JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2402 - 2409
https://publishoa.com
ISSN: 1309-3452

Utilizing Machine Learning Techniques to Classify Network Traffic

Dr.Ch. Ramesh Babu [1], Mohd Farook Ali [2], Aslam Pasha [3], Zulqairnain Arshad [4], Mohammad Asif [5]

[1] *Professor & Head, Department of Computer Science and Engineering, Lords Institute of Engineering and Technology, Hyderabad, Telangana, India.*
[2, 3, 4, 5] *Research Scholar, Department of Computer Science and Engineering, Lords Institute of Engineering and Technology, Hyderabad, Telangana, India.*

Email : [1] *chramesh522@gmail.com*

ABSTRACT

Within community of internet, it may be essential to recognize whatever programs are flowing via the networks in order to execute particular activities. Network traffic categorization is primarily used by Internet service providers (ISPs) to determine the qualities needed to construct a connection, which in turn influences the cable network current effectiveness. Stream, bandwidth, and machine-learning methods were all used to categorise internet protocol, and each has its own benefits and drawbacks. Because of its widespread use across disciplines as well as the increasing awareness between many investigators of its methodology when especially in comparison to everyone else, the Machine learning method [5-9] is popular these times. Naive Bayes as well as K-nearest algorithm results are then compared in this research whenever applied to a networking given dataset taken through live stream feeds using Ethernet program. Python's sklearn module and the pandas and numpy arrays modules are utilised as assist modules to create a machine learning algorithm. Our findings show that K closest approach is more efficient than Nave Bayes, Decision Tree, and Support Vector Machine algorithms.

Index Terms— K-Nearest Neighbours (KNN), Naive Bayes (NB), decision trees (DT) and support vector machines (SVM)

I. INTRODUCTION

Challenges connected using new technologies are a major focus of Network Activity Characterization (NTC). Several machine learning-based identification methods are discussed in [1]. Considers the impact of a specific application protocol's effectiveness on a broadband service supplier's actual quality. In the event that an unfamiliar networks attempts to encroach on the designated traffic lane, it can identify it. They can learn more about its qualities this manner. Utilizing aforesaid ability to identify unfamiliar networks, one may also identify the risks that a network may face owing to specific security breaches. It's also important to control network infrastructure and quality of the service (QoS), and these can be done if we use efficient network classification methods. Classifying our network well allows us to block or allow specific network traffic. Ultimately, networking categorization improves the cable network performance and profitability. Various network activity application submitted have been developed over the past few decades, but they've been able to categorise network traffic. First, the Terminal Driver Distraction Classification technique, that was using terminals to categorize every network, has been used to categorize communications. Initially, it was a very effective method. Results analysis in [2] makes use of the port-based categorization technique. In order to categorise ports, the Internet Assign Number Authority (IANA) was employed to register them [3]. The problem was that so many of the channels weren't registered with IANA systems and weren't eligible for dynamic ports, therefore it eventually failed. [4] also provides a brief overview of Internet traffic classification.

After that, many machine learning techniques are explored for their outcomes analysis [5-9]. A novel machine learning-based model for content rating has been built by the researchers of [11] whereas the Internet traffic is examined in [12]. In order to classify Internet traffic, packets must be matched to the application from which they originated. Network management relies on traffic classification, which is used for things like traffic trying to shape, strategy routing, and packet filtering, among other things. Businesses use it for customer profiling, which gives them valuable marketing information, while scientists and government agencies use it to discover global Internet trends.

A single IP packet can be difficult to classify because the protocol headers do not include an application name. Due to P2P traffic, the communication port number was no longer a reliable way to distinguish between different traffic classes in the early 2000s. DPI (Deep Packet Inspection) is another widely used and accepted method for determining a packet's classification. There are privacy and computational costs to consider despite its accuracy. In addition, traffic encryption has rendered DPI obsolete [6]. Network management relies heavily on traffic classification.

