MACHINE PREDICTIVE MAINTENANCE CLASSIFICATION

Under the guidance of Dr. Ritamshirsa Choudhuri

By
Ayush Taparia
Aashirwad Wardhan
Pritam Chowdhury
Aman Singh
Oishik Mukhopadhyay

B.Tech
CSE (Artificial Intelligence and Machine Learning)

CONTENT

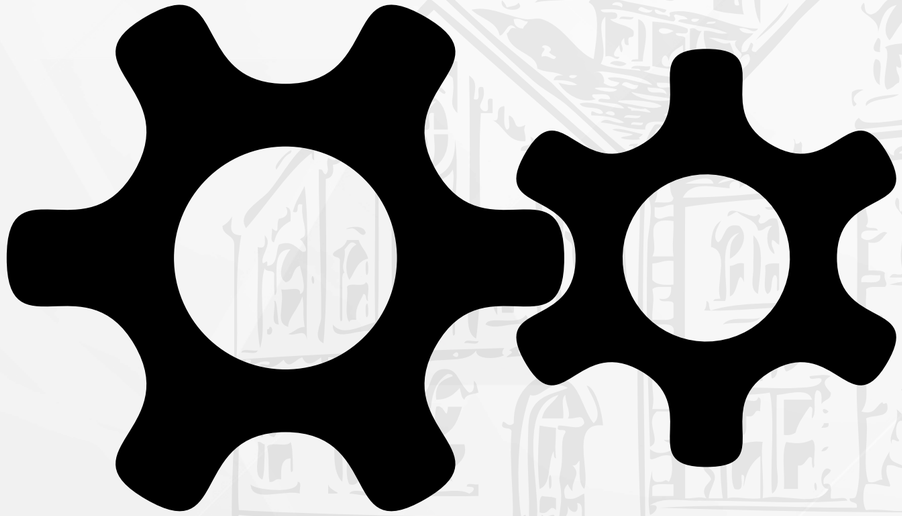
- Aim
- Objectives
- Problem Analysis
- Solutions
- Data Preprocessing
- Exploratory Data Analysis
- Results
- Performance of the Algorithms
- Observations
- Conclusion

WHAT IS OUR MODEL?
AND
WHY IS OUR MODEL?

Under the guidance of Dr. Ritamshirsa Choudhuri

OUR AIM

The need to have a way to determine whether or not a particular machine will fail, as well as the nature of the failure, is essential for generation 4.0 industries



OBJECTIVES

- Way to determine whether or not a particular machine will fail and to repair or replacement of a faulty machine generally requires costs that are much higher than those required for the replacement of a single component.
- The installation of sensors that monitor the state of the machines, collecting the appropriate information, can lead to great savings for industries.
- Machine learning techniques will enhance the productivity of the industries along with the reduction in the input efforts.

Under the guidance of Dr. Ritamshirsa Choudhuri

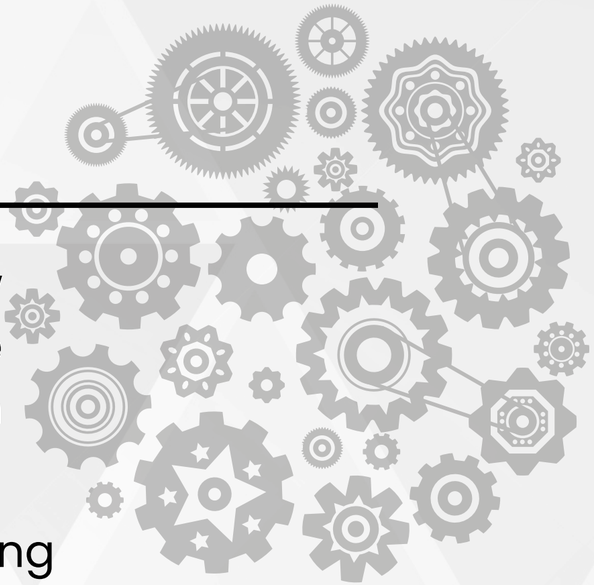
OUR PROBLEMS

- Data Quality and Quantity
- Feature Selection
- Model Complexity
- Task and Data description
- Model Evaluation
- Scale and Accessibility

