PREDICTION OF NETWORK ATTACKS WITH BEST ACCURACY

A PROJECT REPORT

Submitted by

DEEPAN.D (111516205011)

GIRIRAM.S (111516205014)

IYERPRADIP.S (111516205017)

JAYANDHAN.S (111516205019)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

In

INFORMATION TECHNOLOGY

R.M.D ENGINEERING COLLEGE, THIRUVALLUR

ANNA UNIVERSITY :: CHENNAI 600 025

APRIL 2020

BONAFIDE CERTIFICATE

Certified that this project report titled *"Prediction of Network Attacks with Best Accuracy"*, is a *bonafide* work of **DEEPAN.D(111516205014), GIRIRAM.S(111516205014), IYERPRADIP.S(111516205017)** and **JAYANDHAN.S(111516205019)** who carried out the work under my supervision, for the partial fulfillment of the requirements for the award of the degree of *Bachelor of Technology* in *Information Technology*. Certified further that to the best of my knowledge and belief, the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or an award was conferred on an earlier occasion.

SIGNATURE
Dr. Balasubadra K M.E, PhD

HEAD OF THE DEPARTMENT
Dept of Information Technology
RMD Engineering College
R.S.M Nagar
Kavaraipettai – 601 206.

SIGNATURE
Dr. Joe Prathap P M

SUPERVISOR
Assistant Professor
RMD Engineering College
R.S.M Nagar
Kavaraipettai – 601 206.

CERTIFICATION OF EVALUATION

COLLEGE : RMD ENGINEERING COLLEGE

DEPARTMENT : INFORMATION TECHNOLOGY

SEMESTER : 08

Title of the Project	**Name of the Students with Register number**	**Name of the Supervisor with Designation**
Prediction of Network Attacks with Best Accuracy	Deepan D (111516205011) Giriram S (111516205014) IyerPradip S (111516205017) Jayandhan S (111516205019)	Dr. Joe Prathap Assistant Professor RMD Engineering College

The report of the project work submitted by the above students in partial fulfillment for the award of Bachelor of Technology Degree in INFORMATION TECHNOLOGY of Anna University was evaluated and confirmed to be the report of the work done by the above students and then evaluated.

Submitted the project during the viva voce held on ...……

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

A project of this magnitude and nature requires the kind co-operation and support from many, for successful completion. We wish to express our sincere thanks to all those who were involved in the completion of this project.

It is our immense pleasure to express our sense of gratitude to our chairman **Thiru R.S.MUNIRATHINAM,** our vice chairman **Thiru R.M. KISHORE** and our director **Thiru R.JOTHI NAIDU,** for the facilities and support given by them in the college.

We are extremely thankful to our principal **Dr.K.Sivaram, Ph.D.,** for having given us an opportunity to serve the purpose of any education.

We are indebted to **Dr.K.Balasubadra, Ph.D.,** Head of the Department of Information Technology, for her valuable guidance and useful suggestions during the course of the project.

We are thankful to our project supervisor **Dr. Joe Prathap P M** in the Department of Information Technology, R.M.D Engineering College for his/her helpful guidance and valuable support given to us throughout the project.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

To explore all the possibilities of attacks, as it is expensive to create data sets for the Intrusion Detection System (IDS) and software to detect network intrusions and protect a computer network from unauthorized users, including perhaps insiders.

The intrusion detector learning task is to build a predictive model (i.e. a classifier) capable of distinguishing between "bad" connections, called intrusions or attacks, and "good" normal connections. Generally, the network sectors have to predict whether the connection is attacked or not by using the KDDCup99 data set using machine learning techniques.

The aim is to investigate machine learning based techniques for better packet connection transfers and predict the results to obtain best accuracy.

To propose a machine learning-based method to accurately predict the DOS, R2L, UU2R, Probe and overall attacks by using supervised classification machine learning algorithms.

To compare and discuss the performance of various machine learning algorithms for the given data set with an evaluation classification report, identify the confusion matrix and to categorize data in the priority order.

The result shows the effectiveness of the proposed machine learning algorithm technique by obtaining best accuracy and precision through the F1 score calculated by Precision and Recall.

CHAPTER - I : INTRODUCTION

1.1 GENERAL

Machine learning is to predict the future from past data.Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. Process of training and prediction involves use of specialized algorithms. It feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated in to three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and the corresponding labeling to learn data has to be labeled by a human being beforehand. Unsupervised learning is no labels. It provided to the learning algorithm. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance.

1.2 SUPERVISED LEARNING

Supervised learning program occurs when a system is given input and output variables with the intentions of learning how they are mapped together, or related. The goal is to produce an accurate enough mapping function that when new input is given, the algorithm can predict the output.Each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal).

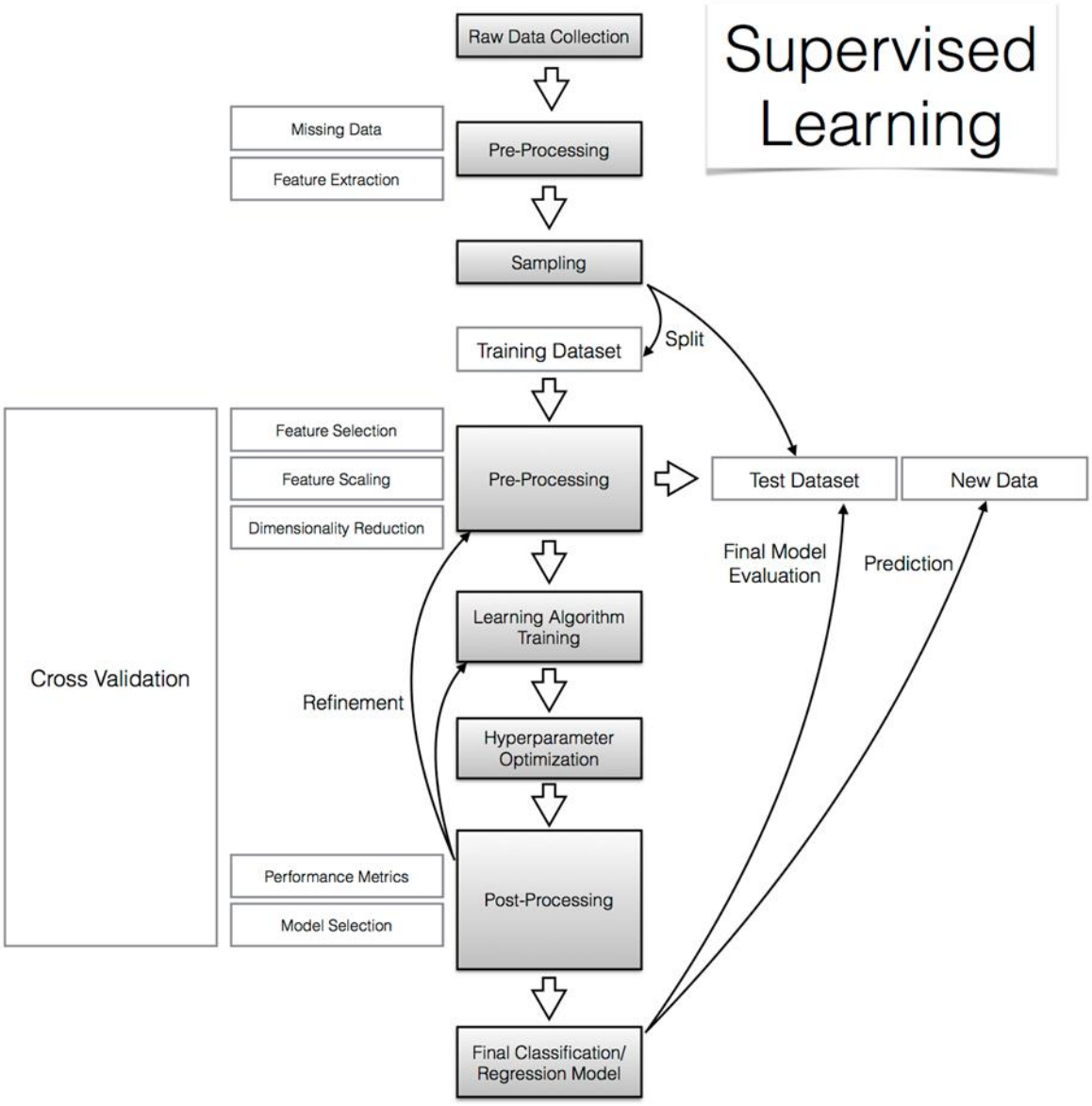


Fig. Work Flow Diagram

1.3 PYTHON

Python is an interpreted high-level programming language for general purpose programming. In python, open cv is to be installed. ‘Open source computer vision library initiated by some enthusiast coders to incorporate Image Processing into a wide variety of coding languages. It has C++, C and Python interfaces running on Windows, Linux, Android, and Mac.

1.4 PROBLEM STATEMENT

To compare and discuss the performance of various machine learning algorithms for the given data set with an evaluation classification report, identify the confusion matrix, to categorize data in the priority order and for better packet connection transfers and predict the results to obtain best accuracy.

1.5 OBJECTIVE

To explore all the possibilities of attacks, as it is expensive to create data sets for the Intrusion Detection System (IDS). To detect network intrusions and protect the computer network from any unauthorized users. The aim is to investigate machine learning based techniques for better packet connection transfers and predict the results to obtain best accuracy. To compare and discuss the performance of various machine learning algorithms from the given data set. To show the effectiveness of the proposed machine learning algorithm technique by obtaining best accuracy and precision through F1 score.

1.6 EXISTING SYSTEM

1.6.1 EXISTING SYSTEM

To focuses on the conception of a monitoring system able to detect and classify jamming and protocol-based attacks and achieve this goal, we proposed to outsource the attack detection function from the network to protect over the time. The frequencies of interest belong to the communication channel between 2.402 and 2.422 GHz. On these frequencies, the proposed estimation model shows good results in the prediction of attacks. The development of connected devices and their daily use is presently at origin of the omnipresence of Wi-Fi wireless networks. In both cases, defense strategies have to be developed to prevent the misuses of the networks. The first objective of this study is to propose a monitoring solution, which is independent of the communication networks, to detect the occurrence of attacks. The second objective is to develop a method that is able to classify attacks of different types: the intentional electromagnetic interference, i.e., jamming attacks and the protocol-based attacks.After that, we build a classification protocol following two steps: the first consists in the construction of a support vector machine (SVM) classification model using the collected spectra, and the second step uses this SVM model to predict the class of the attack.

1.6.2 DISADVANTAGES

1) It can’t discuss to know how our model can evolve in the case where unknown attack occurs with all types of attacks by popular machine learning algorithms.
2) It can’t describe each categorized of DOS attacks like back, Neptune etc. based on the network connections.
3) Algorithm prediction results by best accuracy of classification algorithms with classification report of precision, recall and f1-score.

1.7 PROPOSED SYSTEM

1.7.1 PROPOSED SYSTEM

Exploratory Data Analysis :

Supervised machine learning classification algorithms will be used to obtain data sets and extract patterns, helping to avoid the attacks and make better decisions in the future and to obtain results with maximum accuracy.

Data Wrangling :

Load the data, check for cleanliness and then trim and clean the given data set for analysis.

Data Collection :

The data set collected for predicting the Network attacks is split into Training set and Test set. Training and Test set are split in 7:3 ratio. Decision tree algorithms, K-Nearest Neighbor (KNN) and Support vector classifier (SVC) are applied on the Training set and based on the test result accuracy.

Pre-Processing :

To gain better results data need to be pre processed so as to improve the efficiency of the algorithm. Data integration, data reduction and data transformation are also to be applicable for network connections data set.

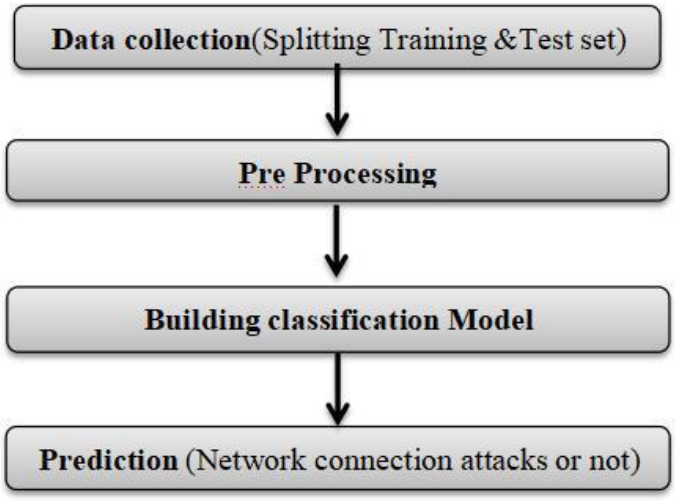


Fig: Proposed Model

1.7.2 ADVANTAGES

1) Improve the efficiency of the algorithm.
2) By comparing and analyze the attacks with the logic it is easy to obtain the best accuracy in each attack.

1.8 APPLICATIONS

1) Extract structural features to learn a mapping function.
2) Predicting values from known targets.
3) Pre-labelled data trains a model to predict new outcomes.
4) Risk evaluation and prevent the access of third party.

1.9 RELATED WORK

In recent year researchers have carried out many researches to develop anomaly-based IDS system using machine learning approach. In contrast to misuse-based intrusion detection that checks for signatures contained in the network packet header, anomaly-based IDS extracts the network data packets to obtain network features/attributes that can be used to detect attacks using machine learning approach. Several machine learning algorithms have been proposed by many researchers. To develop a good machine learning method, some techniques and mechanism need to be implemented in a combination, such as the use of feature and model selection as well as parameter tuning to get the optimal result.

CHAPTER - 2: LITERATURE REVIEW

Title : A Prediction Model of DOS Attack's Distribution Discrete Probability

Author : Wentao Zhao, Jianping Yin and Jun Long

Year : 2008

The process of prediction analysis is a process of using some method or technology to explore or stimulate undiscovered or complicated intermediate processes based on previous and present states and then speculated the results.

In an early warning system, accurate prediction of DOS attacks is the prime aim in the network offence and defense task. Detection based on abnormality is effective to detect DOS attacks. A various studies focused on DOS attacks from different respects.

