IE7374 Final Project Report

AI4I Predictive Maintenance

Group 14
Veer Karadia
Abhishek Taware
Vaibhavee Pulgam
Aabhijatya Somvanshi

karadia.v@northeastern.edu
taware.ab@northeastern.edu
pulgam.v@northeastern.edu
somvanshi.a@northeastern.edu

Percentage of Efforts contributed by Student 1: 25%
Percentage of Efforts contributed by Student 2: 25%
Percentage of Efforts contributed by Student 3: 25%
Percentage of Efforts contributed by Student 4: 25%

Signature of Student 1: *Veer Karadia*
Signature of Student 2: *Abhishek Taware*
Signature of Student 3: *Vaibhavee Pulgam*
Signature of Student 4: *Aabhijatya Somvanshi*

Submission Date: 08/08/2022

Abstract

In this paper, we present to you an analysis and implementation of a dataset known as Artificial Intelligence also known as the AI4I Predictive Maintenance Dataset which was released in the year 2020 to predict because of which mechanical / environmental parameters a machine can fail causing the assembly line to come to a halt which would reduce the efficiency of the line.

To draw our conclusions with accuracy, we have cleaned the data as per our understanding and processed it in such a way in which we can have maximum efficiency with our findings. Some of the things that we have done include changing different Machine Failure Modes to a single column to use them as a target variable, converting categorical variables to numeric to make them usable. After this, we have created some visualizations to analyze the data which can help us to understand the data better and perform some additional analysis to optimize our algorithms. Moving on, we have implemented a few Machine Learning Models such as Naive Bayes, Neural Networks, Logistic Regression and SVM to check which algorithm delivers the result with highest accuracy.

Project Definition and Problem Setting

In machinery heavy industries such as Oil & Gas, Manufacturing, Telecom, Railways etc wherein Automation is used extensively, Predictive Maintenance (PdM) plays an important part in smooth functioning of the machinery so that the system stays in place without disrupting any part of the business.

PdM is a type of condition-based maintenance system that is used to keep an eye on the condition of these machinery devices through sensors and other equipment. These sensors are used to supply data in real-time, which is used to predict when the asset might require maintenance which might avoid any equipment failure if any. The dataset we have at hand consists of 10,000 records with 14 attributes to define these records.

The objective of this study is to understand how we can use the data at hand to predict machine failures using different models such as Logistic Regression, Naive Bayes, Neural Networks, and SVM on the data available to us.

Data Description

In the Dataset, we have 10,000 records, 14 attributes including 6 outcome classes.

Primary Attributes

1) UID: unique identifier ranging from 1 to 10000
2) Product ID: Consisting of a letter L, M, or H for low (50% of all products), medium (30%) and high (20%) as product quality variants and a variant-specific serial number.
3) Air Temperature [K]: Generated using a random walk process later normalized to a standard deviation of 2K around 300K.
4) Process Temperature [K]: Generated using a random walk process normalized to a standard deviation of 1K, added to the air temperature plus 10K.
5) Rotational Speed [rpm]: calculated from a power of 2860 W, overlaid with a normally distributed noise.
6) Torque [Nm]: Torque values are normally distributed around 40 Nm with a Ïƒ = 10 Nm and no negative values.
7) Tool Wear [min]: The quality variants H/M/L add 5/3/2 minutes of tool wear to the used tool in the process.

8) Machine Failure: Indicates whether the machine has failed in this particular datapoint for any of the following failure modes are true.

Independent Failure Modes

If at least one of the below 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

1. Tool Wear Failure (TWF): The tool will be replaced or fail at a randomly selected tool wear time between 200 & 240 mins.
2. 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 tool's rotational speed is below 1380 rpm.
3. Power Failure (PWF): The product of torque and rotational speed (in rad/s) equals the power required for the process. If this power is below 3500W or above 9000W, the process fails.
4. 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.
5. 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 data points, less than could be expected for 10,000 data points in our dataset.

