Arash Mahmoudian
Project: Preventative Maintenance Using AI/ML models
Dataset: Predictive Maintenance Dataset

Business Problem
Predictive Maintenance (PdM) in the industrial context refers to the use of data, analytics, and machine learning techniques to predict when equipment or machinery is likely to fail. By analyzing the condition and performance of equipment in real-time, organizations can take proactive measures to perform maintenance just in time, maximizing asset uptime, and minimizing downtime and associated costs.

Predictive Maintenance is widely adopted in industries with critical and expensive machinery, such as manufacturing, energy, transportation, and healthcare. The goal is to transition from a reactive or preventive maintenance approach to a more proactive and data-driven strategy, ultimately improving reliability, efficiency, and overall operational performance.

In this study Predictive Maintenance Dataset (AI4I 2020) from Kaggle is selected to perform predictive Maintenance Analysis on this. This synthetic dataset is modeled after an existing milling machine and consists of 10 000 data points from a stored as rows with 14 features.

Dataset description:
The dataset consists of 14 columns and 10000 records described below:

1- UID: unique identifier ranging from 1 to 10000.
2- 2- Product ID: consisting of a letters
    - L: low (50% of all products)
    - M: medium (30%)
    - H: high (20%)
    - As product quality variants and a variant-specific serial number.
3- Type: just the product type L, M or H from column 2.
4- Air temperature [K]: generated using a random walk process later normalized to a standard deviation of 2 K around 300 K.
5- Process temperature [K]: generated using a random walk process normalized to a standard deviation of 1 K, added to the air temperature plus 10 K.
6- Rotational speed [rpm]: calculated from a power of 2860 W, overlaid with a normally distributed noise.
7- Torque [Nm]: torque values are normally distributed around 40 Nm with a SD = 10 Nm and no negative values.
8- Tool wear [min]: The quality variants H/M/L add 5/3/2 minutes of tool wear to the used tool in the process.
9- A 'machine failure' label that indicates whether the machine has failed in this datapoint for any of the following failure modes are true.
10- The machine failure consists of five independent failure modes.
11- Tool Wear Failure (TWF): the tool will be replaced of fail at a randomly selected tool wear time between 200 - 240 mins (120 times in our dataset). At this point in time, the tool is replaced 69 times, and fails 51 times (randomly assigned).

12- Heat Dissipation Failure (HDF): heat dissipation causes a process failure, if the difference between air- and process temperature is below 8.6 K and the tools rotational speed is below 1380 rpm. This is the case for 115 data points.
13- below 3500 W or above 9000 W, the process fails, which is the case 95 times in our dataset.
14- Overstrain Failure (OSF): if the product of tool wear and torque exceeds 11,000 minNm for the L product variant (12,000 M, 13,000 H), the process fails due to overstrain. This is true for 98 datapoints.
15- Random Failures (RNF): each process has a chance of 0,1 % to fail regardless of its process parameters. This is the case for only 5 datapoints, less than could be expected for 10,000 datapoints in our dataset.

If at least one of the above failure modes is true, the process fails, and the 'machine failure' label is set to 1. It is therefore not transparent to the machine learning method, which of the failure modes has caused the process to fail.

Objective:
This study intended to offer a practicable model capable of predicting possible failure in machine, to help industries minimize failure risks in their machines.

Analysis:
Figure [1] shows the number of failures occurred in different product qualities. Most of products are labeled "L" indicating low quality.

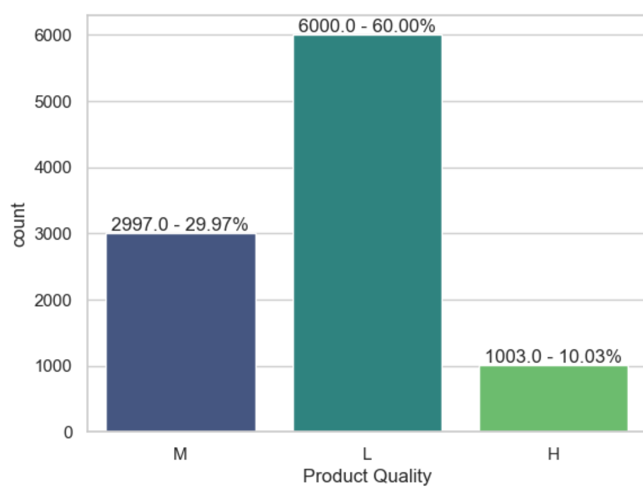


Figure [1]

Figure [2] depicts multiple histograms showing left skewed Rotational Speed and almost normal distribution for Torque.

