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 Quality Prediction Using Machine Learning

Student: Rishabh Kothari

Matrikulation Number: 12519563

Institution: Montanuniversität Leoben

1. Abstract

Steel production is a quite complex industrial process, which is related to a large number of interacting physical, chemical, and thermal factors. Quality stability is a critical issue in steel production because small fluctuations in material quality can easily cause expected and unexpected failures. Conventional quality control systems are based on rule base techniques and physical models, which are not efficient in handling the nonlinear relationships involved in real steel production activities.

The current work examines the possibility of applying supervised machine learning regression algorithms for predicting a continuous steel quality variable based on normalized industrial data. Several regression models have been created, trained, and compared for their performance as regression predictors. The goal would be to determine the best-performing regression model for industrial applications and understand the strengths and weaknesses associated with each regression algorithm. The problem can certainly help in understanding how machine learning algorithms can successfully simulate steel production and contribute towards informed industrial decision-making.

2. Introduction

2.1 Background

Steel production is one of the most prominent industries in modern manufacturing. It is the backbone for several major industries like construction, transport, automobile manufacturing, power generation, heavy machinery, and infrastructure development. Steel items directly influence the strength, functional integrity, durability, and performance of engineering systems. Any anomaly in steel can lead to disastrous failures, increased maintenance costs, or losses.

The steel production process is quite complex and comprises various steps such as melting, refining, casting, rolling, and cooling. It is affected by a number of parameter interactions such as material composition, furnace temperature, pressures, cooling rate, and material treatment chemicals. The parameter interactions are non-linear in nature, making it difficult to model the process effectively through physical approaches.

Conventional steel quality prediction requires rule-based control systems and the domain-specific knowledge of experts. Though it can be very helpful, sometimes it lacks adaptability with respect to the conditions being employed during production and lacks efficiency while attempting to analyze complex patterns within the data being presented to it.

With the development in Industry 4.0, industrial settings are being integrated with new technologies like advanced sensors, automation, as well as tools for real-time monitoring. Such technologies produce enormous amounts of data from industrial processes. The need for high-quality data provides opportunities for machine learning to enter the scene. Machine learning algorithms can find hidden patterns in high volumes of data in an abstract manner.

The reason for undertaking this project is the need to assess the applicability of supervised machine learning regression models in making predictions related to the quality of the steel produced based on the process. Such predictions can help in the early detection of defects, control of the manufacturing process in advance, waste reduction, and efficiency enhancement.

2.1 Objectives

Main goals of the project are

1. To understand the structure and features of industrial steel production data.
2. To perform all data preprocessing and normalization.
3. To carry out exploratory data analysis to look for patterns and relationships.
4. To model various learning paradigms via multiple regression models.
5. To evaluate model performance using statistical metrics.
6. To Compare the Accuracy and Robustness of Models.
7. Analyzing interpretability of the model in an industrial context.
8. To make recommendations deployable in the real world.

3. Methods

3.1 Data Acquisition

The dataset used in this project is provided as part of the coursework in the applied machine learning course. This dataset is from an industrial process where steel is produced. The dataset consists of numerical measurements recorded by the control systems during this process.

Every sample in the dataset is the result of a single process and consists of numerical measurements from the process as well as an accompanying output value of its quality.

This dataset consists of two distinct files:

i) Training dataset - 7,642 samples
ii) Test Dataset - 3337 samples

Every dataset consists of 22 columns features from 21 inputs named input1 to input21 and 1 output feature labeled output.

This is a supervised learning task since it is a regression problem where the output variable is a continuous factor for steel quality. The variables were given in a CSV file that I accessed locally.

3.2 Data Analysis

A few preprocessing steps are necessary for the data before the model can learn. Duplicates are removed for quality and consistency within the data set. The values are standardized for equal weight in the learning process. Standardizing is necessary for models that are dependent on distance and/or kernels, for example, K Neighborhood and SV Regression.

The Exploratory Data Analysis was performed using:

i) Histograms for feature distribution analysis.
ii) Box plots to check for outliers.
iii) Correlation Heatmaps to analyze correlation between variables.

Histograms tended to be mostly unimodal with moderate skewness. There were no extreme data points observed. Correlation matrix analysis showed moderate values among some variables and the target variable. But none of the variables were dominant; hence, more than one parameter affects steel quality.

The problem was formulated as a supervised regression task.