It can be used for a variety of purposes, including network security, traffic visualisation, and quality of service monitoring. The rise of peer-to-peer traffic has led to a rapid change in traffic classification during the past decade. Researchers are constantly looking for innovative ways to keep up with the ever-changing nature of the Internet.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2402 - 2409
https://publishoa.com
ISSN: 1309-3452

Secondly, Payload-based techniques were developed, during which packages of the related networks are analysed and protocols are recognised based on the study. Because it analyses packets, this method is referred to as "Deep Packet Filtering." Unfortunately, this method has flopped given the cost of system implementation and the poor performance it provides for encryption transmissions [1]. These shortcomings lead to the use of machine learning, which has been increasingly popular in recent years because of its precision and efficiency. Labelled classes are transformed into models and then tested using accuracy to ensure their validity.
The paper's contributions are outlined below. Afterwards, we use machine learning approaches to a network information source and conduct a comparison evaluation of multiple algorithms to determine which one is best suited to analyse network traffic. Wire shark is used to capture the features, which are then converted to a csv file format and trained and tested using Python Libraries, which aid in prediction and comparison analysis further. This data is then compared [3]. We use DT, NB, KNN, and SVM as well as Naive Bayes (NB) and K-nearest Neighbour (KNN) approaches. For these applications, we find that the KNN Algorithm outperforms than the alternatives.

II. RELATED WORK

Internet traffic identification utilizing ml algorithms is examined in this paper

IP traffic categorization approaches that don't rely on "fully established" TCP or UDP control signals or interpretation of package contents are increasingly being sought by researchers. The utilisation of traffic information to aid in the recognition & categorization procedure is becoming more common. Natural Language processing (nlp (ML) techniques are employed to IP traffic categorization, which is a cross - functional and cross combination of IP networks and data gathering approaches [5]. ML approaches applied to Netflow segmentation are contextualised and motivated, and 18 major papers through 2004 to early 2007 are reviewed. These papers are sorted and evaluated based on the ML methodologies they employ and the key communities in which they live to the field. ML-based traffic classifications in operating IP networks also need to satisfy a set of core conditions, and the assessed works are rated on how well they achieve those demands. The group also discusses current problems and obstacles in the sector.

A comparison of dynamic queuing schemes low- and medium attack of services assault

The worldwide inter-networking infrastructures are increasingly under risk from Denial of Services (DoS) threats. As an outcome of the underlying premise of terminal collaboration, TCP's congested proposed controller is very resistant to a wide variety of internet states. [12] Low-rate denial - of - service assaults, especially rising threats, are harder for firewalls and neutralise systems to identify. In this work, we study these attacks. We suggest that low DoS traffic conditions that use TCP's restoration duration technique can reduce TCP streams to a quarter of normal appropriate price while evading detection by using analytical modelling, simulators, and Net tests. Due to protocols homogenization threats, we investigate the inherent limitations of randomised time-out strategies in combating similar low-rate Denial of Service (DoS) incidents.

IP Traffic Classification Using Machine Learning Algorithms: A Comparative Study

IP traffic classification is becoming increasingly important to broadband providers and other individuals and international organisations because of the tremendous rise in online bandwidth over through the recent years attributable to the use of a range of online services. Due to the sheer usage of random port numbers rather than of well-known ports in incoming packets and various cryptographic mechanisms, [11] conventional Net flow classification methods including ports total count and bandwidth indirect packet filtering approaches are rarely employed today. Categorization via machine learning (ML) is becoming increasingly popular. As part of this study, a packet capture tool was used to gather regular internet traffic data, which was then reduced using attribute selection methods. With these databases, 5 machine-learning methodologies were also utilised to classify IP traffic: MLP (machine learning), RBF (machine learning), C4.5 (machine learning), and Bayes Net (machine learning). The results of this study reveal that Bayes Net & C4.5 are excellent machine learning algorithms for classifying IP traffic, with an accuracy of about 94%.

