**Darknet Traffic Classification Using Machine Learning Techniques**

**Abstract**

Darknet traffic classification plays an important to categorizing real-time applications; it is an unused address space used on the internet. Analyzing darknet traffic aids in detecting malware and early monitoring of malware before it outbreaks. To identify Darknet traffic, we used machine learning methods. A ROC curve is used to provide a better visual representation of the results, and a feature selection analysis is used for better classifier results. The experiments were carried out on the CIC-Darknet2020 dataset. Traffic is divided into two categories: “Benign” and “Darknet,” where “Tor” and “VPN” are considered into the “Darknet” category and “Non-Tor” and “Non-VPN” are regarded into the “Benign” category. Using several supervised machine learning approaches, like Logistic Regression, Support Vector Machine, Naive Bayes, K-Nearest Neighbors and Decision Tree Classifier, the average prediction accuracy of over 99% was achieved.

## ****Introduction****

## ****Darknet****

A darknet is an Internet overlay network that can only be accessed with specific software, configurations, or authorization and frequently uses a custom communication protocol. “Darknet” this term was originated in the 1970s to describe the networks that are not connected to ARPANET (Advanced Research Projects Agency Network) for security reasons. Darknet addresses could connect to ARPANET and receive data, but they didn’t show up in network listings or respond to pings or other enquiries. Despite receiving communications from ARPANET, they did not respond or recognize them to remain invisible. It is also known as network telescopes, sinkholes, or black holes. Social networks (often used for file hosting over a peer-to-peer connection) and anonymity proxy networks (Tor) are mainly two darknet types.

The act of spreading or providing access to digital media, such as computer programs, multimedia (audio, pictures, and video), documents, or electronic books, is known as File Sharing. File-sharing can be done in a variety of ways. Manual sharing using portable media, centralized servers on computer networks, World Wide Web-based hyperlinked papers and distributed peer-to-peer networking are all common means of storage, transfer, and dispersion.

P2P computing, also known as peer-to-peer networking, is a distributed application architecture in which jobs or workloads are dispersed among peers. Peers are participants in the application who have the same opportunity and are similarly capable. They’re supposed to create a peer-to-peer network of nodes. Peers make a portion of their resources, such as computing power, disc storage, or network bandwidth, directly available to other network members without the need for central coordination by servers or dependable hosts. In contrast to the typical client-server model, peers are both resource suppliers and consumers, which divides resource consumption and supply.

An anonymizer, also called as an anonymous proxy, is a tool that tries to hide your online activity. It’s a proxy server computer that functions as a middleman and a privacy shield between a client computer and the internet. It connects to the internet on the user’s behalf, masking the client computer’s identifying information and protecting its data. Anonymizers are helpful for various reasons, including reducing risk, preventing identity theft, and concealing search records from public disclosure.

TOR and VPN provide users with encrypted entry points and pathways to the darknet. Due to this layered encryption mechanism, darknet users’ identities and locations remain anonymous and cannot be monitored. Users’ data is routed through many intermediate servers using darknet encryption technology, which conceals users’ identities and ensures anonymity. Only a following node in the scheme, which leads to the exit node, may decode the transferred data. The sophisticated mechanism makes duplicating the node path and decrypting the information layer by layer nearly impossible. Websites cannot trace their users’ geo-location and IP addresses because of the high level of encryption, and users are unable to obtain this information about the host. As a result, darknet users’ communication is highly encrypted, allowing them to converse, blog, and share confidential files.

## ****VPN****

A virtual private network connects a private network to a public network, allowing users to send and receive data as if their computers were connected to the private network directly. The functionality, security, and management of a VPN may benefit applications operating over it. It allows telecommuting employees access to resources that are not available on the public network. Encryption is often used, but it is not a requirement for a VPN connection. Dedicated circuits or tunnelling techniques are used to build a virtual point-to-point connection over existing networks, resulting in a VPN. A vast area network’s benefits can be obtained using a VPN accessible via the public Internet (WAN). The resources provided within the private network can be accessed remotely from the user’s perspective. VPN provides confidentiality, authentication and integrity to the transmitted messages. It is mainly classified into three categories:-

* **Remote access:-** Connecting a PC to a local area network is equivalent to a host-to-network configuration. This type allows users to connect to a corporate network, such as an intranet. This could be used by telecommuting workers who require access to private resources or by mobile workers who need access to critical technologies without exposing them to the public internet.
* **Site-to-Site:-** Two networks are connected through a site-to-site arrangement. This setup connects a network to a data centre installation through geographically dispersed offices or a group of offices. A distinct intermediary network, such as two IPv6 networks connected across an IPv4 network, could be used for the interconnecting link.
* **Extra-net-Based Site-to-Site:-** The terms intranet and extranet are used to define two separate use cases in site-to-site deployments. An intranet site-to-site VPN connects all of the same organization, whereas an extranet site-to-site VPN connects all of the different companies.

