Improve Overall Performace Indicators of The Organization using Data Analysis

Dulan Jayasuriya, *dulan.20@cse.mrt.ac.lk* Chanindu Leelananda, *chanindu.20@cse.mrt.ac.lk*

Department of Computer Science & Engineering
Faculty of Engineering
University of Moratuwa
Sri Lanka

***Abstract*—In a business environment, assuring profitability is a key thing to continuous operation. For the manufacturing industry, maintaining Overall Equipment Efficiency(OEE) is a direct indication of profitability. This project is an initiative to launch Data-driven decision Making(D3M) [1] in a manufacturing environment. Data Analysis has been carried out in Descriptive, Diagnosis and Predictive along with the visualisations. The main focus of the descriptive analysis was to give a better representation of data and Diagnostic analysis was used to analyze the relationship between different features. The predictive analysis was done using different regression models such as Linear Regression, Support Vector Machines, Decision Trees, Random Forest and Neural Networks in order to identify the best method to predict the outcome from the selected feature points. We were able to explain the relationships between features and achieve 80% accuracy with the predictive methods. In this approach for data analysis is initiated in a manufacturing environment, it enables the users to visualise the available data in a useful manner in the meantime analyze data and find root causes for efficiency drops. Also by initiating predictive analysis practices, the organisation can get rid of future unfavourable operations in terms of profit.**

[1] ***Index Terms*—performance indicators, data analysis, business environment, manufacturing industry, overall equipment efficiency**

I. Introduction

THE analysis carried out in this report is based on the operational datasets of an Ice cream manufacturing plant. For this analysis, both production-related data and utility-related data is incorporated. The following paragraph shows a brief company introduction.

The Colombo Ice Company(PVT)Ltd owns an aforesaid factory premise which is a modern Ice cream manufacturing facility that meets world-class standards in terms of both process and equipment. It is located at the Seethawaka export processing zone Awissawellla. This factory consists of 18 utility packages and 8 production Machine packages which facilitates two end product lines for finished goods.

The finished goods of this factory are impulse category Ice cream with volumes less than 100 ml which are in common terms Ice Cream Cups, Sticks(Candy) & Ice Cream Cones. There are a number of different products under each category. Giving a brief introduction on ice cream manufacturing, the first step is to process the ice cream mix, then chocolate and edible cones are manufactured as edible material. Afterwards ice cream mix is pumped into the filling and packing sections in order to produce the desired end product.

In terms of financial perspective, the operation of the factory with a competitive advantage is a key parameter in order to survive in the category of Fast Moving Consumer Goods. There is a severe competition within the Sri Lankan market for frozen confectionery(Ice cream) where approximately 40% of Impulse Ice cream market share is held by “Elephant House”. Therefore as a business, there are plans to gain more market share. In simpler terms, one of those strategies is by reducing the conversion costs(manufacturing-related cost) by improving efficiencies in terms of all the controllable parameters.

In a factory environment where production efficiency depends on a number of parameters, it is quite difficult to carry out simple root cause analysis and improve efficiency as it is a simple problem. Now only time-series data is being evaluated weekly basis, how the production quantities and utility consumption with the variation of performance indicators. The issue with this type of illustration is that there are no insights of the day to operations. For example, if OEE of a particular machine has reduced in a day, there is no quantifiable reasoning for the incident. There can be qualitative reasons such as breakdowns have occurred, the material was not delivered on time etc. But there is no indication how much a particular reason affects efficiency. Therefore a systematic approach to analyze the available data set whether it is significant, in order to improve the efficiency of operation.

There are several initiatives launched by operations excellence teams, in Savings projects, Connected factory(IIoT) systems that improve Profitability and Visibility respectively. But proper systematic data analysis initiative with the below-implemented approach will give a next-level competitive advantage for efficiency improvements.

