A MAJOR PROJECT REPORT ON

CYBER CRIMES ANALYSIS AND PREDICTION

Submitted to JNTUH in the partial fulfillment of the Academic Requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

ELECTRONICS AND COMMUNICATION ENGINEERING

BY

AMBATI BHARGAV 18QM1A0406

UNDER THE GUIDANCE OF

Dr. B. Vandana
ASSOCIATE PROFESSOR

KG REDDY
College of Engineering
& Technology
Engineering India's Changemakers

ELECTRONICS AND COMMUNICATION ENGINEERING

KG REDDY COLLEGE OF ENGINEERING AND TECHNOLOGY

(Accredited by NAAC, Approved by AICTE, New Delhi, Affiliated to JNTUH, Hyderabad)

Chilkur (Village), Moinabad (Mandal), R. R Dist, TS-501504

2021-22

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

KG REDDY COLLEGE OF ENGINEERING AND TECHNOLOGY
(Accredited by NAAC, Approved by AICTE, New Delhi, Affiliated to JNTUH, Hyderabad)
Chilkur (Village), Moinabad (Mandal), R. R Dist, TS-501504

KG REDDY
College of Engineering
& Technology
Engineering India's Changemakers

CERTIFICATE

This is to certify that the Project report on "**CYBER CRIMES ANALYSIS AND PREDICTION**" is a bonafide record work carried out by **AMBATI BHARGAV (18QM1A0406),** in partial fulfillment for the requirement for the award of degree of **BACHELOR OF TECHNOLOGY** in "**ELECTRONICS AND COMMUNICATION ENGINEERING", JNTUH,** Hyderabad during the year 2021-2022.

Internal Guide
DR. D. Vandana
ASSOCIATE PROFESSOR

Head of the department
Mr. M.N. NarsaiahM.Tech (P.hd)
HOD

External Examiner

ACKNOWLEDGEMENT

We wish to thank the Management, Sri. **Ln. K. Krishna Reddy, CHAIRMAN** of KG Reddy College of Engineering and Technology for providing the facilities to complete our project work.

We would also especially like to express our extreme gratitude and sincere thanks to our beloved **Director**, **Dr. Rohit Kandakatla**, KG Reddy College of Engineering and Technology, who encouraged and gave support to complete the project successfully.

At the very outset, we wish to place on record our sincere thanks and gratitude to our beloved head of the Institution, **Dr. R.S. Jahagirdar, Principal**, KG Reddy College of Engineering and Technology, who allowed us to complete this endeavor successfully.

We would like to express our sincere deep sense of gratitude and indebtedness to **Mr. M N Narsaiah, HOD**, Department of Electronics and Communication Engineering, for his significant suggestions and help in every respect to accomplish the project work. His persisting encouragement, everlasting patience and keen interest in discussions have benefited us to the extent that cannot be spanned by words. He has been the continuous source of inspiration for us throughout the work.

We would like to sincerely thank our Major **Project coordinator, Mrs. Dr B Vandana, Associate Professor** who stimulated many thoughts for this project and Staff-Members of Department of Electronics and Communication Engineering Department for their goodwill gestures towards us.

We would especially like to express our extreme gratitude and sincere thanks to our project guide **Mrs. Vandana , Associate Professor,** Department of Electronics and Communication Engineering for his enthusiastic and innovative guidance and support.

Last but not least we would also like to thank teaching and non-teaching staff and also all my family, friends without whose support this project work would remain unfulfilled.

AMBATI BHARGAV **18QM1A0406**

ABSTRACT

Cybercrime is a type of crime that involves the use of a computer and a network. As more people rely on technology to carry out their everyday activities, they are prone to getting victimized by cybercrime.

Cybercrime can range from various types such as extortion, cybercrime warfare, spamming, phishing, and various other attacks. Due to the increasing number of cybercrimes in India, the users need to be aware of their online activities and take precautionary measures.

Keywords: *Cyber Forensics, Malicious, Cybersecurity, Tor Browser, Machine learning*

Contents

CHAPTER 1

INTRODUCTION

1.0.What is ML

Machine learning is a subset of artificial intelligence that focuses on learning and analysing vast amounts of data. Through this technology, a computer algorithm can learn and interpret data collected by humans. It can then make recommendations and improve its decisions based on the data it has collected. Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

1.1.Why Is Machine Learning Important?

Data is the lifeblood of all business. Data-driven decisions increasingly make the difference between keeping up with competition or falling further behind. Machine learning can be the key to unlocking the value of corporate and customer data and enacting decisions that keep a company ahead of the competition.

