

**NETWORK INTRUSION DETECTION**

**FOR IOT SECURITY BASED ON**

**LEARNING TECHNIQUES**

**A PROJECT REPORT**

***Submitted by***

**JEEVIDESH O J 712217205026**

**SHOBA A 712217205063**

**LOKESHWARAKALAIVANI.G 712217205031**

**JEYASHALINI.S 712217205027**

***in partial fulfillment for the award of the degree***

***of***

**BACHELOR OF TECHNOLOGY**

**IN**

INFORMATION TECHNOLOGY

**PARK COLLEGE OF ENGINEERING AND TECHNOLOGY, COIMBATORE 641 659**

**ANNA UNIVERSITY: CHENNAI 600 025**

APRIL 2021

**ANNA UNIVERSITY: CHENNAI 600 025**

**BONAFIDE CERTIFICATE**

|  |  |
| --- | --- |
| **SIGNATURE** | **SIGNATURE** |
| DR.VIJAYLAKSHMI  **HEAD OF THE DEPARTMENT** | MR. RAJA.S  **SUPERVISOR** |
|  |  |
| Department Of Information  Technology | Department Of Information  Technology |
| Park College Of Engineering  And Technology ,Kaniyur ,  Coimbatore | Park College Of Engineering  And Technology ,Kaniyur ,  Coimbatore |

Certified that this project report “**DESIGN AND IMPLEMENTATION OF A SYSTEM TO DETECT CHEATING IN EXAM HALLS**” is the bonafide work of “**JEEVIDESH O J (712217205026), SHOBA A (712217205063), LOKESHWARAKALAIVANI.G (712217205031), JEYASHALINI.S (712217205027)**” who carried out the project work under my supervision.

Submitted for the viva voice held on ………………

**INTERNAL EXAMINER EXTERNAL EXAMINER**

**ABSTRACT**

I

**ABSTRACT**

Intrusion detection is one of the important security problems in today’s cyber world. A significant number of techniques have been developed which are based on machine learning approaches. So for identifying the intrusion we have designed the machine learning algorithms. By using the algorithm we find out intrusion and we can identify the attacker’s details also.

In our project we are proposing two types of algorithm (Decision tree algorithm ), (KNN classification algorithm).Using these algorithm we are going to find out which algorithm gives the best result to detect intrusion attack.

**II**

**TABLE OF CONTENT**

**TABLE OF FIGURES**

**FIGURE NO TITLE PAGE NO**

3.1 Classification and Prediction 11

3.2 Building the Model 12

3.3 Using classifier for classification 13

3.4 Data Mining Application 16

3.5 Changes in Data Mining 19

4.1 KNN Classifier 24

4.2 Example Of KNN 24

4.3 New Data Point 25

4.4 Euclidean Distance 26

4.5 Belongs to A Category 27

5.1 Block diagram Of the System 31

7.1 Project Files 40

7.2 Test Dataset 41

7.3 Train Dataset 42

7.4 Training and Testing data rows and columns 43

7.5 Decision Tree Classifier Evaluation 44

7.6 KNN Classifier Evaluation 45

7.7 Heat Map 46

**VI**

**CHAPTER 1**

**INTRODUCTION**

**1**

**CHAPTER 1**

**INTRODUCTION**

One of the leading technical progressions of computing is “Internet of Things” (IOT). The IOT carries in numerous services, promising individuals’ personal lives obtained from the reliable process. It is forecasted that by 2022 trillion IP objects (addresses) will be associated with the internet. Low accessibility and obscurity of many devices in the massive heterogeneous network makes it problematic to observe the flow of data. However, to protect networks, the intruders who are unauthorized should be identified within the limitations of each kind of device before distributing the system information

A detailed investigation and analysis of various machine learning techniques have been carried out for finding the cause of problems associated with various machine learning techniques in detecting intrusive activities. Issues which are related to detecting low-frequency attacks using network attack dataset are also discussed and viable methods are suggested for improvement. Machine learning techniques have been analysed and compared in terms of their detection capability for detecting the various category of attacks.

* 1. **OBJECTIVE OF THE PROJECT**

To effectively classify and predict the data and to decrease problem and enhance the performance of the overall prediction results.

**2**

**CHAPTER 2**

**LITERATURE SURVEY**

3

**1.Title**: The Internet of Things - How the Next Evolution of the Internet is Changing Everything.

**Year**: 2011

**Author**: D. Evans

**Methodology**

This research describes the methodology and the development process of creating an IoT platform. This paper also presents the architecture and implementation for the IoT platform. The goal of this research is to develop an analytics engine which can gather sensor data from different devices and provide the ability to gain meaningful information from IoT data and act on it using machine learning algorithms.

**Advantage**

The proposed system is introducing the use of a messaging system to improve the overall system performance as well as provide easy scalability.

**Disadvantage**

Low cost devices are easily able to connect wirelessly to the Internet, from handhelds to coffee machines, also known as Internet of Things (IoT).

**2.Title:** The Internet of Things: A survey

**Year:** 2010.

**Author:** L. Atzori, A. Iera, and G. Morabito

**Methodology**

The object unique addressing and the representation and storing of the exchanged information become the mostchallenging issues, bringing directly to a third, ‘‘Semantic oriented”, perspective of IoT.

**Advantage**

People are informed of the scope and the way in which their movements are tracked by the system (taking peo-ple informed about possible leaks of their privacy is essential and required by most legislations).

**Disadvantage**

The user can set the preferences of the proxy. When sensor networks and RFID systems are in-cluded in the network, then the proxy operates between them and the services.

**3.Title**: Addressing the Class Imbalance Problem in Medical Datasets

**Year**: 2012

**Author**: M. Mostafizur Rahman and D. N. Davis

**Methodology**

A balanced dataset is very important for creating a good training set. They aim to optimize the overall accuracy without considering the relative distribution of each class. Typically real world data are usually imbalanced and it is one of the main causes for the decrease of generalization in machine learning algorithms.

**Advantage**

The aim was to reduce the ratio gap between the majority classes with the minority class. The proposed method is found to be useful for such datasets where the class labels are not certain and can also help to overcome the class imbalance problem of clinical datasets and also for other data domains.

**Disadvantage**

The outcome labels of most of the clinical datasets are not consistent with the underlying data. The conventional over-sampling and under-sampling technique may not always be appropriate for such datasets.

**4.Title:** A survey on cloud computing security

**Year:** 2010.

**Author:** R. Kanday

**Methodology**

This survey paper provides a general overview on Cloud Computing. The topics that are discussed include characteristics, deployment and service models as well drawbacks.

**Advantage**

The major part of countermeasures focuses on Intrusion Detection Systems. Moving towards Mobile Cloud Computing and Internet of Things, this survey paper gives a general explanation on the applications and potential that comes with the integration of Cloud Computing with any device that has Internet connectivity as well as the challenges that are before it.

**Disadvantage**

Several security issues and countermeasures are also discussed to show the major issues and obstacles that Cloud Computing faces as it is being implemented further.

**5.Title**: Data Mining: Practical Machine Learning Tool and Technique with Java Implementation

**Year**: 2000.

**Author**: Ian H. Witten and Eibe Frank

**Methodology**

The convergence of computing and communication has produced a society that feeds on information. Yet most of the information is in its raw form: data. If data is characterized as recorded facts, then information is the set of patterns, or expectations, that underlie the data. There is a huge amount of information locked up in databases—information that is potentially important but has not yet been discovered or articulated. Our mission is to bring it forth. The weather data (Tables 1.2 and 1.3) presents a set of days together with a decision for each as to whether to play the game or not.

**Advantage**

In these cases the output took the form of decision trees and classification rules, which are basic knowledge representation styles that many machine learning methods used.

**Disadvantage**

The weather problem is a tiny dataset that we will use repeatedly to illustrate machine learning methods.