The other important factor which has not yet been exploited in this field is forecasted approach in production planning. i.e. Usually a production plan is an amended version of a demand plan, where the product quantities are dependent on the macro factors. But micro factors such as profitability indicators are not considered due to the fact, those are difficult to predict

[1]Link to the github page. https://github.com/DulanGit/In20-S1-CS5617-project-1

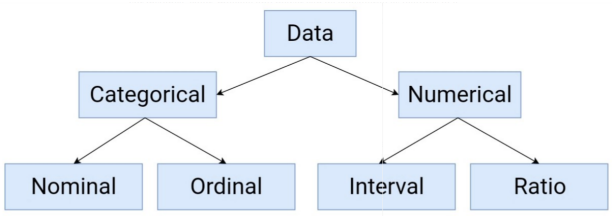

Fig. 1. Different data types in statistics.

due to a large number of dependencies at different weights. Therefore in this paper, some predictive data analysis has been done to classify the plan whether it is profitable or not.

Therefore it is an obvious fact that exploring this uncharted area of data science we are able to dig into deeper insights and discover root possibilities to improve efficiencies thus profitability. Also by predictive analysis, there is a possibility to eliminate the inefficient & production runs which have lower profitability.

Summarising the topic in general concepts, by leading this firm to make better business decisions using analytic methods and create competitive advantages from data then managerial and leadership positions rely on data-driven decision making. Recent studies have shown companies who adopt "Data-Driven Decision Management" achieve significant productivity gains over other firms.

II. Background

In the study of data science, we classify the data into different categories to identify the data type. A simple illustration is shown in Fig. 1 for different categories of data which some types are being included in this project.

In this dataset, there are both categorical and numerical data. Giving an introduction to the tables of this data set, in this factory environment, there are two lines which give the end product. Production and performance-related data for those machines are included in separate tables. The other table consists of all the utility data, All the data is captured daily basis over a time period of one year.

The data sources are a set of spreadsheets which is maintained by respective departments in the company. For clarity below table explains the attributes in one of the tables for the production machine, which we have selected to carry out the analysis in the latter part of the project.

Attribute Name	Description
Date	The date when operation is carried out
Product	The Product that was in manufactured
Planned Start Time	The time production line was expected to start
Actual Start Time	The time when that machine actually started
Stop Time	The time when production of the machine stopped
Total Time	Total time spend for production activities including all the delays
Start-up Delay	The delay occurred to start the production line
Breakdown delays	Total time wasted for machine breakdowns and other operational delays
Effective prod time	The time duration which actual production was carried out
Operating Speed	Output rate of the machine in eaches:It is number of units per hour
Produced Qty	Actual quality approved production quantity in eaches of the product
Rejects Qty	Quality rejected quantity in eaches of the product
Mix Usage	Amount of Ice cream mix(milk) used for filling and packing
Mix Wastage	Amount of Ice cream mix wasted prior to filling and packing
Mix Yield	Percentage of the mix volume used for production out of total mix volume
AR	Percentage: The time machine was ready for production out of total planned time
PR	Quality rate Percentage of good products quantity from total production volume
QR	Quality rate Percentage of good products quantity from total production volume
OEE	Overall equipment efficiency: Product of AR, PR and QR : Overall indicator of the machines performance

As per the end objectives of this project in order to empower Data-Driven Decision Making(D3M), which is the process of making decisions based on the data rather than mere observations and personal experience, this project is carried out as per the steps of D3M which are,

1) Store
   This is how operations data is gathered and stored for an analysis. This part is almost complete since data gathering is completed.
2) Analyse
   This is the key component of this project: using the data set, Descriptive, Diagnosis and Predictive analysis is carried out. Under descriptive analysis simple indications have been given on the parameters answering straight questions on the data. Then under-diagnosis analysis detailed indications have been given for several observations based on the dataset. Then in the predictive analysis part based on the past performance, future production has been forecasted and classified to observe whether those are profitable to operate or not.
3) Visualize
   After analyzing data, several illustrations has been prepared, such a way that users are able to see through the actual picture and what is really happening.
4) Decide
   Several recommendations has been given based on those visualisations.

III. Methodology & Results

A. Pre Processing

Data is often taken from multiple sources which are normally not too reliable and that too in different formats. There may be problems due to human error, limitations of measuring devices, or flaws in the data collection process. Mainly python language [2] and pandas library [3] was used to do the data processing. Preprocessing was done as separate stages.