1.2.Machine Learning Use Cases

Advancements in AI for applications like natural language processing (NLP) and computer vision (CV) are helping industries like financial services, healthcare, and automotive accelerate innovation, improve customer experience, and reduce costs. Machine learning has applications in all types of industries, including manufacturing, retail, healthcare and life sciences, travel and hospitality, financial services, and energy, feedstock, and utilities. Use cases include:

- Manufacturing. Predictive maintenance and condition monitoring
- Retail. Upselling and cross-channel marketing
- Healthcare and life sciences. Disease identification and risk satisfaction

- Travel and hospitality. Dynamic pricing
- Financial services. Risk analytics and regulation
- Energy. Energy demand and supply optimization

1.3.What does Prediction mean in Machine Learning?

The word “prediction” can be misleading. In some cases, it really does mean that you are predicting a future outcome, such as when you’re using machine learning to determine the next best action in a marketing campaign. Other times, though, the “prediction” has to do with, for example, whether or not a transaction that already occurred was fraudulent. In that case, the transaction already happened, but you’re making an educated guess about whether or not it was legitimate, allowing you to take the appropriate action.

1.4.Why are Predictions Important?

Machine learning model predictions allow businesses to make highly accurate guesses as to the likely outcomes of a question based on historical data, which can be about all kinds of things – customer churn likelihood, possible fraudulent activity, and more. These provide the business with insights that result in tangible business value. For example, if a model predicts a customer is likely to churn, the business can target them with specific communications and outreach that will prevent the loss of that customer.

1.5.Classification Model

The classification model is, in some ways, the simplest of the several types of predictive analytics models we’re going to cover. It puts data in categories based on what it learns from historical data. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. Outcomes are labels that can be applied to a dataset. A common job of machine learning algorithms is to recognize objects and being able to separate them into categories. This process is called classification, and it helps us segregate vast quantities of data into discrete values, i.e. distinct, like 0/1, True/False, or a pre-defined output label class.

1.5.1.Clustering Model

The clustering model sorts data into separate, nested smart groups based on similar attributes. If an ecommerce shoe company is looking to implement targeted marketing campaigns for their customers, they could go through the hundreds of thousands of records to create a tailored strategy for each individual. But is this the most efficient use of time? Probably not. Using the clustering model, they can quickly separate customers into similar groups based on common characteristics and devise strategies for each group at a larger scale.

Other use cases of this predictive modelling technique might include grouping loan applicants into “smart buckets” based on loan attributes, identifying areas in a city with a high volume of crime, and benchmarking SaaS customer data into groups to identify global patterns of use.

1.5.2.Forecast Model

One of the most widely used predictive analytics models, the forecast model deals in metric value prediction, estimating numeric value for new data based on learnings from historical data.

This model can be applied wherever historical numerical data is available. Scenarios include:

- A SaaS company can estimate how many customers they are likely to convert within a given week.
- A call centre can predict how many support calls they will receive per hour.
- A shoe store can calculate how much inventory they should keep on hand in order to meet demand during a particular sales period.

The forecast model also considers multiple input parameters.

1.5.3.Outliers Model

The outliers model is oriented around anomalous data entries within a dataset. It can identify anomalous figures either by themselves or in conjunction with other numbers and categories.

The outlier model is particularly useful for predictive analytics in retail and finance. For example, when identifying fraudulent transactions, the model can assess not only amount, but also location, time, purchase history and the nature of a purchase (i.e., a $1000 purchase on electronics is not as likely to be fraudulent as a purchase of the same amount on books or common utilities).

1.5.4.Time Series Model

The time series model comprises a sequence of data points captured, using time as the input parameter. It uses the last year of data to develop a numerical metric and predicts the next three to six weeks of data using that metric. Use cases for this model includes the number of daily calls received in the past three months, sales for the past 20 quarters, or the number of patients who showed up at a given hospital in the past six weeks. It is a potent means of understanding the way a singular metric is developing over time with a level of accuracy beyond simple averages. It also takes into account seasons of the year or events that could impact the metric.

If the owner of a salon wishes to predict how many people are likely to visit his business, he might turn to the crude method of averaging the total number of visitors over the past 90 days. However, growth is not always static or linear, and the time series model can better model exponential growth and better align the model to a company's trend. It can also forecast for multiple projects or multiple regions at the same time instead of just one at a time.

1.6.Common Predictive Algorithms

Overall, predictive analytics algorithms can be separated into two groups: machine learning and deep learning.

- **Machine learning** involves structural data that we see in a table. Algorithms for this comprise both linear and nonlinear varieties. Linear algorithms train more quickly, while nonlinear algorithms are better optimised for the problems they are likely to face (which are often nonlinear).
- **Deep learning** is a subset of machine learning that is more popular to deal with audio, video, text, and images.

