

SUMMER TRAINING REPORT ON

Analysis of Sensor Data for Machine Failure Prediction

Undertaken at
“LaunchED Tech Solutions”

BACHELOR OF TECHNOLOGY (B.tech)

By

“Swastik Pradhan”

*“Kalinga Institute of
Industrial Technology”*



Launched global Ed-tech Solutions

Email used for LaunchED Global: swastikpradhan2003@gmail.com

Contents/Index

Topic	Page No
Project Title	1
Content /Index Page	2
Abstract /Project Summary	3
Introduction	4
Dataset Description	5
Visualizations	7
Methods and Algorithms	12
Project Analysis	14
Final Results	15
Conclusion and Future Scope	18
References	19

Abstract /Project Summary

This internship report is to summarize the data on machine failure based on various different factors through real time sensor reading. Then creating

Different Visuals in Tableau / PowerBI to see patterns and trends in data. Then creating a Machine Learning Model that can effectively predict the new machine data and tell whether it will be failure or not.

Introduction

The analysis of sensor data for machine failure prediction is a critical application of data analytics in modern industrial settings. This Internship project Report, undertaken focuses on leveraging real-time sensor readings to identify patterns and trends associated with machine performance and potential failures. By utilizing a comprehensive dataset comprising variables such as footfall, temperature mode, air quality, ultrasonic sensor data, current sensor readings, volatile organic compounds, rotational position, input pressure, and operating temperature, the study aims to develop predictive Machine learning models for machine failure. Visualizations created using tools like Tableau or PowerBI facilitate the identification of key trends and patterns , while a machine learning model is employed to predict the likelihood of machine failure based on new sensor data. This approach enhances predictive maintenance strategies, enabling proactive interventions to minimize downtime and optimize operational efficiency.

Dataset Description

	A	B	C	D	E	F	G	H	I	J	K
1	Footfall	TempMode	AQ	USS	CS	VOC	RP or RPM	IP	Temperature	Fail	
2	0	7	7	1	6	6	36	3	1	1	
3	190	1	3	3	5	1	20	4	1	0	
4	31	7	2	2	6	1	24	6	1	0	
5	83	4	3	4	5	1	28	6	1	0	
6	640	7	5	6	4	0	68	6	1	0	
7	110	3	3	4	6	1	21	4	1	0	
8	100	7	5	6	4	1	77	4	1	0	
9	31	1	5	4	5	4	21	4	1	0	
10	180	7	4	6	3	3	31	4	1	0	
11	2800	0	3	3	7	0	39	3	1	0	
12	1600	0	3	2	4	4	26	2	1	0	
13	330	5	4	3	6	1	31	4	1	0	
14	190	2	5	4	6	5	22	4	1	1	
15	100	7	4	4	6	0	42	5	1	0	
16	1000	7	5	7	4	0	74	1	1	0	
17	0	7	6	7	5	0	62	3	1	0	
18	130	7	4	4	5	1	58	3	1	0	
19	5	5	3	3	6	1	24	6	1	0	
20	33	7	6	2	6	5	51	4	1	1	
21	19	2	2	1	4	0	36	3	2	0	
22	74	7	4	4	7	2	88	2	2	0	
23	190	0	2	4	6	2	20	4	2	0	

The various attributes of the following dataset are as follows based on the different terms of machine failure prediction for various sensor data:-

Columns Description: -

Footfall: The number of people or objects passing by the machine.

Temp Mode: The temperature mode or setting of the machine.

AQ: Air quality index near the machine.

USS: Ultrasonic sensor data, indicating proximity measurements.

CS: Current sensor readings indicate the machine's electrical current usage.

VOC: Volatile organic compounds level detected near the machine.

RP: The machine parts' rotational position or RPM (revolutions per minute).

IP: Input pressure to the machine.

Temperature: The operating temperature of the machine.

Fail: Binary indicator of machine failure (1 for failure, 0 for no failure).

So based on the above data readings from various machines our ultimate goal is to find if the machine will face failure from various factors like Footfall,Temp Mode,AQ,USS,CS,VOC,RP,IP,Temperature.So our target variable that we wish to find out for given machine data from sensors readings is Fail column (1:Fail,0:Not Fail).This comprehensive set of attributes enables the analysis of machine performance patterns and the development of predictive models for proactive maintenance.

