SLP IDP : TRIDIAGONAL

Industrial Design Problem

ANOMALY DETECTION IN BATCH PROCESSES

FINAL REPORT

Prof. Rajdip Bandhyopadhyaya | Prof. Sanjay Mahajani | Prof. Mani Bhushan

Tapan Kumar Sir, Parth Sinha Sir, Tridiagonal

Krishna Baldwa 210110066

Naufran Neyas 210020083

Pranav Maniyar 210020094

Department of Chemical Engineering, IIT Bombay

TABLE OF CONTENTS

1. Introduction and Problem Scope Definition

1.1 Problem Statement

- List of potential failures in the autoclave machines
- Presentation on relevant research articles focusing on the mathematical modeling approach
- Mapping of relevant process parameters for production Autoclave machines (PFDs)
- Develop a predictive maintenance model: Using AI-ML predict if the next batch is going to fail
- Approach document for modeling
- Development of models using python (JNB interface)
- Causation of the input features (process parameters) on the batch failures

1.2 What is an Autoclave?

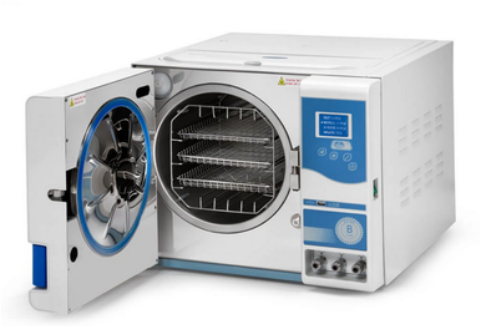


Fig 1 : Autoclave

An autoclave is a machine that uses steam under pressure to kill harmful bacteria, viruses, fungi, and spores on items that are placed inside a pressure vessel. The items are heated to an appropriate sterilization temperature for a given amount of time

1.3 What is an Autoclavable?

Devices must be compatible with the autoclave process. Autoclavable items must be compatible with conditions of high heat and moisture and should be processed per the manufacturer's written instructions for use. Medical devices that have contact with sterile body tissues or fluids are considered critical items. These items may include surgical instruments, implanted medical devices and surgical drapes and linens. These items should be sterile when used because any microbial contamination could result in infection transmission. Steam is often the sterilant of choice for sterilization of heat and moisture stable items because it is reliable, consistent, and lethal to microorganisms while being safe for staff who operates the autoclave.

1.4 **Operating Principles**

- The autoclave works on the principle of moist heat sterilization where steam under pressure is used to sterilize the material present inside the chamber.
- The high pressure increases the boiling point of water and thus helps achieve a higher temperature for sterilization.
- Water usually boils at 100°C under normal atmospheric pressure (760 mm of Hg); however, the boiling point of water increases if the pressure is to be increased.
- Similarly, the high pressure also facilitates the rapid penetration of heat into deeper parts of the material, and moisture present in the steam causes the coagulation of proteins causing an irreversible loss of function and activity of microbes.
- This principle is employed in an autoclave where the water boils at 121°C at the pressure of 15 psi or 775 mm of Hg.
- When this steam comes in contact on the surface, it kills the microbes by giving off latent heat.
- The condensed liquid ensures the moist killing of the microbes.
- Once the sterilization phase is completed (which depends on the level of contamination of material inside), the pressure is released from the inside of the chamber through the whistle.

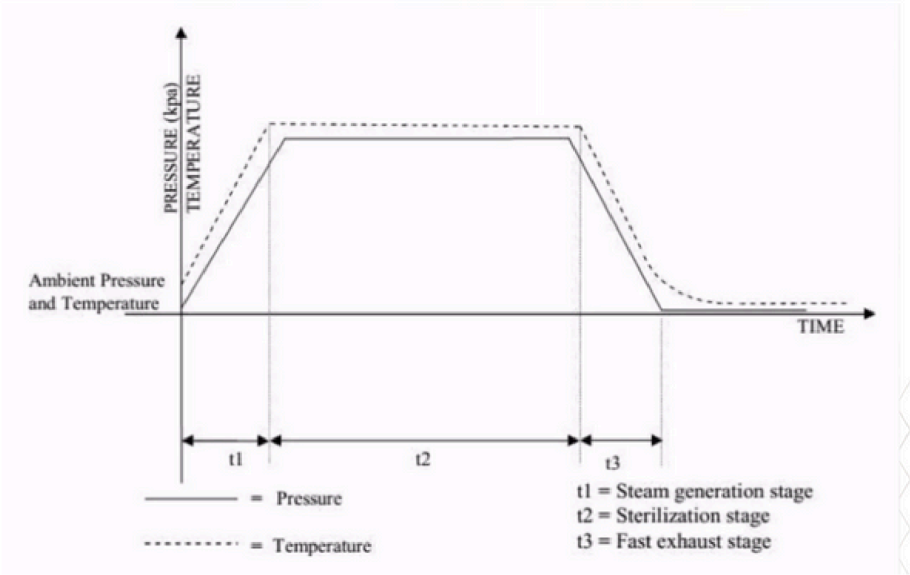


Fig 2 : Sterilization cycle

2. Working

2.1 Working Principle

1. In general, an autoclave is run at a temperature of 121° C for at least 30 minutes by using saturated steam under at least 15 psi of pressure. The following are the steps to be followed while running an autoclave:
2. Before beginning to use the autoclave, it should be checked for any items left from the previous cycle.
3. A sufficient amount of water is then put inside the chamber.
4. Now, the materials to be sterilized are placed inside the chamber.
5. The lid is then closed, and the screws are tightened to ensure an airtight condition, and the electric heater is switched on.
6. The safety valves are adjusted to maintain the required pressure in the chamber.
7. Once the water inside the chamber boils, the air-water mixture is allowed to escape through the discharge tube to let all the air inside to be displaced. The complete displacement can be ensured once the water bubbles cease to come out from the pipe.
8. The drainage pipe is then closed, and the steam inside is allowed to reach the desired levels (15 lbs in most cases).
9. Once the pressure is reached, the whistle blows to remove excess pressure from the chamber.
10. After the whistle, the autoclave is run for a holding period, which is 15 minutes in most cases.
11. Now, the electric heater is switched off, and the autoclave is allowed to cool until the pressure gauge indicates the pressure inside has lowered down to that of the atmospheric pressure.
12. The discharge pipe is then opened to allow the entry of air from the outside into the autoclave.
13. Finally, the lid is opened, and the sterilized materials are taken out of the chamber.

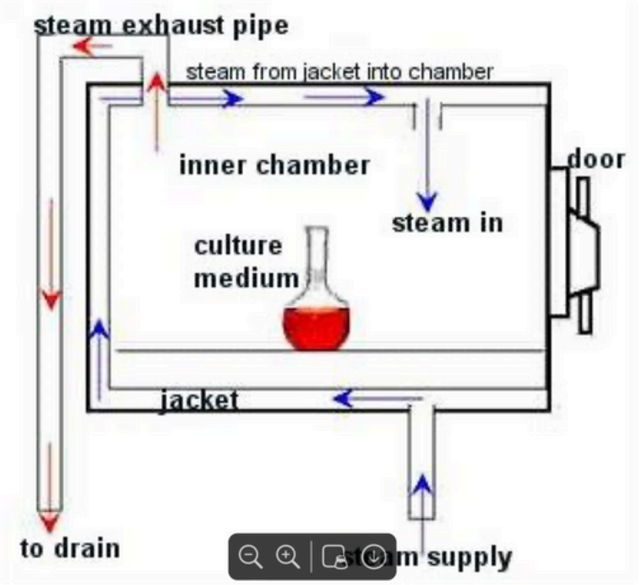


Fig 3 : Sterilization Process

2.2 **Good Manufacturing Practices**

1. Pre-Operational Checks - Ensuring the autoclave is clean, functioning correctly, no moisture remaining from previous cycle
2. Loading the Autoclave: Ensure packets are separated for efficient steam penetration.
3. Choosing appropriate sterilization cycle - eg solid, liquid, medical devices and its validation
4. Monitoring and Documentation - continuously monitoring the cycle parameters help prevent failure.
5. Checking deviations in sterilization process and looking for appropriate correction for the same
6. Maintenance and Calibration from time to time
7. Working personnel should be well equipped & trained using the autoclave
8. Emergency Procedures: Have clear procedures in place for dealing with emergencies or malfunctions of the autoclave.

3. Literature Review

3.1 Articles

3.1.1 Machine Learning for Equipment Failure Prediction and Predictive Maintenance

- ID — ID field that represents a specific machine.
- DATE — The date of the observation.
- MAINTENANCE_VENDOR — a field that represents the company that provides maintenance and service to the machine.
- EQUIPMENT_AGE — Age of the machine, in days.
- S1, S2, S3, S4, S5, S6 etc. — Sensor Values
- EQUIPMENT_FAILURE — A '1' means that the equipment failed. A '0' means the equipment did not fail.

Data transformations and Feature Engineering

Using idea from this paper we will create running summaries of the sensor values. Running summaries of sensor values are often useful in predicting equipment failure. This can also be used to impute the values in empty spaces. For example, if a temperature gauge indicates a machine is warmer than average for the last five days, it may mean something is wrong.

Like in this article we can use idea of feature window, here feature window of last 21 days was taken. We can take feature window of last 21 batches and train it on all the parameters including Program Number, Phase using One-hot-encoding. This field will be called "TIME_SINCE_START" Also, create a variable called "too_soon." When "too_soon" is equal to 1, we have less than 21 days (feature_window) of history for the machine. We will use these new variables to create a running mean, median, max, and min.

We can also -

- Create a running mean, max, min, and median for the sensor variables.
- Another useful transformation is to look for sudden spikes in sensor values. Code creates a value indicating how far the current value is from the immediate norm.
- Create a separate data frame for the training data. We will use this data set to build the model.
- Create a separate data frame for the training and testing data sets. We will use this to tweak our modeling results.

Reference - https://medium.com/swlh/machine-learning-for-equipment-failure-prediction-and-predictive-maintenance-pm-e72b1ce42da1

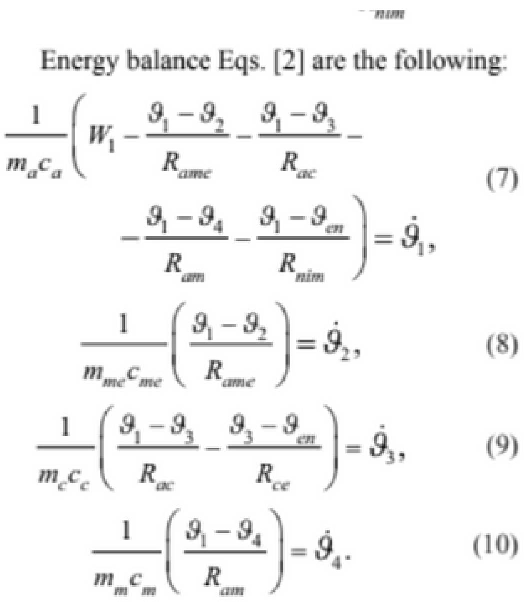
3.2 **Research Papers**

3.2.1 Mathematical modelling of Autoclave

The purpose of this research paper is to build a mathematical model with which the behavior of the processes can be simulated and the **temperature and pressure control in the autoclave can be improved.** This draws a relation between temperature, pressure which should be **different for each program no.** in our case, this modelling should be done properly for each program no. to decrease Failure of Batches.

- The mathematical model is built on the basis of the heat-transfer and pressure-changing theories.
- It aims to improve temperature and pressure control in autoclaves through a mathematical model based on heat-transfer and pressure-changing theories.