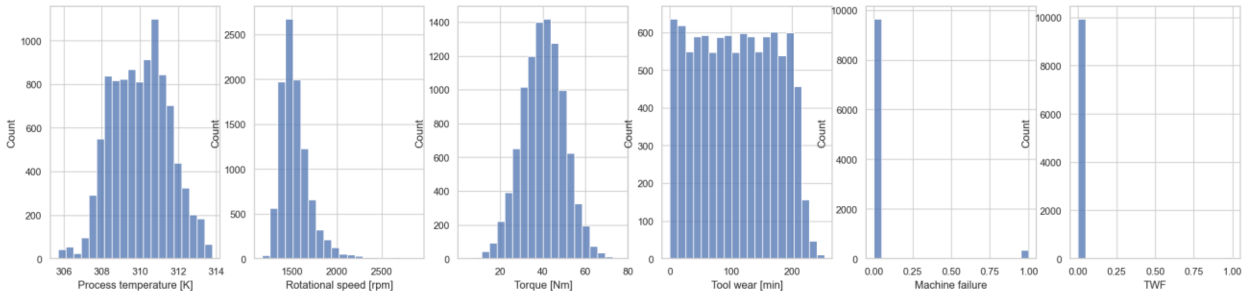


Figure [2]

Consequently, in Figure [3] outliers can be observed in both Rotational Speed and Torque.

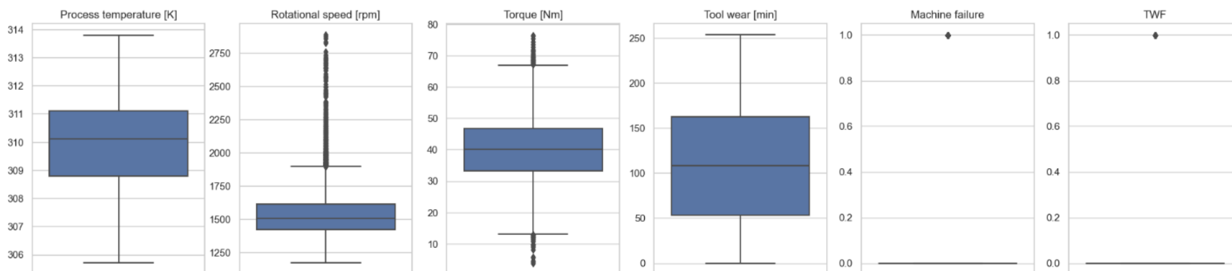


Figure [3]

Correlation analysis shows low quality product failures mostly occur in high and low Rotation speed. Also, it appears failures are mostly taken place when Torque is high and simultaneously when the Rotation speed is either high or low.