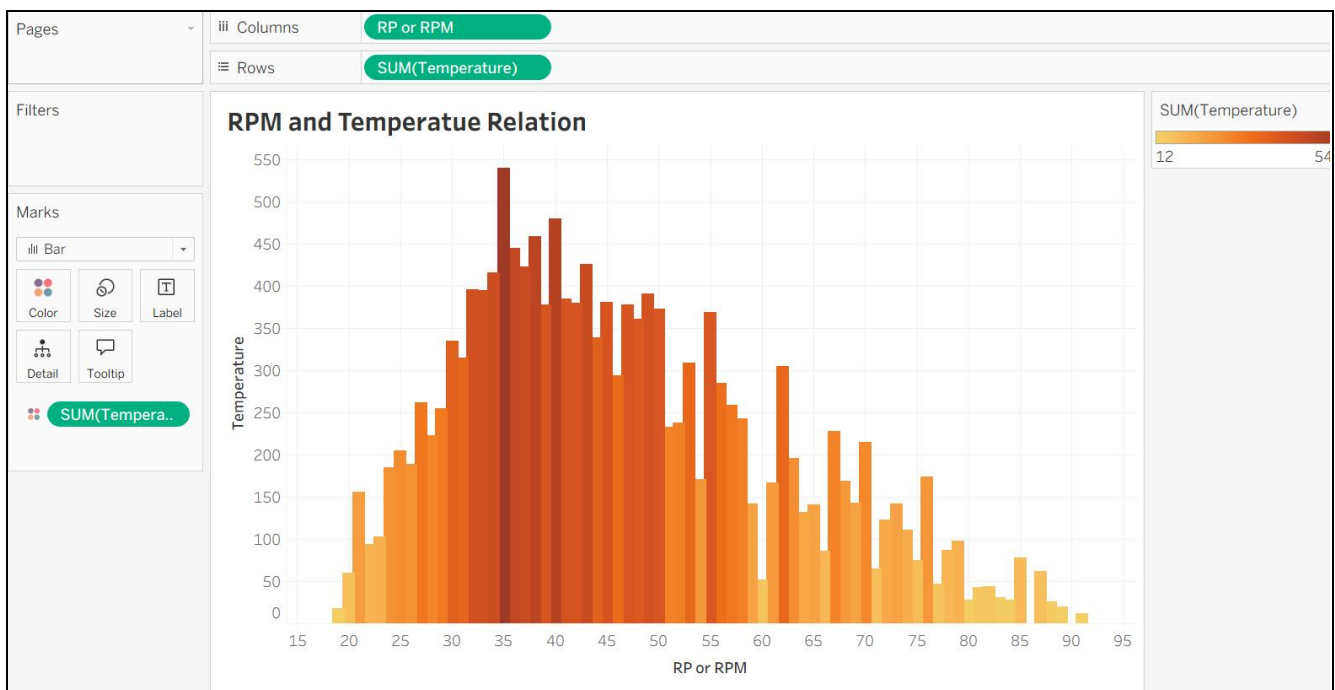
Visualizations

Now on basis of visualizations every parameter must have some relation with each other in order to make a meaningful graph to see patterns/trends. So we have some visuals of the analysis of sensor data where we can use different attributes to make some meaningful visuals to see possible scenarios of possible machine failure, using Tableau(2025.2) visual software.

1) Visual 1

Relation: RPM vs. Temperature

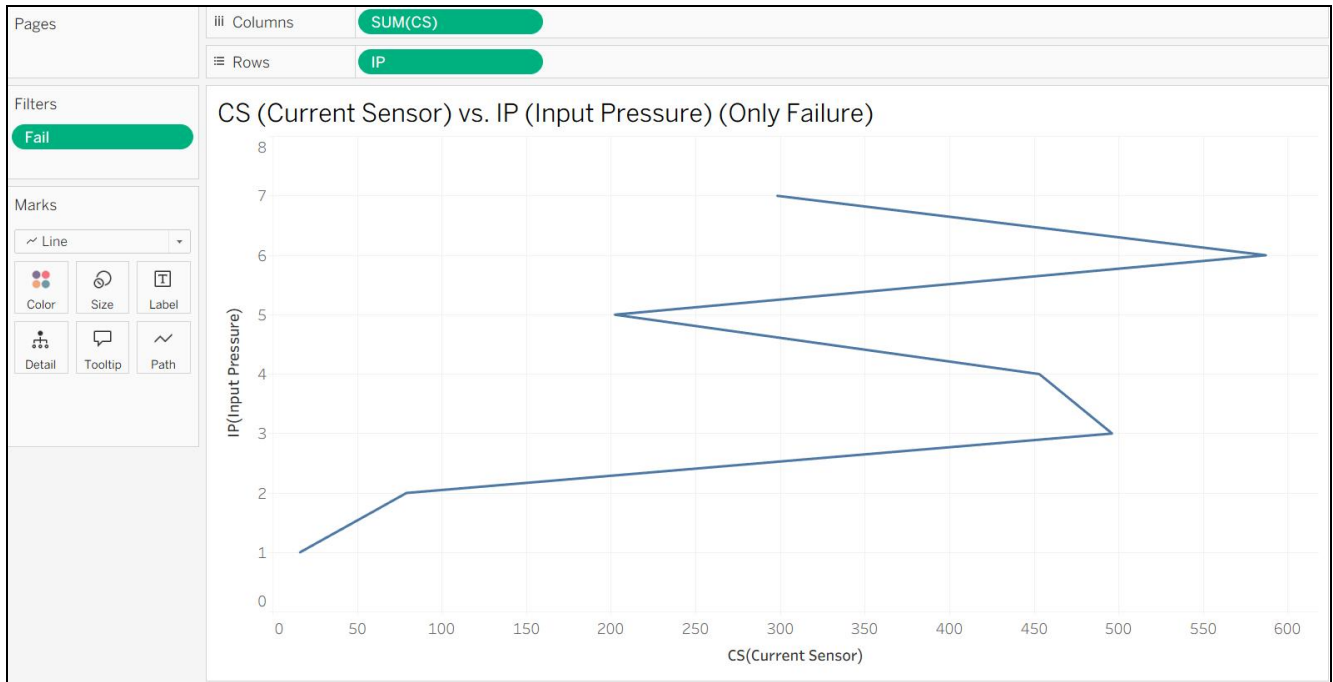
Why relation suitable: High RPM (revolutions per minute) may increase machine temperature due to friction or mechanical stress, potentially leading to failures. For instance, high RPM combined with elevated temperatures might cluster around failure points.



