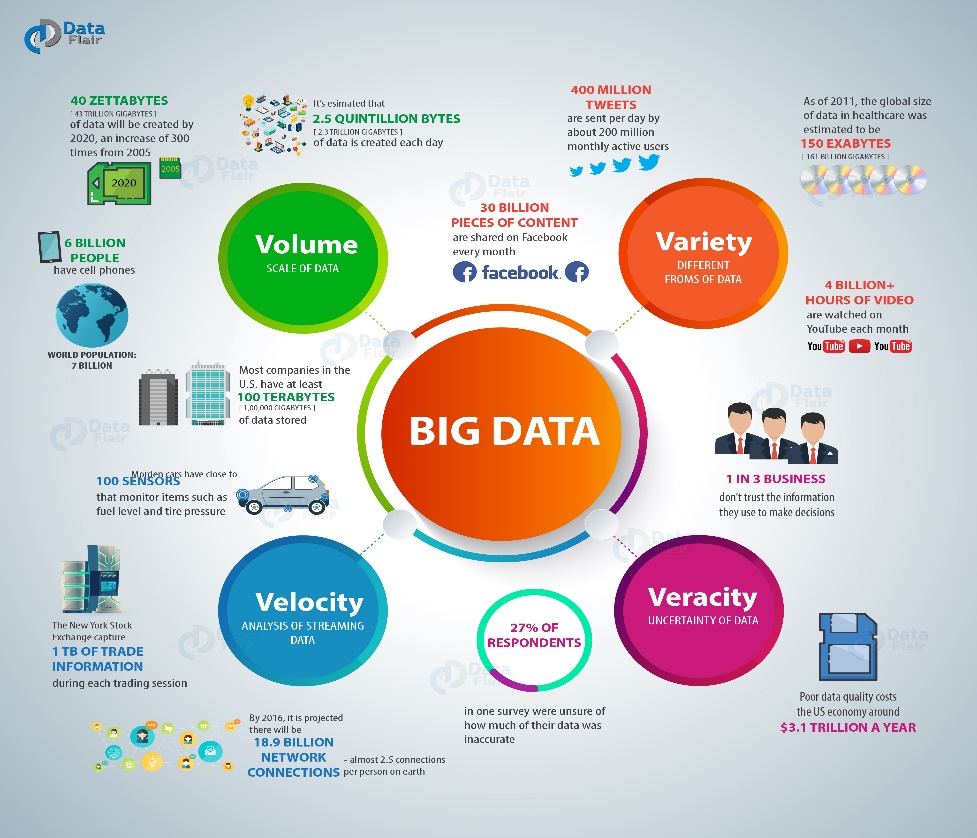
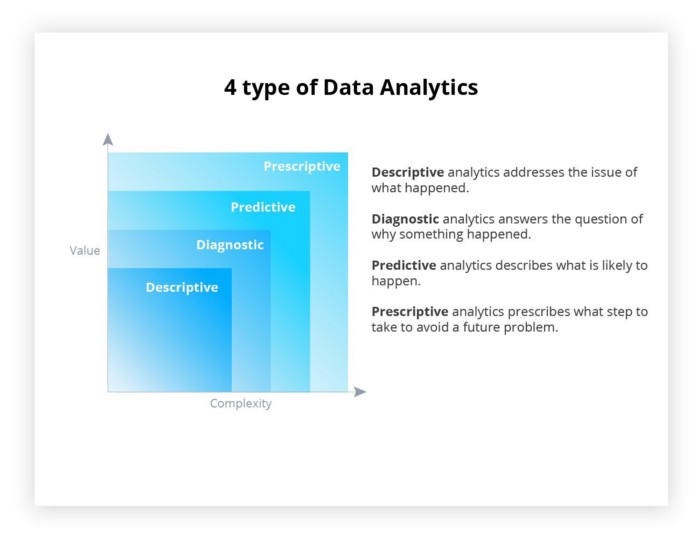
**TASK ONE – NARRATIVE**



5 Vs of bad data (Patidar, 2022)

The best approaches in big data according to me are **Descriptive Analytics, Diagnostic Analytics, Predictive analysis (Machine learning using Support Vector Machine), Prescriptive Analysis**

**Data Science problem:** One of the biggest food brands PepsiCo wanted to improve their supply chain moving the ready products to supermarket shelves quickly and efficiently. Their main motivation was to boost sales numbers in different regions and avoid losses caused due to understocking, overstocking, and product loss.



4 types of Big Data Analysis approaches (onix-systems, 2022)

**Descriptive Analysis:** is used to address the question of ‘what happened’ using data (onix-systems, 2022). This is accomplished using data (current/past). PepsiCo can use the descriptive approach to analyse their past shipping activity, sales activity from retail partners, helping them to understand their market position and KPI’s. This will give PepsiCo insight into how their business is doing. It cannot predict future trends/events, as it only looks at old data on the surface. It can bring attention to factors that might be leading to the company’s success or failure. The analysis can help the sales departments easily visualize the company’s performance, sales number, growth, losses.

Advantages: Analysts can use it without having a complete technical understanding of all the analytical or statistical concepts. Prebuilt resources can be used to perform the analysis, but the data has to be collected by PepsiCo which can be passed into the software. It can run through huge datasets and find the data features the company is interested in without it being hard to understand for the average person (e.g., shipping and sales data).

Disadvantages: Descriptive analysis method works well only on data from the past. The analysed data only answers the “what’s happening” part of the data but does not find the root cause of the issue.

Limitation: This approach cannot predict future events/trends since it only uses data from the past

**Diagnostic Analysis:** Used to determine the cause of an event based on trends and correlations between given variables, using historical data. Problems such as stock being lost on a particular shipping route, then the reason for this could be narrowed down using root cause analysis on the shipping dataset. Essentially this approach can be used to answer “why” something happened. PepsiCo could use this approach to find where most stock is lost and improve it. Or they could use the sales data from its partners to check the region and the flavour of the drink that is not selling well. And based on the data, either choose to advertise it more or improve it in the given region. Analyst must understand some concepts like correlation and causation and diagnostic regression analysis.

Advantage: Companies can with the right data get to the root cause of success or losses in the business. Since machine learning models are used for detecting patterns, they are very efficient and accurate along with eliminating bias and not being misled by causation instead of correlation.

Disadvantage: Is more complex than descriptive analytics, also not all types of business data can be analysed using diagnostic analytics.

Limitations: Machine learning cannot be fully relied on to produce the best results, human intervention is also important to put the output into context.

**Predictive analysis:** Predictive analysis can use big data Pipelines on the dataset using techniques like statistical modelling, data mining and machine learning to predict future trends or outcomes. These insights can give the management the ability to identify future threats to the company and act on them, or make business decisions that will lead to more profits in the future. PepsiCo can use Predictive analysis to learn more about product consumption in different areas of the world around different times of the year. Using this, they can then decide when to manufacture more drinks, which type of drink, and regions to ship them to. This will increase the profitability of the company. It can also be used to predict problems that the company might run into and mitigate them. Predictive analytics uses historical data to perform forecasting the likeliness of an event happening in the future, the magnitude of the events on the business and a time frame of when it might occur. Machine learning models such as Decision trees, Regression, and neural networks can be used as training models and then used to predict the future results

Advantages: Insight gained can be used to increase the company’s growth and revenue. It can help the company management prepare for future changes in shipping lines or manufacturing upgrades to improve productivity. The predictions will give the company a good idea of how much product their shipping line is going to need

Disadvantages: The quality of the data used to train the ML model needs to be accurate and relevant to get the best results.

Limitations: The data that goes into the model needs to be of good quality to get the best predictions.

**Prescriptive analysis:** This is the evolution of Predictive analysis and Descriptive analysis as it can recommend a course of action for the company. This is a complex process so it is PepsiCo could not use it for short term day-to-day operations. To make this work PepsiCo would need data from internal sources and external partners to be used with machine learning models. PepsiCo can use it to get stock numbers they would require for shipping during different times of the year, they ask for recommendations on which region is going to have a higher demand for their products. Also based on current routes data, build more efficient new routes that would be even better.

Advantages: It can be used to get recommendations on a wide variety of business decisions. Since the processing is done by computers the risk of human error is low.

Disadvantages: They cannot replace humans in making the final decision, it is just a tool used to help in the process.

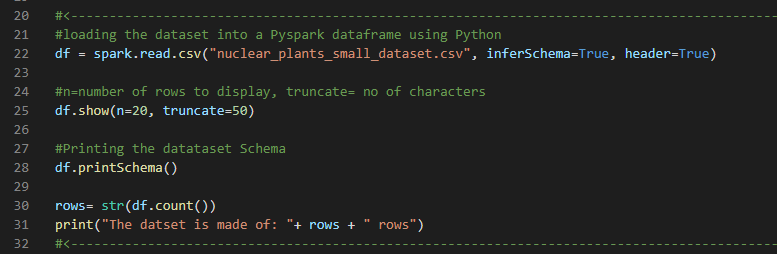
Limitations: Without huge amounts of data to use for training the results can be inaccurate. Sometimes the system cannot account for all variables related to the data.