1) Derive useful features: Some of the useful indicators can be derived from the existing measurements. Rejects percentage is an example. These new rows were added to the existing data frame.

$$Rejects\,Percentage = \frac{Rejects\,Qty}{Produced\,Qty\ +\ Rejects\,Qty} \tag{1}$$

2) Inconsistent columns: Dataset itself contains columns that are irrelevant or useless columns that can drop them to give more focus on the other columns that have meaningful information. First, select the columns which are useful for the analysis. There was a lot of redundant data. From the previous experience, the total time has no effect on the efficiency of the product. Also, some columns are highly dependent on others. As an example, the difference between planned start time and the actual start time was directly reflected in the Startup delays. Most useful columns were selected to analyse the data further.

3) Remove duplicate values (categorical data): The product categories were duplicated so some of them had to be renamed to a common name before processing the data. After removing the duplicates the product column unique products were reduced from eighteen to twelve.

4) Missing Values: Missing values in a dataset are common in data science. This usually happens in the data collection part or when there is some data validation rule. Regardless of that missing values must be taken into consideration. In this dataset, some of the columns had null values. There are two common methods to tackle this problem.

- Estimate missing values
  For this dataset some data could be able to be estimated by looking at other columns.

  $$Total\,time = End\,Time - Start\,Time \tag{2}$$

  Using the above equation null values which appeared in the "Total time" column could be derived from the other columns. Some other values were filled by calculating the mean of that specific c As an example "Mix Yield" column null values are being replaced by the mean of that column.

- Eliminate the rows with missing values
  Missing values were found which cannot recover or reconstruct from other features. Some missing data was removed since those data cannot be estimated.

5) Data Normalization: Data normalization is an important step in most of the analysis problems. Data normalization was done when seeking for relations. Different data were in different ranges, in order to train a regressor, all the data should be normalized. There are different types of data normalization, Z Normalization was used in this study. This transformation sets the mean of the data to 0 and the standard deviation to 1. This was done Feature-wise, for each column.

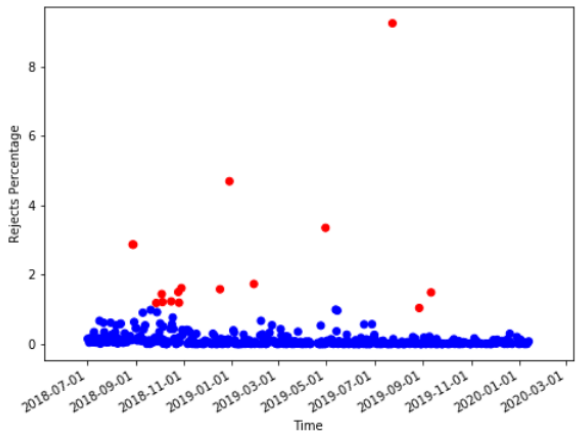

Fig. 2. Identification of Outliers in rejects percentage.

6) One hot encoding: One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. Here a deep neural network was used to predict the data, in order to prepare the data for the input categorized products were converted to one-hot encoding.

7) Remove the outliers: In statistics, an outlier is a data point that differs significantly from other observations. That means an outlier indicates a data point that is significantly different from the other data points in the data set. Outliers can be created due to the errors in the experiments or the variability in the measurements. Some of the values were found in this dataset as abnormal values. Keeping these data will drastically reduce the accuracy of the model. Since some values cannot be estimated by other values, these outliers had to be removed from our dataset.

For instance Fig. 2 was plotted the values of rejected percentage. All the values higher than 1 are represented in red colour dots and which are less than 1 represented by blue colour dots. These rows were removed from the dataset in order to clean the dataset.

B. Descriptive Analysis

A descriptive analysis is an important first step for conducting statistical analyses [4]. Descriptive statistics are used to describe the basic features of the data in a study. In this section, our main focus is to describe the data gathered during the first

phase. Most of the graphs were generated using Seaborn [5] library which is highly focused on visualizations.