2) Visual 2

Relation: CS (Current Sensor) vs. IP (Input Pressure)

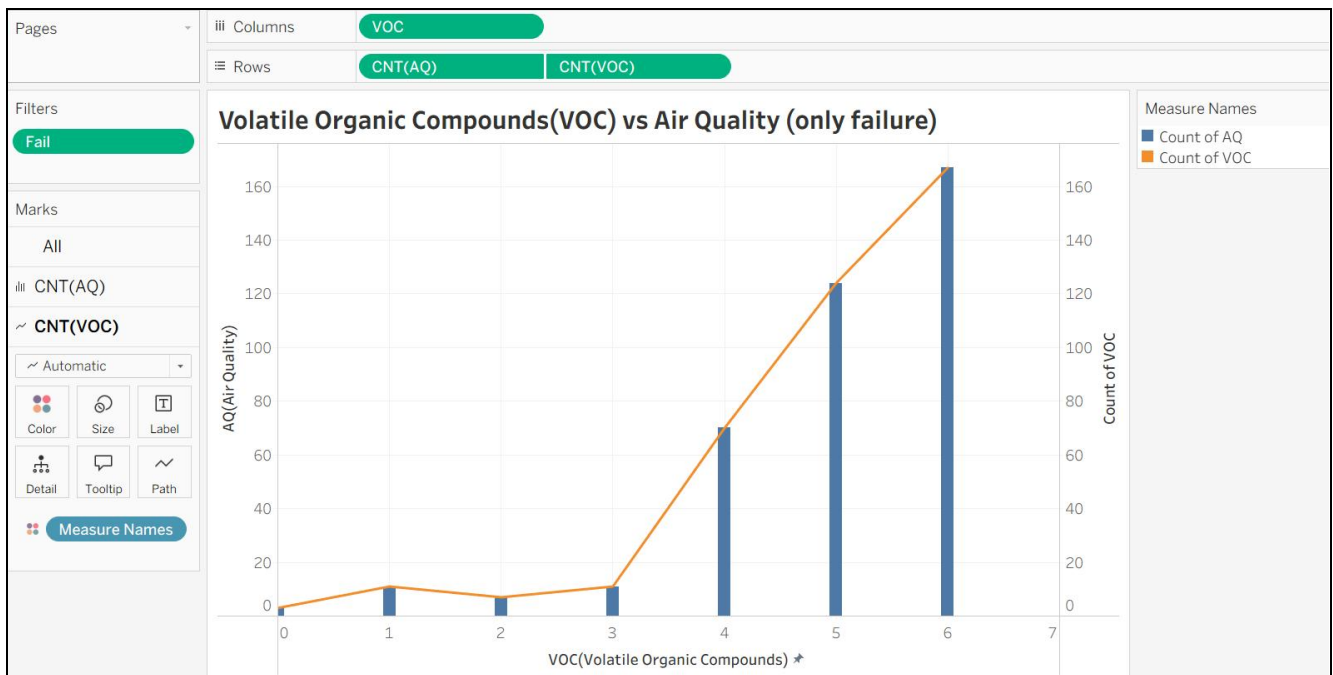
Why relation suitable: Current sensor readings (CS) reflect electrical usage, while input pressure (IP) indicates mechanical stress. operational conditions (e.g., high current and pressure) that are failure-prone. For example, high current with low pressure might indicate motor strain or inefficiency.