## ****TOR****

The Onion Router (TOR) is a free and open-source technology that allows users to communicate anonymously. It hides a user’s location and usage from anyone doing network surveillance or traffic analysis by routing Internet traffic over a free, global volunteer overlay network with over 6,000 relays. Tor makes it more difficult to track an individual’s online behaviour. Tor’s purpose is to safeguard its users’ privacy and their freedom and capacity to communicate in confidence by preventing their Internet activity from being monitored.

Tor allows its users to access the internet, communicate, and send instant messages while remaining anonymous. Many people use it for both lawful and illegal objectives. Tor isn’t intended to be a perfect solution to the problem of online anonymity. Tor isn’t meant to wipe your tracks entirely; rather, it’s intended to make it harder for websites to track your actions and data back to you. Tor is also used for nefarious purposes. Privacy protection or censorship evasion and the spread of child abuse content, drug transactions, or malware distribution are all examples. “Overall, on an average country/day, 6.7 per cent of Tor network users connect to Onion/Hidden Services that are disproportionately used for unlawful activities,” according to one assessment. It is implemented in many programming languages in different ways like:-

* Tor Browser
* Firefox/Tor browser attack
* Tor Messenger
* Third-party applications
* Security-focused operating systems

Tor browser has three levels of security, which can be found under the Security Level (the small grey shield at the top-right of the screen) icon > Advanced Security Settings, depending on the user’s demands. Several extra layers of protection are available to a user, in addition to encrypting data and constantly changing an IP address over a virtual circuit comprised of successive, randomly selected Tor relays:-

* Standard (default) - all browser features are enabled at this security level.
* This level offers the most practical experience while also offering the least security.
* Safer – the following changes apply at this security level:
* On non-HTTPS sites, JavaScript is disabled.
* Performance optimizations are disabled for sites that use JavaScript. Some websites’ scripts may take longer to load.
* Some arithmetic equation display techniques are disabled.
* Click-to-play audio and video (HTML5 media), as well as WebGL.
* Safest - At this level of security, the following additional charges apply:
* On all sites, JavaScript is turned off by default.
* The use of several fonts, icons, math symbols, and images is restricted.
* Click-to-play audio and video (HTML5 media), as well as WebGL.

## ****Importance of machine learning in cyber-attacks detection****

The cyber threat landscape requires the ongoing tracking and correlation of millions of external and internal data points across an organization’s infrastructure and users. It is not possible to manage this volume of data with a small group of individuals. Machine learning excels in this area because it can discover patterns and forecast dangers in large data sets at machine speed. Cyber teams can quickly find threats and isolate instances requiring further human study by automating the analysis.

Machine learning identifies vulnerabilities by continuously monitoring network behaviour for anomalies. Machine learning engines process vast volumes of data in near real-time to detect significant occurrences. Insider threats, undiscovered malware, and policy infractions can all be detected using these methods. Machine learning can help users avoid connecting to harmful websites by predicting “bad neighborhoods” online. Machine learning examines Internet behaviour to detect attack infrastructures ready to respond to existing and emerging threats. Algorithms can detect malware that has never been seen before and is attempting to run on endpoints. It detects new hazardous files and activity based on known malware’s features and behaviour. Machine learning can analyze suspicious cloud app login activity, detect location-based abnormalities, and undertake IP reputation analysis to identify dangers and risks in cloud apps and platforms. Machine learning can detect malware even in encrypted traffic by examining encrypted traffic data pieces in standard network telemetry. Machine learning algorithms, rather than decrypting, identify destructive patterns to uncover risks buried behind encryption.

## ****Supervised Learning****

Machine learning is a task of learning a function that transforms an input to an output based on example input-output pairs is supervised learning. It uses labelled training data and training examples to infer a function. Each instance in supervised learning comprises an input object (usually a vector) and the desired output value (the supervisory signal). A supervised learning algorithm examines the training data and generates an inferred function applied to new cases. The algorithm will accurately determine the class labels for unseen examples in the best-case scenario. This necessitates a “reasonable” generalization of the training data to unknown methods by the learning algorithm. The generalization error is a statistical metric for determining an algorithm’s statistical quality.