Using a "learning-based approach to document ranking"

For the purpose of document similarity, a query document is used to find the most relevant documents. Using a score function, document similarity methods have been shown to initially approximate semantic similarity between such a query and the documents. As according their similarity scores, documents are then ranked. There are three stages to the Text Tiling algorithm in the literature: tokenization into block of text units, scoring, and determining the subtopic boundaries. Throughout this article, we looked at two different methods for document ranking and compared the results to those of a machine learning approach. In the first place, documents are ranked according to the tf-idf concept, which

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2402 - 2409
https://publishoa.com
ISSN: 1309-3452

uses a standard score calculation. Second, the Text ling approach is used to rank documents. Strategies are an important Oriental publication text messages, that really have no paragraph breaks, using Text Tiles is already being implemented in a user interface for an information retrieval system. When summarising documents, textiles are a considerably superior method than ordinary score computation. The system's retrieval performance benefits from this. In this research, we compared and contrasted two methodologies. Our results also included the statistical machine learning techniques, which can be used to solve a wide range of information retrieval issues. Moreover (IR).

A brief summary of a few key articles on the topic of Internet traffic classification Informatics, both theoretical and practical

Network management relies heavily on traffic classification. It can be used for a variety of purposes, including network security, traffic visualisation, and quality of service monitoring. The rise of peer-to-peer traffic has led to a rapid change in traffic classification during the past decade. Researchers are constantly looking for innovative ways to keep up with the ever-changing nature of the Internet. Throughout 2009-2012, a total of 13 publications were published on the subject of traffic classification and associated subjects. Our findings demonstrate the breadth of modern traffic classification algorithms, as well as probable future routes for traffic classification research: the value of multi-level classification, the need for experimental tests, and the importance of common traffic datasets.

III. METHODOLOGY

A variety of machine learning techniques are being used in this research to anticipate traffic or categorise network data including BROWSING, MAIL, and other types of traffic.

A. Naive Bayes Algorithm

His approach is based on the Naive bayes algorithm, which states that almost all training data to learning were independently from each other, since every characteristic of classes was estimated independently when computing the entire. This theory exceeds a number of other Machine Learning techniques, making it especially beneficial when the training dataset is vast. In [8], a full explanation of the Nave Bayes algorithm is provided. For every tuple in the training sample D, there are n "k" attribute values, each of which is represented by the vector V, which is equal to the sum of the tuple's "k" values. C1, C2,...Ck are the "k" classes. As according Naive Bayesian classification, only when it has greater posterior distribution than any class Cy amongst C1, C2,...Ck, where x y, does an input pair I belong to class Cx according to Naive Bayesian classification.

$$P\left[\frac{C_x}{I}\right] > P\left[\frac{C_y}{I}\right], \qquad (1)$$

$$P\left[\frac{C_x}{I}\right] = \frac{P\left[\frac{I}{C_x}\right]P[C_x]}{P[I]}. \qquad (2)$$

B. K – Nearest Neighbours

Simply saving and identifying new examples based on the given parameter is what this technique is built on (mostly distance). When it comes to statistical approaches and predictive modelling, the K-NN technique has been around for a long time. In circumstances where the feature data types are indeed quantitative & categories, the application's computation of distances as a judging parameter yields subpar results. Algorithms that are using Machine Learning can be evaluated by using the test dataset methodology. During in the training stage, the original dataset is tested for outcomes. [4] K-fold Serial Correlation is a machine learning technique that is commonly employed. This method's steps are as follows: It is necessary to partition the original batch across k equal subgroups in order to begin this process. For the sake of clarity, we'll refer to these subsets as "folds," with f1 through fk being assigned to them. For j = 1 to j = k, repeat the loop. Let the fold be the Authentication subset, and the rest of the k-1 sets be cross-validating training sets for confirmation. This cross-validation trained training data set the Machine Learning Technique, and the accuracy is calculated by compared the validating outcomes towards the real amounts in the testing dataset. A machine learning designer's ultimate efficiency is anticipated by averaging the accuracy results from k cross validation tests. The advantages and disadvantages of the method are outlined below. With a huge and noisy dataset, this technique performs better than other methods when it comes to training. The most difficult part of this method is predicting the best K value. Another drawback with using this method is the high processing cost because range is determined for every data input. Another drawback is the inherent uncertainty that comes along with using "length" as a metric to assess performance.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2402 - 2409
https://publishoa.com
ISSN: 1309-3452