**CHAPTER 4**

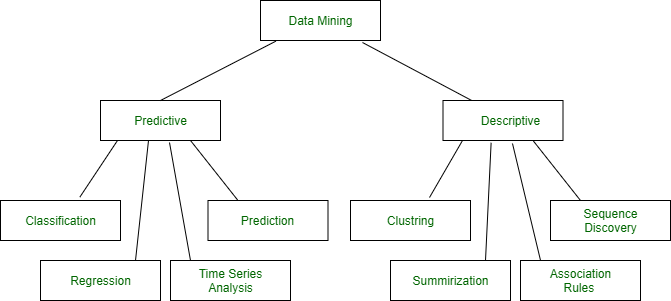
**DATA MINING**

**CHAPTER 3**

**3.1 DATA MINING**

* Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data. Data mining is also called **Knowledge Discovery in Database (KDD)*.*** The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.
* The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining.
* In other words, we can say that Data Mining is the process of investigating hidden patterns of information to various perspectives for categorization into useful data, which is collected and assembled in particular areas such as data warehouses, efficient analysis, data mining algorithm, helping decision making and other data requirement to eventually cost-cutting and generating revenue.
* Data mining is the act of automatically searching for large stores of information to find trends and patterns that go beyond simple analysis procedures. Data mining utilizes complex mathematical algorithms for data segments and evaluates the probability of future events.
* Data Mining is a process used by organizations to extract specific data from huge databases to solve business problems. It primarily turns raw data into useful information.

**10**



**Figure 3.1 Data Mining Classification and Prediction**

**3.2 CLASSIFICATION**:

 It is a data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known

**Example** we can build a classification model to categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.

A bank loan officer wants to analyse the data in order to know which customer (loan applicant) are risky or which are safe.

A marketing manager at a company needs to analyse a customer with a given profile, who will buy a new computer.

In both of the above examples, a model or classifier is constructed to predict the categorical labels. These labels are risky or safe for loan application data and yes or no for marketing data.

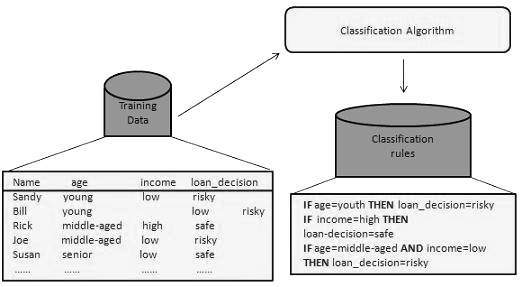
**3.2.1 CLASSIFICATION PROCESS:**

With the help of the bank loan application that we have discussed above, let us understand the working of classification. The Data Classification process includes two steps −

* Building the Classifier or Model
* Using Classifier for Classification

**Building the Classifier or Model**

* This step is the learning step or the learning phase.
* In this step the classification algorithms build the classifier.
* The classifier is built from the training set made up of database tuples and their associated class labels.
* Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points.

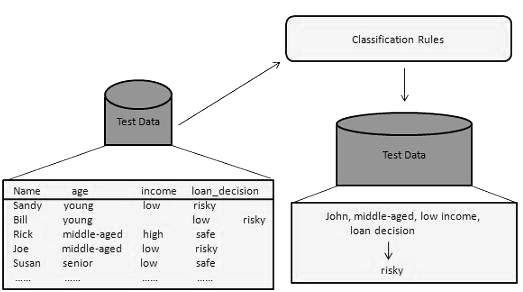


**Figure 3.2 –Building the model**

**12**

**Using Classifier for Classification**

In this step, the classifier is used for classification. Here the test data is used to estimate the accuracy of classification rules. The classification rules can be applied to the new data tuples if the accuracy is considered acceptable.



**Figure 3.3-Using Classifier for Classification**

**3.3 Prediction:**

* The second way to operate data mining is Prediction. It is repeatedly used to detect several data. Same thing as over in classification, the behaviour of data set holds the inputs and similar numerical output values. Compatible with the behaviour of the dataset, the algorithm (division) gets the model or a predictor.
* When the new information is given, the model should detect a numerical output. Despite the classification, this procedure does not have the class label or notes. The model estimates the current valued action or command value.
* Regression (Growth) in most cases is used for Predication. Predicting the price of a house reply on cases such as the number of the apartment, the total region, and so on is an illustration for predication. An organization has the power to find the amount of banknotes payout by the person during a negotiation.

**13**

Examples of cases where the data analysis task is Prediction −

Suppose the marketing manager needs to predict how much a given customer will spend during a sale at his company. In this example we are bothered to predict a numeric value. Therefore the data analysis task is an example of numeric prediction. In this case, a model or a predictor will be constructed that predicts a continuous-valued-function or ordered value.

**3.4 CLASSIFICATION AND PREDICTION ISSUES**

The major issue is preparing the data for Classification and Prediction. Preparing the data involves the following activities −

* **Data Cleaning** − Data cleaning involves removing the noise and treatment of missing values. The noise is removed by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with most commonly occurring value for that attribute.
* **Relevance Analysis** − Database may also have the irrelevant attributes. Correlation analysis is used to know whether any two given attributes are related.
* **Data Transformation and reduction** − The data can be transformed by any of the following methods.
  + **Normalization** − The data is transformed using normalization. Normalization involves scaling all values for given attribute in order to make them fall within a small specified range. Normalization is used when in the learning step, the neural networks or the methods involving measurements are used.
  + **Generalization** − The data can also be transformed by generalizing it to the higher concept. For this purpose we can use the concept hierarchies.

**3.5 COMPARISON OF CLASSIFICATION AND**

**PREDICTION:**

Here is the criteria for comparing the methods of Classification and Prediction.

**14**

* **Accuracy** − Accuracy of classifier refers to the ability of classifier. It predict the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.
* **Speed** − This refers to the computational cost in generating and using the classifier or predictor.
* **Robustness** − It refers to the ability of classifier or predictor to make correct predictions from given noisy data.
* **Scalability** − Scalability refers to the ability to construct the classifier or predictor efficiently; given large amount of data.
* **Interpretability** − It refers to what extent the classifier or predictor understand.

**3.6 ADVANTAGES OF DATAMINING:**

* The Data Mining technique enables organizations to obtain knowledge-based.
* Data mining enables organizations to make lucrative modifications in operation and production.
* Compared with other statistical data applications, data mining is a cost-efficient.
* Data Mining helps the decision-making process of an organization.
* It Facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviours.
* It can be induced in the new system as well as the existing platforms.
* It is a quick process that makes it easy for new users to analyse enormous amounts of data in a short time.

**3.7 DISADVANTAGES OF DATAMINING:**

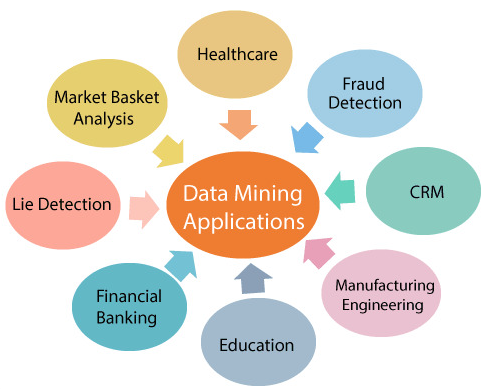
* There is a probability that the organizations may sell useful data of customers to other organizations for money. As per the report, American Express has sold credit card purchases of their customers to other organizations.

**15**

* Many data mining analytics software is difficult to operate and needs advance training to work on.
* Different data mining instruments operate in distinct ways due to the different algorithms used in their design. Therefore, the selection of the right data mining tools is a very challenging task.
* The data mining techniques are not precise, so that it may lead to severe consequences in certain conditions.

**3.8 DATA MINING APPLICATION:**

* Data Mining is primarily used by organizations with intense consumer demands- Retail, Communication, Financial, marketing company, determine price, consumer preferences, product positioning, and impact on sales, customer satisfaction, and corporate profits. Data mining enables a retailer to use point-of-sale records of customer purchases to develop products and promotions that help the organization to attract the customer.



**Figure 3.4 Data Mining Applications**

**16**

**Data Mining in Healthcare:**

Data mining in healthcare has excellent potential to improve the health system. It uses data and analytics for better insights and to identify best practices that will enhance health care services and reduce costs. Analysts use data mining approaches such as Machine learning, Multi-dimensional database, Data visualization, Soft computing, and statistics. Data Mining can be used to forecast patients in each category. The procedures ensure that the patients get intensive care at the right place and at the right time. Data mining also enables healthcare insurers to recognize fraud and abuse.

**Data Mining in Market Basket Analysis:**

Market basket analysis is a modeling method based on a hypothesis. If you buy a specific group of products, then you are more likely to buy another group of products. This technique may enable the retailer to understand the purchase behaviour of a buyer. This data may assist the retailer in understanding the requirements of the buyer and altering the store's layout accordingly. Using a different analytical comparison of results between various stores, between customers in different demographic groups can be done.