- The purpose of the modelling is to simulate and improve the temperature and pressure control in the autoclave, and to test advanced uni- and multi-variable control algorithms.
- The model parameters are obtained from the autoclave manufacturer, the literature, or the model's response fitting to the measured data using the criterion function of the sum of squared errors



$_{nim}$

Energy balance Eqs. [2] are the following:

$$\frac{1}{m_a c_a}\left(W_1 - \frac{\vartheta_1-\vartheta_2}{R_{ame}} - \frac{\vartheta_1-\vartheta_3}{R_{ac}} - \frac{\vartheta_1-\vartheta_4}{R_{am}} - \frac{\vartheta_1-\vartheta_{en}}{R_{nim}}\right) = \dot{\vartheta}_1, \qquad (7)$$

$$\frac{1}{m_{me} c_{me}}\left(\frac{\vartheta_1-\vartheta_2}{R_{ame}}\right) = \dot{\vartheta}_2, \qquad (8)$$

$$\frac{1}{m_c c_c}\left(\frac{\vartheta_1-\vartheta_3}{R_{ac}} - \frac{\vartheta_3-\vartheta_{en}}{R_{ce}}\right) = \dot{\vartheta}_3, \qquad (9)$$

$$\frac{1}{m_m c_m}\left(\frac{\vartheta_1-\vartheta_4}{R_{am}}\right) = \dot{\vartheta}_4. \qquad (10)$$

Fig 4: Energy balance equation

To build this mathematical model of an autoclave considers several key parameters to simulate and improve the temperature and pressure control within the autoclave like -

- Geometry and Structure - cylindrical structure, composite semi-stable form
- Processes - Heating, cooling and Pressure changes
- Heat Transfer Phenomena including conduction and convection
- Data and Parameters - used for model fitting

Reference - https://drive.google.com/file/d/1YFcBRCveGwEoMj5nNEabkA2lBijriEui/view

3.2.2 PREDICTIVE MAINTENANCE IMPLEMENTATION OF AUTOCLAVE REACTOR AGITATOR

The document is a journal article that discusses the implementation of predictive maintenance on a machine used in the food industry to produce sweeteners from tapioca and corn flour.

The authors use OEE (Overall Equipment Effectiveness) as a measure of machine performance and effectiveness, and identify the factors that affect it which are the six big losses to analyze the losses that occur in the machine operation, such as equipment failure, setup and adjustment, reduced speed, idle and minor stoppage, and rework

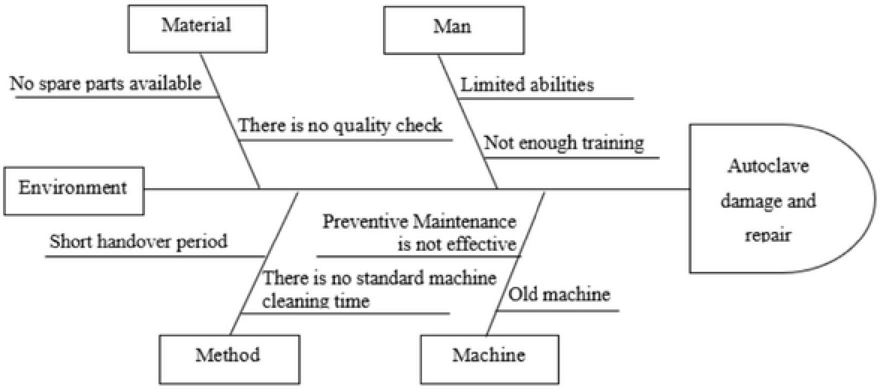

Material
Man
No spare parts available
Limited abilities
There is no quality check
Not enough training
Environment
Autoclave damage and repair
Preventive Maintenance is not effective
Short handover period
There is no standard machine cleaning time
Old machine
Method
Machine

Fig 5 : Fishbone diagram

We can take some hints from this mathematical modelling and use it in our paper like using the Overall Equipment efficiency and also including the factors for failure listed in the Fishbone diagram above

Reference - https://ejournal.itn.ac.id/index.php/JSTAS/article/download/3284/2968/

4. Algorithms of Machine learning to be used

4.1 **Decision Trees**

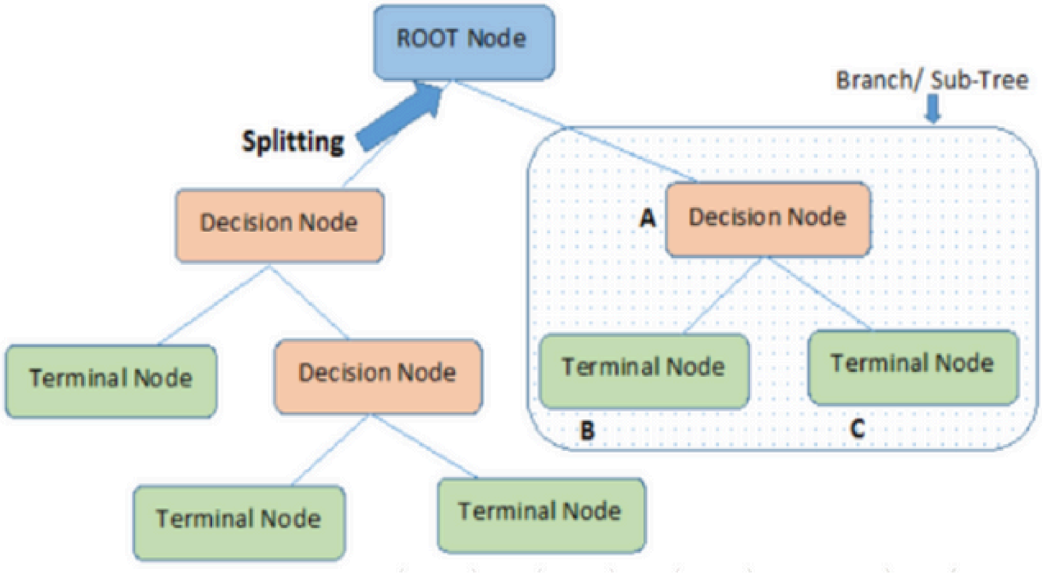


Fig 6 : Decision Tree Classifier

Decision Tree algorithm works in simple steps -

1. Starting at the Root: The algorithm begins at the top, called the "root node," representing the entire dataset.
2. Asking the Best Questions: It looks for the most important feature or question that splits the data into the most distinct groups. This is like asking a question at a fork in the tree.
3. Branching Out: Based on the answer to that question, it divides the data into smaller subsets, creating new branches. Each branch represents a possible route through the tree.
4. Repeating the Process: The algorithm continues asking questions and splitting the data at each branch until it reaches the final "leaf nodes," representing the predicted outcomes or classifications.

https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/

https://colab.research.google.com/drive/1YWqUCkgKeQHXeoY9Dd8ZQF2g-icvBr0d?usp=sharing

4.2 Random Forest

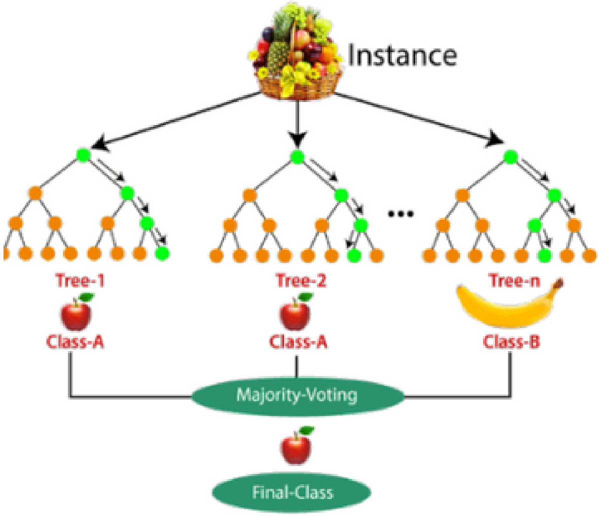


Fig 7: Random Forest Example

A Random Forest is like a group decision-making team in machine learning. It combines the opinions of many “trees” (individual models) to make better predictions, creating a more robust and accurate overall model.

It is an ensemble technique which works on principle of combining multiple models rather than prediction from single model.

It takes into account multiple decision trees from subset of sample and based on majority to select the final predicted value.

https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

https://colab.research.google.com/drive/1Adj2s8A-xKHPBIgubLVj-G_UeEabFXKr?usp=sharing

4.3 LSTM

Long Short-Term Memory Networks is a deep learning, sequential neural network that allows information to persist. It excels at capturing long-term dependencies, making it ideal for sequence prediction tasks.

Unlike traditional neural networks, LSTM incorporates feedback connections, allowing it to process entire sequences of data, not just individual data points. This makes it highly effective in understanding and predicting patterns in sequential data like time series model which is our model in this case. We can use this model and train it on million rows in our dataset

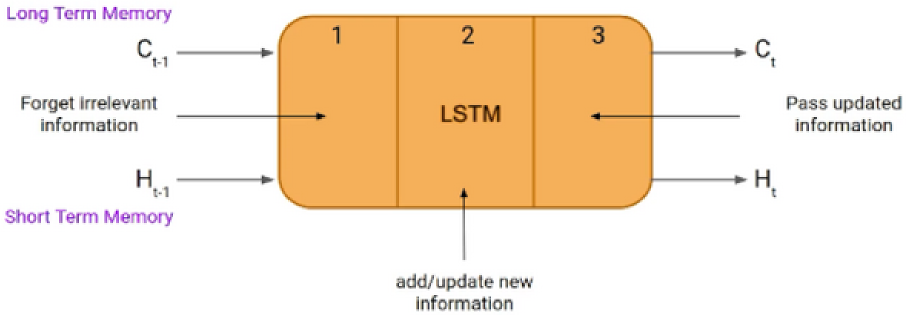


Fig 8: LSTM overview diagram

4.4 Logistic Regression

The machine learning algorithm accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false.

Logical regression analyzes the relationship between one or more independent variables and classifies data into discrete classes. It is extensively used in predictive modeling, where the model estimates the mathematical probability of whether an instance belongs to a specific category or not.

$$f(x) = \frac{1}{1 + e^{-x}}$$

Equation of Logistic Regression

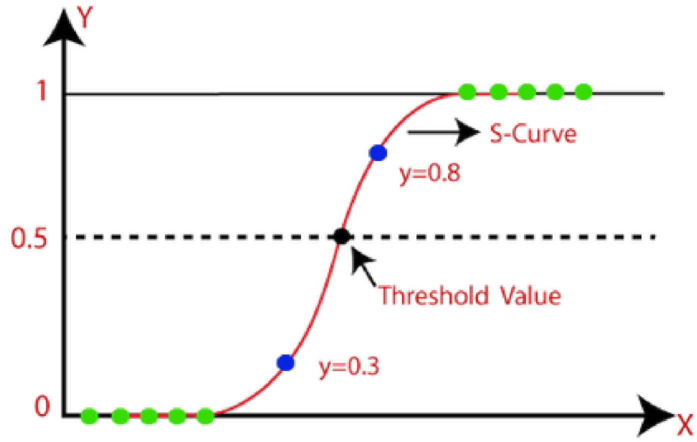


Fig 9: Logistic Regression Example

The threshold of Logistic regression which is by default 0.5 can be changed to classify the data and maximize the accuracy of the function.

https://www.geeksforgeeks.org/understanding-logistic-regression/

5. Model Making Approach

5.1 **Parameters affecting the predictive maintenance model**

- The main parameters that affect the model are **Temperature, pressure** and **time of each cycle** based on the program number and any deviation beyond a range of set point values can be reason for failure of the batch.
- Type and volume of materials sterilized. The **sterilization cycle used** should be in accordance with the material in the autoclave.
- **Operational Data: Frequency of use, cycle interruptions, variances in cycle parameters, details of maintenance activity and part replacement.**
- Other reason for failure might be in the machine operation, setup and adjustment, reduced speed, idle and minor stoppage, and rework or due to Human error.
- Apart from this **temperature gradient** while heating and cooling respectively has to be monitored. It shouldn’t be rapid as it can cause thermal stress on materials.

