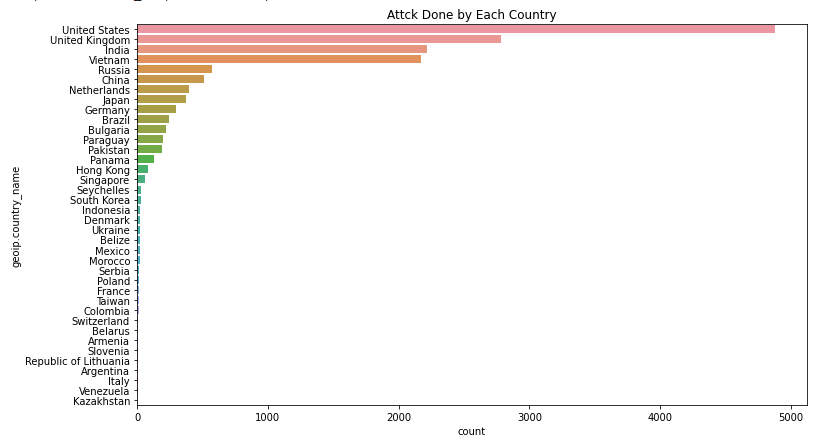
**3Data Analysis and Result:**

Attack Done by Each Country



After successfully plotting graph “Attak done from each country”

We are able to predict Top 5 attack are from US,UK,India,Russia,China

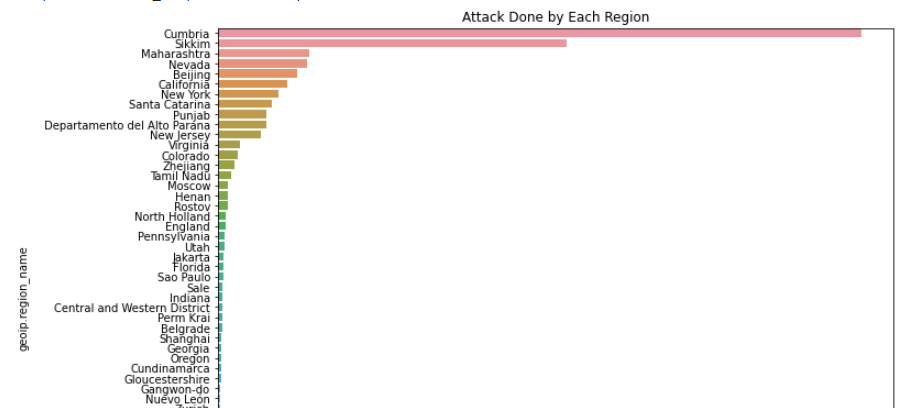
Attacks Done by Each City



After successfully plotting graph “Attak done from each City”

We are able to predict Top 5 attack are from Miliom,Gantok,Mumbai,Las Vegas and Forguilhinha

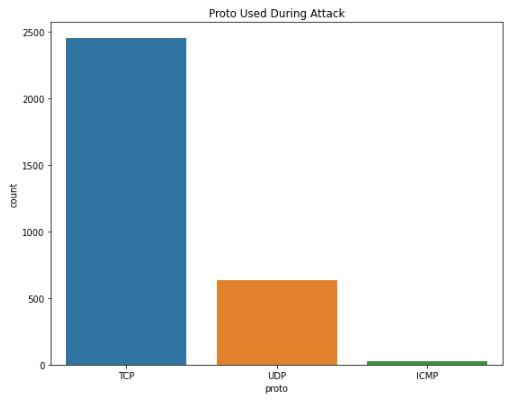
Attack Done by Each Region



After successfully plotting graph “Attak done from each Region”

We are able to predict Top 5 attack are from Cumbria,Sikkim,Maharashtra,Nevada and California.

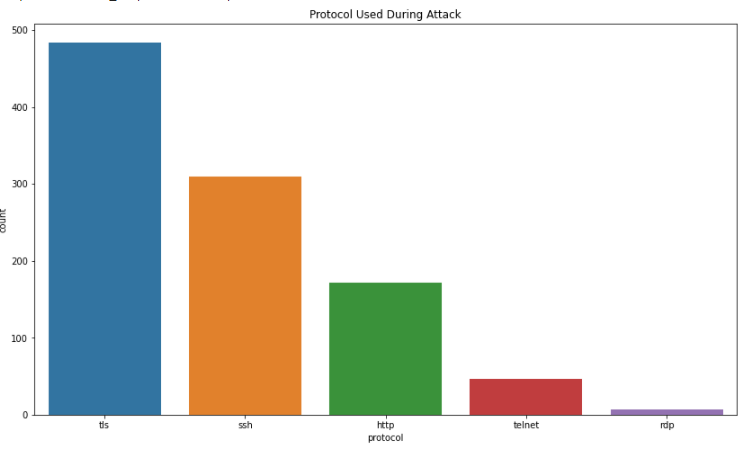
Proto Used During Attack



After successfully plotting graph “Proto Used During Attack”

We are able to get all the protocols used for attack, but from the graph we can clearly identify that TCP protocol have been used from most of the Countries then UDP and then ICMP.