**Data mining in Education:**

Education data mining is a newly emerging field, concerned with developing techniques that explore knowledge from the data generated from educational Environments. EDM objectives are recognized as affirming student's future learning behaviour, studying the impact of educational support, and promoting learning science. An organization can use data mining to make precise decisions and also to predict the results of the student. With the results, the institution can concentrate on what to teach and how to teach.

**Data Mining in Manufacturing Engineering:**

Knowledge is the best asset possessed by a manufacturing company. Data mining tools can be beneficial to find patterns in a complex manufacturing process. Data mining can be used in system-level designing to obtain the relationships between product architecture, product portfolio, and data needs of the customers. It can also be used to forecast the product development period, cost, and expectations among the other tasks.

**17**

**Data Mining in CRM (Customer Relationship Management):**

Customer Relationship Management (CRM) is all about obtaining and holding Customers, also enhancing customer loyalty and implementing customer-oriented strategies. To get a decent relationship with the customer, a business organization needs to collect data and analyse the data. With data mining technologies, the collected data can be used for analytics.

**Data Mining in Fraud detection:**

Billions of dollars are lost to the action of frauds. Traditional methods of fraud detection are a little bit time consuming and sophisticated. Data mining provides meaningful patterns and turning data into information. An ideal fraud detection system should protect the data of all the users. Supervised methods consist of a collection of sample records, and these records are classified as fraudulent or non-fraudulent. A model is constructed using this data, and the technique is made to identify whether the document is fraudulent or not.

**Data Mining in Lie Detection:**

Apprehending a criminal is not a big deal, but bringing out the truth from him is a very challenging task. Law enforcement may use data mining techniques to investigate offenses, monitor suspected terrorist communications, etc. This technique includes text mining also, and it seeks meaningful patterns in data, which is usually unstructured text. The information collected from the previous investigations is compared, and a model for lie detection is constructed.

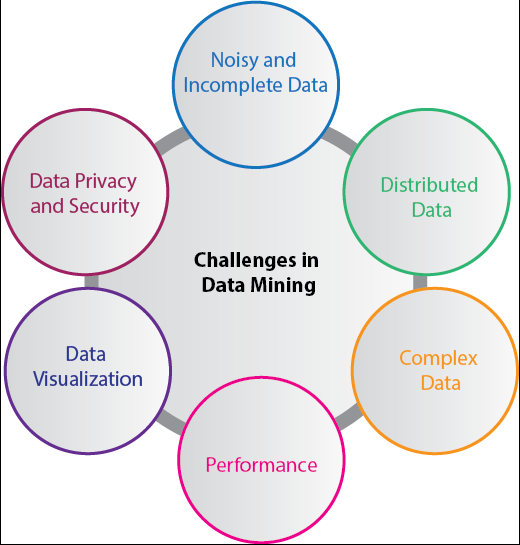
**Data Mining Financial Banking:**

The Digitalization of the banking system is supposed to generate an enormous amount of data with every new transaction. The data mining technique can help bankers by solving business-related problems in banking and finance by identifying trends, casualties, and correlations in business information and market costs that are not instantly evident to managers or executives because the data volume is too large or are produced too rapidly on the screen by experts. The manager may find these data for better targeting, acquiring, retaining, segmenting, and maintain a profitable customer.

**18**

**3.9 CHALLENGES IMPLEMENTING IN DATAMINING**

Although data mining is very powerful, it faces many challenges during its execution. Various challenges could be related to performance, data, methods, and techniques, etc. The process of data mining becomes effective when the challenges or problems are correctly recognized and adequately resolved.



**Figure 3.5 – Changes in data Mining**

**Incomplete and noisy data:**

The process of extracting useful data from large volumes of data is data mining. The data in the real-world is heterogeneous, incomplete, and noisy. Data in huge quantities will usually be inaccurate or unreliable. These problems may occur due to data measuring instrument or because of human errors. Suppose a retail chain collects phone numbers of customers who spend more than $ 500, and the accounting employees put the information into their system. The person may make a digit mistake when entering the phone number, which results in incorrect data. Even some customers may not be willing to disclose their phone numbers, which

**19**

results in incomplete data. The data could get changed due to human or system error. All these consequences (noisy and incomplete data)makes data mining challenging.

**Data Distribution:**

Real-worlds data is usually stored on various platforms in a distributed computing environment. It might be in a database, individual systems, or even on the internet. Practically, It is a quite tough task to make all the data to a centralized data repository mainly due to organizational and technical concerns. For example, various regional offices may have their servers to store their data. It is not feasible to store, all the data from all the offices on a central server. Therefore, data mining requires the development of tools and algorithms that allow the mining of distributed data.

**Complex Data:**

Real-world data is heterogeneous, and it could be multimedia data, including audio and video, images, complex data, spatial data, time series, and so on. Managing these various types of data and extracting useful information is a tough task. Most of the time, new technologies, new tools, and methodologies would have to be refined to obtain specific information.

**Performance:**

The data mining system's performance relies primarily on the efficiency of algorithms and techniques used. If the designed algorithm and techniques are not up to the mark, then the efficiency of the data mining process will be affected adversely.

**Data Privacy and Security:**

Data mining usually leads to serious issues in terms of data security, governance, and privacy. For example, if a retailer analyses the details of the purchased items, then it reveals data about buying habits and preferences of the customers without their permission

**20**

**Data Visualization:**

In data mining, data visualization is a very important process because it is the primary method that shows the output to the user in a presentable way. The extracted data should convey the exact meaning of what it intends to express. But many times, representing the information to the end-user in a precise and easy way is difficult. The input data and the output information being complicated, very efficient, and successful data visualization processes need to be implemented to make it useful .

|  |
| --- |
|  |

**21**

**CHAPTER 4 ALGORITHM**

**22**

**CHAPTER 4**

**ALGORITHM**

**4.1 KNEAREST NEIGHBOUR ALGORITHM**

* K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
* K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
* K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
* K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
* K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
* It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
* KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
* **Example:** Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.

**23**



## Figure 4.1 –KNN Classifier

## 4.1.1 Need of K-NN Algorithm

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram:



**Figure 4.2 –Example of KNN**

**24**

**4.1.2 K-NN working process**

* **Step-1:** Select the number K of the neighbors
* **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
* **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
* **Step-4:** Among these k neighbors, count the number of the data points in each category.
* **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
* **Step-6:** Our model is ready.

Suppose we have a new data point and we need to put it in the required category. Consider the below image:



**Figure 4.3-New data point**

**25**

* Firstly, we will choose the number of neighbors, so we will choose the k=5.
* Next, we will calculate the **Euclidean distance** between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:



**Figure 4.4 – Euclidean distance**

* By calculating the Euclidean distance we got the nearest neighbours, as three nearest neighbours in category A and two nearest neighbours in category B. Consider the below image:

**26**



**Figure 4.5- New data point belong to Category A**

* As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.

## 4.1.3 Value of K in the K-NN Algorithm

Below are some points to remember while selecting the value of K in the K-NN algorithm:

* 8There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.
* A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
* Large values for K are good, but it may find some difficulties.

## 4.2 ADVANTAGES OF KNN ALGORITHM:

* It is simple to implement.
* It is robust to the noisy training data
* It can be more effective if the training data is large.

## 27

## 4.3 DISADVANTAGES OF KNN ALGORITHM:

* Always needs to determine the value of K which may be complex some time.
* The computation cost is high because of calculating the distance between the data
* Points for all the training samples.

**4.4 DECISION TREE ALGORITHM:**

A decision tree is a plan that includes a root node, branches, and leaf nodes. Every internal node characterizes an examination on an attribute, each division characterizes the consequence of an examination, and each leaf node grasps a class tag. The primary node in the tree is the root node.

The subsequent decision tree is for the thought buy a computer that shows whether a purchaser at an enterprise is expected to buy a computer or not. Each internal node characterizes an inspection on an attribute. Each leaf node signifies a class.

**4.4.1** **DECISION TREE ALGORITHM IN DATA MININg:**

## Decision Tree algorithm relates to the persons of directed intelligence techniques. Unlike other-directed education procedures, the decision tree algorithm can be used to answer deterioration and arrangement difficulties.

The objective of using a Decision Tree is to craft a preparation ideal that can use to foresee the class or value of the mark variable by learning easy judgement procedures incidental from previous information (training data).

In Decision Trees, for estimating a class tag for best ever we start with the root of the tree. We make relations with the root attribute to the record’s attribute.