3) Visual 3

Relation: Volatile Organic Compounds(VOC) vs Air Quality (only failure)

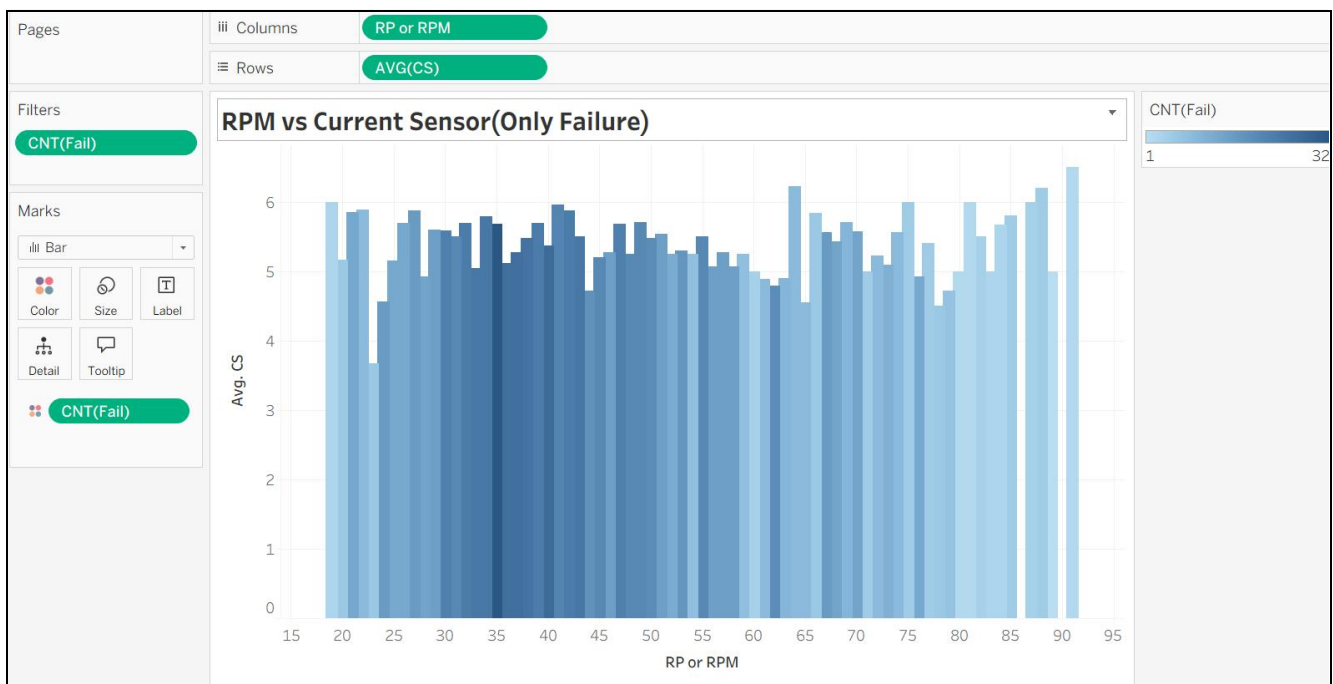
Why relation suitable: Both VOC and AQ relate to environmental conditions around the machine. AQ can reveal if poor air quality (high AQ) coincides with high VOC levels, which might indicate environmental stress contributing to failures.



4) Visual 4

Relation: RPM vs. CS (Current Sensor)(Only Failure)

