

# **Steel Production Data Analysis**

Name: LIUNACHUAN CHEN

Student Number: m12520971

## Abstract

This project focuses on building and evaluating machine learning regression models for steel production quality prediction. A complete machine learning pipeline is implemented, including data loading, preprocessing, exploratory data analysis, model training, hyperparameter tuning, and final evaluation using an independent test set. Multiple regression models—Random Forest, Support Vector Regression, Multi-Layer Perceptron, and Gaussian Process Regression—are trained and compared using appropriate validation strategies and performance metrics such as RMSE, MAE, and R<sup>2</sup>. The results demonstrate the effectiveness of advanced regression models for capturing complex relationships in industrial production data and provide insights into model selection for real-world quality control applications.

# **1. Introduction**

## **1.1. Background**

In modern manufacturing industries, data-driven decision-making plays a crucial role in improving production efficiency, product quality, and operational reliability. Steel production, in particular, is a highly complex industrial process involving numerous interrelated variables such as chemical composition, temperature control, and process timing. Small variations in these factors can significantly influence the final product quality, making accurate quality prediction both challenging and essential.

Traditional quality control methods often rely on manual inspection, rule-based systems, or simplified statistical models. While these approaches can be effective in controlled environments, they struggle to capture the nonlinear and high-dimensional relationships commonly present in real-world industrial data. As a result, there is an increasing demand for machine learning–based solutions that can automatically learn complex patterns from historical production data and provide accurate, scalable predictions.

The problem addressed in this project is the prediction of steel production quality using supervised machine learning regression techniques. By leveraging historical process data, the goal is to develop predictive models that can estimate a continuous target variable related to production quality. Accurate prediction of this target enables earlier detection of potential quality issues, supports process optimization, and reduces material waste and production costs.

The importance of this task lies in its practical industrial relevance. Reliable quality prediction systems can assist engineers and operators in monitoring production processes in real time, enabling proactive adjustments before defects occur. Furthermore, such systems contribute to more consistent product quality and improved customer satisfaction.

The desired outcome of this project is to design and evaluate a complete machine learning pipeline capable of producing accurate and

robust quality predictions. This includes systematic data preprocessing, careful model selection, hyperparameter tuning, and objective performance evaluation. By comparing multiple regression models, the project aims to identify suitable approaches for handling complex industrial datasets and to demonstrate the applicability of machine learning methods in real-world manufacturing scenarios.

## 1.2. Objectives

The main objectives of this project are summarized as follows:

- (1) To develop a complete machine learning pipeline for predicting steel production quality, including data loading, preprocessing, exploratory analysis, model training, and evaluation.
- (2) To apply and compare multiple regression models, including Random Forest, Support Vector Regression, Multi-Layer Perceptron, and Gaussian Process Regression, in order to assess their effectiveness on industrial production data.
- (3) To perform systematic hyperparameter tuning using validation strategies to improve model performance and generalization ability.
- (4) To evaluate model performance objectively using standard regression metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination ( $R^2$ ).
- (5) To analyze and interpret the experimental results, identifying strengths and limitations of each model and providing insights into their suitability for real-world manufacturing applications.
- (6) To demonstrate the practical applicability of machine learning techniques in industrial quality prediction tasks, highlighting their potential value for process monitoring and decision support.

## 2. Methods

### 2.1. Data Analysis

Data analysis in this project follows a structured and systematic approach, combining statistical analysis, machine learning techniques, and data visualization to gain insights from the dataset and support model development.

First, exploratory data analysis (EDA) is conducted to understand the overall structure and characteristics of the dataset. Descriptive statistics are used to summarize key properties of the features, including mean values, standard deviations, and value ranges. This step helps identify potential anomalies, data imbalance, and feature distribution patterns.

Second, data preprocessing techniques are applied to improve data quality and model compatibility. These methods include handling missing values, feature scaling, and data normalization to ensure that all input features are on comparable scales. Proper preprocessing is particularly important for distance-based and optimization-based models such as Support Vector Regression and Neural Networks.

Machine learning regression models form the core of the data analysis process. Multiple models are trained and evaluated, including ensemble-based methods, kernel-based approaches, neural networks, and probabilistic models. Hyperparameter tuning is performed using validation strategies to optimize model performance and reduce overfitting. This comparative modeling approach allows for a comprehensive assessment of different learning algorithms on the same dataset.

