A
Project Report
of
Applied Machine Learning
& Deep Learning

Course: Applied Machine & Deep Learning (190.015)

Semester: Winter Semester 2025/26

Project Category: P1 – Steel Production Data

Student: Rishabh Kothari

Matrikulation Number: 12519563

Institution: Montanuniversität Leoben

Index

1. Topic and Motivation

Steel production is one of the most critical industrial processes in modern manufacturing and forms the backbone of sectors such as construction, automotive engineering, infrastructure development, energy production, and heavy machinery. The quality of steel products directly affects structural integrity, durability, safety, and long-term performance. Even small deviations in material properties can result in significant economic losses, safety hazards, and increased maintenance costs.

The steel manufacturing process involves a large number of interacting physical, chemical, and thermal parameters. These include material composition, furnace temperature, pressure levels, cooling rates, and mechanical processing conditions. Due to this complexity, traditional rule-based control systems and purely physics-driven models often struggle to accurately predict final product quality under varying operational conditions.

With the advent of Industry 4.0 and the increasing deployment of sensors in industrial environments, large volumes of process data are now continuously collected. This data provides an opportunity to apply data-driven methods, particularly machine learning, to model complex relationships between process parameters and quality indicators. Machine learning models can learn these relationships directly from historical data without requiring explicit physical equations.

The motivation of this project is to explore whether supervised machine learning regression techniques can be effectively applied to predict a continuous steel quality factor from normalized production data. Such predictive capabilities can support early quality assessment, enable proactive process adjustments, reduce material waste, and improve overall production efficiency. Furthermore, this study aims to compare different regression algorithms to understand their strengths and limitations in an industrial context.

2. Related Work

The application of machine learning in industrial process monitoring and quality prediction has been widely studied in recent years. Early approaches focused on statistical regression methods, such as linear and multivariate regression, due to their simplicity and interpretability. These methods are still widely used in manufacturing environments where transparency and explainability are essential.

However, as industrial systems became more complex, researchers began exploring nonlinear machine learning models. Support Vector Regression (SVR) has been frequently applied to manufacturing datasets due to its ability to handle nonlinear relationships through kernel functions. Several studies report that SVR performs well on medium-sized datasets with normalized features and limited noise.

Decision Tree-based models offer another popular approach, particularly in industrial settings. Decision Trees are capable of modelling nonlinear interactions between variables and provide intuitive decision rules that can be interpreted by domain experts. Nevertheless, single decision trees are prone to overfitting, especially when the dataset contains noise or redundant features.

Ensemble learning methods, such as Random Forests, have emerged as powerful tools for industrial prediction tasks. Random Forests combine multiple decision trees trained on random subsets of data and features, resulting in improved generalization performance and robustness. Numerous studies in steel quality prediction and metallurgical process optimization demonstrate that Random Forest models outperform individual learners in terms of accuracy and stability.

In addition to tree-based models, instance-based learning techniques such as K-Nearest Neighbors (KNN) have also been explored. While KNN is conceptually simple, its performance strongly depends on feature scaling, distance metrics, and data distribution. In high-dimensional industrial datasets, KNN often struggles due to the curse of dimensionality.

This project builds upon existing research by implementing and systematically comparing multiple regression algorithms on the same normalized steel production dataset. The goal is not only to identify the best-performing model but also to provide insights into how different learning paradigms behave when applied to industrial process data.

3. Dataset Description

3.1 Dataset Overview

The dataset used in this project originates from a steel production process and consists of numerical measurements collected from sensors and control systems. Each sample represents a single production instance described by a set of input features and an associated output value.

The dataset is provided in two separate files: -

Training dataset: 7,642 samples - **Test dataset:** 3,337 samples

Each dataset contains 22 columns in total: - 21 input features (input1 to input21) - 1 output feature (output)

The output variable represents a continuous steel quality factor, making the task suitable for supervised regression.

3.2 Data Characteristics

All features in the dataset are continuous and have been normalized prior to analysis. Normalization ensures that all features contribute equally during model training and prevents dominance of features with larger numerical ranges. This is particularly important for distance-based and kernel-based learning algorithms such as KNN and SVR.

The dataset does not contain missing values or duplicate samples. This indicates a high level of data quality and allows the focus of the project to remain on model development and evaluation rather than extensive data cleaning.