C. Decision Tree Algorithm

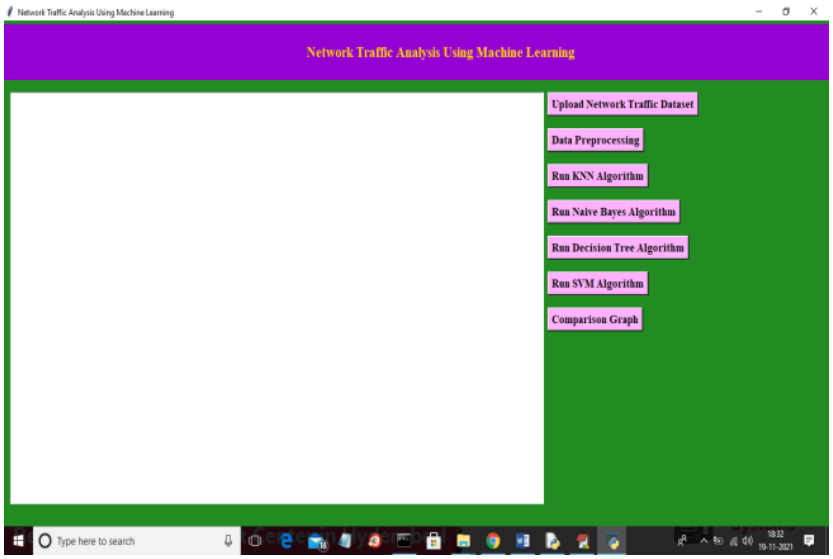
Using the principle of "training," a Decision Tree performs classification that could anticipate category targeted parameters values by presuming to the determination standards set and determined during the training sample phase. It has advantages since it is easy to explain [6] because it makes judgments just like a person. However, if there are numerous classification identifiers, the computations can become complicated. A decision tree could be built using a variety of methods. A few of those are indeed the ID3, C4.5, CHAID, and MARS (Categorisation of Regression Model). The CART method is used to make predictions. The Decision Tree algorithm's transparency is a well-known benefit. It implies that the findings are being provided after careful consideration of the best method to obtain the best outcomes. The situation's specificity and the value assigned to it make it a significant advantage. Because of its comprehensiveness, it is able to consider every alternative. In addition to being more visually appealing, this tree is much more accessible to the average user. This also performs well when dealing with a mix of quantitative and qualitative records. Even a small change in the input information can cause these algorithms to be even more volatile, which can lead to subpar outcomes. Complex trees may perform worse when tested with a variety huge datasets. Because when algorithms picks up on disturbance inside the information, it results in imbalanced datasets.

D. Support Vector Machine

Many different types of classification and regression problems can be addressed with SVMs, which are a set of similar supervised learning techniques. Linear classification family members include them. SVM's unique feature is its ability to minimise empirical classification while simultaneously increasing geometrical margin. So SVMs are also known as Makes More sense Classification models. Structural risk minimization is referred to as SVM (SRM). Through the use of SVM [9] , a hyper plane with the greatest possible separation can be generated from the input vector. Along one edge of the information higher dimensional space, two parallel hyper planes are constructed. Splitting hyper - parameters increases the separation between two perpendicular hyper planes.
Because it has no local minima, the SVM is an effective approach for convex combinatorial optimization.
Most analysts are attracted to SVM since it is predicated on an estimation of a limit on the test error rate.