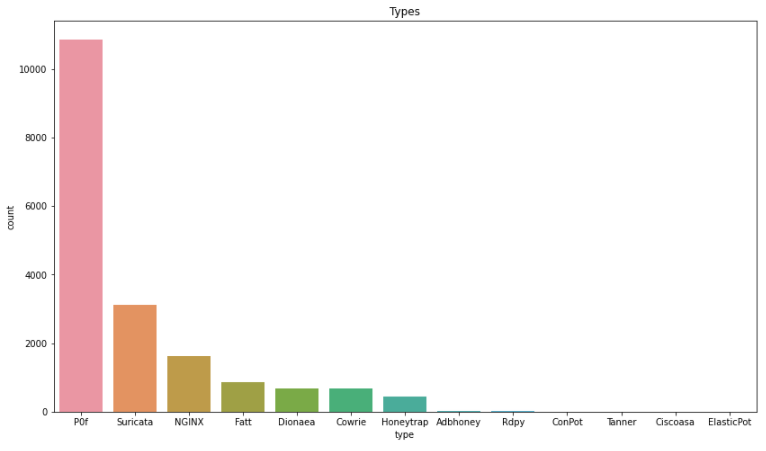
Protocol Used During Attack



After successfully plotting graph “Protocol Used During Attack”

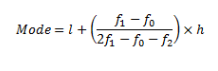
We are able to get all the protocols used for attack, but from the graph we can clearly identify that tls protocol have been used from most of the Countries then ssh then http protocol then telnet and then rdp.

HoneyPot Type



**Imputing Categorical missing values by using Mode**

Mode means a value or a number that appears most frequently in a dataset. Sometimes we may need to find the value, which is occurring more frequently in the dataset.



**Imputing Numerical missing values using random values.**

We can fill the Missing Values using Basics Mathods like Mean,Median,Mode and Std Dev

Here we have Many Missing Data available so it is not recomended to go with Measure of central Tendency and Std dev, If we consider this methods basically it will badily impact on Distribution.

When we have Huge number of Data are missing. In such case if you are going to consider Mean at that time your Distribution gets badly impacted and thaht's what we dont want because you dont want to change the disrtibution of the data because whatever distribution is given to you, your Normal Distribution of your data is most suitable to your ML Model

Whever we have Missing values , Just replace with some Random value in that particular column

**Handling IP Address**

The IP address information was one of the most important features in the databases; it is strongly related to network attacks and is collected as an IPv4 address. An IPv4 address is a 32-bit number that uniquely identifies a network interface on a machine, and it is displayed as four numbers and split by three dots. The IPv4 address is one of the core protocols of standards-based internetworking methods in the Internet and other packet-switched networks.

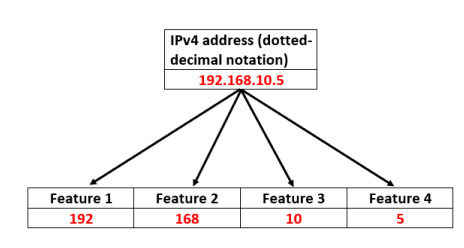
Even though it is displayed as numbers split by three dots, an IPv4 address cannot be easily operated by machine learning algorithms.

This research involved two kinds of IP addresses: source IPs and destination IPs. When a device initiates communication with servers or other devices, its IP address is called a source IP, which also sends IP packets. The device’s IP address that receives the packets is called a destination IP

we will extract source and destination IP addresses of packets from a huge data set of network packets and cluster the IPv4 addresses in to different groups based on their four octets, using unsupervised ML method of clustering.

**Splitting an IP Address to Four Number**s

Convert the IP addresses was to split the 32-bit address into four separate numbers. These four numbers were considered as four individual features. After converting the source IP and the destination IP, these IP address became eight different features in the database. As shows the process of splitting an IP address.



# Features extraction from IP addresses

For clustering of data, we need to choose particular features of the data based on which the clustering algorithm can divide the data in to clusters of similar members.