OEE (Overall equipment efficiency) is the main focus of this dataset Since it is the main measurement of the efficiency of this entire production line. A couple of graphs were used to analyse this OEE to check whether how it changes.

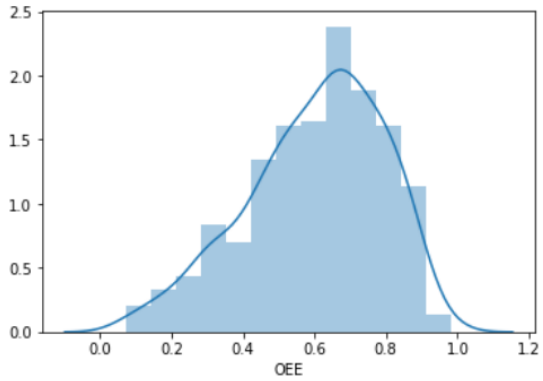

Fig. 3. Normal Distribution of OEE

When checking the distribution of OEE Fig. 3 , it was mainly oscillating around 70%. Organization's target is to improve this and move this distribution more towards higher percentages.

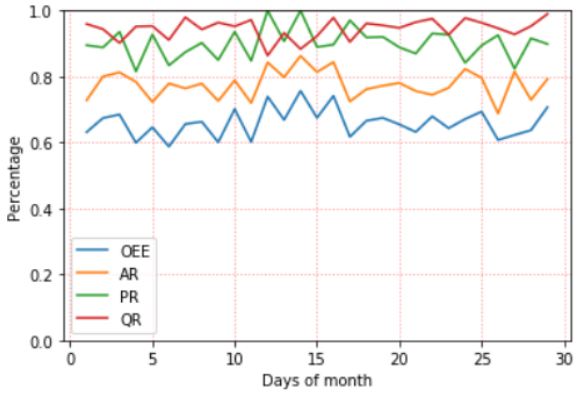

Fig. 4. Daily OEE, AR, PR,QR variation for a month

The following Fig. 4 is the OEE(Overall Equipment Efficiency), AR(Availability Rate), PR(Performance Rate), QR(Quality Rate) variation throughout the month. All these are the final indicators of the efficiency of the production. High efficiencies were observed during the middle of the month (12-17). The reason higher volumes are produced in during the middle of the month in order to cater the sales targets.

Another example of descriptive analysis is how the same parameters are shown in Fig. 5 graph has a varied month on month during the year 2019. In this graph, it's clear that the indicators are improving with time. Last year January OEE started around 60% and it improved to 70% at the end of 2019.

The visualisation in Fig. 6 illustrates the variation of OEE for different products manufactured in the stick line. In

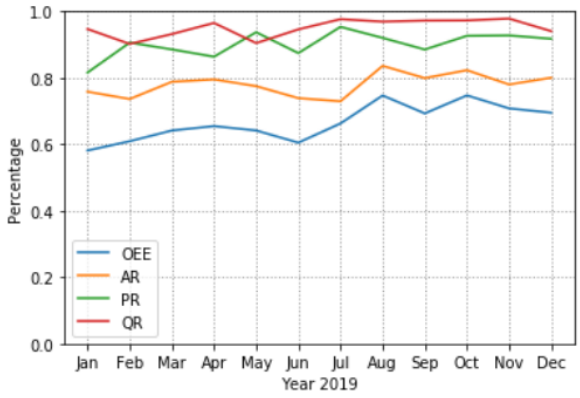

Fig. 5. Month on Month Average OEE, AR, PR and QR

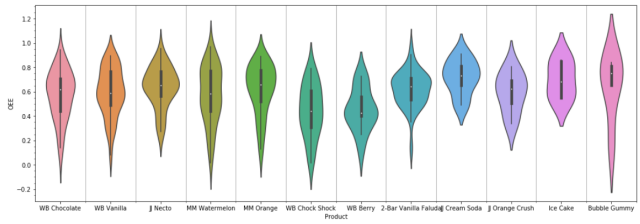

Fig. 6. Distribution of OEE product-wise; The highlighted vertically black component is Q1 Q3 range and white dot indicated the mean. Higher the width of the shape the number of higher the number of occurrences.

this violin plot, the white dot represents the mean and the black bar represents the interquartile range. And the entire distribution is represented by the coloured area for each product. As a descriptive indicator, it is obvious that WB Chock Shock and WB Berry have a lower mean of OEE compared to other products. While Bubble gummy has a wide distribution, Cream soda has the narrowest. We can conclude that Bubble Gummy can be controlled and bring up the OEE, there are some facts that changing with time to time which make it vary the OEE job to job.