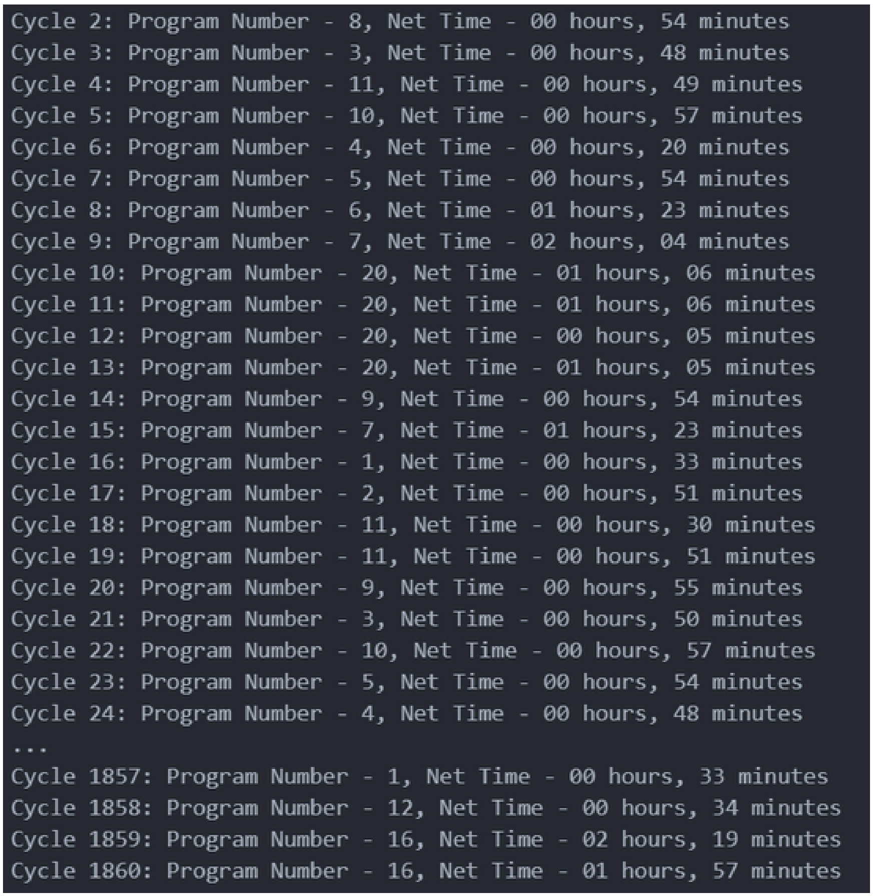
5.2 **Exploratory Data Analysis**

5.2.1 Batches and Phases

The number of Total Batches in the dataset with million rows are - 1860 Batches
Number of Batches passed: 1754
Number of Batches failed: 106

Total number of Phases - 35
These phases are same for same Program number and change across different program numbers unless a batch has failed before reaching the end Phase the count of Phases is same.
['PREPARE AUTOCLAVE' 'DEPRESSURIZE BY VACUUM PUMP' 'DYNAMIC STEAM RISING PULSE' 'FALLING PULSE' 'HEATING' 'EMERGENCY' 'STERILIZATION' 'TIMED VACUUM' 'CYCLE END' 'AIR TO CHAMBER' 'RETURN TO ATMOSPHERIC PRESSURE' 'TIMED VACUUM, STEAM INJECTION' 'AIR DISCHARGE BY GRAVITY' 'PRESSURIZE CHAMBER BY AIR' 'CONTROLLED RATE COOLING' 'COOLING EXTENSION' 'WATER DRAIN' 'REDUCED RATE DEPRESSURIZE' 'NORMAL RATE DEPRESSURIZE' 'MODULATED STEAM RISING PULSE' 'DYNAMIC STEAM PRESSURE HOLD' 'FALLING PULSE SLOW' 'FALLING PULSE NORMAL' 'STEAM TO CHAMBER, MODULATED' 'MODULATED VACUUM BALANCE' 'DEPRESSURIZE BY' 'CHAMBER VACUUM' 'RETURN TO' 'PRESSURIZE CHAMBER' 'PRESSURE STABILIZATION' 'CHAMBER PRESSURE' 'MODULATED DEPRESSURIZATION' 'RISING AIR PULSE' 'AIR PRESSURE HOLD' 'MODULATED FALLING PULSE']

5.2.2 Program No. vs Cycle Time Analysis



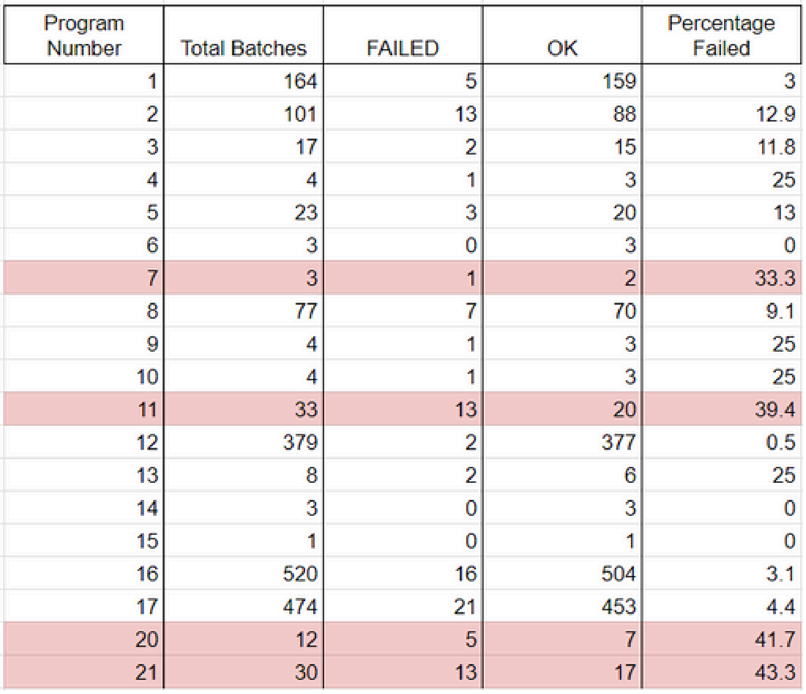
```
Cycle 2: Program Number - 8, Net Time - 00 hours, 54 minutes
Cycle 3: Program Number - 3, Net Time - 00 hours, 48 minutes
Cycle 4: Program Number - 11, Net Time - 00 hours, 49 minutes
Cycle 5: Program Number - 10, Net Time - 00 hours, 57 minutes
Cycle 6: Program Number - 4, Net Time - 00 hours, 20 minutes
Cycle 7: Program Number - 5, Net Time - 00 hours, 54 minutes
Cycle 8: Program Number - 6, Net Time - 01 hours, 23 minutes
Cycle 9: Program Number - 7, Net Time - 02 hours, 04 minutes
Cycle 10: Program Number - 20, Net Time - 01 hours, 06 minutes
Cycle 11: Program Number - 20, Net Time - 01 hours, 06 minutes
Cycle 12: Program Number - 20, Net Time - 00 hours, 05 minutes
Cycle 13: Program Number - 20, Net Time - 01 hours, 05 minutes
Cycle 14: Program Number - 9, Net Time - 00 hours, 54 minutes
Cycle 15: Program Number - 7, Net Time - 01 hours, 23 minutes
Cycle 16: Program Number - 1, Net Time - 00 hours, 33 minutes
Cycle 17: Program Number - 2, Net Time - 00 hours, 51 minutes
Cycle 18: Program Number - 11, Net Time - 00 hours, 30 minutes
Cycle 19: Program Number - 11, Net Time - 00 hours, 51 minutes
Cycle 20: Program Number - 9, Net Time - 00 hours, 55 minutes
Cycle 21: Program Number - 3, Net Time - 00 hours, 50 minutes
Cycle 22: Program Number - 10, Net Time - 00 hours, 57 minutes
Cycle 23: Program Number - 5, Net Time - 00 hours, 54 minutes
Cycle 24: Program Number - 4, Net Time - 00 hours, 48 minutes
...
Cycle 1857: Program Number - 1, Net Time - 00 hours, 33 minutes
Cycle 1858: Program Number - 12, Net Time - 00 hours, 34 minutes
Cycle 1859: Program Number - 16, Net Time - 02 hours, 19 minutes
Cycle 1860: Program Number - 16, Net Time - 01 hours, 57 minutes
```

Fig 10: Program no. vs Cycle time

Failed Batches are generally seen to have smaller Batch times, the reason might be worker's lethargic behavior or even problem in Autoclave reactor

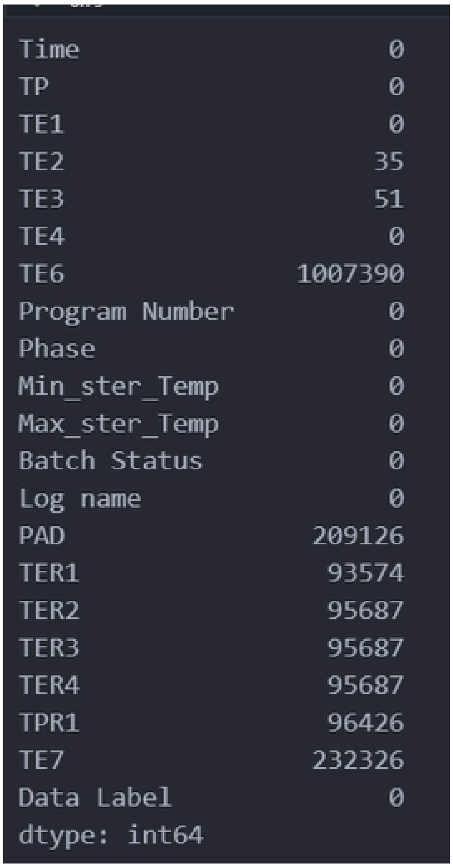
5.2.3 Program No. and Batch Failure analysis

A concerning fact here is that for **Program number 11, 20 and 21** the failure rates are very high. The modelling of Temperature and Pressure might not have been done properly so the company should look at this data again



Program Number	Total Batches	FAILED	OK	Percentage Failed
1	164	5	159	3
2	101	13	88	12.9
3	17	2	15	11.8
4	4	1	3	25
5	23	3	20	13
6	3	0	3	0
7	3	1	2	33.3
8	77	7	70	9.1
9	4	1	3	25
10	4	1	3	25
11	33	13	20	39.4
12	379	2	377	0.5
13	8	2	6	25
14	3	0	3	0
15	1	0	1	0
16	520	16	504	3.1
17	474	21	453	4.4
20	12	5	7	41.7
21	30	13	17	43.3

5.2.4 Imputing values and Idea of Moving averages



```
Time                   0
TP                     0
TE1                    0
TE2                   35
TE3                   51
TE4                    0
TE6              1007390
Program Number         0
Phase                  0
Min_ster_Temp          0
Max_ster_Temp          0
Batch Status           0
Log name               0
PAD               209126
TER1               93574
TER2               95687
TER3               95687
TER4               95687
TPR1               96426
TE7               232326
Data Label             0
dtype: int64
```

Fig 11: Imputing null values with avg

Imputing null values and dropping columns with too many null values for model to perform accurately

To achieve the goal of replacing missing values in TE1 and TE4 with the average of TE2 and TE3, we can adopt two strategies.

Firstly, we can directly compute the average of TE2 and TE3 and fill the null values in TE1 and TE4 accordingly.

Secondly, we can take a more dynamic approach by computing moving averages based on the Program Number and Phase of each batch. This involves grouping the data by Program Number and Phase, and then calculating the rolling mean to impute missing values. By incorporating the characteristics of each batch, such as the Program Number and Phase, we can capture more nuanced trends in the data, potentially leading to more accurate imputations.