However, these methods required a priori knowledge being a necessity and were difficult to discriminate between normal burst traffics and flux of DOS attacks.

Based on data from flux inspecting and intrusion detection, it proposed a prediction model of DOS attack's distribution discrete probability based on clustering method of genetic algorithm and Bayesian method and the clustering problem first, and then utilizes the genetic algorithm to implement the optimization of clustering methods.

Based on the optimized clustering on the sample data, we get various categories of the relation between traffics and attack amounts, and then builds up several prediction sub-models about DOS attack.

This paper begins with the relation exists between network traffic data and the amount of DOS attack, and then proposes a clustering method based on the genetic optimization algorithm to implement the classification of DOS attack data. This method first gets the proper partition of the relation between the network traffic and the amount of DOS attack based on the optimized clustering and builds the prediction sub-models of DOS attack.

Title : Apriori Viterbi Model for Prior Detection of Socio-Technical Attacks in a Social Network

Author : Preetish Ranjan, Abhishek Vaish

Year : 2014

Socio-technical attack is an organized approach which is defined by the interaction among people through maltreatment of technology with some of the malicious intent to attack the social structure based on trust and faith. Awful advertisement over internet and mobile phones may defame a person, organization, group and brand value in society which may be proved to be fatal.

People are always very sensitive towards their religion therefore mass spread of manipulated information against their religious belief may create pandemonium in the society and can be one of the reasons for social riots, political mis-balance etc.

Cyber-attack on water, electricity, finance, health care, food and transportation system are may create chaos in society within few minutes and may prove even more destructive than that of a bomb as it does not attack physically but it attacks on the faith and trust which is the basic pillar of our social structure.

It tries to find the hidden information in huge social network by compressing it in small networks through apriori algorithm and then diagnosed using viterbi algorithm to predict the most probable pattern of conversation to be followed in the network.

From the above simulation done on CDR, it may be concluded that if this kind of simulation applied on networks based on the internet and if we are in the position to get the data which could be transformed in transition and emission matrix then several kind of prediction may be drawn which will be helpful to take our decisions.

Title : New Attack Scenario Prediction Methodology

Author : Seraj Fayyad, Cristoph Meinel

Year : 2013

Intrusion detection systems (IDS) are used to detect the occurrence of malicious activities against IT system. Through monitoring and analyzing of IT system activities the malicious activities will be detected. In ideal case IDS generate alert(s) for each detected malicious activity and store it in IDS database.

Duplication relation means that the two alerts generated as a result of same malicious activity. Where same attack scenario relation means that the two related alert are generated as a result of related malicious activities.

Attack scenario or multi step attack is a set of related malicious activities run by same attacker to reach specific goal. Possible multi step attack against a network start with information gathering about network and the information gathering is done through network Reconnaissance and fingerprinting process.

Through fingerprint process Operating system type and version are identified. propose a real time prediction methodology for predicting most possible attack steps and attack scenarios.

Proposed methodology benefits from attacks history against network and from attack graph source data it comes without considerable computation overload such as checking of attack plans library.

Possible third attack step is to identify attack plan based on the modeled attack graph in the past step. The attack plan usually will include the exploiting of a sequence of founded vulnerabilities.

This sequence of nodes vulnerabilities is related through causal relation and connectivity. Lastly Attacker start orderly exploits the attack scenario sequences till reaching his/her goal. Attack plan consist of many correlated malicious activities end up with attacking goal.

Title : Cyber Attacks Prediction Model Based on Bayesian Network

Author : Jinyu W1, Lihua Yin and Yunchuan Guo

Year : 2012

The prediction results reflect the security situation of the target network in the future, and security administrators can take corresponding measures to enhance network security according to the results.

To quantitatively predict the possible attack of the network in the future, attack probability plays a significant role.It can be used to indicate the possibility of invasion by intruders. As an important kind of network security quantitative evaluation measure, attack probability and its computing methods has been studied for a long time. Many models have been proposed for performing evaluation of network security.

Graphical models such as attack graphs become the main-stream approach. Attack graphs which capture the relationships among vulnerabilities and exploits show us all the possible attack paths that an attacker can take to intrude all the targets in the network.

In our cyber-attacks prediction model, they used attack graph to capture the vulnerabilities in the network. In addition we consider 3 environment factors that are the major impact factors of the cyber-attacks in the future. They are the value of assets in the network, the usage condition of the network and the attack history of the network.

Cyber attacks prediction is an important part of risk management. Existing cyber attacks prediction methods did not fully consider the specific environment factors of the target network, which may make the results deviate from the true situation. In this paper, we propose a Cyber attacks prediction model based on Bayesian network. Then we capture the using environment factors using Bayesian network model. Cyber attacks predictions are performed on the constructed Bayesian network.

Title : Adversarial Examples: Attacks and Defenses for Deep Learning

Author : Xiaoyong Yuan , Pan He, Qile Zhu, and Xiaolin Li

Year : 2019

It reviewed the recent findings of adversarial examples in DNNs. We investigated the existing methods for generating adversarial examples.A taxonomy of adversarial examples was proposed.

We also explored the applications and countermeasures for adversarial examples. This paper attempted to cover the state-of-the-art studies for adversarial examples in the DL domain.

Compared with recent work on adversarial examples, we analyzed and discussed the current challenges and potential solutions in adversarial examples. However, deep neural networks (DNNs) have been recently found vulnerable to well-designed input samples called adversarial examples.

Adversarial perturbations are imperceptible to human but can easily fool DNNs in the testing/deploying stage. The vulnerability to adversarial examples becomes one of the major risks for applying DNNs in safety-critical environments.

Therefore, attacks and defenses on adversarial examples draw great attention. In this paper, we review recent findings on adversarial examples for DNNs, summarize the methods for generating adversarial examples, and propose taxonomy of these methods.

Under the taxonomy, applications for adversarial examples are investigated. We further elaborate on countermeasures for adversarial examples. In addition, three major challenges in adversarial examples and the potential solutions are discussed.

CHAPTER - 3: REQUIREMENT ANALYSIS

3.1 HARDWARE REQUIREMENTS

Processor **:** Pentium IV/III

Hard disk **:** minimum 80 GB

RAM **:** minimum 2 GB

3.2 SOFTWARE REQUIREMENTS

Operating system **:** Windows

Tool **:** Anaconda with Jupyter Notebook

3.3 FUNCTIONAL SPECIFICATIONS

The software requirements specification is a technical specification of requirements for the software product. It is the first step in the requirements analysis process. It lists requirements of a particular software system. The following details to follow the special libraries like sk-learn, pandas, numpy, mat-plot-lib and sea-born.

Process of functional steps,

Problem define

Preparing data

Evaluating algorithms

Improving results

Prediction the result

3.4 NON-FUNCTIONAL SPECIFICATIONS

Machine needs the enough hard disk space to install the software and run our project. The software must be installed correctly in a good working system. Assurance is the level of guarantee that the system application will behave as excepted.A risk is a possible event which could cause a loss.

3.5 SUPERVISED LEARNING TECHNOLOGY

3.5.1 ANACONDA

Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing that aims to simplify package management and deployment. The Anaconda distribution is used by over 12 million users and includes more than 1400 popular data-science packages suitable for Windows, Linux, and Mac OS. So, Anaconda distribution comes with more than 1,400 packages as well as the Conda package and virtual environment manager called Anaconda Navigator and it eliminates the need to learn to install each library independently. The open source packages can be individually installed from the Anaconda repository with the conda install command or using the pip install command that is installed with Anaconda. Pip packages provide many of the features of conda packages and in most cases they can work together. Custom packages can be made using the conda build command, and can be shared with others by uploading them to Anaconda Cloud, PyPI or other repositories. The default installation of Anaconda2 includes Python 2.7 and Anaconda3 includes Python 3.7. However, you can create new environments that include any version of Python packaged with conda.

3.5.2 ANACONDA NAVIGATOR

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows users to launch applications and manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository, install them in an environment, run the packages and update them. It is available for Windows, macOS and Linux.

The following applications are available by default in Navigator:

JupyterLab	Glueviz
Jupyter Notebook	Orange
QtConsole	Rstudio
Spyder	Visual Studio Code

3.5.3 THE JUPYTER NOTEBOOK

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

3.5.4 NOTEBOOK DOCUMENT

Notebook documents (or “notebooks”, all lower case) are documents produced by the Jupyter Notebook App, which contain both computer code (e.g. python) and rich text elements (paragraph, equations, figures, links, etc…). Notebook documents are both human-readable documents containing the analysis description and the results (figures, tables, etc.) as well as executable documents which can be run to perform data analysis.

CHAPTER - 4: SYSTEM DESIGN

Design is meaningful engineering representation of something that is to be built. Software design is a process design is the perfect way to accurately translate requirements in to a finished software product. Design creates a model, provides detail about software data structure, architecture, interfaces and components that are necessary to implement a system.

If the broader topic of product development "blends the perspective of change in settings, design, and integrating maps into a single approach to product development," then design is the act of taking the user profile settings values information and creating the design of the product to be developed.

Systems design is therefore the process of defining and developing systems to satisfy specified requirements of the user.