Data Source

Link to the UCI Repository:
https://archive.ics.uci.edu/ml/datasets/AI4I+2020+Predictive+Maintenance+Dataset

Methods used

Principal Component Analysis

In today’s world, the size of datasets have grown exponentially. What earlier used to be a few hundred ows, have grown multifolds today making them difficult to interpret. Principal Component Analysis is one such technique using which we can reduce the dimensions of the dataset due to which its readability can increase without having to incur any loss of data. PCA creates new uncorrelated variables which maximize the variance of the dataset. By doing this, PCA reduces to solving eigenvalue or eigenvector problem as the new variables are defined by the dataset during runtime.

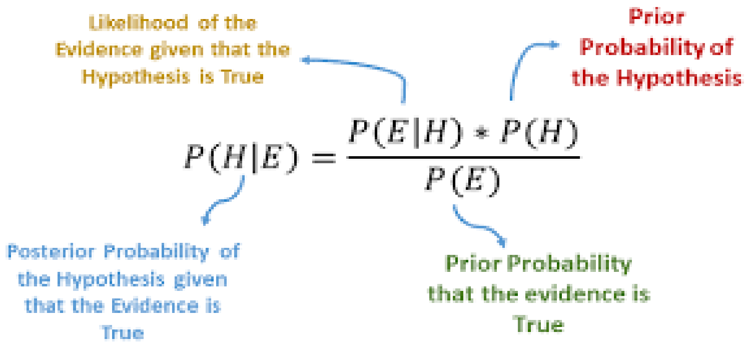
Logistic Regression

Logistic Regression was used in the early twentieth century in the biological sciences, later used in many social science applications. Predictive analytics and categorization frequently make use of this kind of statistical model. Based on a given dataset of independent variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. In logistic regression, the odds—that is, the probability of success divided by the probability of failure—are transformed using the logit formula. It is also referred to as the log odds or the natural logarithm of odds.

$$h\theta(X) = \frac{1}{1 + e^{-\left(\beta_0 + \beta_1 X\right)}}$$

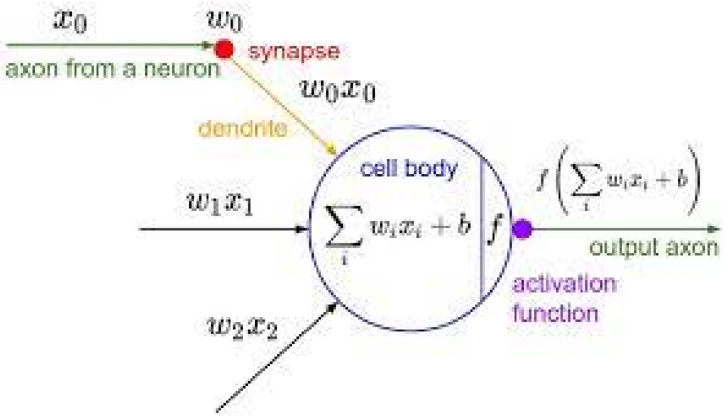
Naive Bayes

Naive Bayes is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. It is a family of algorithms rather than a single method, and they are all based on the idea that every pair of features being classified is independent of the other. Because they are probabilistic, they determine the probabilities of each tag for a given text and output the tag with the highest likelihood.The way they get these probabilities is by using Bayes' Theorem, which describes the probability of a feature, based on prior knowledge of conditions that might be related to that features.