OUR SOLUTIONS

- Ensure that you have high-quality and sufficient data. Collect data from reliable sources.
- Conduct a thorough analysis to identify the most relevant features affecting machines . Use domain knowledge and statistical methods to select the best features
- Choose a model that suits your dataset size and complexity.
- Since real predictive maintenance datasets are generally difficult to obtain the data provided by the UCI repository is a synthetic dataset that reflects real predictive maintenance encountered in industry to the best of their knowledge.
- With the aim of obtaining a ratio of 80 to 20 between functioning and faulty observations and the same percentage of occurrence between the causes involved in the failures.
- The need to have a way to determine whether or not a particular machine will fail, as well as the nature of the failure, is essential for generation 4.0 industries.

DATA PREPROCESSING

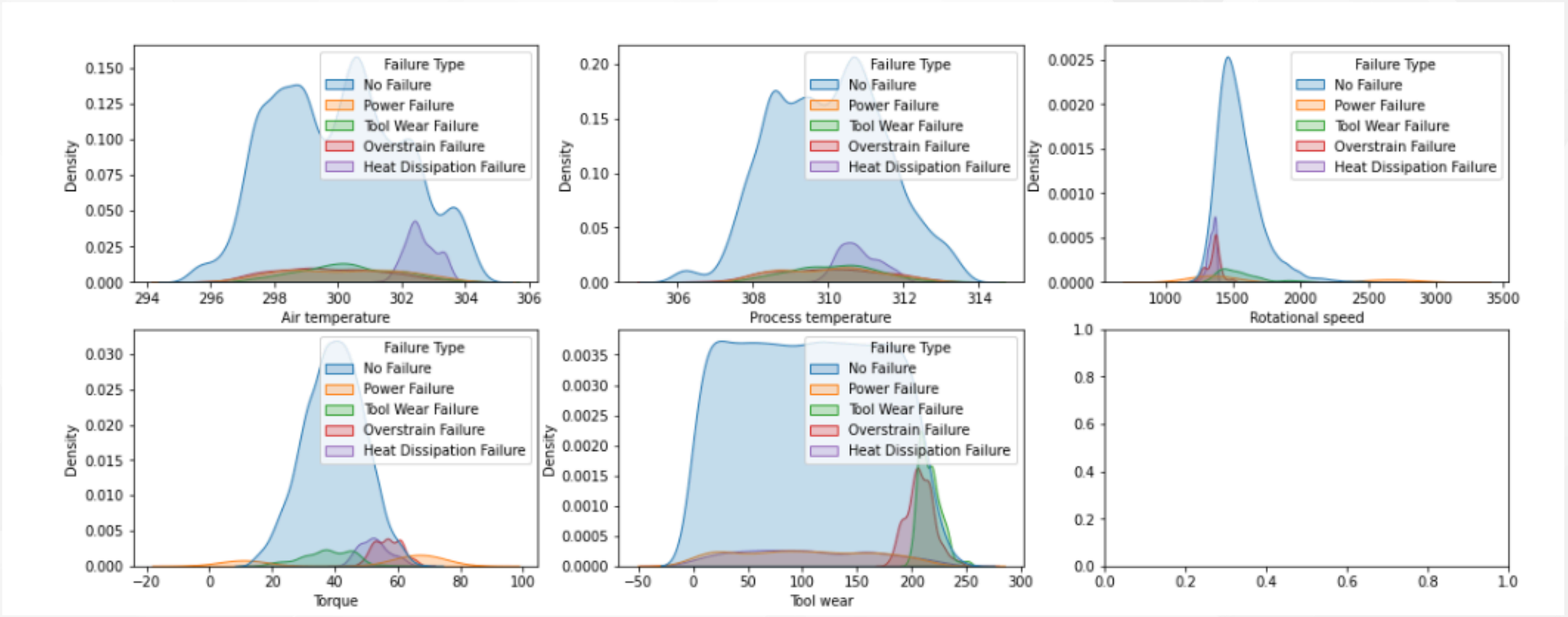


Data preprocessing is a vital process that involves preparing raw data and transforming it into a format suitable for a machine learning model. It serves as the initial step in developing an effective machine learning model.