5.2.5 Run Length Analysis

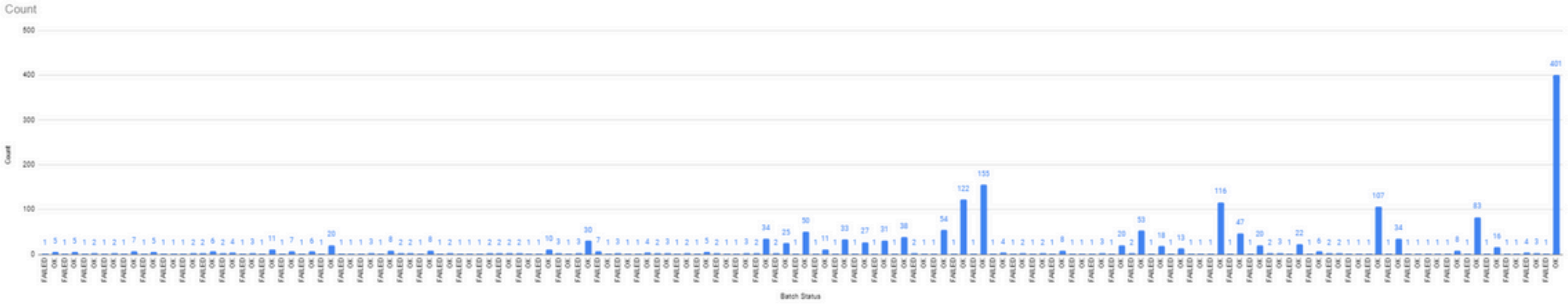


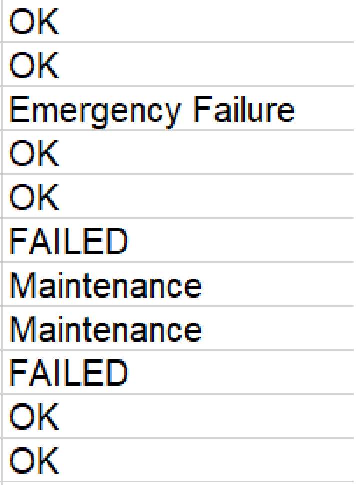
Fig 12: Run length of OK and failed batches

- The data is from 01-09-2019 to 04-05-2021. There has been no failed batch from 29-12-2020. All the batches (401 in total) are passed batches. This would hamper the model a lot, as suddenly the batches have stopped failing.
- Maybe the reason is Preventive maintenance.
- For other values, we can first find the parameters that majorly affect the batches, if there is some significant difference in temperatures, pressure of various sensors.
- This depends on cycle time, program no., temperatures, pressure data.
- An idea is there to take a frequency window of batches lets say last 21 batches and put that in LSTM model to get whether the next batch is going to fail or not.

5.2.6 New Dataset discovery for Sterilization Phase

Sterilization Phase is the most important phase, batches in which sterilization phase in an Autoclave process. If Sterilization Phase is not there in a batch, it is labelled as maintenance if the Batch is OK, else if the Batch in which Sterilization Phase is not there and Batch status is FAILED, it is labelled as Emergency Failure.

The other batches in which sterilization phase is present are labelled as OK and FAILED based on passing of the batch.

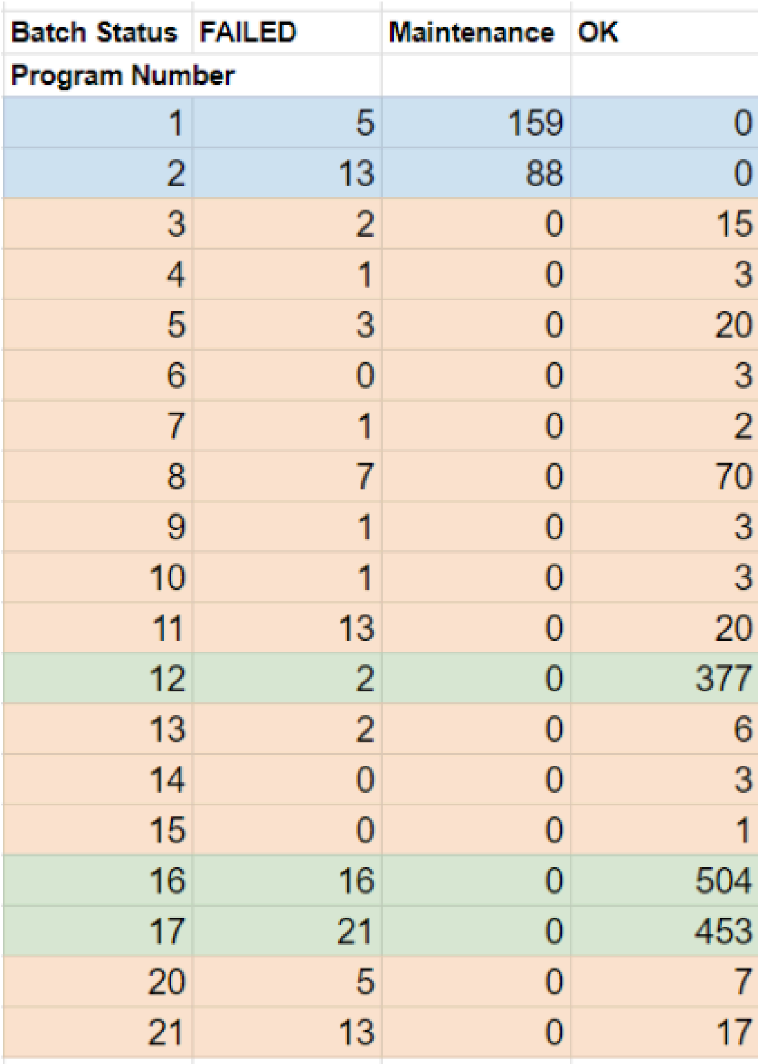


OK
OK
Emergency Failure
OK
OK
FAILED
Maintenance
Maintenance
FAILED
OK
OK

Fig 13: OK and FAILED batches run length

5.2.6 Program no. divided into categories

The data highlighted below will be used for Coin Toss model, divided into three categories for different values of variable a by which the probability decreases for a type of batch. There is a different label for the maintenance batch, different label for for the program no. that have high failure rtates and other with low failure rates.



Batch Status	FAILED	Maintenance	OK
Program Number			
1	5	159	0
2	13	88	0
3	2	0	15
4	1	0	3
5	3	0	20
6	0	0	3
7	1	0	2
8	7	0	70
9	1	0	3
10	1	0	3
11	13	0	20
12	2	0	377
13	2	0	6
14	0	0	3
15	0	0	1
16	16	0	504
17	21	0	453
20	5	0	7
21	13	0	17

6. Coin toss model

6.1 Overview of the Model

In the coin toss model we compare the batch status of the autoclave to probability of getting suppose heads in the next event of tossing a biased coin. The model gives a probability to classify batches as OK, FAILED based on historical data. The initial probability for batch status to be OK is set to p, it then decreases by alpha times for the next run. Whenever the status would be fail, probability would be 1-(alpha*probability) of OK status of the previous batch).

When maintenance is performed, the probability is reset to p.

The results can be determined by seeing the probability of OK status for the next batch.If it lies within the threshold of passing then the machine can be used else maintenance may be needed.

Here, p is probability of autoclave not failing and alpha is the factor of decrease in probability of success after each trial.

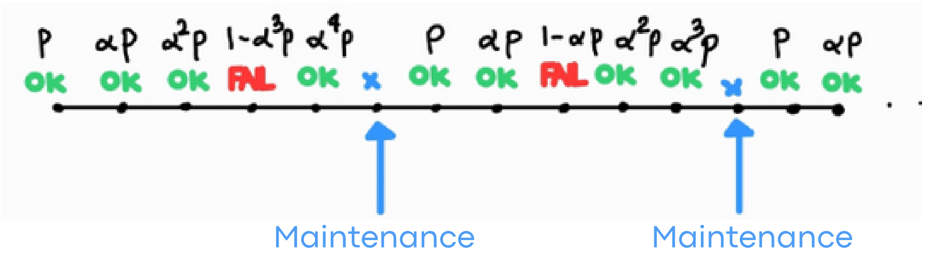


Fig 14: Probability of each batch using MLE to maximize

6.2 Applying model on entire dataset

Model Training:

The training dataset is utilized to estimate the optimal parameters for the predictive model. Parameters include the initial probability (p) and the decay rate (alpha) for different program categories. Negative log-likelihood is employed as the optimization metric, computed over a range of parameter values (p and alpha).

Model Evaluation:

- A threshold value was selected to classify batch statuses based on their probabilities.
- The model's performance evaluated using the test dataset to assess the alignment between predicted and actual batch statuses based on the threshold and based on this the accuracy can be found for which good amount of True Negatives are identified correctly.

Results:

On applying the model on entire dataset together at once the results obtained were:

- Minimum Negative log-likelihood: 359.3315898567369
- Best Parameters (p,a): (0,950000000001, 0.99) (a=alpha)

- The heat map of given probability distribution is shown as below -

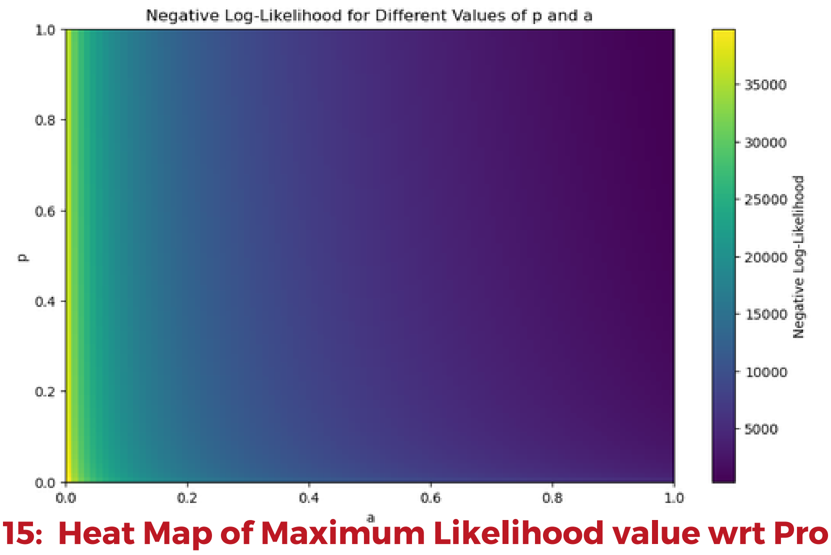
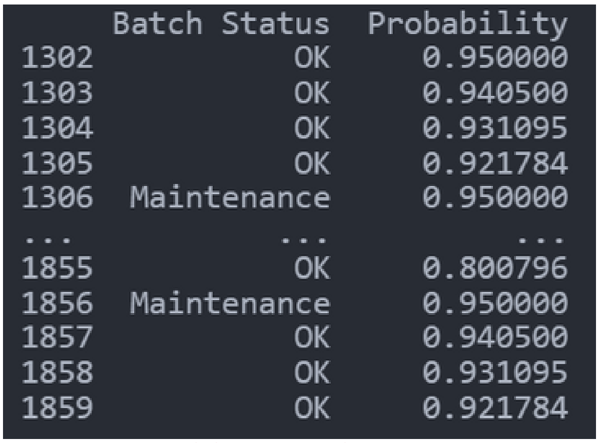


Fig 15: Heat Map of Maximum Likelihood value wrt Probability

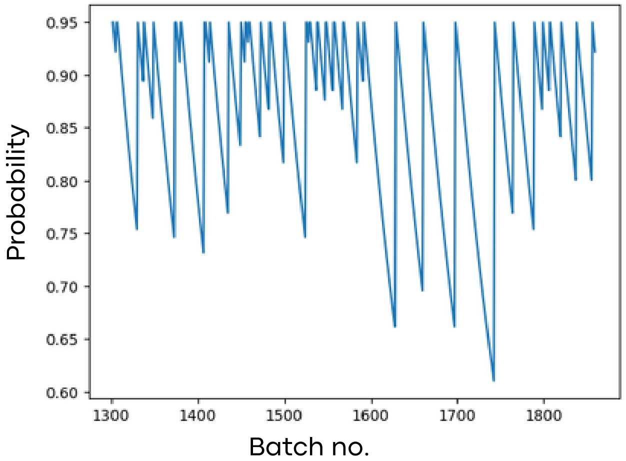
- Testing the model on test data to get the probabilities and from those probabilities we can keep a threshold to find the accuracy -

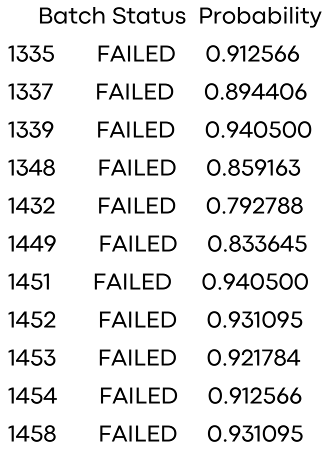


```
     Batch Status  Probability
1302           OK     0.950000
1303           OK     0.940500
1304           OK     0.931095
1305           OK     0.921784
1306  Maintenance     0.950000
...           ...          ...
1855           OK     0.800796
1856  Maintenance     0.950000
1857           OK     0.940500
1858           OK     0.931095
1859           OK     0.921784
```

Fig 16: Probability of Failed and OK Batches

- The probability distribution obtained is shown below:





