**OPTIMIZATION NETWORK TRAFFIC SECURITY**

**K. Sai Kumar1,** **D. Raju2, K. Prudhvi3**

1-4Department Of Cyber Security, Malla Reddy University, Hyderabad, Telangana, India.

5Assistant Professor, Department of Cyber Security, Malla Reddy University, Hyderabad, Telangana, India.

[**2111cs040088@mallareddyuniversity.ac.in**](mailto:2111cs040088@mallareddyuniversity.ac.in)**1,**[**2111cs040078@mallareddyuniversity.ac.in**](mailto:2111cs040078@mallareddyuniversity.ac.in)**2,** [**2111cs040076@mallareddyuniversity.ac.in**](mailto:2111cs040076@mallareddyuniversity.ac.in)**3**

Abstract: In the world of networking, it sometimes becomes essential to know what types of applications flow through the network for performance of certain tasks. Network traffic classification sees its main usage among ISP’s to analyze the characteristics required to design the network and hence affects the overall performance of a network. There are various techniques adopted to classify network protocols, such as port-based, pay-load based and Machine Learning based, all of them have their own pros and cons. Prominent nowadays is Machine Learning technique due to its vastness in usage in other fields and growing knowledge among researchers of its better accuracy among others when compared. In this paper, we compare two of the basic algorithms, Naïve Bayes and K nearest algorithm results when employed to networking data set extracted from live video feed using Wireshark software. For an implementation of Machine learning algorithm, python sklearn library is used with numpy and pandas library used as helper libraries. Finally, we observe that K nearest algorithm gives more accurate prediction than Naïve Bayes Algorithm, Decision Tree Algorithm and Support Vector Machine.

Keywords—K-Nearest Neighbors (KNN), Naïve Bayes, Decision Tree, Support Vector Machine (SVM)

Introduction: The Network Traffic Classification plays a dynamic role pertaining to threats associated with the emerging technology nowadays. The various machine learning based classification techniques are presented in [1]. It helps internet service providers to manage overall performance of the network by considering the factors associated with certain application protocol. It has its usage in recognizing the unknown network if any tries to intrude the specified traffic lane. By this way we get to study its properties as well. Using the above property of recognizing unknown network, one can also recognize the potential threats that a network can suffer due to certain security attacks. Management of network security and Quality of Service (QoS) is also essential task and can be achieved if we have good techniques to classify network. Blocking or allowance of certain network traffic can also be achieved if we classify our network well. Overall, classification of network helps in overall growth of the network and its efficiency.

EXISTING PROBLEM**:**

Various Network Classification Techniques.

1. Port Based–All the ports are registered with IANA, by hashing the application protocol with the ports registered from IANA, one can easily identify the traffic in the network. For instance, the standardized port numbers assigned for sending and receiving E-mail are 25(SMTP) and 110(POP3), respectively. Now, these port numbers are universal standard for all the networks all over the globe. There are certain drawbacks associated with this technique. The major drawback felt in this type of classification is detection of only wellknown port numbers. Another drawback is in cases of unregistered or dynamic port numbers that are not mentioned by IANA. Below is the table that illustrates the currently registered port numbers with IANA

Payload Based Technique -This technique is based on Payload to classify network traffic. As the name suggests, this technique majorly uses “payload” i.e the distinguishable contents of data packets of different applications. Due to the tendency to “inspect” the protocol, it is also referred as Deep Packet Inspection (DPI) technique. This technique is usually employed in the P2P (Peer to Peer) networks, which generally use dynamic port numbers.

2)Machine Learning Techniques - Machine learning has been used to classify various forms of data. In this project, this technique is employed to classify the application protocols. According to this technique, the machine (learning classifier) is trained, by given certain collection of data, in order to achieve maximum success rate, usually a large amount of data is employed for better training. After the training process, certain sample of data class is supplied to the classifier, so as to check how well the machine is trained, by evaluating the output of the machine and comparing them to original output. There are two types of machine learning techniques: Unsupervised and Supervised Machine Learning. i)Unsupervised Technique: In this type of technique, A raw dataset is provided to the learning classifier, by raw we mean the data without any predefined labels or tags. This method of machine learning is also known as clustering. This technique divided the dataset into set of clusters with each data entry belonging to a specific cluster and hence can be used to further predict cluster to which future data entry would correspond to. But we cannot use this technique for network classification as it cannot cluster according to predefined class variables.

* 1. Ii)Supervised Technique Yet another classification technique and a useful part of machine learning is Supervised Technique. Presence of a well labeled complete data set is requirement if one wishes to employ any of supervised learning method. The working process of this method is a two steps procedure, first the data sample is trained using the specified labels, and then it is then employed to test in a new data sample [2]. Hence an useful technique is used to classify the network traffic data.
  2. Proposed solution:Naïve Bayes Algorithm – Working in relation with the Bayes theorem, his algorithm mainly works on the assumption that all the input features put to training are independent of each other i.e while calculating the whole, each feature of class is taken as separate entity and hence computed individually. This theorem is particularly useful in the cases where the data set is large and also outperforms several other Machine Learning algorithms.
  3. The advantage of using this technique is its working of fast and easier to predict the class of training data set; moreover it works better for multiclass predictions. When most of the features are independent of each other, the classification method performs better than many other regression models. It also handles categorical inputs well enough so as not to form a disadvantage if more categorical inputs are there.
  4. The drawbacks are explained as follows. If any input in test data test is absent in training dataset, then probability will be 0 and hence this algorithm would be unable to make any predictions and would fail in this particular scenario. Another disadvantage of this algorithm is that in practical real life situations, one feature is never fully independent of each other and hence prediction fails practically.
  5. K – Nearest Neighbors– This algorithm works on the simple fact and process of storing all the available cases and predicting new case depending upon the chosen parameter,(mostly distance). K-NN technique is traditional in terms of application for the people in 1970‟s as for statistical measurements and pattern recognition. Major Drawback faced by this algorithm is when the data types of features is both numerical and categorical, as calculation of distance as parameter to judge doesn‟t give good result in such cases. The cross validation technique is useful in accessing performance.
  6. Decision Tree Algorithm–Decision Tree algorithm works on the principal of “learning” in which the model is created that can then predict the class target variables value by inferring to the decision rules employed and decided during the training data stage. Its ease of explanation as it makes the decisions just like humans makes it easier to understand and hence an advantage. But sometimes, calculations can get complex if it has many class labels. There are several algorithms that can be used to construct a decision tree. Some of them are ID3, C4.5, CART (Classification of Regression Tree), CHAID and MARS. We employ the CART technique to perform the task of prediction.
  7. Support Vector Machine –The SVMs are the collection of some of the related supervised learning methods that can be employed for classification and regression. They are a part of generalized linear classification family. Simultaneously minimizing the empirical classification and maximizing the geometry margin is the special property of SVM. So SVM is also called Maximum Margin Classifiers. SVM is alias of the Structural risk Minimization (SRM). SVM links the input vector to a higher dimensional plane thus a maximal separating hyper plane is constructed. The construction of two parallel hyper planes on either side of the data hyper plane occurs. The maximization of distance between 2 parallel hyper planes is achieved by separating hyper planes.