- **Data Cleaning -** From duplicates and outliers to missing numbers, it fixes them all.
- **Data Integration-** This process integrates (merges) information extracted from multiple sources to outline and create a single dataset.
- **Data Transformation-** This process entails putting the data in a format that will allow for analysis. Normalization, standardization, and discretisation are common data transformation procedures.
- **Data Reduction-** Data reduction is the process of lowering the dataset's size while maintaining crucial information.

EXPLORATORY DATA ANALYSIS

After performing cleaning and normalizing the datasets we performed Exploratory Data Analysis to get a better understanding . Through plotting and analyzing the graph we came to know a lot of useful information and could understand the dataset better.

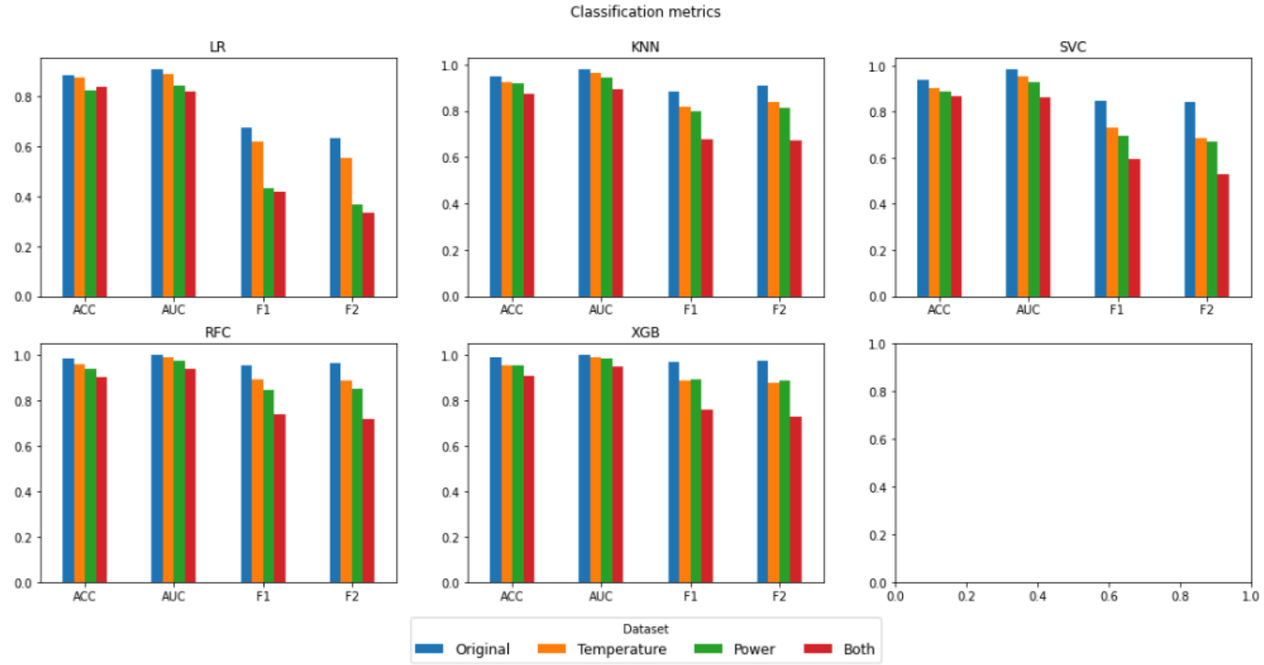


RESULTS

After our Exploratory Analysis of the dataset we arrived at various facts about the Machine failure scenario as mentioned in observation. We trained 4-5 Machine Learning Models on the final dataset. We have calculated the score and results are as follows:

PERFORMANCE OF THE ALGORITHMS

- **Logistic Regression**
- **K-nearest neighbors (K-NN)**
- **Random Forest Classification**
- **Support Vector Machine**
- **XGBoost**