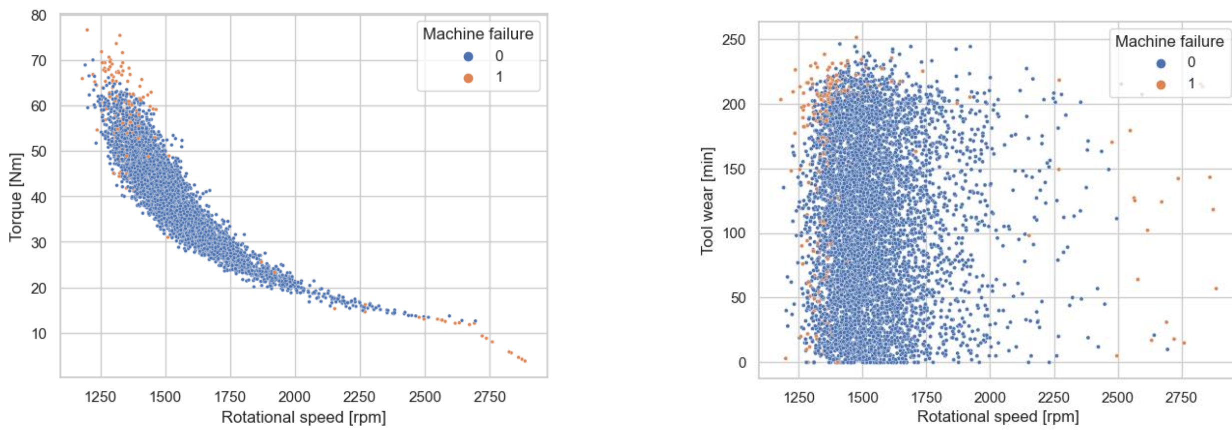


Figure [4]

Studying correlation matrix we can see that Machine Failure has the highest correlation of 0.19 to Torque [Nm] and 0.11 to Tool Wear[min]. On the other hand, the lowest correlation of -0.044 to Rotational Speed [rpm]

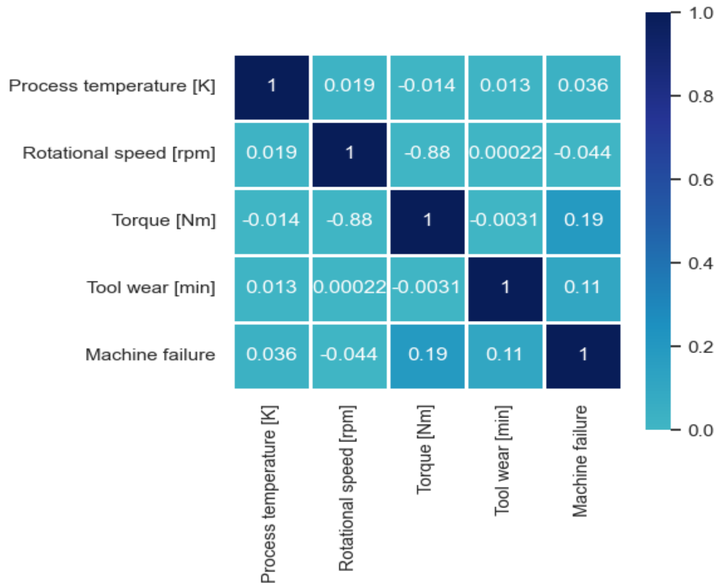


Figure [5]

Imbalance:
From number of failures standpoint our dataset shows an extreme imbalance behaviour as only 3.48% of records are labeled with “failure” Figure [6]

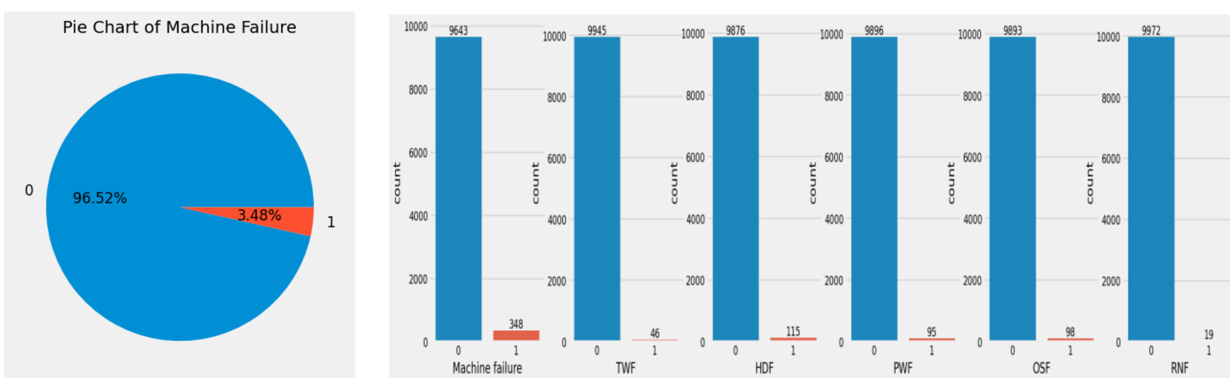


Figure [6]

SMOTE technique has been used to handle imbalanced dataset by generating synthetic samples for the minority class (Machine Failure). Figure [7] shows new dataset after removing data imbalances.