Literature Review: The term network security is widespread and is used to describe the actions the administrator takes to avoid network violations or illegal access to information by hackers[1, 2]. Today’s online criminals are highly skilled at attacking enterprises. Moreover, it is difficult for businesses to tell the difference between legitimate and fraudulent activity [3]. Meanwhile, flow analysis is based on the identification of anonymity networks [4][5]. Meanwhile, other ML algorithms are build models of behaviors and use those models as a basis for making future predictions based on new input data [6]. Machine learning can aid in solving the most common tasks including regression, prediction, and classification in the era of extremely large amount of data and cybersecurity talent shortage [7]. Moreover, clustering and classification extract patterns out of data packets which can be used in many applications such as security analysis and user profiling [8]. Furthermore, there are many applications for analyzing traffic based on the ML algorithms such as identifying anomalies through discovery-based workbooks or features that describe user behaviour [9]. The NTA tools have been commonly used to analyze and identify network security and performance issues[10]. Traffic analysis is important in network management and operations for several purposes[11]. Attackers are always inventing new methods and modifying methods to avoid detection by the network defence systems. Traffic analysis products have emerged in response to ongoing updates that provide ways to combat these attackers [12]. Wireless Sensor Networks WSN: It is a technology that can be used in large systems such as commercial applications, where security is vital for their applicability. Classify attacks in wireless sensor networks to explore patterns and possible countermeasures is widely used to processing and online processing [15].

Existing system:

Feasibility Study: The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of themajor requirements for the system is essential.

Economical Feasibility**:** This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies usedare freely available. Only the customized products had to be purchased.

Technical Feasibility:This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modestrequirement, as only minimal or null changes are required for implementing this system.

Social Feasibility: The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it.

ALGORITHMS:

1)SUPPORT VECTOR MACHINE

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot).Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line).More formally, a support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite- dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space.

2)NAÏVE BAYES ALGORITHM: Working in relation with the Bayes theorem, his algorithm mainly works on the assumption that all the input features put to training are independent of each other i.e while calculating the whole, each feature of class is taken as separate entity and hence computed individually. This theorem is particularly useful in the cases where the data set is large and also outperforms several other Machine Learning algorithms. The detailed working of Naïve Bayes algorithm is explained in [8] is as follows. Let D be the training set attribute with n tuples where each tuple has „k‟ attribute values and represented by vector V, where V={V1 ,V2 ,…Vk }. Let there be „k‟ classes C1, C2,…Ck. According to the Naive Bayesian classifier, an input tuple I belongs to class Cx only if it has higher posterior probability than any of the classes Cy among C1, C2,…Ck, where x ≠ y.

The advantage of using this technique is its working of fast and easier to predict the class of training data set; moreover it works better for multiclass predictions. When most of the features are independent of each other, the classification method performs better than many other regression models. It also handles categorical inputs well enough so as not to form a disadvantage if more categorical inputs are there.

3)K-NEAREST NEIGHBORS:This algorithm works on the simple fact and process of storing all the available cases and predicting new case depending upon the chosen parameter,(mostly distance). K-NN technique is traditional in terms of application for the people in 1970‟s as for statistical measurements and pattern recognition. Major Drawback faced by this algorithm is when the data types of features is both numerical and categorical, as calculation of distance as parameter to judge doesn‟t give good result in such cases. The cross validation technique is useful in accessing performance of Machine Learning Models. It is basically testing the data set for outputs during the training phase itself. K-fold Cross Validation is the technique which is basically used for Machine Learning. The steps followed in this technique are:

First is to divide the initial set into k equally divided subsets. Let these subsets be called the folds and assigning f1, f2, …,fk to them.

Loop for j = 1 to j = k

Let the foldfj be the Validation set and let all the remaining k-1 sets to be cross validating training sets. The training of Machine Learning Model would be done by this cross validation training set and the computation of accuracy is done by comparing the validation results with the actual values in the validation set.

The final accuracy of Machine Learning Model can be predicted by taking the average of accuracies derived in all cases of k cross validation cases. The positive and negatives aspects of the technique are presented as follows. This method works well when training dataset is large and also gives comparatively better results when data set is of noisy nature. The major challenge in this algorithm is to predict the value of K that would give the best results. One more minus to use this technique is that computational cost is high as distance has to be computed for each entry in data set. The ambiguity associated with “distance” being the parameter for determining results also adds to its disadvantages.

4)DECISION TREE ALGORITHM: Decision Tree algorithm works on the principal of “learning” in which the model is created that can then predict the class target variables value by inferring to the decision rules employed and decided during the training data stage. Its ease of explanation as it makes the decisions just like humans makes it easier to understand and hence an advantage. But sometimes, calculations can get complex if it has many class labels. There are several algorithms that can be used to construct a decision tree. Some of them are ID3, C4.5, CART (Classification of Regression Tree), CHAID and MARS. We employ the CART technique to perform the task of prediction. One of the well known advantages of Decision Tree algorithm is its transparent nature. It means giving the results having considered the best possible way to get the best results. Its specific nature and assigning specific value to the problem adds to its big plus. Its comprehensive nature allows analysis of each and every possible decision. This tree is also user friendly as it can better be illustrated. It also handles the mixture of categorical data and numerical data well and gives a good result.

This algorithm becomes unstable with a very little change in the data and hence sometimes lead to poorer results. If a tree is complicated, then testing with different set of data sets can give poorer results. Over fittings observed when the algorithm catches noise in the dataset.

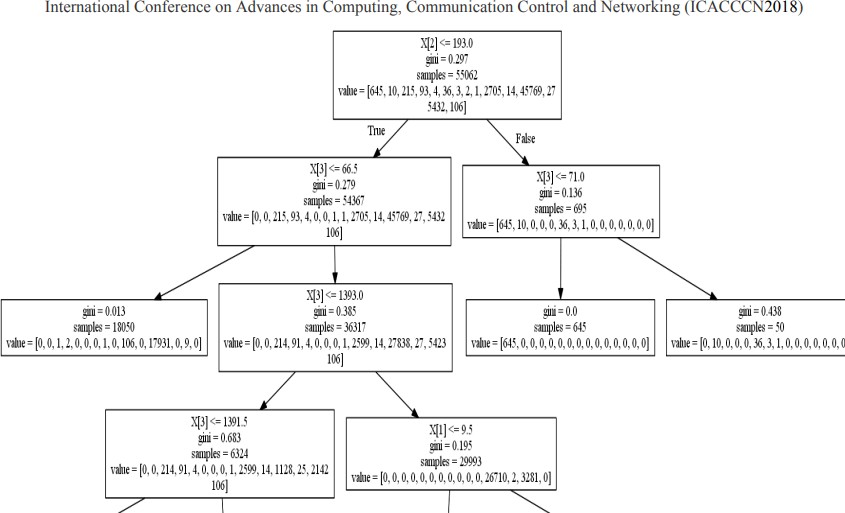


Fig.1.The constructed decision tree from the training data set.

