IST 718 – M003

Project Abstract – Group 4

**Malware Detection on Microsoft Systems**

* **Team Member Names:**

1. Pranav Kottoli Radhakrishna
2. Prankur Garg
3. Sarthak Tandon
4. Tejas Patil

* **Brief Description of Dataset:**

This dataset contains 82 independent assessment parameters which are a mix of categorical and numerical type. The target variable ‘HasDetections’ is a binary attribute indicating whether that machine was infected by malware or not. Malware detection is inherently a time-series problem, but it is made complicated by the introduction of different kinds of machines, for e.g. new machines, machines that receive new operating systems, online-offline machines, etc.

Link to the dataset: [Microsoft Malware Detection Dataset](https://www.kaggle.com/c/microsoft-malware-prediction/data)

* **Business Problem:**

Malware is any software intentionally designed to cause damage to any computer, server, client or computer network. Malware exploits security defects in the design of operating system, in applications or in vulnerable versions of browser plugins. The dataset contains the information about different machines with windows operating systems and their attributes that will help in determining the likelihood of that machine getting infected by malware. For preventing such instances happening in future, it is utmost important to find out whether the computer is infected by malware and resolve the issue accordingly.

* **Data Science Methodology:**

1. The training dataset here contains more than 8.9 million instances. Hence, we will be using stratified sampling to pick a sample that will represent the entire dataset.
2. We will begin with data cleaning which includes treating for missing values, special characters and outliers.
3. Then we’ll perform exploratory data analysis by statically and numerically analyzing the data using data visualizations.
4. Next, we will split the dataset into training and testing datasets. The training set will be used to build machine learning models and the test set will be used to validate the output of the models.
5. We will use classification algorithms such as logistic regression, decision tree and random forest to predict if the machine will be infected by malware.
6. Finally, we will use the test dataset to calculate evaluation parameters to assess and compare the performance of different models.