Likelihood of the Evidence given that the Hypothesis is True
Prior Probability of the Hypothesis
$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$
Posterior Probability of the Hypothesis given that the Evidence is True
Prior Probability that the evidence is True

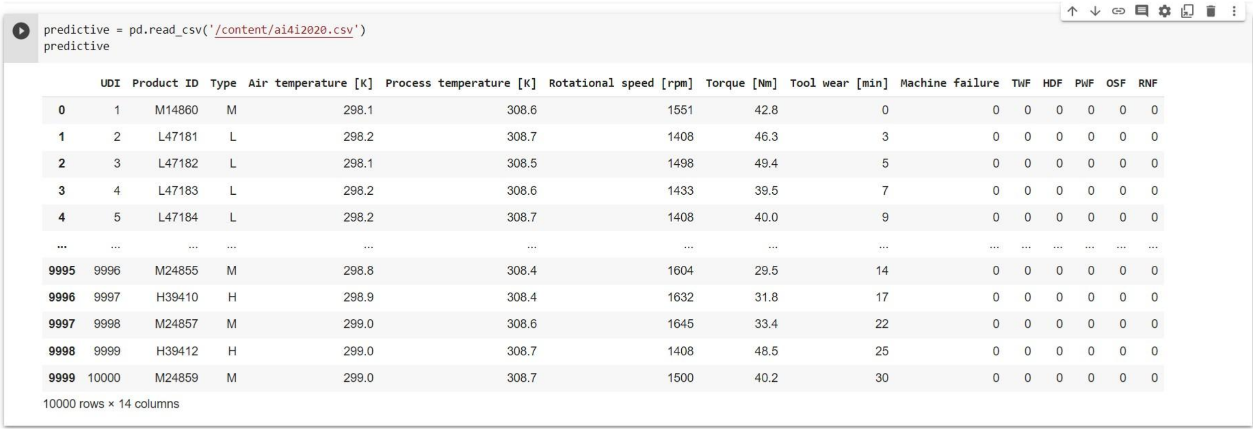
Neural Networks

Neural networks, which are a subset of machine learning and are at the core of deep learning algorithms, are also known as artificial neural networks (ANNs) or simulated neural networks (SNNs).Their name and structure are derived from the human brain, and they replicate the way organic neurons communicate with one another.Training data is essential for neural networks to develop and enhance their accuracy over time. However, these learning algorithms become effective tools in computer science and artificial intelligence once they are adjusted for accuracy, enabling us to quickly classify and cluster data.



Exploratory Data Analysis

Overview of the dataset

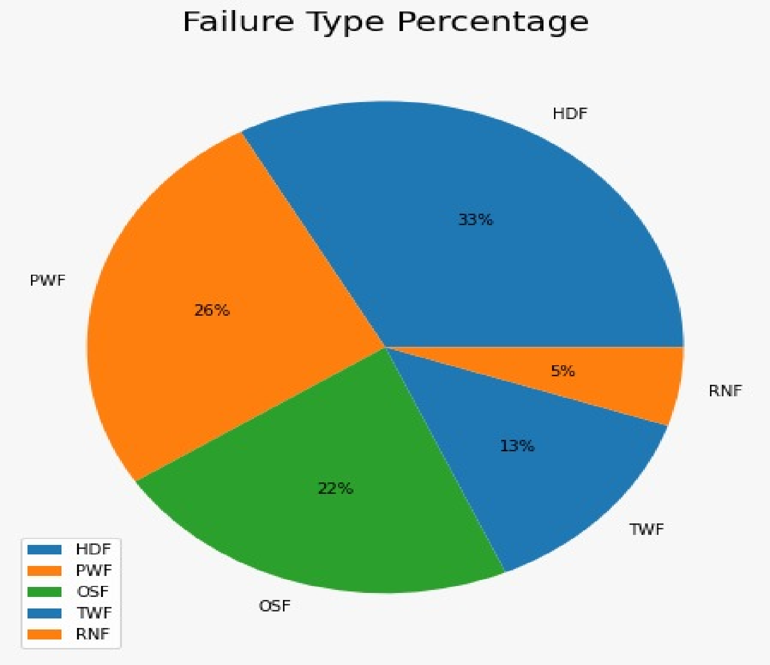


```
predictive = pd.read_csv('/content/ai4i2020.csv')
predictive
```