3.3 Assumptions

Several assumptions are made in this study: - The training and test datasets are representative of the same underlying steel production process. - The relationships between input features and the output variable remain stable over time. - The normalized data adequately captures the relevant information needed for quality prediction.

4. Exploratory Data Analysis

4.1 Distribution Analysis

Exploratory Data Analysis (EDA) was conducted to gain an initial understanding of the dataset. Histograms were generated for all input features and the output variable. The majority of features exhibit unimodal distributions with varying degrees of skewness. Some features are concentrated within narrow ranges, while others display broader distributions.

The output variable shows a continuous distribution, further confirming that the problem is a regression task. No evidence of extreme outliers or abnormal distributions was observed, indicating that the dataset is well-suited for regression modeling.

4.2 Correlation Analysis

A feature correlation matrix was computed to examine linear relationships between variables. The correlation heatmap reveals moderate correlations among certain input features, suggesting potential interactions and shared information. Additionally, some input features show noticeable correlation with the output variable, indicating their relevance for quality prediction.

However, no single input feature demonstrates a dominant correlation with the output. This suggests that the steel quality factor is influenced by a combination of multiple process parameters rather than a single controlling variable. As a result, multivariate regression models are necessary to capture these relationships effectively.

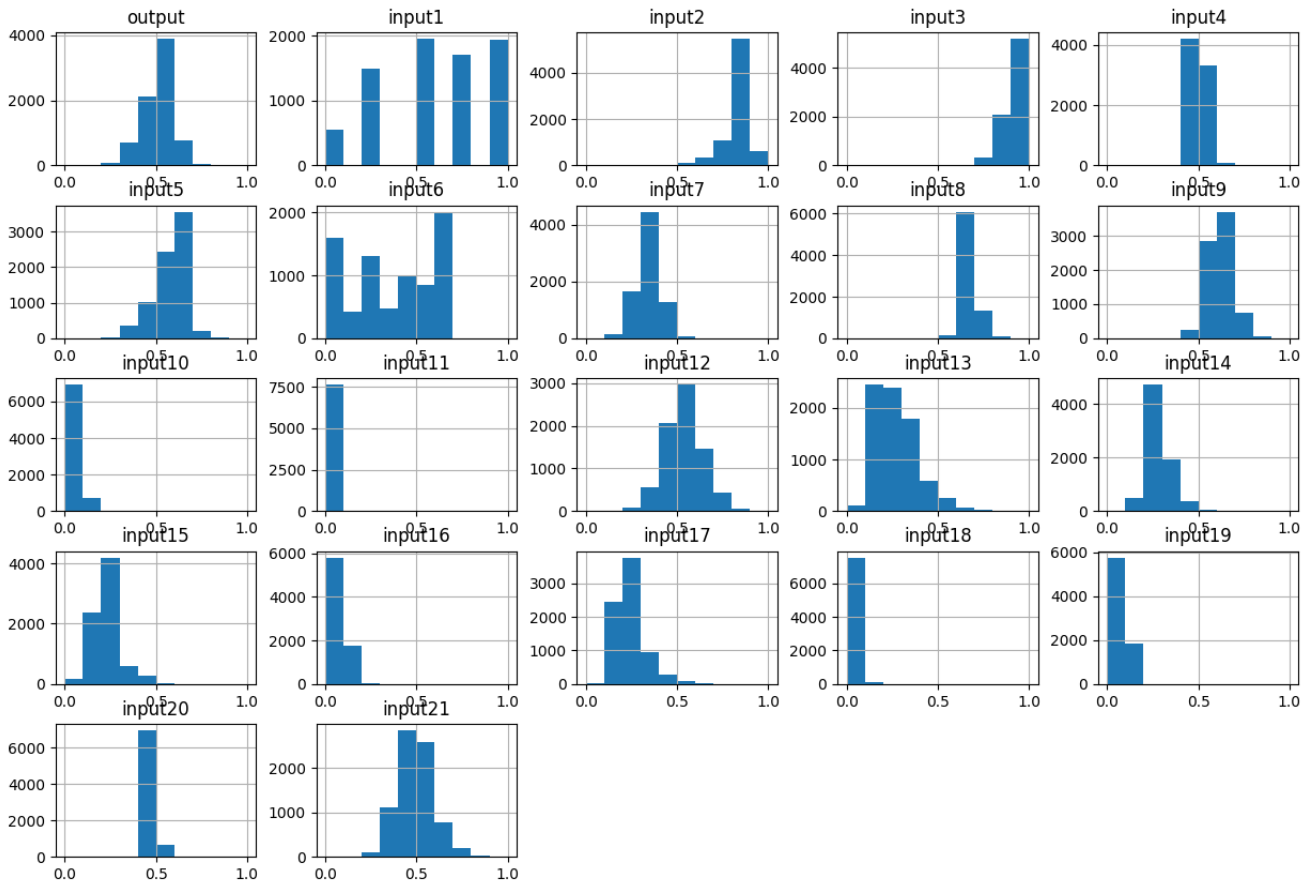

Fig. 1: Feature Distributions

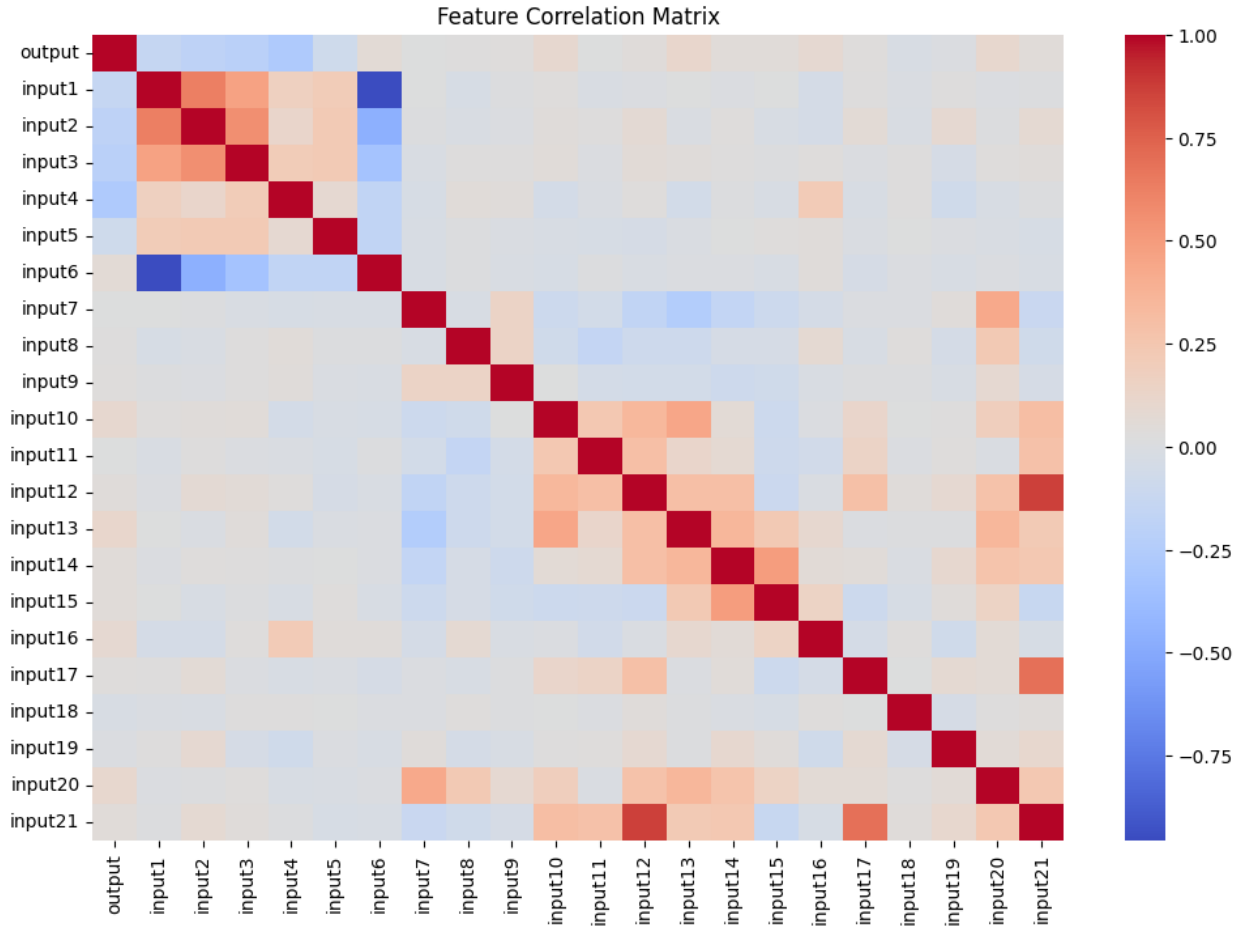

Fig. 2: Feature Correlation Matrix

5. Methodology

5.1 Problem Formulation

The problem is formulated as a supervised regression task. Given an input vector consisting of 21 normalized features, the objective is to predict a continuous output value representing steel quality. The goal is to minimize the prediction error on unseen test data.

5.2 Model Selection Rationale

To ensure a comprehensive evaluation, five regression models representing different learning paradigms were selected:

- **Linear Regression:** Serves as a baseline model and provides insight into linear relationships.
- **Support Vector Regression (SVR):** Captures nonlinear patterns using kernel functions.
- **K-Nearest Neighbors Regressor:** Relies on similarity-based predictions.
- **Decision Tree Regressor:** Models nonlinear feature interactions through hierarchical splits.
- **Random Forest Regressor:** Combines multiple decision trees to improve robustness and generalization.

5.3 Training Procedure

All models were trained using the normalized training dataset. Default hyperparameters were employed to ensure a fair comparison between methods. Each model was evaluated on the independent test dataset to assess generalization performance.

6. Evaluation Metrics

Since the target variable is continuous, regression-specific metrics were used for evaluation:

- **Mean Absolute Error (MAE):** Represents the average absolute difference between predicted and actual values.
- **Mean Squared Error (MSE):** Measures the average squared prediction error and penalizes large deviations.