ARCHITECTURE

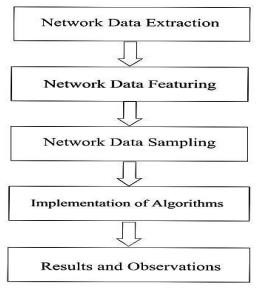


Fig 2. System Architecture

This section deals with description of entire process of Network Traffic Classification; we display step-step approach to classify unknown network data into certain protocol. The Network classification we used is of machine learning technique. Fig. 1 below explains the process of classification of Network Traffic.

Network Data Extraction:To start working with the algorithms, we first need to get the data on which we need to employ it. This is base of every algorithm employment and known as data collection step. The real time network data traffic capturing is carried out in subsection. Numerous tools are employed for capturing network data, popular ones being Wireshark and tcpdump tool [10]. We use Wire Shark tool for capturing the network packet which is further analyzed [2]. We analyze the real time data of Internet surfing for duration of 1 hour and produce around 85000 entries.

Network Data Featuring: Wireshark provides an easy method to convert data to csv format for analysis further. Anyhow, identifying the correct attribute required to classify the protocol is a challenging task that can be achieved only through extensive study of the topic.We have identified four features that can help in accurate prediction of the protocol and classify network traffic well. We are working to extract 19 more features. These features are Source IP address, Destination IP address, interval between packets, and length of packet. Target Variable is Protocol.

Network Data Sampling: This process is carried to sample the data. Sampling the data in the supervised technique refers to label the data to achieve the purpose of classifying the unknown network classification. The table below depicts a small part of the sampled data set Employed for making predictions.

DATA FLOW DIAGRAM:

1)The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.

2)The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.

3)DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.

4)DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail

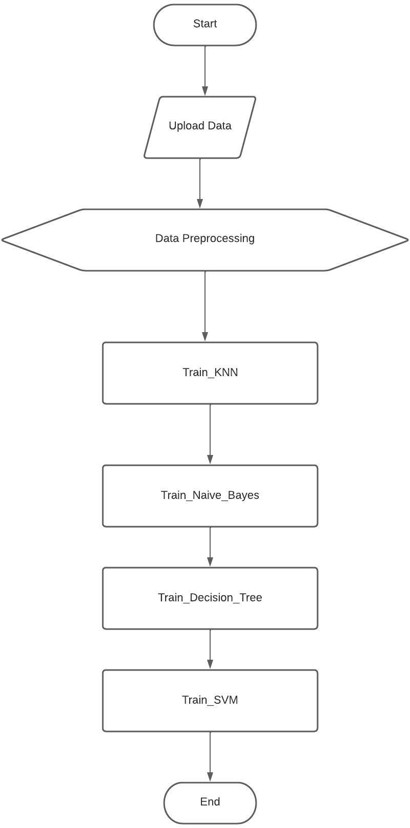


Fig 3. Data Flow Diagrams

CLASS DIAGRAM:

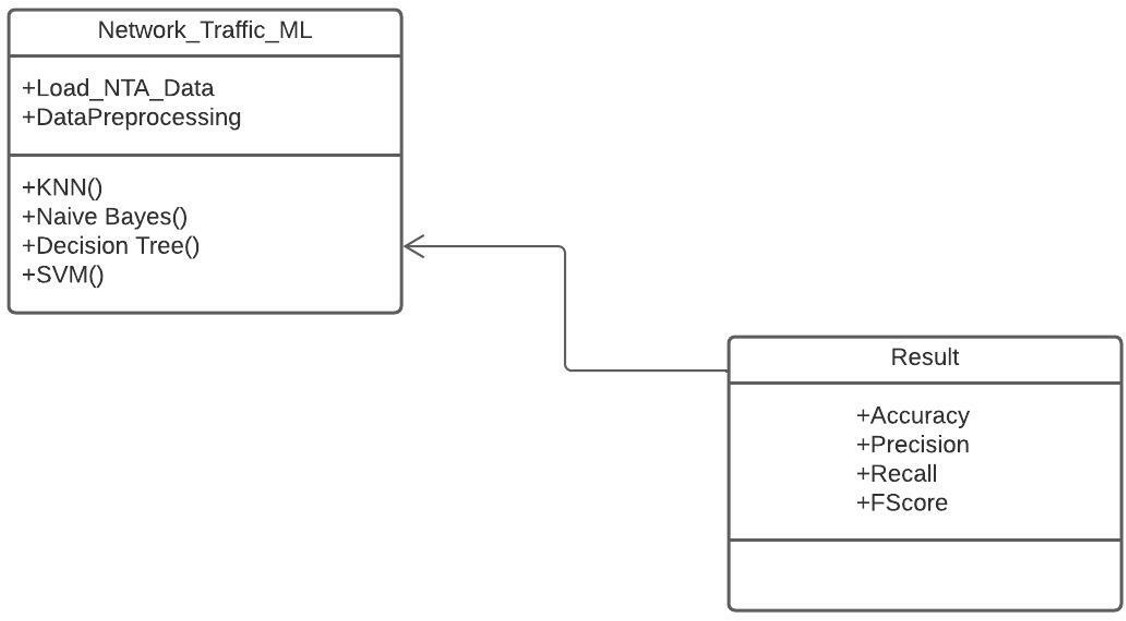


Fig 4.Class Diagram

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

USE CASE DIAGRAM: A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

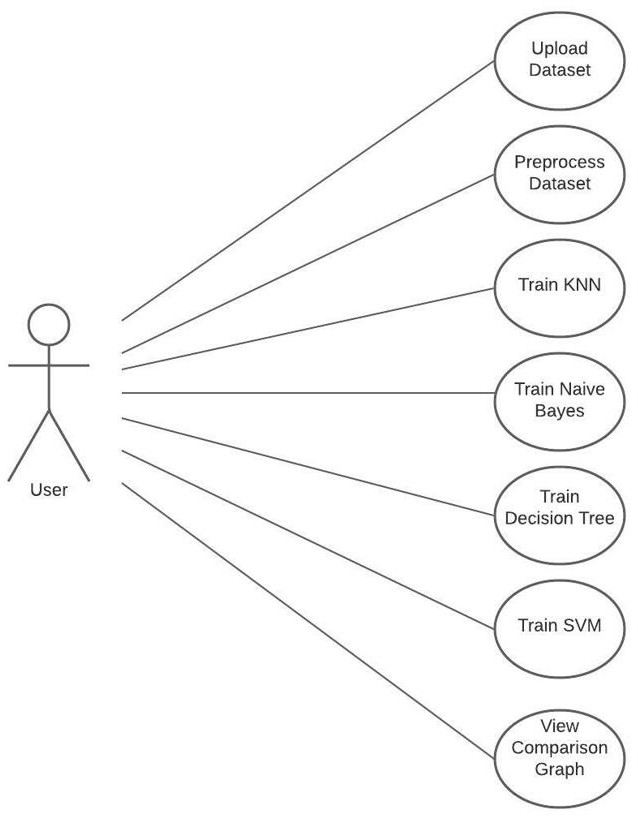


Fig 5. Use Case Diagram

ACTIVITY DIAGRAM: Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

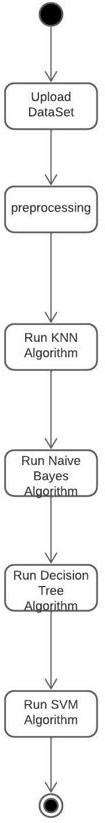


Fig .6 Activity Diagram

TESTING: The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirementsand user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

Functional test: Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effectivevalue of current tests is determined.

