horizontal line

**Dhanumitha Vijayakumar**

Data Science, January 2025 Batch

Machine Failure Prediction Report

**4th March 2025**

# OVERVIEW

Machine failures can lead to operational downtime, increased maintenance costs, and reduced efficiency. This report analyzes machine sensor data to predict future failures, allowing for preventive maintenance and optimized performance.

## OBJECTIVE

The primary goal of this project is to build a predictive model that can assess the likelihood of a machine failure based on real-time sensor data. The objectives include:

* Analyzing historical sensor data to identify key indicators of machine failure.
* Training machine learning models for failure prediction.
* Evaluating model performance using standard metrics.
* Deploying the best model for real-time monitoring.

# PARAMETERS INFLUENCING THE MACHINE FAILURE

1. Footfall: The number of people or objects passing by the machine.
2. Temp Mode: The temperature mode or setting of the machine.
3. AQ: Air quality index near the machine.
4. USS: Ultrasonic sensor data, indicating proximity measurements.
5. CS: Current sensor readings, indicating the electrical current usage of the machine.
6. VOC: Volatile organic compounds level detected near the machine.
7. RP: Rotational position or RPM (revolutions per minute) of the machine parts.
8. IP: Input pressure to the machine.
9. Temperature: The operating temperature of the machine.

# GOALS

The main goal of this project is to analyze the following aspects:

1. **Explore the data** (summary statistics, missing values, correlations)
2. **Visualize the data** (plots, distributions)
3. **Train a machine learning model** to predict failure.

# DATA ANALYSIS

#### **1. Basic Information**

* Number of rows-944
* Number of columns-10
* All columns are of integer type (int64)
* No missing values in any column

#### **2. Summary Statistics**

* footfall: Ranges from 0 to 7300, with a high standard deviation (1082.61), indicating variability.
* tempMode: Has values between 0 and 7.
* AQ (Air Quality Index): Values range from 1 to 7.
* USS (Ultrasonic Sensor): Values between 1 and 7.
* CS (Current Sensor): Values between 1 and 7.
* VOC (Volatile Organic Compounds): Ranges from 0 to 6.
* RP (Rotational Position/RPM): Between 19 and 91.
* IP (Input Pressure): Between 1 and 7.
* Temperature: Between 1°C and 24°C.
* fail (Machine Failure Indicator): Binary (0 or 1), with 41.63% of records indicating failure.

**3. Project Analysis**

* From the given dataset, it can be concluded that ‘fail’ is the dependent variable with discrete values. Since the data is numerical, it is a supervised data. It falls under the classification for discontinuous data.
* The non-failure count is higher than the failure count.
* Feature Importance Ranking

| **Features** | **Priority Percentage** |
| --- | --- |
| VOC | 41.8596% |
| AQ | 20.3282% |
| USS | 11.9737% |
| CS | 5.9648% |
| RP | 5.0531% |
| footfall | 4.9972% |
| Temperature | 4.5628% |
| IP | 2.6764% |
| tempMode | 2.5840% |

* Correlation of Features with Machine

| **Features** | **Correlation** | **Correlation in percentage** | **Analysis** |
| --- | --- | --- | --- |
| VOC | 0.797329 | 79.73% | High positive correlation |
| AQ | 0.583238 | 58.32% | Positive correlation |
| Temperature | 0.190257 | 19.02% | Positive correlation |
| IP | 0.085624 | 8.56% | Positive correlation |
| RP | 0.053668 | 5.36% | Positive correlation |
| CS | 0.018855 | 1.88% | Slightly neutral correlation |
| tempMode | -0.014462 | -1.44% | Negative correlation |
| footfall | -0.073066 | -7.30% | Negative correlation |
| USS | -0.466574 | -46.65% | High negative correlation |

1. **Training the Model**

Gradient boosting gives the most accuracy at 92%.

| **Model** | **Accuracy** |
| --- | --- |
| Logistic Regression | ~85% |
| Random Forest | ~90% |
| Gradient Boosting | ~92% |
| SVM | ~88% |

**ALGORITHMS USED IN MACHINE FAILURE PREDICTION**

## 1. Logistic Regression

## Concept: Logistic Regression is a statistical method used for binary classification problems. It estimates the probability of an event occurring (machine failure in this case) using the sigmoid function.

## Why Used?

## Simple and interpretable model

## Works well with linearly separable data

## Outputs probabilities, helping in risk assessment

## Limitations:

## Assumes linear decision boundary

## Performance drops with highly complex data

## 2. Random Forest Classifier

## Concept: Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions. Each tree is trained on a random subset of the dataset (bagging technique). The final prediction is made based on the majority vote from all trees.

## Why Used?

## Handles non-linear relationships well

## Reduces overfitting due to ensemble averaging

## Provides feature importance scores

## Limitations:

## Computationally expensive

## Difficult to interpret compared to simpler models

## 3. Gradient Boosting Classifier

## Concept: Gradient Boosting is another ensemble learning method that builds trees sequentially. Unlike Random Forest, where trees are trained independently, Gradient Boosting trains each tree to correct errors made by the previous ones. This is done by minimizing a loss function through gradient descent.

## Why Used?

## Highly accurate for structured datasets

## Handles complex relationships in data

## Works well with small to medium-sized datasets

## Limitations:

## Training can be slow

## Prone to overfitting if not carefully tuned

## 4. Support Vector Machine (SVM)

## Concept: SVM is a powerful classifier that finds the optimal hyperplane that best separates data points of different classes. It uses kernel functions to transform data into higher dimensions when needed.

## Why Used?

## Effective for high-dimensional spaces

## Works well when data is not linearly separable

## Can handle noisy data with soft margin techniques

## Limitations:

## Slower training time for large datasets

## Choosing the right kernel is crucial

## CONCLUSION AND INSIGHTS

* The Gradient Boosting Classifier emerged as the most effective model with 92% accuracy.
* Feature importance analysis highlighted CS (Current Sensor), Temperature, and VOC as the strongest predictors of failure.
* The developed model can be used to monitor real-time data and predict machine failures, allowing industries to take proactive maintenance measures.

The machine failure prediction model will classify whether a machine **passes (does not fail) or fails** based on the given sensor data.

### **How is the decision made?**

1. The trained **Gradient Boosting Classifier (92% accuracy)** or another chosen model will take new sensor readings as input.
2. The model will output **1 (Fail) or 0 (Pass/No Failure)** based on the learned patterns from historical data.
3. The probability of failure can also be calculated to determine how likely the machine is to fail.

It can be concluded that the machine does not fail, as the model predicts its result to be 1.