	UDI	Product ID	Type	Air temperature [K]	Process temperature [K]	Rotational speed [rpm]	Torque [Nm]	Tool wear [min]	Machine failure	TWF	HDF	PWF	OSF	RNF
0	1	M14860	M	298.1	308.6	1551	42.8	0	0	0	0	0	0	0
1	2	L47181	L	298.2	308.7	1408	46.3	3	0	0	0	0	0	0
2	3	L47182	L	298.1	308.5	1498	49.4	5	0	0	0	0	0	0
3	4	L47183	L	298.2	308.6	1433	39.5	7	0	0	0	0	0	0
4	5	L47184	L	298.2	308.7	1408	40.0	9	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9995	9996	M24855	M	298.8	308.4	1604	29.5	14	0	0	0	0	0	0
9996	9997	H39410	H	298.9	308.4	1632	31.8	17	0	0	0	0	0	0
9997	9998	M24857	M	299.0	308.6	1645	33.4	22	0	0	0	0	0	0
9998	9999	H39412	H	299.0	308.7	1408	48.5	25	0	0	0	0	0	0
9999	10000	M24859	M	299.0	308.7	1500	40.2	30	0	0	0	0	0	0

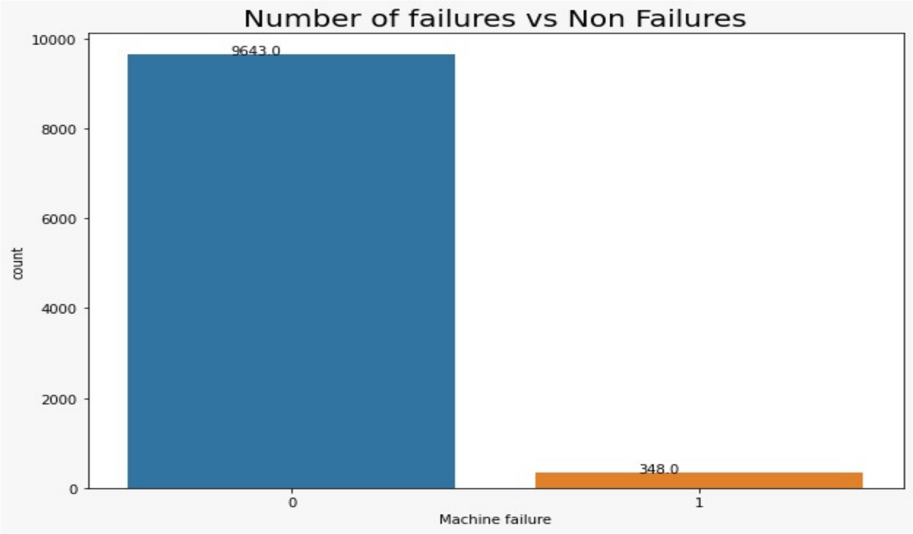
10000 rows × 14 columns