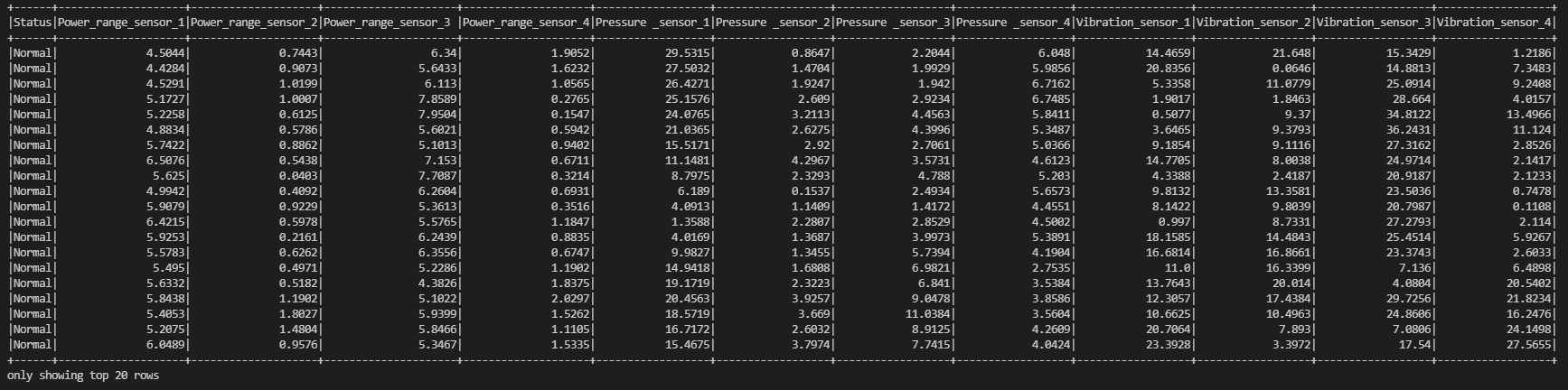
|  |  |  |  |
| --- | --- | --- | --- |
| Big Data Approach | Advantages | Disadvantages | Limitations |
| Descriptive Analysis | * Analysts can perform descriptive analysis without having the complete technical knowhow of all the analytical or statistical concepts * It can answer most queries about the business KPI and performance * It is very really well with huge datasets | * This method works well on the data only from the past * Cannot find the root cause of the cause of an event or trend | * This method cannot provide establish relationships for data pertaining to future events * Only works with data from the past |
| Diagnostic analysis | * Can find correlations and relationships between given variables * Uses a range of techniques to work such as data mining and discovery, statistical analysis, algorithms sensitivity and error analysis * Transforms complex huge datasets into manageable and visualisable data for the average person | * Cannot be used to predict future trends, since it mainly focuses on the data from the past and its relationship * Not all business use cases can rely on diagnostic analysis to base their decisions on | * Diagnostic analysis focuses on past events; hence it is not able to deliver any useful future insights * Some human intervention is needed to give context to the output data |
| Predictive analysis | * Information can be used to form a strong business strategy leading to the company’s growth * Hep the company’s management be better prepared for future events and be better prepared for future challenges * Can predict how much product the company is going to need to supply in the future * Can be used to perform risk and loss reduction * Help to forecast inventory needs | * The quality of the data used is very crucial to the quality of the insight received from the analysed data * The data collection process is very unreliable as some sources of data might not have consistent standards, also people might lie when asked certain types of questions | * If the data used for training the model is bad then the Predicted analysis will be very inaccurate and cannot be relied on * It is not 100% accurate since it is based on probabilities |
| Prescriptive Analysis | * Low risk of human error * It can provide advice on possibilities in many areas of the business * Its output can give management insight into making good decisions for the uncertain future | * Uses complex machine learning algorithms which increases its complexity * Still requires humans to make the final decision * Poor quality input data can result in forecasts that are faulty | * Needs large amounts of data to be fed into the system to get useful results * The data need to be of high quality |

**TASK TWO – ANALYSIS**

**Section 1: Data summary, Understanding and Visualisation**

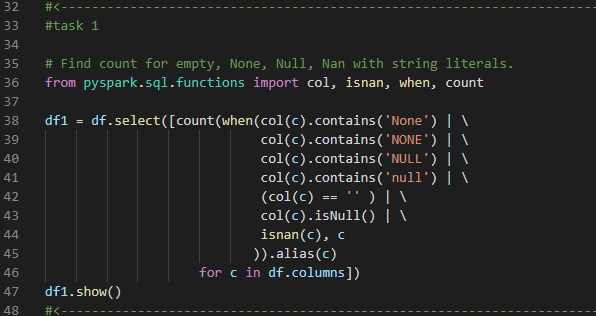
Read csv file, shown below:



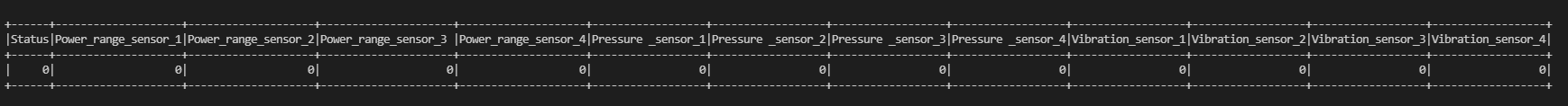


**Task 1**

For the provided reactor dataset, **there were no missing values** in it. Using the code shown in the image below the data set was checked for “None”, “NONE”, “NULL”, “null”, and blank values(“ ”). The missing data vales were counted for each column using the count(when(col(c).contains(“values”)))



The code above checks for missing values and displays it as a table shown below. All Columns have no missing/invalid values since they are all displaying 0.

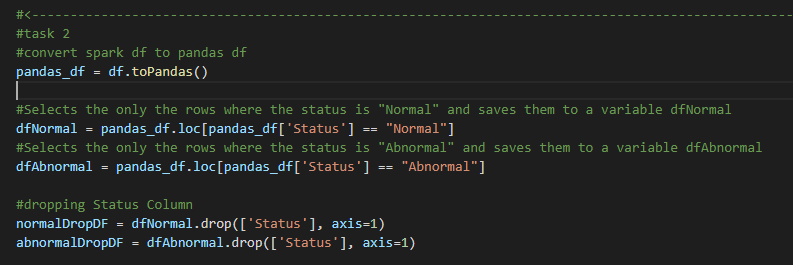


One way to deal with missing values is to remove the rows or columns that have any or more than half of the values missing (Kumar, 2022). This will improve the model accuracy but this could lead to loss of information if there are excessive missing values in comparison to the complete dataset. Generally, it is better to avoid this method but when there are few missing values it can be used.

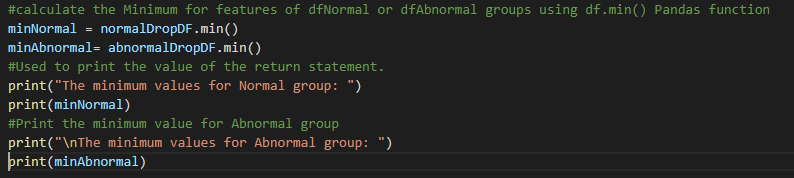
Another method is to replace the missing value with either the mean or median value of the column. The advantage here is that it prevents loss of data. But, it can cause data leakage,also it only works with numerical values and not with string type values.

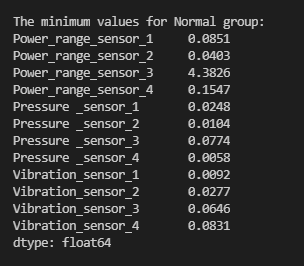
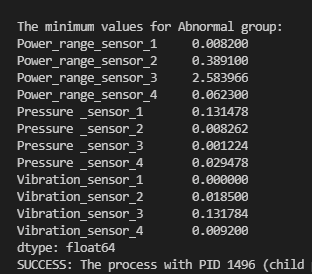
**Task 2**

For task 2 I have used Pandas for calculating and printing values to the terminal since its presentation of data is better than spark. Next, split the data based on the value of the “Status” column into two.

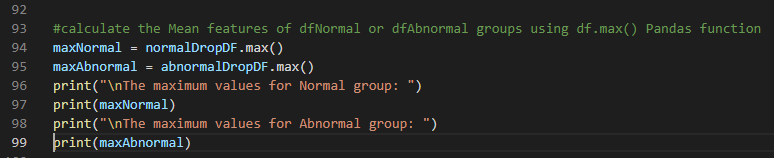


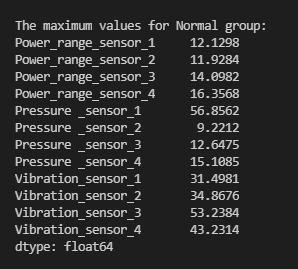
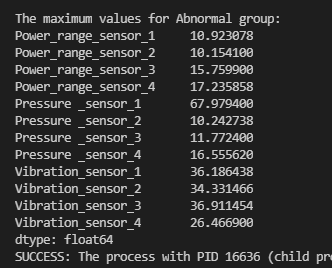
To calculate the minimum values the DataFrame should only consist of numeric values. So, the ‘Status column is dropped’ for both group’s DataFrame’s and they are saved to **normalDropDF** and **abnormalDropDF** df variables. To calculate the minimum for features of both DataFrame’s, the dataframe.min() Pandas function. The minimum values for both groups are shown below:



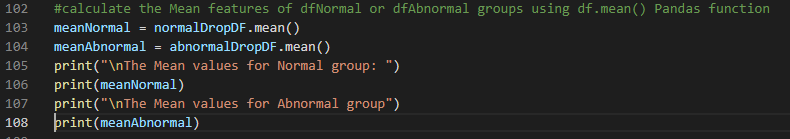
 

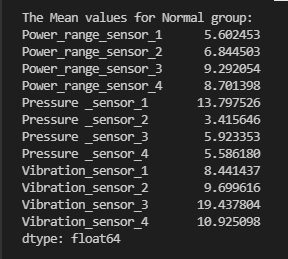
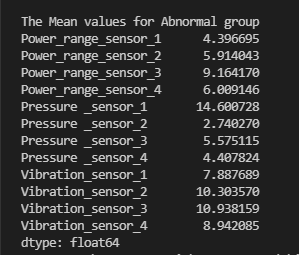
Calculating the Maximum values is accomplished by using dataframe.max() Pandas df function that return the maximum of the values in each feature:



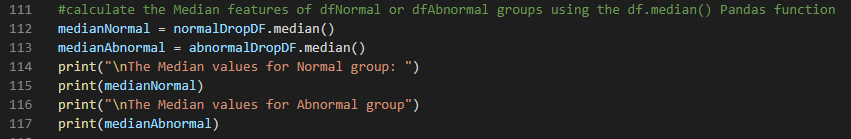
 

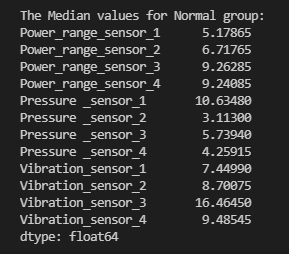
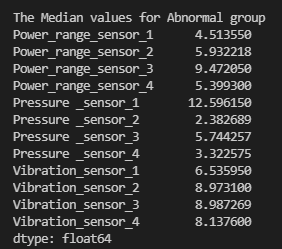
The mean of the features were calculated by used the dataframe.mean() function. Its good practice to remove any outliers in the data as they can affect the mean value. Mean values shown below:



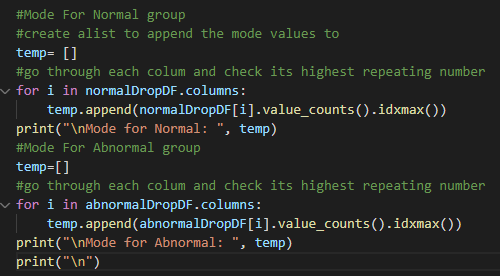
 

Median is the middle value (by magnitude) for the given column, and is calculated by using dataframe.median() function. Median values are shown below:

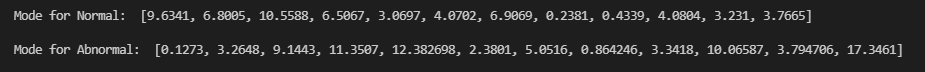


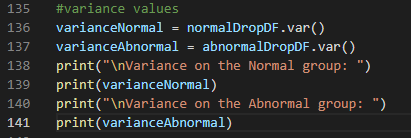
To calculate the mode (most repeated value) of the features, a List called “temp” has been created to store the mode values. Then, using a for loop we iterate through the columns of both DataFrame’s and use dataframe[i].value\_counts().idxmax() where value\_counts() gets the most repeated values in each column and a randomly selected value among these is appended to temp, since some of our columns have more than one mode.