**28**

We make division agreeing to that value and jump to the subsequent node on the base of choice.

 4.5 **ADVANTAGES OF DECISION TREE**

* In comparison to other procedures, decision trees need not as much energy for information training during pre-processing.
* A decision tree does not involve stabilization of information.
* A decision tree does not need scaling of information as well.
* Omitted values in the information also do not disturb the procedure of constructing a decision tree to any substantial degree.
* A Decision tree model is identical instinctive and stress-free to describe to practical squads as well as investors.

## ****4.6 DISADVANTAGES OF DECISION TREE****

* A minor variation in the information can cause a huge variation in the configuration of the decision tree triggering unpredictability.
* For a Decision tree occasionally calculation can go far extra multifaceted in comparison to other procedures.
* Decision tree repeatedly takes greater time to train the model.
* Decision tree preparation is comparatively lavish as the difficulty and period have taken are extra.
* The Decision Tree procedure is insufficient for relating deterioration and forecasting uninterrupted values

**29**

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

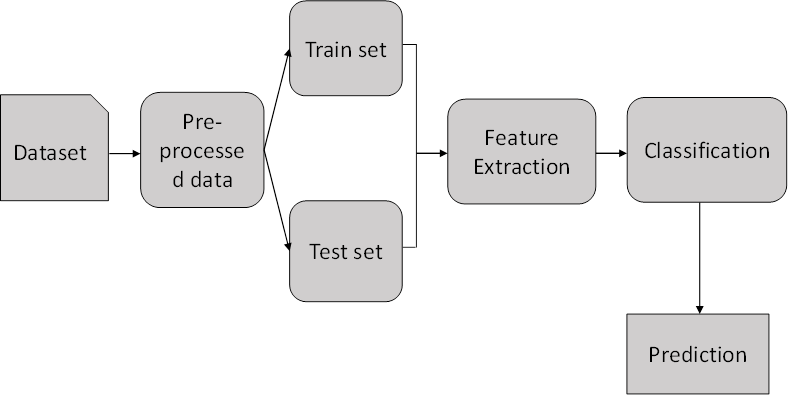
**CHAPTER 5**

**SYSTEM MODUES**

**30**

**CHAPTER 5**

**BLOCK DIAGRAM**



**Figure 5.1-Block Diagram of the System**

**5.1 DATA SELECTION AND LOADING**

* Data selection is the process of determining the appropriate data type and source, as well as suitable instruments to collect data.
* Data selection precedes the actual practice of data collection and it is the process where data relevant to the analysis is decided and retrieved from the data collection.
* Data loading refers to the "load" component.
* After data is retrieved and combined from multiple sources, cleaned and formatted, it is then loaded into a storage system, such as a cloud data warehouse.
* In this project, the NSL-KDD dataset is used for detecting attacks.

**31**

**5.2 DATA PREPROCESSING**

* The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.
* **Missing Data:**  
  This situation arises when some data is missing in the data. It can be handled in various ways.
  + - Ignore the tuples:  
      This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
    - Fill the Missing values:  
      There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.
* **Regression:**  
  Data can be made smooth by fitting it to a regression function. The regression used may be linear or multiple .

**5.3 SPLITTING DATA SET INTO TRAIN AND TEST**

* Data splitting is the act of partitioning available data into two portions, usually for cross-validator purposes.
* One Portion of the data is used to develop a predictive model and the other to evaluate the model's performance.
* Separating data into training and testing sets is an important part of evaluating data mining models.

**32**

* Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing.
* To train any machine learning model irrespective what type of dataset is being used you have to split the dataset into training data and testing data.

**5.4 FEATURE EXTRACTION**

* Feature scaling. Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.
* Feature Scaling or Standardization: It is a step of Data Pre Processing which is applied to independent variables or features of data. It basically helps to normalise the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

**5.5 CLASSIFICATION**

* Classification is the problem of identifying to which of a set of categories, a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.
* **K-NN algorithm** assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data

**33**

**5.6 PREDICTION**

Predictive analytics algorithms try to achieve the lowest error possible by either using “boosting” or “bagging”.

**Accuracy** − Accuracy of classifier refers to the ability of classifier. It predict the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

**Speed** − Refers to the computational cost in generating and using the classifier or predictor.

**Robustness** − It refers to the ability of classifier or predictor to make correct predictions from given noisy data.

**Scalability** − Scalability refers to the ability to construct the classifier or predictor efficiently; given large amount of data.

**Interpretability** − It refers to what extent the classifier or predictor understands.

**5.7RESULT GENERATION**

The Final Result will get generated based on the overall classification and prediction. The performance of this proposed approach is evaluated using some measures like,

* Accuracy

**Accuracy** of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

**34**

AC=

* Precision

**Precision** is defined as the number of true positives divided by the number of true positives plus the number of false positives.

Precision=

* Recall

**Recall** is the number of correct results divided by the number of results that should have been returned. In binary classification, recall is called sensitivity. It can be viewed as the probability that a relevant document is retrieved by the query.

Recall=

* F-Measure

**F measure** (F1 score or F score) is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test.

**35**

**CHAPTER 6**

**SOFTWARE IMPLEMENTATION**

**36**

**CHAPTER 6**

**SOFTWARE IMPLEMENTATION**

**6.1 PYTHON**

* **Python**  is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. It was created by Guido van Rossum during 1985- 1990. Like Perl, Python source code is also available under the GNU General Public License (GPL). This **tutorial** gives enough understanding on **Python programming** language.
* **Python** is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.
* **Python** is a MUST for students and working professionals to become a great Software Engineer specially when they are working in Web Development Domain. I will list down some of the key advantages of learning Python:
* **Python is Interpreted** − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
* **Python is Interactive** − You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
* **Python is Object-Oriented** − Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
* **Python is a Beginner's Language** − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

**37**

**6.3** **PACKAGE –ANACONDA**

**Anaconda** is a [distribution](https://en.wikipedia.org/wiki/Software_distribution) of the [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) and [R](https://en.wikipedia.org/wiki/R_(programming_language)) programming languages for [scientific computing](https://en.wikipedia.org/wiki/Scientific_computing) ([data science](https://en.wikipedia.org/wiki/Data_science), [machine learning](https://en.wikipedia.org/wiki/Machine_learning) applications, large-scale data processing, [predictive analytics](https://en.wikipedia.org/wiki/Predictive_analytics), etc.), that aims to simplify [package management](https://en.wikipedia.org/wiki/Package_management) and deployment. The distribution includes data-science packages suitable for Windows, Linux, and macOS. It is developed and maintained by Anaconda, Inc., which was founded by Peter Wang and [Travis Oliphant](https://en.wikipedia.org/wiki/Travis_Oliphant) in 2012. As an Anaconda, Inc. product, it is also known as **Anaconda Distribution** or **Anaconda Individual Edition**, while other products from the company are Anaconda Team Edition and Anaconda Enterprise Edition, both of which are not free.

Package versions in Anaconda are managed by the [package management system](https://en.wikipedia.org/wiki/Package_manager) [conda](https://en.wikipedia.org/wiki/Conda_(package_manager)). This package manager was spun out as a separate open-source package as it ended up being useful on its own and for other things than Python.[[10]](https://en.wikipedia.org/wiki/Anaconda_(Python_distribution)#cite_note-10) There is also a small, bootstrap version of Anaconda called **Miniconda**, which includes only conda, Python, the packages they depend on, and a small number of other packages.

**6.3 SOFTWARE SPYDER**

**Spyder**  is an [open-source](https://en.wikipedia.org/wiki/Open-source_software) cross-platform [integrated development environment](https://en.wikipedia.org/wiki/Integrated_development_environment) (IDE) for scientific programming in the [Python language](https://en.wikipedia.org/wiki/Python_(programming_language)). Spyder integrates with a number of prominent packages in the scientific Python stack, including [NumPy](https://en.wikipedia.org/wiki/NumPy), [SciPy](https://en.wikipedia.org/wiki/SciPy), [Matplotlib](https://en.wikipedia.org/wiki/Matplotlib), [pandas](https://en.wikipedia.org/wiki/Pandas_(software)), [IPython](https://en.wikipedia.org/wiki/IPython), [SymPy](https://en.wikipedia.org/wiki/SymPy) and [Cython](https://en.wikipedia.org/wiki/Cython), as well as other open-source software. It is released under the [MIT license](https://en.wikipedia.org/wiki/MIT_license).