Pie chart of the failure share of each Failure Type



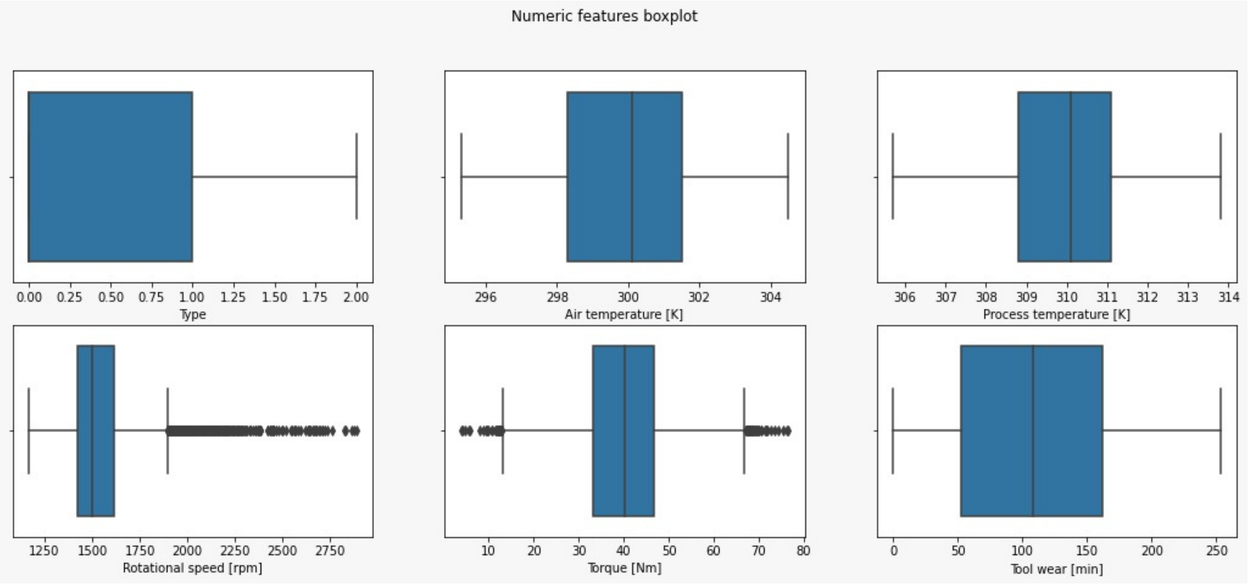
The number of failures vary depending on the type of the failure. The HDF class has the highest failure of 33% compared to other failures whereas RNF has the least failures with 5%.

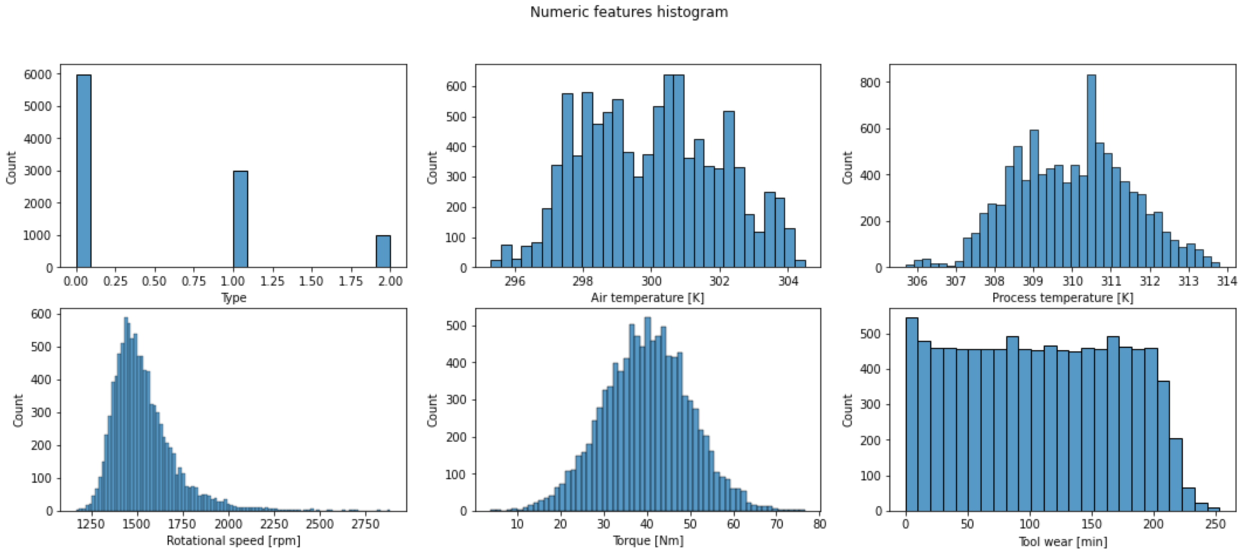
Comparison of Failure and Non Failure rows



As we can see in the graph, the number of rows having failure is just 348 which is approximately 3.49% of the dataset, we can conclude that the dataset is imbalanced. To overcome this imbalance, we have resampled the data so that each Failure Type is represented equally in the dataset. In order to achieve this, we have used the SMOTE Technique which works by randomly picking a point from the minority class and compute the k nearest neighbors to the data point.

