TEAMMATES:

- **Saransh Agarwal(2019MT60763)**
- **Avadhesh prasad(2019MT60747)**
- **Aman Chauhan(2019MT60742)**

CHARACTERIZATION OF THE DATASET

What the data is about?

The dataset is about sensorless drive diagnosis. The sensors are too expensive to monitor every drive unit in a single factory. In this dataset the phase currents are used to predict the state of a drive and allow for predictive maintenance. Features are extracted from electric current drive signals. The drive has intact and defective components. This results in 11 different classes with different conditions. Each condition has been measured several times by 12 different operating conditions, this means by different speeds, load moments and load forces. The current signals are measured with a current probe and an oscilloscope on two phases. There is total 49 attributes and 58509 instances in the dataset.

Covariance and correlation are calculated for the dataset. Boxplot is also plotted for the dataset.

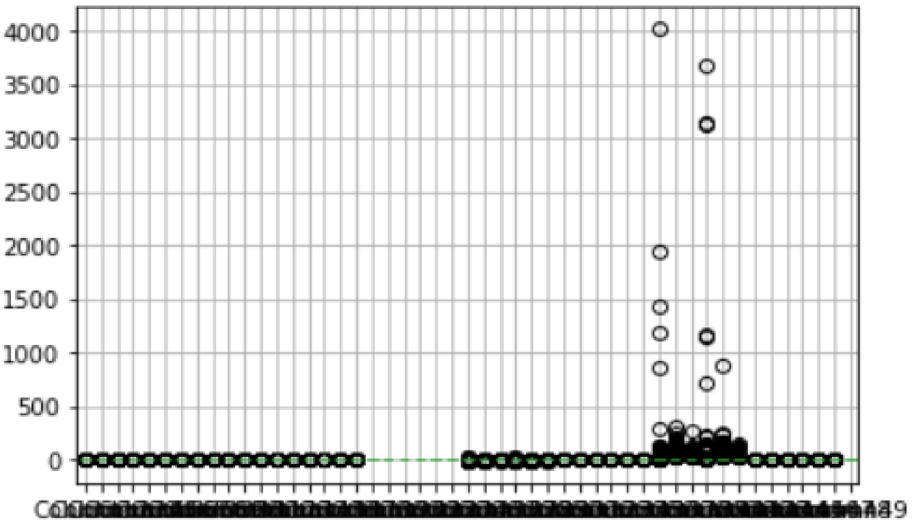
What type of benefit you might hope to get from data mining?

Data mining will be helpful to predict future trends and obtaining new information and creating value from present measurements without introducing additional sensors is cost-efficient and mitigates data that is collected and stored by information systems but not used.

Discussing Data quality issues

- Problems with the dataset

  The variation in values of attributes is quite significant that can be observed from the boxplot. Some attributes like attribute number from 1-6 have values in order of 10^{-6} while some attribute like attribute number 38,39 have huge values.



Boxplot

- Appropriate response to the data quality issues.

To overcome the above stated problem, standardization of the dataset is used.

<u>Comparing The Performances Using k-fold cross validation</u>

CLASSIFIER	TRAINING SCORE	CROSS VALIDATION SCORE
DECISION TREE	0.9965	0.9972
RANDOM FOREST	0.9999	0.9991
NAIVE BAYES	0.7562	0.7312
KNN	0.9927	0.9802

Comparing the performance using 5-fold cross validation

CONCLUSION:

We have tuned the hyperparameters for all the classifiers and then Training score and cross validation score is calculated using the best hyperparameters obtained by the hyperparameter tuning. We can observe that for our dataset Random Forest give the highest accuracy while naïve bayes classifier perform poorly on our dataset. The naive bayes classifier assume independence of the attributes while in our datasets the attributes are not independent, which can be observed form the correlation values due to which naive bayes classifier gives low accuracy.