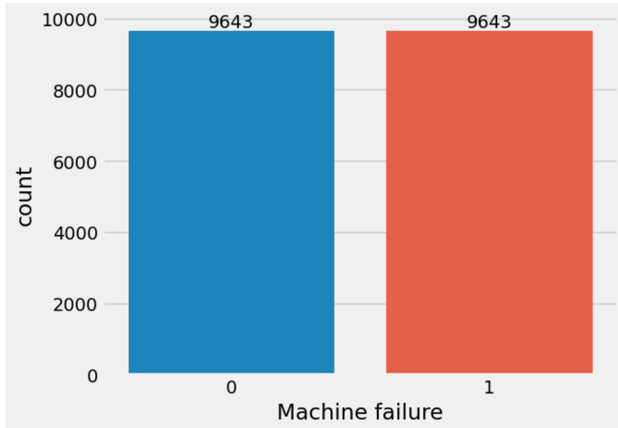


Figure [7]

Methods:

To detect failures following three Machine Learning techniques including Logistic Regression, Decision Tree, and Random Forest along with a Deep Learning model has been created evaluated for their performance comparision.

Row	**Model**	**Score**
1	LogisticRegression	0.836595
2	DecisionTree	0.976536
3	RandomForest	0.982370
4	Deep Learning	0.969500

In addition, the ROC curve of three machine learning models is plotted along with learning curve of DL model. ROC shows a lazy learning behavior whereas Random Forest has the sharpest learning curve.

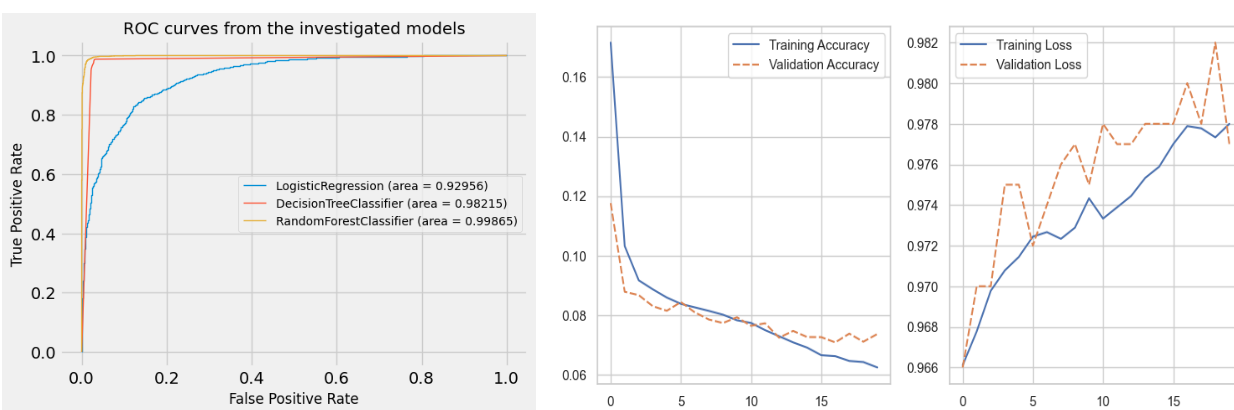
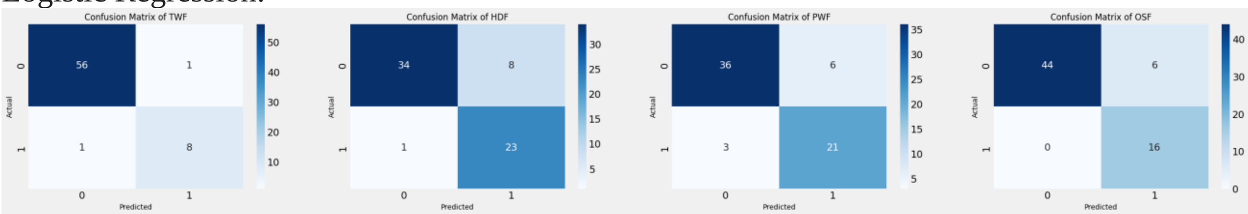