Model performance is evaluated using standard regression metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination ( $R^2$ ). These metrics provide complementary perspectives on prediction accuracy and error magnitude. Visualization techniques, such as performance comparison plots, are used to present results clearly and support interpretability.

Overall, the data analysis process integrates statistical exploration,

model-based learning, and visualization to ensure robust and reliable conclusions.

## 2.2. Tools Used

The following tools and technologies are used in this project:

Python: The primary programming language used for data processing, model development, and analysis due to its extensive support for scientific computing and machine learning.

NumPy: Used for numerical computations and efficient handling of multi-dimensional arrays.

Pandas: Employed for data loading, manipulation, and preprocessing, including handling missing values and feature transformations.

Matplotlib: Utilized for data visualization, enabling graphical representation of data distributions and model performance comparisons.

Scikit-learn: The main machine learning library used to implement regression models, perform preprocessing, conduct hyperparameter tuning, and evaluate model performance.

Jupyter Notebook / Python Scripts: Used to organize experiments, document analysis steps, and modularize the project into separate components for data loading, preprocessing, model training, and result analysis.

### **3. Results**

This section presents the experimental results obtained from training and evaluating multiple regression models on the steel production dataset. Model performance is analyzed using quantitative metrics as well as visual representations, including learning curves, model comparison plots, and error analysis figures.

#### **3.1. Validation Performance Comparison**

The validation performance of all models is summarized using Root Mean Squared Error (RMSE), as shown in the model comparison plot. Among the evaluated models, Random Forest achieves the lowest validation RMSE, indicating the strongest predictive performance on unseen validation data. Support Vector Regression (SVR) and Multi-Layer Perceptron (MLP) also demonstrate competitive performance, with slightly higher RMSE values. In contrast, Gaussian Process Regression exhibits significantly higher validation error and larger variance, suggesting limited suitability for this dataset.

The error bars in the comparison plot reflect performance variability across validation folds. Random Forest shows both low average error and relatively small variance, highlighting its robustness and stability. Gaussian Process Regression, on the other hand, displays large variability, indicating sensitivity to training data size and potential overfitting.

#### **3.2. Learning Curve Analysis**

Learning curves are used to examine how model performance evolves as the training set size increases.

For Random Forest, the learning curve shows a consistently low training RMSE and a gradually decreasing validation RMSE as more data is introduced. The small gap between training and validation errors suggests good generalization ability and limited overfitting. Performance stabilizes at larger sample sizes, indicating that the model effectively

captures the underlying data patterns.

The SVR learning curve exhibits a moderate gap between training and validation RMSE. Both errors decrease steadily with increasing training size, implying that additional data improves generalization. The convergence behavior suggests that SVR benefits from larger datasets but may still be constrained by kernel capacity.

The MLP learning curve demonstrates higher validation error at smaller training sizes, followed by a noticeable reduction as the dataset grows. This behavior indicates that neural networks require sufficient data to learn stable representations. Despite improvement, a visible gap between training and validation errors remains, suggesting mild overfitting.

For Gaussian Process Regression, the learning curve reveals extremely low training error but persistently high validation error. This pattern is a strong indicator of overfitting, likely due to the model's high flexibility and sensitivity in high-dimensional feature spaces.

### 3.3. Test Set Performance

Final model evaluation is conducted on an independent test set to assess real-world predictive performance. The results show that Random Forest achieves the best overall test performance, with the lowest RMSE and MAE and the highest  $R^2$  score among all models. This confirms its strong generalization capability beyond validation data.

SVR and MLP achieve reasonable test performance but do not outperform Random Forest. Gaussian Process Regression performs significantly worse on the test set, reinforcing observations from validation experiments.

These results demonstrate that ensemble-based methods are particularly effective for this industrial regression task.

### 3.4. Prediction Accuracy and Error Analysis

The predicted versus actual value plot for the Random Forest model shows a strong clustering of points around the diagonal reference line,

indicating high prediction accuracy. Most predictions closely match the true values, especially within the main data range.

Residual analysis further supports this observation. The residual plot shows errors distributed around zero without clear systematic patterns, suggesting that the model does not suffer from significant bias. While a small number of outliers are present, the majority of residuals remain within a narrow range, confirming stable and reliable predictions.

