MUNI
FI

Wine quality classification

Finished implementation

Miroslav Mažgut (525136), Filip Gregora (525265)

Faculty of Informatics, Masaryk University

April 23, 2025

What was implemented from the last presentation

- Sampling:
  - Random sampler
  - SMOTE
  - KMeanSMOTE
  - Large language model (prompting ChatGPT)
- Training model on sampled data

Used sampling methods

- *Random Sampling*: Selects data points randomly from a dataset without considering class distribution.
- *SMOTE (Synthetic Minority Over-sampling Technique)*: Balances class distribution by generating synthetic examples of the minority class using feature-space interpolation.
- *KMeans-SMOTE*: Enhances SMOTE by first clustering the data (typically with KMeans) and then applying SMOTE within clusters to preserve data structure and reduce noise.
- *ChatGPT Prompting*: Uses a ChatGPT to guide or generate samples based on prompts, enabling controlled data generation aligned with desired attributes or classes.

ChatGPT prompting

For the uploaded dataset, do upsampling so that the number of samples is balanced for the quality attribute. The dataset contains physicochemical analysis of northern Portuguese wines.

For upsampling interpret each sample as row and predict the data only based on your opinion. For upsampling use deep AI generative approach. Provide me upsampled dataset.

`https://chatgpt.com/share/68053f08-2e4c-8011-a21c-1b7f9f6cacd5`

Final Training and Evaluation

In this phase, we will train each of the selected machine learning models — Naive Bayes, Decision Tree, Linear Regression, Support Vector Machine (SVM), and Neural Network — using the sampled dataset.

To assess model performance, we will apply the same evaluation metrics introduced in the previous presentation: Mean Squared Error (MSE), Accuracy (A), and Balanced Accuracy (BA).

Project Contribution

An existing project working with the same dataset can be found here:

- `https://www.scirp.org/journal/paperinformation?paperid=107796`

Our main contribution lies in the application of sampling techniques to address class imbalance. By increasing the representation of underrepresented classes, we aim to enhance the performance and reliability of machine learning models.

Division work

Filip Gregora: Data sampling implementation, Model training
Miroslav Mažgut: Model evaluation, 3th Presentation

Thank You for Your Attention!

MUNI
FACULTY
OF INFORMATICS