Project Report: Machine Failure Prediction

Github Repo: https://github.com/LeLuke007/Machine-Failure-Prediction

1. Introduction

When machines used in industrial and manufacturing processes break down unexpectedly there is a high chance of wasting both time in production and money as well as safety risks. Predictive maintenance also targets to address them by predicting the failures to be experienced in advance and conducting them periodically, thereby minimizing the impending downtime.

This project aims at designing a sound predictive system of the Machine Failure based on sensor data. Using machine learning, the system simplifies both real-time machine operational data analysis, and the prediction of potential failures by analysing operational data of machinery (e.g. temperature, rotational speed, torque). The final app is a web-based solution to make predictions in real-time in the form of a user-friendly application that is practical to industrial operators and maintenance teams.

2. Objectives

The primary goals of this project were:

- To perform a comprehensive **exploratory data analysis (EDA)** to understand the relationships between various sensor readings and machine failures
- To preprocess and prepare the sensor data for training
- To train and evaluate multiple machine learning models to identify the most effective algorithm for failure prediction
- To select the best-performing model and build a predictive system based on it

3. Dataset

The project utilizes the "Machine Failure Predictions" dataset sourced from Kaggle. This dataset contains 10,000 data points with 14 features, simulating sensor readings from machinery.

Features:

- **Type:** The quality variant of the machine (L: Low, M: Medium, H: High)
- **Air temperature [K]:** The ambient air temperature in Kelvin
- **Process temperature [K]:** The temperature of the machine during the process in Kelvin
- **Rotational speed [rpm]:** The speed at which the machine is rotating in revolutions per minute

- **Torque [Nm]:** The torque applied by the machine in Newton-meters
- **Tool wear [min]:** The wear on the tool in minutes

Target Variable:

- **Machine failure:** A binary indicator where 1 signifies a failure and 0 signifies no failure

The dataset also includes specific failure type indicators (TWF, HDF, PWF, OSF, RNF), which were used for analysis but not as the primary prediction target.

4. Methodology

Technologies Used: Python, Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Streamlit

4.1 Data Preprocessing and EDA

The first step focused on the data loading and EDA in order to discover insights. This includes:

- Loading the data
- Visualizing the distribution of numerical features (Shown in Fig. 4.1, Fig. 4.2)
- Analyzing the correlation between different sensor readings (Shown in Fig. 4.3)
- Examining the relationship between each feature and the Machine failure target variable (Shown in Fig. 4.4)
- The categorical feature Type was converted into a numerical format using one-hot encoding

4.2 Model Training and Comparison

Five different machine learning models were trained and evaluated to determine the best fit for the prediction task:

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Gradient Boosting
5. Support Vector Machine (SVM)

The dataset was split into training and validation sets. Each model was trained on the training data and its performance was measured on the unseen validation data to ensure unbiased evaluation

