**AI & Cybersecurity**

Midterm Project Report

DSCI6672 – spring 2020

**A close up of a logo

Description automatically generated**

**BY**

**Greeshma Bachala**

**Introduction:**

In this project network intrusion detection, the intrusion or attack in the network connection are anomaly or normal. The network connection is delivering the packets from one source to another, the environment created is to acquire raw TCP/IP dump data for a network. An association is a grouping of TCP packets starting to end duration in between which the data packets streams to and from source IP to destination IP address under some defined set of protocols. Also, each connection is labelled as either normal or as an attack with exactly one specific type. Each connection record consists of about 100 bytes.

For each TCP/IP connection, 41 quantitative and qualitative features are obtained from normal and attack data (3 qualitative and 38 quantitative features). The class variable has two categories are normal and anomalous.

To solve this, implement the machine learning techniques to predict the attack whether it is anomaly or normal in this network connection. The KNN classifier, logistic regression and Decision tree classifier are used to predict the best model with high accuracy and predict which attack is it.

**Data preparation steps:**

* Import the required libraries
* Load the data from source
* Create a data frame with the training set
* Statistical summary of the data (exploratory data analysis)
* Dropping the specified columns from the data frame
* Scaling the numerical features by preprocessing
* Encoding the categorical features by standard scalar
* Selecting the important features using RFE.

Exploratory data analysis is performed to level up all the features and to avoid the variations in the results. dropping the specified columns from the data frame with infinite values.

Recursive feature elimination to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through a coef attribute or through a feature importance attribute. least important features in data frame are pruned from the current set of features.

**Standardization** of datasets is a **common requirement for many machine learning estimators** implemented in scikit-learn; they might behave badly if the individual features do not more or less look like standard normally distributed data with **zero mean and unit variance**.

A screenshot of a social media post

Description automatically generated

The above picture tells you the important features after the scaling and encoding the numerical and categorical features. It removes all the unwanted features and plotted the graph with the range of the sorted values based on their importance.

**Splitting the data:**

Splitting up the data with function from sklearn test\_train\_split with train size 0.70. splitting the data into training and testing data by 70 and 30 percent. To work on training set and test on test set that model is worked or fitted.

**Implementation:**

The implementation is done with the machine learning classifiers, KNN classifiers, Random forest classifier, logistic regression, and decision tree classifier. I also tried to implement the keras model by adding layers using the sequential model but shown up error with the input dimensions. So, I got sticked to implement the machine learning models to predict the best fit model to predict the high accuracy. Evaluating the model by cross validation score, confusion matrix and classification report includes macro average, weighted average, and accuracy. Predicting the best model on all three and predicts the type of attack. The evaluation models validate the models on the cross-validation score and accuracy.

A screenshot of a computer

Description automatically generated

Random forest classifier gives the best and high accuracy it is the best fit model compared to all other models.

A screenshot of a cell phone

Description automatically generated

Test results shown the best fit model as random forest classifier with 99.64 percent as model accuracy. Classification report gives you the precision 1.0, recall 1.0, f1-score 1.0, macro average and weighted average as 1.0. this shows you the best algorithm for intrusion detection is random forest classifier.

A screenshot of a social media post

Description automatically generated

Prediction of which type of intrusion in this network. The picture shows the anomaly and normal which class it belongs to.

**Conclusion:**

This project is about the network intrusion detection which type is the attack in the transfer of data packets in TCP/IP connection. The anomaly attack, which is not normal, unusual, and not as expected. In this I have learned the small details involved in the preprocessing steps. The imblearn combine provides methods which combine over-sampling and under-sampling. The scaling which is performed by using standard scalar function. Using the importance feature by recursive feature elimination. This helps in pruning the less important features in the data frame. Random forest classifier is the best fit model as it performs into the deeper subset of features, splitting on a random subset of features. This gave the which type of attack is in the network connection.