System Test:

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing:

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose.It is purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing:

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the softwareunder test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing:

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Integration Testing:

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects. The task of the integration test is to check that components or software applications, e.g. components in a software system level.

Acceptance Testing:

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results**:** All the test cases mentioned above passed successfully. No defects encountered.

RESULT:

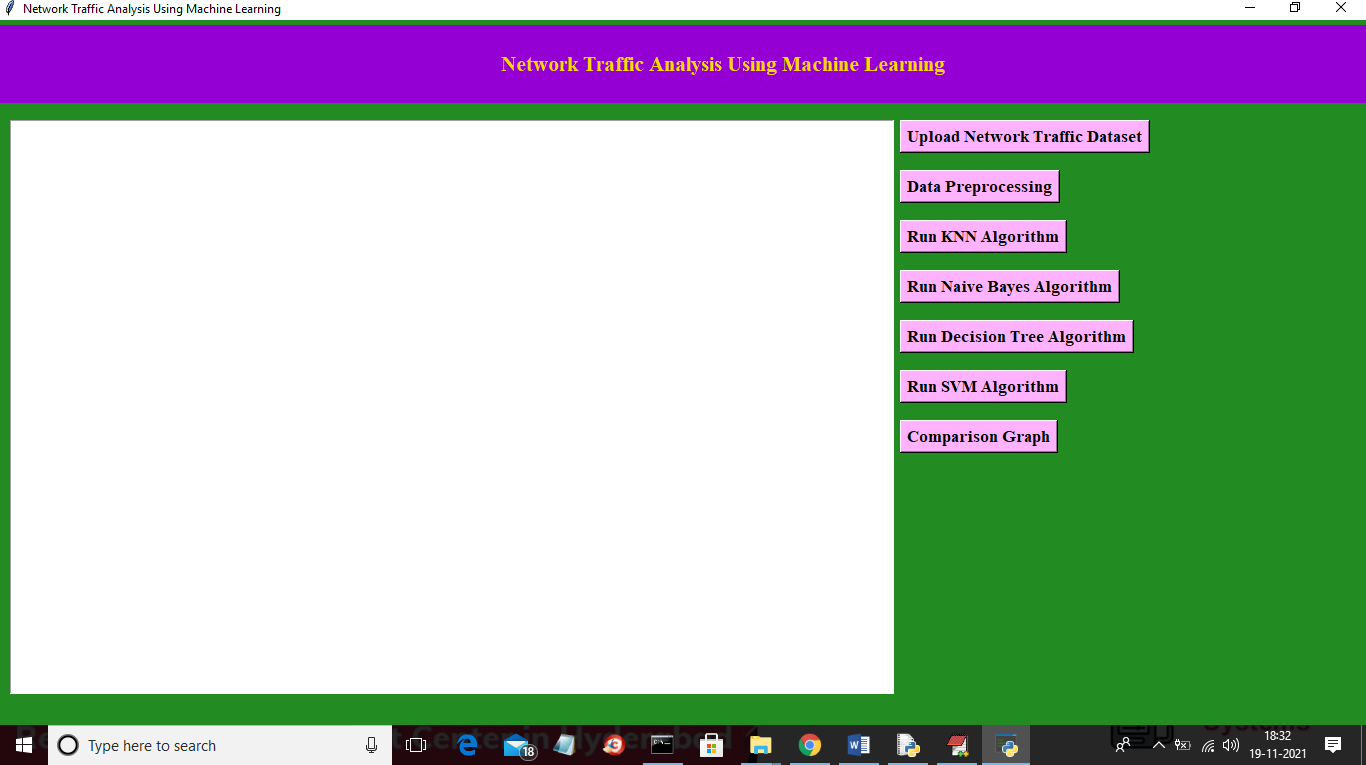


Fig.7 In above screen click on ‘upload Network Traffic Dataset’ button to upload dataset