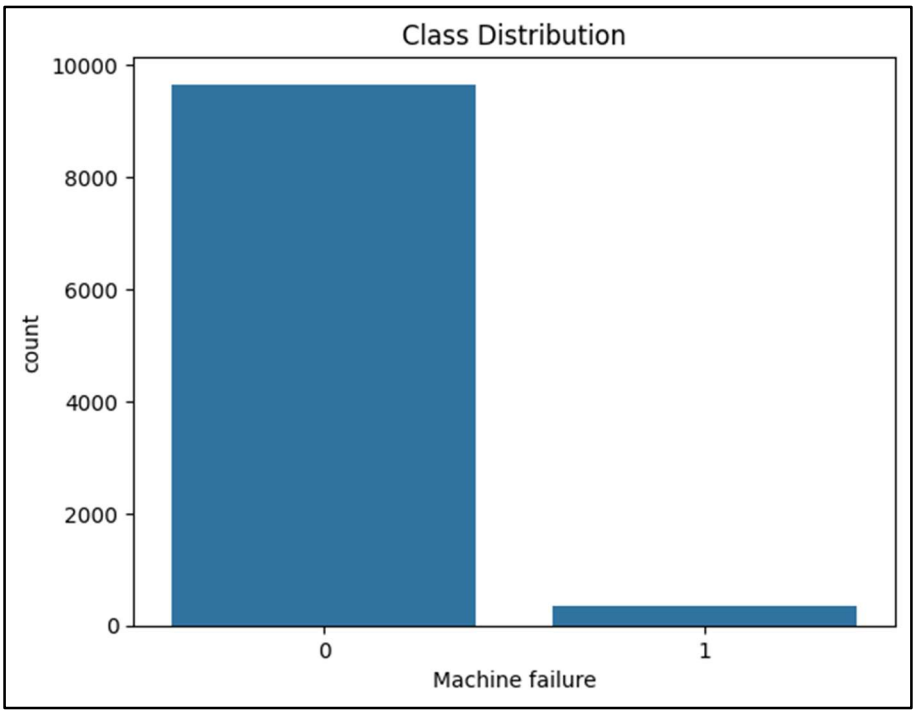

Fig 4.1: Distribution of Target Variable

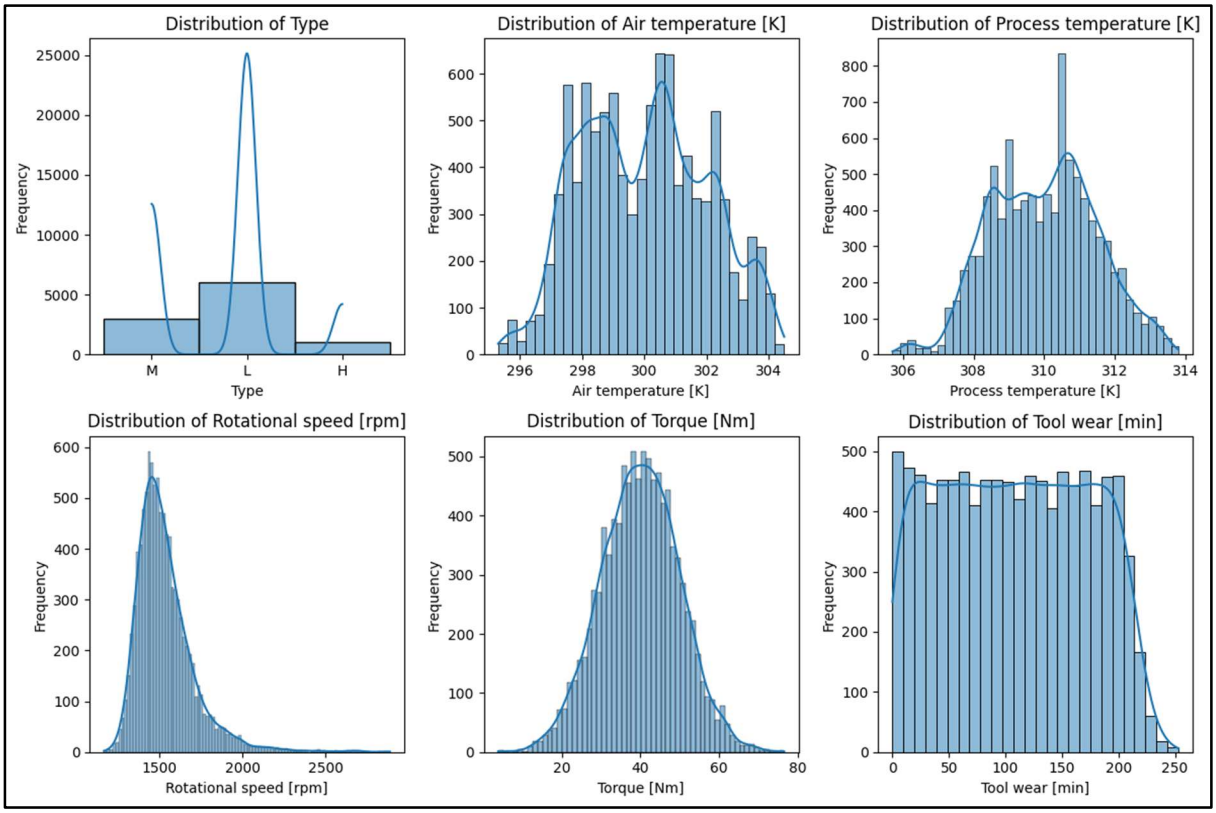

Fig 4.2: Distribution of Features

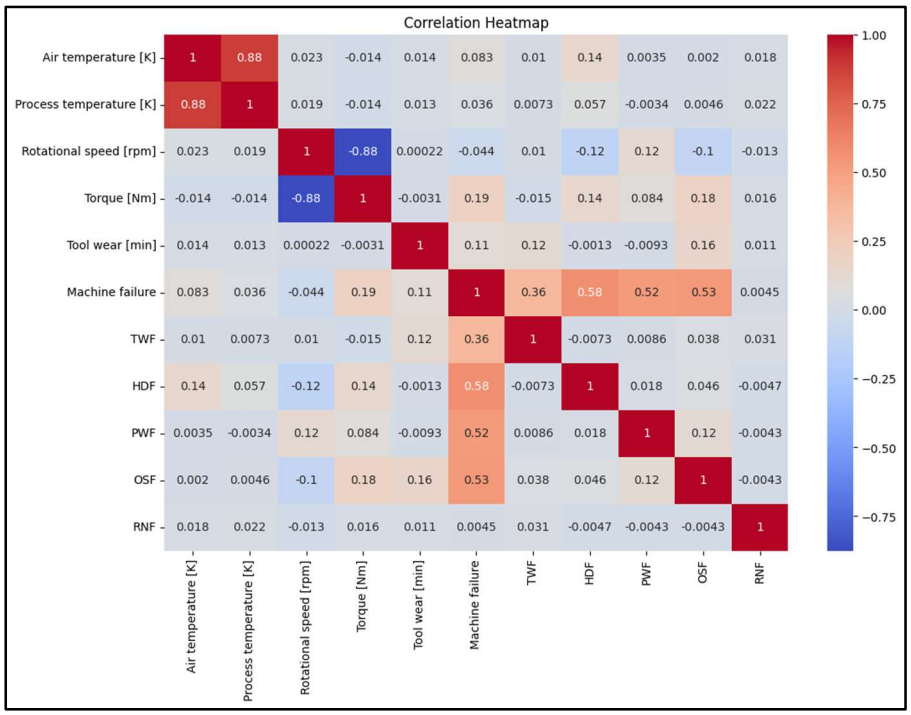

Fig 4.3: Correlation Heatmap between Features

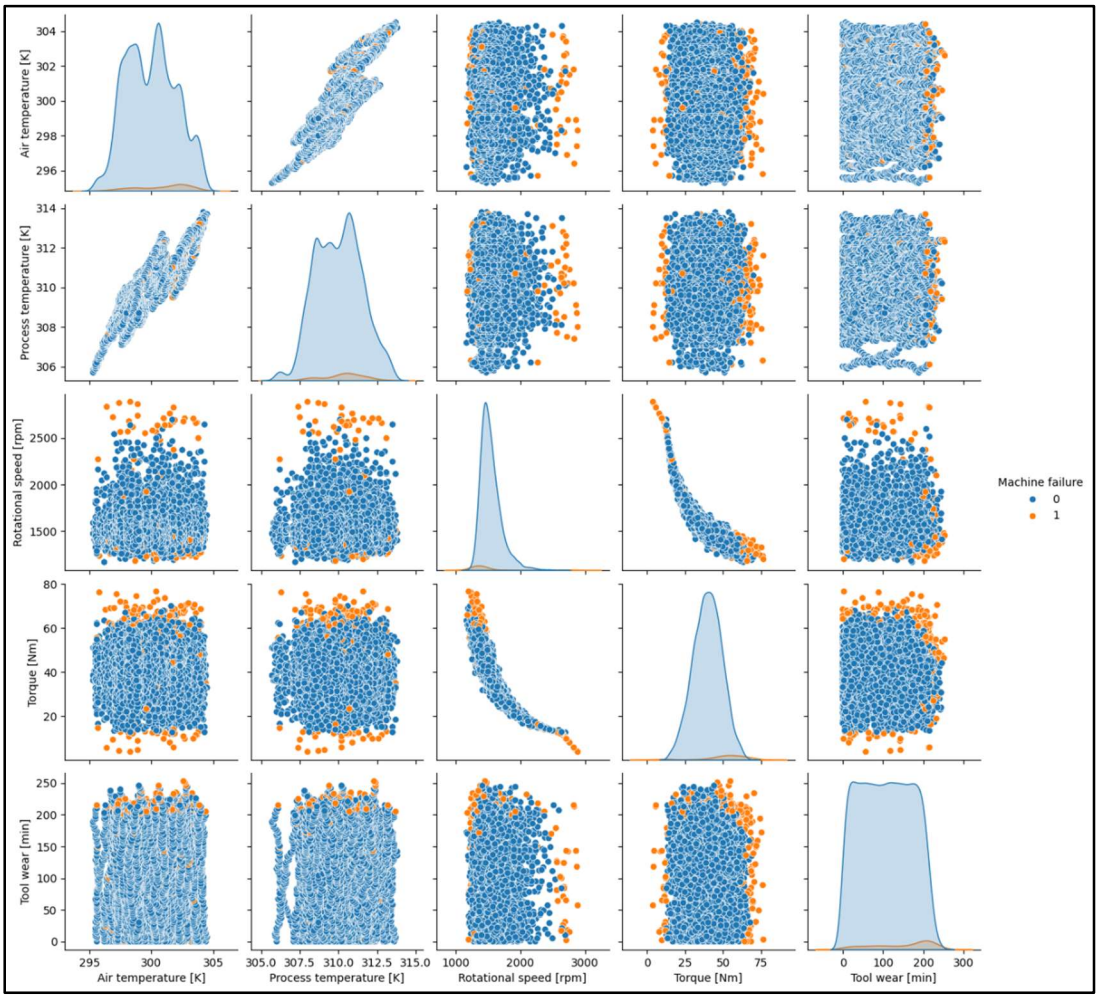

Fig 4.4: PairPlot between Features and Target Variable

5. Results and Model Evaluation

The models were evaluated based on their **validation accuracy**.

Model	Validation Accuracy
Logistic Regression	0.973
Decision Tree	0.9805
Random Forest	0.9835
Gradient Boosting	0.983
Support Vector Machine	0.9770

The **Random Forest** model achieved the highest accuracy of 98.35% and was therefore selected for deployment.

A confusion matrix for the Random Forest model (Shown in Fig. 5.1) provided a deeper look into its performance, showing a very low number of false positives and false negatives, which is crucial for a reliable prediction system in a real-world scenario.