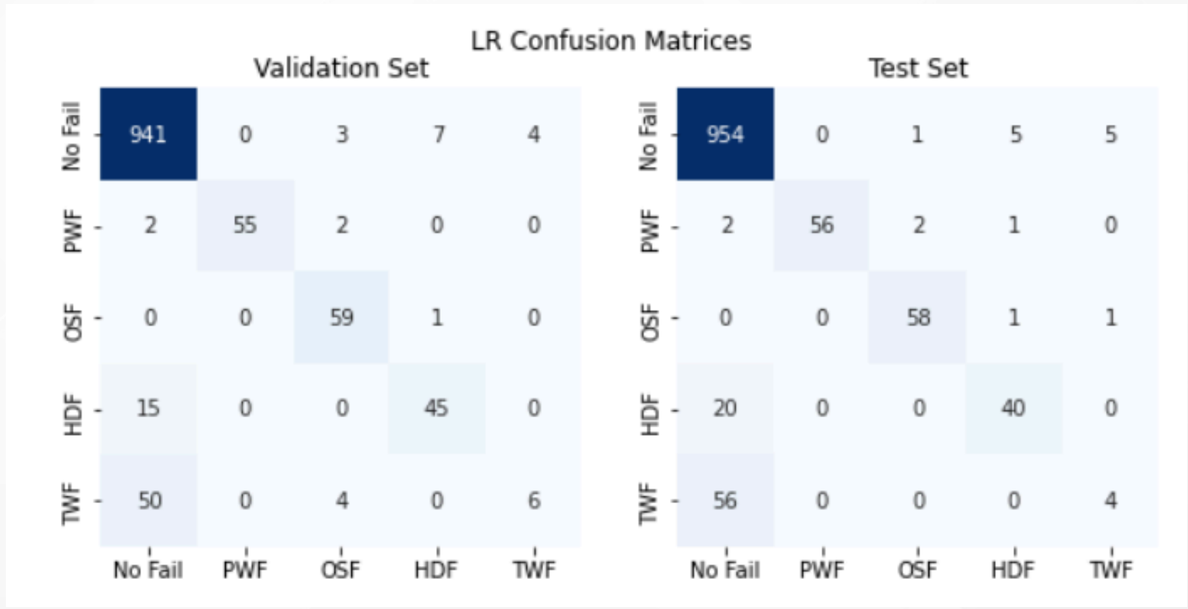


ALGORITHMS

Logistic Regression

lit estimates the probability of a dependent variable as a function of independent variables. The dependent variable is the output that we are trying to predict while the independent variables or explanatory variables are the factors that we feel could influence the output. For its simplicity and interpretability, we decide to use Logistic Regression as a Benchmark model, a basic model that represents the starting point for comparing the results obtained from other models.

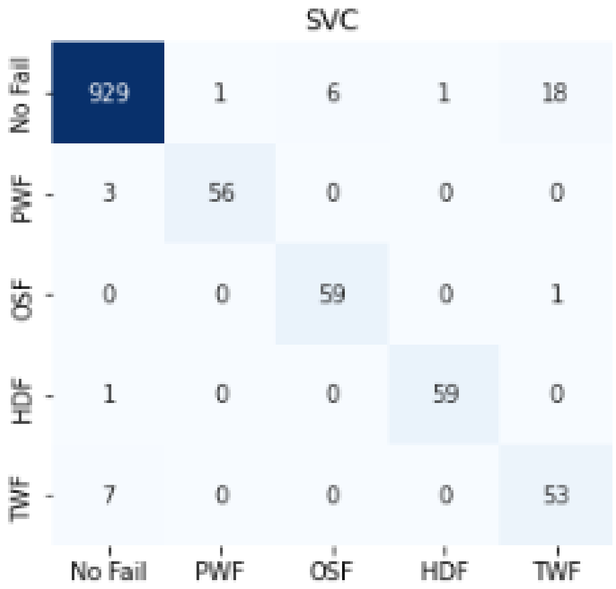
```
Validation set metrics:
ACC     0.926
AUC     0.983
F1      0.909
F2      0.919
dtype: float64
Test set metrics:
ACC     0.922
AUC     0.982
F1      0.904
F2      0.914
dtype: float64
```



ALGORITHMS

Support Vector Machine

its aim is to find a hyperplane in an N-dimensional space (N—the number of features) that distinctly classifies the data points while maximizing the margin distance, i.e. the distance between data points of both classes.