C. Diagnostic Analysis

Following Fig. 7 is a sample diagnosis analysis done for the variation of the month on month OEE, reasons for OEE reduction in a particular month is quantifiably analysed. Below figure is the OEE variation for the last year.

Observing the Fig. 7 graph it highlights the fact that OEE has recorded a minimum during January and June also maximized in August and October. The typical question for this kind of scenario is why OEE has been changed so far. In order to identify the quantifiable reasons for the reduction of OEE, the following Fig. 8 correlation matrix is generated between the factors that affect OEE. Without the implication of subjective reasons such as breakdowns has occurred, Start-up has been delayed for the reduction of OEE, by providing a correlation matrix, reasons can be provided in an accurate manner. Below is the correlation matrix for the year.

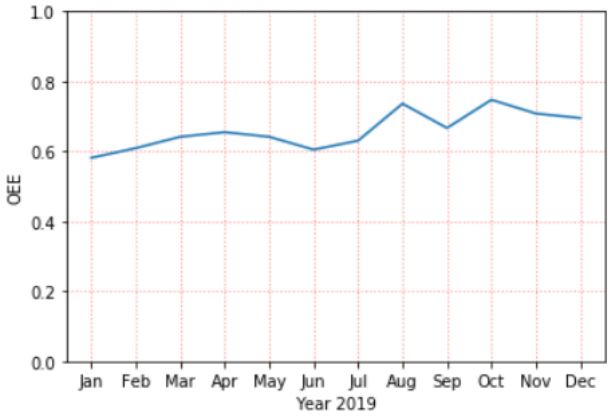

Fig. 7. OEE change during last year

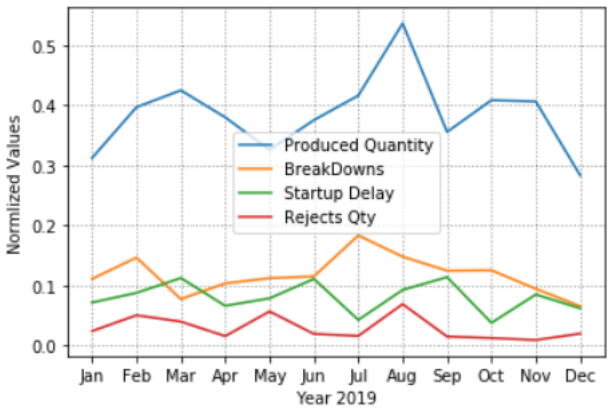

Fig. 9. Month on Month variation on Produced Quantity Breakdowns Start-up Delay & Rejects

- Pearson Correlation Coefficient

Pearson's correlation coefficient (Equ. 3) is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (3)$$

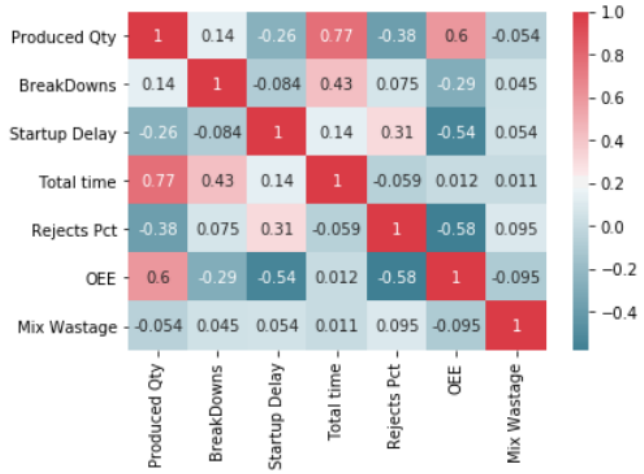

Fig. 8. Correlation matrix against main attributes of the relation

By analyzing the correlation matrix it is highlighted that OEE has a strong positive correlation with Produced Quantity and Negative correlation with BreakDown Delays, Startup delays and Rejects percentage. Also, the Total time and the Mix wastage correlation can be negligible. To analyse the effects, the relevant correlated features were plotted with respect to the time. Fig. 9 shows the graph.

