ABSTRACT

In today's world hackers use different types of attacks for getting the valuable information. Network Intrusion Detection System (NIDS) is a software application that monitors the network or system activities for malicious activities and unauthorized access to devices. The objective of designing NIDS is to protect the data confidential. Our project mainly focuses on these issues with the help of Data Mining. This project includes the implementation of different data mining algorithms including Logistic regression, Gaussian NB and Decision tree to generate the rules for classify network activities. A comparative analysis of these techniques to detect intrusions has also been made.UNSW-NB15 dataset has been used for implementation.

LITERATURE SURVEY

As the attacks are difficult to prevent with firewalls, security policies, or other mechanisms due to rapid change of technology both system and application and due to unknown weakness or bugs, we need monitoring and evaluation periodically. Intrusion detection systems are designed to detect those attacks even with respect of security precautions. Some intrusion detection systems detect attacks in real time and can be used to stop an attack during the course. Others provide the detection techniques, after understanding the nature of attack and thereafter reduce the occurrence of attacks in the future. Many people are working for the enhancement of intrusion detection systems, which includes research organization, educational institutions, and software companies. Even the Department of Defense also works on these security issues. As they explore different techniques and develop various new systems for intrusion detection.

INTRODUCTION

Protecting the credential data from the intruders became the challenging task for organizations. Cyber security or information technology security are the techniques of protecting computers, networks, programs and data from unauthorized access or attacks that are aimed for exploitation. The IDS helps the network administrator to detect any malicious activity on the network and alerts the administrator to get the data secured by taking the appropriate actions against those attacks. Network security is one of its area. Network security includes activities to protect the usability, reliability, integrity and safety of the network. Effective network security targets a variety of threats and stops them from entering or spreading on the network. Network security components includes Anti-virus and anti-spyware, Firewall to block unauthorized access to your network, Intrusion prevention systems (IPS) to identify fast-spreading threats, such as zero-day or zero-hour attacks and Virtual Private Networks (VPN) to provide secure remote access. Intrusion detection is a relatively new addition to set of security technologies. IDS is an evolution which enhance the network security and safeguarding the data of the organization. An intrusion refers to any unauthorized access or malicious utilization of information resources. An intruder tries to gain unauthorized access to information that will cause harm in other malicious activities. The Intrusion detection system is about the firewall security. The firewall protects an organization from the malicious attacks from the Internet and the IDS detects if someone tries to access in through the firewall or manages to break in the firewall security and tries to have an access on any system in the organization and alerts the system administrator if there is an undesired activity in the firewall. Network Based IDS (NIDS) present in a computer or device connected to a segment of an organization's network and monitors network traffic on that network segment looking for ongoing attacks. When any disturbances occur then the network based IDS is planned to know an attack and it responds by sending notifications to administrators. NIDS looks for attack patterns within network traffic.

MODEL ARCHITECTURE

Logistic Regression is a regression model.The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as one. Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem.The decision for the value of the threshold value is majorly affected by the values of precision and recall.

1. Low Precision/High Recall.- we want to reduce the number of false negatives without necessarily reducing the number false positives, we choose a decision value which has a low value of Precision or high value of Recall.

2. High Precision/Low Recall.- we want to reduce the number of false positives without necessarily reducing the number false negatives, we choose a decision value which has a high value of Precision or low value of Recall.

Gaussian Naive-Bayes continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution. A Gaussian distribution is also called Normal distribution. When plotted, it gives a bell shaped curve which is symmetric about the mean of the feature values.

Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems.It uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

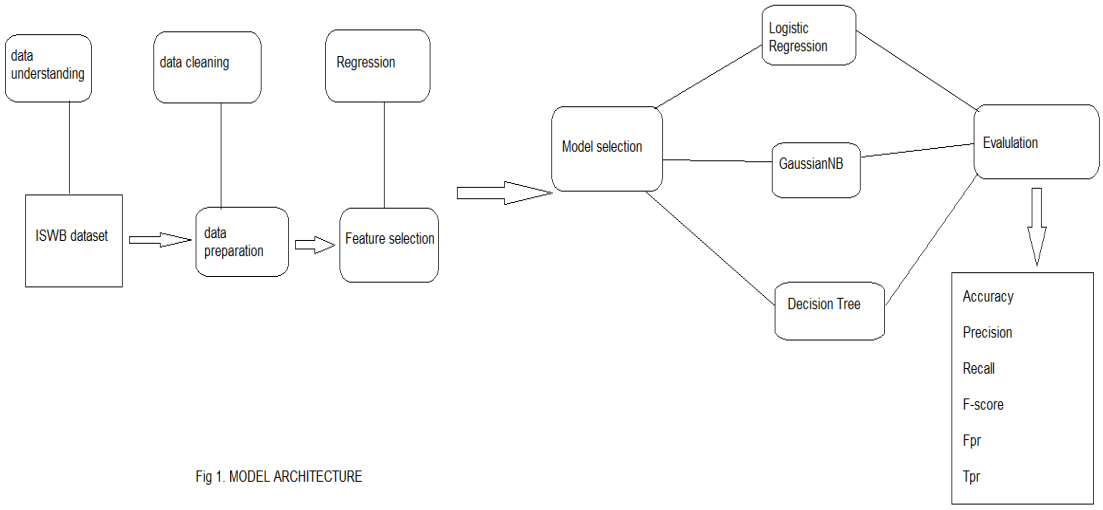

Fig 1. MODEL ARCHITECTURE

METHODOLOGY

LOGISTIC REGRESSION :

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes) or 0 (no). t logistic regression model predicts P(Y=1) as a function of X. Following are the steps for logistic regression algorithm on a dataset.