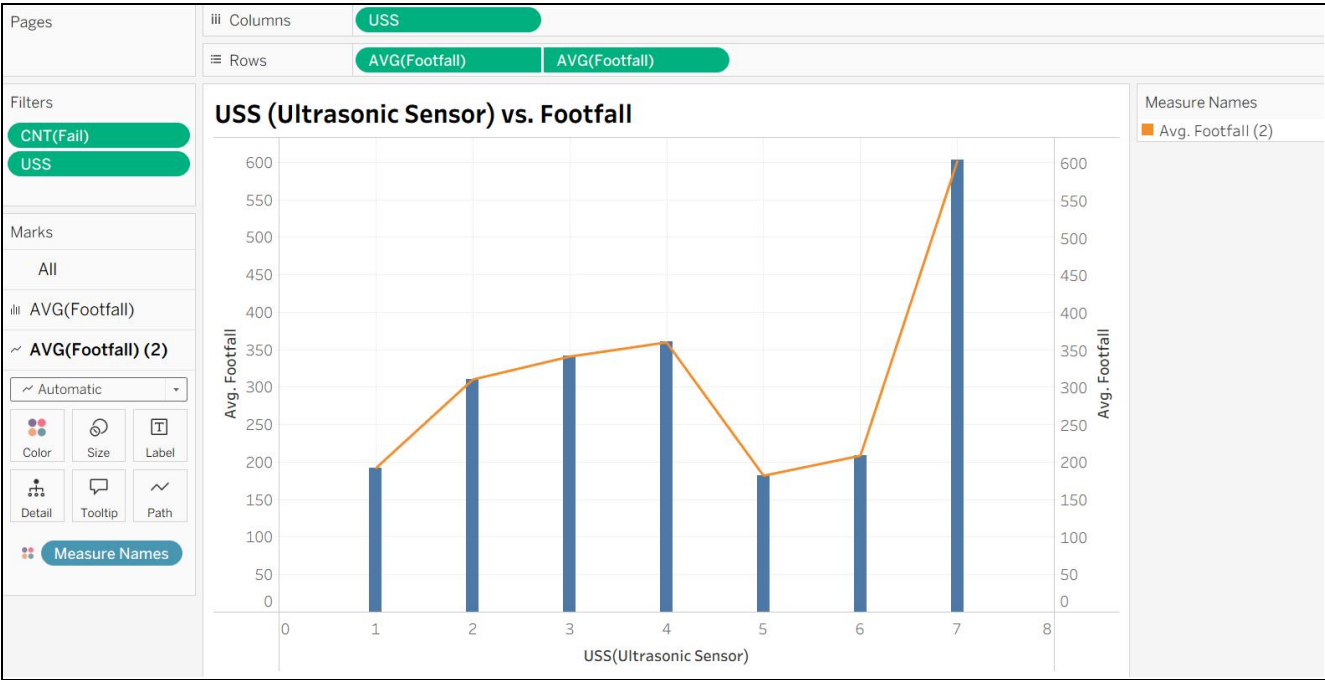
Why relation suitable: RPM reflects mechanical activity, while CS indicates electrical consumption. High RPM with low current might suggest mechanical inefficiency or motor issues.