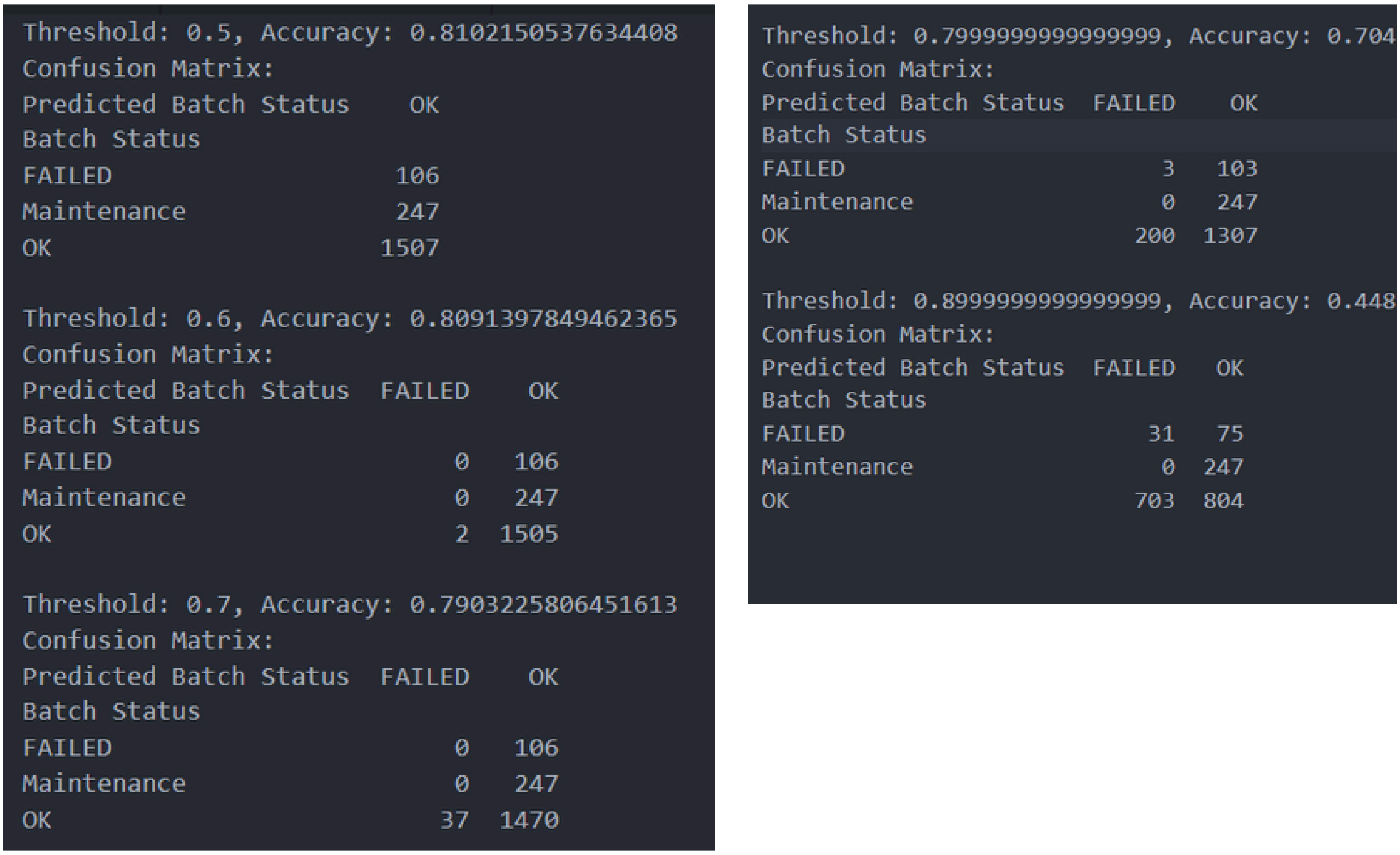
	Batch Status	Probability
1335	FAILED	0.912566
1337	FAILED	0.894406
1339	FAILED	0.940500
1348	FAILED	0.859163
1432	FAILED	0.792788
1449	FAILED	0.833645
1451	FAILED	0.940500
1452	FAILED	0.931095
1453	FAILED	0.921784
1454	FAILED	0.912566
1458	FAILED	0.931095

Fig 17: Line chart of Probability

Fig 18: Failed Batches in last 500 batches

As the no. of failed batches in last 500 batches is very less with no batch failed from 1460 to 1860 Batch no., the data is much more biased and will affect the threshold probability value

Other idea was to test the best parameters on entire dataset -
Varying the threshold between 0.5 and 1 to maximize the accuracy. The result can be seen below of confusion matrix at different threshold values -



```
Threshold: 0.5, Accuracy: 0.8102150537634408
Confusion Matrix:
Predicted Batch Status    OK
Batch Status
FAILED                   106
Maintenance              247
OK                      1507

Threshold: 0.6, Accuracy: 0.8091397849462365
Confusion Matrix:
Predicted Batch Status  FAILED    OK
Batch Status
FAILED                       0   106
Maintenance                  0   247
OK                           2  1505

Threshold: 0.7, Accuracy: 0.7903225806451613
Confusion Matrix:
Predicted Batch Status  FAILED    OK
Batch Status
FAILED                       0   106
Maintenance                  0   247
OK                          37  1470
```

```
Threshold: 0.7999999999999999, Accuracy: 0.704
Confusion Matrix:
Predicted Batch Status  FAILED    OK
Batch Status
FAILED                       3   103
Maintenance                  0   247
OK                         200  1307

Threshold: 0.8999999999999999, Accuracy: 0.448
Confusion Matrix:
Predicted Batch Status  FAILED   OK
Batch Status
FAILED                      31   75
Maintenance                  0  247
OK                         703  804
```

Fig 19: Changing threshold to maximize accuracy of Coin toss model

6.3 Year-wise cointoss model for same a and p

Obtimization:

The current values of a and p were evaluated by taking entire data together. Better parameters may be obtained when we segregate the data. We have segregated the data on the basis of year and then on the basis of program number.

1.For the year 2019,2020 and 2021, the values of best parameters obtained is as shown below:

Year: 2019 Best Parameters (p a): (0.8300000000000001 0.99)
Year: 2020 Best Parameters (p a): (0.97 0.99)
Year: 2021 Best Parameters (p a): (0.99 0.99)

The probability p gradually keeps on increasing as the years pass by because of the no. of Failed batches keeps on decreasing.

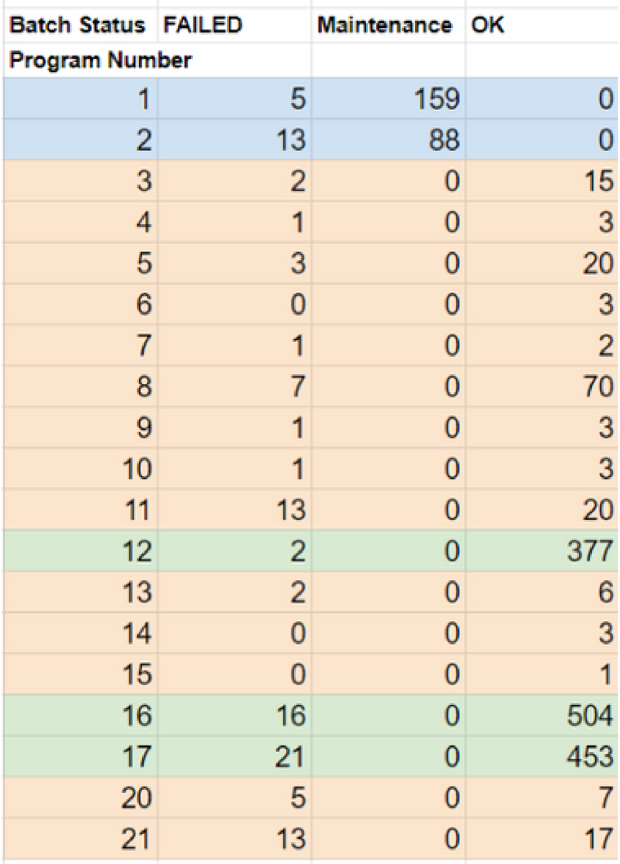
6.4 Cointoss model with a values changing with program no.

2. On the basis of program number, the data was divided into three groups: first with Program Number 1 and 2(group M), second with Program Number 12,16 and 17(group B) and third group consisting the rest (group A) as shown in the table which was shown above and reproduced below.

The results obtained on such division of the data is:

Minimum Negative Log-Likelihood: **359.3315898567369**

Best Parameters: (p, a_values): (**0.9500000000000001**, {'A': **0.99**, 'B': 0.99})



Batch Status	FAILED	Maintenance	OK
Program Number			
1	5	159	0
2	13	88	0
3	2	0	15
4	1	0	3
5	3	0	20
6	0	0	3
7	1	0	2
8	7	0	70
9	1	0	3
10	1	0	3
11	13	0	20
12	2	0	377
13	2	0	6
14	0	0	3
15	0	0	1
16	16	0	504
17	21	0	453
20	5	0	7
21	13	0	17

6.5 Year-wise Cointoss model with a values changing with program no.

3. Doing year-wise calculation for p, a1, a2 and getting the optimum value of these values for every year so that we can look at how improvements in autoclave has resulted in increasing of probability p.

The results obtained on such division of the data year-wise is:

Year: 2019, Best Parameters (p, a_values): (0.82, {'A': 0.97, 'B': 0.99})

Year: 2020, Best Parameters (p, a_values): (0.96, {'A': 0.99, 'B': 0.99})

Year: 2021, Best Parameters (p, a_values): (0.99, {'A': 0.99, 'B': 0.99})

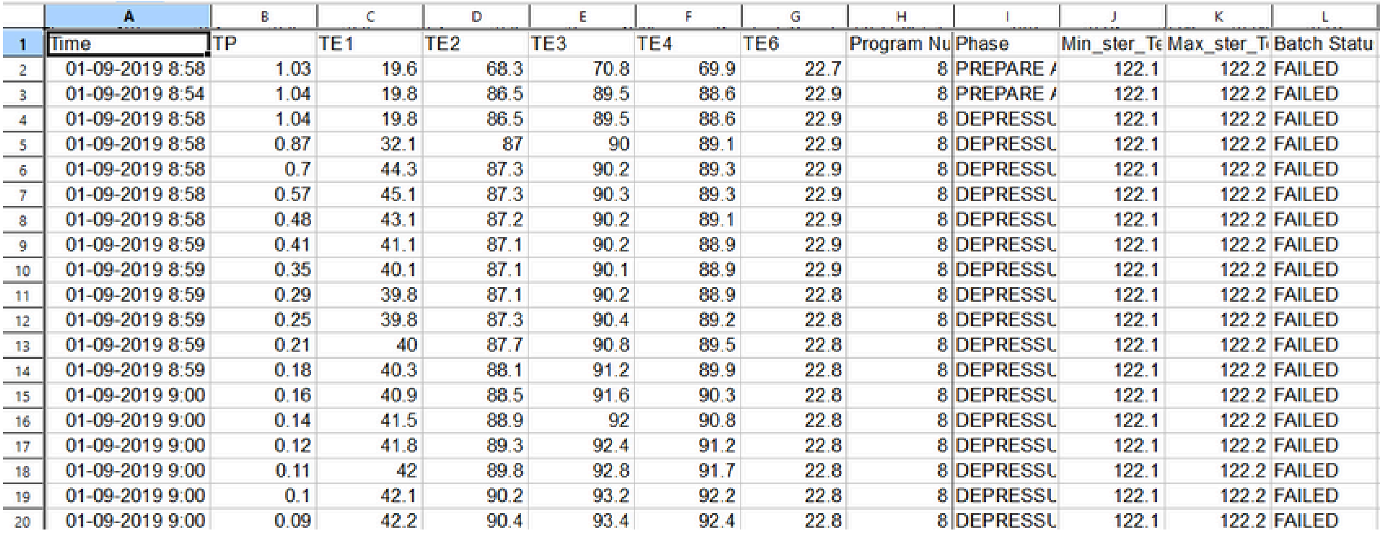
We can keep a threshold on this and test the year-wise accuracy like we have done in the normal case. We can get the probability of Batch status on entire dataset using these valued and then keep a threshold to maximize the values.