Validation scores:

	KNN	SVC	RFC	XGB
ACC	0.959	0.966	0.977	0.987
AUC	0.954	0.987	0.997	0.999
F1	0.902	0.916	0.943	0.969
F2	0.928	0.931	0.954	0.970

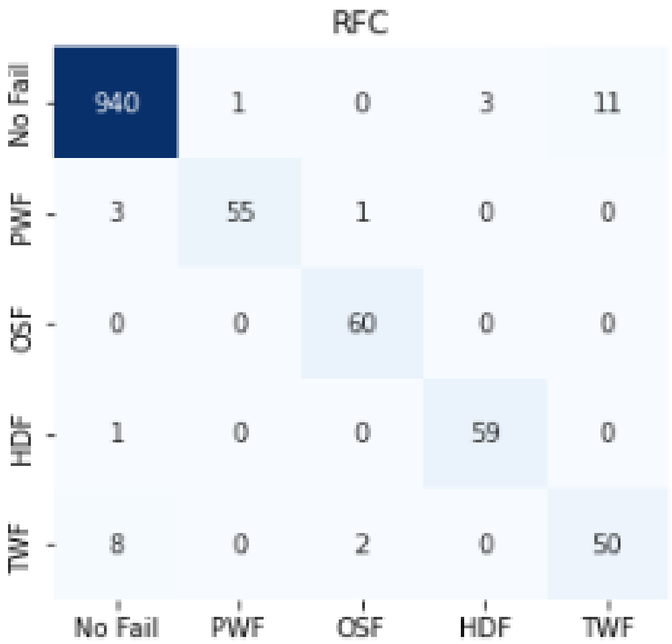
Test scores:

	KNN	SVC	RFC	XGB
ACC	0.966	0.973	0.972	0.983
AUC	0.954	0.992	0.997	0.998
F1	0.916	0.934	0.931	0.958
F2	0.927	0.941	0.945	0.956

ALGORITHMS

Random Forest Algorithm

it uses ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. Random Forest uses bagging technique: it constructs a multitude of decision trees in parallel, all with the same importance, and the output is the class selected by most trees.



Validation scores:

	KNN	SVC	RFC	XGB
ACC	0.959	0.966	0.977	0.987
AUC	0.954	0.987	0.997	0.999
F1	0.902	0.916	0.943	0.969
F2	0.928	0.931	0.954	0.970

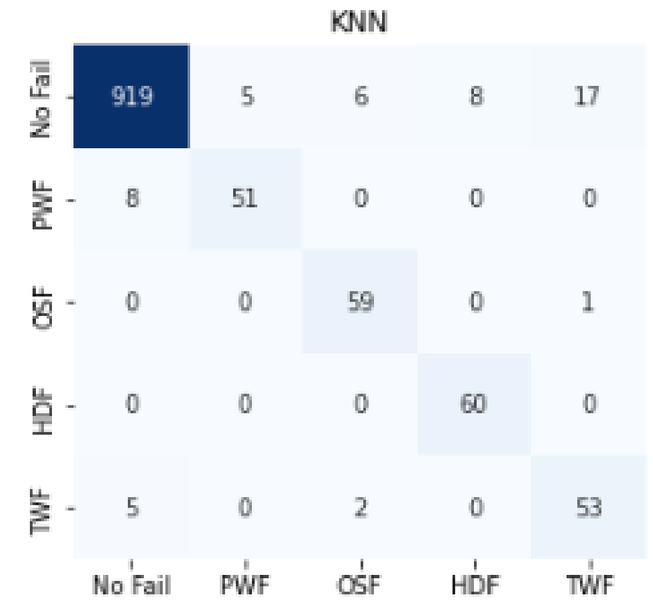
Test scores:

	KNN	SVC	RFC	XGB
ACC	0.966	0.973	0.972	0.983
AUC	0.954	0.992	0.997	0.998
F1	0.916	0.934	0.931	0.958
F2	0.927	0.941	0.945	0.956

ALGORITHMS

K-Nearest Neighbors Algorithm

KNN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. In order to determine which data points are closest to a given query point, the distance between the query point and the other data points will need to be calculated.