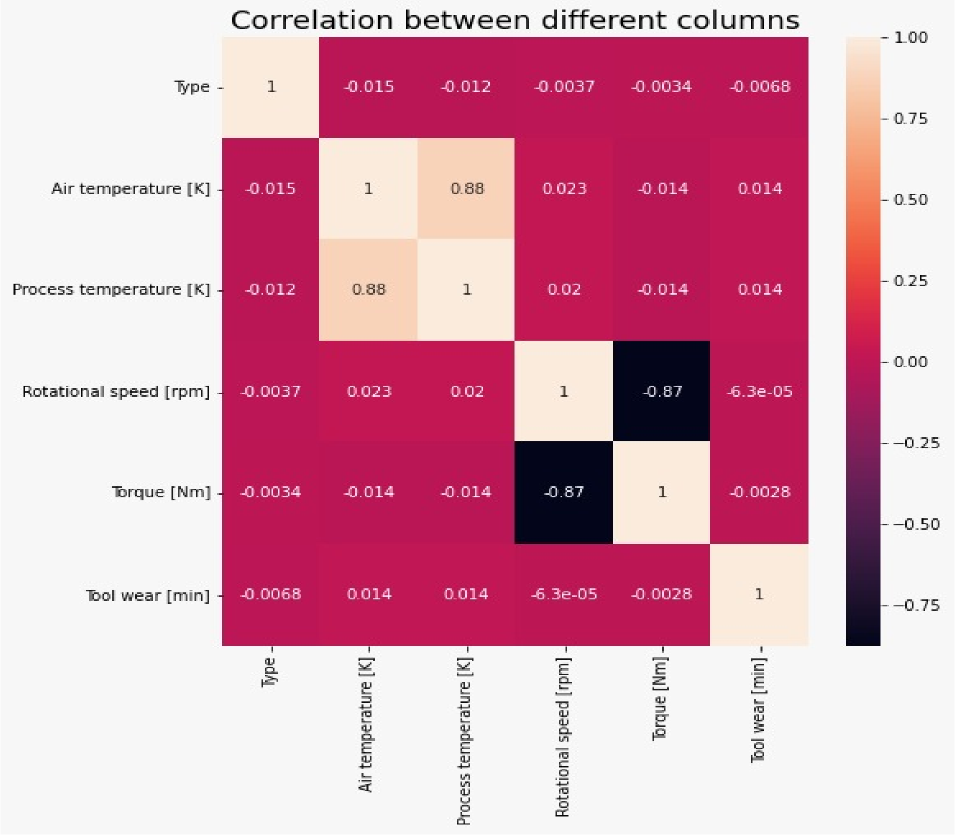
Box Plot & Histogram of Numeric Features





As per the boxplot we have highlighted the possible features of the dataset, however in the Torque feature the boxplot skewed compared to Rotational speed, which is also a huge difference between both the features. Therefore we have kept the outliers for now and later decide whether to act on or not after considering other aspects.

Correlation Matrix for Feature Selection



According to the correlation graph, the features related to temperature and the features related to power, are highly correlated. The Torque and Rotational speed are not correlating well as per the observation. But since we have only a few selectable input columns, we've decided to proceed with the dataset as it is.