4.1 OVERALL ARCHITECTURE DIAGRAM

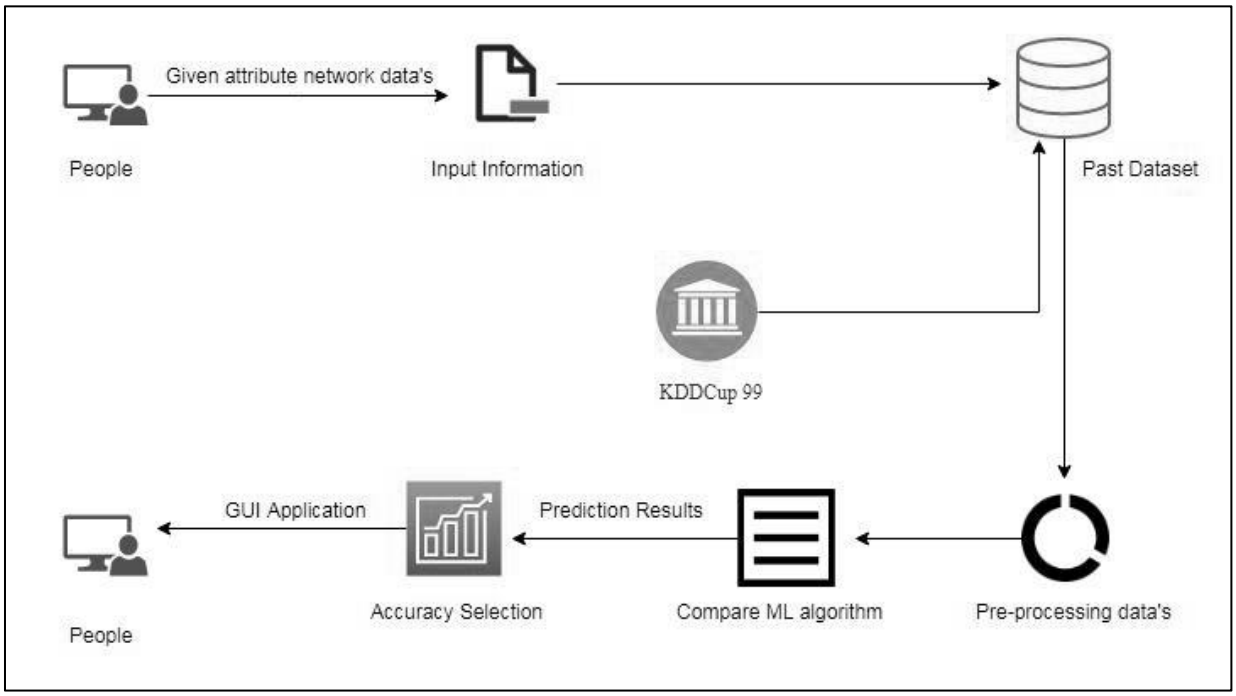


Fig. 4.1.1 Overall Architecture Diagram

4.2 DATA FLOW DIAGRAM

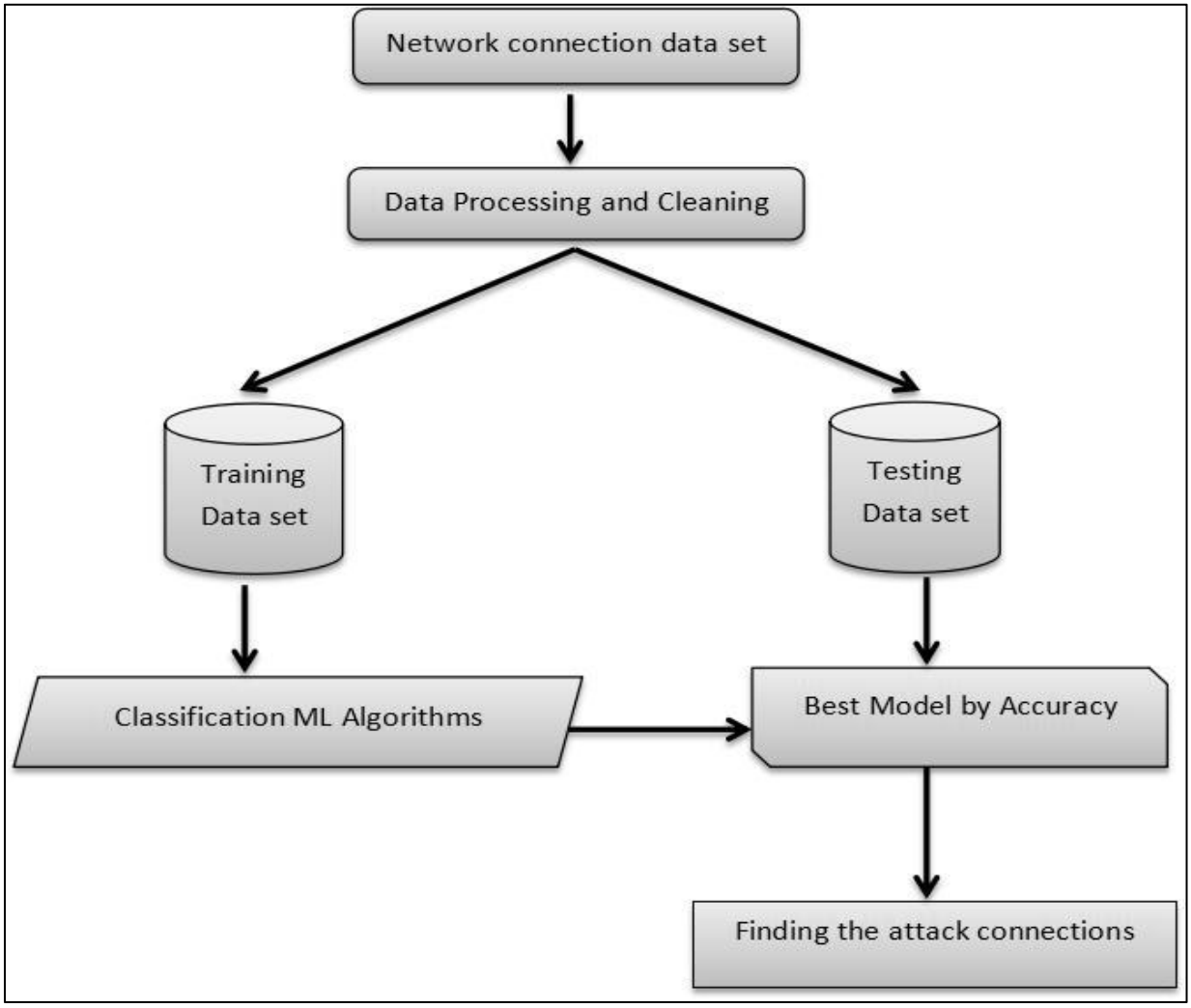


Fig. 4.2.1 Data Flow Diagram

Data Flow Diagrams are considered as high level requirement analysis of the system. Load the data, check for cleanliness and then trim and clean the given data set for analysis. The data set collected for predicting the network attacks is split into Training set and Test set. Training and Test set are split in 7:3 ratio. Decision tree algorithms, K-Nearest Neighbor (KNN) and Support vector classifier (SVC) are applied on the Training set and based on the test result accuracy. Data integration, data reduction and data transformation are also to be applicable for network connections data set.To gain better results data need to be pre-processed so as to improve the efficiency of the algorithm.

4.3 USE CASE DIAGRAM

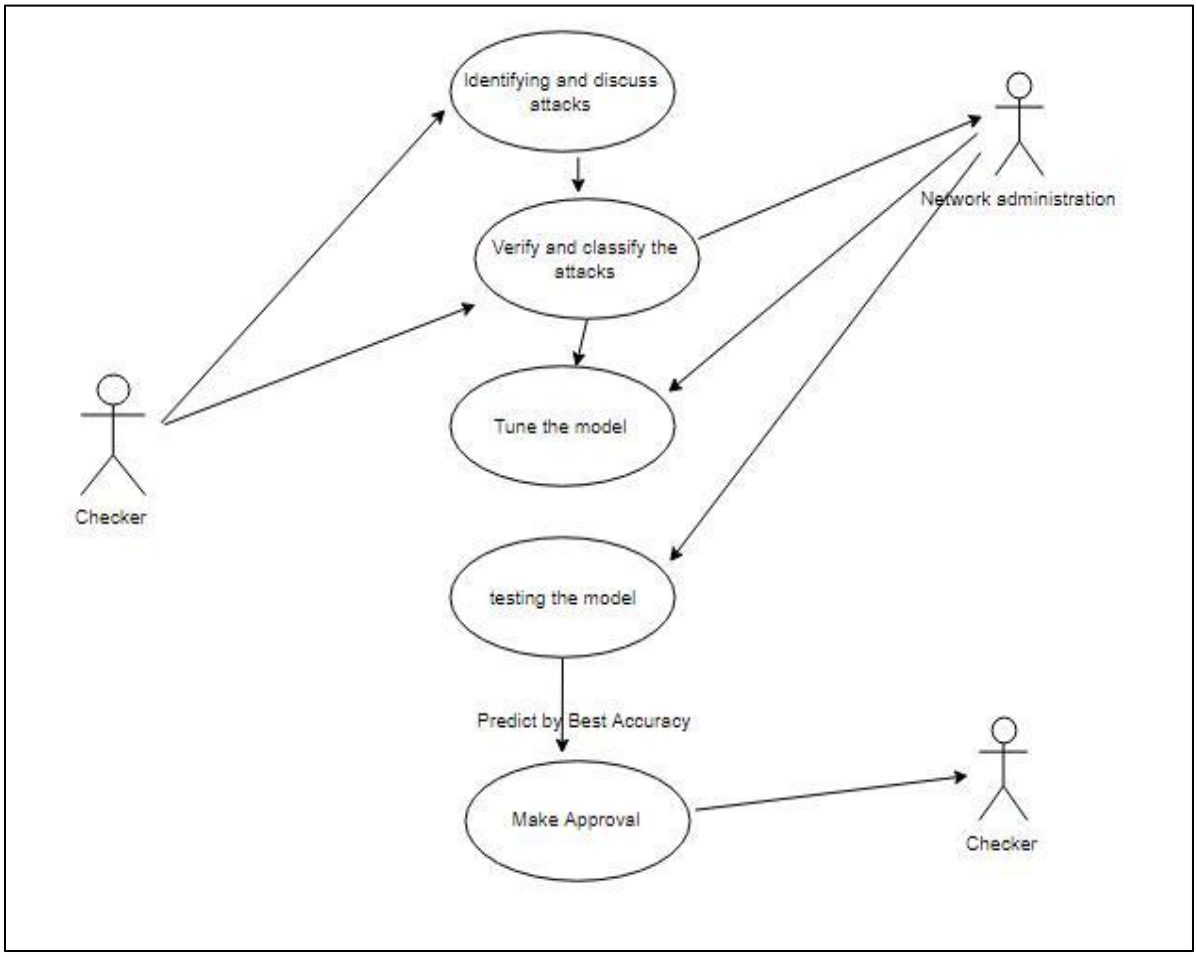


Fig. 4.3.1 Use Case Diagram

Use case diagrams are considered for high level requirement analysis of a system. So when the requirements of a system are analyzed the functionalities are captured in use cases. So, it can say that uses cases are nothing but the system functionalities written in an organized manner. Now the second things which are relevant to the use cases are the actors.

4.4 ENTITY RELATIONSHIP DIAGRAM

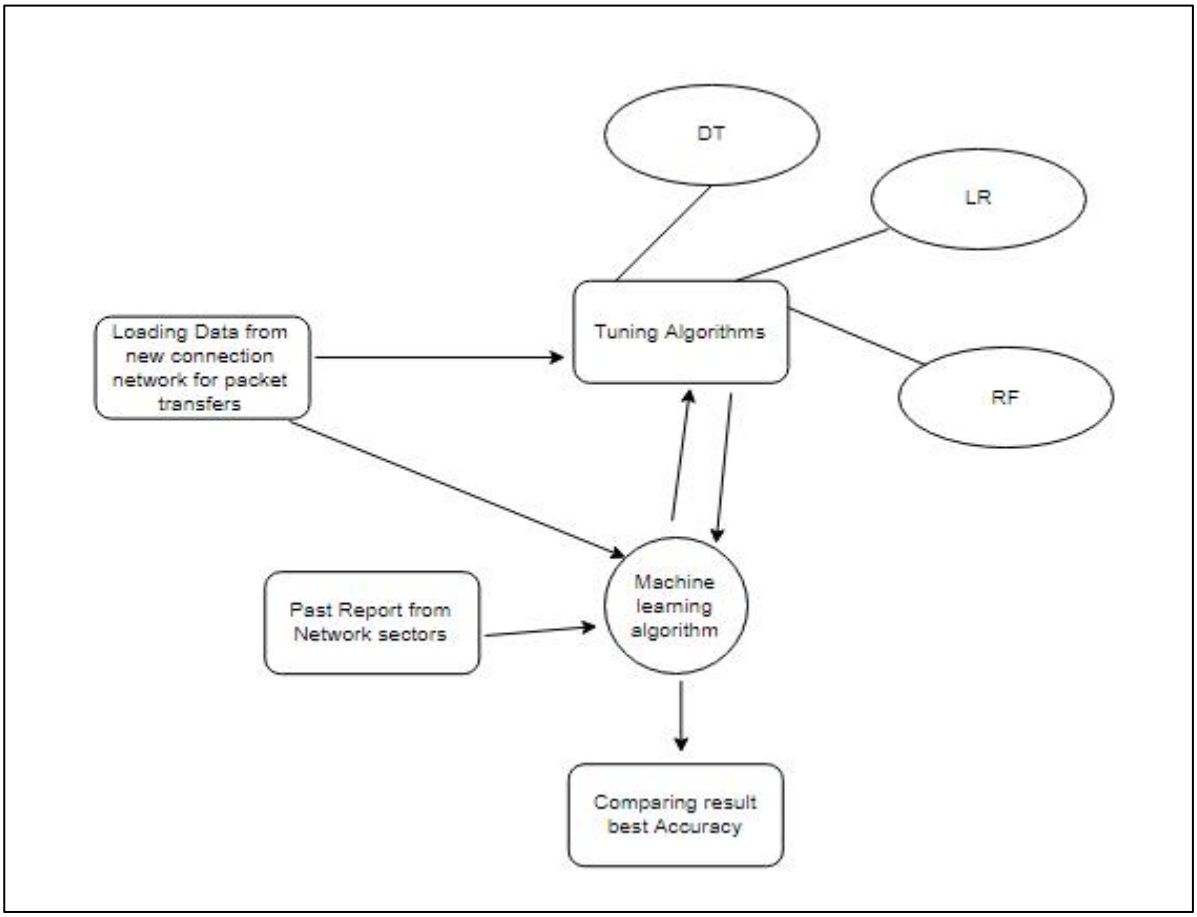


Fig. 4.4.1 Entity Relationship Diagram

An entity relationship diagram (ERD), also known as an entity relationship model, is a graphical representation of an information system that depicts the relationships among people, objects, places, concepts or events within that system. An ERD is a data modeling technique that can help define business processes and be used as the foundation for a relational database. Entity relationship diagrams provide a visual starting point for database design that can also be used to help determine information system requirements throughout an organization. After a relational database is rolled out, an ERD can still serve as a referral point, should any debugging or business process re-engineering be needed later.

4.5 SEQUENCE DIAGRAM

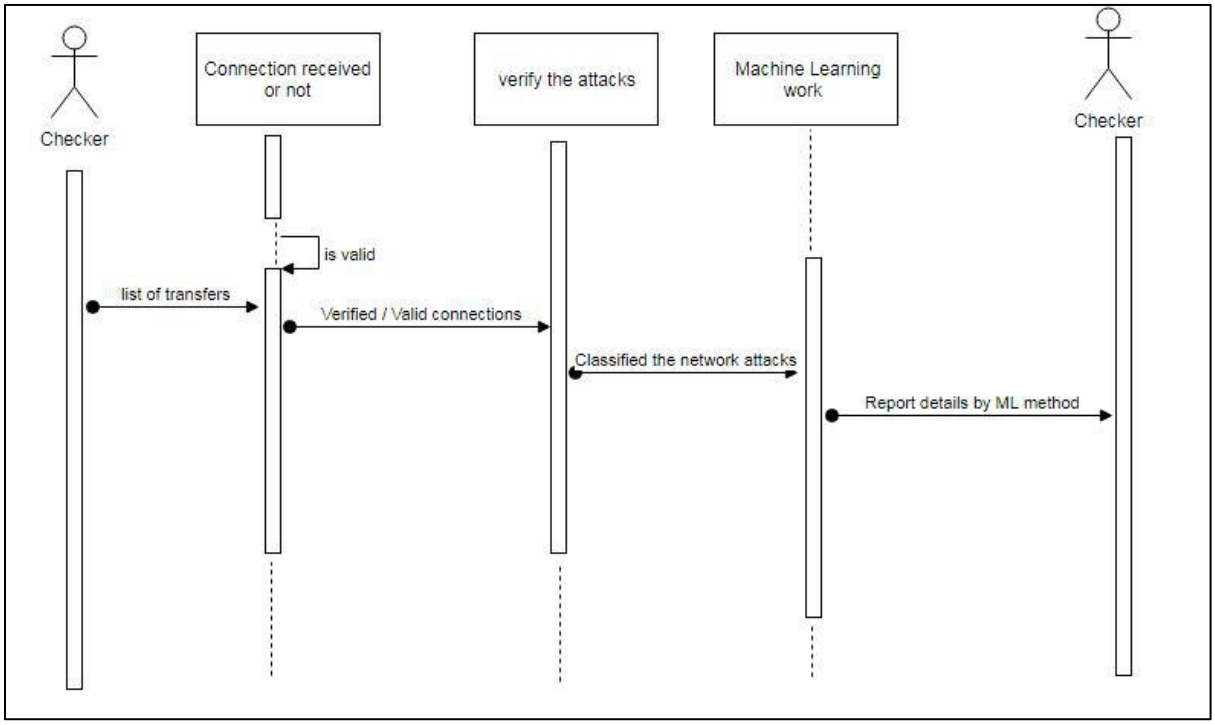


Fig. 4.5.1 Sequence Diagram

Sequence diagrams model the flow of logic within your system in a visual manner, enabling you both to document and validate your logic, and are commonly used for both analysis and design purposes. Sequence diagrams are the most popular UML artifact for dynamic modeling, which focuses on identifying the behavior within your system. Other dynamic modeling techniques include activity diagramming, communication diagramming, timing diagramming, and interaction overview diagramming. Sequence diagrams, along with class diagrams and physical data models are in my opinion the most important design-level models for modern business application development.