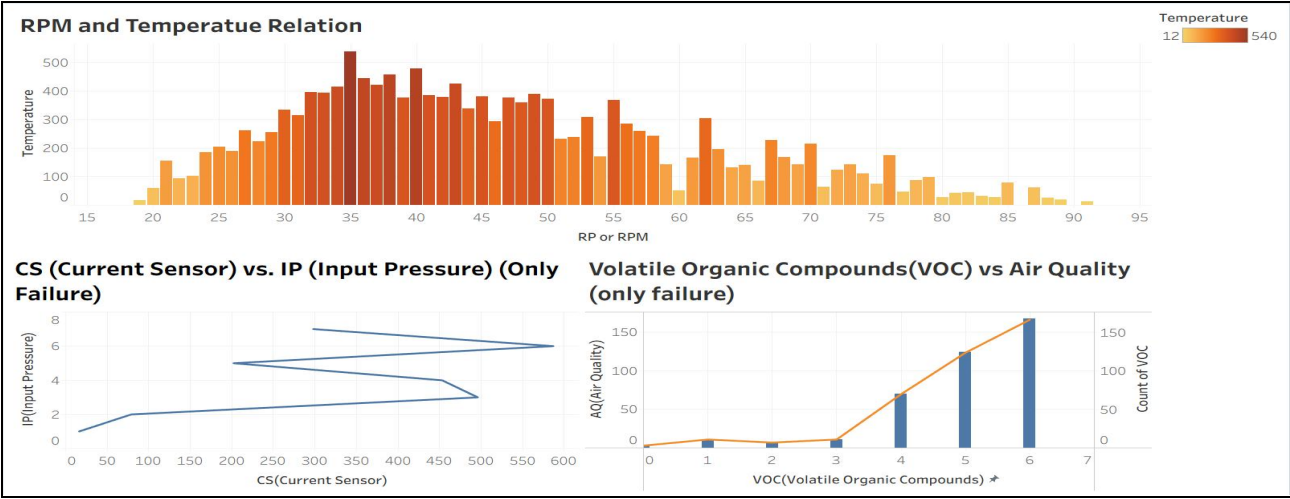
5) Visual 5

Relation: USS (Ultrasonic Sensor) vs. Footfall

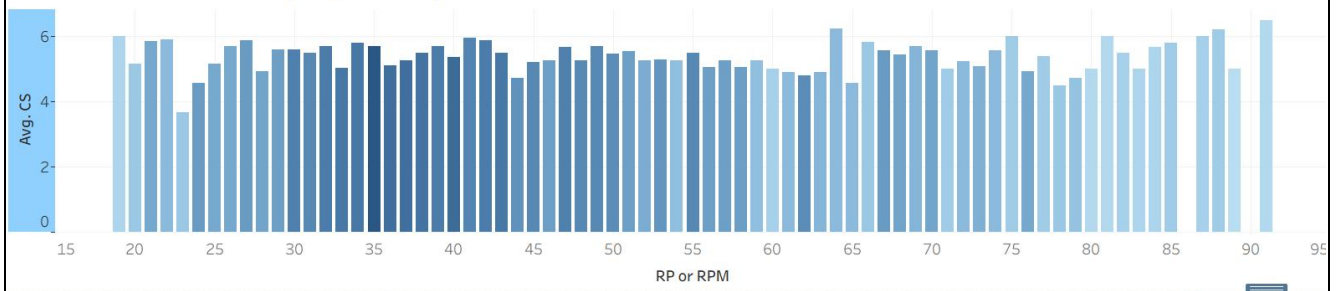
Why relation suitable: Ultrasonic sensor data (USS) measures proximity, which may correlate with footfall, as more objects or people nearby could trigger closer proximity readings. High footfall corresponds to specific USS values, potentially indicating crowded or obstructed environments that stress the machine.



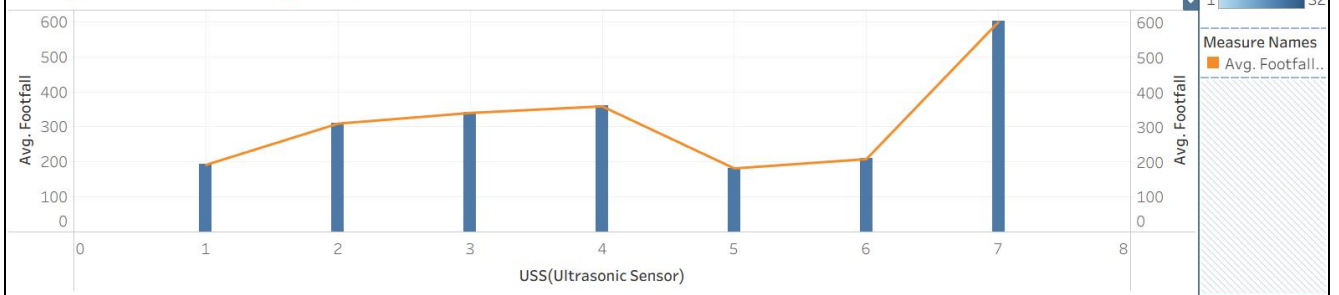
Dashboards



RPM vs Current Sensor(Only Failure)