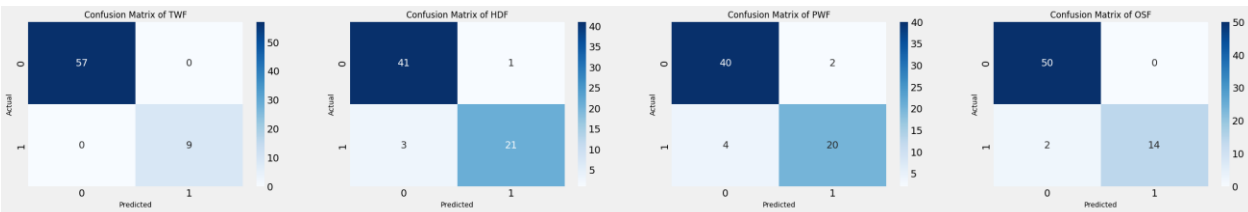
Figure [8]

Following confusion matrix plots show how different models are performing. Interestingly the Deep Learning model shows %100 accuracy on test dataset. However, this model is computationally expensive compared to the Machine Learning models.

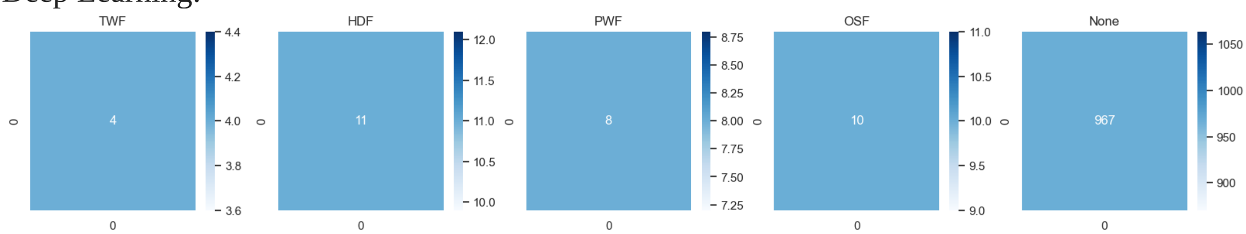
Logistic Regression:



Random Forest:



Deep Learning:



Conclusion:

Maintenance can be done both proactively and reactively, however waiting for a failure to occur in a machine can be expensive and time consuming. For this proactively monitoring machines can give us the ability to predict its future status and prevent possible failure in them. In this study a sample machine failure dataset has been selected then using measurements like rotation speed and torque and environmental temperature multiple models have been created to predict if a process with a given condition will result to a failure or not. A total of three machine learning models including Logistic Regression, Random Forest, and Decision Tree classification along with a simple Deep Learning model has been created and examined through this study.

Logistic Regression classification showed the worst performance with an accuracy of %83 whereas Random Forest had highest accuracy of %98. Additionally, although Deep Learning model showed accuracy level of %97, its performance on test data was %100.

Even though Deep Learning Model accuracy level was %100 on test data, it is an expensive method from computation standpoint. For this, for a such applications Random Forest Classification seems to be an appropriate approach as it is easy to make and doesn’t need to be trained for a longer time.

Finally, it is suggested to gather more failure data points and repeat this process again to make sure the model is not outperforming.

Limitations:
The dataset was imbalanced from number failure samples as compared to the conditions that doesn’t result in any failures.

Challenge:
Distributing failure points into both training and test sets was challenging, there were no enough failure samples.

Assumption:
Samples are collected via accurate measurement sensors and are labeled accurately.

Future Uses/Additional Applications:
With a robust and accurate model encompassing all kinds of failure types on different machines and conditions existing in an industry, companies can prevent malfunctioning and save on their maintenance expenditures.

