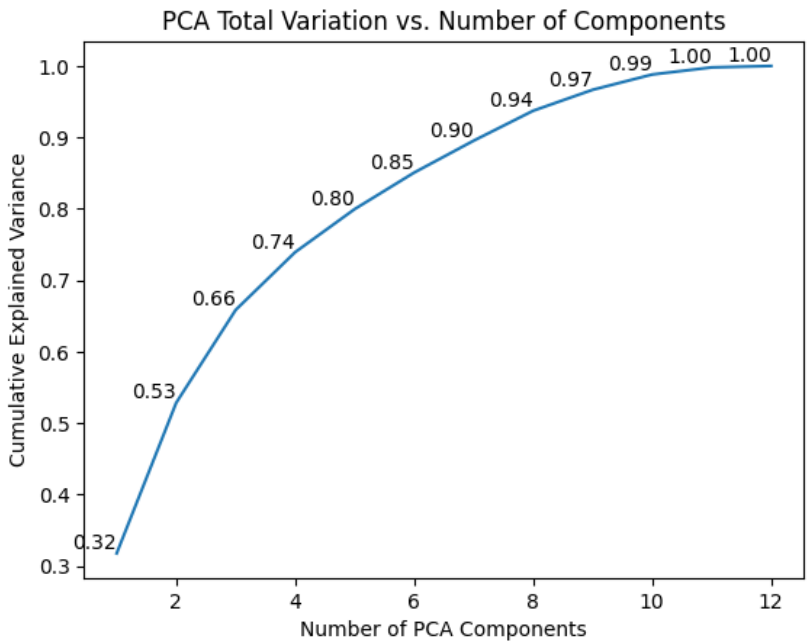


Figure: Data distribution in individual classes concerning wine type

Preprocessing

- Number of attributes: 12
- Some attributes negatively correlated
- Normalization of data
- Reduction of dimension using PCA:



Trained models and evaluation

We used a grid search to find hyperparameters. Trained models:

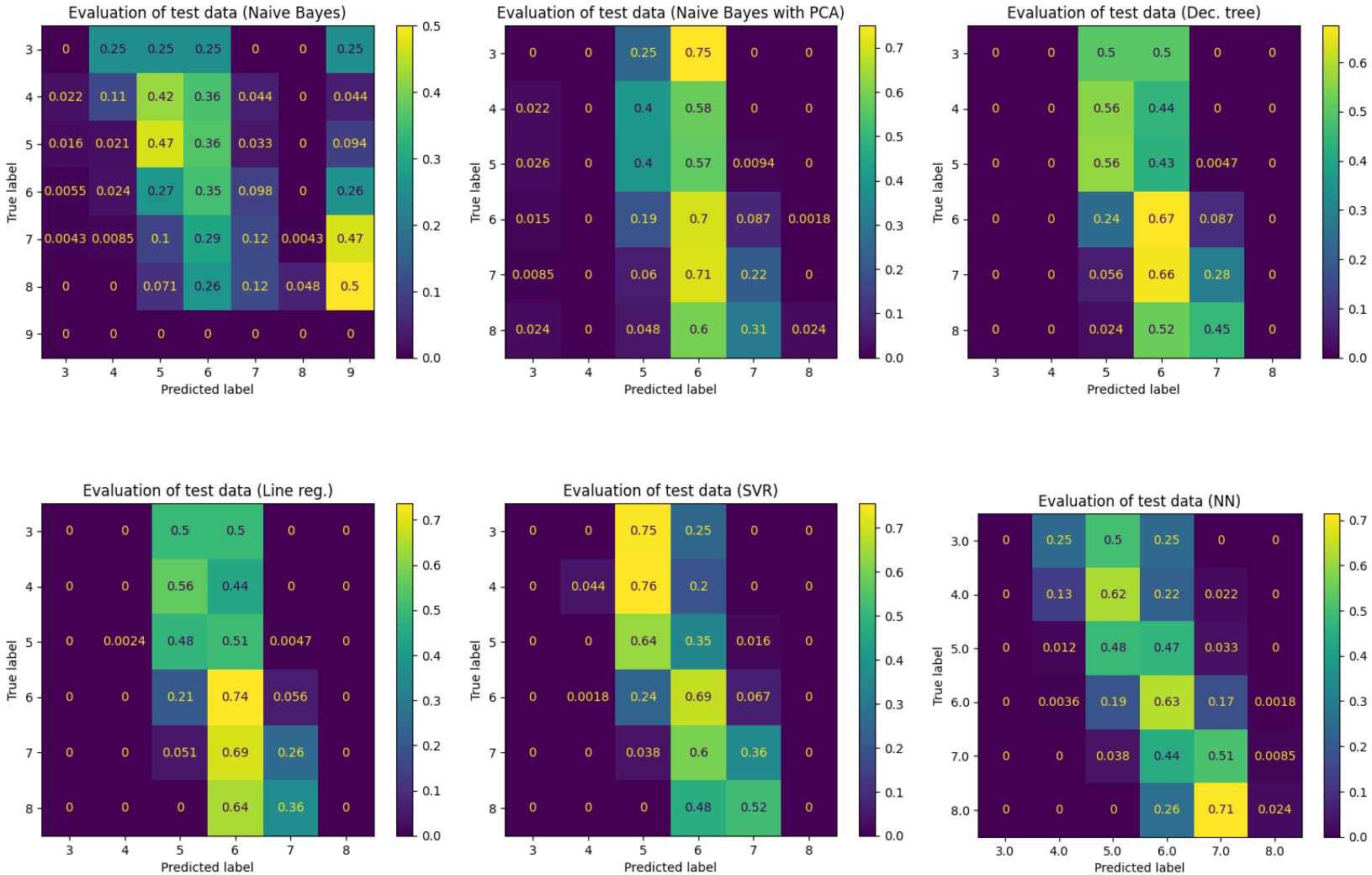
- Naive Bayes without any preprocessing
- Naive Bates on reduced data (PCA components: 8)
- Decision Tree (Max depth: 4)
- Linear regression (PCA components: 12, Ridge alpha: 42)
- SVM (Polynomial degree: 2, Epsilon: 0.1, Kernel: RBF
- Neural network (MLP with five layers: (32, 128, 64, 16, 8))

Evaluation parameters: Mean square error (MSE), Accuracy (A), Balanced accuracy (BA)

Model	MSE Train	MSE Test	A Train (%)	A Test (%)	BA Train (%)	BA Test (%)
Naive Bayes (raw)	2.687	2.596	34	33	35	18
Naive Bayes with PCA	0.783	0.843	51	47	25	22
Regression Tree	0.529	0.536	54	52	22	25
Linear Regression	0.535	0.543	54	51	21	24
SVM	0.399	0.468	63	56	27	29
Neural Network	0.344	0.501	64	55	34	31

Table: Model evaluation results

Prediction distribution



Conclusion and Future work

The poor results were due to a highly imbalanced dataset with unrepresented classes (especially wines of quality 3 and 9). Due to data anonymization, we could not obtain more information on the dataset. Therefore, we would like to try the following upsampling methods:

- naive upsampling (RandomOverSampler)
- chatGPT balancing
- SMOTENC
- KMeansSMOTE

Thank You for Your Attention!

MUNI
FACULTY
OF INFORMATICS