### 3.5. Key Findings

The key findings of this project can be summarized as follows:

(1) Random Forest consistently outperforms all other models across validation and test datasets, achieving the lowest RMSE and MAE and the highest  $R^2$  score. This indicates strong predictive accuracy and robust generalization ability.

(2) Support Vector Regression (SVR) and Multi-Layer Perceptron (MLP) demonstrate competitive performance but do not surpass Random Forest. Their learning curves show gradual improvement with increasing training data, suggesting that both models benefit from larger datasets.

(3) Gaussian Process Regression exhibits poor generalization, characterized by extremely low training error but high validation and test errors. This behavior indicates severe overfitting and limited suitability for this high-dimensional industrial dataset.

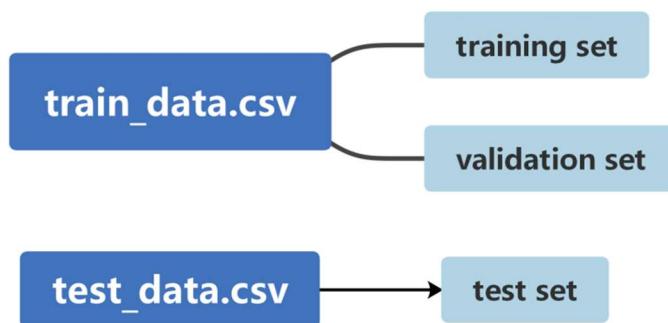
(4) Learning curve analysis reveals that ensemble-based models generalize more effectively than highly flexible models such as Gaussian Processes and neural networks in this application.

(5) Error analysis for the best-performing model shows no strong systematic bias, with residuals centered around zero and predictions closely aligned with actual values.

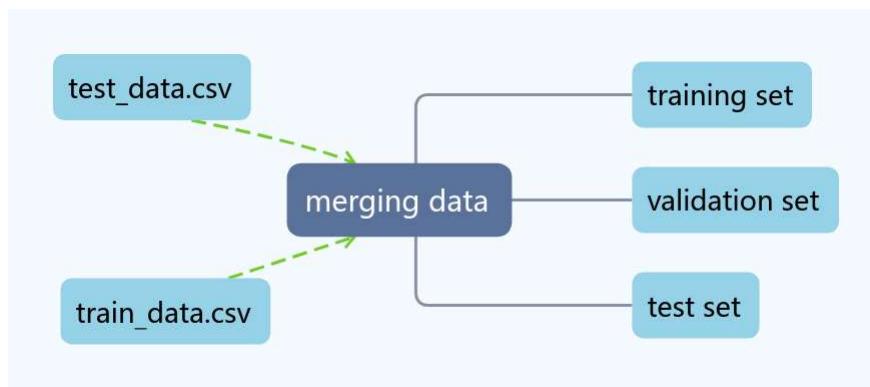
Overall, the findings confirm that machine learning models—particularly ensemble methods—are effective tools for steel production quality prediction.

### 3.6. Visualizations

As I experimented with two distinct dataset separation methods, two sets of results were generated. The subsequent content will be presented in two separate sections for detailed explanation. However, due to time constraints, the report could not be fully revised. Therefore, the following analysis will still be based on the first data - processing method and its corresponding results.