Results

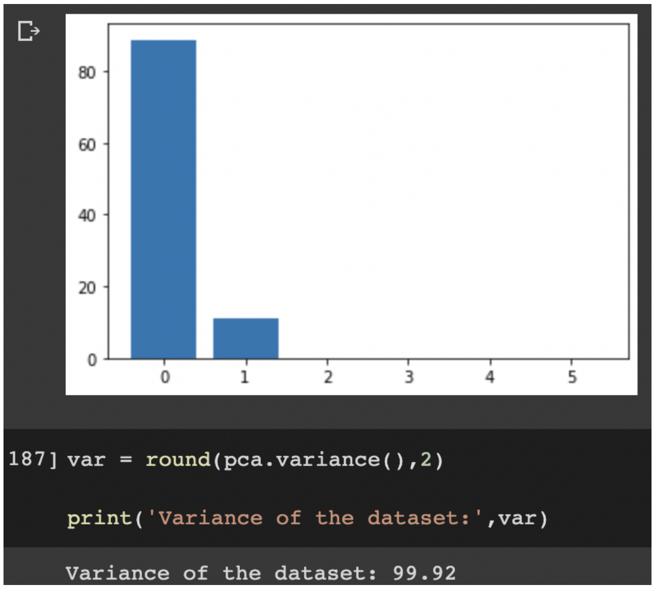
Confusion Matrix (Train)

	Logistic Regression	G Naive Bayes	Neural Networks
Accuracy	0.82	0.945	0.853
F1 Score	0.17	0.848	0.629
Precision	0.17	0.794	0.546
Recall	0.17	0.91	0.742

Confusion Matrix (Test)

	Logistic Regression	G Naive Bayes	Neural Networks
Accuracy	0.83	0.948	0.899
F1 Score	0.17	0.852	0.745
Precision	0.17	0.792	0.632
Recall	0.17	0.922	0.908

Principal Component Analysis

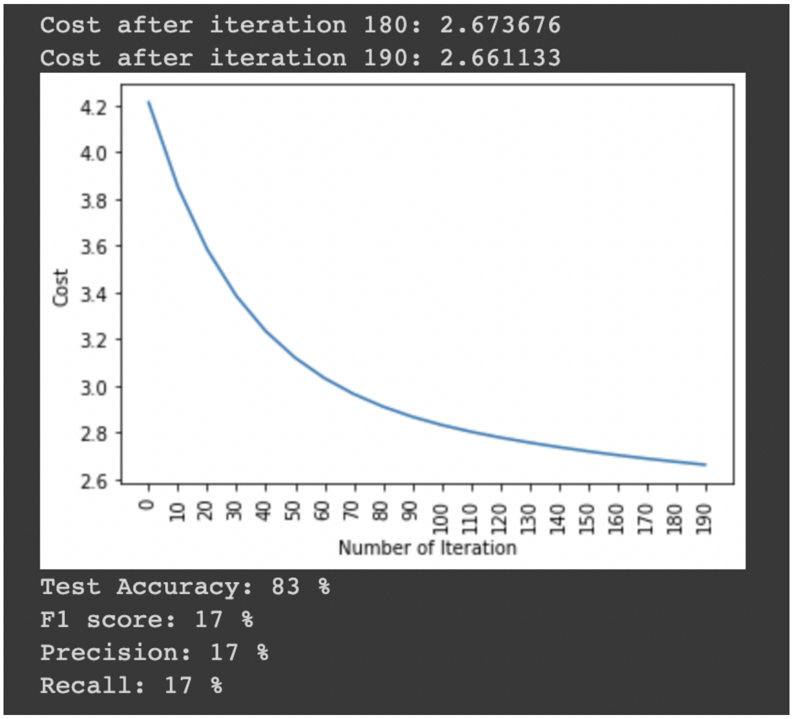


```
187] var = round(pca.variance(),2)

    print('Variance of the dataset:',var)
```

```
    Variance of the dataset: 99.92
```

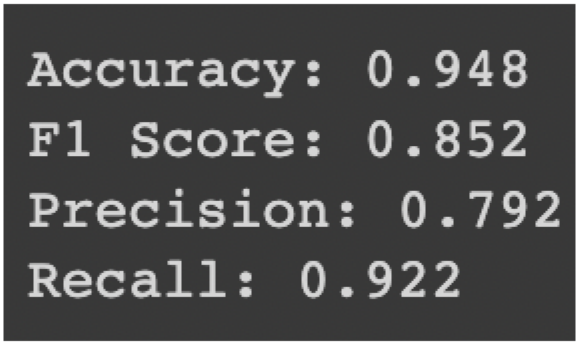
Logistic Regression



```
Cost after iteration 180: 2.673676
Cost after iteration 190: 2.661133
```