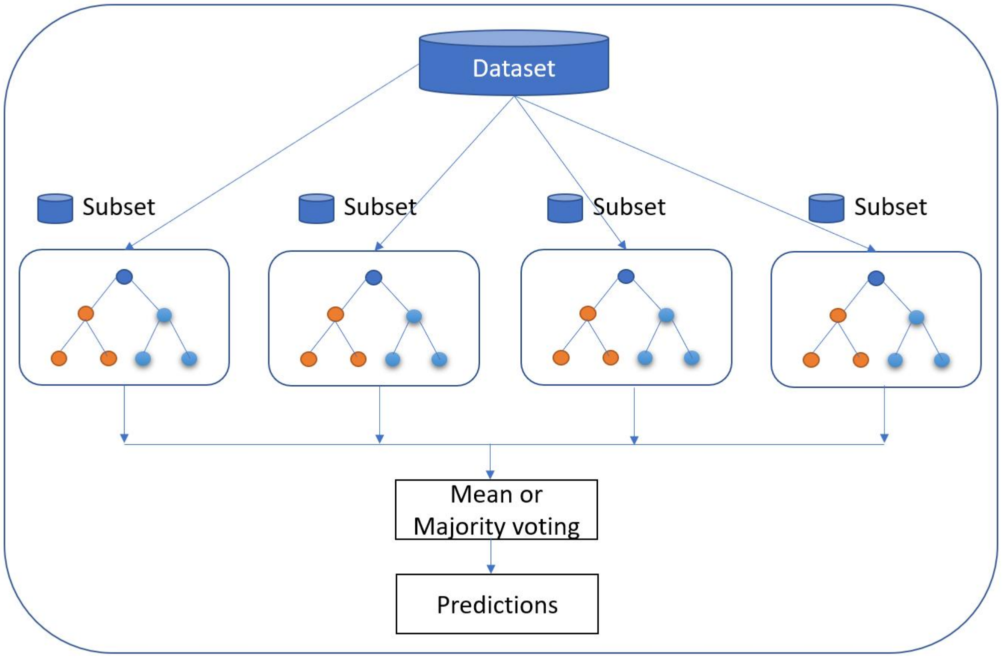
With machine learning predictive modelling, there are several different algorithms that can be applied. Below are some of the most common algorithms that are being used to power the predictive analytics models described above.

1.6.1.Random Forest

Random Forest is perhaps the most popular classification algorithm, capable of both classification and regression. It can accurately classify large volumes of data.

The name “Random Forest” is derived from the fact that the algorithm is a combination of decision trees. Each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the “forest.” Each one is grown to the largest extent possible.

Predictive analytics algorithms try to achieve the lowest error possible by either using “boosting” (a technique which adjusts the weight of an observation based on the last classification) or “bagging” (which creates subsets of data from training samples, chosen randomly with replacement). Random Forest uses bagging. If you have a lot of sample data, instead of training with all of them, you can take a subset and train on that, and take another subset and train on that (overlap is allowed). All of this can be done in parallel. Multiple samples are taken from your data to create an average.



While individual trees might be “weak learners,” the principle of Random Forest is that together they can comprise a single “strong learner.”

The popularity of the Random Forest model is explained by its various advantages:

- Accurate and efficient when running on large databases
- Multiple trees reduce the variance and bias of a smaller set or single tree
- Resistant to overfitting
- Can handle thousands of input variables without variable deletion
- Can estimate what variables are important in classification
- Provides effective methods for estimating missing data
- Maintains accuracy when a large proportion of the data is missing

1.6.2.Generalised Linear Model (GLM) for Two Values

The Generalised Linear Model (GLM) is a more complex variant of the General Linear Model. It takes the latter model’s comparison of the effects of multiple variables on continuous variables before drawing from an array of different distributions to find the “best fit” model.

Let’s say you are interested in learning customer purchase behaviour for winter coats. A regular linear regression might reveal that for every negative degree difference in temperature, an additional 300 winter coats are purchased. While it seems logical that another 2,100 coats might be sold if the temperature goes from 9 degrees to 3, it seems less logical that if it goes down to -20, we’ll see the number increase to the exact same degree.