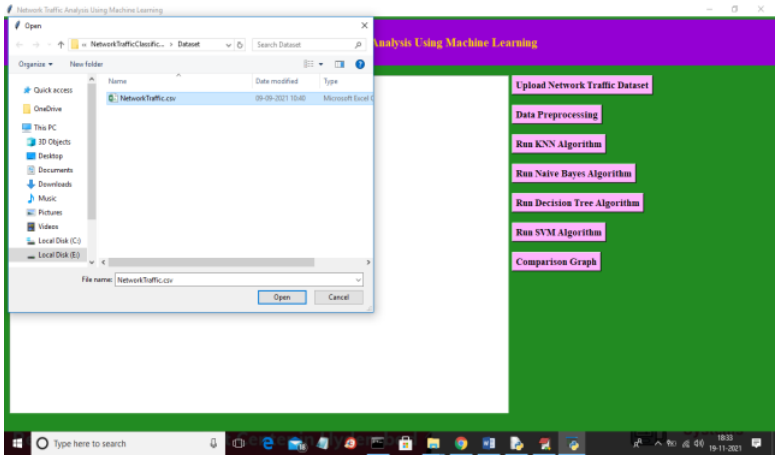
IV. Result and Discussion

In this project we are using various machine learning algorithms such as KNN, SVM, Decision Tree and Naïve Bayes to predict traffic or classify type of network data such as BROWSING traffic, MAIL traffic etc. Lots of network traffic type of data is available but in this project we are training ML algorithms to predict or classify 14 different types of traffic. To run the project to get below result

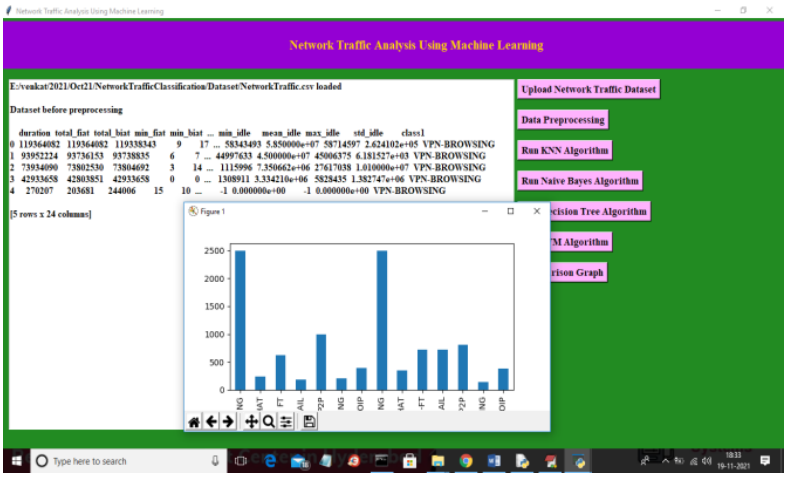


Uploaded Network Traffic Database by clicking upload Network Traffic Dataset in the above result.

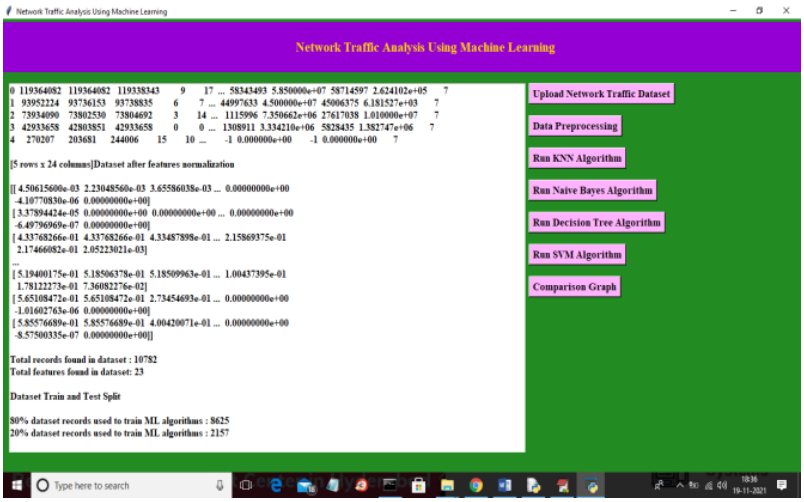
JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2402 - 2409
https://publishoa.com
ISSN: 1309-3452



Choose NetworkTraffic.csv and then click "Open" to load the dataset. The outcome is shown below.