- **Root Mean Squared Error (RMSE):** Provides error magnitude in the same unit as the output variable.
- **R^2 Score:** Indicates the proportion of variance in the output variable explained by the model.

These metrics collectively provide a comprehensive view of model accuracy, robustness, and goodness-of-fit.

7. Experimental Results

7.1 Quantitative Results

The quantitative performance of all regression models is summarized in Table 1.

Table 1: Performance Comparison of Regression Models

Model	MAE	MSE	RMSE	R^2 Score
Linear Regression	0.0493	0.00331	0.0576	0.493
SVR	0.0601	0.00576	0.0759	0.118
KNN Regressor	0.1028	0.01347	0.1161	-1.064
Decision Tree	0.0533	0.00454	0.0674	0.304
Random Forest	0.0500	0.00399	0.0631	0.390

7.2 Interpretation of Results

Linear Regression achieved the highest R^2 score, indicating that a substantial portion of the variance in the steel quality factor can be explained by a linear combination of input features. This suggests that the underlying process contains strong linear relationships.

Random Forest Regressor performed competitively, demonstrating its ability to capture nonlinear feature interactions and improve robustness. Decision Tree Regressor achieved moderate performance but exhibited higher error compared to ensemble methods.

The KNN Regressor produced a negative R^2 score, indicating poor generalization performance. This highlights the limitations of distance-based methods in high-dimensional industrial datasets.

7.3 Visual Analysis

A bar plot comparing R^2 scores across models was generated to visualize relative performance. Additionally, an Actual vs. Predicted scatter plot for the Random Forest Regressor shows that most predictions align closely with the ideal diagonal line, confirming good predictive accuracy.