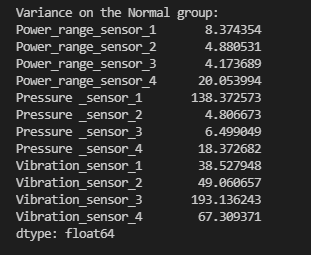
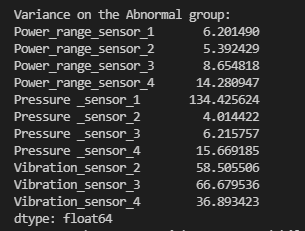


Mode for the DataFrame’s columns is shown below:

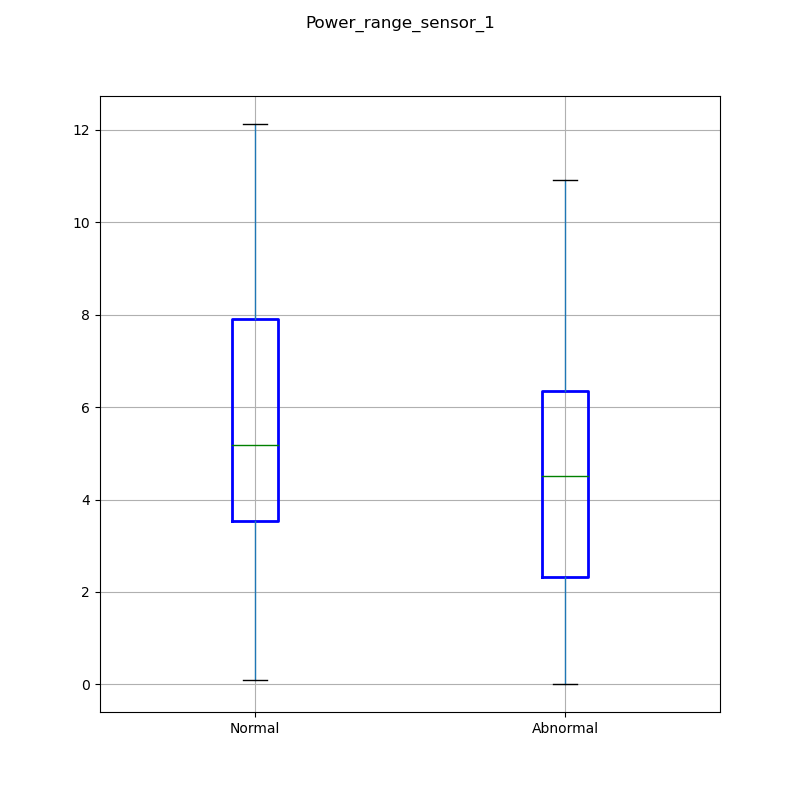


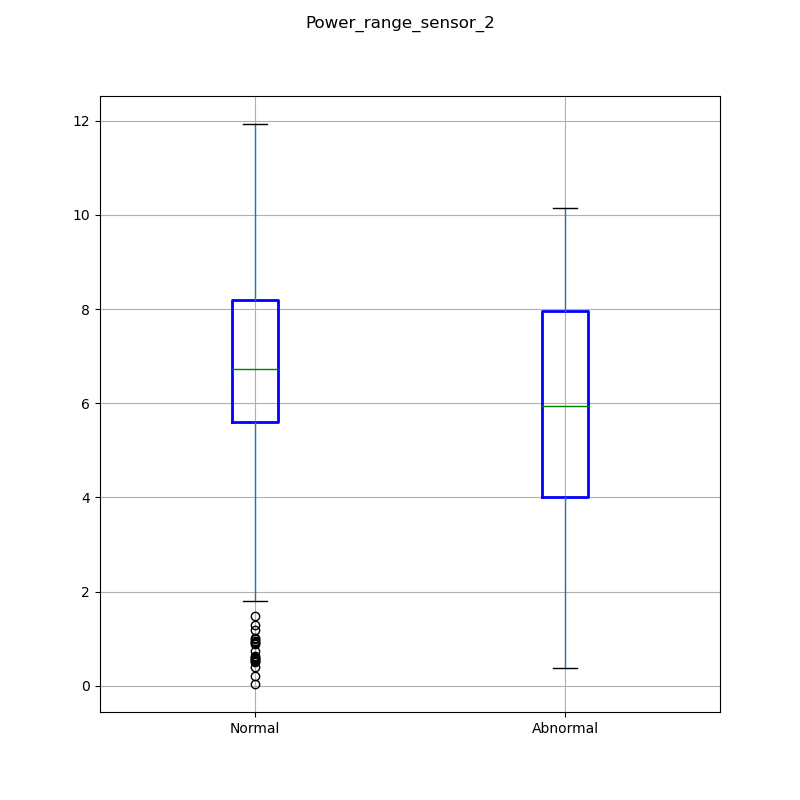
The variance of the DataFrame is calculated using dataframe.var() function. It is used to calculate the degree of spread in the dataset, the larger the variance the bigger the bigger the data spread is going to be.

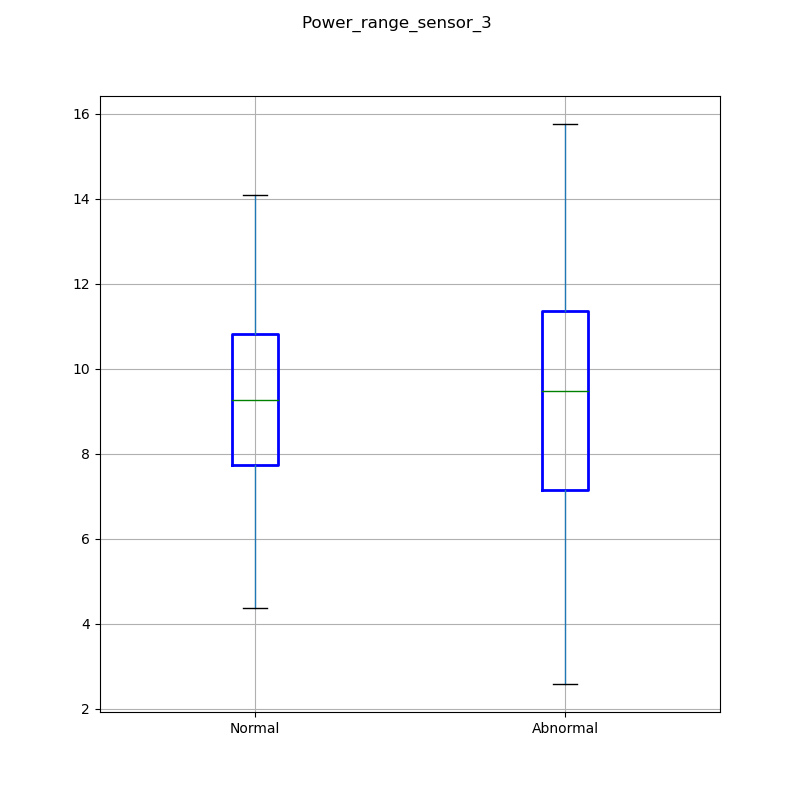


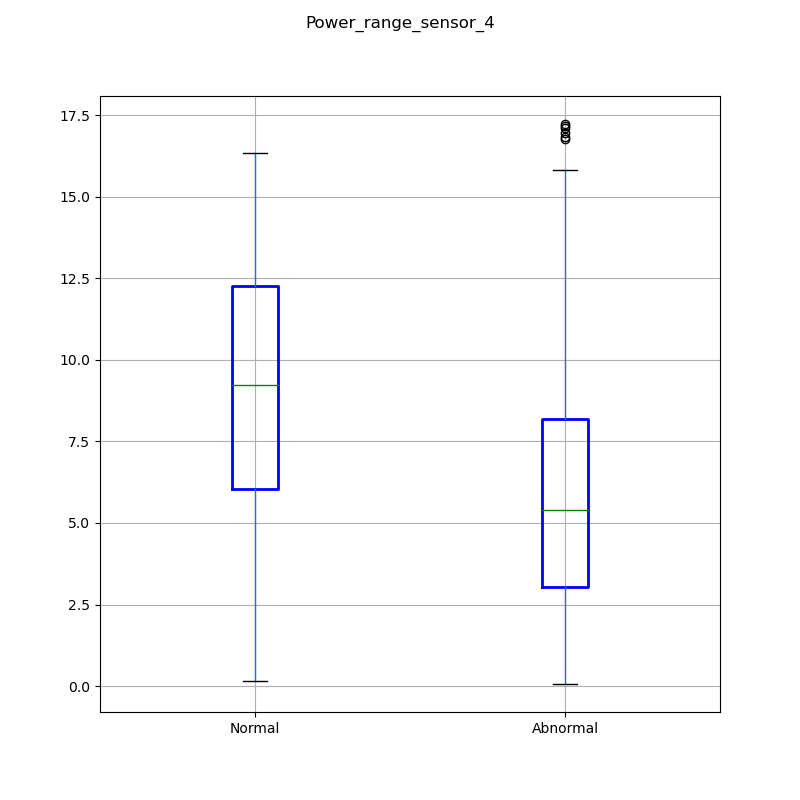
 

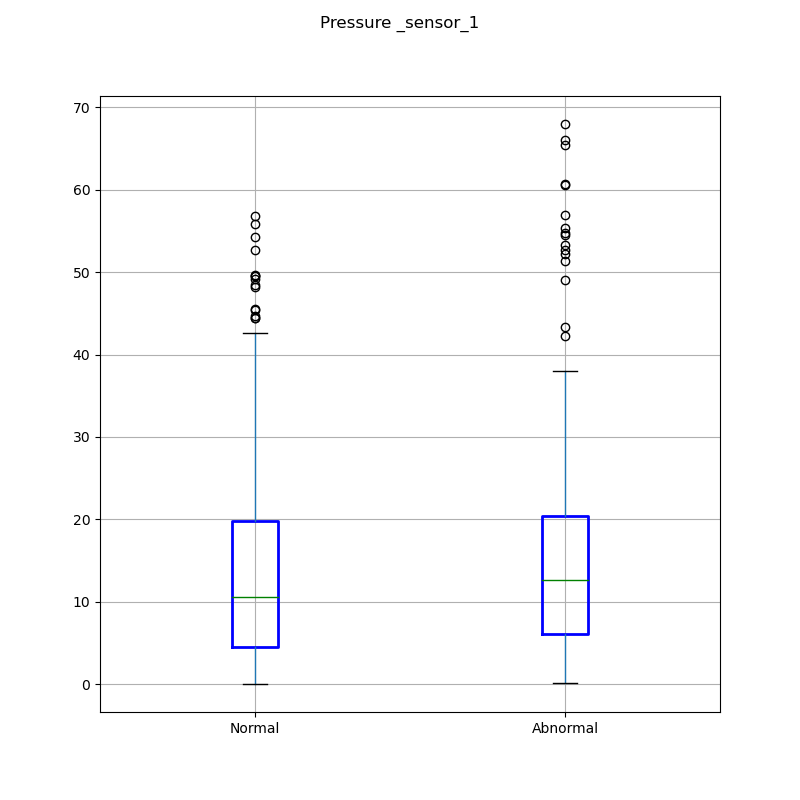
Box plots for features of the datasets: The boxplots show the value distribution for the given columns in the “Normal” and “Abnormal” groups DataFrame’s side by side for comparison.

