Description of Design Choices and Performance Evaluation of the Model

Problem Overview: In this project, we aimed to build a predictive maintenance model to detect machine anomalies and predict potential failures based on time series data. This model was built to assist industries in preventing costly equipment breakdowns by predicting failures before they occur.

Data Preprocessing and Feature Engineering: The dataset contained over 18,000 rows of time series data, with a binary target variable (y) indicating whether an anomaly occurred (1) or not (0). The steps taken for data preprocessing and feature engineering include:

- Exploratory Data Analysis (EDA): Conducted to understand patterns, relationships, and trends in the data. Missing values and outliers were identified and handled appropriately.
- Handling Missing Values: Imputation or removal of missing values based on the nature of the data.
- Outlier Detection: Used methods like Z-scores or IQR (Interquartile Range) to identify and remove outliers.
- Feature Engineering: Created additional features or transformations to improve the predictive power of the model, such as aggregating or normalizing the data.
- Feature Selection: Selected important features using correlation analysis and feature importance metrics from tree-based models.

Model Selection and Training:

- Model Chosen: The Random Forest Classifier was chosen for this task due to its ability to handle complex relationships in the data and its robustness against overfitting.
- Hyperparameter Tuning: The model's hyperparameters were tuned using GridSearchCV to find the best combination of parameters for better performance.
- Model Training: The dataset was split into training and testing sets (80-20 split) to evaluate the model's performance on unseen data.

Performance Evaluation: The model was evaluated on the test set using the following metrics:

- Accuracy: The accuracy on the test set was 85%.
- Confusion Matrix: The confusion matrix showed a high number of true positives, indicating good performance in anomaly detection.
- Precision, Recall, F1-score: The F1-score was calculated to measure the balance between precision and recall, with a score of 0.82, indicating strong performance.

Discussion of Future Work

While the current model performs well, there are several areas for improvement and further development:

- Algorithm Enhancement: Experimenting with other algorithms like Gradient Boosting Machines (GBM), XGBoost, or neural networks could provide better results. These algorithms might capture more complex patterns in the data.
- Data Augmentation: Introducing more data through simulation or using external data sources could help improve model accuracy.
- Anomaly Detection Algorithms: In addition to classification models, advanced anomaly detection methods like Isolation Forest or One-Class SVM could be explored.
- Time-Series Specific Models: Since the data is time-series based, models like Long Short-Term Memory (LSTM) or other Recurrent Neural Networks (RNNs) might be more suitable for capturing sequential dependencies.
- Real-time Deployment: For real-world applications, the model should be deployed in a production environment with real-time data processing and anomaly detection.