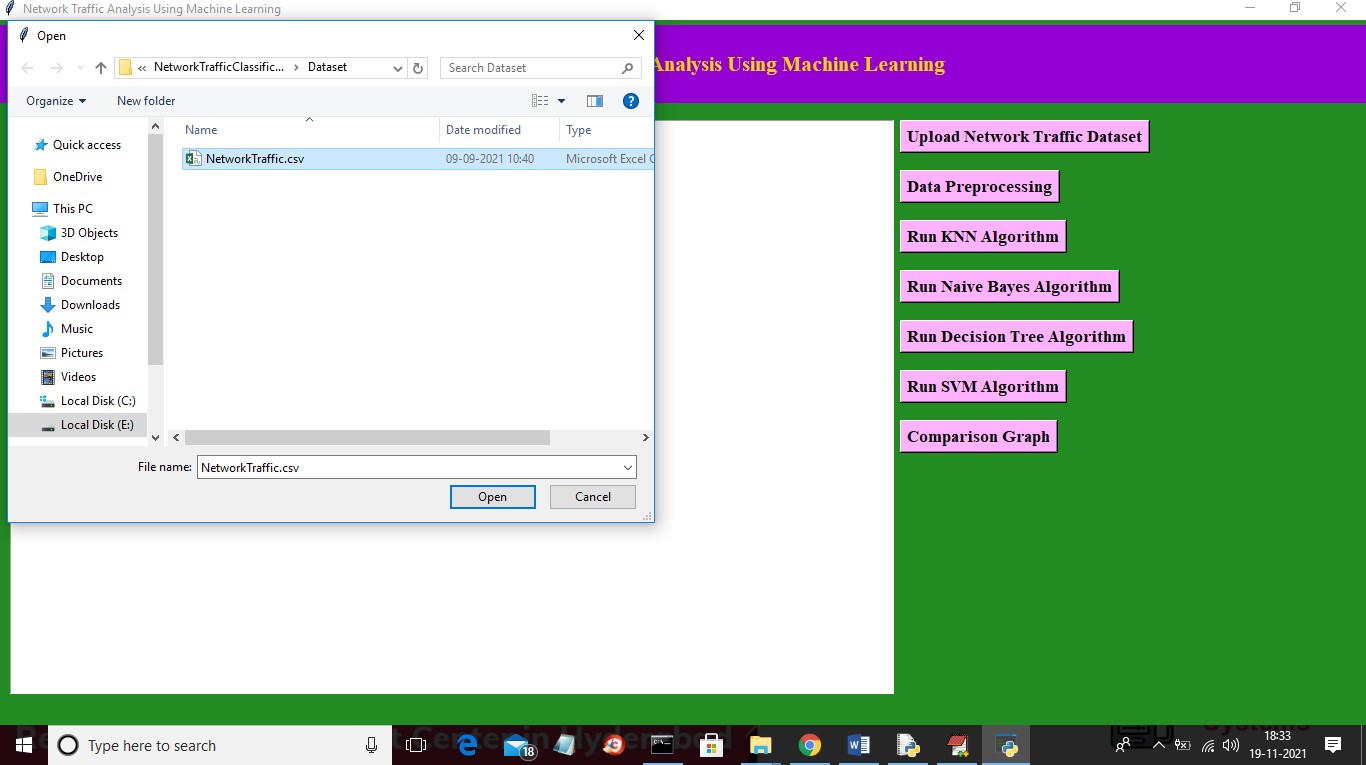


Fig.8 In above screen selecting and uploading ‘NetworkTraffic.csv’ file and then click on ‘Open’ button to load dataset and to get below screen

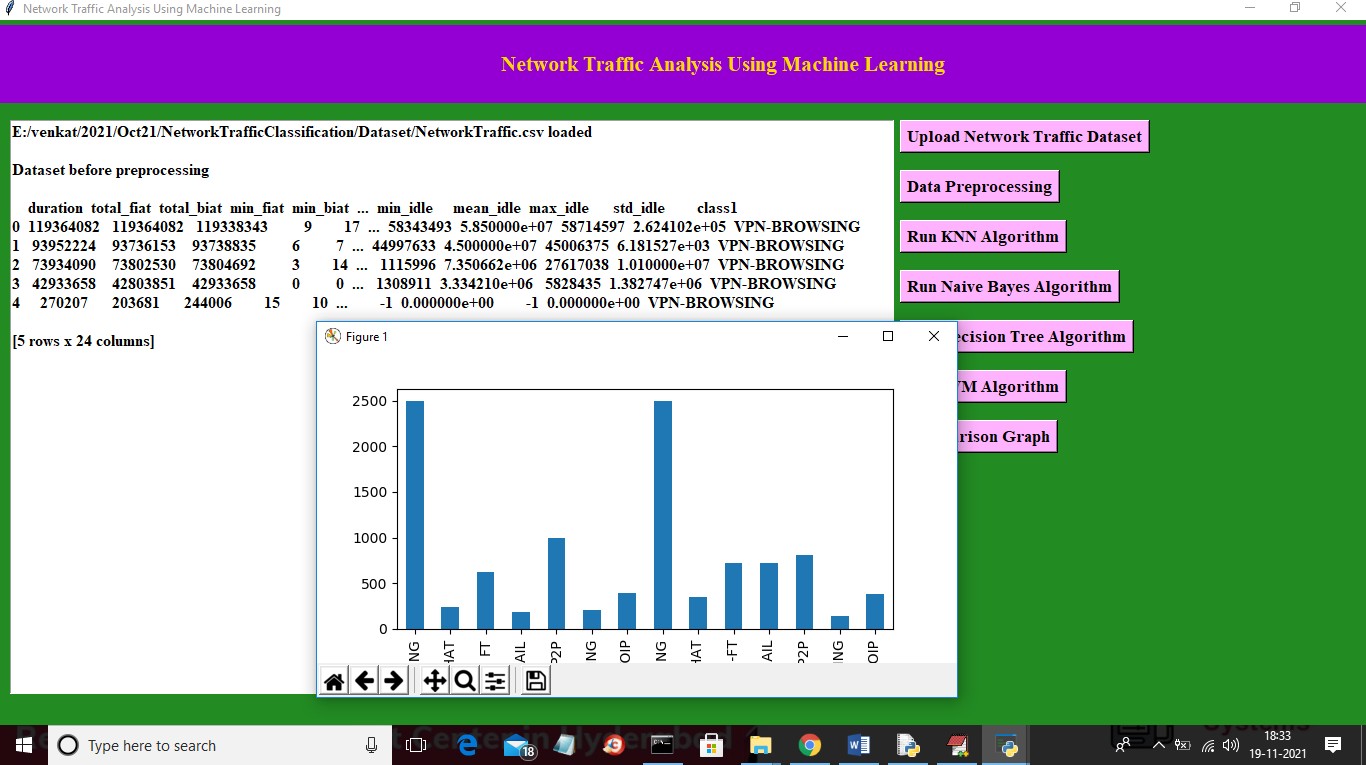


Fig.9 In above screen we can see dataset loaded and dataset contains lots of non-numeric values so we need to process it and in graph x-axis we can see traffic type and y-axis represents total records in dataset for that traffic. Now close above graph and then click on ‘Data Preprocessing’ to clean dataset.