ALGORITHM :

Step 1: Collecting data.

Step 2: Analyze the data.

Step 3: Data wrangling.

Step 4: Train and Test data.

Step 5: Accuracy checking.

Initially collect the input data from data set and after that perform the data analizations.After completing the two steps perform data cleaning in order to remove unwanted attributes from the data set.Training and Testing is the important step that is to be performed after data cleaning .Finally check the accuracy of a model on a particular dataset by applying confusion matrix and record those values .

GAUSSIAN NB (NAIVE-BAYES)

GaussianNB implements the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian.A Gaussian Naive Bayes algorithm is a special type of NB algorithm. It's specifically used when the features have continuous values. It's also assumed that all the features are following a gaussian distribution i.e; normal distribution.

$$P(X\ i \mid y\ ) = (1/\sqrt{2}\pi\sigma y).\exp\left(-(xi-u)\right)^2/2\sigma\text{^}2$$

The parameters μy and σy are estimated using maximum likelihood.

ALGORITHM:

Step 1: Data import.

Step 2: Data preprocessing.

Step 3: Import GaussianNB module.

Step 4: Find Accuracy score.

First step is to import the dataset after that perform data preprocessing where all missing values are to be handled .Import the sklearn GaussianNB module in order to calculate the model accuracy on a particular dataset. Accuracy can be increased by performing cross-validation.

DECISION TREE:

Decision tree is one of the predictive modelling approaches used in Statistics,Machine learning and data-mining. Decision tree is used for classification and prediction. It is a flowchart like tree structure where each internal node denotes a test on an attribute and also each branch represents an outcome of the test and each leaf node holds a class label. Decision trees can handle high dimensional data.It gives the best accuracy when compared with other models.

ALGORITHM:

Step 1: Get list of rows (dataset) which are taken into consideration for making decision tree .

Step 2: Calculate uncertainty of dataset .

Step 3: Partition rows into True rows and False rows .

Step 4: Divide the node on best question. Repeat again from step 1 again until we get leaf nodes.

Step 5: Find the Accuracy score of the model.

Classification is a two-step process, learning step and prediction step in machine learning. In the learning step, the model is developed based on given training data. In the prediction step, the model is used to predict the response for given data. Decision Tree is one of the easiest and popular classification algorithms to understand and interpret. For predicting a class label for a record start from the root of the tree. compare the values of the root attribute with the record's attribute. On the basis of comparison , Follow the branch corresponding to that value and jump to the next node. Finally find the accuracy of model using coefficient matrix.

EXPERIMENTATION AND RESULTS :

After applying the above algorithms on the dataset the following values are obtained.

ALGORITHM	ACCURACY	PRECISION	RECALL	F SCORE	FPR	TPR
Logistic Regression	0.85	0.99	0.82	0.90	0.82	0.96
Gaussian NB	0.93	0.99	0.92	0.95	0.92	0.96
Decision Tree	0.93	0.99	0.91	0.95	0.91	0.99

CONCLUSION :

The main objective of the NIDS is to develop data mining algorithms for detecting attacks and threats against computer systems .For applying the models such as logistic regression, Gaussian NB and Decision Tree on a dataset initially some predefined libraries must be imported. Inorder to execute the source code which is written in python language in jupyter notebook. From the above results decision tree gives the more accurate value when compared with remaining models. The objective of designing NIDS is to protect the data confidential. This project mainly focuses on these issues with the help of Data Mining.

REFERENCES :

[1]. Kapil Wankhade, Sadia Patka and Ravindra Thools, “An Efficient Approach for Intrusion Detection Using Data Mining Methods”, IEEE 2013.

[2]. S. A. Joshi and Varsha S. Pimprale, “Network Intrusion Detection System (NIDS) based on data mining,” International Journal of Engineering Science and Innovative Technology (IJESIT), Vol. 2, No. 1, pp. 95-98, 2013.

[3]. Deepthy K Denatious & Anita John, “Survey on Data Mining Techniques to Enhance Intrusion Detection”, International Conference on Computer Communication and Informatics (ICCCI - 2012), Jan. 10 – 12, 2012, Coimbatore, INDIA.

[4]. Mrutyunjaya Panda and Manas Ranjan Patra, “A Comparative Study Of Data Mining Algorithms For Network Intrusion Detection”, First International Conference on Emerging Trends in Engineering and Technology, pp 504-507, IEEE, 2008.