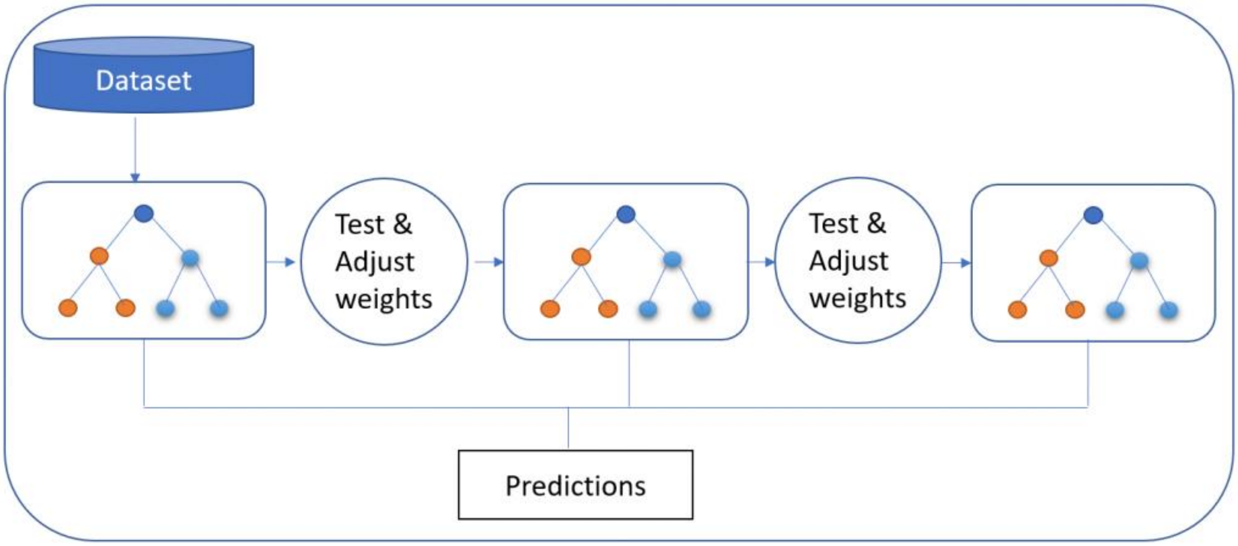
The Generalised Linear Model would narrow down the list of variables, likely suggesting that there is an increase in sales beyond a certain temperature and a decrease or flattening in sales once another temperature is reached.

The advantage of this algorithm is that it trains very quickly. The response variable can have any form of exponential distribution type. The Generalised Linear Model is also able to deal with categorical predictors, while being relatively straightforward to interpret. On top of this, it provides a clear understanding of how each of the predictors is influencing the outcome, and is fairly resistant to overfitting. However, it requires relatively large data sets and is susceptible to outliers

1.6.3.Gradient Boosted Model (GBM)

The Gradient Boosted Model produces a prediction model composed of an ensemble of decision trees (each one of them a “weak learner,” as was the case with Random Forest), before generalising. As its name suggests, it uses the “boosted” machine learning technique, as opposed to the bagging used by Random Forest. It is used for the classification model.

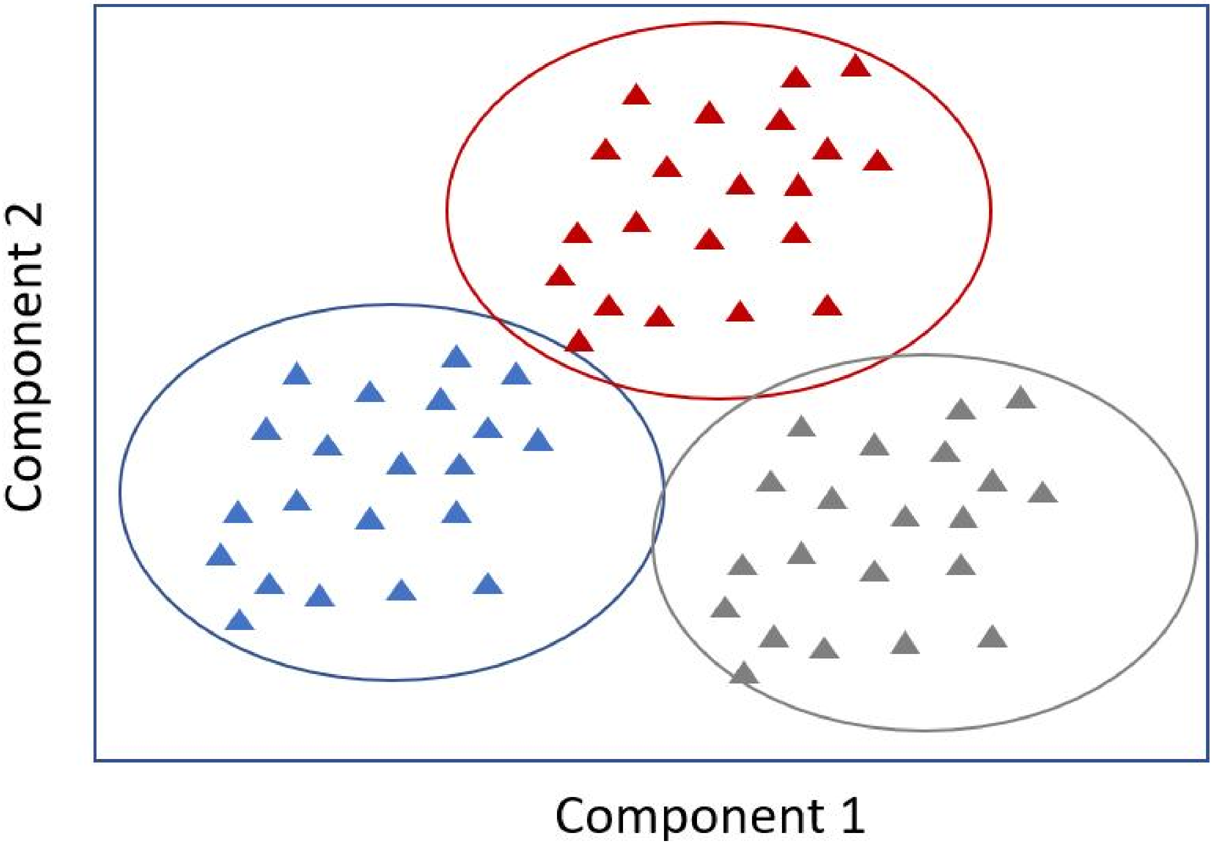
The distinguishing characteristic of the GBM is that it builds its trees one tree at a time. Each new tree helps to correct errors made by the previously trained tree —unlike in the Random Forest model, in which the trees bear no relation. It is very often used in machine-learned ranking, as in the search engines Yahoo and Yandex.