Recommendations:
Besides gathering more failure samples from all forms, it is also recommended to gather more attributes like their user of machine and their skills. These features can be added to environmental conditions and be fed to the AI/ML models.

Questions:

1-What is preventative maintenance?

Preventive maintenance (PM) refers to the systematic care and servicing of equipment or machinery to keep it in good working condition and prevent potential breakdowns. The primary goal of preventive maintenance is to perform regular inspections, repairs, and replacements of components before they fail or cause larger issues. This proactive approach helps in avoiding unplanned downtime, reducing the likelihood of equipment failures, and extending the lifespan of assets.

2-Why preventative maintenance is important?

- Minimizes Downtime
- Reduces Repair Costs
- Extends Equipment Lifespan
- Improves Equipment Reliability
- Enhances Safety
- Optimizes Performance

3-How does AI/ML assist in preventative maintenance?

- Predictive Analytics
- Condition Monitoring
- Failure Mode and Effect Analysis (FMEA)
- Prescriptive Maintenance
- Dynamic Scheduling

4-Is machine learning better than deep learning?

The choice between machine learning (ML) and deep learning (DL) for preventative maintenance depends on various factors, including the nature of the data, the complexity of the problem, and the availability of resources. Both ML and DL are subsets of artificial intelligence, but they have different strengths and applications.

5-Why is correlation chart important in EDA?

- Identifying Relationships
- Strength of Relationships
- Variable Selection
- Multicollinearity Detection
- Pattern Recognition
- Data Quality Check

6-What industries can benefit from AI/ML preventative maintenance?
AI/ML preventative maintenance can provide significant benefits across a wide range of industries where equipment, machinery, and infrastructure play a crucial role. Here are some industries that can benefit from the implementation of AI and ML in preventative maintenance:
Manufacturing, Energy and Utilities, Transportation and Logistics, Aerospace, Oil and Gas, Healthcare, Telecommunications, etc.

7-Can we say Random Forest can be uniformly used for preventative maintenance?
While Random Forest is a powerful machine learning algorithm and can be applied to various predictive maintenance tasks, it's important to note that the suitability of any algorithm, including Random Forest, depends on the specific characteristics of the data and the goals of the predictive maintenance application. Here are some considerations: Versatility, Handles Non-linearity, Ensemble Learning, Hyperparameter Tuning

8-What extra features can be collected for such a purpose?
There are common types of features that are often valuable for building effective predictive maintenance models. Here are some categories of features you might consider collecting for preventative maintenance purposes:
Vibration, Pressure, Noise Levels, Run Time, Start/Stop Frequency, Humidity Levels, Dust and Particle Levels, Geospatial Data

9-What is the use case for correlation matrix?
A correlation matrix is commonly used in data analysis and statistics to examine the relationships between variables in a dataset. It's particularly useful for exploring the strength and direction of linear relationships between pairs of variables.

10-What is the difference between ROC and Deep Learning trend chart?
ROC (Receiver Operating Characteristic) curve and a "Deep Learning Trend Chart" are different types of visualizations used in different contexts, particularly in the evaluation of machine learning models. The ROC curve is used to evaluate the performance of a classification model, particularly binary classifiers (models that classify instances into one of two classes). The term "Deep Learning Trend Chart" is not a standard term in the context of model evaluation. However, assuming it's used to describe a trend chart related to deep learning models, it might refer to tracking the performance of a deep learning model over time or epochs during training.

Ethical Consideration:

The dataset reference indicates that dataset was collected from real data and there are no scrambling and fake data in it. Additionally, as the companies and manufactures try to keep their data confidential, there are no links/refers to the origin of this dataset. Moreover, this dataset is part of the following publication; therefore, any citation should go to this publication:

Reference:

1- S. Matzka, "Explainable Artificial Intelligence for Predictive Maintenance Applications," 2020 Third International Conference on Artificial Intelligence for Industries (AI4I), 2020, pp. 69-74, doi: 10.1109/AI4I49448.2020.00023.

2- https://www.kaggle.com/datasets/stephanmatzka/predictive-maintenance-dataset-ai4i-2020