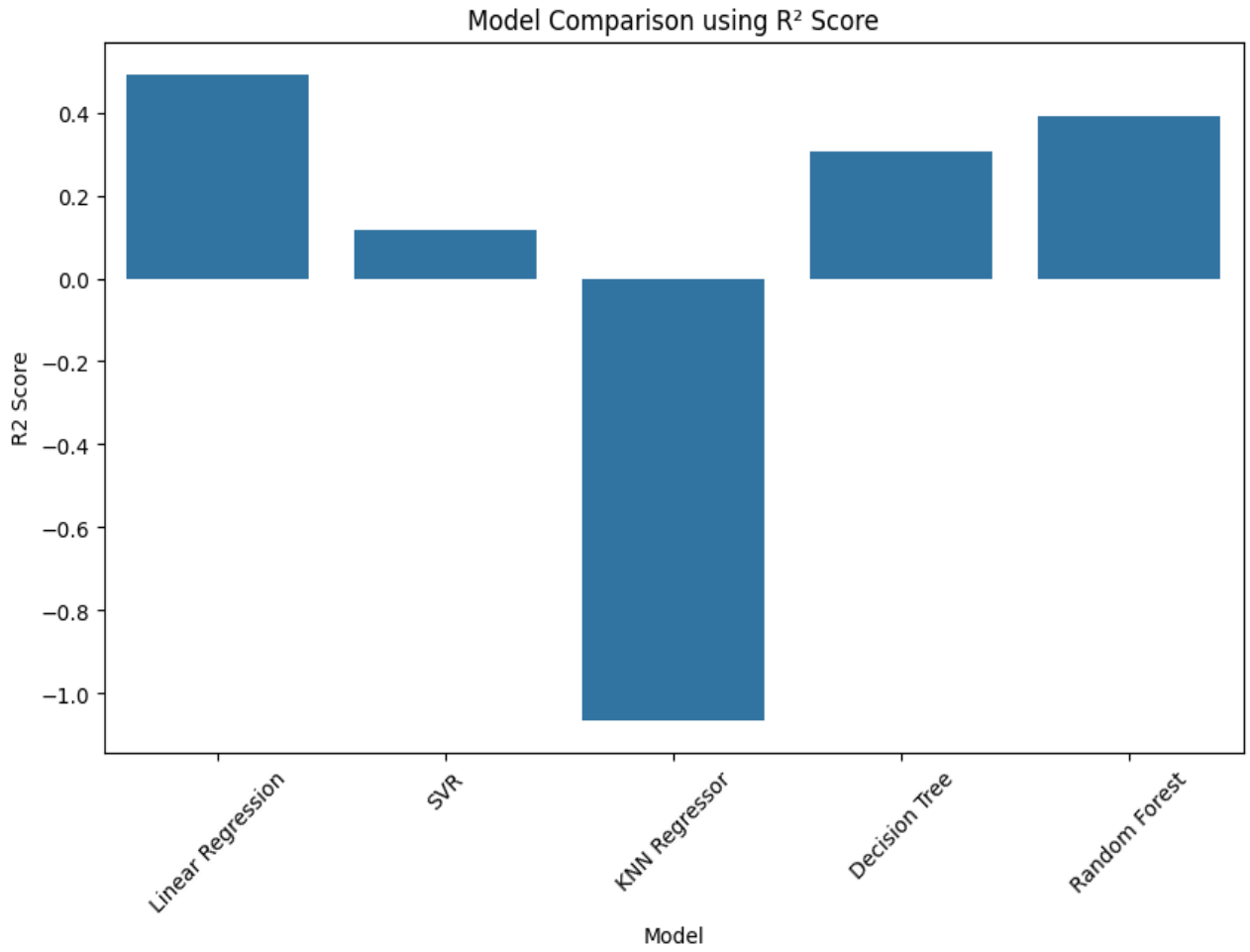

Fig. 3: Model Comparison using R^2 Score

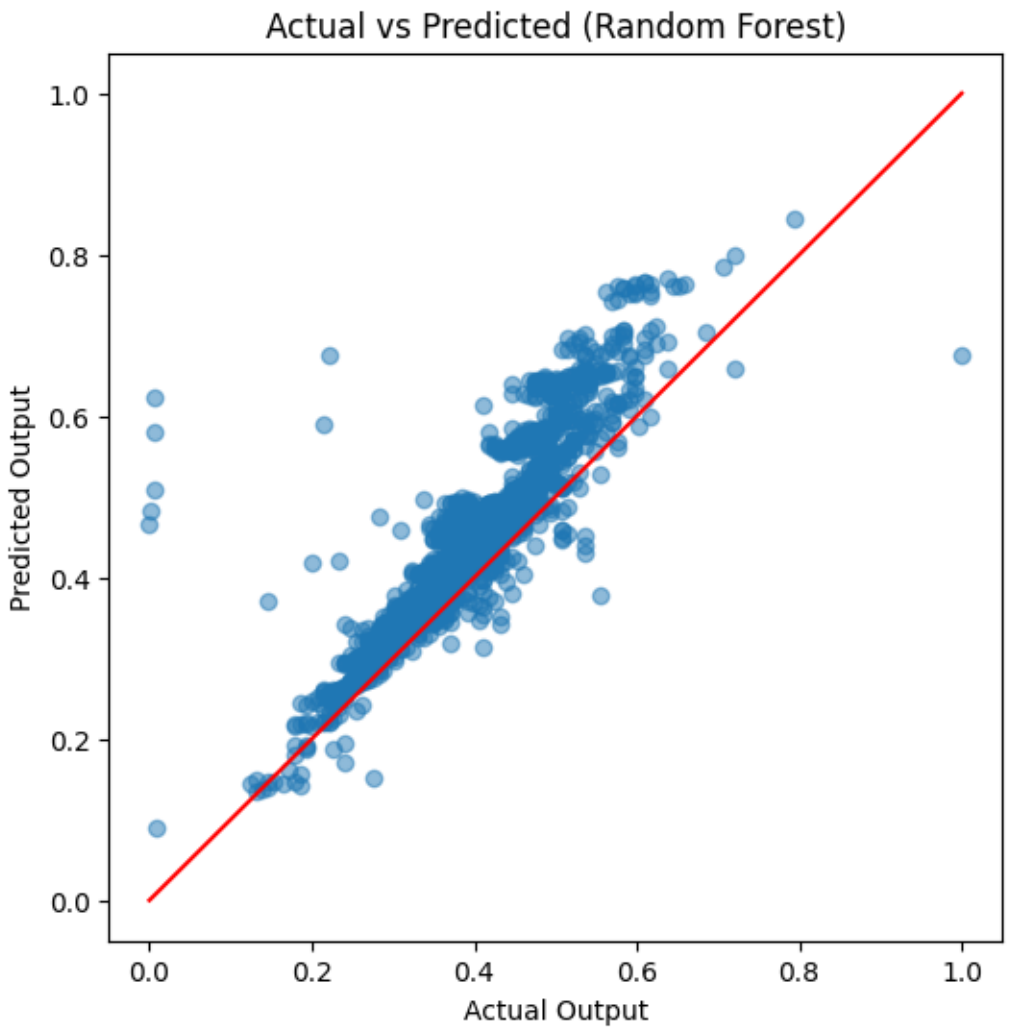

Fig. 4: Actual vs Predicted (Random Forest)

8. Discussion

8.1 Interpretation of Model Performance

The experimental results indicate that Linear Regression achieved the highest R^2 score among all evaluated models. This outcome suggests that a substantial portion of the variability in the steel quality factor can be explained through linear relationships between the input process parameters and the output variable. Such behavior is realistic in industrial processes where many control variables influence quality in an approximately linear manner within normal operating ranges.

The strong performance of Linear Regression also highlights the importance of selecting simple baseline models before applying more complex algorithms. In industrial environments, simpler models are often preferred due to their interpretability, ease of deployment, and lower computational requirements.

Random Forest Regressor achieved competitive performance and demonstrated robustness across the dataset. Although its R^2 score was slightly lower than that of Linear Regression, Random Forest showed stable predictions across different regions of the input space. This indicates that ensemble-based models are effective in capturing nonlinear interactions and compensating for noise present in industrial measurements.

8.2 Analysis of Poorly Performing Models

The KNN Regressor produced a negative R^2 score, indicating that its predictions were worse than simply predicting the mean of the output variable. This behavior can be attributed to the high dimensionality of the feature space and the relatively uniform distribution of normalized features. In such scenarios, distance-based similarity measures become less meaningful, a phenomenon commonly referred to as the *curse of dimensionality*.

Similarly, Support Vector Regression exhibited limited performance improvement compared to Linear Regression. This may be due to suboptimal hyperparameter settings or the fact that the dataset already exhibits strong linear characteristics, reducing the benefit of nonlinear kernel functions.

8.3 Implications for Industrial Applications

From an industrial perspective, the results suggest that machine learning-based quality prediction systems can be effectively integrated into steel production pipelines. Linear Regression models, in particular, offer a balance between predictive accuracy and interpretability, making them suitable for real-time decision support systems.

Random Forest models, while slightly more complex, provide resilience against noise and process variability, which is valuable in real-world production environments. These models could be used in offline analysis or as secondary validation tools to support operational decisions.