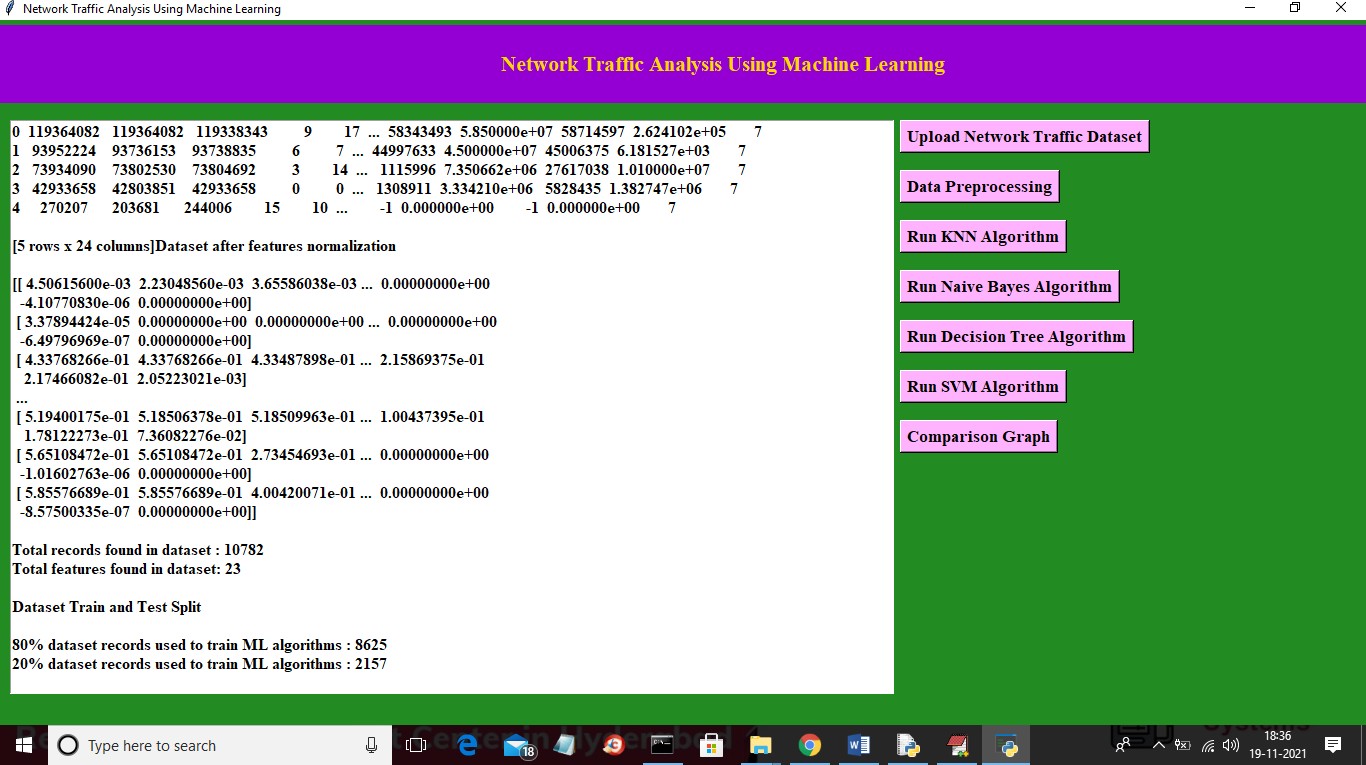


Fig.10.In above screen we can see all dataset converted to numeric format and in last we can see total records and columns found in dataset and then splitted dataset percentage for train and test records. Now train and test data is ready and now click on ‘Run KNN Algorithm’ button to train KNN and get below result

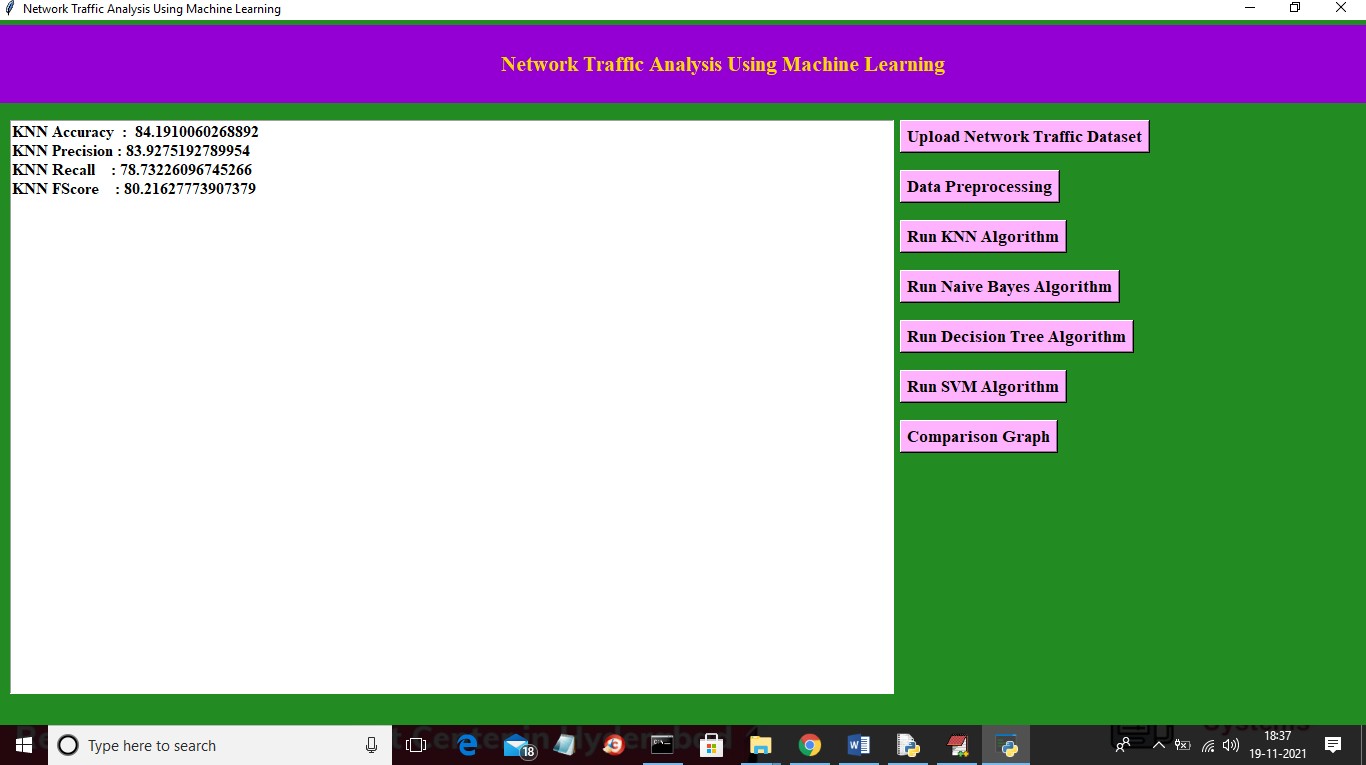


Fig.11.In above screen with KNN we got 84% accuracy and now click on ‘Run Naïve Bayes Algorithm’ button to train it