Initially created and developed by Pierre Raybaut in 2009, since 2012 Spyder has been maintained and continuously improved by a team of scientific Python developers and the community.

Spyder is extensible with first-party and third-party plugins,[ includes support for interactive tools for data inspection and embeds Python-specific code quality assurance and introspection instruments, such as Pyflakes, [Pylint](https://en.wikipedia.org/wiki/Pylint) and Rope. It is available cross-platform through [Anaconda](https://en.wikipedia.org/wiki/Anaconda_(Python_distribution)), on Windows, on macOS through [MacPorts](https://en.wikipedia.org/wiki/MacPorts), and on major Linux distributions such as [Arch Linux](https://en.wikipedia.org/wiki/Arch_Linux), [Debian](https://en.wikipedia.org/wiki/Debian), [Fedora](https://en.wikipedia.org/wiki/Fedora_(operating_system)), [Gentoo Linux](https://en.wikipedia.org/wiki/Gentoo_Linux), [openSUSE](https://en.wikipedia.org/wiki/OpenSUSE) and [Ubuntu](https://en.wikipedia.org/wiki/Ubuntu_(operating_system)).

**38**

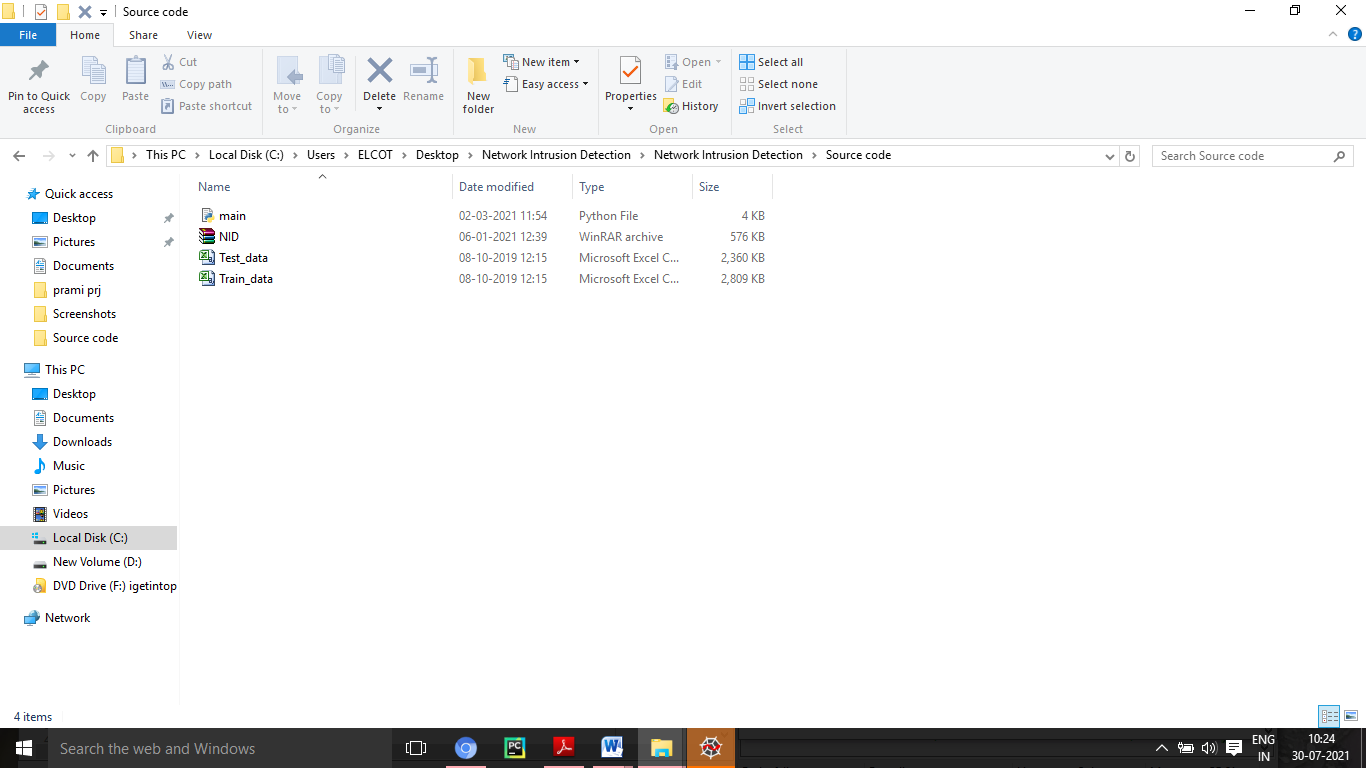
**CHAPTER 7**

**SCREENSHOT**

**39**

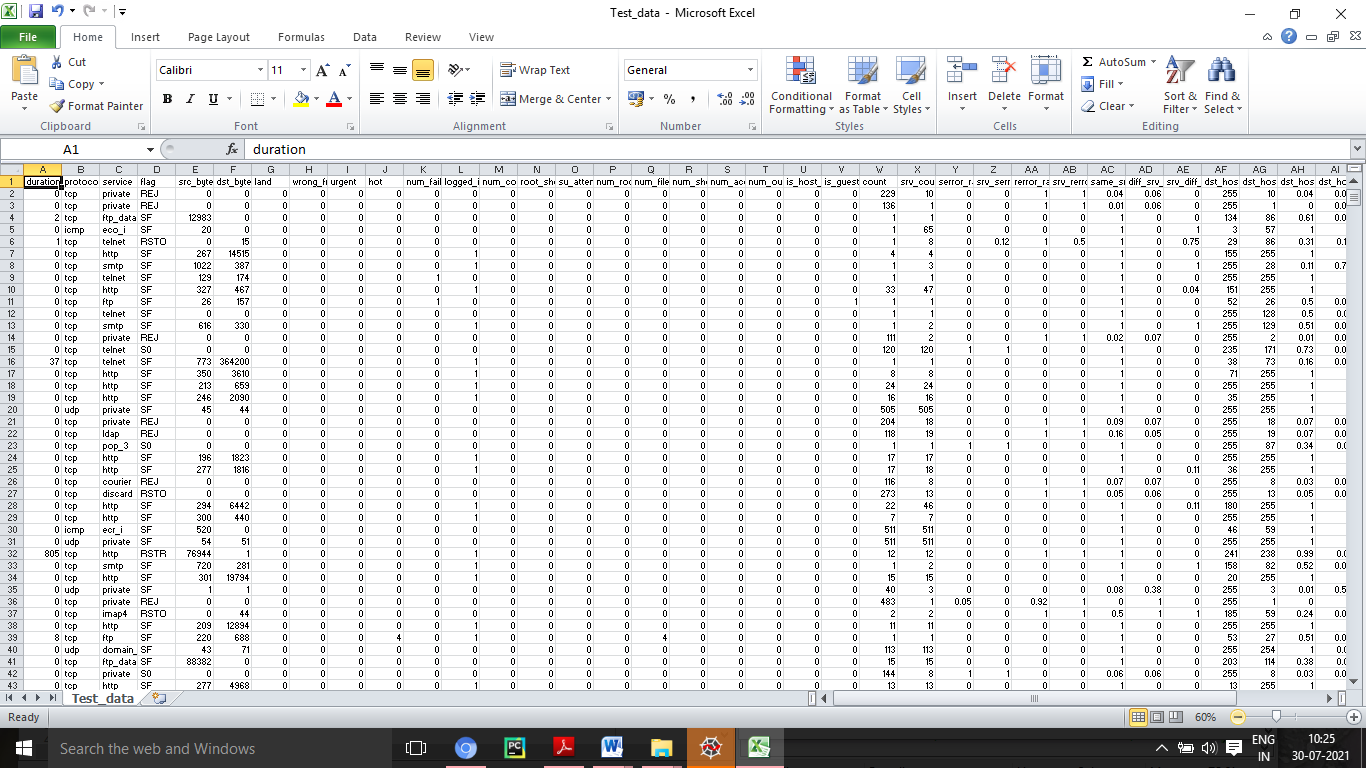
**CHAPTER 7**

All are project related files are together hold in single folder ,it includes test and train dataset and source code.