Via the GBM approach, data is more expressive, and benchmarked results show that the GBM method is preferable in terms of the overall thoroughness of the data. However, as it builds each tree sequentially, it also takes longer. That said, its slower performance is considered to lead to better generalisation.

1.6.4.K-Means

A highly popular, high-speed algorithm, K-means involves placing unlabeled data points in separate groups based on similarities. This algorithm is used for the clustering model. For example, Tom and Rebecca are in group one and John and Henry are in group two. Tom and Rebecca have very similar characteristics but Rebecca and John have very different characteristics. K-means tries to figure out what the common characteristics are for individuals and groups them together. This is particularly helpful when you have a large data set and are looking to implement a personalised plan—this is very difficult to do with one million people.



In the context of predictive analytics for healthcare, a sample size of patients might be placed into five separate clusters by the algorithm. One particular group shares multiple characteristics: they don't exercise, they have an increasing hospital attendance record (three times one year and then ten times the next year), and they are all at risk for diabetes. Based on the similarities, we can proactively recommend a diet and exercise plan for this group.

1.6.5. Linear Regression

Linear regression is perhaps one of the most well-known and well-understood algorithms in statistics and machine learning.

Predictive modelling is primarily concerned with minimising the error of a model or making the most accurate predictions possible, at the expense of explainability. We will borrow, reuse and steal algorithms from many different fields, including statistics and use them towards these ends.

The representation of linear regression is an equation that describes a line that best fits the relationship between the input variables (x) and the output variables (y), by finding specific weightings for the input variables called coefficients (B).
We will predict y given the input x and the goal of the linear regression learning algorithm is to find the values for the coefficients B0 and B1.

Different techniques can be used to learn the linear regression model from data, such as a linear algebra solution for ordinary least squares and gradient descent optimization.

Linear regression has been around for more than 200 years and has been extensively studied. Some good rules of thumb when using this technique are to remove variables that are very similar (correlated) and to remove noise from your data, if possible. It is a fast and simple technique and a good first algorithm to try.

1.6.6.Logistic Regression

Logistic regression is another technique borrowed by machine learning from the field of statistics. It is the go-to method for binary classification problems (problems with two class values).

Logistic regression is like linear regression in that the goal is to find the values for the coefficients that weight each input variable. Unlike linear regression, the prediction for the output is transformed using a nonlinear function called the logistic function.

The logistic function looks like a big S and will transform any value into the range 0 to 1. This is useful because we can apply a rule to the output of the logistic function to snap values to 0 and 1 (e.g. IF less than 0.5 then output 1) and predict a class value.

Because of the way that the model is learned, the predictions made by logistic regression can also be used as the probability of a given data instance belonging to class 0 or class 1. This can be useful for problems where you need to give more rationale for a prediction.

Like linear regression, logistic regression does work better when you remove attributes that are unrelated to the output variable as well as attributes that are very similar (correlated) to each other. It’s a fast model to learn and effective on binary classification problems.

1.6.7.Linear Discriminant Analysis

Logistic Regression is a classification algorithm traditionally limited to only two-class classification problems. If you have more than two classes then the Linear Discriminant Analysis algorithm is the preferred linear classification technique.

The representation of LDA is pretty straight forward. It consists of statistical properties of your data, calculated for each class.

