

Applied Machine Learning

P1- Steel Production Analysis



**Montanuniversität
Leoben**

Nikhil Yadav

M12525177

Abstract

This project focuses on analyzing steel production data using machine learning techniques to predict production performance and understand key influencing factors. The dataset was preprocessed, normalized, and analyzed through exploratory data analysis to uncover underlying patterns and correlations. Multiple regression-based machine learning models were trained and evaluated to compare their predictive performance. The goal of this study is to identify the most suitable model for steel production prediction and provide data-driven insights for industrial decision-making.

Introduction

Background

Steel production is a critical industrial process that plays a major role in infrastructure, manufacturing, and economic development. Accurate prediction and analysis of steel production parameters can help improve efficiency, reduce operational costs, and optimize resource utilization. With the increasing availability of industrial data, machine learning techniques provide powerful tools for analyzing complex, multivariate datasets generated from manufacturing processes.

Objectives

- 1.) To perform exploratory data analysis (EDA) to understand data distribution, correlations, and outliers.
- 2.) To preprocess and normalize the dataset for effective model training.
- 3.) To implement and compare multiple machine learning models for steel production prediction.
- 4.) To evaluate models using standard performance metrics such as RMSE, MAE, and R².
- 5.) To identify the most suitable model and draw meaningful insights from the results.

Methods

Data Acquisition

The dataset used in this project consists of steel production-related variables collected from an industrial process dataset provided for academic analysis. The data includes multiple numerical features representing process parameters and one target variable representing production output. The dataset was provided in CSV format and loaded using Python's pandas library.

Data Analysis

The data analysis process included:

- Handling missing values and ensuring data consistency
- Feature normalization using StandardScaler
- Splitting the dataset into training, validation, and testing sets

- Exploratory Data Analysis (EDA), including:
 - Correlation matrix heatmaps
 - Feature distributions
 - Box plots for outlier detection
 - Target variable distribution analysis

Machine learning models implemented:

- Random Forest Regressor
- Support Vector Machine (SVM)
- Multi-Layer Perceptron (MLP)
- Gaussian Process Regressor

Each model was trained on the normalized training dataset and evaluated on the validation set.

Tools Used

- . **Programming Language:** Python
- . **Libraries:** pandas, numpy, matplotlib, seaborn, scikit-learn
- . **Development Environment:** Visual Studio Code
- . **Version Control:** Git & GitHub
- . **Documentation:** Overleaf (LaTeX)

Results

The models were evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R² score, training time, and inference time.

Findings

- The Random Forest model achieved the best overall performance with the lowest RMSE and highest R² score.
- The SVM model showed moderate predictive performance but required higher inference time.
- The MLP model struggled with generalization, resulting in a negative R² score.
- The Gaussian Process model performed poorly due to high variance and computational complexity.

Overall, ensemble-based models proved more effective for this dataset compared to neural networks and probabilistic models.

Visualizations

The following visualizations were generated:

- Correlation heatmap showing feature relationships
- Feature distribution histograms
- Box plots for outlier detection
- Bar plots comparing model performance
- Prediction vs Actual scatter plots
- Residual analysis plots

(All figures are stored in the figures/ directory.)

Conclusion

This project demonstrated the effectiveness of machine learning techniques in analyzing and predicting steel production data. Among the evaluated models, the Random Forest Regressor provided the most reliable and accurate predictions. The study highlights the importance of proper data preprocessing and model selection in industrial data analysis.

Limitations of this work include reliance on a single dataset and lack of real-time process variables. Future work may involve incorporating time-series data, deploying models for real-time monitoring, and exploring deep learning architectures for improved performance.

License

The dataset used in this project is intended for academic and educational purposes only.

This project is licensed under the **GNU General Public License v3.0**

License – see the [LICENSE](#) file for details.

Acknowledgments

- Montan University of Leoben (MUL) for academic guidance and resources
- Course instructors for project framework and evaluation criteria
- Scikit-learn and open-source Python community
- ChatGPT was used for:
 - Code structuring and debugging
 - Documentation drafting
 - Report formatting guidance