Steps involved in supervised learning:-

* Determine the type of training dataset you’ll be using.
* Gather the training data that has been labelled.
* Divide the data into three sections: training, testing, and validation.
* Determine the training dataset’s input characteristics, which should contain enough information for the model to predict the output accurately.
* Choose an appropriate algorithm for the model, such as a support vector machine or a decision tree.
* Use the training dataset to run the algorithm. Validation sets, a subset of training datasets, are sometimes required as control parameters.
* By giving the test set, you may assess the model’s correctness. If the model correctly predicts the outcome, then our model is accurate.

It is of two types:-

1) Classification

2) Regression

### ****A. Classification****

The technique of predicting the class of given data points is known as classification. Targets, labels, and categories are all terms used to describe types. The task of approximating a mapping function (f) from input variables (X) to discrete output variables is known as classification predictive modelling (y). Classification is a type of supervised machine learning in which the input data is also delivered to the objectives. Classification has numerous uses in various fields, including credit approval, medical diagnosis, and target marketing.

It is divided into two categories:

i. Lazy learners:-

Lazy learners simply save the training data and wait for the testing data. When this happens, classification is performed using the most closely related data from the stored training data. It has less training time but more predicting time.

Ex. K-Nearest Neighbors

ii. Eager learners:-

Before receiving data for classification, eager learners develop a classification model based on the available training data. It must commit to a hypothesis that encompasses all possible instances. Enthusiastic learners take a long time to train and a short time to predict due to the model’s structure.

Ex. Decision Tree, Naive Bayes

### ****B. Regression****

The technique of assessing the relationship between a dependent variable and independent factors is known as regression analysis, i.e. fitting a function from a selected family of operations to the sampled data while accounting for some inaccuracy. Regression analysis is one of the strategies for prediction in machine learning. You fit a function to the available information and forecast the future outcome or hold out data points using regression. Some of the regression methods are:-

1. [Linear Regression](https://www.ibm.com/analytics/learn/linear-regression" \t "_blank)
2. [Logistical Regression](https://www.ibm.com/analytics/learn/logistic-regression" \t "_blank)
3. Polynomial Regression

## ****Unsupervised Learning****

Unsupervised learning is a machine learning technique in which models are not supervised using a training dataset, as the name suggests. On the other hand, models use the data to uncover hidden patterns and insights. It is comparable to the learning in the human brain while learning new things. Unsupervised learning can be defined as “a sort of machine learning in which models are taught using unlabeled datasets and then allowed to act on that data without supervision.”

Because, unlike supervised learning, we have the input data but no corresponding output data, unsupervised learning cannot be immediately applied to a regression or classification task. Unsupervised learning aims to uncover a dataset’s underlying structure, categorize data based on similarities, and compactly display the dataset. It is of two types:-

1. Clustering: Clustering is a way of organizing things into clusters so that those with the most similarities stay in one group. In contrast, those with less or no similarities stay in another. Cluster analysis identifies commonalities among data objects and classifies them according to the presence or absence of such commonalities.

2. An association rule is an unsupervised learning strategy used to discover links between variables in an extensive database. It identifies the group of items that appear in the dataset together. The association rule improves the effectiveness of marketing strategies. People who buy X items are more likely to purchase Y items.

Ex:- Market Basket Analysis

## ****Supervised Learning vs Unsupervised Learning****

The difference between the two techniques is the use of tagged datasets. To put it another way, supervised learning algorithms use tagged input and output data, but unsupervised learning algorithms do not.

In supervised learning, the algorithm “learns” from the training dataset by iteratively generating predictions on the data and adjusting for the correct response. While supervised learning models are more accurate than unsupervised learning models, they still require human input in order to accurately recognize the data. A supervised learning model, for example, can forecast the length of your commute based on the time of day, weather conditions, and other factors. But first, you’ll have to teach it that driving in rainy weather takes longer.

On the other hand, Unsupervised learning models function independently to uncover the structure of unlabeled data. It’s worth noting that validating output variables still necessitate human intervention.

Goals:- The purpose of supervised learning is to predict new data results. You know exactly what to expect from the start. An unsupervised learning algorithm aims to derive insights from enormous amounts of further data. What is unusual or exciting from the dataset is determined by machine learning.