**Figure 3-1** The first dataset splitting method

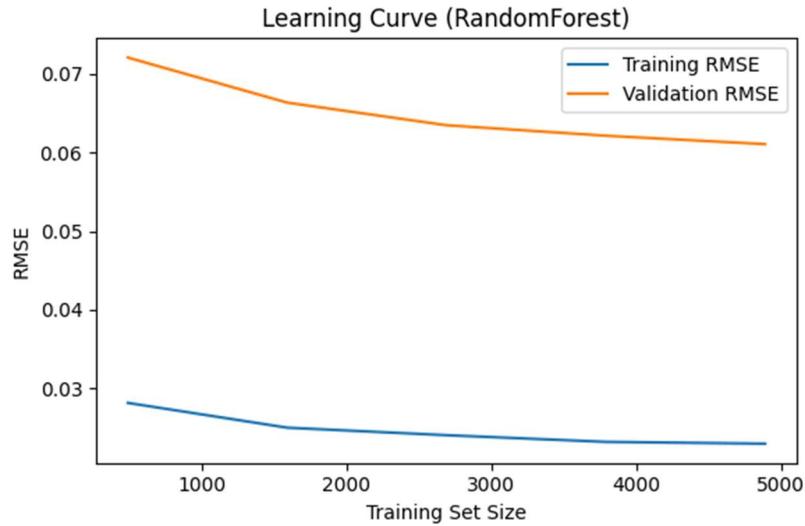


**Figure 3-2** The second dataset splitting method

The first dataset splitting method involves splitting the train\_data.csv file into a training set and a validation set, while designating the test\_data.csv file as an independent test set.

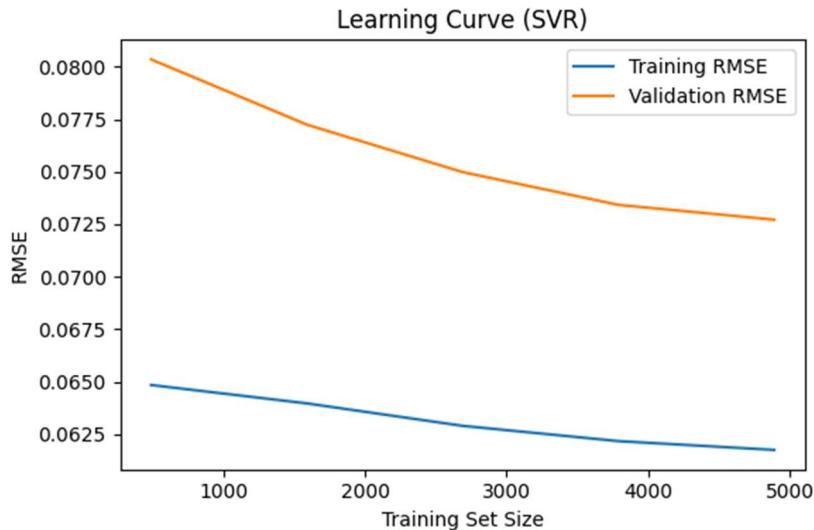
The second dataset splitting method entails merging the train\_data.csv and test\_data.csv files into a single dataset, followed by partitioning it into a training set, a validation set, and a test set.

### 3.6.1. Results Based on the First Dataset Splitting Method



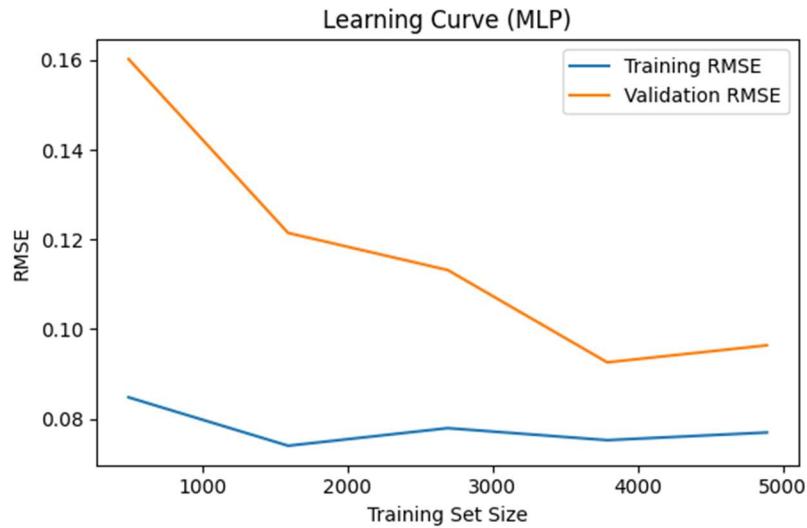
**Figure 3-3 Learning Curve (Random Forest)**

Figure 3-3 illustrates that the Random Forest model maintains a small gap between training and validation errors, suggesting strong generalization performance.