Below are some machine learning models that were used:

i) LinearRegression
ii) Support Vector Regression
iii) K Nearest Neighbors
iv) Decision Tree Regressor
v) Random Forest Regression

All the models were trained using normalized training data with their default hyperparameters and were compared fairly. Evaluation was performed on an independent test dataset.

3.3 Tools Used:

1. Python language was used for implementation.

2. Data handling and preprocessing were done using Pandas and NumPy.

3. Visualization was performed with Matplotlib and Seaborn.

4. Model training and evaluation were done with the Scikit-learn library.

5. Experiments were conducted and analyzed with Jupyter Notebook.

6. Git and GitHub were used for version control and project management.

4. Results

4.1 Quantitative Findings:

Model performance was evaluated using

i) Mean Absolute Error
ii) Mean Squared Error
iii) Root Mean Squared Error
iv) R squared score

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

Results summary:

i) Linear Regression achieved the highest R squared score of 0.493
ii) Random Forest achieved 0.390
iii) Decision Tree achieved moderate performance
iv) SVR showed limited improvement
v) KNN produced negative R squared indicating poor performance

4.2 Interpretation:

The strong performance of Linear Regression suggests that the dataset contains strong linear relationships between process parameters and steel quality.

Random Forest performed well due to its ability to capture nonlinear interactions and reduce overfitting.

KNN performed poorly due to high dimensional feature space and the curse of dimensionality.

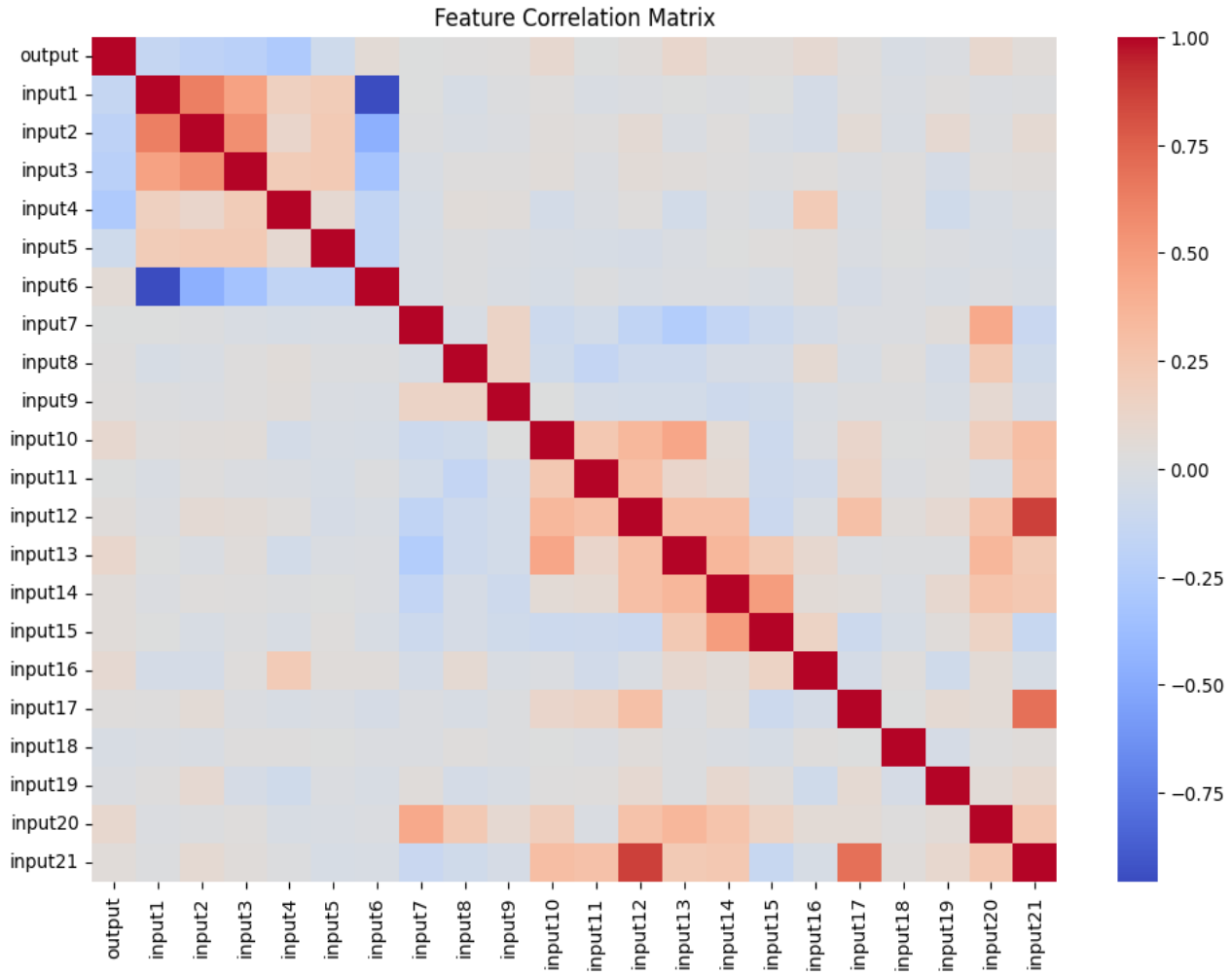
SVR showed limited benefit because the data already exhibits linear characteristics.