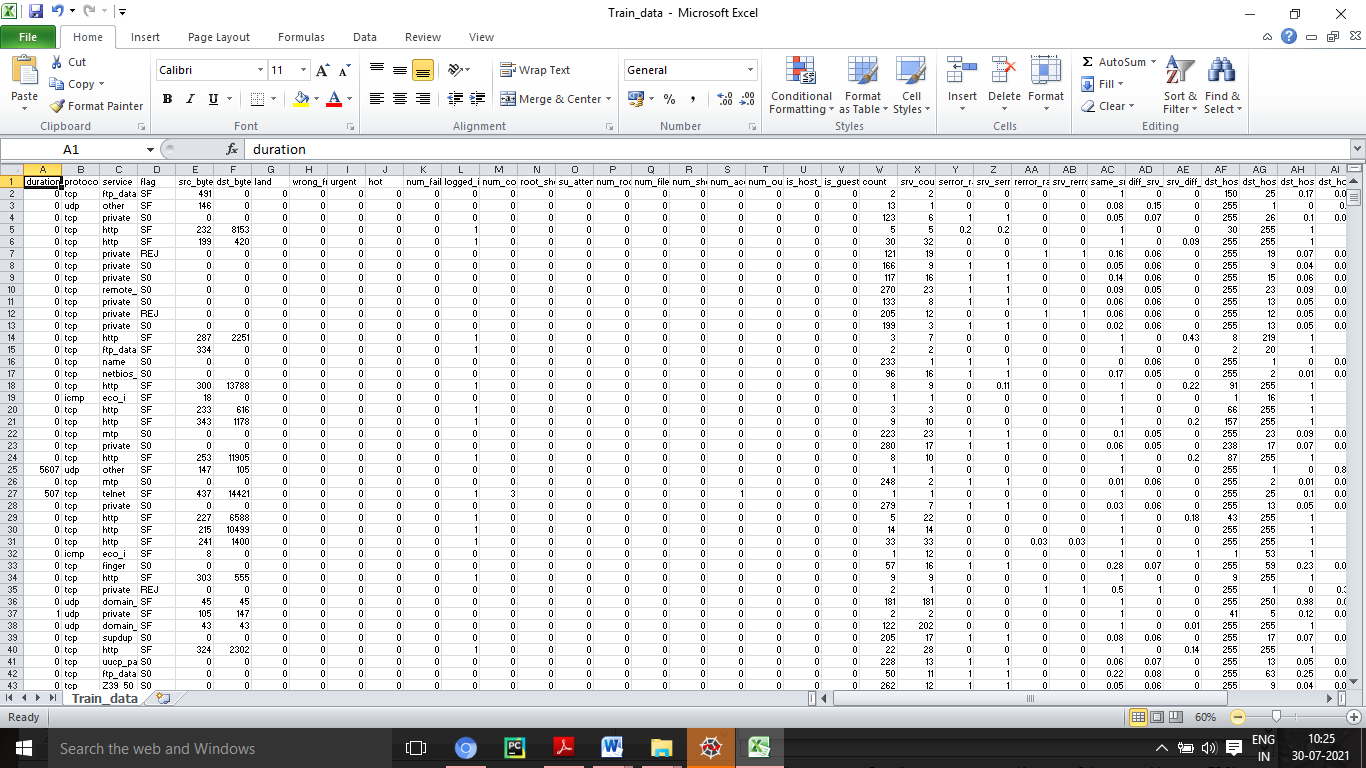
**Figure 7.1 Project Files**

**40**



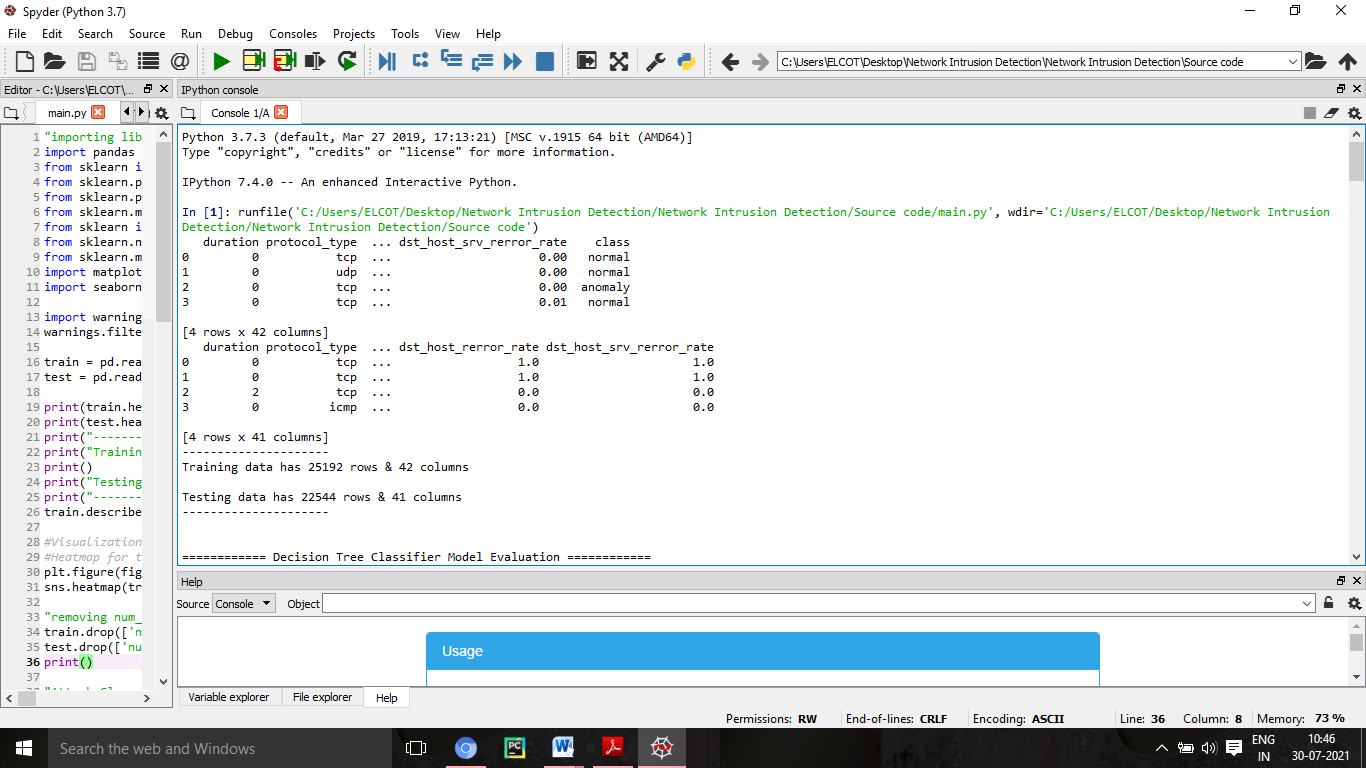
**Figure 7.2 Predefined Test Data**

**41**

****

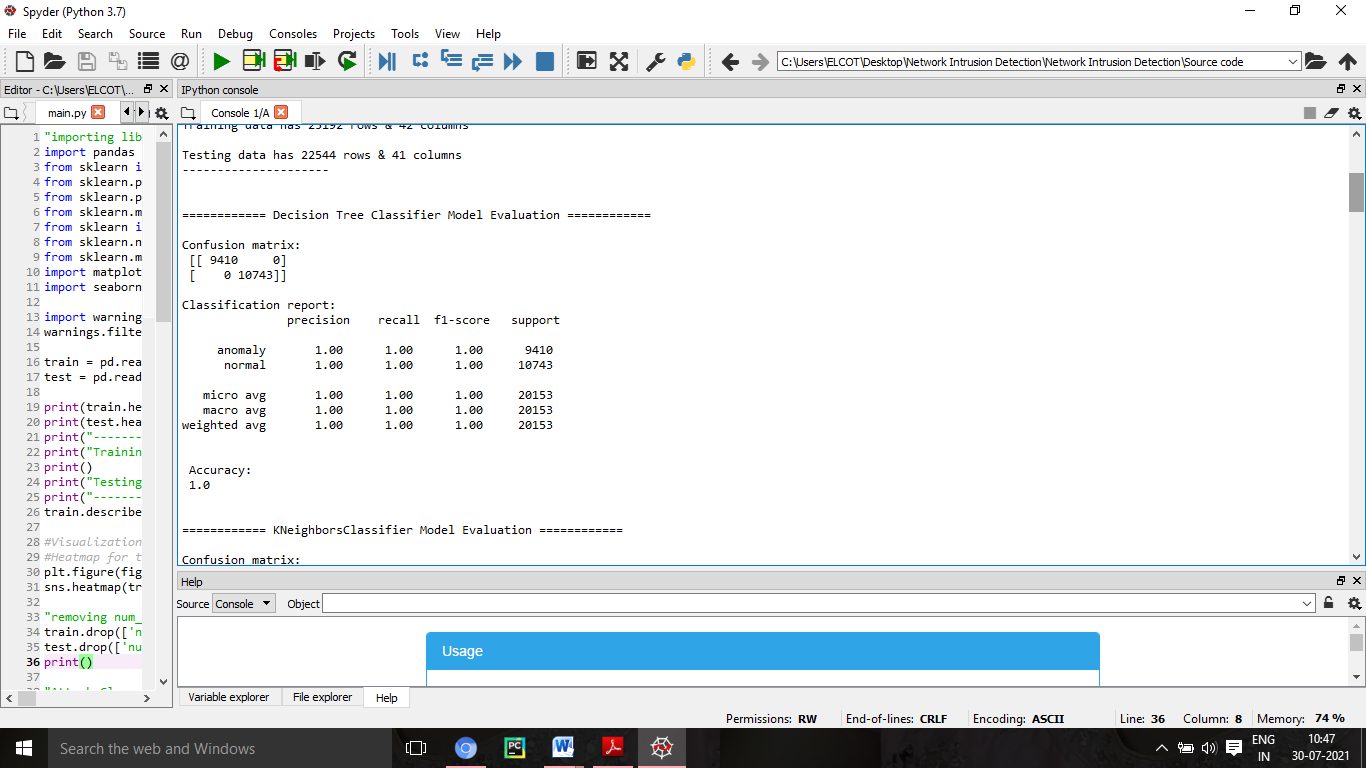
**Figure 7.3 Predefined Train Data**

**42**

****

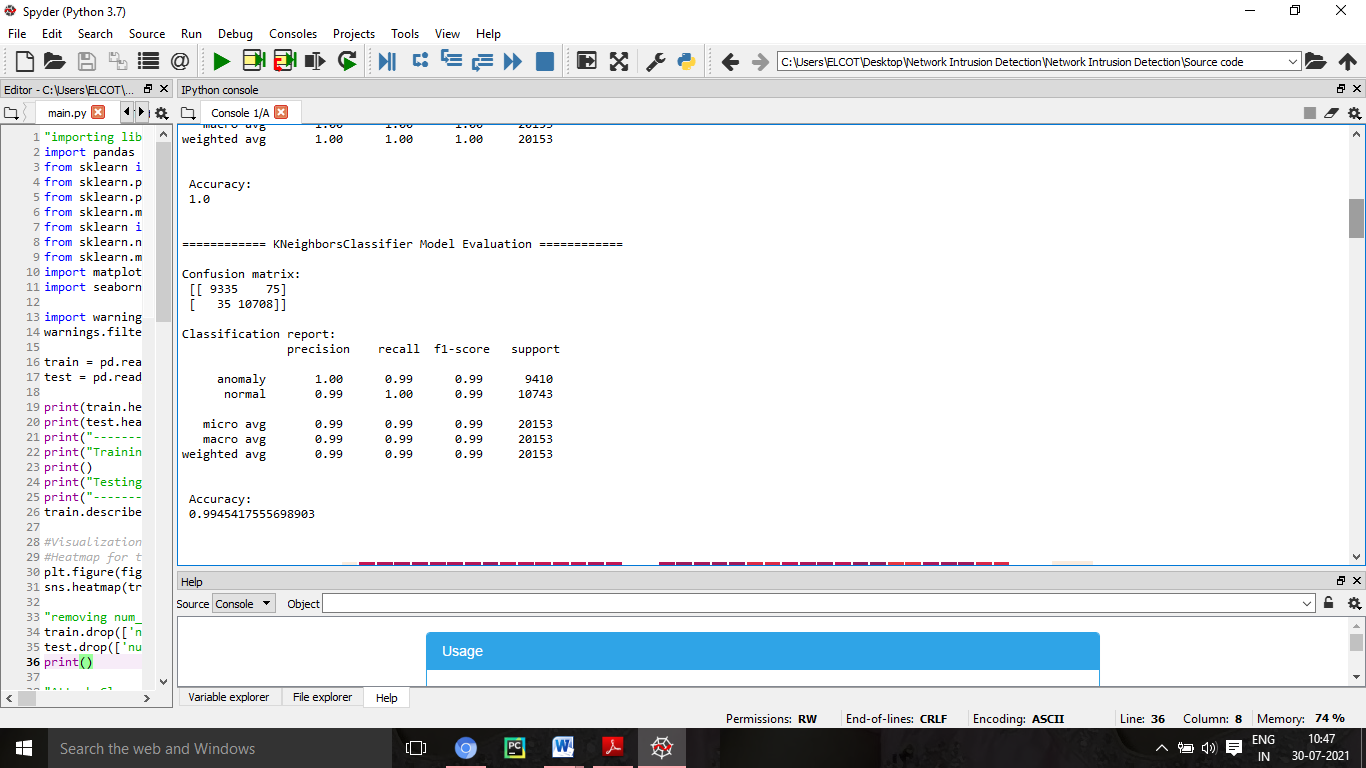
**Figure 7.4 Training and Testing Data Rows and Columns**

**43**

****

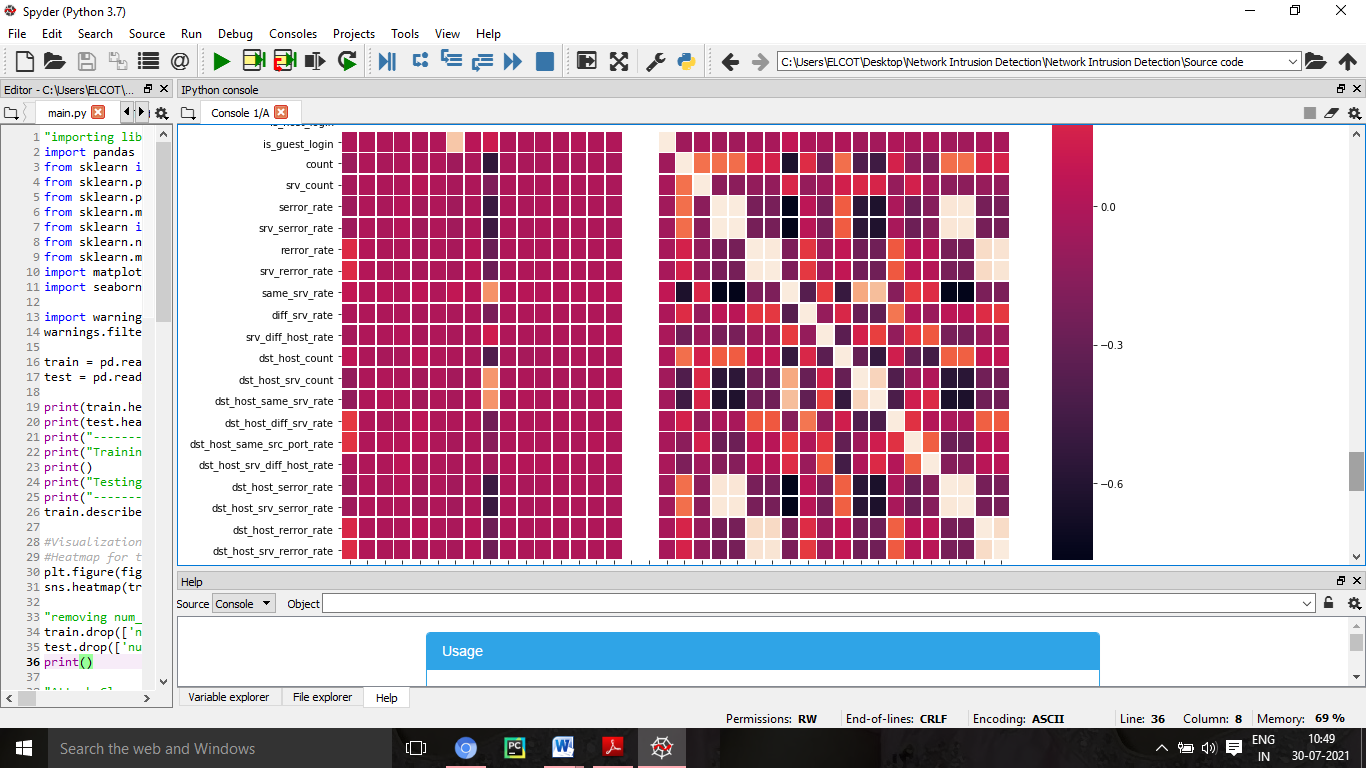
**Figure 7.5 Decision tree Classifier Model Evaluation**

**44**

****

**Figure 7.6 KNN Classifier Evaluation**

**45**

****

**Figure 7.7 Heat Map**

**46**

**CHAPTER 8**

**CONCLUSION AND FUTURE WORK**

**47**

**CHAPTER 8**

**CONCLUSION**

We reviewed several inﬂuential algorithms for intrusion detection based on various machine learning techniques. Characteristics of ML techniques makes it possible to design IDS that have high detection rates and low false positive rates while the system quickly adapts itself to changing malicious behaviours. We divided these algorithms into two types of ML-based schemes: Artiﬁcial Intelligence (AI) and Computational Intelligence (CI). Although these two categories of algorithms share many similarities, several features of CI-based techniques, such as adaptation, fault tolerance, high computational speed and error resilience in the face of noisy information, conform the requirement of building eﬃcient intrusion detection systems.