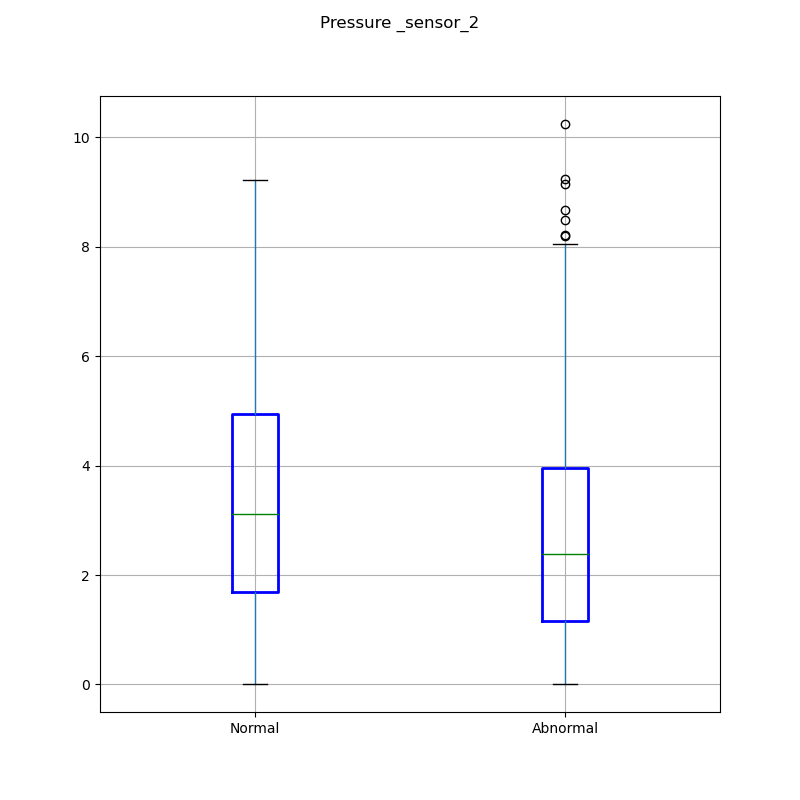


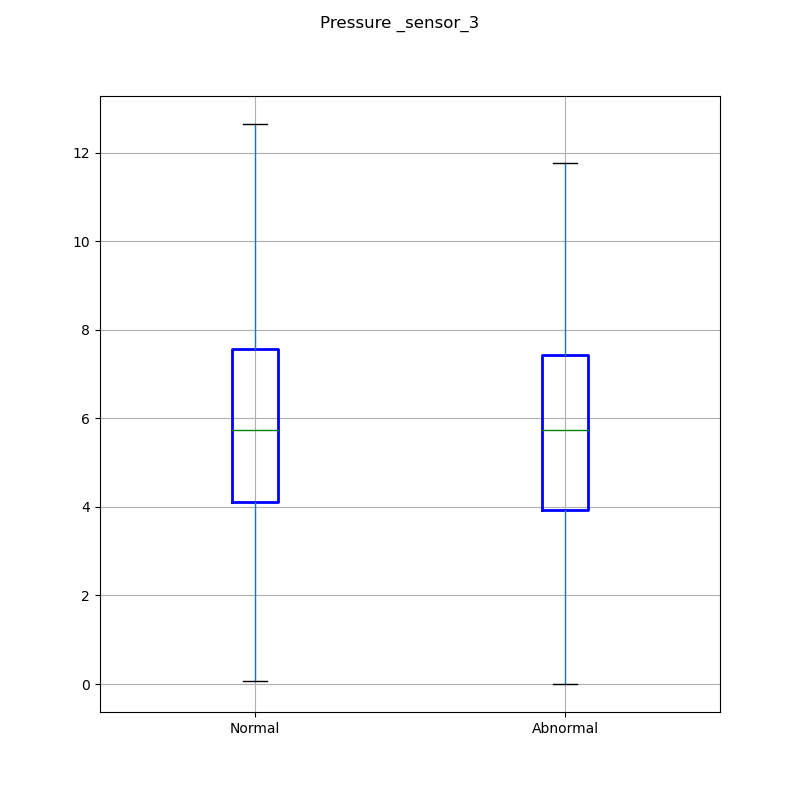


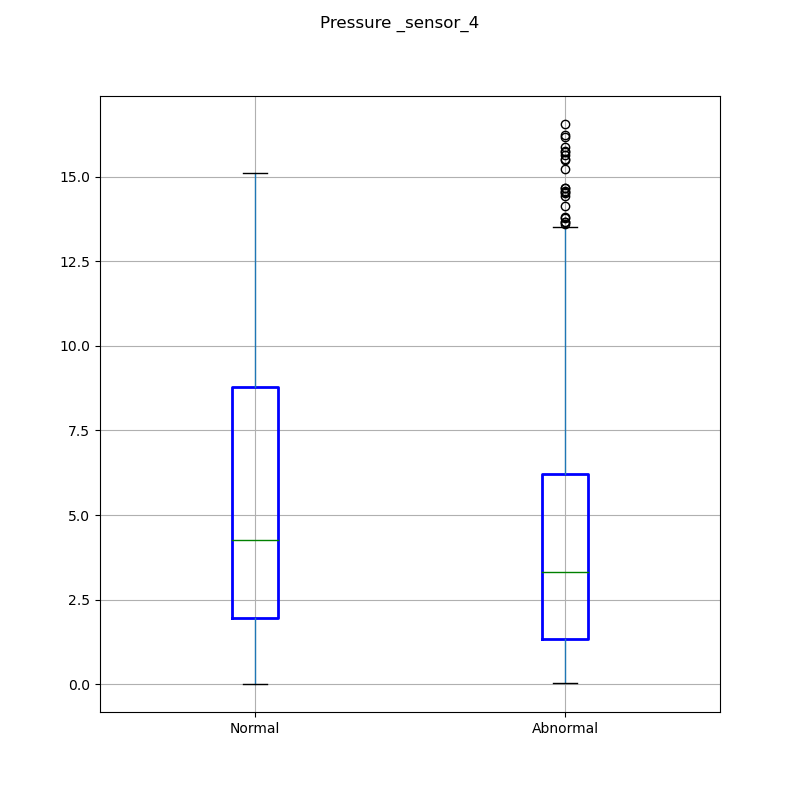


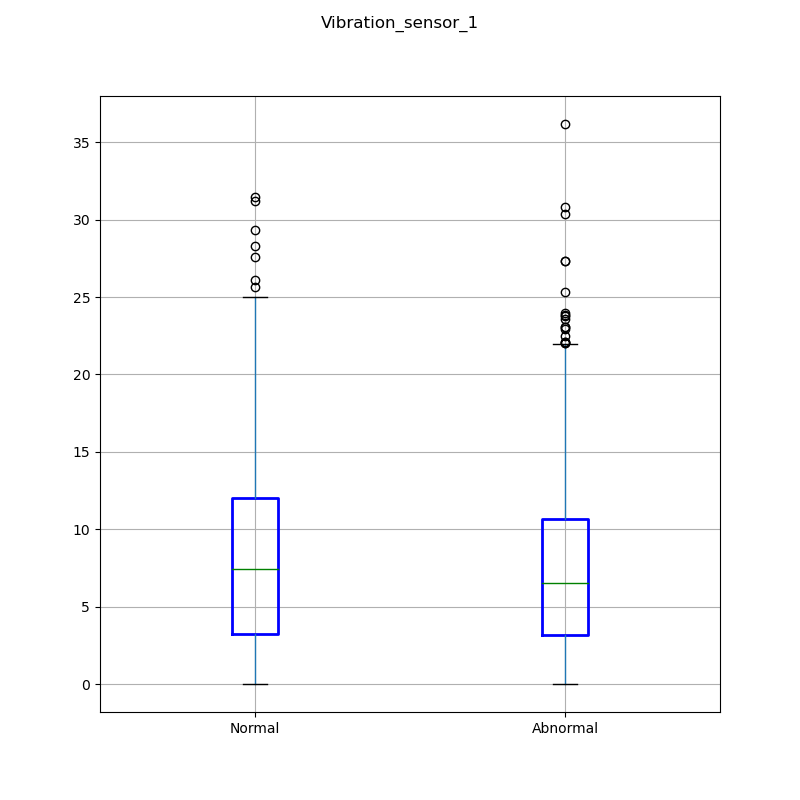


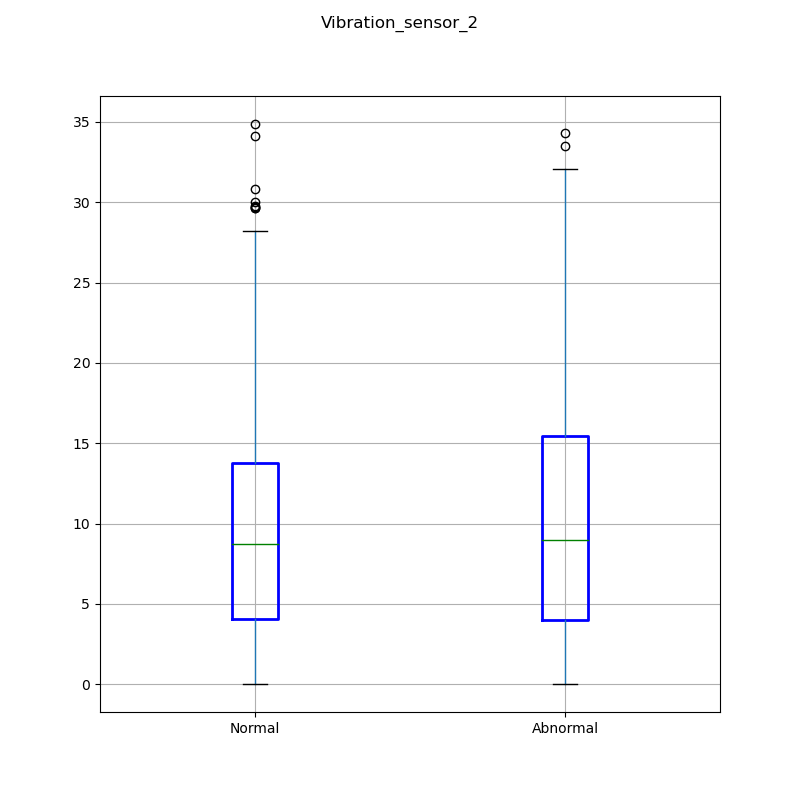


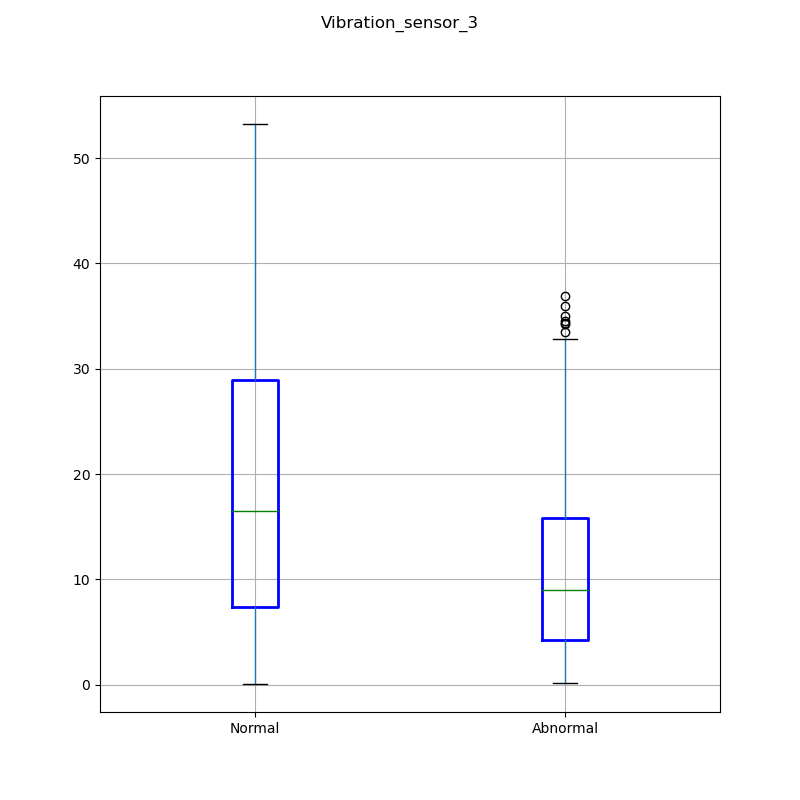


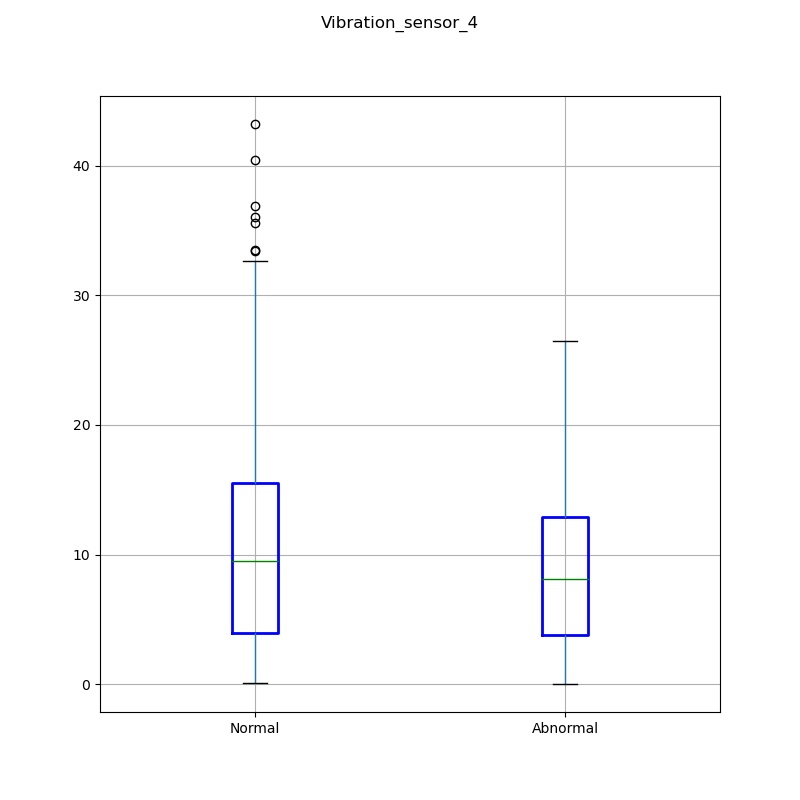








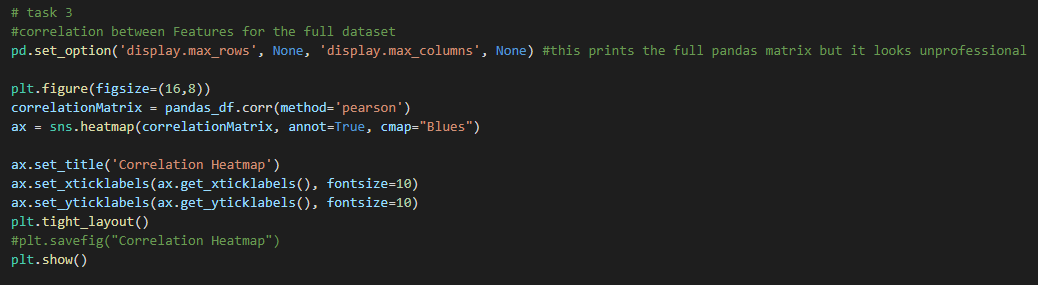




When looking at the summary statistics- they all need to be taken into consideration. If the user picks and chooses which statistics metric they want to look at it might skew their understanding of the data in a bad way.

**Task 3**

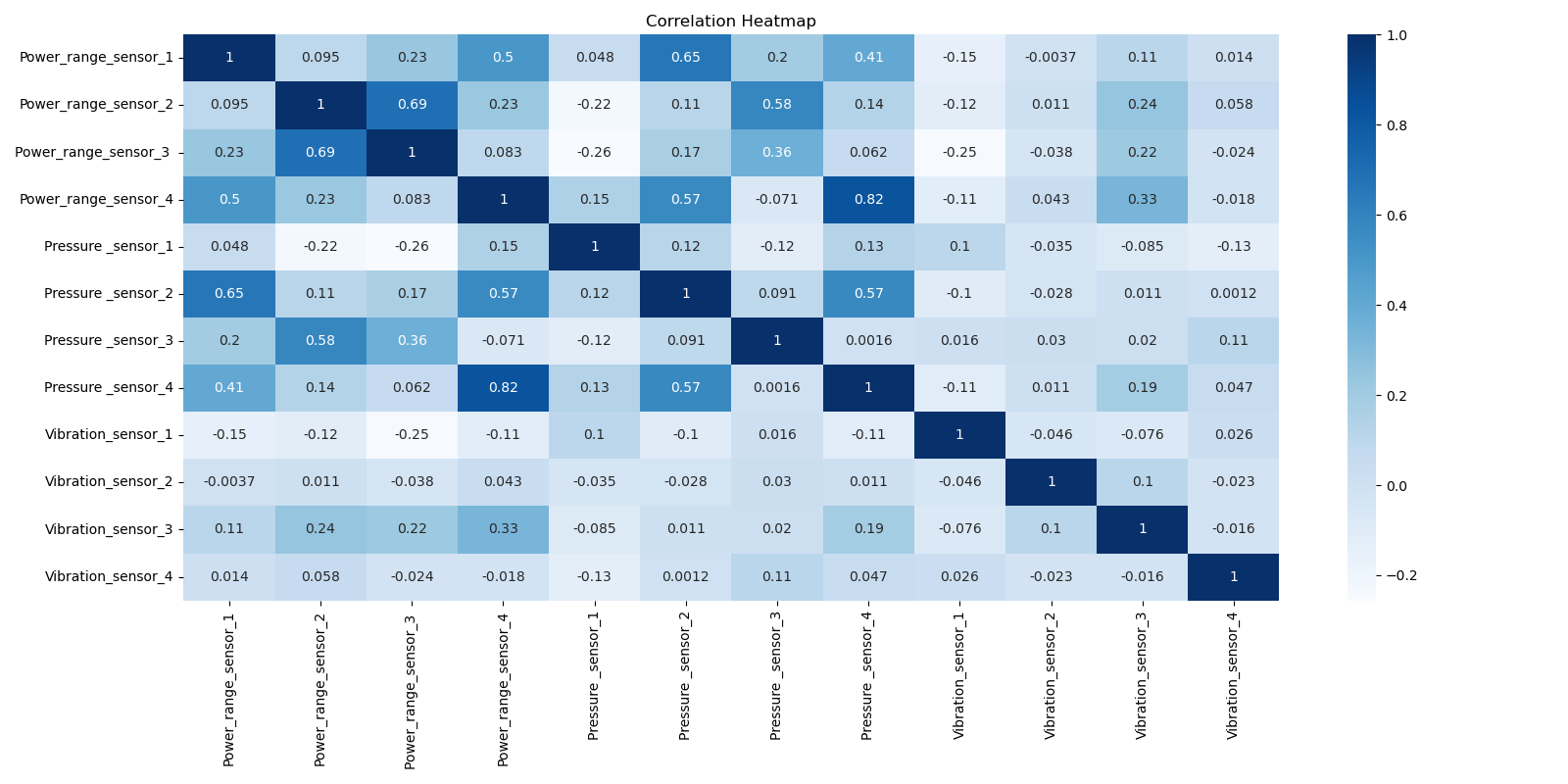
Correlation matrix table is used to show the correlation between two features in the dataset are related (Lutes, 2022). It makes it easier to understand their statistical dependence by providing quantitative measurement. The correlation measure goes from 0 to 1 where: 0= there is no correlation at all, 1= they are perfectly correlated



Correlation matrix has been calculated using df.corr(method=’pearson’) and then plotted using a heatmap to help visualise it better.(pearson=r\_value)

As seen in the image below the diagonal line of 1’s is because, the two same features always perfectly correlate with itself. Highest correlated features:

1. “Power\_range\_sensor\_4” and “Pressure\_sensor\_4” has the strongest positive correlation at **0.82**
2. “Power\_range\_sensor\_2” and “Power\_range\_sensor\_3” has the second highest correlation at **0.69**
3. “Power\_range\_sensor\_3” and “Pressure\_sensor\_1” has strongest negative correlation at -**0.26**