4.6 CLASS DIAGRAM

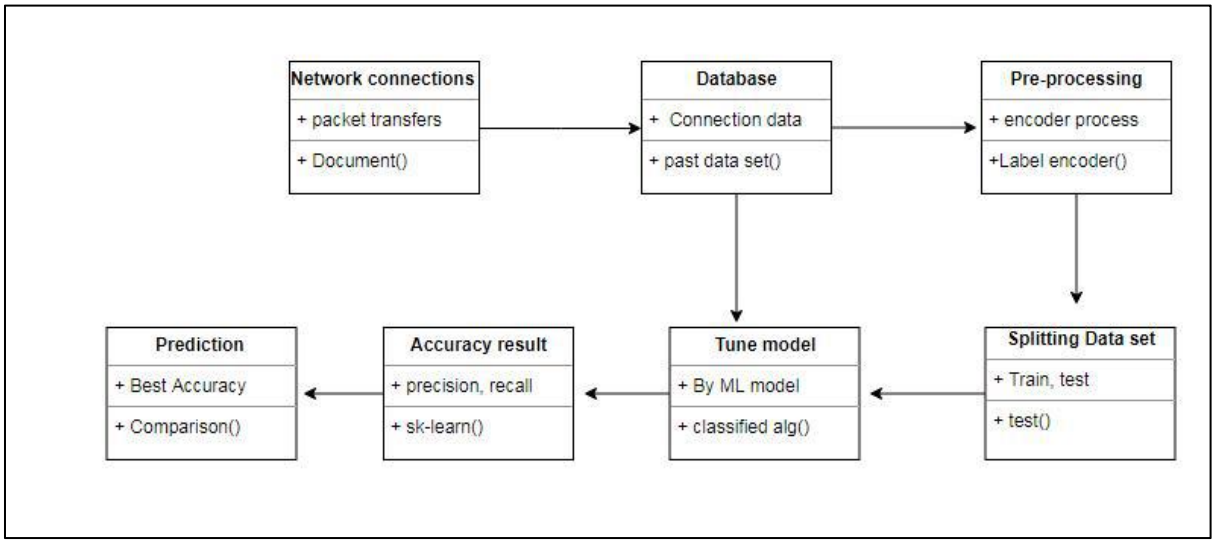


Fig. 4.6.1 Class Diagram

Class diagram is basically a graphical representation of the static view of the system and represents different aspects of the application. So a collection of class diagrams represent the whole system. The name of the class diagram should be meaningful to describe the aspect of the system. Each element and their relationships should be identified in advance Responsibility (attributes and methods) of each class should be clearly identified for each class minimum number of properties should be specified and because, unnecessary properties will make the diagram complicated. Use notes whenever required to describe some aspect of the diagram and at the end of the drawing it should be understandable to the developer/coder. Finally, before making the final version, the diagram should be drawn on plain paper and rework as many times as possible to make it correct.

4.7 ACTIVITY DIAGRAM

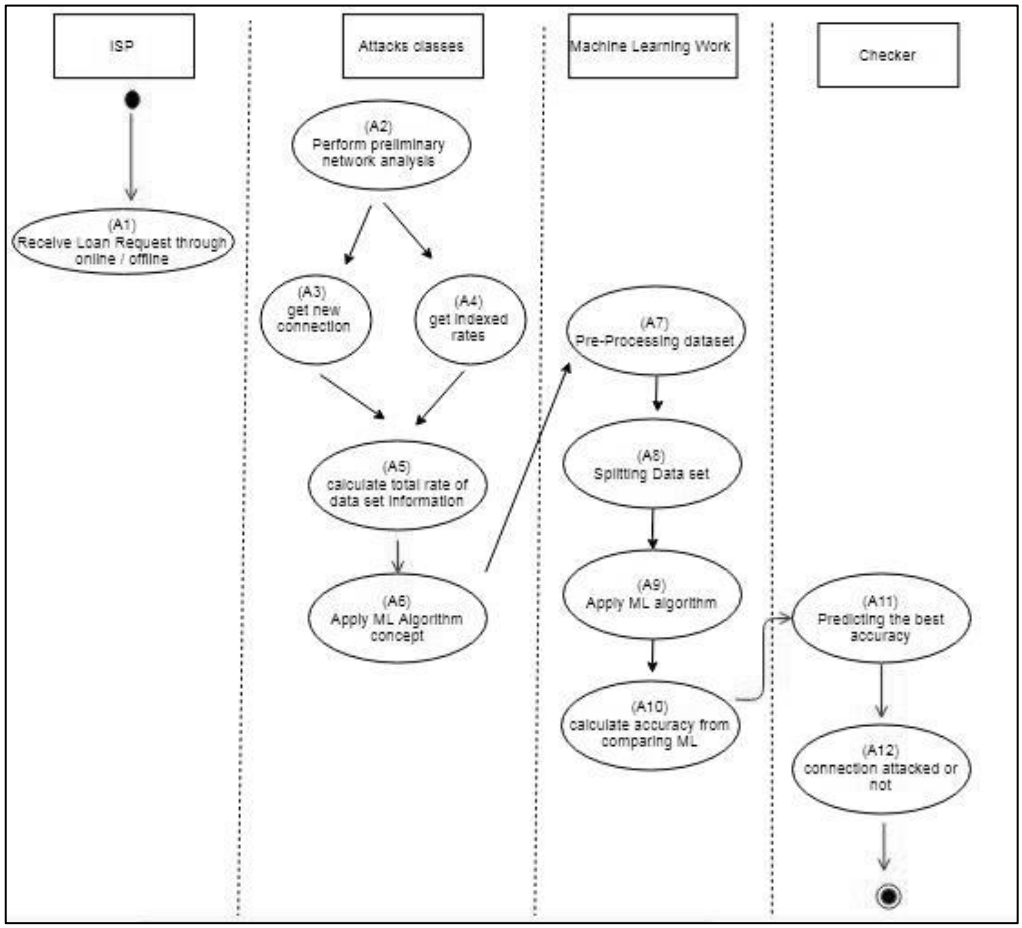


Fig. 4.7.1 Activity Diagram

Activity is a particular operation of the system. Activity diagrams are not only used for visualizing dynamic nature of a system but they are also used to construct the executable system by using forward and reverse engineering techniques. The only missing thing in activity diagram is the message part. It does not show any message flow from one activity to another. Activity diagram is some time considered as the flow chart. Although the diagrams looks like a flow chart but it is not. It shows different flow like parallel, branched, concurrent and single.

CHAPTER - 5: IMPLEMENTATION

5.1 GENERAL

When you are applying machine learning to your own data sets, you are working on a project.A machine learning project may not be linear, but it has a number of well-known steps:Define Problem, Prepare Data, Evaluate Algorithms, Improve Results, Present Results. The best way to really come to terms with a new platform or tool is to work through a machine learning project end-to-end and cover the key steps. Installing the Python anaconda platform, Loading the data set, Summarizing the data set, Visualizing the data set, Evaluating some algorithms and Making some predictions.

5.2 ALGORITHM EXPLANATION

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class or it may be multi-class too. In Supervised Learning, algorithms learn from labeled data. After understanding the data, the algorithm determines which label should be given to new data based on pattern and associating the patterns to the unlabeled new data.

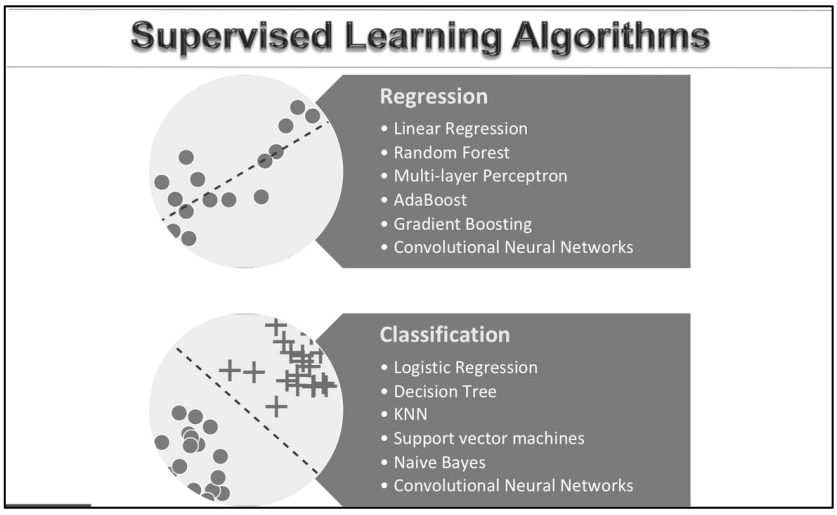


Fig. 5.2.a.Supervised Learning Algorithms

5.2.1 LOGISTIC REGRESSION(LR)

It is a statistical method for analyzing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest and a set of independent (predictor or explanatory) variables. Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable.

Dependent variable (binary variable**) = Response or Outcome Variable**

Dependent variable = 0 (no, failure)

Dependent variable = 1 (yes, success)

Assumptions:

- The logistic regression model predicts P(Y=1) as a function of X.
- Binary logistic regression requires the dependent variable to be binary.For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
- Only the meaningful variables should be included.The independent variables should be independent of each other.
- That is, the model should have little.
- The independent variables are linearly related to the log odds.
- Logistic regression requires quite large sample sizes.

5.2.2 RANDOM FOREST(RF)

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees habit of over fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision *trees*, resulting in a *forest of trees*, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

Basic steps involved in performing the random forest algorithm:

- Pick N random records from the data set.
- Build a decision tree based on these N records.
- Choose the number of trees you want in your algorithm and repeat steps 1 and 2.
- In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output).
- The final value can be calculated by taking the average of all the values predicted by all the trees in forest.
- Or, in case of a classification problem, each tree in the forest predicts the category to which the new record belongs.
- Finally, the new record is assigned to the category that wins the majority vote.

5.2.3 K-NEAREST NEIGHBORS(KNN)

K-Nearest Neighbor is a supervised machine learning algorithm which stores all instances correspond to training data points in n-dimensional space. When an unknown discrete data is received, it analyzes the closest k number of instances saved (nearest neighbors) and returns the most common class as the prediction and for real-valued data it returns the mean of k nearest neighbors. In the distance-weighted nearest neighbor algorithm, it weights the contribution of each of the k neighbors according to their distance using the following query giving greater weight to the closest neighbors. Usually KNN is robust to noisy data since it is averaging the k-nearest neighbors.

Basic steps involved in performing the k-nearest-neighbors algorithm:

- It takes a bunch of labeled points and uses them to learn how to label other points.
- To label a new point, it looks at the labeled points closest to that new point (those are its nearest neighbors), and has those neighbors vote, so whichever label the most of the neighbors have is the label for the new point (the “k” is the number of neighbors it checks).
- Makes predictions about the validation set using the entire training set.
- KNN makes a prediction about a new instance by searching through the entire set to find the k “closest” instances.
- “Closeness” is determined using a proximity measurement (Euclidean) across all features.

5.2.4 DECISION TREE(DT)

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. Decision tree builds classification or regression models in the form of a tree structure. It utilizes an if-then rule set which is mutually exclusive and exhaustive for classification. The rules are learned sequentially using the training data one at a time. Each time a rule is learned, the tuples covered by the rules are removed.

This process is continued on the training set until meeting a termination condition. It is constructed in a top-down recursive divide-and-conquer manner. All the attributes should be categorical. Attributes in the top of the tree have more impact towards in the classification and they are identified using the information gain concept. A decision tree can be easily over-fitted generating too many branches and may reflect anomalies due to noise.

Assumptions:

- At the beginning, we consider the whole training set as the root.
- Attributes are assumed to be categorical for information gain, attributes are assumed to be continuous.
- On the basis of attribute values records are distributed recursively.
- We use statistical methods for ordering attributes as root or internal node.

5.2.5 SUPPORT VECTOR CLASSIFIER(SVC)

A classifier that categorizes the data set by setting an optimal hyper plane between data. I chose this classifier as it is incredibly versatile in the number of different kernel functions that can be applied and this model can yield a high predictability rate. Support Vector Machines are perhaps one of the most popular and talked about machine learning algorithms.

5.2.6 NAIVE BAYES(NB)

Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large data sets. Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features.The Naive Bayes algorithm is an intuitive method that uses the probabilities of each attribute belonging to each class to make a prediction.Naive bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method.

Basic steps involved in performing the naive bayes algorithm:

- The probability of a class value given a value of an attribute is called the conditional probability.

- By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class. To make a prediction we can calculate probabilities of the instance belonging to each class and select the class value with the highest probability.

5.3 MODULE EXPLANATION

5.3.1 DATA VALIDATION

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, as close to the true error rate of the data set. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Pre-processing is a technique that is used to convert the raw data into a clean data set.

```
In [4]: data.head(10)
```

Out[4]:

	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	Wrong_fragment	Urgent	hot	...	dst_host_srv_count	dst_host_same_srv_rate	dst_host_ srv_
0	0	icmp	ecr_i	SF	1032	0	0	0	0	0	...	255	1.0	
1	0	icmp	ecr_i	SF	1032	0	0	0	0	0	...	255	1.0	
2	0	icmp	ecr_i	SF	1032	0	0	0	0	0	...	255	1.0	
3	0	icmp	ecr_i	SF	1032	0	0	0	0	0	...	255	1.0	
4	0	icmp	ecr_i	SF	1032	0	0	0	0	0	...	255	1.0	
5	0	icmp	ecr_i	SF	1032	0	0	0	0	0	...	255	1.0	
6	0	icmp	ecr_i	SF	1032	0	0	0	0	0	...	255	1.0	
7	0	icmp	ecr_i	SF	1032	0	0	0	0	0	...	255	1.0	
8	0	icmp	ecr_i	SF	1032	0	0	0	0	0	...	255	1.0	
9	0	icmp	ecr_i	SF	1032	0	0	0	0	0	...	255	1.0	

10 rows × 42 columns

Fig. 5.3.1.a Data set

Before Pre-Processing:

```
In [31]: df.head()
```

Out[31]:

	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	Wrong_fragment	Urgent	hot	...	dst_host_rerror_rate	dst_host_srv_rerror_rate	class
0	0	icmp	ecr_i	SF	1032	0	0	0	0	0	...	0.0	0.0	perl.
1	0	icmp	ecr_i	SF	1032	0	0	0	0	0	...	0.0	0.0	pod.
2	0	icmp	ecr_i	SF	1032	0	0	0	0	0	...	0.0	0.0	mscan.
3	0	icmp	ecr_i	SF	1032	0	0	0	0	0	...	0.0	0.0	xsnoop.
4	0	icmp	ecr_i	SF	1032	0	0	0	0	0	...	0.0	0.0	named.

