Big data processing

Project presentation

Group members

Dataset description

- Overview of data features
- Description of target variable
- Data pre-processing:
  - Number of missing entries (before and after)
  - Encoding of categorical data
  - Feature scaling
  - Feature transformations
  - Splitting of the dataset

Comparison of machine learning models

- Model 1: Linear regression
- Model 2: XGBoost
- Comparison metrics:
  - Precision
  - Accuracy
  - F1 Score
  - …
- Which model performs better?

Data usage for further analysis

- Data streaming:
  - Why process/monitor the data in real-time?
  - What decision can be made in close to real time?
- Graph representation:
  - What are nodes of the graph?
  - What are edges?
  - Do edges have a direction? Are they weighted?

Conclusions