Applications:- Spam detection, sentiment analysis, weather forecasting, and pricing forecasts are just a few of the applications for supervised learning models. On the other hand, unsupervised learning is well suited to anomaly detection, recommendation engines, customer personas, and medical imaging.

Complexity:- Supervised learning is a straightforward machine learning method usually calculated using languages like R or Python. You’ll need robust tools for working with vast amounts of unclassified data in unsupervised learning. Unsupervised learning models are computationally complex because they require an extensive training set to obtain the desired results.

Drawbacks:- Training supervised learning models takes time, and the labels for input and output variables require knowledge. Meanwhile, unless human intervention is used to evaluate the output variables, unsupervised learning algorithms might produce radically erroneous findings.

# 2. Methodology

Data Acquisition

Data Pre-processing

Feature Selection

Building Supervised Models

Comparative Analysis

Outcomes

1. **Stage 1**

### ****Data Acquisition****

The term “data acquisition” refers to the process of gathering data from relevant sources before it is stored, cleaned, preprocessed, and used in other methods. It is the process of gathering essential business data, translating it into the appropriate business form, and loading it into the right system. It is categorized into three main segments:-

1. Data Exploration:- Data discovery is the initial step in acquiring data. This is a critical step when indexing, distributing, and searching for new datasets available on the web and incorporating data lakes. There are two parts to it: searching and sharing. To begin, the data must be categorized or indexed, then published for sharing via one of the various collaborative platforms available.

2. Data Augmentation:- Data augmentation is the next step in the data collecting process. We are essentially enriching the existing data by adding more external data in the context of data acquisition. Augment is to make something more outstanding by adding to it. Therefore we are essentially enriching the existing data by adding more external data. Using pre-trained models and embeddings to expand the number of training features is frequent in deep and machine learning.

3. Data Generation:- The data is created, as the name implies. If we don’t have enough and don’t have access to any external data, we can make the datasets manually or automatically. Crowd-sourcing is a standard method for manually collecting data, in which people are assigned tasks to gather the information needed to create a dataset. Automatic methods for creating synthetic datasets are also available. Also, where data is accessible but missing values that need to be credited, the data production process can be viewed as data augmentation.

### ****Collection Of Datasets****

VPN and Tor applications are combined in the CICDarknet2020 dataset to detect and characterize the genuine representation of darknet traffic by combining two public datasets, namely ISCXTor2016 and ISCXVPN2016, to build a complete darknet dataset encompassing Tor and VPN traffic, respectively. A two-layered method is employed at the first layer to generate benign and darknet traffic. The second layer of the darknet traffic comprises Audio-Stream, Browsing, Chat, Email, P2P, Transfer, Video-Stream, and VOIP. Table 1 lists the different types of darknet traffic and the applications used to generate it.

1. **Stage 2**

### ****Data Pre-processing****

The transformations we apply to our data before feeding it to the algorithm are preprocessing. Data preprocessing is a method for converting unclean data into a clean data set, i.e. anytime data is received from various sources, collected in raw format, making analysis impossible.

Need of Data Preprocessing

Ø The data must be formatted correctly to achieve better outcomes from the used model in Machine Learning applications.

Ø The data set should be organized so that it can run many Machine Learning and Deep Learning algorithms in parallel and choose the best one.

1. **Stage 3**

**Feature Selection**

Feature selection is the process of selecting the features that contribute the most to the prediction variable or output that you are interested in, either automatically or manually. The presence of irrelevant characteristics in your data can reduce model accuracy and cause your model to train based on exterior features. Some of the feature selection techniques are:-

1. Filter Method:- This method employs the variable ranking strategy to choose the variables for ordering, and the features chosen are unaffected by the classifiers utilized. When we talk about ranking, we’re talking about how valuable and crucial each attribute is for classification. As a preprocessing step, it picks subsets of variables independent of the predictor. Before classification, the ranking approach can filter out the less essential features in filtering. It performs feature selection as a preprocessing phase without an induction approach. Some of the filter methods are:-

a. Chi-Square Test:- This method is used to test the independence of two events in general. We can acquire the observed count and the predicted count from a dataset for two occurrences, and this test assesses how much both counts are derived from one other.

b. Variance Threshold:- This feature selection method eliminates any features whose variance falls below a certain threshold. In general, it eliminates all zero-variance characteristics, which have the same value across all samples.

c. Information Gain:- Information gain (IG) refers to the amount of information a feature provides about a class. As a result, we can determine which attribute in a set of training features is useful for distinguishing between the classes to be lean.