Predictions are made by calculating a discriminant value for each class and making a prediction for the class with the largest value. The technique assumes that the data has a Gaussian distribution (bell curve), so it is a good idea to remove outliers from your data beforehand. It’s a simple and powerful method for classification predictive modelling problems.

1.6.8.Classification and Regression Trees

Decision Trees are an important type of algorithm for predictive modelling machine learning.

The representation of the decision tree model is a binary tree. This is your binary tree from algorithms and data structures, nothing too fancy. Each node represents a single input variable (x) and a split point on that variable (assuming the variable is numeric).

The leaf nodes of the tree contain an output variable (y) which is used to make a prediction. Predictions are made by walking the splits of the tree until arriving at a leaf node and output the class value at that leaf node.

Trees are fast to learn and very fast for making predictions. They are also often accurate for a broad range of problems and do not require any special preparation for your data.

1.6.9.K-Nearest Neighbours

The KNN algorithm is very simple and very effective. The model representation for KNN is the entire training dataset. Simple right?

Predictions are made for a new data point by searching through the entire training set for the K most similar instances (the neighbours) and summarising the output variable for those K instances. For regression problems, this might be the mean output variable, for classification problems this might be the mode (or most common) class value.

The trick is in how to determine the similarity between the data instances. The simplest technique if your attributes are all of the same scale (all in inches for example) is to use the Euclidean distance, a number you can calculate directly based on the differences between each input variable.

KNN can require a lot of memory or space to store all of the data, but only performs a calculation (or learn) when a prediction is needed, just in time. You can also update and curate your training instances over time to keep predictions accurate.

The idea of distance or closeness can break down in very high dimensions (lots of input variables) which can negatively affect the performance of the algorithm on your problem. This is called the curse of dimensionality. It suggests you only use those input variables that are most relevant to predicting the output variable.

1.6.10.Learning Vector Quantization

A downside of K-Nearest Neighbours is that you need to hang on to your entire training dataset. The Learning Vector Quantization algorithm (or LVQ for short) is an artificial neural network algorithm that allows you to choose how many training instances to hang onto and learns exactly what those instances should look like.

The representation for LVQ is a collection of codebook vectors. These are selected randomly in the beginning and adapted to best summarise the training dataset over a number of iterations of the learning algorithm. After learning, the codebook vectors can be used to make predictions just like K-Nearest Neighbours. The most similar neighbour (best matching codebook vector) is found by calculating the distance between each codebook vector and the new data instance. The class value or (real value in the case of regression) for the best matching unit is then returned as the prediction. Best results are achieved if you rescale your data to have the same range, such as between 0 and 1.

If you discover that KNN gives good results on your dataset try using LVQ to reduce the memory requirements of storing the entire training dataset.

2.0.What is Cyber Security

Cyber security is the practice of defending computers, servers, mobile devices, electronic systems, networks, and data from malicious attacks. It's also known as information technology security or electronic information security. The term applies in a variety of contexts, from business to mobile computing, and can be divided into a few common categories.

- **Network security** is the practice of securing a computer network from intruders, whether targeted attackers or opportunistic malware. Network security is a broad term that covers a multitude of technologies, devices and processes. In its simplest term, it is a set of rules and configurations designed to protect the integrity, confidentiality and accessibility of computer networks and data using both software and hardware technologies. Network security consists of the policies, processes and practices adopted to prevent, detect and monitor unauthorised access, misuse, modification, or denial of a computer network and network-accessible resources.

- **Application security** focuses on keeping software and devices free of threats. A compromised application could provide access to the data it's designed to protect. Successful security begins in the design stage, well before a program or device is deployed. Application security is the process of developing, adding, and testing security features within applications to prevent security vulnerabilities against threats such as unauthorised access and modification. Application security includes all tasks that introduce a secure software development life cycle to development teams. Its final goal is to improve security practices and, through that, to find, fix and preferably prevent security issues within applications.

- **Information security** protects the integrity and privacy of data, both in storage and in transit. Information Security refers to the processes and methodologies which are designed and implemented to protect print, electronic, or any other form of confidential, private and sensitive information or data from unauthorised access, use, misuse, disclosure, destruction, modification, or disruption. Information security, sometimes shortened to InfoSec, is the practice of protecting information by mitigating information risks. It is part of information risk management.

- **Operational security** includes the processes and decisions for handling and protecting data assets. The permissions users have when accessing a network and the procedures that determine how and where data may be stored or shared all fall under this umbrella. Operations security (OPSEC) is a process that identifies critical information to determine if friendly actions can be observed by enemy intelligence, determines if information obtained by adversaries could be interpreted to be useful to them, and then executes selected measures that eliminate or reduce adversary exploitation of friendly critical information. Operational security (OPSEC) began as a military process but is now commonly used in business as a risk management strategy for protecting data from unintentional leaks. This article explains why OPSEC is important for business security and offers best practice tips for implementation, such as keeping your system up to date with comprehensive antivirus.