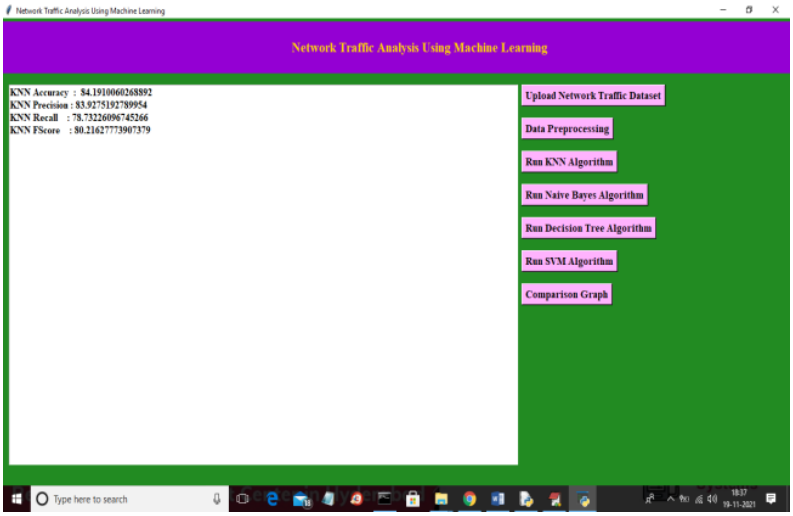


There are a lot of non-numeric values in the dataset that we need to analyse, therefore we can see that in the graph x-axis (traffic type) and y-axis (the number of entries in the dataset for that traffic) are shown. To tidy up the dataset, click 'Data Pre-processing' after closing the graph above.

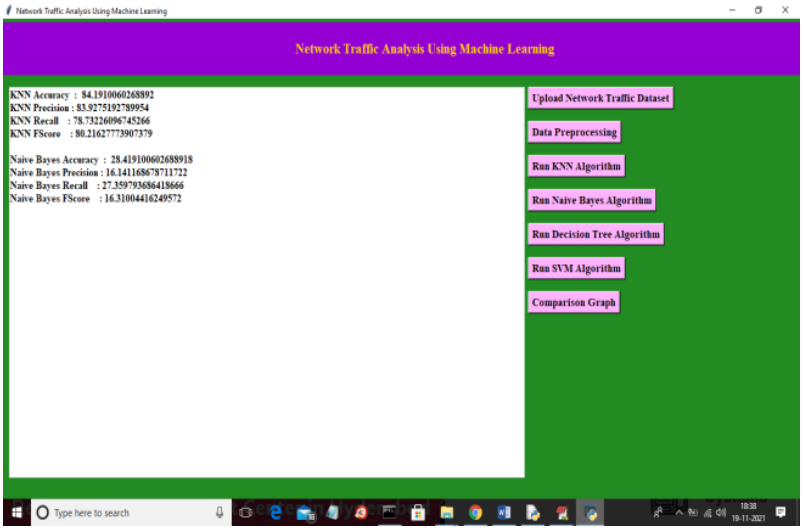


To summarise, we can see the total number of observations & fields found in the dataset, as well as a percentage breakdown of the dataset into train and test records, in the results shown above. When all of the training and testing data is in place, click on the 'Run KNN Algorithm' tab to begin training KNN.