2. Wrapper Method:- The learning machine of interest is used as a black box in this method to score subsets of variables based on their prediction capability. The induction technique is illustrated with a collection of training cases in the above picture. Each instance is described by a vector of feature values and a class label in supervised machine learning. The induction algorithm, often known as the black box, is used to create a classifier that may be used to classify data. The feature subset selection technique is used as a wrapper around the induction process in the wrapper approach. One of the most significant disadvantages of this method is the many computations necessary to produce the feature subset. Some of the wrapper methods are:-

a. Genetic Algorithms:- A subset of features can be found using this technique. CHCGA is a modified version of this algorithm that converges faster and produces a more practical search by preserving population diversity and avoiding stagnation.

b. Recursive Feature Elimination:- RFE is a feature selection method that fits a model and removes the weakest feature (or features) until the desired amount of features is reached. The model’s coef\_ or feature importances\_ attributes rank features, and RFE seeks to minimize dependencies and collinearity in the model by recursively deleting a small number of features per loop. RFE requires that a certain number of features be kept, although useful features are frequently unknown in advance. Cross-validation is used with RFE to score several feature subsets and pick the top-scoring collection of features to determine the optimal amount of features.

c. Sequential Feature Selection:- This naive method starts with a null set, then adds one feature to the first step that represents the maximum value for the objective function, and then adds the remaining features individually to the current subset from the second step onwards, resulting in the new subgroup being assessed. This approach is continued until all of the essential features have been included.

3. Embedded Method:- This method seeks to combine the efficiency of both preceding methods and performs variable selection throughout the training process. It is usually particular to specific learning machines. This method determines which attribute contributes the most to the model’s accuracy. Some of the embedded method techniques are:-

a. L1 Stabilization:- LASSO (Least Absolute Shrinkage and Selection Operator) is a linear model that estimates sparse coefficients and is effective in specific situations since it prefers solutions with fewer parameter values.

b. Ridge Regression:- The L2 Regularization, also known as Ridge Regression or Tikhonov Regularization, solves a regression model with a linear least-squares function as the loss function and regularization.

c. Elastic Net:- This linear regression model is trained with L1 and L2 as regularizes, allowing it to learn a sparse model with few non-zero weights, similar to Lasso, but yet keeping Ridge’s regularization qualities.

1. **Stage 4**

### ****Building Supervised Models****

**Naive Bayes**

The Bayes theorem inspired Naive Bayes, a probabilistic classifier that works under the assumption that the qualities are conditionally independent. With the above assumption applied to Bayes theory, the classification is done by obtaining the maximum posterior, the maximal P(Ci|X). This assumption drastically minimizes the computational cost by merely counting the class distribution. Even though the premise is not valid in most circumstances because the qualities are dependent, Naive Bayes has performed well. Naive Bayes is a straightforward algorithm to develop, and it has produced good results in the majority of applications. Because it requires linear time rather than the expensive iterative approximation employed by many other types of classifiers, it can quickly scale to larger datasets. The zero probability problem can be a difficulty with naive Bayes. The prediction is invalid when the conditional probability for a given property is zero. Using a Laplacian estimator, this must be addressed explicitly.

There are three types of Naive Bayes classifiers:-

ü Multinomial Naïve Bayes

ü Bernoulli Naïve Bayes

ü Gaussian Naïve Bayes

### ****Support Vector Machine (SVM)****

The SVM algorithm’s purpose is to find the best line or decision boundary that can divide n-dimensional space into classes so that new data points can be readily placed in the correct category in the future. A hyperplane denotes the optimal choice boundary. SVM selects the hyperplane helping extreme points/vectors. Support vectors are extreme situations, and the Support Vector Machine algorithm is named after them. It is of 2 types:-

i. Linear SVM:- It is a classifier used for linearly separable data, implying that if a dataset can be classified into two types using a single straight line, it is called linearly separable data, and the classifier is named Linear SVM.

ii. Non-linear SVM:- It is used for non-linearly separated data, which implies that if a dataset can’t be classified using a straight line, it’s non-linear data, and the classifier employed is called Non-linear SVM.