**FUTURE WORK**

In future, it is possible to provide extensions or modifications to the proposed clustering and classification algorithms using intelligent agents to achieve further increased performance. Apart from the experimented combination of data mining techniques, further combinations such as artificial intelligence, soft computing and other clustering algorithms can be used to improve the detection accuracy and to reduce the rate of false negative alarm and false positive alarm. Finally, the intrusion detection system can be extended as an intrusion prevention system to enhance the performance of the system.

**48**

**REFERENCES**

1. S. Yinbiao and K. Lee, “Internet of Things: Wireless Sensor Networks Executive summary,” 2014. [5] F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci
2. “A survey on sensor networks,” IEEE Commun. Mag., vol. 40, no. 8, pp. 102–105, 2002.
3. X. Chen, K. Makki, K. Yen, and N. Pissinou, “Sensor network security: a survey,” IEEE Commun. Surv. Tutorials, vol. 11, no. 2, pp. 52–73, 2009.
4. A.-S. K. Pathan, H.-W. Lee, and C. S. Hong, “Security in wireless sensor networks: issues and challenges,” 2006 8th Int. Conf. Adv. Commun. Technol., vol. 2, p. 6 pp.-pp.1048, 2006.
5. P. Yi, Y. Jiang, Y. Zhong, and S. Zhang, “Distributed Intrusion Detection for Mobile Ad Hoc Networks,” 2005 Symp. Appl. Internet Work. (SAINT 2005 Work., pp. 94–97, 2005.
6. H. Sedjelmaci and M. Feham, “Novel Hybrid Intrusion Detection System for Clustered Wireless Sensor Network,” Int. J. Netw. Secur. Its Appl. (IJNSA), Vol.3, No.4, July 2011, vol. 3, no. 4, pp. 1–14, 2011.
7. L. Khan, M. Awad, and B. Thuraisingham, “A new intrusion detection system using support vector machines and hierarchical clustering,” VLDB J., vol. 16, no. 4, pp. 507–521, 2007.
8. S. K. Sahu, S. Sarangi, and S. K. Jena, “A detail analysis on intrusion detection datasets,” Souvenir 2014 IEEE Int. Adv. Comput. Conf. IACC 2014, pp. 1348–1353, 2014.

**49**

1. O. Can, C. Turguner, and O. K. Sahingoz, “A Neural Network Based Intrusion Detection System For Wireless Sensor Networks,” Signal Process. Commun. Appl. Conf. (SIU), 2015 23th, pp. 2302–2305, 2015.
2. F. Lu and L. Wang, “Intrusion Detection System Based on Integration of Neural Network for Wireless Sensor Network,” J. Softw. Eng. 2014.
3. Y. Y. Li and L. E. Parker, “Intruder detection using a wireless sensor network with an intelligent mobile robot response,” Southeastcon, 2008. IEEE, pp. 37–42, 2008. [15] A. Kulakov and D. Davcev, “Tracking of unusual events in wireless sensor networks based on artificial neural-networks algorithms,” Inf. Technol. Coding Comput. 2005. ITCC 2005. Int. Conf., pp. 534–539, 2005.
4. M. Panda, “Security Threats at Each Layer of Wireless Sensor Networks,” Int. J. Adv. Res. Comput.Sci. Softw. Eng., vol. 3, no. 11, pp. 61–67, 2013.
5. Karlof and D. Wagner, “Secure routing in wireless sensor networks: attacks and countermeasures,” Proc. First IEEE Int. Work. Sens. Netw. Protoc. Appl. 2003., pp. 113–127, 2003.

**50**

**CHAPTER 9**

**APPENDIX**

**51**

"importing libraries"

import pandas as pd

from sklearn import metrics

from sklearn.preprocessing import StandardScaler

from sklearn.preprocessing import LabelEncoder

from sklearn.model\_selection import train\_test\_split

from sklearn import tree

from sklearn.neighbors import KNeighborsClassifier

from sklearn.model\_selection import cross\_val\_score

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

warnings.filterwarnings('ignore')

train = pd.read\_csv("Train\_data.csv")

test = pd.read\_csv("Test\_data.csv")

print(train.head(4))

print(test.head(4))

**52**

print("---------------------")

print("Training data has {} rows & {} columns".format(train.shape[0],train.shape[1]))

print()

print("Testing data has {} rows & {} columns".format(test.shape[0],test.shape[1]))

print("---------------------")

train.describe()

#Visualization

#Heatmap for train data

plt.figure(figsize=(15, 15))

sns.heatmap(train.corr(), linewidths=.8)

"removing num\_outbound\_cmds column from datasets"

train.drop(['num\_outbound\_cmds'], axis=1, inplace=True)

test.drop(['num\_outbound\_cmds'], axis=1, inplace=True)

print()

"Attack Class Distribution count"

train['class'].value\_counts()

"Standard scaler function"

**53**

scaler = StandardScaler()

"extract numerical attributes and scale it to have zero mean and unit variance "

cols = train.select\_dtypes(include=['float64','int64']).columns

sc\_train = scaler.fit\_transform(train.select\_dtypes(include=['float64','int64']))

sc\_test = scaler.fit\_transform(test.select\_dtypes(include=['float64','int64']))

"turn the result back to a dataframe"

sc\_traindf = pd.DataFrame(sc\_train, columns = cols)

sc\_testdf = pd.DataFrame(sc\_test, columns = cols)

"Label encoding"

encoder = LabelEncoder()

"extract categorical attributes from both training and test sets "

cattrain = train.select\_dtypes(include=['object']).copy()

cattest = test.select\_dtypes(include=['object']).copy()

"encode the categorical attributes"

traincat = cattrain.apply(encoder.fit\_transform)

testcat = cattest.apply(encoder.fit\_transform)

"separate target column from encoded data "

enctrain = traincat.drop(['class'], axis=1)

**54**

cat\_Ytrain = traincat[['class']].copy()

train\_x = pd.concat([sc\_traindf,enctrain],axis=1)

train\_y = train['class']

train\_x.shape

test\_df = pd.concat([sc\_testdf,testcat],axis=1)

test\_df.shape

"Training and testing sets"

X\_train,X\_test,Y\_train,Y\_test = train\_test\_split(train\_x,train\_y,train\_size=0.80, random\_state=2)

" KNeighborsClassifier Model"

KNN\_Classifier = KNeighborsClassifier(n\_jobs=-1)

KNN\_Classifier.fit(X\_train, Y\_train);

" Decision Tree Model"

DTC\_Classifier = tree.DecisionTreeClassifier(criterion='entropy', random\_state=0)

DTC\_Classifier.fit(X\_train, Y\_train)

models = []

**55**

models.append(('Decision Tree Classifier', DTC\_Classifier))

models.append(('KNeighborsClassifier', KNN\_Classifier))

for i, v in models:

scores = cross\_val\_score(v, X\_train, Y\_train, cv=10)

accuracy = metrics.accuracy\_score(Y\_train, v.predict(X\_train))

confusion\_matrix = metrics.confusion\_matrix(Y\_train, v.predict(X\_train))

classification = metrics.classification\_report(Y\_train, v.predict(X\_train))

print()

print('============ {} Model Evaluation ============'.format(i))

print()

print("Confusion matrix:" "\n", confusion\_matrix)

print()

print("Classification report:" "\n", classification)

print()

print (" Accuracy:" "\n", accuracy)

print()

**56**