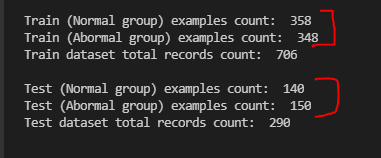


Correlation can help the machine learning model by checking which features are highly correlated so one of them can be dropped, preventing the duplication of data and improving the performance. While correlation has proven to be useful, it is important to understand that it may not always be meaningful or substantial. In a nutshell, correlation does not always imply causation

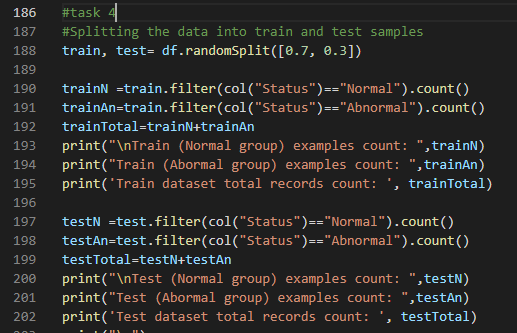
**Section 2: Classification & Big data analysis**

**Task 4**

**My examples values count:**

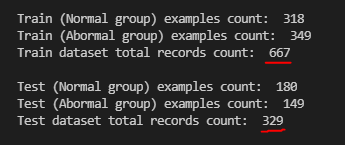


Here we split the Spark DataFrame(df) into train and test DataFrame’s with a random split of 70% to 30% respectively. We use the train, test = dataframe.randomsplit([0.7,0.3]) to do the above and then print the number of records in train and test. The count of records in each (train and test) is going to be randomly split which means records are selected at random and added to train and test, so for each run the size of records is going to be different.

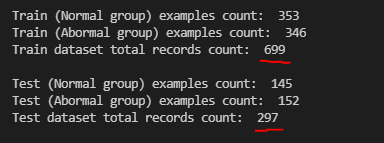


When the above code is run it print the outputs given below. Since it is random the record count is different on each try.

Try (1)



Try (2)

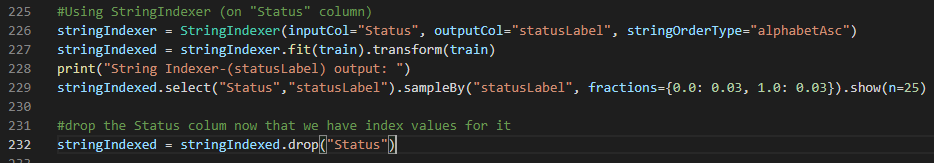


**Task 5**

Transformers are spark API algorithms that can take a DataFrame and transform it onto another DataFrame. It uses a transform() method to do the df conversion.

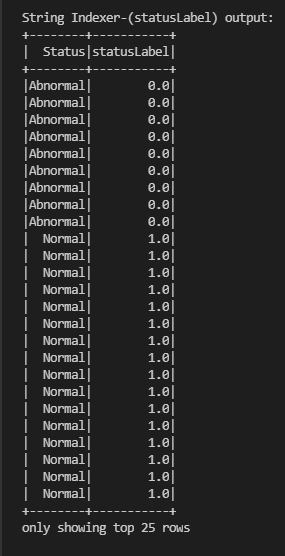
1. String Indexer

To prepare the feature data to be used with the classifiers, it needs to be transformed first. To start with the DataFrame does not have a label column yet. StringIndexer will take the input column “Status” from the train DataFrame, and outputs a column named “statusLabel”.



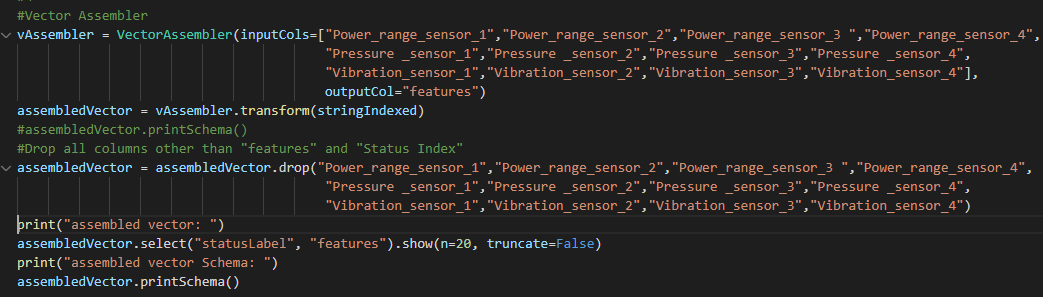
The picture below shows the “Status” Colum and the label(“statusLabel”), String indexer assigns to each element in the DataFrame. Every “statusLabel” element corresponding to “Abnormal” will have the value “0.0” and “Normal” will have the value “1.0”

Now the “Status” column is not needed so we can drop it now. And pass the stringIndexed DataFrame to the vector assembler. The stringIndexer transformer can only output double type values as output and not string, so the index is always going to be numeric values, not strings



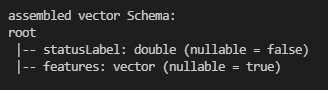
1. Vector Assembler

The step is to combine all the feature columns (see the image below) into a single vector column called features. The DataFrame is classified using the “statusLabel” column, the rest of the feature columns are passed as input columns to the vectorAssembler and it returns “features” column.

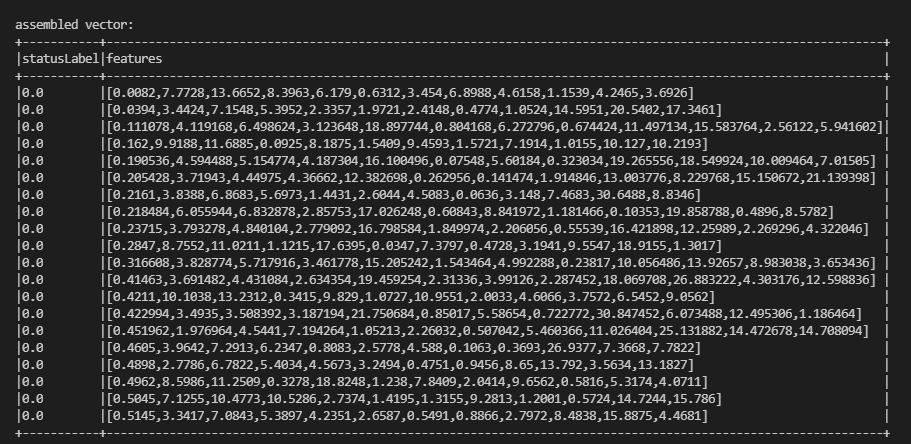


Next drop all other columns other than “statusLabel”(our Label column) and “features”(vectorAssembler output column). Pass the assembledVector DataFrame to the Normaliser. Passing vectorAssembler output features to ML models like Decision trees and regression models can improve their accuracy.

VectorAssembler schema:



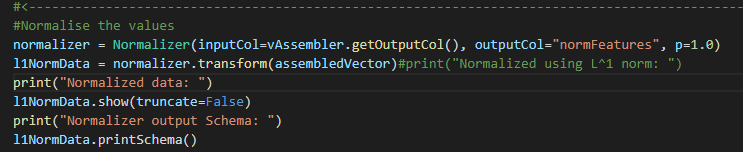
VectorAssembler Output:



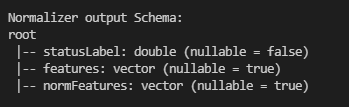
1. Normalizer

This is a transformer that takes in a row of vectors and normalizes each vector to the specified unit form. p is used to specify the p-norm used for normalising.

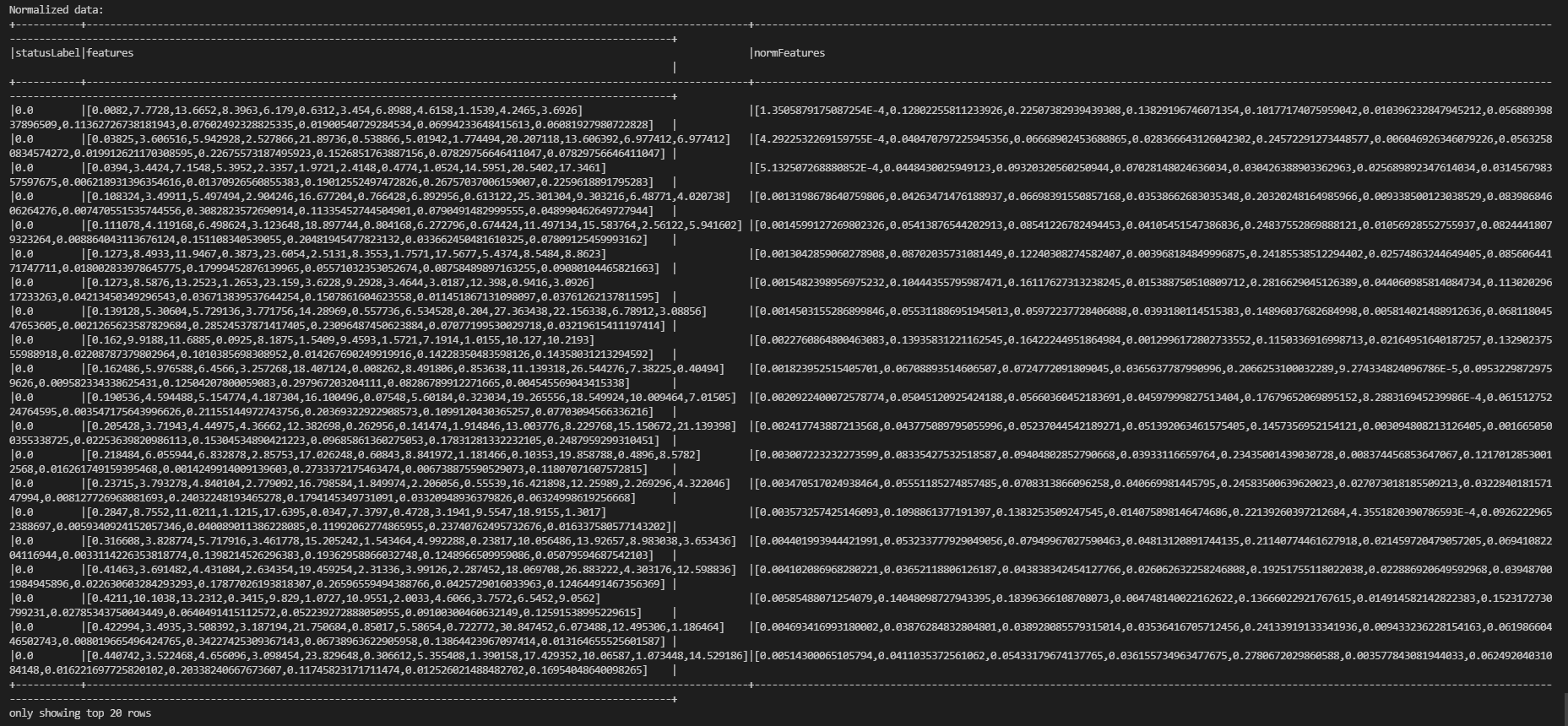
Here we pass the features output column received from vectorAssembler and pass it as the input column, and it returns normalised values of the input which improves the accuracy of Artificial neural network (MPC) algorithms.



