### **Frameworks of Approach:**

1. **Problem statement**: - Predict breakdown (if the machine will breakdown or not)-classification problem and preventive maintenance to reduce downtime.
2. **Solution Identification**: - Since the dependent variable is categorical variable (dichotomous) the solution would be a building an algorithm which will correctly classify if the machine will break down or not. (use of decision tree)
3. **Solution Implication / Data Exploration and Extraction: -**

* Tools Used: R and r studio.
* Data Extraction: Import data into R using a read.csv () since the file given is a csv file.
* Data integration: since only one Csv file is provided not need of merging csv files.
* Data Assessment: good quality of Data.
* **EDA Analysis:**

1. Looking into the structure of the data using str()
2. Get a brief view of the data and columns available using summary () function.
3. Using Amelia package to visually analyses the missingness and quality of data.
4. If there is any missing variable impute/ignore the column/remove the rows depending on the % of missingness but in this file no missing data.
5. Look for potential outliers using boxplot () function.
6. Univariate and Bivariate analysis since:
7. Dependent variable: broken, independent variable: lifetime, pressure1, pressure2, pressure3, team, provider.
8. Univariate Analysis: team and provider – categorical data (using count and bar chart to visualise), Lifetime, P1, P2, P3- numerical data (using histogram, summary (),boxplot to analyse the distribution of these variables.)
9. Bivariate analysis: - dependent vs independent variables- (using mosaic plot (), boxplots to analyse the bivariate analysis)
10. Feature Engineering: - create calculated field based on the already available variables.
11. **Model Creation and Validation:**

* Splitting the Dataset into Train and test set approx. ratio (70:30)
* Model Used: Decision tree to establish rules, graph.
* Reasons of Choosing Decision tree: Easily interpretable, can establish rules and graphs for decision making and preventive maintenance.
* Higher accuracy.
* Validation: using caret package, to validate the findings.
* Use of confusion matrix to check the actual vs predicted
* Use of ROCR curve to check the area under curve the strength of prediction.
* Enhancing the accuracy using Ensemble methods- Random forest.
* Use of confusion Matrix and ROCR curve to validate the predictions.

By Sidharth Khatua (ID: Jig 16447)