**K-Nearest Neighbor (KNN)**

The k-Nearest Neighbor algorithm is a lazy learning algorithm that stores all instances in n-dimensional space corresponding to training data points. When an unknown discrete data is received, it examines the nearest k number of saved models (nearest neighbours) and returns the most common class as the prediction. In contrast, real-valued data returns the mean of k nearest neighbours. The distance-weighted most immediate neighbour method uses the following query to weight the contributions of each of the k neighbours based on their distance, providing more significant weight to the closest neighbours.

### ****Decision Trees****

A decision tree builds regression or classification models in a tree structure. It uses mutually exclusive and exhaustive if-then rules to classify data. The rules are learned at a time, one by one, from the training data. The tuples covered by a rule are eliminated each time it is learned. On the training set, this process is repeated until a termination condition is satisfied. Top-down recursive divide-and-conquer is used to build the tree. All of the characteristics must be categorical. They should be discretized ahead of time if not. The information gain concept is used to identify attributes at the top of the tree that significantly impact classification. A decision tree can easily be over-fitted, resulting in an excessive number of branches, revealing anomalies due to noise or outliers. The performance of an over-fitted model on unseen data is terrible, despite its outstanding performance on training data. Pre-pruning, which stops tree growth early, or post-pruning, which removes branches from a fully grown tree, can help avoid this.

### ****Random Forest****

Random forest is a supervised machine learning technique used for classification and regression. The “forest” refers to a group of uncorrelated decision trees combined to reduce variation and generate accurate data predictions. Any of the individual constituent models will outperform many reasonably uncorrelated models (trees) working as a committee. The key is the low correlation between models.

### ****Logistic Regression****

Whenever the dependent variable is categorical, like when it has binary outputs, such as “true” and “false” or “yes” and “no,” logistic regression is used. Although regression models aim to understand correlations between data inputs, logistic regression is mainly utilized to solve binary classification problems. The Sigmoid function is used to convert predicted values to probabilities. The logistic regression hypothesis suggests that the cost function be limited to a value between 0 and 1.

### ****Evaluation Parameters****

A classification algorithm’s success is determined by its overall accuracy, recall, precision, F-measure, and false-positive rate (FPR). Overall accuracy, detection rate, and false-positive rate are used to evaluate IDS performance.

Ø Accuracy is the percentage of total records adequately identified as positive or negative.

Ø Recall Rate: The percentage of positive records accurately identified compared to the total number of negative records correctly classified as positive or wrongly classified as negative.

Ø The True Positive Rate (TPR) measures how many positive records are correctly identified.

Ø The False Positive Rate (FPR) measures how many negative records are correctly identified.

Ø The True Negative Rate (TNR) measures how many positive records are wrongly identified.

Ø The False Negative Rate (FNR) measures how many negative records are wrongly identified.

Ø Precision is the ratio of accurately classified positive documents to the total number of records labelled as positive.

Ø The F-Measure is the harmonic mean of recall and precision, and it provides a better indicator of an unbalanced dataset’s performance.

### ****Receiver Operating Characteristics Curve (ROC Curve)****

It’s a graphical representation of a binary classifier system’s diagnostic capabilities as its discriminating threshold changes. Starting in 1941, the approach was created for operators of military radar receivers, hence the name. Plotting the actual positive rate (TPR) against the false positive rate (FPR) at various threshold levels yields the ROC curve. It’s also known as a power plot as a function of the decision rule’s Type I Error. As a result, the ROC curve represents sensitivity or recall as a function of fall-out. In general, the ROC curve can be created by plotting the probability distributions for both detection and false alarm.

1. **Stage 5**

## ****Comparative Analysis****

For the training data sets of darknet network traffic, the five models (DT, KNN, LR, NB, and SVM) were compared in accuracy, precision, sensitivity, and specificity metrics. The average accuracy, precision, sensitivity, and specificity across all categories are shown in Table 1.

**Table 1: Comparative Analysis**

1. **Stage 6**

**Outcomes**

Table 1 shows that the accuracy metrics for Logistic Regression and GaussianNB models were reasonably constant across categories, while the other metrics (precision, sensitivity, and specificity) were inconsistent. Two models have (relatively) low metrics, indicating an under-fitting issue that isn’t worth optimizing further. The Decision Tree, Support Vector Machine, and KNN all performed well on all measures (an average of 99 per cent accuracy for both models). These models are pretty comparable and perform admirably in other areas.