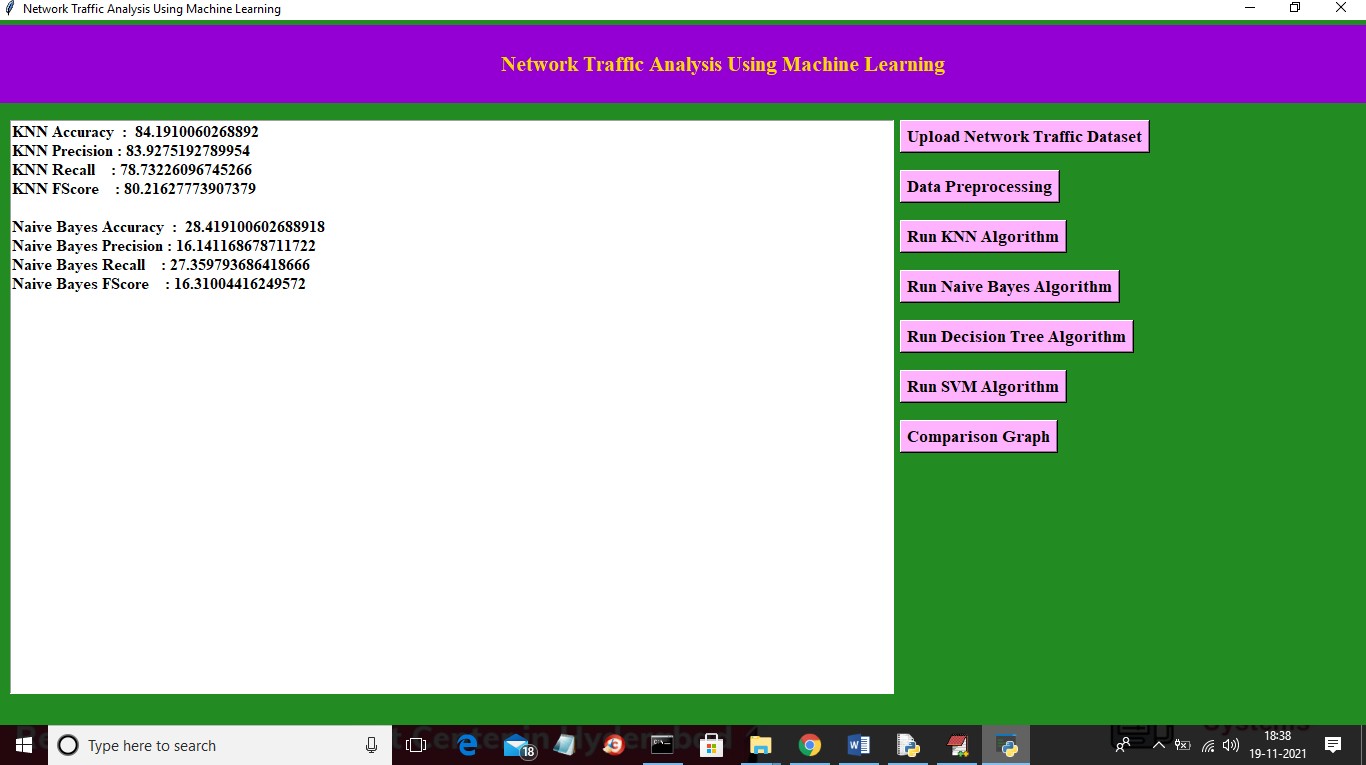
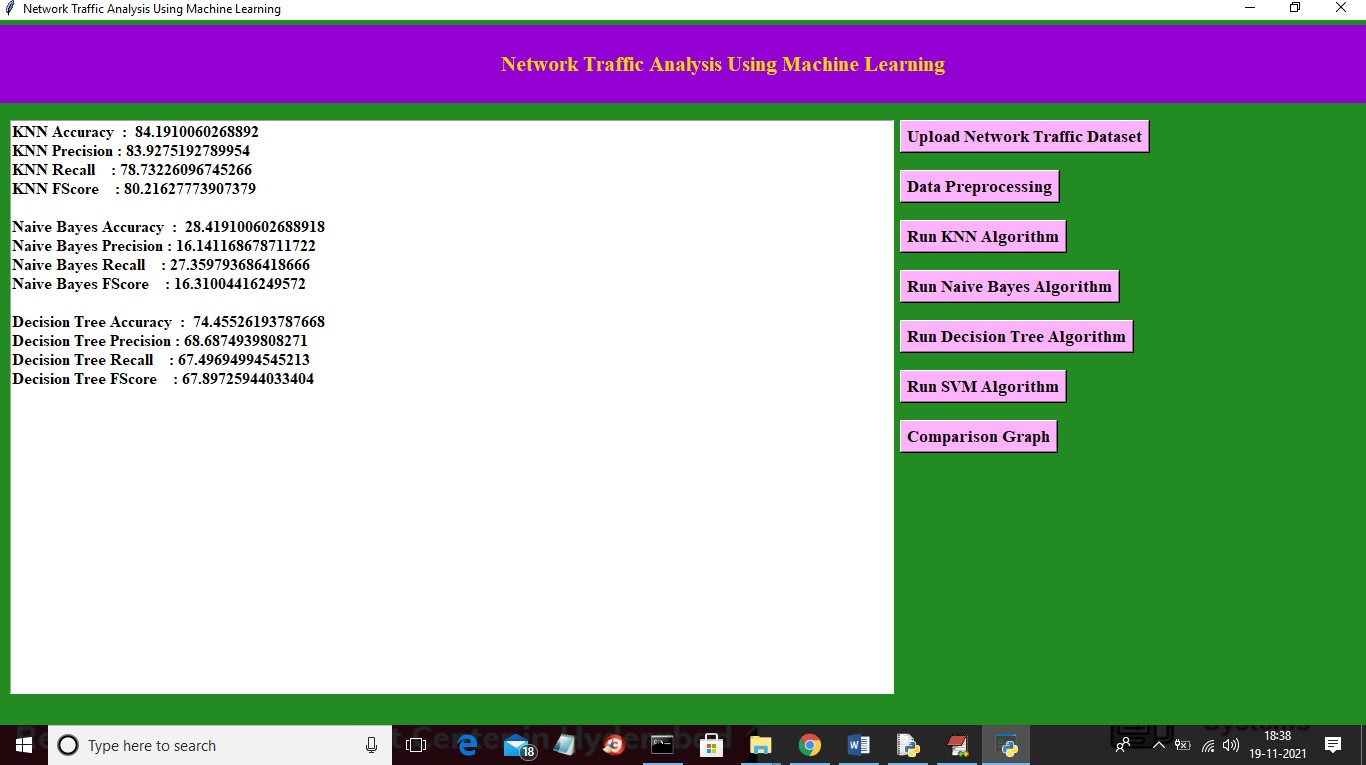


Fig.12 In above screen for same dataset with naïve bayes we got 28% accuracy and now click on ‘Run Decision Tree Algorithm’ button to get below result



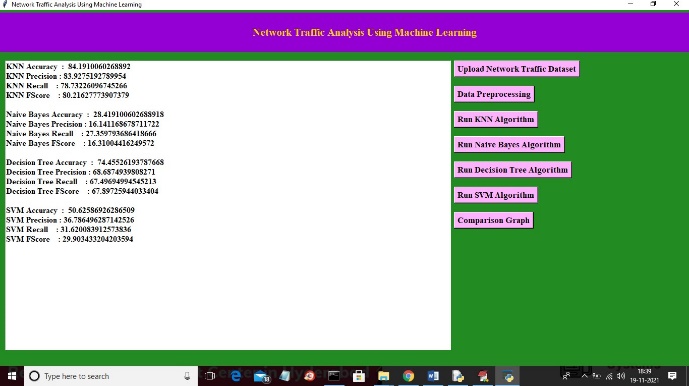
Fig.13 In above screen for same dataset with SVM we got 74% accuracy and now click on ‘Run SVM Algorithm’ button to get below result

Fig.14 In above screen with SVM we got 50% accuracy and now click on ‘Comparison Graph’ button to get below result