4.2 Figures

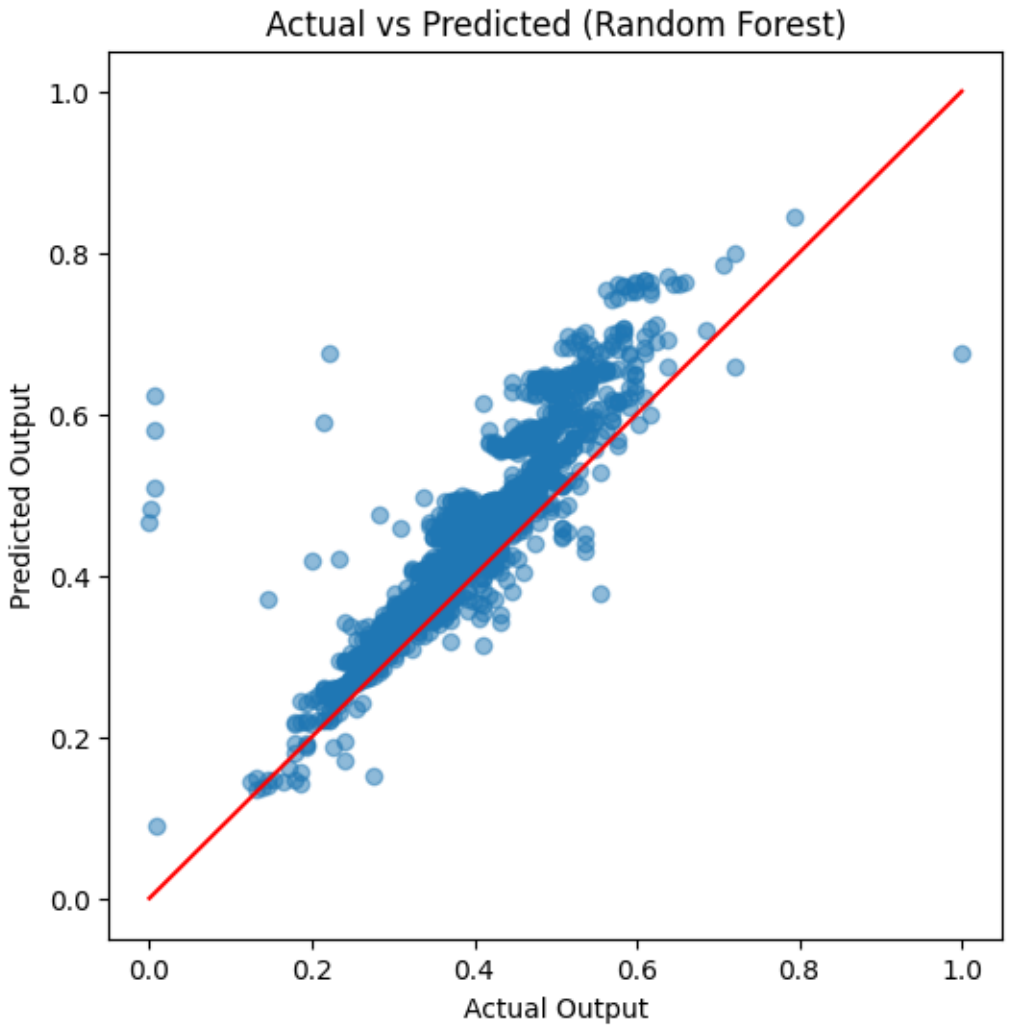
i) Feature correlation matrix.

ii) Scatter plot of actual vs predicted values for Random Forest, showing alignment along the diagonal.