**Figure 1: ROC curve before feature selection**

**Figure 2: ROC Curve after feature selection**

1. **Conclusion**

Traffic classification using machine learning algorithms was not previously considered relevant. With the rise of traffic encryption and anonymity services like Tor and VPN in the darknet, machine learning techniques for encrypted traffic classification should be viewed as one of the essential methods for identifying this traffic. We outlined machine learning categorization for darknet traffic networks in this research. We will start with some technical background on darknet and machine learning. To summarize, many aspects of the machine learning classification process could be researched and improved to reveal the truth about network privacy protection.

1. **References**

1) [http://205.174.165.80/CICDataset/CICDarknet2020/Dataset/](http://205.174.165.80/CICDataset/CICDarknet2020/Dataset/" \t "_blank)

2) [https://www.unb.ca/cic/datasets/darknet2020.html](https://www.unb.ca/cic/datasets/darknet2020.html" \t "_blank)

3) [https://en.wikipedia.org/wiki/Machine\_learning](https://en.wikipedia.org/wiki/Machine_learning" \t "_blank)

4) [https://www.ibm.com/cloud/learn/machine-learning#:~:text=Machine%20learning%20is%20a%20branch,learn%2C%20gradually%20improving%20its%20accuracy.&text=Machine%20learning%20is%20an%20important,growing%20field%20of%20data%20science.](https://www.ibm.com/cloud/learn/machine-learning" \l ":~:text=Machine learning is a branch,learn, gradually improving its accuracy.&text=Machine learning is an important,growing field of data science." \t "_blank)

5) [https://en.wikipedia.org/wiki/Darknet](https://en.wikipedia.org/wiki/Darknet" \t "_blank)

6) [https://en.wikipedia.org/wiki/Virtual\_private\_network](https://en.wikipedia.org/wiki/Virtual_private_network" \t "_blank)

7) [https://en.wikipedia.org/wiki/Tor\_(anonymity\_network)](https://en.wikipedia.org/wiki/Tor_(anonymity_network)" \t "_blank)

8) [https://en.wikipedia.org/wiki/Supervised\_learning](https://en.wikipedia.org/wiki/Supervised_learning" \t "_blank)

9) [https://en.wikipedia.org/wiki/Unsupervised\_learning](https://en.wikipedia.org/wiki/Unsupervised_learning" \t "_blank)

10) [https://www.analytixlabs.co.in/blog/data-acquisition/](https://www.analytixlabs.co.in/blog/data-acquisition/" \t "_blank)

11) [https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825](https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825" \t "_blank)

12) [https://en.wikipedia.org/wiki/Decision\_tree\_learning](https://en.wikipedia.org/wiki/Decision_tree_learning" \t "_blank)

13) [https://en.wikipedia.org/wiki/K-nearest\_neighbors\_algorithm](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm" \t "_blank)

14) [https://en.wikipedia.org/wiki/Logistic\_regression](https://en.wikipedia.org/wiki/Logistic_regression" \t "_blank)

15) [https://en.wikipedia.org/wiki/Support-vector\_machine](https://en.wikipedia.org/wiki/Support-vector_machine" \t "_blank)

16) [https://en.wikipedia.org/wiki/Naive\_Bayes\_classifier](https://en.wikipedia.org/wiki/Naive_Bayes_classifier" \t "_blank)

17) [https://machinelearningmastery.com/rfe-feature-selection-in-python/](https://machinelearningmastery.com/rfe-feature-selection-in-python/" \t "_blank)

18) [https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623](https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623" \t "_blank)

19) [https://builtin.com/data-science/regression-machine-learning#:~:text=Regression%20is%20a%20supervised%20machine,are%20variance%2C%20bias%20and%20error.](https://builtin.com/data-science/regression-machine-learning" \l ":~:text=Regression is a supervised machine,are variance, bias and error." \t "_blank)

20) [https://en.wikipedia.org/wiki/Feature\_selection](https://en.wikipedia.org/wiki/Feature_selection" \t "_blank)

21) [https://www.upgrad.com/blog/naive-bayes-classifier/#Advantages\_of\_Naive\_Bayes](https://www.upgrad.com/blog/naive-bayes-classifier/" \l "Advantages_of_Naive_Bayes" \t "_blank)

22) [https://en.wikipedia.org/wiki/Random\_forest](https://en.wikipedia.org/wiki/Random_forest" \t "_blank)

23) [https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226](https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226" \t "_blank)

24) [https://en.wikipedia.org/wiki/Receiver\_operating\_characteristic](https://en.wikipedia.org/wiki/Receiver_operating_characteristic" \t "_blank)