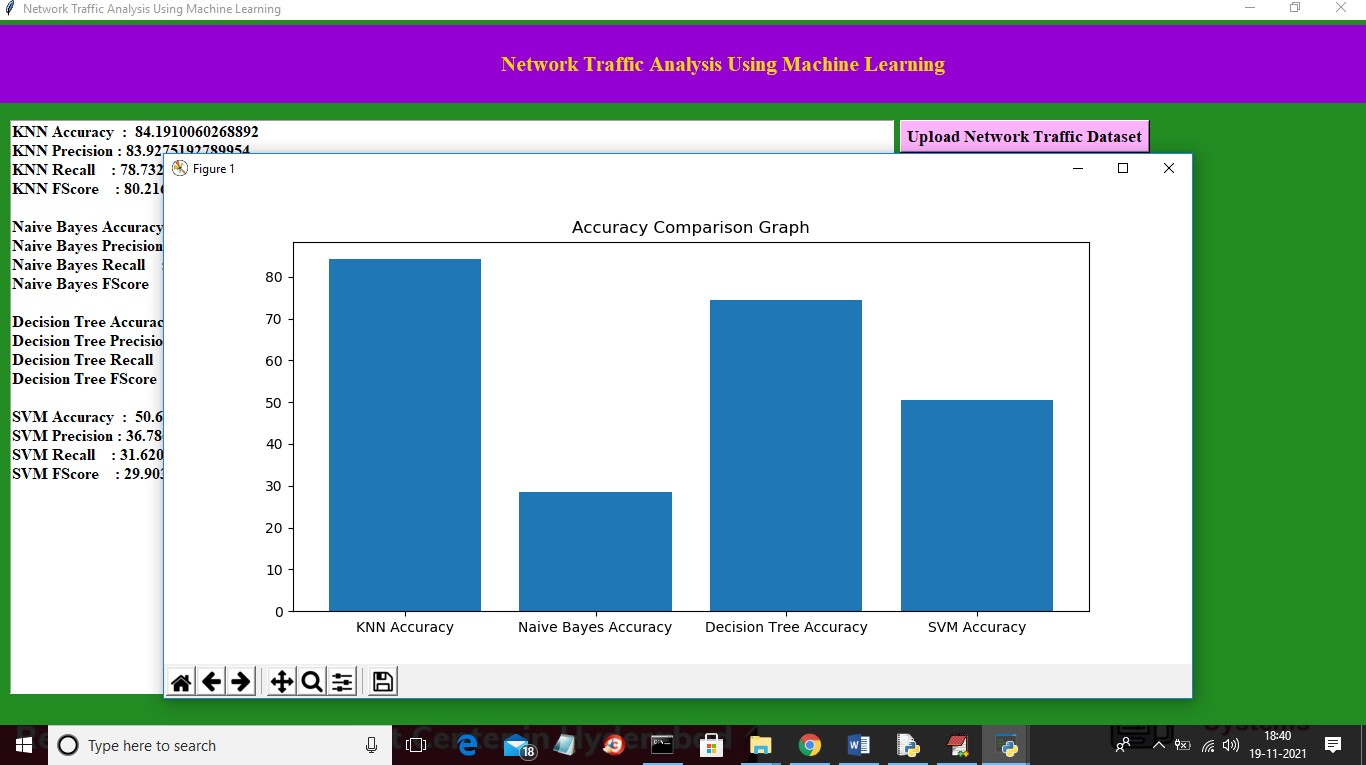


Fig.15 In above graph x-axis represents algorithm name and y-axis represents accuracy of those algorithms and in all algorithms KNN shows better result

CONCLUSION**:**

The Network traffic classification techniques are discussed in this paper to enhance some idea about Machine Learning algorithms for network traffic data. The analysis carried out definitely helps to a new analyst to make the decision about which Machine Learning algorithm is more appropriate for this application. Initially, the network traffic extraction is carried out to evaluate the different Machine Learning algorithm which is trained in later phase. The Machine Learning algorithms are used for managing the performance of network and classification of unknown applications.We then employ four basic Machine Learning algorithms to analyze the protocol. Further, the classifiers using different Machine Learning algorithms are developed to compare the accuracy for this network traffic data. We find that K-nearest neighbor (KNN) algorithm outperforms Naïve Bayes algorithm, Decision Tree and Support Vector Techniques in terms of accuracy which is due to the fact that KNN uses better classification criterion than Naïve Bayes and Decision Tree Algorithm. We find that KNN is most robust among the algorithms: NB, DT, and SVM for out training data set. It is also able to maintain highest mean for accuracy.

REFERENCES:

[1] Chakraborty, A., J.S. Banerjee, and A. Chattopadhyay. Non-uniform quantized data fusion rule alleviating control channel overhead for cooperative spectrum sensing in cognitive radio networks. in 2017 IEEE 7th International Advance Computing Conference (IACC). 2017. IEEE.

[2] Chakraborty, A., J.S. Banerjee, and A. Chattopadhyay, Non-uniform quantized data fusion rule for data rate saving and reducing control channel overhead for cooperative spectrum sensing in cognitive radio networks. Wireless Personal Communications, 2019. 104(2): p. 837-851.

[3] Rueda, A. A survey of traffic characterization techniques in telecommunication networks. in Proceedings of 1996 Canadian Conference on Electrical and Computer Engineering. 1996. IEEE.

[4] Shahbar, K. and A.N. Zincir-Heywood. How far can we push flow analysis to identify encrypted anonymity network traffic? in NOMS 2018 -2018 IEEE/IFIP Network Operations and Management Symposium. 2018. IEEE.

[5] Axelsson, S., Intrusion detection systems: A survey and taxonomy. 2000, Technical report.

[6] Wang, P., Y. Li, and C.K. Reddy, Machine learning for survival analysis: A survey. ACM Computing Surveys (CSUR), 2019. 51(6): p. 110.

[7] Namdev, N., S. Agrawal, and S. Silkari, Recent advancement in machine learning based internet traffic classification. Procedia Computer Science, 2015. 60: p. 784-791.

[8] Cheng, Y., et al., Bridging machine learning and computer network research: a survey. CCF Transactions on Networking, 2019. 1(1- 4): p. 1-15.

[9] Mukkamala, S., G. Janoski, and A. Sung. Intrusion detection: support vector machines and neural networks. in proceedings of the IEEE International Joint Conference on Neural Networks (ANNIE), St. Louis, MO. 2002.

[10] Taylor, V.F., et al., Robust smartphone app identification via encrypted network traffic analysis. IEEE Transactions on Information Forensics and Security, 2017. 13(1): p. 63-78. [11] Kim, J., et al., Multivariate network traffic analysis using clustered patterns. Computing, 2019. 101(4): p. 339-361.

[12] Shafiq, M., et al. Network traffic classification techniques and comparative analysis using machine learning algorithms. in 2016 2nd IEEE International Conference on Computer and Communications (ICCC). 2016. IEEE.

[13] Sommer, R. and V. Paxson. Outside the closed world: On using machine learning for network intrusion detection. in 2010 IEEE symposium on security and privacy. 2010. IEEE.