We will use the four octets of IP address as features for clustering

We have chosen to split the four octets of each IPv4 address and use them as a vector of length 4 to be fed to the clustering algorithm to represent each data (IP address ). The reasoning behind choosing this approach was taken from [this research paper.](https://hammer.purdue.edu/articles/thesis/Encoding_IP_Address_as_a_Feature_for_Network_Intrusion_Detection/11307287)

**Applying Anomaly Detection in Source IP Address and Destination IP Address for predicting Malicious IP address and Non-** **Malicious IP address.**

Anomaly Detection it is kind of identifying the or suspicious data that are present in the original dataset. This is a simple technique in the unsupervised Machine Learning.

Detecting the Anomaly is very different for different dataset. If you have the less noise that is present in the dataset then you can find the Anomaly, but if you have high noise it is a very hard process for detecting the anomaly.

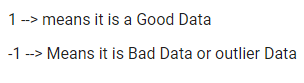
**Applying Isolation Forest Algorithm in Source IP Address and Destination IP Address.**

Isolation Forest is used to create Tree like Structure based on the data points we have.It build the tree like structure based on the closest relationship between the root value and child value. This module will form the tree according to their relation.

In the Isolation Forest you can decide how many tree values we need to form. Forming the tree is the itterative process in this library.

From Sklearn libraries we import the Isolation forest module.

After prediction we successfully classify Malicious and Non-Malicious Source IP Address.

This is the prediction Output w.r.t Source IP Address we get.

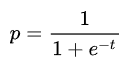
After successfully finding Anomaly from the Dataset specifically w.r.t Source IP Address then applying Machine Learning Classification Algorithm.

Logistic Regression

Logistic Regression or Logit Model is used to model the Probability of a certain class or event. Conceptually, the algorithm analyses the association between multiple independent variables and a categorical dependent variable, which the probability of occurrence of an event is estimated by fitting data to a logistic curve. Logistic regression consists of two models includes binary logistic regression and multinomial logistic. Binary logistic regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No.

To implement classification Logistic Regression in the present study, the LogisticRegression

function was imported from the Scikit-Learn library.

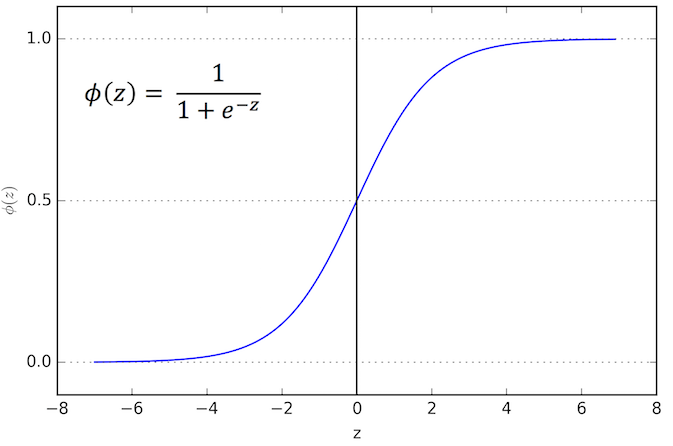


Were,

t=b0+b1x

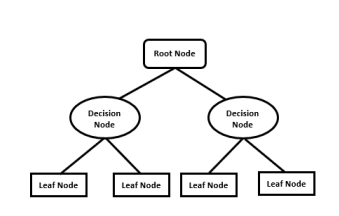
b0 is the intercept

b1 is the coefficient



Decision Tree

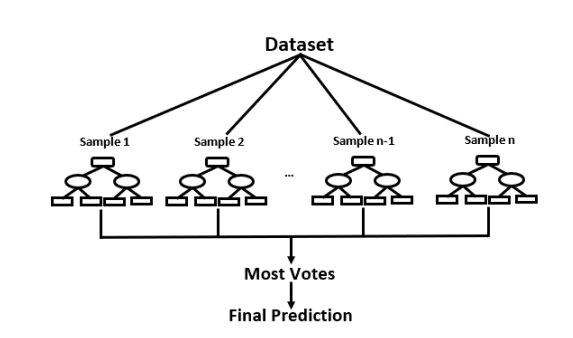
A decision tree, also known as classification tree, is a tree-based algorithm, and it is considered one of the most popular supervised learning methods. The concept of the decision tree model is very simple, and it has three main components: the model of decisions, a tree-like graph, and possible consequences using if-then and yes-no logic. Figure represents a decision tree model. The tree structure brakes the data set into different subsets. The root node is the first decision node and represents the entire sample. The decision node partitions the data, and it is split sub-node by sub-node. Eventually, predictions of the problem are given by leaf nodes, and the leaf nodes cannot be further split.