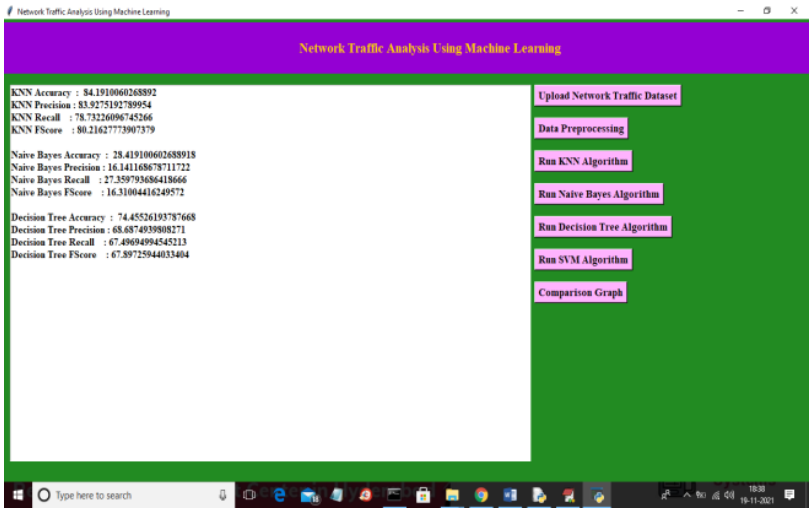
JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2402 - 2409
https://publishoa.com
ISSN: 1309-3452



KNN accuracy is shown in the above results. With KNN, we achieved an accuracy rate of 84%, and the Run Nave Bayes Algorithm tab will be used to begin training the model.

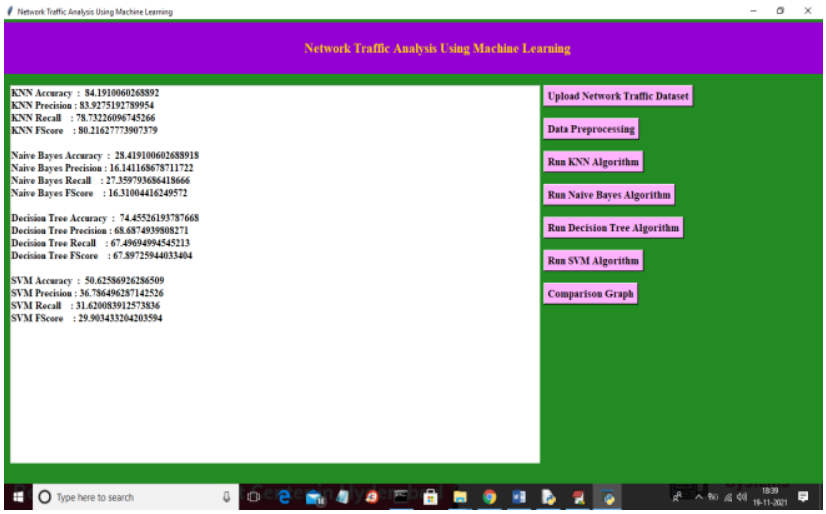


Using naive bayes, we were able to acquire a 28 percent accuracy rate for the same dataset, but when we clicked the 'Run Decision Tree Algorithm' button, we were able to see the results below.

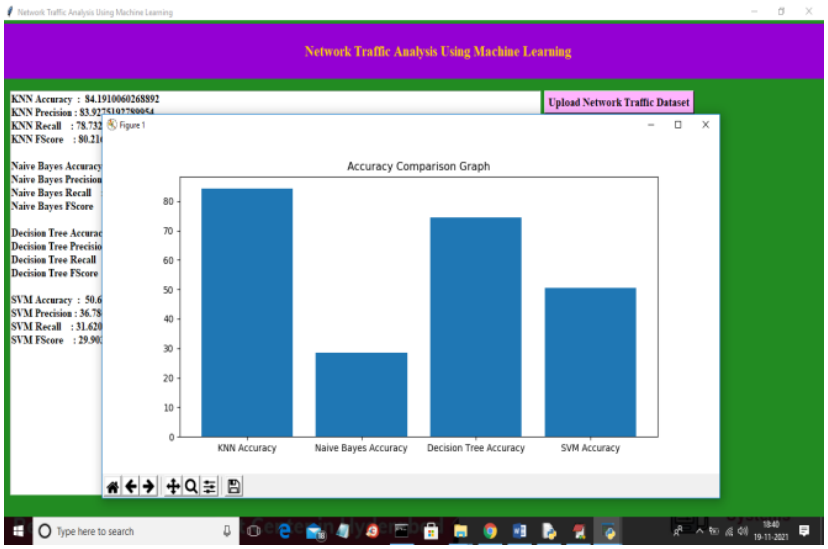


JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2402 - 2409
https://publishoa.com
ISSN: 1309-3452