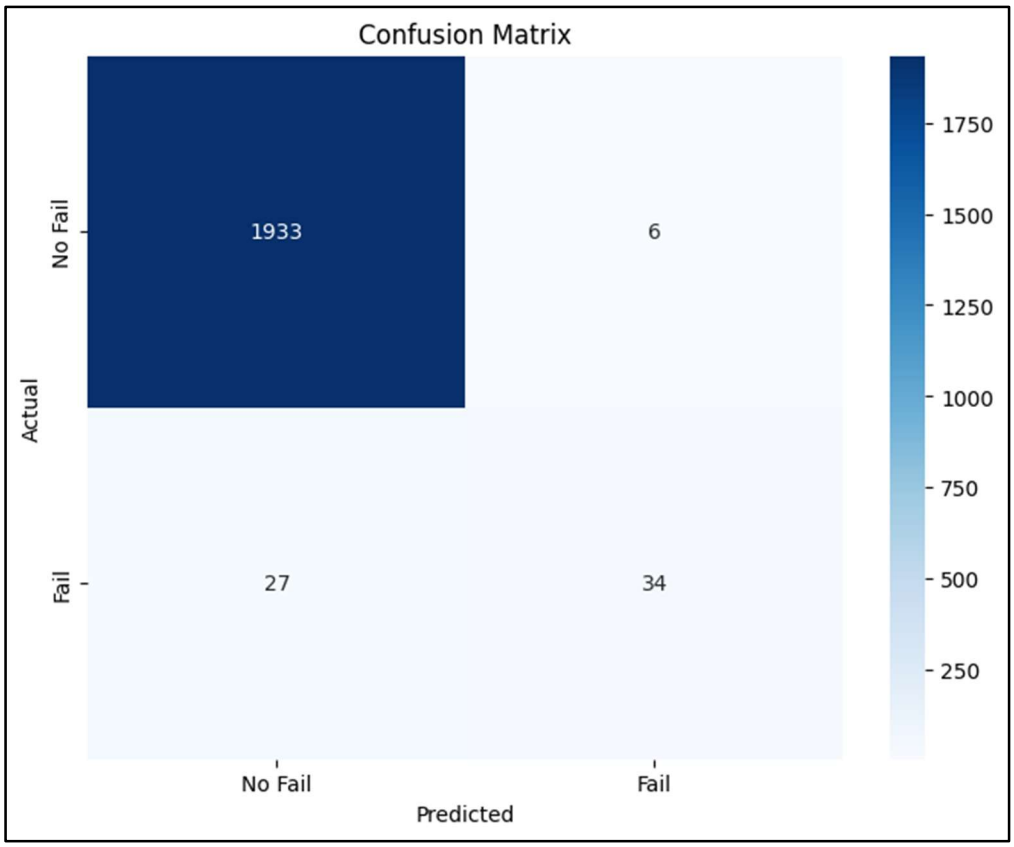

Fig 5.1: Confusion Matrix for Random Forest

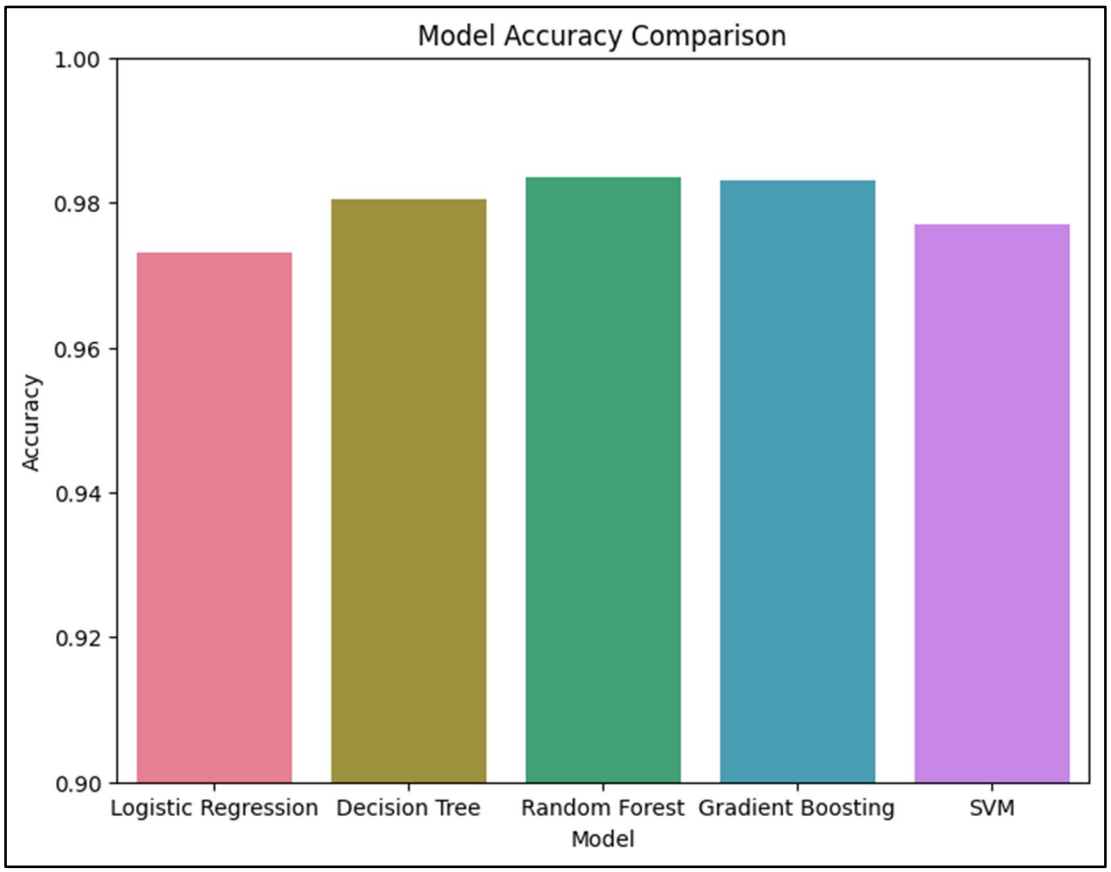

Fig 5.2: Accuracy comparison between different models

6. Conclusion

This project successfully demonstrates the effectiveness of machine learning for predictive maintenance. A highly accurate Machine Failure Prediction system was developed, with the Random Forest model achieving an **98.35% accuracy**.

The final deployed Streamlit application provides an intuitive and accessible way to leverage this powerful model, allowing users to input sensor data and receive instant failure predictions. Such a tool can significantly reduce maintenance costs, prevent unplanned downtime, and improve overall operational efficiency in an industrial setting.

7. References and Appendix

- Dataset Link: https://www.kaggle.com/datasets/shashanknecrothapa/machine-failure-predictions
- Streamlit App: https://machine-failure-prediction-leluke007.streamlit.app/
- https://colab.research.google.com/
- https://machinelearningmastery.com/
- https://www.youtube.com/playlist?list=PLg8h8Ej1e8l2u2Hdt2EIX86SFJpgvUUs3
- https://www.geeksforgeeks.org/