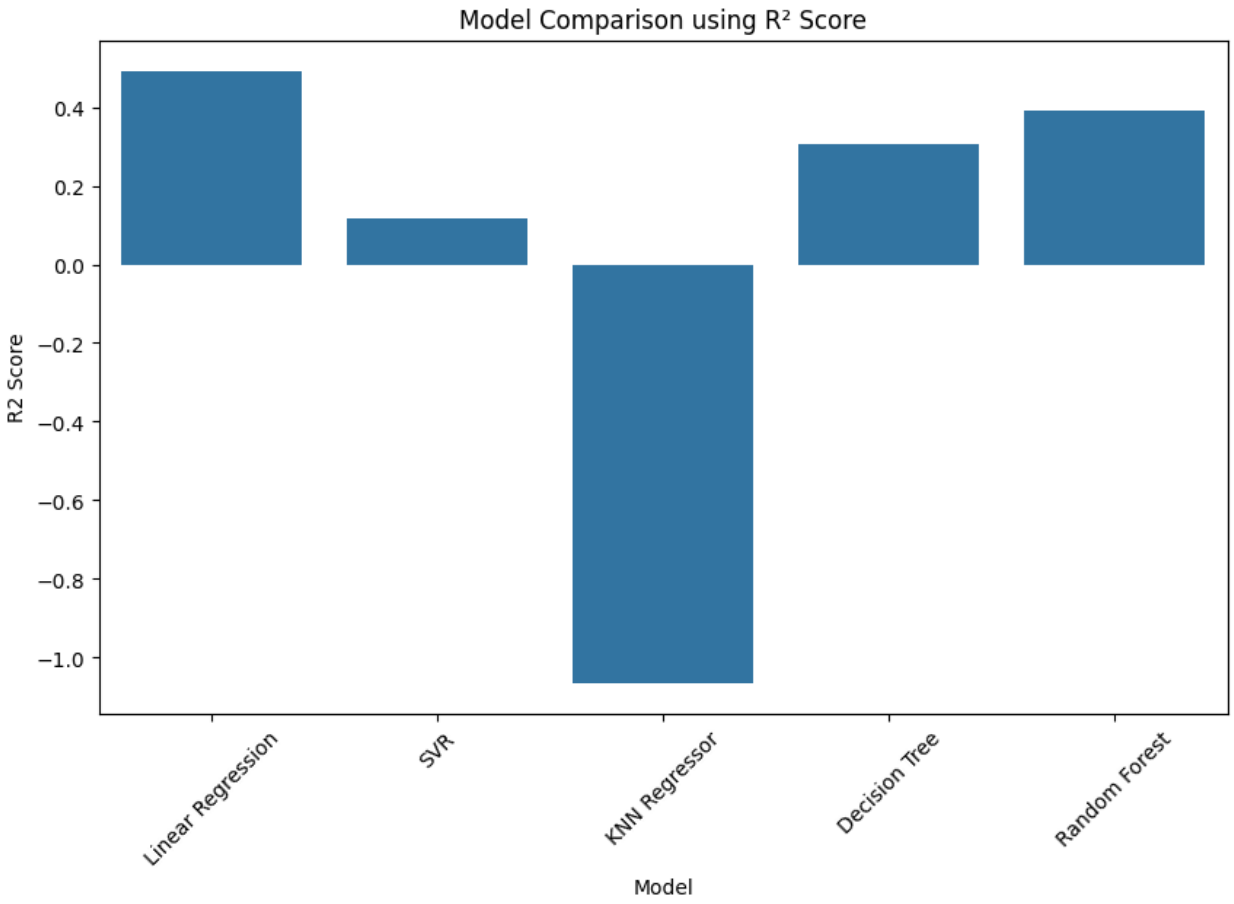
iii) Bar chart to compare model comparison results based on the R-squared statistic.