On January Production Quantity is low and all other factors are average. Since the production quantity having a positive correlation with the OEE its obvious that that the OEE is reduced because of the quantity. July there is a huge increase in breakdowns the highest reported in 2019, which is highly negatively correlated with the OEE as the correlation graph shows. Its cleat that the increase in breakdowns brings the OEE down in July.

When checking the August there is a huge increase in produced quantity so this is positively effected for the OEE to go up and on October the startup delay was minimum which results in a higher OEE on that month.

D. Predictive Analysis

This is the scenario where the output or the end result is being forecast based on the controllable or parameters that can be provided as targets. In this project, the focus will be to consider the set of parameters that are taken into consideration prior to planning a production run. Conventional production planning is carried out taken into consideration of two factors.

- Demand for the period
- Finished good stock level

Although this approach does the job of producing enough quantities for the period to satisfy the market demand, the efficiency factors are not considered in this process. Therefore this kind of planning might lead to in-efficient production runs. Therefore before starting manufacturing a product some targets have to be set. The following information was taken into consideration to predict whether a production run will be with acceptable OEE or not. By implementing this classifier, it enables to eliminate the unfavourable OEE thus making the plannings make sense in terms of profitability.

- Product
  What are the products need to be manufactured: This is an input from the market condition and current stock levels: Macro Parameter
- Production Quantity
  How much products from each quantity should be produced: Also a macro factor which depends on the market conditions.

- Down Time
  Internal factor which can be assigned to a person as a target: This is the expected time that can be allowed for breakdowns.
- Start-up delay
  Internal factor which can be assigned to a person as a target. This is the expected maximum allowable time delay that can be there during the start-up process.
- Total Time
  Planned time to carry out the production-run: Internal parameter that is determined in planning stage.
- Rejects percentage
  Internal factor which can be assigned to a person as a target: This is the maximum allowable rejects(As a Percentage) that can happen during a production run.

So for predictive analysis, the above variables were taken as features and the final prediction is the OEE. For these predictions, different methods were used. Following are the methods used and results obtained.

1) Linear Regression: The term "linearity" in algebra refers to a linear relationship between two or more variables. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

Linear regression was obtained for the selected feature set and tested for the test dataset. It achieved 86% accuracy which is really good.

2) Support Vector Machines: . This supervised machine learning algorithm has strong regularization and can be leveraged both for classification or regression challenges. In the SVM algorithm, each data item is a point in n-dimensional space (where n is the number of features) with the value of each feature being the value of a particular coordinate. Then, perform classification by finding the hyper-plane that differentiates the two classes very well. In regression, Support Vector Machines algorithms use epsilon-insensitivity (margin of tolerance) loss function to solve regression problems.

The next approach was to use the Support Vector machines as a regressor for this problem. All these six features were selected and trained by this regressor. It gave 82.5% accuracy for this. But this was lower than linear regression may be because SVM accuracy is lower for regression problems, unlike classification.

3) Decision Trees: . A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Decision tree classifier was chosen to train the dataset initially. Sklearn library in python was used to train the regressor. It was given the accuracy of 76.03% for the test dataset.

Then hyperparameter tuning was used to further increase this value. Following parameters were obtained from the tuning

Matrix	Value
Splitter	Best
Criterion	MAE
Min Samples split	5
Min Samples leaf	2
Max Features	None
Max Depth	50

Using hyperparameter tuning, it was able to predict for the same dataset and reach 83.43% of accuracy. Check the decision tree image using this link.