7. Modifying Dataset for Model

7.1 Changing dataset to be appropriate input to model

7.1.1 Changing each Batch to a single row

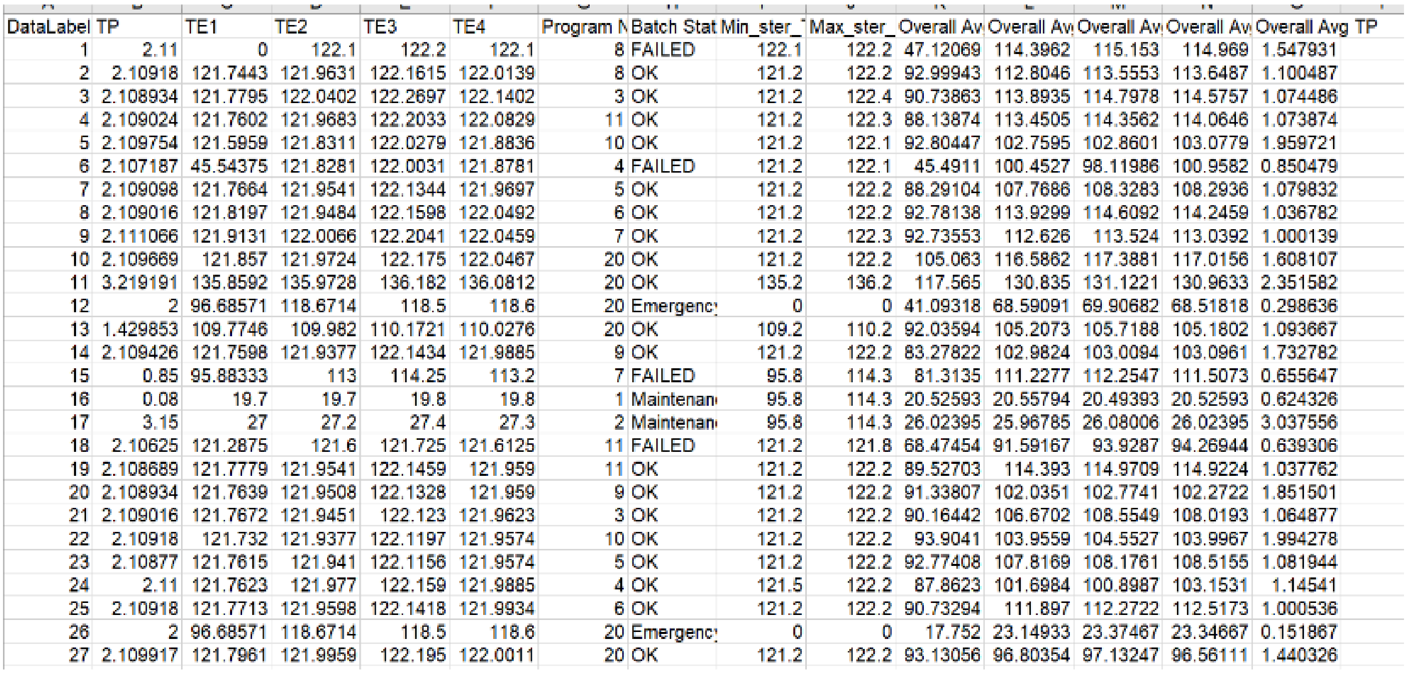
The provided data looks as shown below:



	A	B	C	D	E	F	G	H	I	J	K	L
1	Time	TP	TE1	TE2	TE3	TE4	TE6	Program Nu	Phase	Min_ster_Te	Max_ster_Te	Batch Statu
2	01-09-2019 8:58	1.03	19.6	68.3	70.8	69.9	22.7	8	PREPARE /	122.1	122.2	FAILED
3	01-09-2019 8:54	1.04	19.8	86.5	89.5	88.6	22.9	8	PREPARE /	122.1	122.2	FAILED
4	01-09-2019 8:58	1.04	19.8	86.5	89.5	88.6	22.9	8	DEPRESSU	122.1	122.2	FAILED
5	01-09-2019 8:58	0.87	32.1	87	90	89.1	22.9	8	DEPRESSU	122.1	122.2	FAILED
6	01-09-2019 8:58	0.7	44.3	87.3	90.2	89.3	22.9	8	DEPRESSU	122.1	122.2	FAILED
7	01-09-2019 8:58	0.57	45.1	87.3	90.3	89.3	22.9	8	DEPRESSU	122.1	122.2	FAILED
8	01-09-2019 8:58	0.48	43.1	87.2	90.2	89.1	22.9	8	DEPRESSU	122.1	122.2	FAILED
9	01-09-2019 8:59	0.41	41.1	87.1	90.2	88.9	22.9	8	DEPRESSU	122.1	122.2	FAILED
10	01-09-2019 8:59	0.35	40.1	87.1	90.1	88.9	22.9	8	DEPRESSU	122.1	122.2	FAILED
11	01-09-2019 8:59	0.29	39.8	87.1	90.2	88.9	22.8	8	DEPRESSU	122.1	122.2	FAILED
12	01-09-2019 8:59	0.25	39.8	87.3	90.4	89.2	22.8	8	DEPRESSU	122.1	122.2	FAILED
13	01-09-2019 8:59	0.21	40	87.7	90.8	89.5	22.8	8	DEPRESSU	122.1	122.2	FAILED
14	01-09-2019 8:59	0.18	40.3	88.1	91.2	89.9	22.8	8	DEPRESSU	122.1	122.2	FAILED
15	01-09-2019 9:00	0.16	40.9	88.5	91.6	90.3	22.8	8	DEPRESSU	122.1	122.2	FAILED
16	01-09-2019 9:00	0.14	41.5	88.9	92	90.8	22.8	8	DEPRESSU	122.1	122.2	FAILED
17	01-09-2019 9:00	0.12	41.8	89.3	92.4	91.2	22.8	8	DEPRESSU	122.1	122.2	FAILED
18	01-09-2019 9:00	0.11	42	89.8	92.8	91.7	22.8	8	DEPRESSU	122.1	122.2	FAILED
19	01-09-2019 9:00	0.1	42.1	90.2	93.2	92.2	22.8	8	DEPRESSU	122.1	122.2	FAILED
20	01-09-2019 9:00	0.09	42.2	90.4	93.4	92.4	22.8	8	DEPRESSU	122.1	122.2	FAILED

It consists of rows of Time, TP, TE1, TE2, TE3, TE4, TE6, Program Number, Phase, Min_ster_Temp, Max_ster_Temp, Batch Status.TP is reading for pressure and TEs represent different temperature ssensors.Batch status is either OK or FAILED.

This data is modified to the following format:



DataLabel	TP	TE1	TE2	TE3	TE4	Program N	Batch Stat	Min_ster_	Max_ster_	Overall Av	Overall Av	Overall Av	Overall Av	Overall Avg TP
1	2.11	0	122.1	122.2	122.1	8	FAILED	122.1	122.2	47.12069	114.3962	115.153	114.969	1.547931
2	2.10918	121.7443	121.9631	122.1615	122.0139	8	OK	121.2	122.2	92.99943	112.8046	113.5553	113.6487	1.100487
3	2.108934	121.7795	122.0402	122.2697	122.1402	3	OK	121.2	122.4	90.73863	113.8935	114.7978	114.5757	1.074486
4	2.109024	121.7602	121.9683	122.2033	122.0829	11	OK	121.2	122.3	88.13874	113.4505	114.3562	114.0646	1.073874
5	2.109754	121.5959	121.8311	122.0279	121.8836	10	OK	121.2	122.1	92.80447	102.7595	102.8601	103.0779	1.959721
6	2.107187	45.54375	121.8281	122.0031	121.8781	4	FAILED	121.2	122.1	45.4911	100.4527	98.11986	100.9582	0.850479
7	2.109098	121.7664	121.9541	122.1344	121.9697	5	OK	121.2	122.2	88.29104	107.7686	108.3283	108.2936	1.079832
8	2.109016	121.8197	121.9484	122.1598	122.0492	6	OK	121.2	122.2	92.78138	113.9299	114.6092	114.2459	1.036782
9	2.111066	121.9131	122.0066	122.2041	122.0459	7	OK	121.2	122.3	92.73553	112.626	113.524	113.0392	1.000139
10	2.109669	121.857	121.9724	122.175	122.0467	20	OK	121.2	122.2	105.063	116.5862	117.3881	117.0156	1.608107
11	3.219191	135.8592	135.9728	136.182	136.0812	20	OK	135.2	136.2	117.565	130.835	131.1221	130.9633	2.351582
12	2	96.68571	118.6714	118.5	118.6	20	Emergency	0	0	41.09318	68.59091	69.90682	68.51818	0.298636
13	1.429853	109.7746	109.982	110.1721	110.0276	20	OK	109.2	110.2	92.03594	105.2073	105.7188	105.1802	1.093667
14	2.109426	121.7598	121.9377	122.1434	121.9885	9	OK	121.2	122.2	83.27822	102.9824	103.0094	103.0961	1.732782
15	0.85	95.88333	113	114.25	113.2	7	FAILED	95.8	114.3	81.3135	111.2277	112.2547	111.5073	0.655647
16	0.08	19.7	19.7	19.8	19.8	1	Maintenan	95.8	114.3	20.52593	20.55794	20.49393	20.52593	0.624326
17	3.15	27	27.2	27.4	27.3	2	Maintenan	95.8	114.3	26.02395	25.96785	26.08006	26.02395	3.037556
18	2.10625	121.2875	121.6	121.725	121.6125	11	FAILED	121.2	121.8	68.47454	91.59167	93.9287	94.26944	0.639306
19	2.108689	121.7779	121.9541	122.1459	121.959	11	OK	121.2	122.2	89.52703	114.393	114.9709	114.9224	1.037762
20	2.108934	121.7639	121.9508	122.1328	121.959	9	OK	121.2	122.2	91.33807	102.0351	102.7741	102.2722	1.851501
21	2.109016	121.7672	121.9451	122.123	121.9623	3	OK	121.2	122.2	90.16442	106.6702	108.5549	108.0193	1.064877
22	2.10918	121.732	121.9377	122.1197	121.9574	10	OK	121.2	122.2	93.9041	103.9559	104.5527	103.9967	1.994278
23	2.10877	121.7615	121.941	122.1156	121.9574	5	OK	121.2	122.2	92.77408	107.8169	108.1761	108.5155	1.081944
24	2.11	121.7623	121.977	122.159	121.9885	4	OK	121.5	122.2	87.8623	101.6984	100.8987	103.1531	1.14541
25	2.10918	121.7713	121.9598	122.1418	121.9934	6	OK	121.2	122.2	90.73294	111.897	112.2722	112.5173	1.000536
26	2	96.68571	118.6714	118.5	118.6	20	Emergency	0	0	17.752	23.14933	23.37467	23.34667	0.151867
27	2.109917	121.7961	121.9959	122.195	122.0011	20	OK	121.2	122.2	93.13056	96.80354	97.13247	96.56111	1.440326

It adds rows of Overall Averages of TE1, TE2, TE3, TE4 and TP. We aren't considering TE5 and TE6 in our modified data as these sensors values aren't available for most of the batches. Batch Status now consists of OK, FAILED, MAINTENACE and EMERGENCY FAILURE. Whenever sterilization phase isn't present in an batch, that batch is assumed to be Maintenance if it's provided status was OK and Emergency Failure if it's provided status was FAILED.

Each Batch has multiple Phases and various Temperature sensors, we have to convert this into a single row to be able to apply the model.

To convert hundred's of rows of each of the Batch into a single row we have used strategy of using Average of Sterilization Phase of 4 sensors, Overall Average of the 4 temperature sensors and Overall Variance of the 4 sensors as shown in the Table below.

7.1.2 Dealing with Average Sterilization Phase temp.

There are 1860 such rows composition of which is given below -

OK	1507
Maintenance	247
FAILED	83
Emergency Failure	23

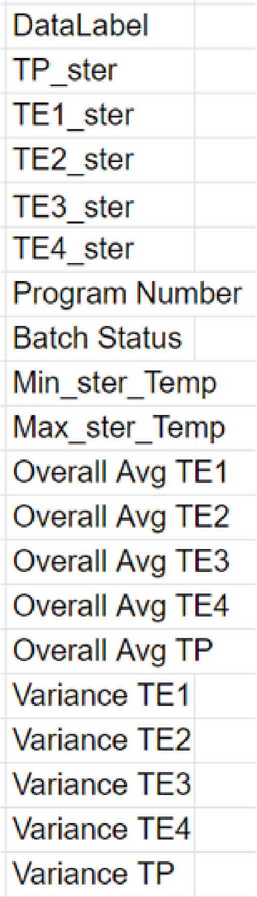
For the OK and FAILED batches we have the value of sterilization Phase temperatures and pressure but for the Maintenance and Emergency Failure Batches we don't have any data for these temperature values.