Fig. 5.3.1.b Before Pre-Processing

After Pre-Processing:

```
In [32]: df.columns

Out[32]: Index(['duration', 'protocol_type', 'service', 'flag', 'src_bytes',
                'dst_bytes', 'land', 'Wrong_fragment', 'Urgent', 'hot',
                'num_failed_login', 'logged_in', 'num_compromised', 'root_shell',
                'su_attempted', 'num_root', 'num_file_creations', 'num_shells',
                'num_access_files', 'num_outbound_cmds', 'is_host_login',
                'is_guest_login', 'count', 'srv_count', 'serror_rate',
                'srv_serror_rate', 'rerror_rate', 'srv_rerror_rate', 'same_srv_rate',
                'diff_srv_rate', 'srv_diff_host_rate', 'dst_host_count',
                'dst_host_srv_count', 'dst_host_same_srv_rate',
                'dst_host_diff_ srv_rate', 'dst_host_same_src_port_rate',
                'dst_host_srv_diff_host _rate', 'dst_host_serror_rate',
                'dst_host_srv_serror_rate', 'dst_host_rerror_rate',
                'dst_host_srv_rerror_rate', 'class', 'DOSland', 'DOSlandclass', 'DOS',
                'R2L', 'U2R', 'Probe', 'attack'],
               dtype='object')
```

Fig.5.3.1.c After Pre-Processing

5.3.2 DOS ATTACK

A denial-of-service attack (DOS attack) is a cyber-attack in which the perpetrator seeks to make a machine or network resource unavailable to its intended users by temporarily or indefinitely disrupting services of a host connected to the Internet. Denial of service is typically accomplished by flooding the targeted machine or resource with superfluous requests in an attempt to overload systems and prevent some or all legitimate requests from being fulfilled. In a distributed denial-of-service attack (DDOS attack), the incoming traffic flooding the victim originates from many different sources. This effectively makes it impossible to stop the attack simply by blocking a single source.

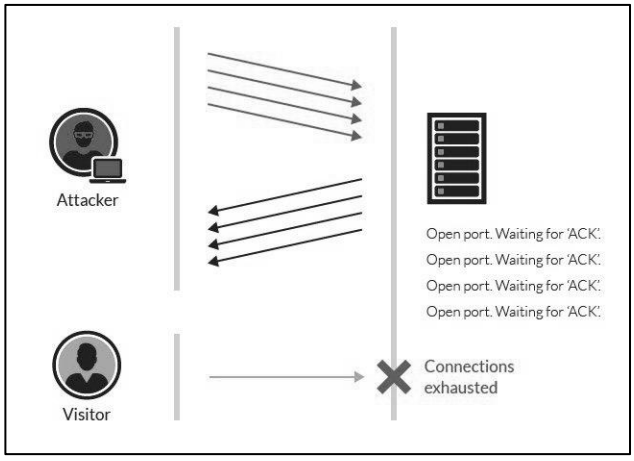


Fig. 5.3.2.a DOS Attack

5.3.3 R2L ATTACK

Malicious action to view or steal data illegally, introduce viruses or other malicious software to another computer or network or system, and cause damage to the targeted computer or network.A type of attack that is performed to access a particular network address remotely illegally.

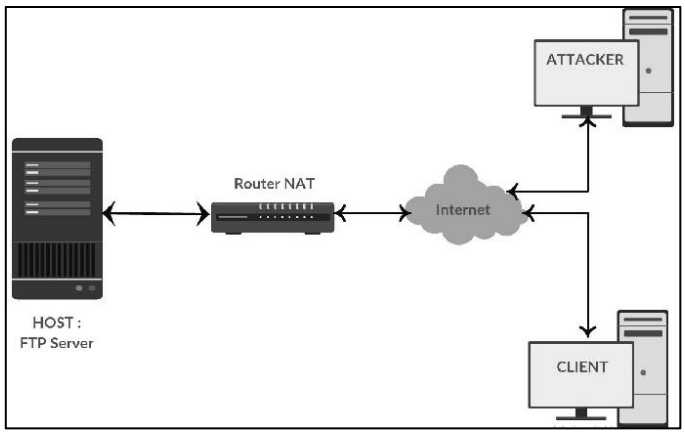


Fig.5.3.3.a R2L Attack

5.3.4 U2R ATTACK

An attacker or a hacker tries to get the access rights from a normal host in order to gain the root access to the system.These attacks are exploitations in which the hacker starts off on the system with a normal user account and attempts to abuse vulnerabilities in the system in order to gain super user privileges.

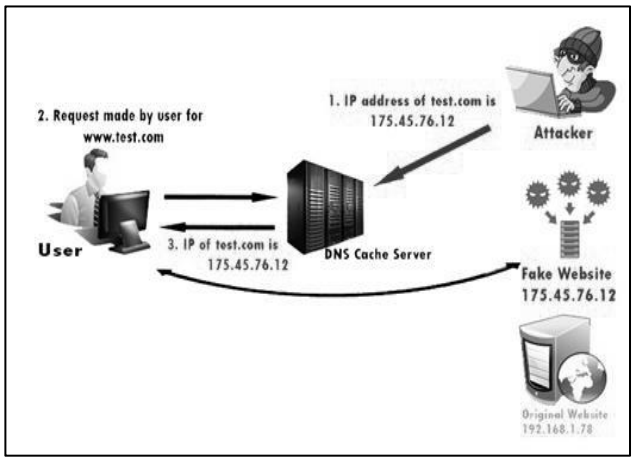


Fig.5.3.4.a U2R Attack

5.3.5 PROBE ATTACK

The hacker scans a machine or a networking device in order to determine weaknesses that may later be exploited so as to compromise the system.Probing is an attack in which the hacker scans a machine or a networking device in order

to determine weaknesses or vulnerabilities that may later be exploited so as to compromise the system.

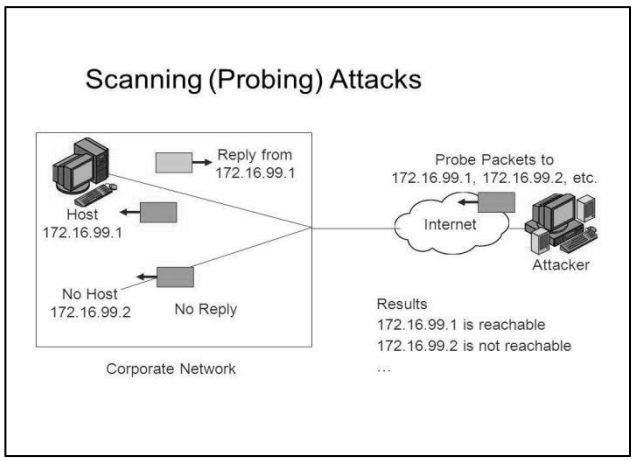


Fig.5.3.5.a Probe Attack

5.3.6 OVERALL NETWORK ATTACKS

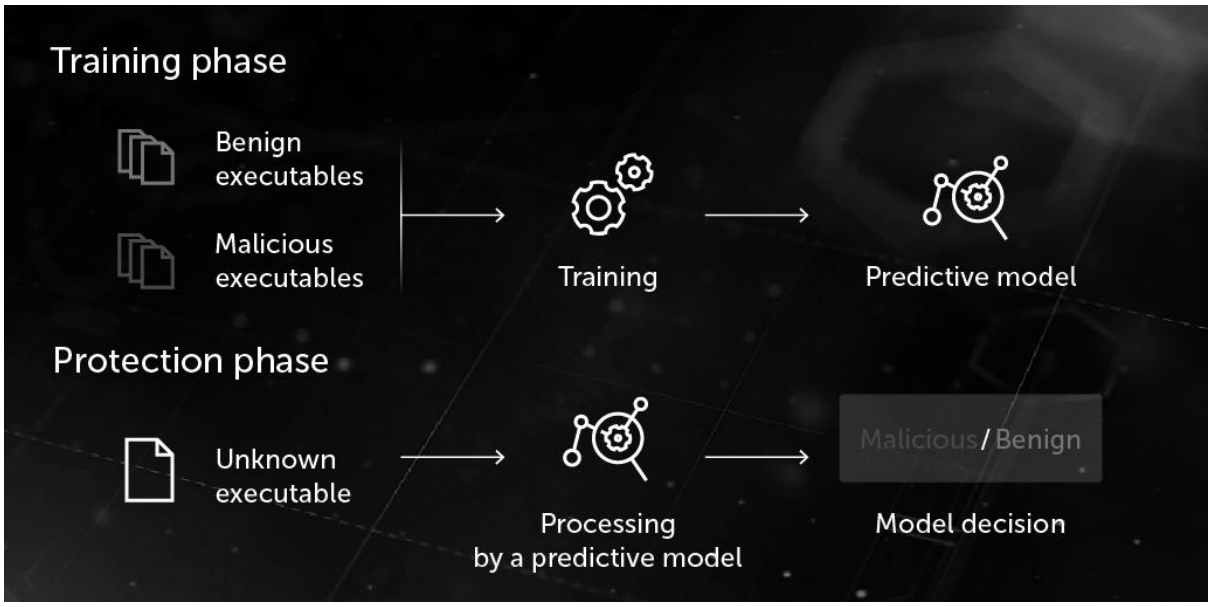


Fig.5.3.6.a Overall Attacks

5.3.7 GUI BASED PREDICTION

Tkinter is a python library for developing GUI(Graphical User Interfaces). We use the tkinter library for creating an application of UI (User Interface), to create windows and all other graphical user interface and Tkinter will come with Python as a standard package, it can be used for security purpose of each users or accountants.

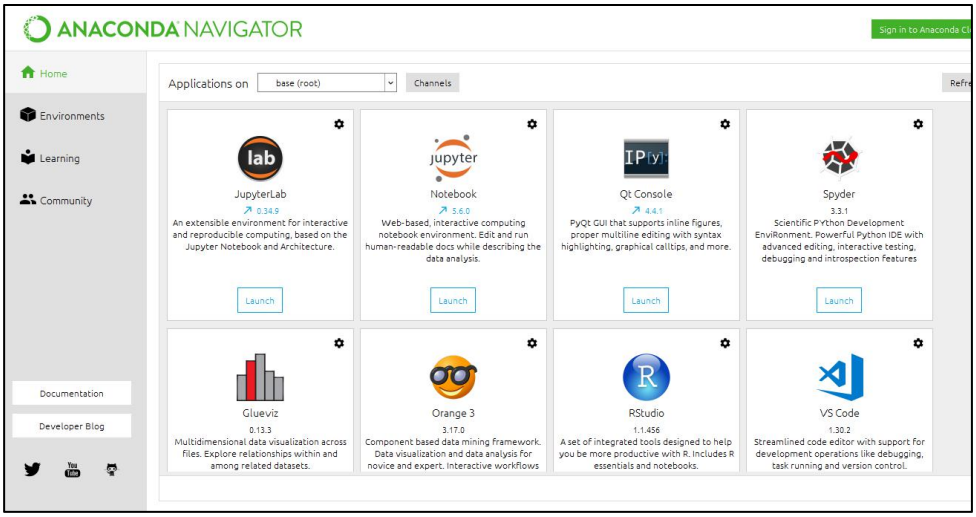


Fig.5.3.7.a Anaconda Navigator

Fig.5.3.7.b Jupyter Notebook

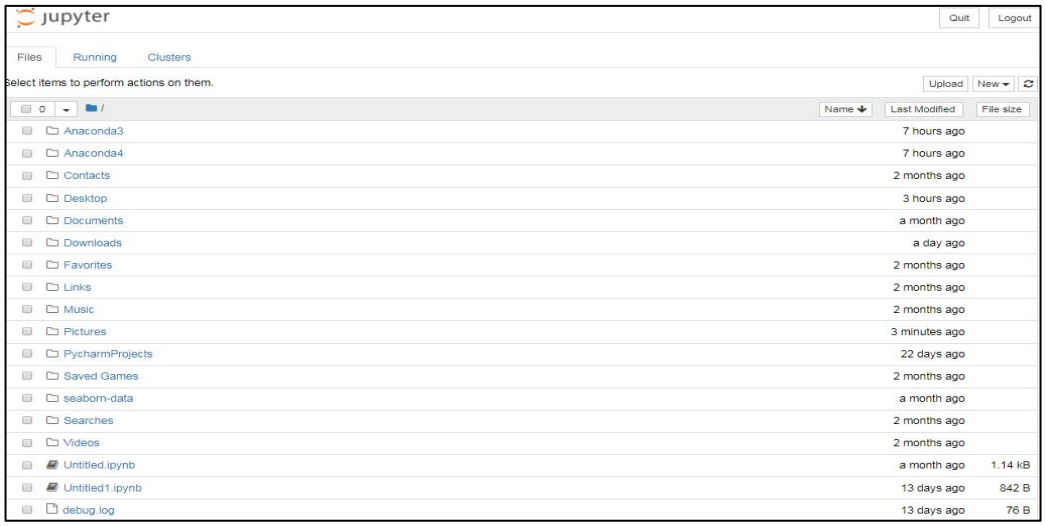


Fig.5.3.7.b Open the correspondent result folder

CHAPTER - 6: TESTING

Steps involved in accuracy calculation:

Training the data set.

Testing the data set.

Sensitivity:

Sensitivity = True Positive(TP)/(True Positive + False Negative(FN))

Specificity:

Specificity = True Negative(TN)/(True Negative + False Positive(FN))

True Positive Rate(TPR) = TP/(TP+FN)

False Positive Rate(FPR) = FP/(FP+TN)

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

F-Measure = 2TP/(2TP+FP+FN)

F1 Score = 2*(Recall*Precision)/(Recall+Precision)

True Positive(TP) : Number of persons who are unhealthy and are predicted as unhealthy.

False Negative(FN) : Number of persons who are unhealthy and got predicted as healthy.

True Negative(TN) : Number of persons who are healthy and are predicted as healthy.

