**Executive Summary – Group B**

**Microsoft Malware Detection**

Anchal Atlani – [aatlani@unc.edu](mailto:aatlani@unc.edu) – 801084129  
Anushka Kamalapurkar – [akamala1@uncc.edu](mailto:akamala1@uncc.edu) – 801073705  
Anup Bharade – [abharade@uncc.edu](mailto:abharade@uncc.edu) – 801077163  
Apurva Nanajkar – [ananajka@uncc.edu](mailto:ananajka@uncc.edu) – 801081289  
Mandar Phapale – [mphapale@uncc.edu](mailto:mphapale@uncc.edu) – 801075522  
Rithesh Shetty – [rshetty2@uncc.edu](mailto:rshetty2@uncc.edu) – 801058230  
Nikhil Jadhav – [njadhav@uncc.edu](mailto:njadhav@uncc.edu) – 801075504  
Sagar Kulkarni – [skulka23@uncc.edu](mailto:skulka23@uncc.edu) – 801073394  
Soham Dhodapkar – [sdhodapk@uncc.edu](mailto:sdhodapk@uncc.edu) – 801075881  
Shravanti Bugade – [sbugade@uncc.edu](mailto:sbugade@uncc.edu) – 801076033  
Omkar Marne – [omarne@uncc.edu](mailto:omarne@uncc.edu) - 801076695

Malware attack is one of the most sever security threat in today’s computerized world. Malware attack spreads by exposing system vulnerabilities or by taking advantage of negligence of users responsible for using or administering the system. To protect a computer system or to eliminate malware from an existing compromised system, it is extremely necessary to detect malware. The most practical method in malware detection process can detect evident classes of malware. But this process various theoretical limitation which is the process is unable to detect new type of malware. Besides this limitation there are couple of other challenges: 1) The poly and metamorphism techniques used to dodge malware detection system, 2) The speed with which the malware is generated and distributed across the network.

In this Paper, we have taken a dataset provided by Microsoft which has unprecedented malware data in tabular format. We have shown predictions using different Analysis techniques like Classification, Clustering, etc. We have used feature selection and dimensionality reduction techniques to eliminate the features which are not required for further analysis. We intend to check the prediction accuracy and error rate of our prediction to analyze which model is most suitable for this data.

The dataset used is from Microsoft competition held on Kaggle. Most of the variables in this dataset are binary variables, also variable ‘HasDetections’ is used as the output variable. This output variable is also a binary variable. We will be running different models of classification like classification tree, KNN, and logistic regression.

Using different performance metric like accuracy, precision, confusion matrix etc. we intend to find the best possible model for this dataset. We will run the classification tree algorithm in XLMiner and will identify the top features which will affect our classification. We are planning to run the KNN classification algorithm and Logistic Regression and going to compare these methods on the basis of error and accuracy. Apart from classifying the data, we are planning to find the association rules between the categorical variables which will give more insight about the data and can be used to infer the inherent correlation between the variables. Lastly, we will apply the k-means clustering technique using Census\_PowerPlatformRoleName” independent variable and our output variable “HasDetections” and find the best cluster.