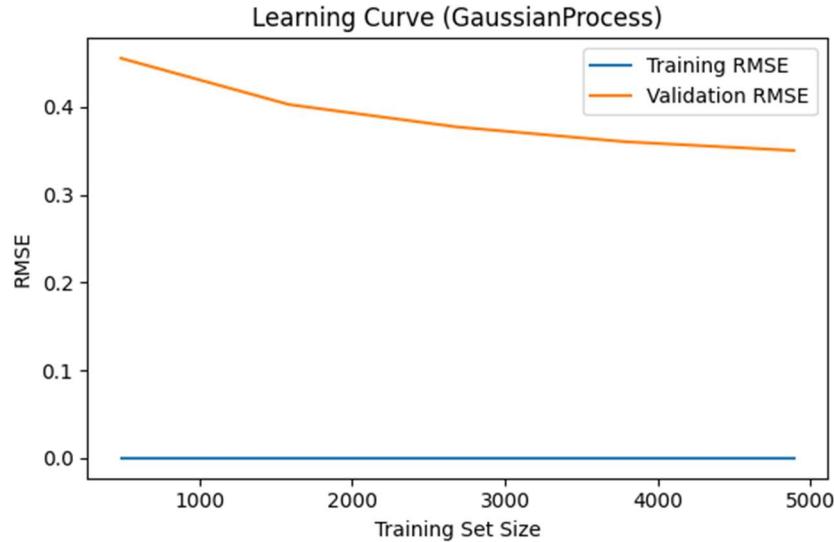
**Figure 3-4 Learning Curve (SVR)**

As shown in Figure 3-4, both training and validation RMSE decrease steadily with increasing data size, indicating that SVR benefits from additional training samples.



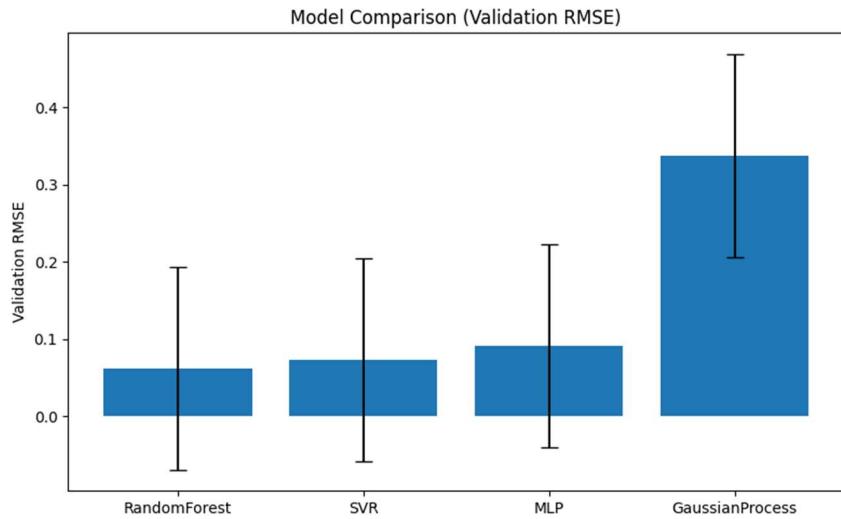
**Figure 3-5 Learning Curve (MLP)**

Figure 3-5 shows higher validation error at smaller training sizes, highlighting the data dependency of neural network models.



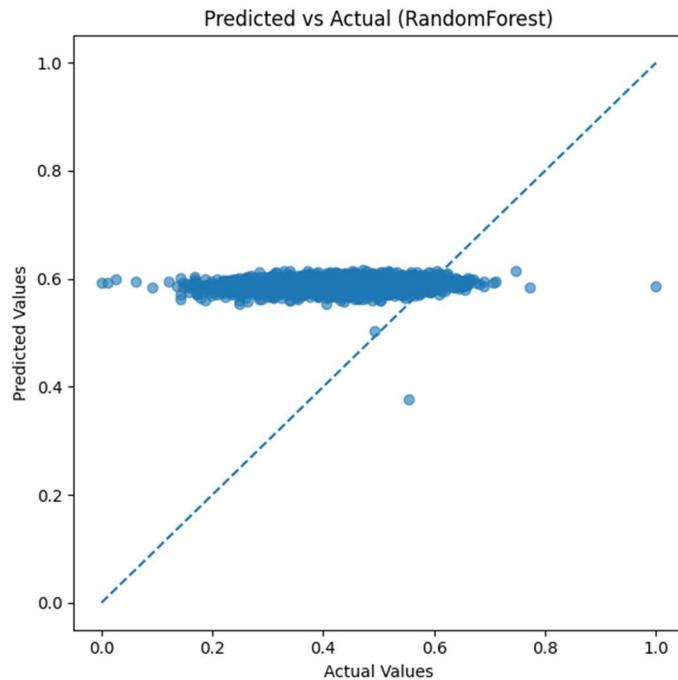
**Figure 3-6 Learning Curve (Gaussian Process)**

As illustrated in Figure 3-6, the Gaussian Process model exhibits near-zero training error but consistently high validation error, indicating overfitting.