Note that model cannot take empty values in the dataset.

So instead of leaving those values as Nan we had to fill those values with dummy values.

For **Emergency Failure** values, we replaced them with FAILED values Batch average values.

For the **Maintenance** Batches we have used a standard template of other overall average temperature and pressure values to be replaced in these columns as well.



DataLabel
TP_ster
TE1_ster
TE2_ster
TE3_ster
TE4_ster
Program Number
Batch Status
Min_ster_Temp
Max_ster_Temp
Overall Avg TE1
Overall Avg TE2
Overall Avg TE3
Overall Avg TE4
Overall Avg TP
Variance TE1
Variance TE2
Variance TE3
Variance TE4
Variance TP

Fig 20: Different Variables of new dataset

7.1.2 Converting data to input to the model considering Historical data to predict future Batch status

These rows converted to 1 single as input and Batch status of 21st row is given as output to predict future maintenance even before the Batch is failed. This is the Predictive Maintenance Model we have applied in this report.

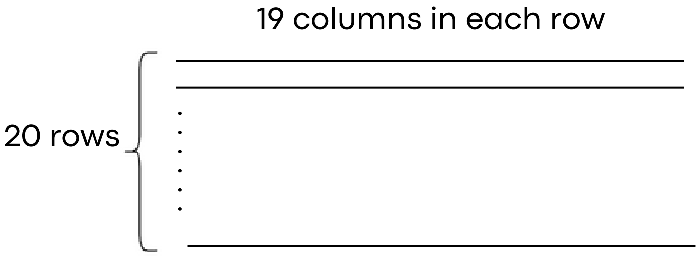


Fig 21: Converting multiple batches into single vector as input

As you will see the no. of Batches to be taken in consideration to predict status of next Batch has been optimized for each model differently

Shape of new vector list formed is (1840, 380).

7.2 Method to not make data biased

7.2.1 Splitting the data into train and test set

Data is split into train and test set such that the testing set also contains 20% of the total failed Batches and 20% of the passed batches. Remaining is given to the train set.

This is done so that their isn't much biasedness towards the passed Batches and to not make the test set biased.

7.2.2 Accuracy of this data on different models when only sterilization Phase averages and Overall Average were included

Accuracy of model with only including averages is for the **decision trees model**

```
Accuracy: 0.7317073170731707
Mean Squared Error: 0.2682926829268293
Classification Report:
              precision    recall   f1-score    support
        0       0.28         0.30      0.29         70
        1      0.83         0.82      0.82          299
```

The Confusion Matrix is -

```
[[ 21  49]
 [ 55 244]]
```

For the **Random Forest model**, the accuracy is -

```
Accuracy: 0.8184281842818428
Mean Squared Error: 0.18157181571815717
          precision    recall   f1-score    support


        0       0.41       0.33      0.37         70
        1      0.85        0.89      0.87         299
```

Confusion Matrix -

```
[[ 23  47]
 [ 33 266]]
```

For the **Logistic Regression model**, the accuracy is -

```
Accuracy: 0.8265582655826558
Mean Squared Error: 0.17344173441734417
```

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


       0      0.57         0.33       0.42        70
       1      0.86         0.94       0.90        299
```

```
Confusion Matrix -
[[ 23  47]
 [ 17 282]]
```

Optimizing the threshold values to maximize accuracy, but do note that False Negatives changes with accuracy and prediction is not that good.

```
Threshold | Accuracy | MSE
-----------------------------------
0.0     | 0.8103   | 0.1897
0.1     | 0.8103   | 0.1897
0.2     | 0.8103   | 0.1897
0.3     | 0.7913   | 0.2087
0.4     | 0.7832   | 0.2168
0.5     | 0.7724   | 0.2276
0.6     | 0.7507   | 0.2493
0.7     | 0.7182   | 0.2818
0.8     | 0.6721   | 0.3279
0.9     | 0.5881   | 0.4119
1.0     | 0.1897   | 0.8103
```

```
Threshold: 0.0
Confusion Matrix:
[[  0  70]
 [  0 299]]

Threshold: 0.1
Confusion Matrix:
[[ 12  58]
...
Threshold: 1.0
Confusion Matrix:
[[ 70   0]
 [299   0]]
```

7.2.3 Accuracy when variance was also included in the above data

Accuracy of decision trees model with only including variances and averages in data for the **decision trees model**

```
Accuracy: 0.7208672086720868
Mean Squared Error: 0.2791327913279133
Classification Report:
            precision   recall  f1-score   support

         0      0.27      0.29      0.28       70
         1      0.83      0.82      0.83       299

The Confusion Matrix is -
[[ 20  50]
 [ 53 246]]
```

For the **Random Forest model**, the accuracy & confusion matrix is -

```
Accuracy: 0.7452574525745257
Mean Squared Error: 0.2547425474254742 5
            precision   recall   f1-score   support

         0     0.41      0.33      0.37        70
         1     0.85      0.89      0.87        299

Confusion Matrix -
[[ 23  47]
 [ 33 266]]
```

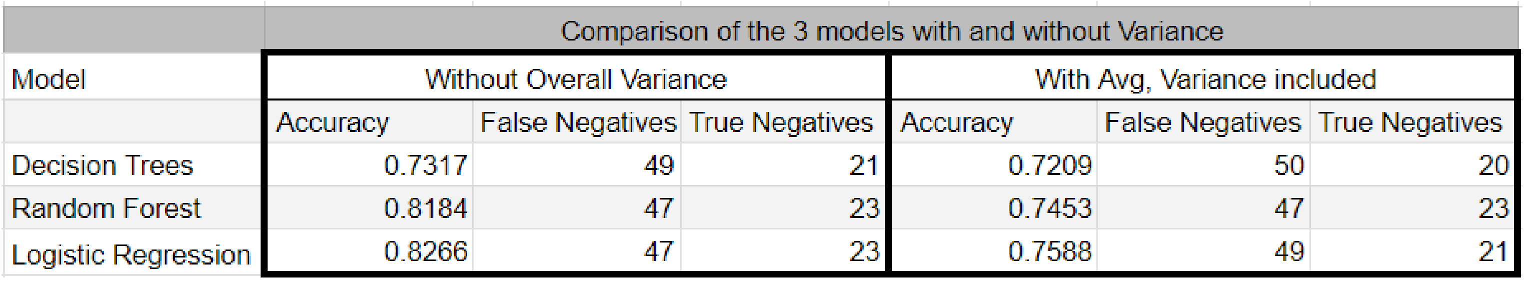
For the **Logistic Regression model**, the result parameters are given below -

```
Accuracy: 0.7588075880758808
Mean Squared Error: 0.24119241192411925

Classification Report:
             precision    recall   f1-score   support

         0      0.34       0.30      0.32       70
         1      0.84       0.87      0.85       299
Confusion Matrix:
[[ 21  49]
 [ 40 259]]
```

7.2.4 Comparison of different models



	Comparison of the 3 models with and without Variance					
Model	Without Overall Variance			With Avg, Variance included		
	Accuracy	False Negatives	True Negatives	Accuracy	False Negatives	True Negatives
Decision Trees	0.7317	49	21	0.7209	50	20
Random Forest	0.8184	47	23	0.7453	47	23
Logistic Regression	0.8266	47	23	0.7588	49	21

7.3 Method to not make data biased using SMOTE

7.3.1 Use SMOTE to do oversampling of Failed batches in train set

Synthetic Minority Oversampling Technique

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class.

Here, we have given 30% of data to test set and remaining to train set

The method automatically makes number of datapoints for Failed and OK batches equal in the dataset and we have applied this to the train set.

Initially the no. of rows in trainset were -

OK	1045
FAILED	242

Resampled class distribution:

OK	1045
FAILED	1045

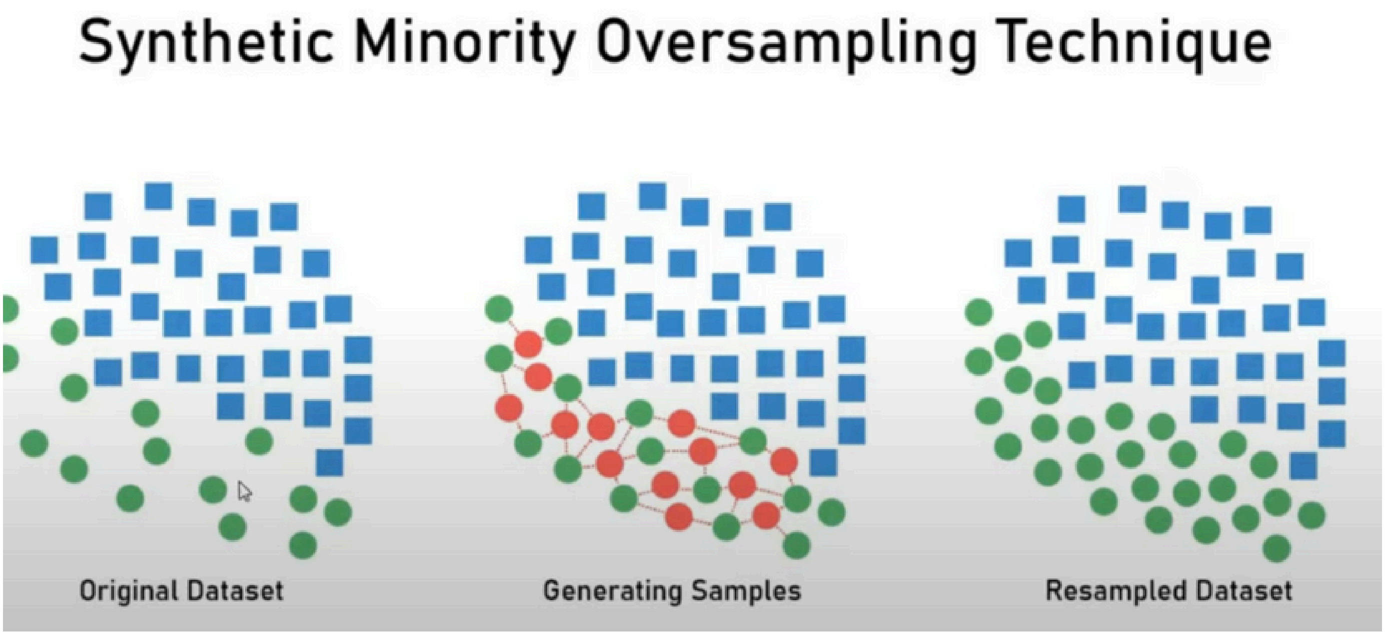


Fig 22: Synthetic Minority Oversampling technique method

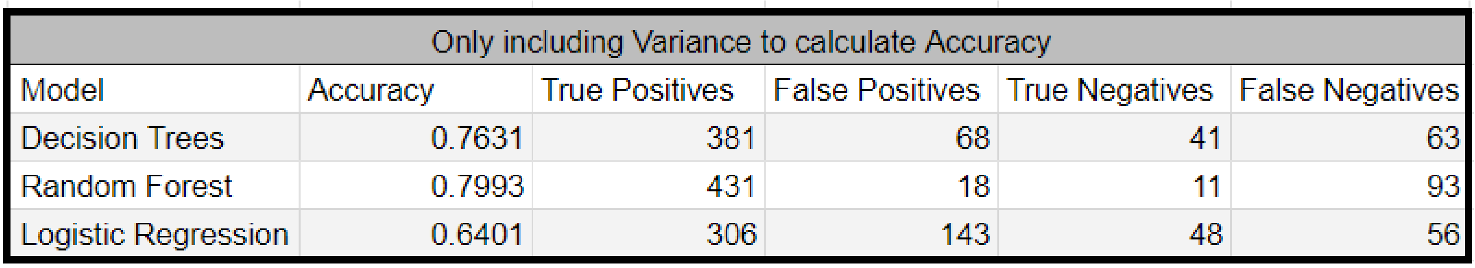
7.3.2 Using SMOTE including only Variance the different parameters for different models are listed below