8.4 Model Interpretability and Explainability

An important consideration in industrial machine learning applications is model interpretability. Linear Regression offers clear insights into the influence of individual input features through regression coefficients. Decision Tree-based models also provide partial interpretability through hierarchical decision rules.

While Random Forest models are less interpretable than linear models, feature importance measures can still be extracted to identify key process parameters influencing steel quality. This interpretability is crucial for gaining trust from domain experts and facilitating adoption in industrial settings.

9. Limitations

Despite the promising results obtained in this study, several limitations should be acknowledged.

First, the evaluation strategy relied on a single predefined train–test split. While this approach provides an unbiased estimate of performance on unseen data, it does not fully capture the variability that may arise from different data partitions. As a result, the reported metrics may be sensitive to the specific split used. Employing cross-validation techniques would allow a more statistically robust assessment of model performance.

Second, the dataset used in this project represents a specific steel production process under controlled conditions. Industrial processes can vary significantly across production lines, facilities, and operating regimes. Therefore, the trained models may not generalize directly to other steel plants or production setups without retraining or adaptation.

Third, hyperparameter tuning was intentionally limited in order to maintain fairness and comparability between models. While this approach is suitable for methodological comparison, it may not yield the optimal performance achievable by each model. More extensive tuning could potentially improve prediction accuracy, particularly for nonlinear models such as SVR and Random Forest.

Finally, although the dataset was normalized and free from missing values, the analysis did not explicitly account for measurement noise, sensor drift, or data imbalance that commonly occur in real-world industrial environments. These factors could influence model stability and performance when deployed in practice.

10. Conclusion

This project investigated the application of supervised machine learning regression techniques for predicting a continuous steel quality factor using normalized industrial process data. A comprehensive experimental study was conducted, involving multiple regression models representing different learning paradigms, including linear, kernel-based, instance-based, tree-based, and ensemble approaches.

The results demonstrate that machine learning models are capable of capturing meaningful relationships between process parameters and steel quality. Among the evaluated models, Linear Regression achieved the highest R^2 score, indicating the presence of strong linear dependencies within the dataset. Random Forest Regressor also performed competitively, highlighting the benefit of ensemble learning in modelling nonlinear feature interactions and improving robustness.

An important outcome of this study is the observation that increased model complexity does not necessarily lead to superior performance. The strong results obtained by a simple linear model emphasize the importance of data understanding and baseline evaluation before adopting more complex techniques. This insight is particularly relevant for industrial applications, where interpretability, computational efficiency, and ease of deployment are critical considerations.

Overall, this work confirms the effectiveness of regression-based machine learning methods for industrial quality prediction tasks. The findings provide a solid foundation for further research and practical deployment of data-driven quality monitoring systems in steel production environments.

11. Future Work

Several promising directions can be explored to extend this work and improve predictive performance.

11.1 Hyperparameter Optimization

In this project, default hyperparameters were used to ensure fair comparison between models. Future work could involve systematic hyperparameter optimization using techniques such as Grid Search or Random Search. Optimizing parameters such as tree depth, number of estimators, kernel parameters, and neighbourhood size may lead to improved performance.

11.2 Cross-Validation and Statistical Robustness

The evaluation was conducted using a single train-test split. To obtain more reliable performance estimates, k-fold cross-validation can be employed. Cross-validation would reduce variance in evaluation metrics and provide a more robust assessment of model generalization capabilities.

11.3 Feature Engineering and Selection

Future studies could explore feature engineering techniques such as polynomial feature expansion, interaction terms, or dimensionality reduction methods like Principal Component Analysis (PCA). Feature selection techniques could also be applied to identify the most influential process parameters and reduce model complexity.

11.4 Advanced Machine Learning Models

More advanced machine learning models could be investigated, including multilayer perceptron (MLPs), deep neural networks, and recurrent architectures. These models may capture complex nonlinear relationships that are not fully addressed by classical regression techniques.

11.5 Real-Time Deployment and Industrial Integration

A significant extension of this work would be the integration of predictive models into real-time steel production systems. This would involve deploying trained models within control architectures, enabling online quality prediction, anomaly detection, and proactive process optimization.

12. References

Course material and project guidelines provided for Applied Machine & Deep Learning (190.015).