False Positive(FP) : Number of persons who are healthy and got predicted as unhealthy.

Accuracy : The Proportion of the total number of predictions that is correct.

Precision : The proportion of positive predictions that are actually correct.

Recall : The proportion of positive observed values correctly predicted.

F1 Score : The weighted average of Precision and Recall.

6.1 PREDICTION OF DOS ATTACK

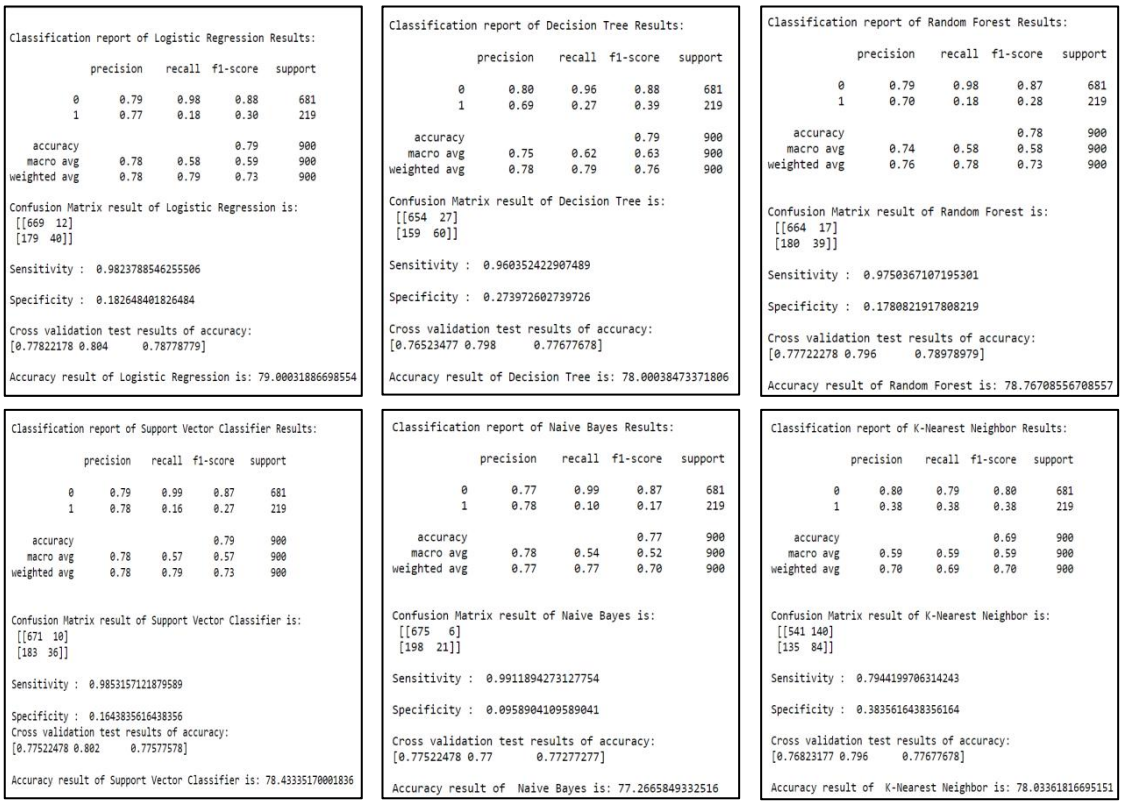
Main goal:

The primary intention of Denial-of-Service is to disable those functions or features used against financial institutions to distract IT and security personnel from security breaches.

Solution:

Denial of service is typically accomplished by flooding the targeted machine or resource with superfluous requests in an attempt to overload systems and prevent all legitimate requests from being fulfilled. Since the incoming traffic flooding the victim originates from different sources, it may be impossible to stop the attack simply by using ingress filtering.

Compare and analyse the Algorithms to predict the best accuracy from DOS attack:



Classification report of Logistic Regression Results:

	precision	recall	f1-score	support
0	0.79	0.98	0.88	681
1	0.77	0.18	0.30	219
accuracy			0.79	900
macro avg	0.78	0.58	0.59	900
weighted avg	0.78	0.79	0.73	900

Confusion Matrix result of Logistic Regression is:
[[669 12]
[179 40]]

Sensitivity : 0.9823788546255506

Specificity : 0.182648401826484

Cross validation test results of accuracy:
[0.77822178 0.804 0.78778779]

Accuracy result of Logistic Regression is: 79.00031886698554

Classification report of Decision Tree Results:

	precision	recall	f1-score	support
0	0.80	0.96	0.88	681
1	0.69	0.27	0.39	219
accuracy			0.79	900
macro avg	0.75	0.62	0.63	900
weighted avg	0.78	0.79	0.76	900

Confusion Matrix result of Decision Tree is:
[[654 27]
[159 60]]

Sensitivity : 0.960352422907489

Specificity : 0.273972602739726

Cross validation test results of accuracy:
[0.76523477 0.798 0.77677678]

Accuracy result of Decision Tree is: 78.00038473371806

Classification report of Random Forest Results:

	precision	recall	f1-score	support
0	0.79	0.98	0.87	681
1	0.70	0.18	0.28	219
accuracy			0.78	900
macro avg	0.74	0.58	0.58	900
weighted avg	0.76	0.78	0.73	900

Confusion Matrix result of Random Forest is:
[[664 17]
[180 39]]

Sensitivity : 0.9750367107195301

Specificity : 0.1780821917808219

Cross validation test results of accuracy:
[0.77722278 0.796 0.78978979]

Accuracy result of Random Forest is: 78.76708556708557

Classification report of Support Vector Classifier Results:

	precision	recall	f1-score	support
0	0.79	0.99	0.87	681
1	0.78	0.16	0.27	219
accuracy			0.79	900
macro avg	0.78	0.57	0.57	900
weighted avg	0.78	0.79	0.73	900

Confusion Matrix result of Support Vector Classifier is:
[[671 10]
[183 36]]

Sensitivity : 0.9853157121879589

Specificity : 0.1643835616438356
Cross validation test results of accuracy:
[0.77522478 0.802 0.77577578]

Accuracy result of Support Vector Classifier is: 78.43335170001836

Classification report of Naive Bayes Results:

	precision	recall	f1-score	support
0	0.77	0.99	0.87	681
1	0.78	0.10	0.17	219
accuracy			0.77	900
macro avg	0.78	0.54	0.52	900
weighted avg	0.77	0.77	0.70	900

Confusion Matrix result of Naive Bayes is:
[[675 6]
[198 21]]

Sensitivity : 0.9911894273127754

Specificity : 0.0958904109589041

Cross validation test results of accuracy:
[0.77522478 0.77 0.77277277]

Accuracy result of Naive Bayes is: 77.2665849332516

Classification report of K-Nearest Neighbor Results:

	precision	recall	f1-score	support
0	0.80	0.79	0.80	681
1	0.38	0.38	0.38	219
accuracy			0.69	900
macro avg	0.59	0.59	0.59	900
weighted avg	0.70	0.69	0.70	900

Confusion Matrix result of K-Nearest Neighbor is:
[[541 140]
[135 84]]

Sensitivity : 0.7944199706314243

Specificity : 0.3835616438356164

Cross validation test results of accuracy:
[0.76823177 0.796 0.77677678]

Accuracy result of K-Nearest Neighbor is: 78.03361816695151

Fig.Classification reports of each algorithm

DOS attack types:

Back, Neptune, Land, Pod, Smurf, Teardrop, Apache2, Mail bomb, Process table, UDP Storm.

Performance measurements of DOS attack prediction:

Parameters	LR	DT	RF	SVC	KNN	NB
Precision	0.79	0.80	0.79	0.79	0.80	0.77
Recall	0.98	0.96	0.98	0.99	0.79	0.99
F1-Score	0.88	0.88	0.87	0.87	0.80	0.87
Sensitivity	0.98	0.96	0.97	0.98	0.79	0.99
Specificity	0.18	0.27	0.17	0.16	0.38	0.09
Accuracy (%)	**79.00**	**78.00**	**78.76**	**78.43**	**78.03**	**77.26**

Graphical Representation of DOS attack prediction:

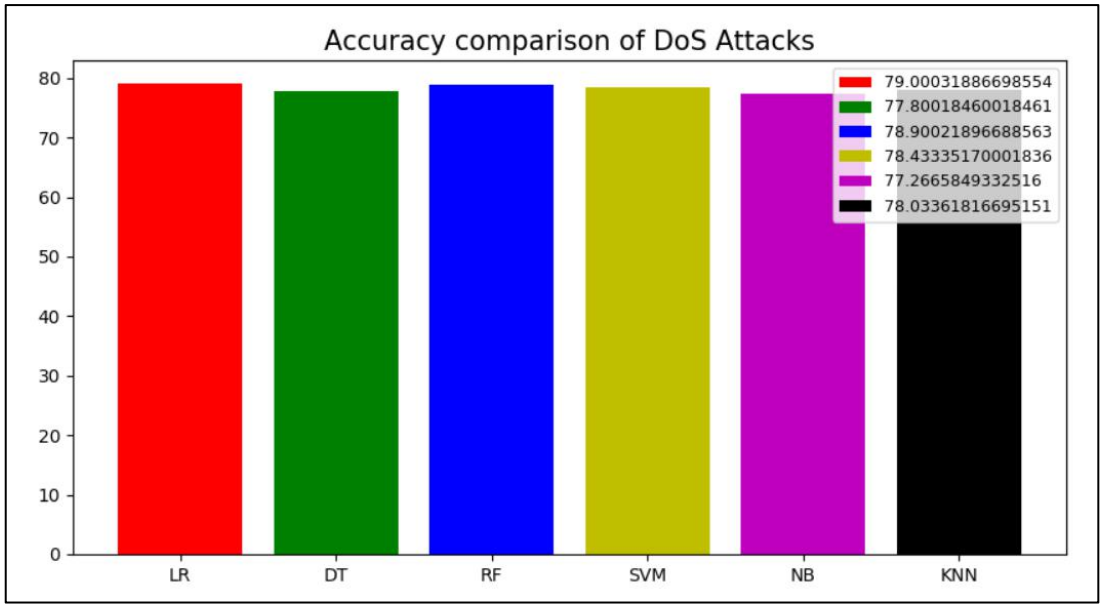


Fig.6.1.a DOS attack prediction

Result:

The highest accuracy for DOS attack is **Logistic Regression algorithm.**

6.2 PREDICTION OF R2L ATTACK

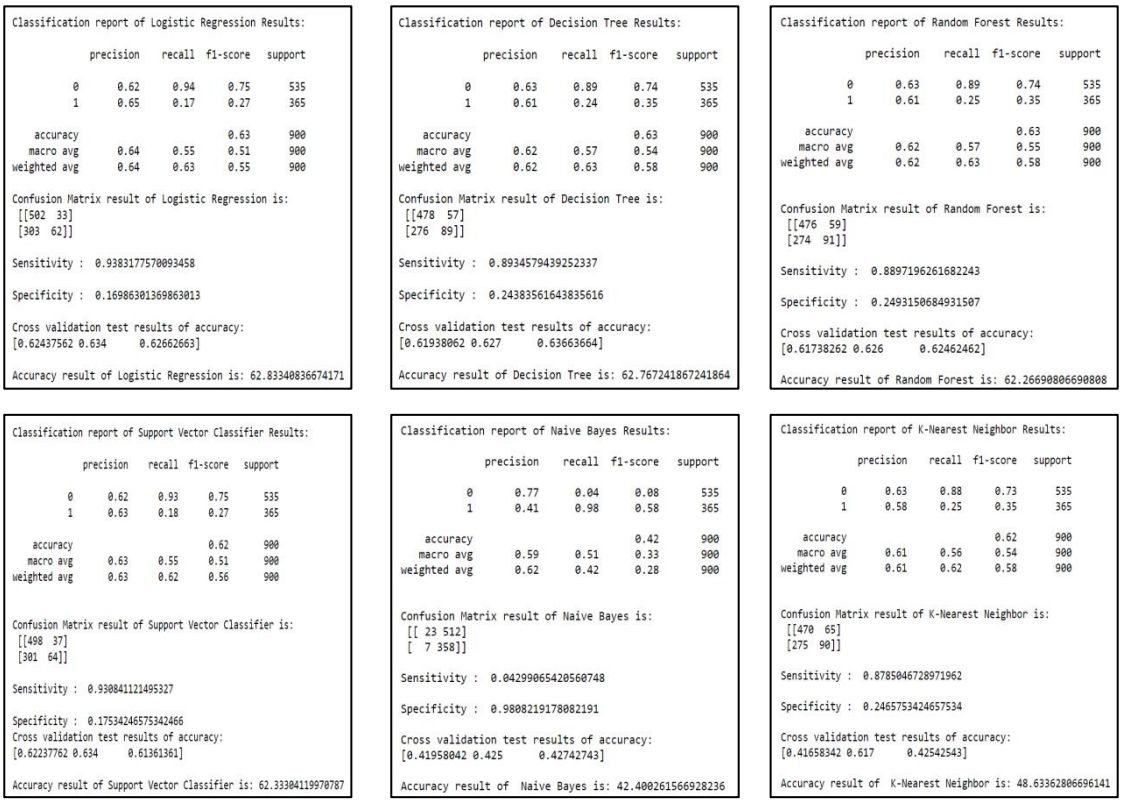
Main goal:

The primary intention of Remote to User attack is to maintain a high level security to ensure safe and trusted communication of information between various organizations.

Solution:

Secured data communication over internet and any other network is always under threat of intrusions and misuse.