- **Disaster recovery and business continuity** define how an organisation responds to a cyber-security incident or any other event that causes the loss of operations or data. Disaster recovery policies dictate how the organisation restores its operations and information to return to the same operating capacity as before the event. Business continuity is the plan the organisation falls back on while trying to operate without certain resources. Business continuity outlines exactly how a business will proceed during and following a disaster. It may provide contingency plans, outlining how the business will continue to operate even if it has to move to an alternate location. Business continuity planning may also take into account smaller interruptions or minor disasters, such as extended power outages. Disaster recovery refers to the plans a business puts into place for responding to a catastrophic event, such as a natural disaster, fire, act of terror, active shooter or cybercrime. Disaster recovery involves the measures a business takes to respond to an event and return to safe, normal operation as quickly as possible.

- **End-user education** addresses the most unpredictable cyber-security factor: people. Anyone can accidentally introduce a virus to an otherwise secure system by failing to follow good security practices. Teaching users to delete suspicious email attachments, not plug in unidentified USB drives, and various other important lessons is vital for the security of any organisation. End-user education is building awareness among employees by equipping them with the necessary tools and skills required to protect themselves and the company data from loss or attack.

2.1.Types of cyber threats

The threats countered by cyber-security are three-fold:

1. Cybercrime includes single actors or groups targeting systems for financial gain or to cause disruption.

2. Cyber-attack often involves politically motivated information gathering.

3. Cyberterrorism is intended to undermine electronic systems to cause panic or fear.

2.1.1.Malware

Malware means malicious software. One of the most common cyber threats, malware is software that a cybercriminal or hacker has created to disrupt or damage a legitimate user's computer. Often spread via an unsolicited email attachment or legitimate-looking download, malware may be used by cybercriminals to make money or in politically motivated cyber-attacks.

2.1.2.Virus

A self-replicating program that attaches itself to clean file and spreads throughout a computer system, infecting files with malicious code.A virus is a computer code or programme that has the capability of damaging or deleting your computer data.Computer viruses have a proclivity for rapidly making duplicate copies of themselves, spreading them throughout all folders and causing data loss on your computer system.A computer virus is a harmful software programme (also known as "malware") that repeats itself by altering other computer programmes and injecting its own code when it infects your system.

2.1.3.Trojans

A type of malware that is disguised as legitimate software. Cybercriminals trick users into uploading Trojans onto their computer where they cause damage or collect data.A Trojan horse, or Trojan, is a type of malicious code or software that looks legitimate but can take control of your computer. A Trojan is designed to damage, disrupt, steal, or in general inflict some other harmful action on your data or network. A Trojan acts like a bona fide application or file to trick you.

2.1.4.Spyware

A program that secretly records what a user does, so that cybercriminals can make use of this information. For example, spyware could capture credit card details.Spyware is any software that uninstalls itself on your computer and starts covertly monitoring your online behaviour without your knowledge or permission. Spyware is a kind of malware that secretly gathers information about a person or organisation and relays this data to other parties.

2.1.5.Ransomware

Malware which locks down a user's files and data, with the threat of erasing it unless a ransom is paid.Ransomware is malware that employs encryption to hold a victim's information at ransom. A user or organisation's critical data is encrypted so that they cannot access files, databases, or applications. A ransom is then demanded to provide access.

2.1.6.Adware

Advertising software which can be used to spread malware.Adware (or advertising software) is the term used for various pop-up advertisements that show up on your computer or mobile device. Adware has the potential to become malicious and harm your device by slowing it down, hijacking your browser and installing viruses and/or spyware.

2.1.7.Botnets

Networks of malware infected computers which cybercriminals use to perform tasks online without the user’s permission.A botnet (derived from 'robot network') is a large group of malware-infected internet-connected devices and computers controlled by a single operator. Attackers use these compromised devices to launch large-scale attacks to disrupt services, steal credentials and gain unauthorised access to critical systems.

2.1.8.SQL injection

An SQL (structured language query) injection is a type of cyber-attack used to take control of and steal data from a database. Cybercriminals exploit vulnerabilities in data-driven applications to insert malicious code into a database via a malicious SQL statement. This gives them access to the sensitive information contained in the database.

SQL injection (SQLi) is a type of cyberattack against web applications that use SQL databases such as IBM Db2, Oracle, MySQL, and MariaDB. As the name suggests, the attack involves the injection of malicious SQL statements to interfere with the queries sent by a web application to its database.

2.1.9.Phishing

Phishing is when cybercriminals target victims with emails that appear to be from a legitimate company asking for sensitive information. Phishing attacks are often used to dupe people into handing over credit card data and other personal information.