Normalizer Schema:



Normaliser output:



The next step is to apply Estimators, which are algorithms that can be fit on a DataFrame’s to produce a Transformer.

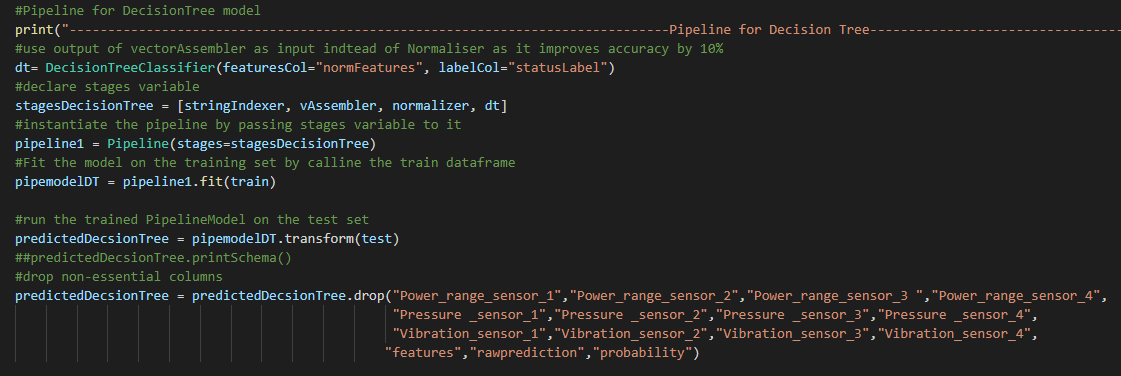
1. Decision Tree classifier

A decision tree is a machine learning classifier that works with categorical features. The decision tree takes as input featuresCol= ”normFeatures”(output of Normalizer) and lableCol= ”statusLabel” from the same.

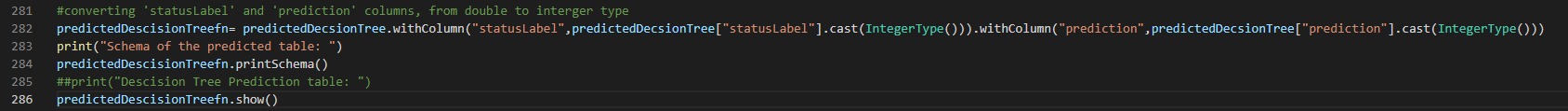
Pipelines are used to chain multiple transformers and Estimators together to specify the ML workflow for the project

To start declaring the stages for the pipeline, used for declaring the order in which the transformer algorithms are going to run in stages, Create the pipeline model for the model and then fit the pipeline on the train dataset.

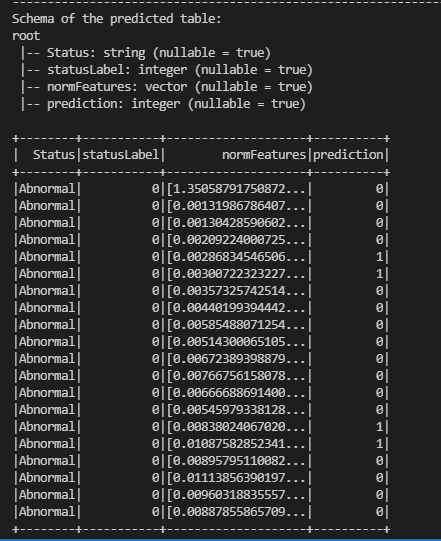
Lastly, run the trained pipeline to the test set using pipemodel.transform(test)



Drop all the non-essential columns and change the prediction value types from double to int.



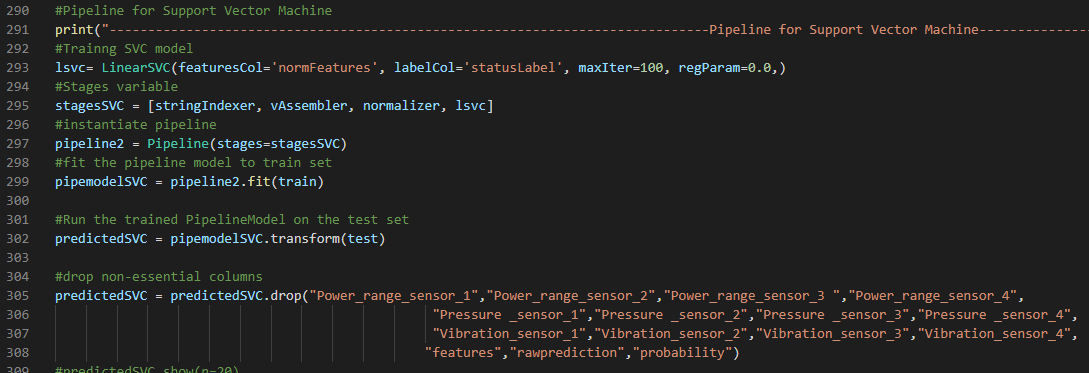
The prediction table for Decision tree is shown below

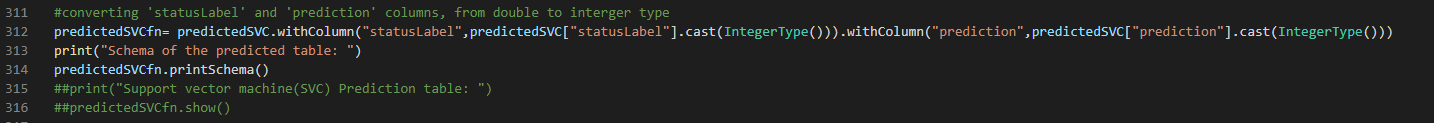


1. Support Vector Machine classifier

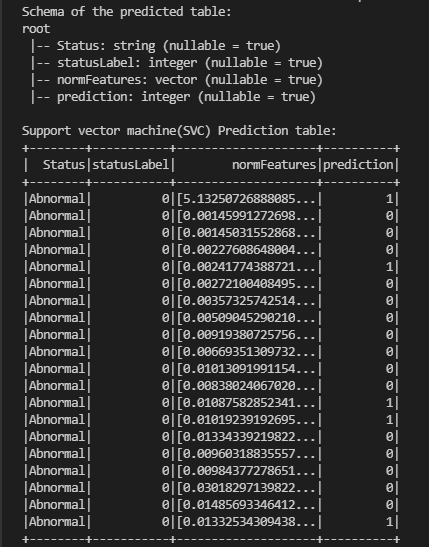
Linear support vector machine does classification by constructing a hyperplane to separate the data classes. SVM takes as input featuresCol = “normFeatures”, lableCol=”StatusLabel”, maxIter and regParam=0.1

The next few steps are the same as in the previous. Declare the stages, create a pipeline with the stages variable, fit the train data to the pipeline and lastly fit the trained pipeline model on the test data, using pipelinemodel.transform(test) method.





Prediction table for the SVM classifier:

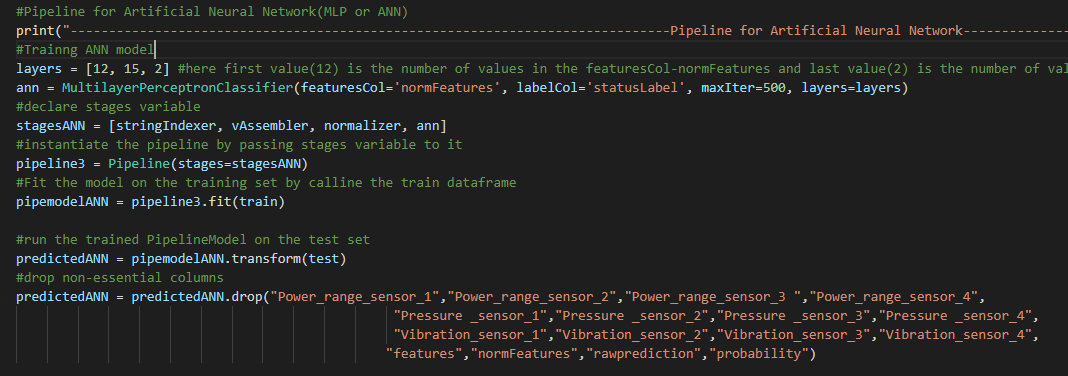


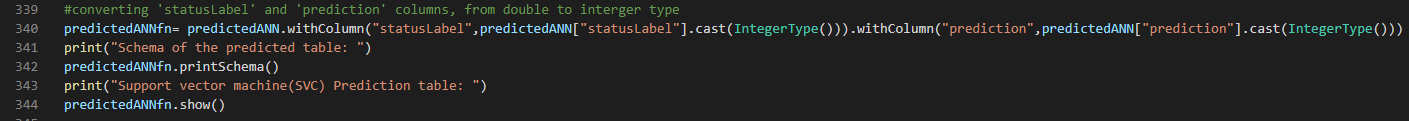
1. Artificial Neural Network:

MPC is a classifier on a feedforward artificial neural network containing multiple layers of nodes that are connected to the next layer of nodes in the network. Nodes in the input layer relate to the input data, the rest of the nodes map inputs to outputs.

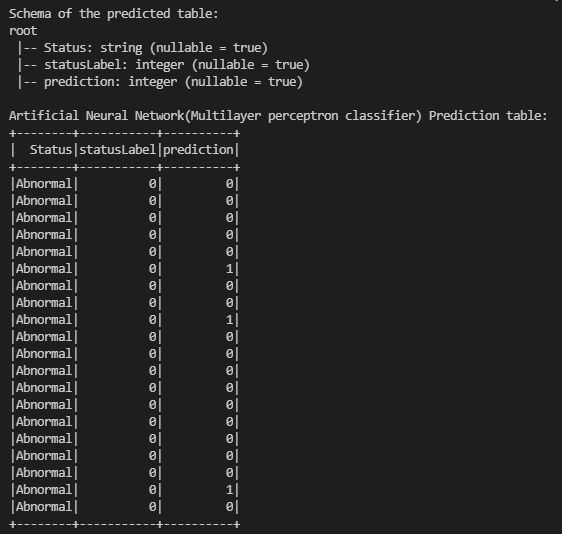
For ANN, we start with defining layers the MPC take “normFeatures” as featuresCol and “statusLabel” as labels and layers= [12, 15, 2].

