IITD-AIA FOUNDATION FOR SMART FACTORY AUTOMATION

MACHINE LEARNING WEEK 1 REPORT

Name: Arun kumar D

College: PSG College of Technology

Domain: Machine Learning

Project code: INTP2022-ML-2

Mentor: Devesh taraisa

My project was to identify the Remaining Usable Life Estimation from **the NASA Turbine Dataset**. The first week of the internship involved an informative ride down the data exploratory. Initially the first few days involved around the understanding of the concept why this project is being implemented. I was able to understand that investment in aircrafts is very high. So therefore, the prediction of the remaining useful life this would be beneficiary for them to prevent the complete damage that will be caused. It is also the at most necessary that knowing the remaining useful life of the aircraft turbine will save many people life.

I was able to understand that every machine learning project has 5 phases, they are

- Understanding data and Exploratory Data Analysis
- Model Building and Training
- Model Hyperparameter tuning
- Model comparison and testing
- Model Deployment

So based on this said phases I created a Gantt chart to make myself a timeline of the internship proceeding.

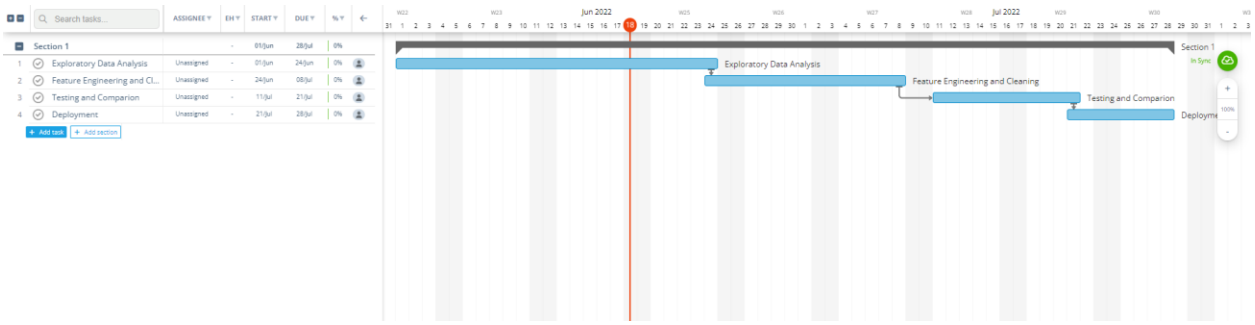

Figure 1: GANTT CHART

Initially I considered only the **FD001** dataset. The dataset shared was in text file format using the Ms Excel I converted them to comma separated value (.csv) format. In that concentrated on the data analysis of the train data. Initially there were 26 columns. The column names had to be named for each column. The columns were

- Unit Number
- Time (in Cycles)
- Operational Setting 1
- Operational Setting 2
- Operational Setting 3
- Sensor-1 Measurement
- Sensor-2 Measurement

  .

  .

- Sensor-20 Measurement

So, when I uploaded the dataset onto the google colab (platform I'm using), there was slight issue where the dataset had 28 columns soi had to drop the last two columns which only had null values. Finally, my dataset had 26 columns for which I had named the column names. So next up to check once again I used **shape ()** function to check the dimensions of the data frame. I took the train dataset directly as 'df'. Next up, I checked for the presence of any null values in the dataset. There weren't any null values in the dataset. I tried to take the data of the maximum cycle each engine failed. But after getting the analyses, it made me understand that it was not necessary for my proceeding. Next, I used the command **describe ()** to know more about the dataset. There I was able to infer

that certain column had same repeated values throughout them. This was understood by comparing the mean, maximum value, and minimum value for which all of them were same. The columns were sensor 1, sensor 5, sensor 6, sensor 10, sensor 16, sensor 18, sensor 19, and operational Setting 3. So to confirm them various analysis where done. First, I plotted the box plot for the sensor values. From the box plot I was able to infer that sensor 1, sensor 5, sensor 6, sensor 10, sensor 16, sensor 18, and sensor 19 have constant values. So confirmed to drop them. Next, I checked for unique entries in the operational setting 3 column. I was able to understand that only one value '100' was recorded so decided to drop them. Next, I checked for the shape of the data frame which led to know the number columns present currently are 18. To find the correlation I checked for the types of the data that are available in the data frame. I was able to find only 2 data types were there that are **int** and **float**. So next found the correlation for the data frame. To analyze relation of the variables with each other. So, after finding the correlation a heat map was drawn to understand the relationship. I was able to understand that the column engine number, operational setting 1and operational setting 2 are weakly related to other columns to dropped them . so finally, I checked the shape of our data frame which was found to be 15. After removing them again a heat map was drawn to analyze.

To conclude this first week was filled with lots of learning about the project, data analysis and to be precise the whole internship itself. So, for the next week I have planned to do the scaling and then to decide which model will be suitable for my statement. My understanding is that after normalization it would model training and then testing so once I have completed all these process for the **FDOO1.** It'll be easy for me to do the same for the other dataset provided.