USS (Ultrasonic Sensor) vs. Footfall



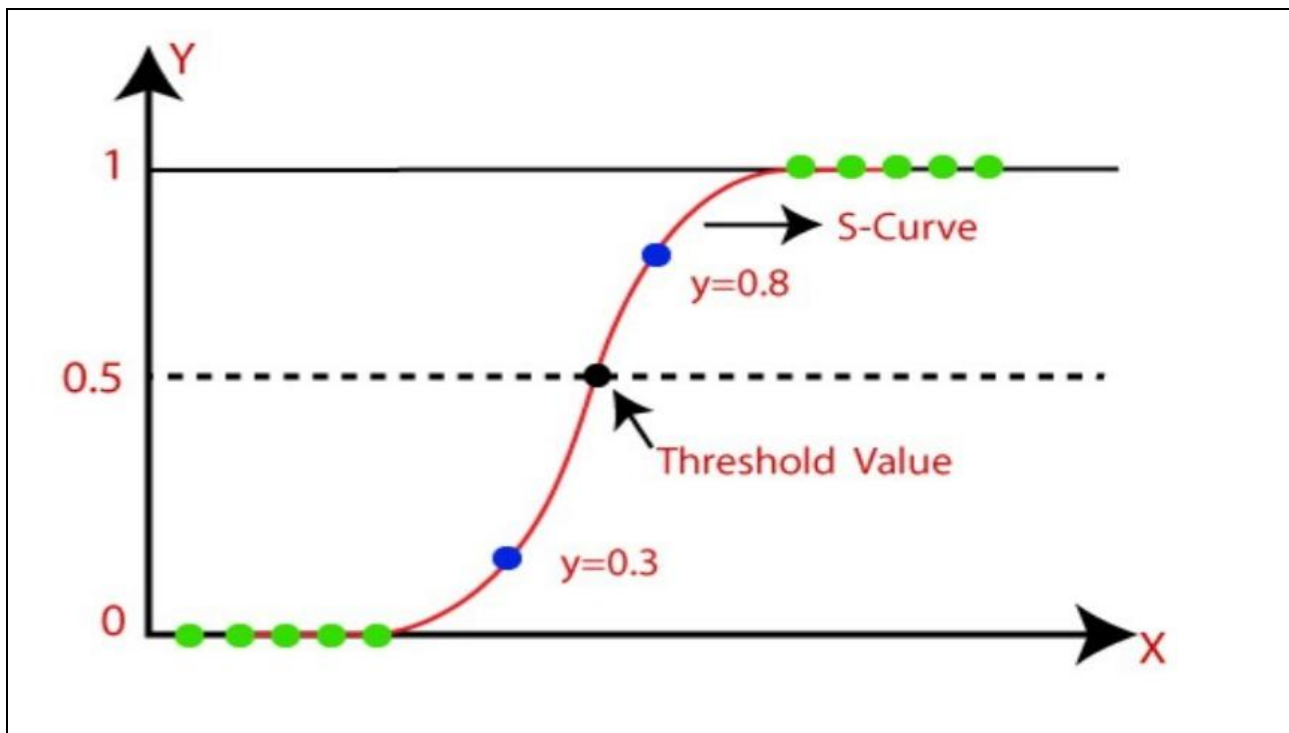
Methods and Algorithms

As Per the Data set we have received,we can clearly say that it is classification type problem.

In a Classification type problem we classify/distinguish between various different classes to which a particular data instance may belong to.

For the data where we have for Machine failure where the target label/column is Fail (1=Fail,0=Not Fail),we can clearly see that it is binary classification where we have to distinguish the data points into 2 classes of Fail and Not Fail.

So for binary classification one of the best algorithms that is used in this project is the Supervised learning of Logistic regression.Where in Logistic Regression use of sigmoid function equation to classify a data instance in such a way that it falls into one of the 2 classes.



So the Threshold value is is set to 0.5 and any other data point when it will be put in the equation of the sigmoid function,and any other value which will obtained if the value is such that it is greater than 0.5 then it belongs to one class

and below 0.5 it belongs to another class.

Logistic regression is a highly suitable method for the analysis of the sensor data in this project due to its ability to predict binary outcomes, such as the "Fail" variable (0 for no failure, 1 for failure), which is the primary focus of the machine failure prediction task.

This statistical technique models the relationship between a set of independent variables—such as Footfall, TempMode, AQ, USS, CS, VOC, RP or RPM, IP, and Temperature—and the probability of a binary outcome. For the provided dataset, logistic regression can effectively identify how combinations of these sensor readings influence the likelihood of machine failure, enabling the development of a predictive model to classify new data points.

The utility of logistic regression in this context lies in its ability to handle both continuous and categorical variables, making it adaptable to the diverse range of attributes in the dataset. It provides interpretable coefficients that indicate the strength and direction of each variable's impact on failure probability (e.g., higher VOC or Temperature might increase the odds of failure).

Additionally, logistic regression assumes a linear relationship between the log-odds of the outcome and the predictors, which can be a reasonable approximation for the sensor data trends observed. By training on the labeled dataset, it can generate a robust model to predict failure risks proactively, supporting maintenance strategies and minimizing downtime.

The link for the working Logistic regression model is as follows done in the Google collab:

<https://drive.google.com/drive/folders/16kaGB839OIkkMoX5rtcpgzsFGX8r1JCm?usp=sharing>

Project Analysis

This section of the project report will describe the project in step wise details of how we approach it.

Steps Involved :-

- 1) Go through the data and get to know the terms /definitions.
- 2) Clean the data to remove Outliers /Anomalies if needed.
- 3) Analyze the data.
- 4) Based on the data after cleaning load the data into any visualization software (Tableau/PowerBI).
- 5) Find out meaningful relations in the data and try to make different charts/graphs/figures etc.
- 6) Then after the visualization ,prepare a Machine Learning model (in this case:Logistic Regression) to properly classify the data instances.
- 7) For the Logistic Model to classify the data instances properly we need to first split the data into training and testing data.
- 8) Train the Logistic Regression Model using the training data by seting appropriate parameters like learning rate ,train_test split,Max iterations etc.
- 9) After the Model is trained with the training data ,test the model with unseen/new data instances to see if the model has been trained properly or not.
- 10) Then to test the model's correctness/Precision in the training and testing,evaluate model performance using Confusion matrix,ROC/AUC Curve to check model classification performance.