```
Shape of train and test set is -
Shape of X_train: (1287, 160)
Shape of y_train: (1287,)
Shape of X_test: (553, 160)
Shape of y_test: (553,)
```

```
Shape of train and test set is -
Shape of X_train: (1287, 160)
Shape of y_train: (1287,)
Shape of X_test: (553, 160)
Shape of y_test: (553,)
```

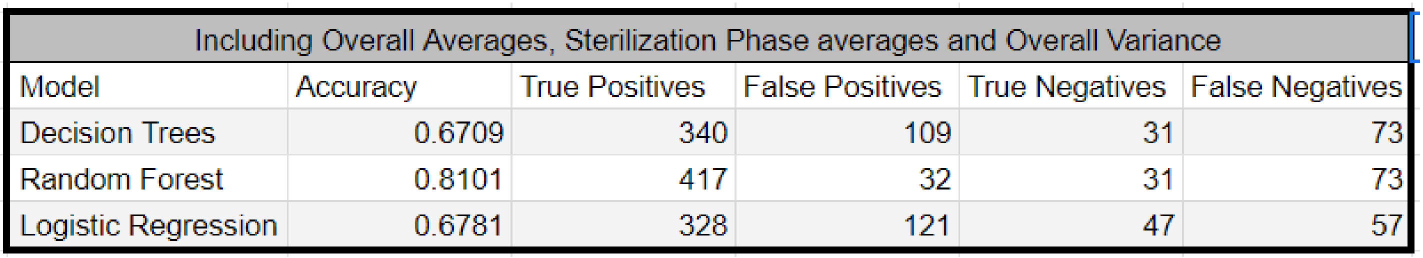
7.3.3 Comparison of Models including and exculding Variance

Results only including the Overall average and Sterilization Phase average to calculate accuracy



Only including Variance to calculate Accuracy					
Model	Accuracy	True Positives	False Positives	True Negatives	False Negatives
Decision Trees	0.7631	381	68	41	63
Random Forest	0.7993	431	18	11	93
Logistic Regression	0.6401	306	143	48	56

Results including the Overall average, Overall Variance and Sterilization Phase average to calculate accuracy

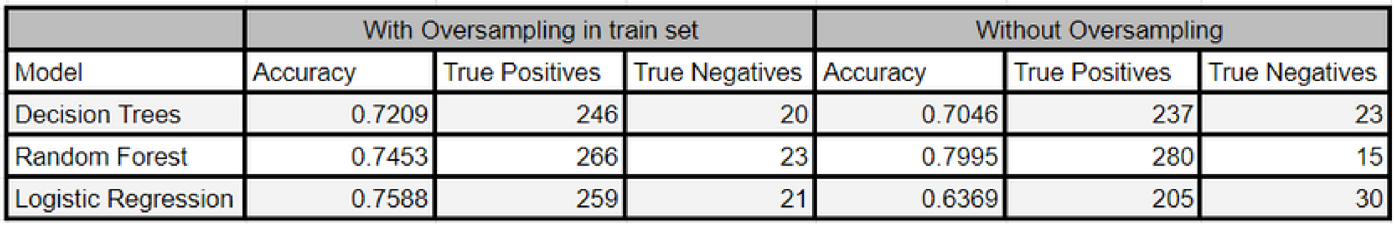


Including Overall Averages, Sterilization Phase averages and Overall Variance					
Model	Accuracy	True Positives	False Positives	True Negatives	False Negatives
Decision Trees	0.6709	340	109	31	73
Random Forest	0.8101	417	32	31	73
Logistic Regression	0.6781	328	121	47	57

These are the results for SMOTE used

It can be seen that accuracy does change by some amount for each of the models when variance are included and not included respectively.

7.3.4 Comparing the Models with and without oversampling



	With Oversampling in train set			Without Oversampling		
Model	Accuracy	True Positives	True Negatives	Accuracy	True Positives	True Negatives
Decision Trees	0.7209	246	20	0.7046	237	23
Random Forest	0.7453	266	23	0.7995	280	15
Logistic Regression	0.7588	259	21	0.6369	205	30

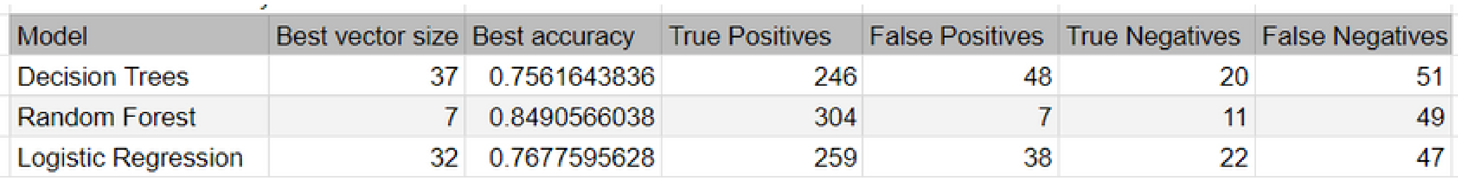
The models have take the Variance data as well into account

7.4 Optimizing no. of previous batches taken into consideration to predict next batch

The no. of previous batches taken into consideration would change with the model and thus no. of previous batches taken into consideration are optimized separately for each model..

The values of previous batch taken into consideration are varied from 5 to 40 in each case and the true negatives have been tried to maximize here with the accuracy so that the model predicts maximum True negatives correctly.
Recognizing the failed batches before it fails is very important for us to do the predictive maintenance.

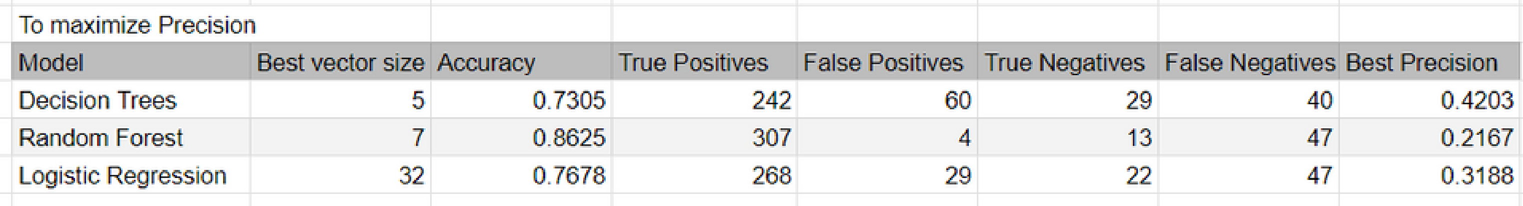
7.4.1 Finding best Vector size for maximum accuracy



Model	Best vector size	Best accuracy	True Positives	False Positives	True Negatives	False Negatives
Decision Trees	37	0.7561643836	246	48	20	51
Random Forest	7	0.8490566038	304	7	11	49
Logistic Regression	32	0.7677595628	259	38	22	47

7.4.2 Finding best Vector size for maximum Precision

Note that here precision is defined as TN/(TN+FN). Here, TN is True Negatives and FN stands for False Negatives.
Here we tried to keep 20% equal in train and test set but the error was not resolvable, so the train-test split is random



To maximize Precision							
Model	Best vector size	Accuracy	True Positives	False Positives	True Negatives	False Negatives	Best Precision
Decision Trees	5	0.7305	242	60	29	40	0.4203
Random Forest	7	0.8625	307	4	13	47	0.2167
Logistic Regression	32	0.7678	268	29	22	47	0.3188

7.4.3 Conclusion

From all the above models it is quite evident that Random Forest fits quite well for the data with oversampled trained data and 20% false values contained in the test data to not make the data biased.

In some cases Logistic Regression is also working good, so the order of accuracy of these models are Random Forest > Logistic Regression > Decision Trees Model.

8. Future Prospects

8.1 **Relating Variance**

8.1.1 Variance of previous batch affecting next batch failure

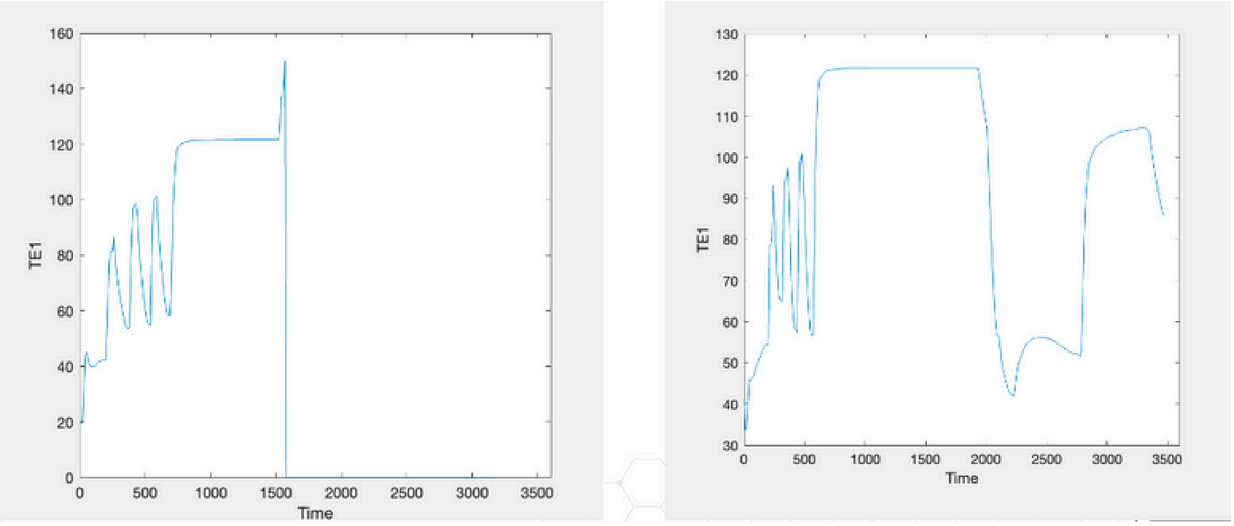
One thing we tried in this model and can be looked at in future work is -
If variance of previous batch is very high, the next batch should ideally fail. We did get some satisfactory results which gives the no. of batches in which Variance in previous batch is above a threshold and the next batch has failed but this has to be optimized.

Now this can be done for various temperature sensors, choosing the best and most appropriate among these is important and a threshold value has to be optimized for this sensor. It is a tedious task to check accuracy by changing the threshold values and optimizing the accuracy but this can be done to just check effect of Variance in previous batch with failure.

This can later be included as one of the main model parameters to be taken into consideration, telling that the previous batch variance depicts future batch failure.

8.1.2 Inclusion of Variance of different phases

Overall variance does give some idea but it isn't that much relevant. There are multiple profiles of temperature and pressure, variance of which can be looked at based on the phases.
If variance for a Phase profile is quite high compared to the average variance in that phase then the next batch is bound to fail.
Example of the same is shown below -



Failed Batch

OK Batch

Fig 23: Failed and OK batches of same program number

At around 1500 s, the Variance in Failed Batch is quite high. Somehow this has to be scaled up for the entire dataset, which in itself is a big task as all phases are not present in every batch.

8.1.3 Using more accurate technique to find failed batches

Instead of just taking average and variance, more accurate ways to detect failed batches have to be found out whether it be variance in previous batch included with probability of failing of next batch.

Variance and average not considering the different phases but at different scaled timepoints can also be a good way to find the values which can be input to the Decision Trees, Random Forest and the Logistic Regression model.

8.2 **Conclusion**

A better way has to be looked at to handle the enormous amount of null values, a rolling mean is a good strategy that can be included if entries in a batch are not entirely NULL values,

By integrating predictive modeling with different data aggregation techniques like including average variance at different timepoints for example, we can develop a holistic approach to predicting future batch failure in autoclave processes.

Somehow we have to further reduce biasedness of the test data to predict the future batch failure, the no. of Failed Batches are quite less and there isn't much difference that can be seen in average values in general of Failed and Passes Batches for different phases, although the overall values differ in both the batches.

Cycle time of FAILED batch is lower as compared to the OK batch which can also be included in the model

★★Thank you ★★