The default classifier model of the decision tree was used in the present thesis. The tree package was imported from Scikit-Learn, the fit method from the DecisionTreeClassifier class was called to train the data, and the predict method was used to make predictions.

Random Forest

Random Forest is operated by building a large number of unique decision trees, and it is an ensemble learning method of decision tree algorithms. There are three main types of ensemble learning: bagging, boosting and stacking. The random forest training process normally applies bagging to the decision tree. Bagging is another name for bootstrap aggregating and is often used to reduce variances. Random forests mainly consist of four steps. First, the algorithms randomly select samples from a data set. Second, a decision tree is created for each sample, and a prediction is generated from each decision tree. After that, voting occurs for each prediction result. The prediction of the random forest is the class with most votes based on the class prediction, which is split from each decision tree created in the random forest. Figure gives an example of how random forest is used for a prediction.



To implement classification random forest in the present study, the RandomForestClassifier function was imported from the Scikit-Learn library, and most parameters of random forest were the same as the decision tree. The fit method from RandomForestClassifier class was called to train the data, and the predict method was used to make predictions.

# Detection Source and Destination IP Address.By applying PCA,GMM and K-Means Clustering

# Applying Principal Component Analysis(PCA) ****to reduce the dimensions for cluster visualization****

PCA (Principle Component Analysis) it is unsupervised Machine Learning Algorithm. It is technique which is used for reducing the dimensionality. Any number of Independent feature and if you are trying to actually reduce these number of feature, to the number that you want you can actually use this PCA

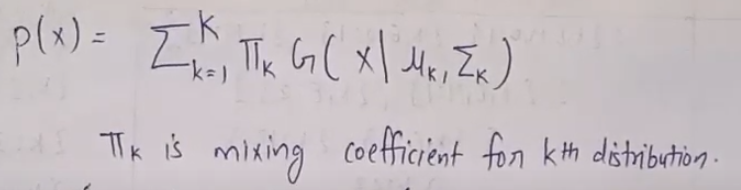
This technique is usually involves creating lot of vector space or number of independent components or the number of dimensions that you want, we can convert any dimension into that particular dimension

As the number of dimension increases there will be always a curse. Always your Accuracy get impacted with the number of dimension. Your Dimension increases your accuracy get impacted.

## **Use Gaussian Mixture Modeling (GMM) Clustering algorithm to group similar IP addresses.**

Gaussian Mixture Model (GMM) clustering suits ellipsoidal shaped clusters. GMM also is a probablistic clustering algorithm and provides easier way to detect anomalies.

Ii is a probabilistic model for representing normally distributed subpopulation with an overall population. Mixture model in general don’t require knowing which subpopulation a data point belongs to, allowing the model to learn subpopulation automatically. Since subpopulation assignment is not known, this constitutes a form of unsupervised learning.



Implemented GMM to cluster the source IP addresses and Destination IP Address Separately. And finding the anomalies.

Find Anomalies/Outliers from the clusters of IP addresses We use score\_samples() method of GMM to estimate the density at the location of each data point in the clusters. The greater the score, higher the density of the cluster at the location of that data point. Any instance located in a low-density region can be considered an anomaly. We can define a density threshold of say 4%, and consider all the data lying in areas of density below 4th percentile of the range of densities values, as anomalies. This threshold is arbitrary and we can choose it according to our discretion

**OPTIONAL**

# Applying K Means Clustering ****Algorithm****

# ****Clustering**: Clustering is the method of identify similar instances and keeping them together. In other words, Clustering identifies homogeneous subgroups among the observation.**

# ****K- Means Clustering**: K- Means Clustering approach in which the data is grouped into k distinct non-overlapping clusters based on their distances from the k cluster. The value of k needs to be specified first and then the algorithm assigns the points to exactly one cluster. In K-Means Clustering we used to build a random k number of cluster and then one by one, one by one so we can keep on like adding the dataset and then we keep on changing the centroid.**

# **So, we used to plot WCSS (Within cluster Summation of Square) Vs k, we used to plot it and then this graph is going to tell you that, wherever we are able to see a smooth dispersion of this graph, so that value of k we are supposed to select. Assuming k=5, this is showing the smooth dispersion there is no use decline into this one WCSS we will try to select that value of K.**

# **We choose the value of K based on WCSS and elbow method.An elbow plot shows at what value of K, the distance between the mean of a cluster and other data points in the cluster.**

# 

# 

**Best inovation in the field of AI for last 2 years**