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



SVM Accuracy is shown in the results above. Using SVM, we were able to get a 50% accuracy rate. Click on 'Compare Graph' to see the following result.



This graph indicates that KNN outperforms all other algorithms on the x-axis and the y-axis.

V. CONCLUSION

In order to better understand machine learning techniques for dataset, this study explores network traffic evaluation measures. A novice analysts can benefit greatly from the study done, as it allows them determine whichever machine learning model is most suited for this particular procedure. As a first step, the internet traffic separation is used to assess the various Machine Learning techniques that will be trained in the subsequent phase. Algorithms for machine learning are employed to control system performance and classify unidentified activities.

Machine Learning methods are then used to study the protocols. This data is also used to construct classifications utilizing various Machine Learning techniques to examine their efficiency. Due to KNN's better classifications than Nave Bayes and Decision Tree Method; we discover that K-nearest neighbour (KNN) algorithm surpasses Nave Bayes methodology, Decision Tree and Support Vector Methodology through high accuracy. To test our training data set [5], we found that KNN was the most stable method out of the three other algorithms: NB, DT and SVM. Maintaining the excellence of precision is also possible.

REFERENCES

1. Nguyen, Thuy TT, and Grenville Armitage. "A survey of techniques for internet traffic classification using machine learning." IEEE Communications Surveys & Tutorials, vol. 10, no. 4, pp. 56-76, 2008.
2. M. Shafiq, X. Yu, A. A. Laghari, L. Yao, N. K. Karn, and F. Abdessamia, "Network traffic classification techniques and comparative analysis using machine learning algorithms," in Proc. IEEE International Conference on Computer and Communications (ICCC-2016), pp. 2451- 2455, 2016.
3. Internet Assigned Numbers Authority (lANA), http://www.iana.orglassignments/port-numbers, as of August 12, 2008.
4. Pawel Foremski, "On different ways to classify Internet traffic: a short review of selected publications Theoretical and Applied Informatics, " 2013.
5. K. Sing and S. Agrawal, "Comparative Analyssis of Five Machine Learning Algorithms for IP Traffic Classification," in Proc. IEEE International Conference on Emerging Trends in Network and Computer Communication (ETNCC-2011), pp. 33-38, 2011.
6. Q. Dai, C. Zhang and H. Wu, "Research of Decision Tree Classification Algorithm in Data Mining," International Journal of Database Theory and Application, vol. 9, no.5, pp. 1-8, 2016.
7. Cristina Petri, "Decision Trees", Cluj Napoca, 2010.
8. S. Karthika and N. Sairam, "A Naïve Bayesian Classifier for Educational Qualification," Indian Journal of Science and Technology, vol. 8, no. 16, Jul. 2015; DOI: 10.17485/ijst/2015/v8i16/62055.
9. D. K. Srivastava and L. Bhambhu, "Data Classification Using Support Vector Machine," Journal of Theoretical and Applied Information Technology, vol. 12, no. 1, Feb. 2010.
10. Wireshark tool: https://www.wireshark.org/docs/dfref/.
11. S. Patel and A. Sharma, "The low-rate denial of service attack based comparative study of active queue management scheme," in Proc. 2017 Tenth International Conference on Contemporary Computing (IC3- 2017), pp. 1-3, 10-12 Aug. 2017.
12. S. Patel, K. Khanna, and V. Sharma, "Documents ranking using learning approach," in Proc. 2016 International Conference on Computing, Communication and Automation (ICCCA-2016) , pp. 65-70, 29-30 Apr. 2016.