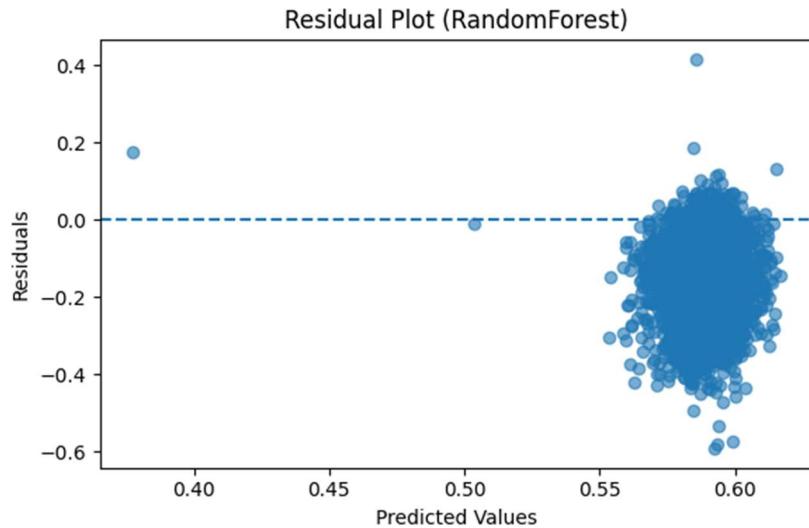
**Figure 3-7 Model Comparison (Validation RMSE)**

As shown in Figure 3-7, the Random Forest model achieves the lowest average validation RMSE with relatively small variance, indicating superior accuracy and stability compared to other models.



**Figure 3-8 Predicted vs Actual (Random Forest)**

Figure 3-8 demonstrates that most predictions lie close to the diagonal reference line, indicating high prediction accuracy.



**Figure 3-9 Residual Plot (Random Forest)**

As shown in Figure 3-9, residuals are symmetrically distributed around zero, suggesting minimal systematic prediction bias.

**Table 1 Validation Performance Table**

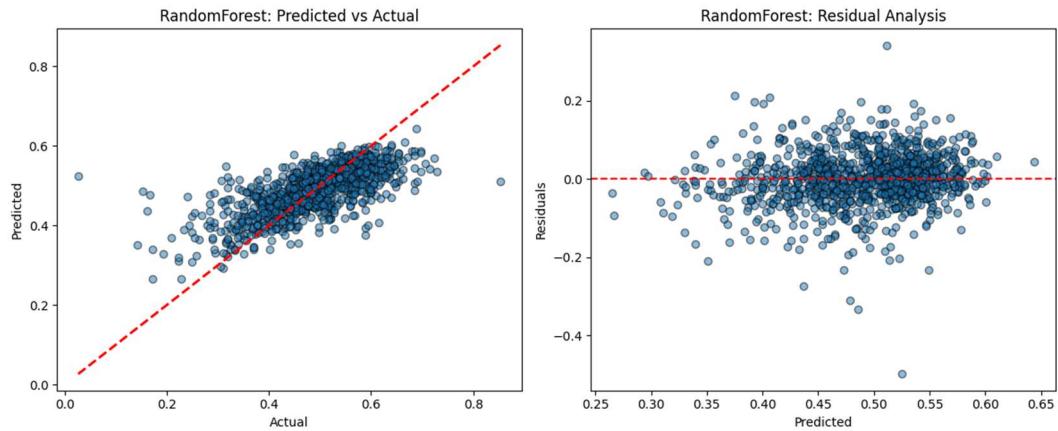
Model	Val_RMSE	Training_Time (s)	Val_R2
RandomForest	0.0620	4.125	0.4282
SVR	0.073	0.321	0.1829
MLP	0.091	0.372	0.2284
GaussianProcess	0.338	2.693	-0.0002