Compare and analyse the Algorithms to predict the best accuracy from R2L attack:



Classification report of Logistic Regression Results:

	precision	recall	f1-score	support
0	0.62	0.94	0.75	535
1	0.65	0.17	0.27	365
accuracy			0.63	900
macro avg	0.64	0.55	0.51	900
weighted avg	0.64	0.63	0.55	900

Confusion Matrix result of Logistic Regression is:
[[502 33]
[303 62]]

Sensitivity : 0.9383177570093458

Specificity : 0.16986301369863013

Cross validation test results of accuracy:
[0.62437562 0.634 0.62662663]

Accuracy result of Logistic Regression is: 62.83340836674171

Classification report of Decision Tree Results:

	precision	recall	f1-score	support
0	0.63	0.89	0.74	535
1	0.61	0.24	0.35	365
accuracy			0.63	900
macro avg	0.62	0.57	0.54	900
weighted avg	0.62	0.63	0.58	900

Confusion Matrix result of Decision Tree is:
[[478 57]
[276 89]]

Sensitivity : 0.8934579439252337

Specificity : 0.24383561643835616

Cross validation test results of accuracy:
[0.61938062 0.627 0.63663664]

Accuracy result of Decision Tree is: 62.767241867241864

Classification report of Random Forest Results:

	precision	recall	f1-score	support
0	0.63	0.89	0.74	535
1	0.61	0.25	0.35	365
accuracy			0.63	900
macro avg	0.62	0.57	0.55	900
weighted avg	0.62	0.63	0.58	900

Confusion Matrix result of Random Forest is:
[[476 59]
[274 91]]

Sensitivity : 0.8897196261682243

Specificity : 0.2493150684931507

Cross validation test results of accuracy:
[0.61738262 0.626 0.62462462]

Accuracy result of Random Forest is: 62.26690806690808

Classification report of Support Vector Classifier Results:

	precision	recall	f1-score	support
0	0.62	0.93	0.75	535
1	0.63	0.18	0.27	365
accuracy			0.62	900
macro avg	0.63	0.55	0.51	900
weighted avg	0.63	0.62	0.56	900

Confusion Matrix result of Support Vector Classifier is:
[[498 37]
[301 64]]

Sensitivity : 0.930841121495327

Specificity : 0.17534246575342466
Cross validation test results of accuracy:
[0.62237762 0.634 0.61361361]

Accuracy result of Support Vector Classifier is: 62.33304119970787

Classification report of Naive Bayes Results:

	precision	recall	f1-score	support
0	0.77	0.04	0.08	535
1	0.41	0.98	0.58	365
accuracy			0.42	900
macro avg	0.59	0.51	0.33	900
weighted avg	0.62	0.42	0.28	900

Confusion Matrix result of Naive Bayes is:
[[23 512]
[7 358]]

Sensitivity : 0.04299065420560748

Specificity : 0.9808219178082191

Cross validation test results of accuracy:
[0.41958042 0.425 0.42742743]

Accuracy result of Naive Bayes is: 42.400261566928236

Classification report of K-Nearest Neighbor Results:

	precision	recall	f1-score	support
0	0.63	0.88	0.73	535
1	0.58	0.25	0.35	365
accuracy			0.62	900
macro avg	0.61	0.56	0.54	900
weighted avg	0.61	0.62	0.58	900

Confusion Matrix result of K-Nearest Neighbor is:
[[470 65]
[275 90]]

Sensitivity : 0.8785046728971962

Specificity : 0.2465753424657534

Cross validation test results of accuracy:
[0.41658342 0.617 0.42542543]

Accuracy result of K-Nearest Neighbor is: 48.63362806696141

Fig.Classification reports of each algorithm

R2L attack types:

FTP Write, Multihop, Phf, Spy, Warezclient, Warezmaster, Imap, Guess password, http tunnel, named, send mail, snmpget attack, snmp guess, worm, xlock, xsnoop

Performance measurements of R2L attack prediction:

Parameters	LR	DT	RF	SVC	KNN	NB
Precision	0.62	0.63	0.63	0.62	0.63	0.77
Recall	0.94	0.89	0.89	0.93	0.88	0.04
F1-Score	0.75	0.74	0.74	0.75	0.73	0.08
Sensitivity	0.93	0.89	0.88	0.93	0.87	0.04
Specificity	0.16	0.24	0.24	0.17	0.24	0.98
Accuracy (%)	**62.83**	**62.76**	**62.26**	**62.33**	**48.63**	**42.40**

Graphical Representation of R2L attack prediction:

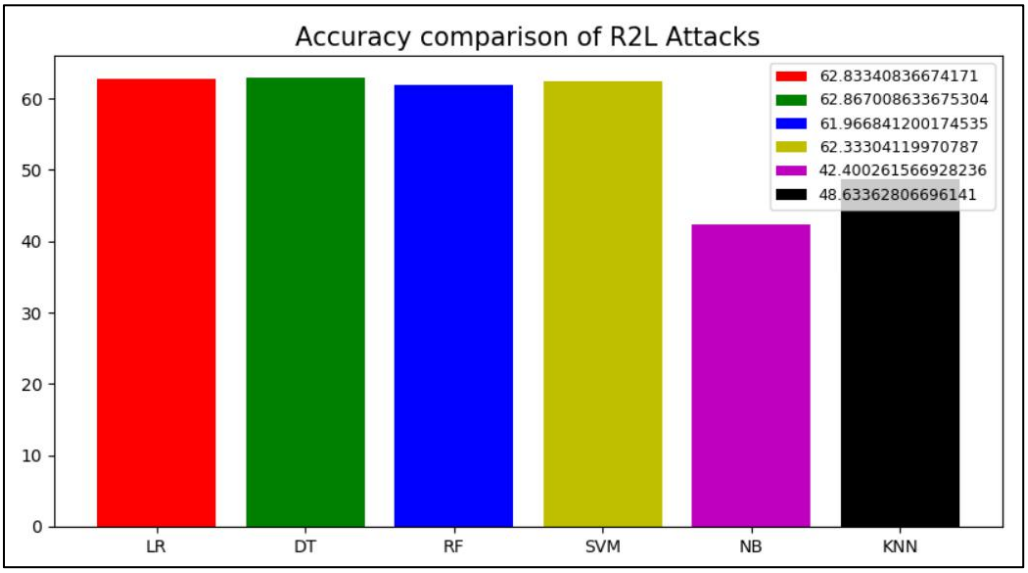


Fig.6.2.a R2L attack prediction

Result:

The highest accuracy for R2L attack is **Logistic Regression algorithm.**

6.3 PREDICTION OF U2R ATTACK

Main goal:

The primary intention of User to Root attack is to identify important features to improve the detection rate of U2R attack class.

Solution:

An attacker or a hacker tries to get the access rights from a normal host in order to gain the root access to the system.

Compare and analyze the Algorithms to predict the best accuracy from U2R attack:

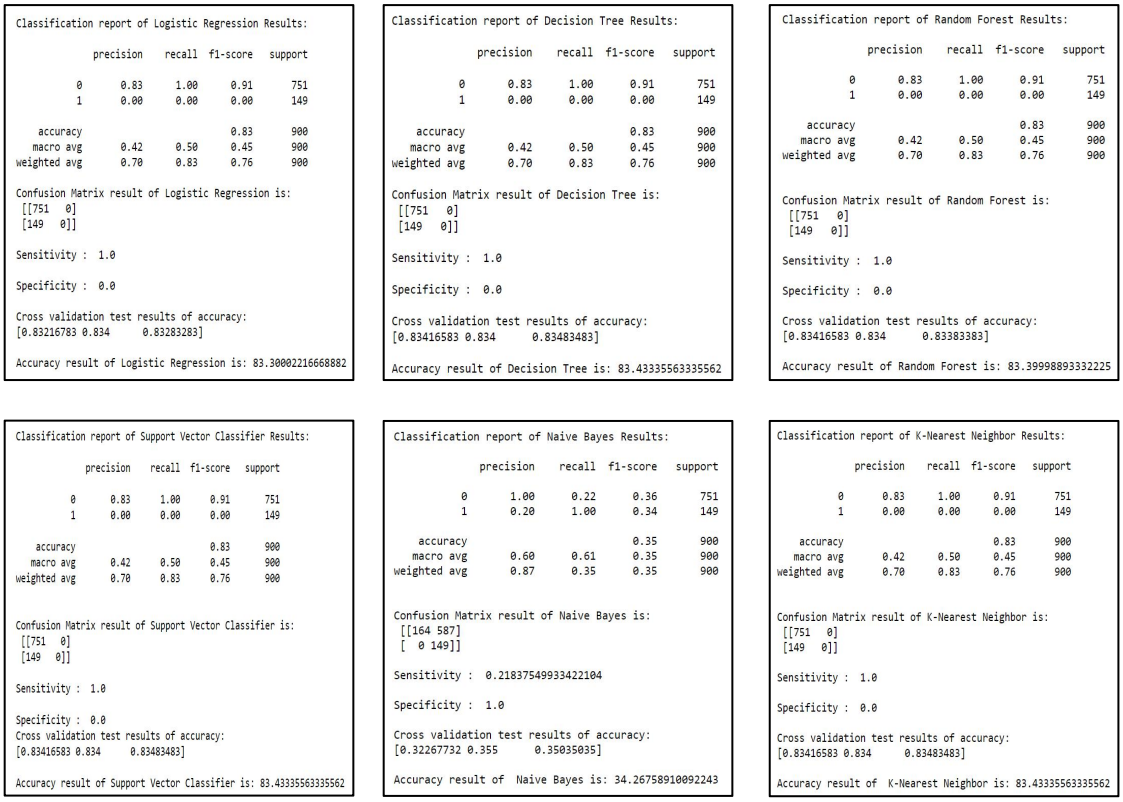


Fig.Classification reports of each algorithm

U2R attack types:

Load mosule, Rerl, Rootkit, Buffer overflow, Ps, Sql attack, xterm.

Performance measurements of U2R attack prediction:

Parameters	LR	DT	RF	SVC	KNN	NB
Precision	0.83	0.83	0.83	0.83	0.83	1
Recall	1	1	1	1	1	0.22
F1-Score	0.91	0.91	0.91	0.91	0.91	0.36
Sensitivity	1	1	1	1	1	0.21
Specificity	0	0	0	0	0	1
Accuracy (%)	**83.30**	**83.43**	**83.39**	**83.43**	**83.43**	**34.26**

Graphical representation of U2R attack prediction:

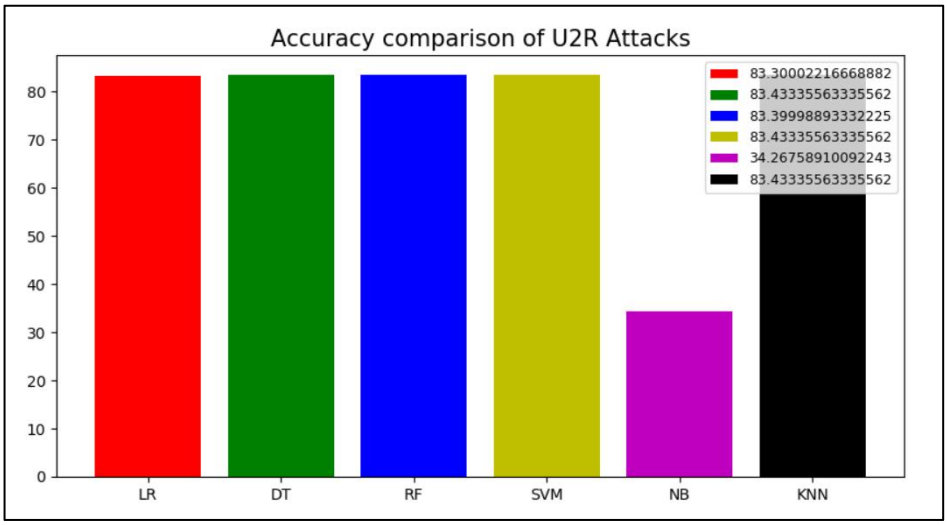


Fig.6.3.a U2R attack prediction

Result:

The highest accuracy for U2R attack is **Decision Tree, Support Vector Classifier** and **K-Nearest Neighbors algorithms.**

6.4 PREDICTION OF PROBE ATTACK

Main goal:

The primary intention of Probe Attack is to learn the detector's location and defensive capabilities and to report it.

Solution:

Analyze the fundamental trade off between the ability of a collaborative network to detect epidemic threats and security of individual participants against probe attacks. Then design and evaluate a collaborative detection system which provides protection against probe attacks.

Compare and analyze the Algorithms to predict the best accuracy from Probe attack:

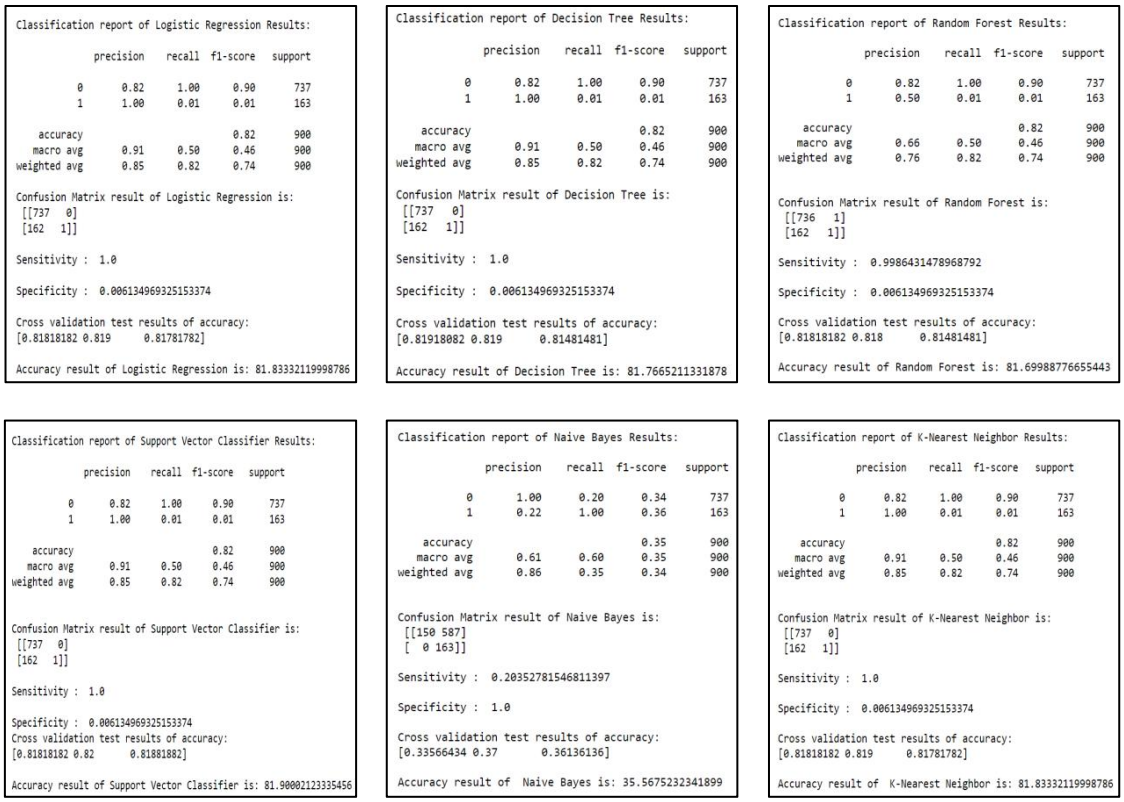


Fig.Classification reports of each algorithm

Probe attack types:

Ip sweep, Nmap, Satan, Port sweep, Msscan, saint.