Validation scores:

	KNN	SVC	RFC	XGB
ACC	0.959	0.966	0.977	0.987
AUC	0.954	0.987	0.997	0.999
F1	0.902	0.916	0.943	0.969
F2	0.928	0.931	0.954	0.970

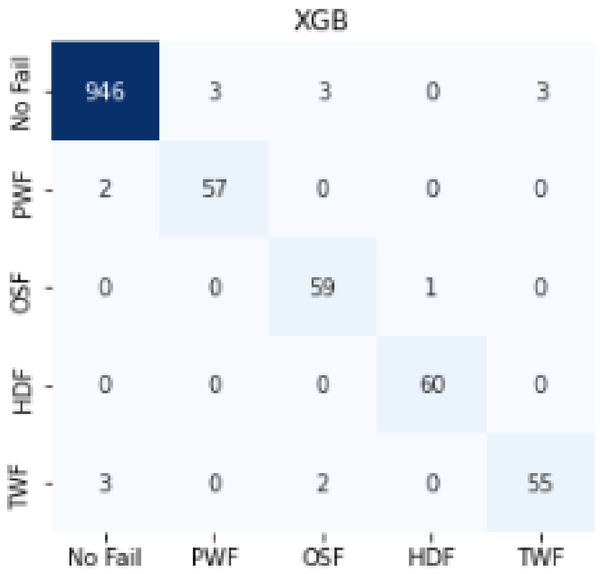
Test scores:

	KNN	SVC	RFC	XGB
ACC	0.966	0.973	0.972	0.983
AUC	0.954	0.992	0.997	0.998
F1	0.916	0.934	0.931	0.958
F2	0.927	0.941	0.945	0.956

ALGORITHMS

XGBoost

is a gradient-boosted decision tree (GBDT) machine learning library. A Gradient Boosting Decision Tree (GBDT) is a decision tree ensemble learning algorithm similar to Random Forest, from which differs because it uses a boosting technique: it iteratively trains an ensemble of shallow decision trees, with each iteration using the error residuals of the previous model to fit the next model. The final prediction is a weighted sum of all of the tree predictions.



Validation scores:

	KNN	SVC	RFC	XGB
ACC	0.959	0.966	0.977	0.987
AUC	0.954	0.987	0.997	0.999
F1	0.902	0.916	0.943	0.969
F2	0.928	0.931	0.954	0.970

Test scores:

	KNN	SVC	RFC	XGB
ACC	0.966	0.973	0.972	0.983
AUC	0.954	0.992	0.997	0.998
F1	0.916	0.934	0.931	0.958
F2	0.927	0.941	0.945	0.956

OBSERVATIONS

- By comparing the results obtained, we see that K-NN is the model that performs the worst and its accuracy is a little lower than Logistic Regression's one. Despite this, we cannot exclude it a priori, as it still reaches high values for the metrics and, moreover, gives an immediate response. So, we can use it whenever we need to get an idea quickly about the situation and, then apply other models when we have more time.
- SVC and RFC's performances are very similar each other and XGB performs better than them
- If we look at the training phase, SVC and RFC take the same time, while XGB takes more than four times as much as them. So, since, the improvement obtained with XGB is only 1.5%, one can choose which model he prefers according to his needs.
- While the best parameters for multiclass K-NN and SVC are the same as binary classification, for XGB and RFC the Gridsearch for the two types of task returns different parameters..

CONCLUSION

According to the analyses carried out and the results obtained, it is possible to make some conclusive considerations related to this project.
We decided to tackle two tasks: predict whether a machine will fail or not and predict the type of failure that will occur. . Briefly, in preprocessing phase we have deleted some ambiguous samples, we applied a label encoding to the categorical columns and then we performed the scaling of the columns with StandardScaler. We also noticed the presence of some data points which at first we referred as outliers but later turned out to be part of the natural variance of the data and played an important role in the classification task. Then we ran PCA and found that most of the variance is explained by the first three components, that can be represented as the following features: combination of the two Temperatures, Machine Power (product of Rotational Speed and Torque) and Tool Wear.
At the end, we can conclude that for both task the chosen models perform very well. For both tasks the best model is XGBoost and the worst is KNN; however the response time of KNN is instant while XGBoost takes more time and this further increase when we proceed with the multi-class classification task. The choice of the model depends on the needs of the company: for faster application one can use KNN while if one cares more about accuracy one can use XGBoost.

THANK YOU!

Under the guidance of Dr. Ritamshirsa Choudhuri