Feature Correlation Matrix



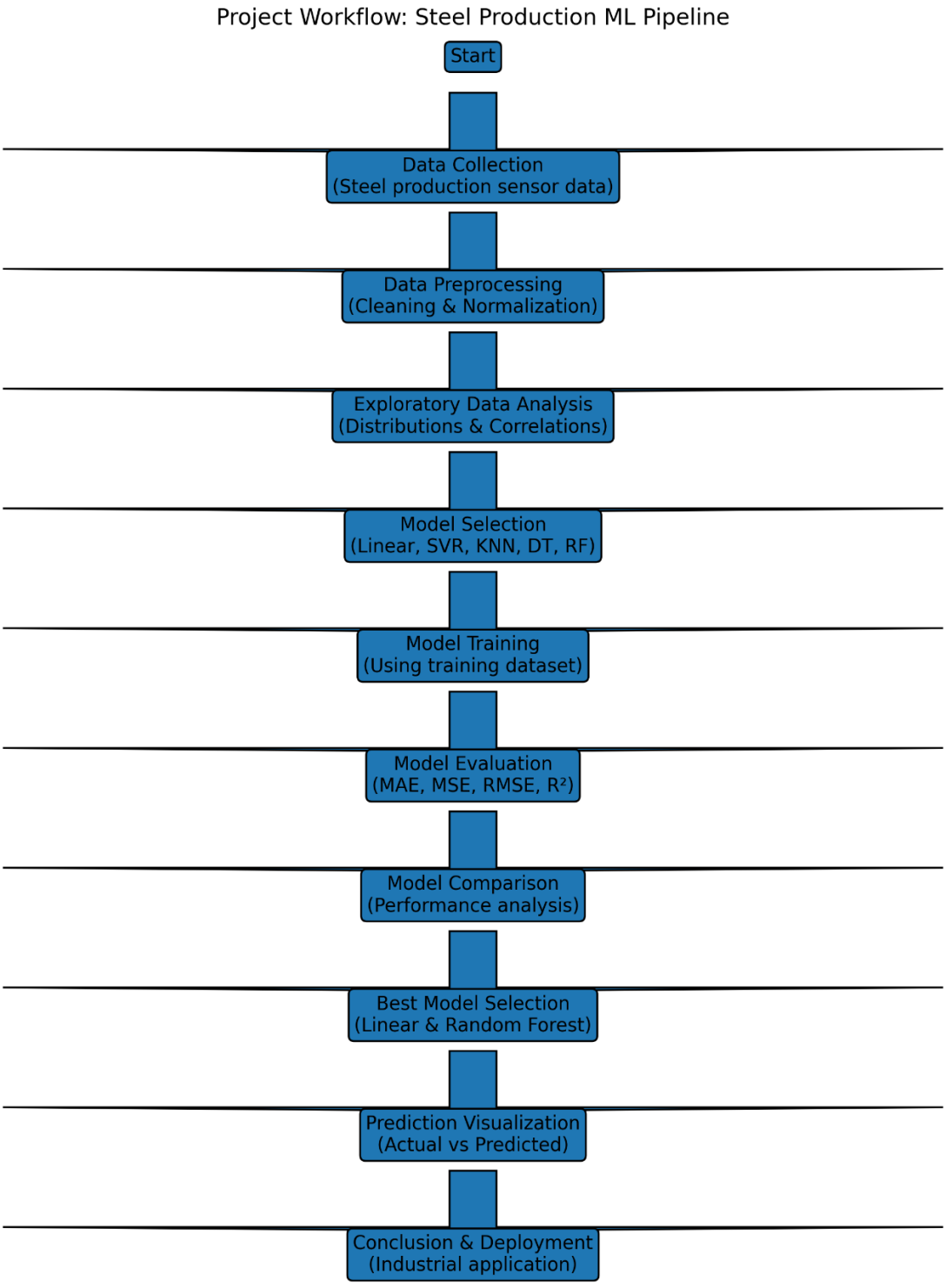
Actual vs Predicted plot for Random Forest



Model Comparison using R^2 score

These plots confirm the numerical findings and support interpretation.

5. Conclusion



End to end workflow of the machine learning pipeline for steel quality prediction

A flowchart of the complete pipeline followed for the project is given above,

Data Collection: Data is collected from sensors on steel production. Here, the raw dataset is preprocessed, cleansed, and normalized in order to make it consistent and reliable. Then, exploratory data analysis should be performed to know the distribution of features and how they relate to each other.

The data understanding provides a scope to select multiple regression models that represent different learning paradigms. The models are trained using a normalized training dataset. The model performance is evaluated on standard regression metrics such as MAE, MSE, RMSE, and R^2 score.

Among the various tested models, comparative analysis identifies the best-performing ones. Actual versus predicted plots, among other prediction visualizations, serve as model behavioral validations. At the end of this entire process, it concludes an industrially applicable and deployable project in a real-time production environment.

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.

6. Acknowledgments

I would like to thank my course instructor for guidance and dataset access.

I acknowledge the use of the following resources:

i) Scikit learn documentation

ii) Matplotlib documentation

iii) Research papers on machine learning in manufacturing

ChatGPT was used for

i) Debugging Python code

ii) Improving data preprocessing logic

iii) Assisting with report writing

iv) Preparing presentation content