Performance measurements of Probe attack prediction:

Parameters	LR	DT	RF	SVC	KNN	NB
Precision	0.82	0.82	0.82	0.82	0.82	1
Recall	1	1	1	1	1	0.20
F1-Score	0.91	0.90	0.90	0.90	0.90	0.34
Sensitivity	1	1	0.99	1	1	0.20
Specificity	0	0	0	0	0	1
Accuracy (%)	**81.83**	**81.76**	**81.69**	**81.90**	**81.83**	**35.56**

Graphical representation of Probe attack prediction:

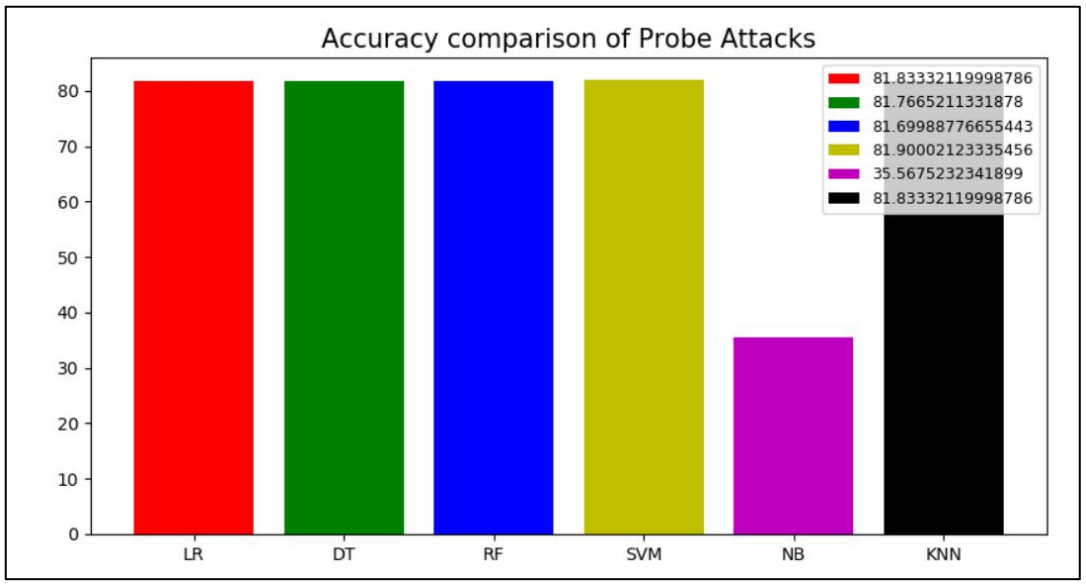


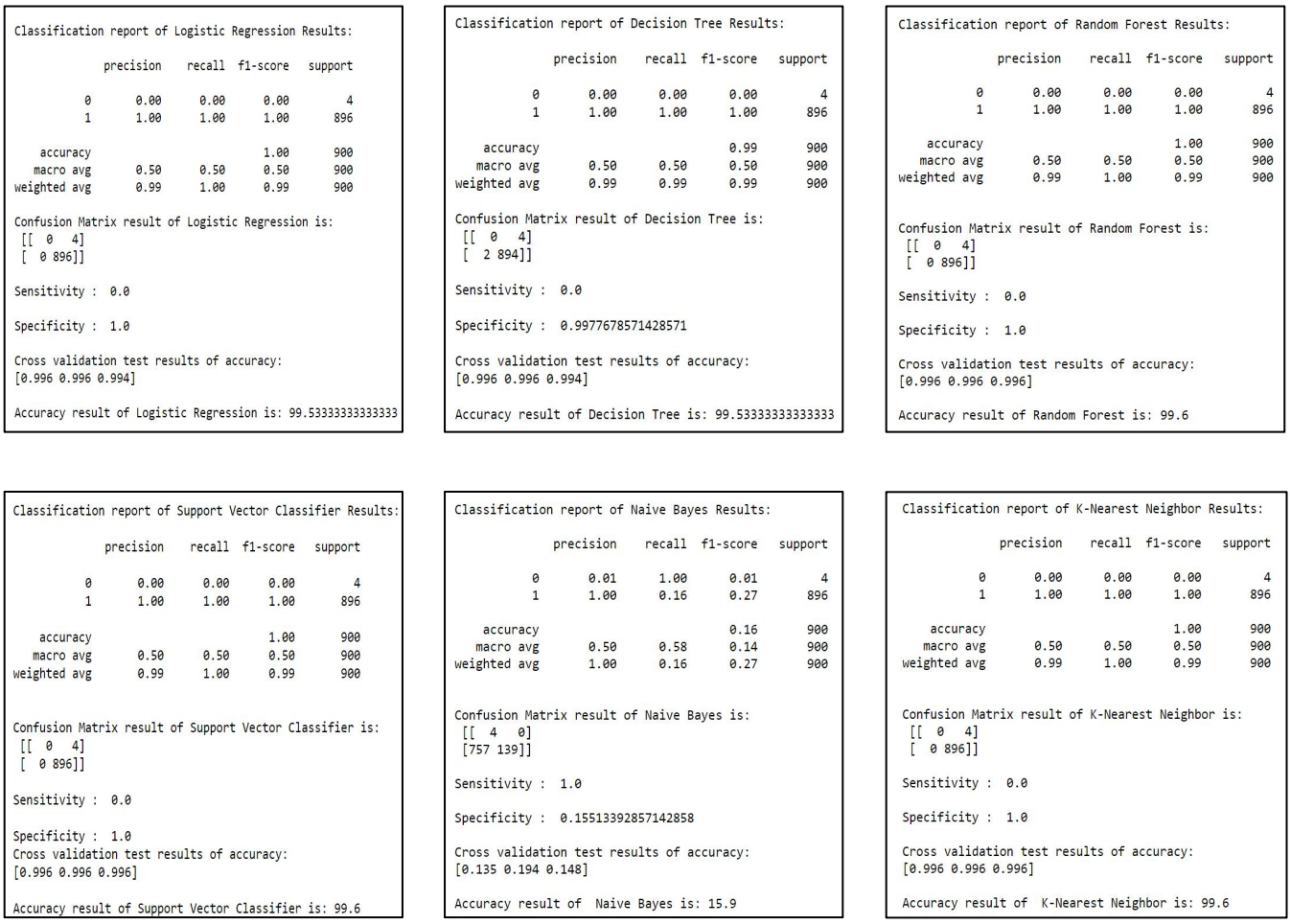
Fig.6.4.a Probe attack prediction

Result:

The highest accuracy for U2R attack is **Support Vector Classifier algorithm.**

6.5 PREDICTION OF OVERALL ATTACKS

Compare and analyze the Algorithms to predict the best accuracy from Overall Attacks:



Classification report of Logistic Regression Results:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	4
1	1.00	1.00	1.00	896
accuracy			1.00	900
macro avg	0.50	0.50	0.50	900
weighted avg	0.99	1.00	0.99	900

Confusion Matrix result of Logistic Regression is:
[[0 4]
[0 896]]

Sensitivity : 0.0

Specificity : 1.0

Cross validation test results of accuracy:
[0.996 0.996 0.994]

Accuracy result of Logistic Regression is: 99.53333333333333

Classification report of Decision Tree Results:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	4
1	1.00	1.00	1.00	896
accuracy			0.99	900
macro avg	0.50	0.50	0.50	900
weighted avg	0.99	0.99	0.99	900

Confusion Matrix result of Decision Tree is:
[[0 4]
[2 894]]

Sensitivity : 0.0

Specificity : 0.9977678571428571

Cross validation test results of accuracy:
[0.996 0.996 0.994]

Accuracy result of Decision Tree is: 99.53333333333333

Classification report of Random Forest Results:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	4
1	1.00	1.00	1.00	896
accuracy			1.00	900
macro avg	0.50	0.50	0.50	900
weighted avg	0.99	1.00	0.99	900

Confusion Matrix result of Random Forest is:
[[0 4]
[0 896]]

Sensitivity : 0.0

Specificity : 1.0

Cross validation test results of accuracy:
[0.996 0.996 0.996]

Accuracy result of Random Forest is: 99.6

Classification report of Support Vector Classifier Results:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	4
1	1.00	1.00	1.00	896
accuracy			1.00	900
macro avg	0.50	0.50	0.50	900
weighted avg	0.99	1.00	0.99	900

Confusion Matrix result of Support Vector Classifier is:
[[0 4]
[0 896]]

Sensitivity : 0.0

Specificity : 1.0
Cross validation test results of accuracy:
[0.996 0.996 0.996]

Accuracy result of Support Vector Classifier is: 99.6

Classification report of Naive Bayes Results:

	precision	recall	f1-score	support
0	0.01	1.00	0.01	4
1	1.00	0.16	0.27	896
accuracy			0.16	900
macro avg	0.50	0.58	0.14	900
weighted avg	1.00	0.16	0.27	900

Confusion Matrix result of Naive Bayes is:
[[4 0]
[757 139]]

Sensitivity : 1.0

Specificity : 0.15513392857142858

Cross validation test results of accuracy:
[0.135 0.194 0.148]

Accuracy result of Naive Bayes is: 15.9

Classification report of K-Nearest Neighbor Results:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	4
1	1.00	1.00	1.00	896
accuracy			1.00	900
macro avg	0.50	0.50	0.50	900
weighted avg	0.99	1.00	0.99	900

Confusion Matrix result of K-Nearest Neighbor is:
[[0 4]
[0 896]]

Sensitivity : 0.0

Specificity : 1.0

Cross validation test results of accuracy:
[0.996 0.996 0.996]

Accuracy result of K-Nearest Neighbor is: 99.6

Fig.Classification reports of each algorithm

Overall attack types:

Back, Neptune, Land, Pod, Smurf, Teardrop, Apache2, Mail bomb, Process table, UDP Storm, FTP Write, Multihop, Phf, Spy, Warezclient, Warezmaster, Imap, Guess password, http tunnel, named, send mail, snmpget attack, snmp guess, worm, xlock, xsnoop, Load mosule, Rerl, Rootkit, Buffer overflow, Ps, Sql attack, xterm, Ip sweep, Nmap, Satan, Port sweep, Msscan, saint.

Performance measurements of Probe attack prediction:

Parameters	LR	DT	RF	SVC	KNN	NB
Precision	1	1	1	1	1	1
Recall	1	1	1	1	1	0.16
F1-Score	1	1	1	1	1	0.27
Sensitivity	0	0	0	0	0	1
Specificity	1	0.99	1	1	1	0.15
Accuracy (%)	**99.53**	**99.53**	**99.6**	**99.6**	**99.6**	**15.9**

Graphical representation of Probe attack prediction:

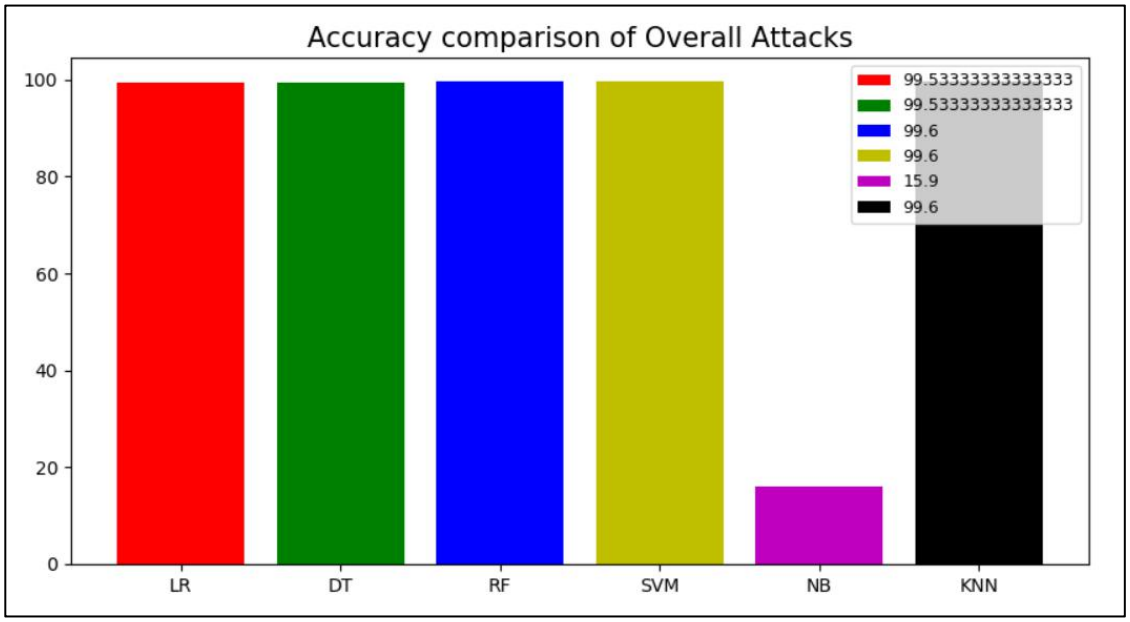


Fig.6.5.a Overall attacks prediction

Result:

The highest accuracy from Overall Attacks are **Random Forest, Support Vector Classifier,** and **K-Nearest Neighbors algorithm.**