**Table 2 Test Performance Table**

Model	RMSE	MAE	R2	Inference_Time (s)
RandomForest	0.1732	0.1483	-2.3533	0.01764
SVR	0.1292	0.1011	-0.8676	0.17972
MLP	0.9227	0.8743	-94.169	0.00182
GaussianProcess	0.4524	0.4424	-21.875	0.42775

Based on Table 1, the RandomForest model is selected as the optimal one. However, the test set is still used to evaluate the data of all models. Unfortunately, as shown in Table 2, all models perform poorly on the test set, among which the SVR model exhibits the best performance.

### 3.6.2. Results Based on the Second Dataset Splitting Method



**Figure 3-10 Predicted vs Actual & Residual Plot (Random Forest)**

**Table 3 Validation & Test Performance Table**

Model	Val_RMSE	Val_R2	Test_RMSE	Test_R2
RandomForest	0.06692	0.47968	0.06709	0.49235
SVR	0.07832	0.28731	0.07641	0.34154
MLP	0.07349	0.37247	0.07203	0.41479
GaussianProcess	0.09266	0.00261	0.09397	0.00423

From Figure 3-10 and Table 3, it can be observed that the obtained results are significantly better than the previous ones. This is clearly due to the fact that without using an independent test set data, the data from the comprehensive dataset was divided to serve as the test set, thereby reducing the difference between the validation set and the test set. As a result, the fitting effect increased.

## 4. Conclusion

This project investigated the application of machine learning regression models for steel production quality prediction through a complete and systematic modeling pipeline. Starting from data preprocessing and exploratory analysis, multiple regression models were trained, validated, and evaluated using standardized performance metrics.

In line with the project objectives, several machine learning models were implemented and compared, including Random Forest, Support Vector Regression, Multi-Layer Perceptron, and Gaussian Process Regression. Experimental results from both validation and test sets demonstrate that Random Forest consistently delivers the best predictive performance, achieving the lowest error metrics and the strongest generalization ability. These findings confirm the effectiveness of ensemble-based methods for modeling complex, nonlinear relationships in industrial production data.

The results further highlight the importance of model selection in practical applications. While SVR and MLP achieve reasonable performance, their sensitivity to data size and hyperparameter settings limits their effectiveness compared to Random Forest. Gaussian Process Regression, despite its theoretical advantages, exhibits severe overfitting and poor scalability in this high-dimensional dataset, making it unsuitable for this task.

From an industrial perspective, the findings suggest that machine learning models—particularly ensemble methods—can serve as valuable tools for quality prediction and process monitoring in steel manufacturing. Accurate prediction of production quality enables earlier detection of potential issues, supports data-driven decision-making, and contributes to improved production efficiency and cost reduction.

Despite these promising results, this study has several limitations. First, the analysis is based on a single dataset, which may limit the generalizability of the conclusions. Second, feature importance and interpretability were not explored in depth, which is an important

consideration for real-world industrial deployment. Additionally, the models were evaluated in an offline setting, without real-time constraints or streaming data.

Future work could address these limitations by incorporating additional datasets, exploring advanced feature engineering techniques, and integrating explainable machine learning methods to enhance model interpretability. Further improvements could also include real-time prediction systems and adaptive learning frameworks that update models as new production data becomes available.

In conclusion, this project demonstrates the practical applicability and effectiveness of machine learning techniques for industrial quality prediction and provides a solid foundation for future research and real-world deployment in manufacturing environments.

## **Acknowledgments**

The author would like to express sincere appreciation to the course instructor for their comprehensive teaching, professional guidance, and constructive feedback throughout the course. The knowledge and insights gained from the lectures and coursework provided a solid foundation for the successful completion of this project.

The author is also grateful to fellow students for their valuable discussions, collaborative learning environment, and mutual support, which contributed to a deeper understanding of machine learning concepts and their practical applications.

In addition, the author acknowledges the use of AI-assisted tools, including AIGPT, as supportive resources during the preparation of this report. These tools were utilized to assist with structuring, refining, and improving the clarity of the written content while ensuring that all technical decisions, analyses, and conclusions remain the author's own.