**MINI PROJECT (SEM IV)**

**OUTLIER DETECTION**

**NAME: TANYA CHETNA VAISH**

**ROLL NO: 2013553**

**PROBLEM STATEMENT:**

Involves *detection of outliers* from the dataset.

**MOTIVATION:**

A subset of artificial intelligence, machine learning allows systems to make predictions and crucial business decisions, driven by data and pattern-based experiences. Without humans having to intervene, the algorithms that are fed to the systems are helping them develop and improve their own models and understanding of a certain use-case. Over the course of my engineering, I have a built a keen interest in this field, and hoping to build career in this field. Since, I am pursuing specialisations in *Machine Learning*, I thought of choosing this topic and experiment my knowledge regarding *outliers in dataset* and their *detection and removal.*

**WHAT ARE OUTLIERS?**

* A statistical outlier is any datapoint in a dataset that is beyond a pre-defined distribution range, usually representing a measurement error or abnormal data that should not be included.
* Outliers can be visually determined based on a plotted graph of the data samples.

Since the root of all deep learning training techniques is analysing vast amounts of data to find some sort of mathematical pattern or relationship, outliers can produce all sorts of “ghosts” in a machine program if not weeded out early.

* Outliers can be of two kinds: **univariate** and **multivariate**.

**HOW TO MAKE OUR MACHINE LEARNING MODELS ROBUST TO OUTLIERS?**

We can use descriptive analytics-based techniques to start with. For visualization box plot is a good starting point to identify outliers. For more complex problems using density and distance-based methods help in identifying outliers - basic idea is outliers are farther than their neighbours than other normal points. based on same principle you can use nearest neighbour-based machine learning techniques or clustering techniques also.

Some of the most popular techniques for outlier detection are-zscore or extreme value analysis, Linear Regression Models (PCA, LMS), Probabilistic and Statistical Modeling (parametric), Proximity Based Models (non-parametric).

**DATASET USED: *Glass Identification Data Set from UCI*** **Content**

(*Source:* [*https://archive.ics.uci.edu/ml/datasets/Glass+Identification*](https://archive.ics.uci.edu/ml/datasets/Glass+Identification)*)*

**METHODLOGY:**

*Link to the code on my GitHub repository:* [*https://github.com/TanyaChetnaVaish/MiniProjectSem4*](https://github.com/TanyaChetnaVaish/MiniProjectSem4)

In this dataset, I used Box-Plot method to visualise dataset for outliers. A **boxplot**, sometimes called a box and whisker plot, is a type of graph used to display patterns of quantitative data.

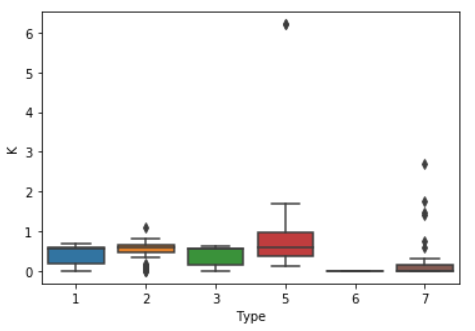
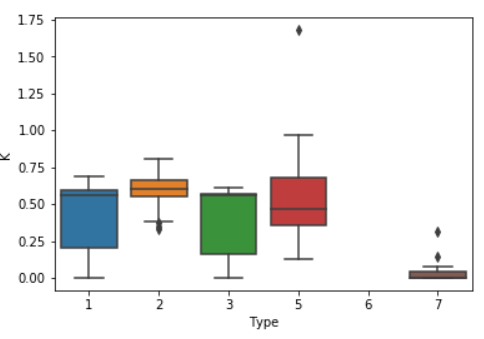
* Imported major libraries-pandas, numpy, seaborn, matplotlib. pyplot, skelearn, sklearn. model\_selection, sklearn.preprocessing.I imported sklearn.model\_selection to implement GridSearchCV
* Plot the Boxplot to check for outliers, for each column against the target variable.

The diamond shaped dots in the boxplot represents Outliers. And there are extreme cases in K, Ba and Fe.

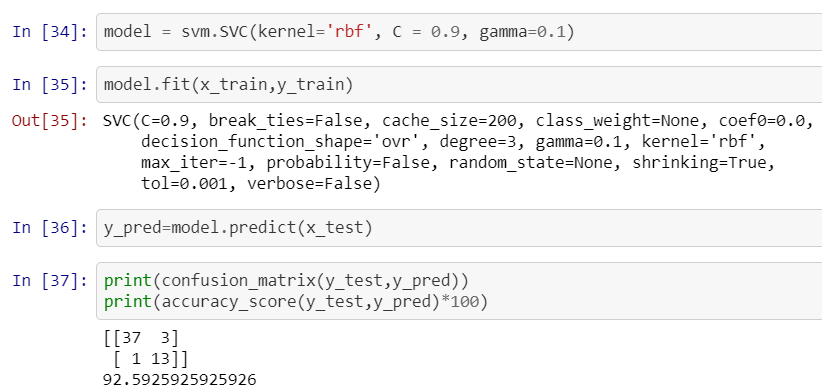
* In this project I used two methods for removing outliers-a) quantile method and b) dropping columns depending on more or less outliers and splitting the data.

Here is the data visualisation for each column, before and after removing outliers (using Boxplot) for K column

Before: After:

* Further, for prediction I used GridSearchCv. Grid search is the process of performing hyper parameter tuning in order to determine the optimal values for a given model. This is significant as the performance of the entire model is based on the hyper parameter values specified. We can also use other algorithms like xgboost, Random forest, KNN and so on.



I accuracy score came out to be 92.5%. I leant to implement GridSearchCv for accuracy calculation. I wanted to learn this algorithm so I learnt and implemented it.I hope I did it right. Thank you!