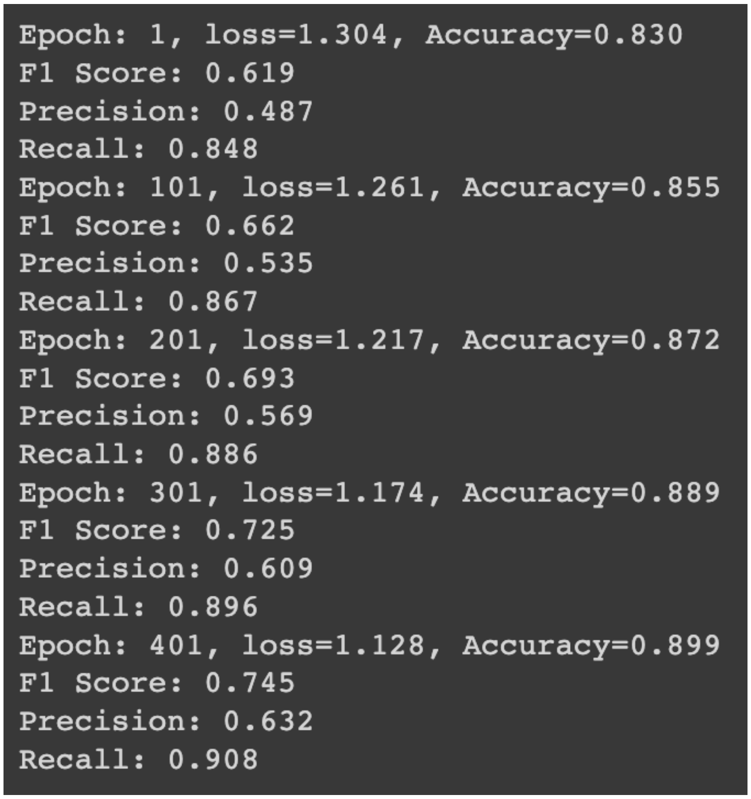
```
Test Accuracy: 83 %
F1 score: 17 %
Precision: 17 %
Recall: 17 %
```

Gaussian Naive Bayes



```
Accuracy: 0.948
F1 Score: 0.852
Precision: 0.792
Recall: 0.922
```

Neural Networks



```
Epoch: 1, loss=1.304, Accuracy=0.830
F1 Score: 0.619
Precision: 0.487
Recall: 0.848
Epoch: 101, loss=1.261, Accuracy=0.855
F1 Score: 0.662
Precision: 0.535
Recall: 0.867
Epoch: 201, loss=1.217, Accuracy=0.872
F1 Score: 0.693
Precision: 0.569
Recall: 0.886
Epoch: 301, loss=1.174, Accuracy=0.889
F1 Score: 0.725
Precision: 0.609
Recall: 0.896
Epoch: 401, loss=1.128, Accuracy=0.899
F1 Score: 0.745
Precision: 0.632
Recall: 0.908
```

Discussion

As per the results shared above, we first ran PCA on the dataset which in turn gave out a very high prediction rate. Following which, we performed logistic regression, Gaussian Naive Bayes, and Neural Networks Model. When we execute Logistic Regression, we come to notice that in PCA, only the variance of the independent variables and not of the response variable is considered. Logistic Regression considers all the independent variables' properties on the response variable. Gaussian Naive Bayes which gave us the highest F1 Score and Accuracy outperforming all the other algorithms. In the end, we run the algorithm for Neural Networks which outperforms other algorithms except Gaussian Bayes with a high accuracy of 89% and a F1 Score of 0.75