Final Results

For the Results we need to understand the Evaluation matrices used in the model for determining the performance of model:

1) Accuracy: The proportion of correct predictions (both true positives and true negatives) out of all predictions made.

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{TP} + \text{TN} + \text{False Positives (FP)} + \text{False Negatives (FN)}}$$

It measures overall correctness but can be misleading for imbalanced datasets.

2) Precision: The proportion of true positive predictions out of all positive predictions made by the model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

It answers: "Of all instances predicted as positive, how many are actually positive?"
High precision indicates fewer false positives.

3) Recall (Sensitivity or True Positive Rate): The proportion of true positives correctly identified out of all actual positive instances.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

It answers: "Of all actual positive instances, how many did the model correctly identify?" High recall indicates fewer false negatives.

4) F1 Score: The harmonic mean of precision and recall, providing a single metric that balances both. Useful for imbalanced datasets.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

It ranges from 0 to 1, with 1 being perfect precision and recall.

5) Confusion Matrix: A table summarizing the performance of a classification model by comparing predicted and actual labels. For a binary classifier, it includes:

- True Positives (TP): Correctly predicted positive cases.
- True Negatives (TN): Correctly predicted negative cases.
- False Positives (FP): Negative cases incorrectly predicted as positive (Type I error).
- False Negatives (FN): Positive cases incorrectly predicted as negative (Type II error).

6) ROC-AUC Curve:

- **ROC (Receiver Operating Characteristic) Curve:** A plot of the True Positive Rate (Recall) against the False Positive Rate ($FPR = \frac{FP}{FP+TN}$) at various classification thresholds. It shows the trade-off between sensitivity and specificity.
- **AUC (Area Under the Curve):** A single scalar value summarizing the ROC curve, ranging from 0 to 1.
 - AUC = 1: Perfect model.
 - AUC = 0.5: Random guessing (no discriminative power).
 - AUC < 0.5: Worse than random.A higher AUC indicates better model performance in distinguishing between classes.

So the Results of the Machine learning model of Logistic regression for machine failure analysis of current sensor data reading is as follows :

```
Available sheets: ['Capstone Project', 'Details of Attributes']
Confusion Matrix:
[[100  10]
 [  7  72]]
```

```
Classification Report:
              precision    recall  f1-score   support

      0       0.93      0.91      0.92      110
      1       0.88      0.91      0.89       79

 accuracy      0.91
 macro avg     0.91
weighted avg     0.91
```

```
ROC-AUC Score: 0.9649021864211738
```

Conclusion and Future Scope

Conclusion:

The analysis of sensor data for machine failure prediction, conducted at LaunchED Tech Solutions, has demonstrated the potential of leveraging real-time sensor readings and advanced analytics to enhance predictive maintenance. By utilizing a comprehensive dataset encompassing variables such as Footfall, TempMode, AQ, USS, CS, VOC, RP or RPM, IP, and Temperature, the project successfully identified key patterns and trends associated with machine failures. Visualizations created in Tableau, such as those exploring VOC vs. Air Quality and RPM vs. Current Sensor, provided valuable insights into failure-prone conditions. The application of logistic regression further enabled the development of a robust model to predict failure probabilities, offering a proactive approach to minimize downtime and optimize operational efficiency. This work underscores the effectiveness of integrating data analytics and machine learning in industrial settings to support reliable machine performance.

Future Scope:

The project opens several avenues for future enhancement and application. Expanding the dataset with additional sensors or longitudinal data could improve the accuracy and generalizability of the predictive model. Incorporating advanced machine learning techniques, such as random forests or neural networks, may capture more complex relationships within the data. Real-time implementation of the model through an API or IoT integration could enable immediate alerts for maintenance teams. Additionally, exploring the impact of environmental factors (e.g., seasonal variations in AQ or VOC) and integrating cost-benefit analyses for maintenance strategies could further refine the approach. Future research could also extend this methodology to other industrial machines or sectors, broadening its impact on predictive maintenance practices.

References

- 1) <https://www.projectpro.io/article/data-science-project-report/620>
- 2) <https://grok.com/chat/4dd46249-572b-4298-9bad-286cd37f3ac4>
- 3) <https://github.com/perborgen/LogisticRegression>
- 4) <https://github.com/search?q=logistic+regression&type=repositories&p=2>
- 5) <https://www.scribd.com/document/684582732/Internship-report>
- 6) <https://www.perplexity.ai/search/can-you-suggest-me-a-website-w-3AazZMGSTgyqrp0vPAEQsg>
- 7) <https://www.perplexity.ai/search/import-the-neccesary-modules-t-OLGVAac3Q2GSGiIw36oamg>