4) Random Forest Method: Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

Then hyperparameter tuning was used to further increase this value. Following parameters were obtained from the tuning

Matrix	Value
Num estimators	800
Criterion	MAE
Max features	auto
Max depth	80
Min samples split	5
Min samples leaf	1

After hyperparameter tuning Random forest accuracy was improved from 88.27% to 88.67%.

5) Deep Neural Networks: A deep neural network (DNN) is an artificial neural network (ANN) with multiple layers between the input and output layers. The DNN finds the correct mathematical manipulation to turn the input into the output, whether it be a linear relationship or a non-linear relationship. Keras deep learning neural network [6] library was used to create and train a neural network. Network architecture is shown in Fig. 10

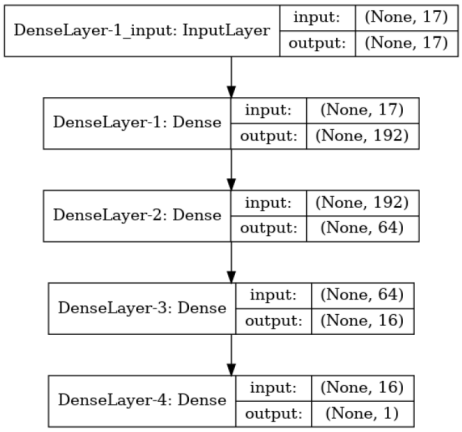

Fig. 10. Deep Neural Network Architecture

Hereafter several changes to the network architecture finally

came up with this model which is not overfitting or underfitting. It gave a pretty good test accuracy for the current dataset after 100 epochs. The values after 100 epochs as below.

Matrix	Value
Training Loss	0.0138
Validation Loss	0.0344
Test Dataset accuracy	93.04%

Accuracies for all the above methods are summarized in the below table.

Regression Method	Accuracy
Linear Regression	84.48%
Support Vector Machine	80.43%
Decision Tree	83.43%
Random Forest	88.52%
Neural Network	94.21%

Different methods were taken to accurately predict the OEE with the available targets beforehand. Comparing the accuracy values for the above methods it's clear that the best accuracy was given by the NN model. It predicted the OEE with the accuracy of 94.21% percent.

These predictions are very valuable for a company like this. Starting a production line is a big task, its time and resource consuming. There is no easy way to make a decision when to start a new batch of products, most of the time with the experience one might take the decision based on that. With this approach, there is a clear prediction about how the outcome would be given the targets. Even a non-experienced person can now decide to start a new production line or not based on the information that is provided by these models.

From these predictions, we are clear that we can predict the Overall Equipment Efficiency beforehand. This is very useful for the organization since it can foresee the outcome before even starting the production.

IV. Conclusion

The main objective of this project is to introduce Data-Driven Decision making for a manufacturing environment where more weight is imposed on the data analysis and visualization steps. Data Analysis has been carried out in a descriptive, diagnosis & predictive manner by providing different answers on the data set which is taken into consideration. For the sake of summarising the work only a few aspects, each analysis criteria is illustrated in this report. But there are other possibilities that same analysis methods can be implemented in order to get in-depth insights.

Acknowledgment

The authors would like to thank Dr. Uthayasanker Thayasivam and Dr. A Shehan Perera at University of Moratuwa for providing guidance to complete this project.

References

[1] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," *Big data*, vol. 1, no. 1, pp. 51–59, 2013.
[2] Python official page. [Online]. Available: https://www.python.org/
[3] Pandas documentation. [Online]. Available: https://pandas.pydata.org/docs/
[4] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 4–37, 2000.
[5] seaborn: statistical data visualization. [Online]. Available: https://seaborn.pydata.org/
[6] Keras: The python deep learning library. [Online]. Available: https://keras.io/

Dulan Jayasuriya received the BSc in Engineering specialized in Electronic and Telecommunication Engineering from the University of Moratuwa in 2017, currently following an MSc program in Computer science in University of Moratuwa.

Chanindu Leelananda received the BSc in Engineering specialized in Electrical Engineering from the University of Moratuwa in 2017, currently following an MSc program in Computer science in University of Moratuwa.