Predictive Quality

1 Task Description

1.1 Objective: Prediction of the part quality based on production line data

The objective of the group work is to build a classification model that can accurately predict the product quality based on production line data. To win practitioners' trust in the model, expert-based features have to be used and the model complexity should be kept on a low to medium level.

Due to the size of the data set, teamwork is required to succeed by splitting up tasks wherever possible!

1.2 Domain-specific knowledge and literature: measuring principle

Along the production lines, data is collected for each product. This data represents in process sensor values and sensor values from quality assurance steps.

The data format can be binary or continuous and usually comes with a timestamp of the measurement. Besides the values, the measurement timestamp can also be of interest for quality prediction (e.g. bad influences only appear during night shifts). Furthermore, the timestamps can be used to derive further features such as the processing time or storage time of a product.

1.3 Data: Numerical, date, categorical features and labels

The provided data consists of three files with a table-like structure:

- Numeric.hdf: This file includes per row numeric information about one product referenced by its ID. Each column represents a sensor in one station of a production line. The name convention for columns is as follows: L2_S0_D1 → Sensor 1 at station 0 at line 2
  The last column 'Response' holds the labels for each product (0: No defect in the product, 1: defect in the product)
- Categorical.hdf: This file includes per row categorical information about one product referenced by its ID. Each column represents a

sensor in one station of a production line (same column name convention as before).

- Date.hdf: This file includes the timestamps when each numerical or categorical sensor value was taken for each product per row and each sensor per column (same column name convention as before).

1.4 Evaluation: Quality prediction accuracy

As the task is to classify products regarding their quality, have a look at classical metrics to evaluate classification performance. Hint: Have a look at the class distribution!

1.5 Research questions

As the provided data set allows a variety of analysis, it is recommended to answer the questions below to develop the required classification model.

Week 1: Data exploration

Which structure has the production system? Are all features present for all products? Which product clusters can be identified regarding the material flow through the production system? Do features correlate? Are there outliers in the features? Which features can be derived from the timestamp information?

Week 2: Feature engineering

Which features are relevant for the classification task? Which further preprocessing steps are necessary before modelling? How does dimensionality reduction increase the separability of classes?

Week 3: Modeling

Which models can be used for the data? What are benchmark models for predictive quality tasks found in the literature (have a look at Google Scholar, sciencedirect.com or ieeexplore.ieee.org)?

Week 4: Results and model insights

How good does the feature-model-combination perform on test data? What are levers to increase model performance? Is hyperparameter tuning possible? Which are the most important features for the model?

Week 5: Report writing

Please consider the report structure and remarks presented on the next pages.

2 Report

- One report per group in German or English
- Identify, which group member contributed to each chapter in which share (for example: Introduction: All 20 %, Methods: Person A 50 %, Person B 50 % …)
- 10 – 12 pages (for 5 group members)

Structure of the report:

1. Introduction (15 %)
   - Why is this topic dealt with?
2. Methods/Experiments (20 %)
   - How was this topic dealt with?
3. Results (30 %)
   - What were the outcome and results?
4. Discussion (30 %)
   - How can the results be interpreted, what are the consequences and limitations?
5. Summary (5 %)
   - What could be next steps and key learnings?

Introduction

- Introducing the reader to the topic
- The reader should understand the general topic and the motivation.

Methods/Experiments

- Description of the entire data-handling and analysis process (for example application of the KDD process)
- Description of the applied methods and models
- Descriptions should enable the reproduction off the results presented in the next chapter
- Why did you chose the respective methods and models?

Results

- Clear presentation of the results.
- Precise and meaningful labeling of illustrations and diagrams

Discussion

- Core of the report
- Interpretation and evaluation of the results
- Direct references to the results in previous section

Summary

- Indicate the key findings you had and the future research you would conduct if you had more time

Annex

- Additional plots
- Code

General remarks – Citations

- No copyright infringement, will be counted as fraud
- Indication in the text with first author and year, e.g. MUSTERMANN 2009
- Add references to illustrations from other authors

General remarks – Bibliography

- Required for all sources used
- Information in the bibliography with authors, title, journal, publisher, edition, year of publication and page number
- For online sources, the main page with date of the last access

General remarks – Language and expression

- Short and meaningful sentences
- Precise writing style
- Avoidance of unnecessary filler words
- No first person
- Cover page without page number

General remarks – Content

- Methodological procedure
- Originality
- Complexity
- Integration of domain knowledge
- Interpretation of results and outlook
- …

3 Presentation

- 20 minutes presentation
- Q&A session. Everyone should answer at least once.