Next, we define stages, create a pipeline with stages, fit the pipeline to train data, and lastly transform the pipe model to the test set.



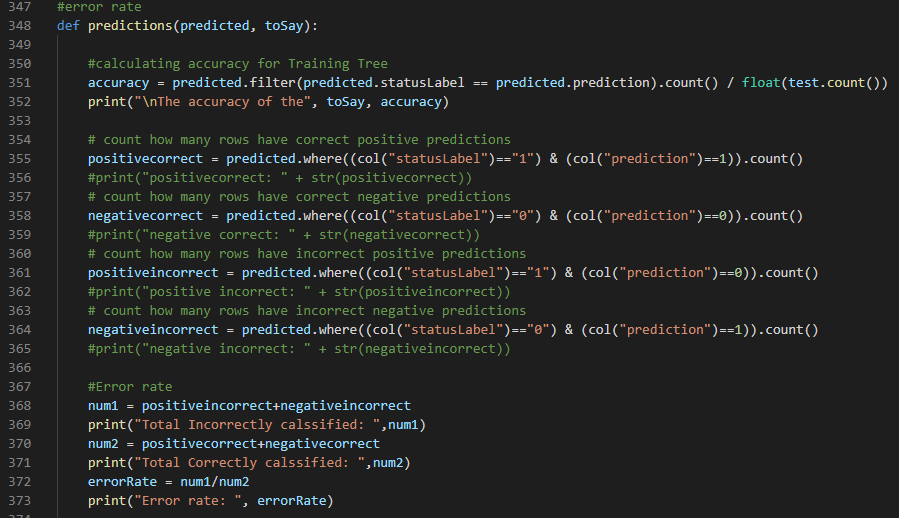


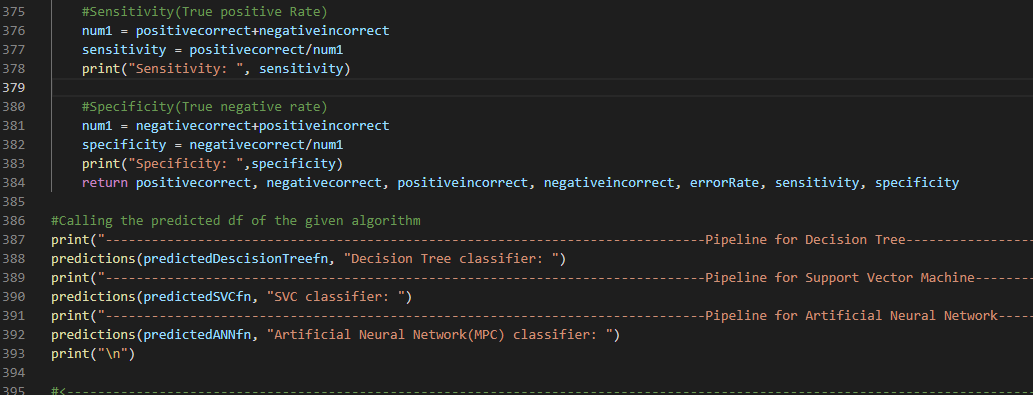
The Schema and prediction table for ANN:



1. Calculate:

To calculate these for all three classifiers function was the best approach all the required values are calculated and printed inside the function.





My Calculation Results:

**Decision Tree**

Error rate: 0.28027681660899656

Sensitivity: 0.821917808219178

Specificity: 0.6153846153846154

**Support Vector Machine**

Error rate: 0.34256055363321797

Sensitivity: 0.6575342465753424

Specificity: 0.6573426573426573

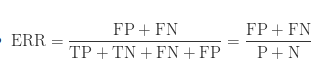
**Artificial Neural Network**

Error rate: 0.19377162629757785

Sensitivity: 0.8013698630136986

Specificity: 0.8111888111888111

Error rate: Error rate is calculated as the number of incorrect predictions divided by the total number of predictions. The best error rate is 0.0 ad it gets worse as it goes up to 1.0 (Basic evaluation measures from the confusion matrix, 2022)



Sensitivity (True Positive Rate): it’s the ratio of true positives (predicted positive) divided by the actual positives in the data. This is important when False Negative (actually positive, predicted negative) is of higher concern than False Positive (actually negative, predicted positive). The best sensitivity is 1, whereas the worst sensitivity is 0.

Crucial when: when getting false negatives are unacceptable



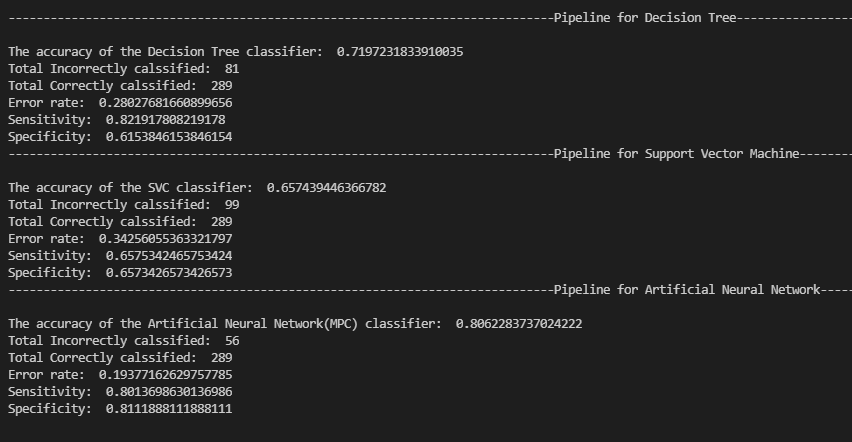
Specificity (true negative rate): is the ratio of true negatives (predicted negative) to the actual negative values in the data. The best specificity is 1, and worst is 0.

important: all true negatives (predicted negative) are needed

Eg: it is important to not raise false alarms. E.g., a Drug test where the report falsely turns out positive could lead to jail time



The below image shows all the calculations the function performs:



**Task 6**

Sensitivity (TPR): having high sensitivity is important because false negative (actually abnormal, predicted normal) predictions in our case will be very dangerous since abnormal states might go unnoticed and lead to disaster. As opposed to having high specificity, would result in only minor inconvenience.

Comparing:

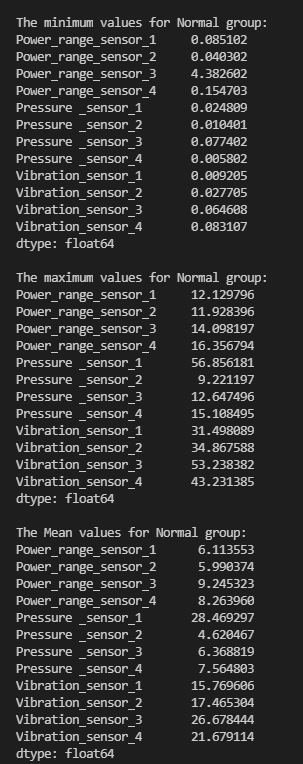
|  |  |  |
| --- | --- | --- |
| **ANN** | **SVM** | **Decision Tree** |
| Has the lowest/best error value | Has the highest/worst error value | Has the second-best error value |
| Has the second-highest sensitivity value (TPR) | Has the highest/worst sensitivity value (TPR) | Has the highest/best sensitivity value (TPR) |
| Has the highest/best Specificity (TNR) value | Has the second-highest Specificity (TNR) value | Has the lowest/best Specificity (TNR) value |

Based on the error rate, sensitivity readings, and specificity. ANN classification is the best for the given dataset as it had the lowest error rate (0.19) along with the highest Specificity value (0.81). At the same time, it has the second-best sensitivity value (0.80) that is very close to the highest value (0.82). Overall, ANN will be the best fit for the dataset.

**Task 7**

When the features are trained using the ANN(MPC) classifier method its accuracy is 80% which can work for predicting the status of the reactor 8 times out of 10. Accuracy can be improved by adding more data to the dataset or fine-tuning the algorithm. SVC and Decision tree methods have lower accuracy than ANN (around 70), so they should be avoided and Ann should be used instead, it would also be recommended to try and improve its accuracy to better predict abnormalities in reactors.

**Task 8**



**References**

sparkbyexamples.com (2022) *PySpark – Find Count of null, None, NaN Values*. [online] Available from https://sparkbyexamples.com/pyspark/pyspark-find-count-of-null-none-nan-values/ [Accessed 24 January 2022].

Kumar, S. (2022) *7 Ways to Handle Missing Values in Machine Learning*. [online] Medium. Available from https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e [Accessed 24 January 2022].

Burris, P., Finos, R., Kobielus, J. and Gilbert, G. (2022). *Wikibon's 2018 Big Data and Analytics Market Share Report - Wikibon Research*. [online] Wikibon Research. Available from https://wikibon.com/wikibons-2018-big-data-analytics-market-share-report/ [Accessed 29 January 2022].

GetSmarter Blog. 2022. *Big Data Analysis Techniques - GetSmarter Blog*. [online] Available from https://www.getsmarter.com/blog/career-advice/big-data-analysis-techniques/ [Accessed 29 January 2022].

Patidar, S. (2022). [image] Available from https://dz2cdn1.dzone.com/storage/temp/9159302-big-data-infographic.jpg [Accessed 29 January 2022].

Durcevic, S. (2022). *What is Ad Hoc Reporting and Analysis? Meaning & Examples*. [online] BI Blog | Data Visualization & Analytics Blog | datapine. Available from https://www.datapine.com/blog/ad-hoc-reporting-analysis-meaning-benefits-examples/ [Accessed 29 January 2022].

Kopanakis, J. (2022). *5 Real-World Examples of How Brands are Using Big Data Analytics | Mentionlytics Blog*. [online] Mentionlytics. Available from https://www.mentionlytics.com/blog/5-real-world-examples-of-how-brands-are-using-big-data-analytics/ [Accessed 29 January 2022].

onix-systems. (2022). *4 types of data analytics Approaches*. [image] Available from https://onix-systems.medium.com/things-you-should-know-about-types-of-data-analysis-40aacf06f13d [Accessed 30 January 2022].

Bock, T. (2022). *What is a Correlation Matrix?*. [online] Displayr. Available from https://www.displayr.com/what-is-a-correlation-matrix/ [Accessed 3 February 2022].

Lutes, J. (2022). *Correlation Is Simple With Seaborn And Pandas*. [online] Medium. Available from https://towardsdatascience.com/correlation-is-simple-with-seaborn-and-pandas-28c28e92701e [Accessed 3 February 2022].

Classifier evaluation with imbalanced datasets. (2022). *Basic evaluation measures from the confusion matrix*. [online] Available from https://classeval.wordpress.com/introduction/basic-evaluation-measures/ [Accessed 3 February 2022].