Phishing attacks are the practice of sending fraudulent communications that appear to come from a reputable source. It is usually done through email. The goal is to steal sensitive data like credit card and login information, or to install malware on the victim's machine.

2.1.10.Man-in-the-middle attack

A man-in-the-middle attack is a type of cyber threat where a cybercriminal intercepts communication between two individuals in order to steal data. For example, on an unsecure WiFi network, an attacker could intercept data being passed from the victim's device and the network.

A man in the middle (MITM) attack is a general term for when a perpetrator positions himself in a conversation between a user and an application—either to eavesdrop or to impersonate one of the parties, making it appear as if a normal exchange of information is underway.

A man-in-the-middle attack is a type of cyberattack in which an attacker eavesdrops on a conversation between two targets. The attacker may try to "listen" to a conversation between two people, two systems, or a person and a system.

2.1.11.Denial-of-service attack

A denial-of-service attack is where cybercriminals prevent a computer system from fulfilling legitimate requests by overwhelming the networks and servers with traffic. This renders the system unusable, preventing an organisation from carrying out vital functions.

A Denial-of-Service (DoS) attack is an attack meant to shut down a machine or network, making it inaccessible to its intended users. DoS attacks accomplish this by flooding the target with traffic, or sending it information that triggers a crash

3.0.Latest cyber threats

What are the latest cyber threats that individuals and organisations need to guard against? Here are some of the most recent cyber threats that the U.K., U.S., and Australian governments have reported on.

3.1.Dridex malware

In December 2019, the U.S. The Department of Justice (DoJ) charged the leader of an organised cyber-criminal group for their part in a global Dridex malware attack. This malicious campaign affected the public, government, infrastructure and business worldwide.

Dridex is a financial trojan with a range of capabilities. Affecting victims since 2014, it infects computers though phishing emails or existing malware. Capable of stealing passwords, banking details and personal data which can be used in fraudulent transactions, it has caused massive financial losses amounting to hundreds of millions.

In response to the Dridex attacks, the U.K.'s National Cyber Security Centre advises the public to "ensure devices are patched, antivirus is turned on and up to date and files are backed up".

3.2.Romance scams

In February 2020, the FBI warned U.S. citizens to be aware of confidence fraud that cybercriminals commit using dating sites, chat rooms and apps. Perpetrators take advantage of people seeking new partners, duping victims into giving away personal data.

The FBI reports that romance cyber threats affected 114 victims in New Mexico in 2019, with financial losses amounting to $1.6 million.

3.3.Emotet malware

In late 2019, The Australian Cyber Security Centre warned national organisations about a widespread global cyber threat from Emotet malware.

Emotet is a sophisticated trojan that can steal data and also load other malware. Emotet thrives on unsophisticated passwords, a reminder of the importance of creating a secure password to guard against cyber threats.

4.0.Cyber safety tips - protect yourself against cyberattacks

1. Update your software and operating system:This means you benefit from the latest security patches.

2. Use anti-virus software:Security solutions like Kaspersky Total Security will detect and remove threats. Keep your software updated for the best level of protection.

3. Use strong passwords:Ensure your passwords are not easily guessable.

4. Do not open email attachments from unknown senders:These could be infected with malware.

5. Do not click on links in emails from unknown senders or unfamiliar websites:This is a common way that malware is spread.

6. Avoid using unsecured WiFi networks in public places:Unsecure networks leave you vulnerable to man-in-the-middle attacks.

5.0.What is Cybercrime ?

Cybercrime is a crime that involves a computer and a network.The computer may have been used in the commission of a crime, or it may be the target.There are many privacy concerns surrounding Cybercrime when confidential information is intercepted or disclosed, lawfully or otherwise. Internationally, both governmental and non-state actors engage in cybercrimes, including espionage, financial theft, and other cross-border crimes. Cybercrimes crossing international borders and involving the actions of at least one nation-state are sometimes referred to as cyberwarfare.

5.1.Cyberterrorism

Government officials and information technology security specialists have documented a significant increase in Internet problems and server scams since early 2001. There is a growing concern among government agencies such as the Federal Bureau of Investigation (FBI) and the Central Intelligence Agency (CIA) that such intrusions are part of an organised effort by cyberterrorist foreign intelligence services, or other groups to map potential security holes in critical systems.A cyberterrorist is someone who intimidates or coerces a government or an organisation to advance his or her political or social objectives by launching a computer-based attack against computers, networks, or the information stored on them.

5.2.Cyber Extortion

Cyber Extortion occurs when a website, e-mail server, or computer system is subjected to or threatened with repeated denial of service or other attacks by malicious hackers. These hackers demand money in return for promising to stop the attacks and to offer "protection". According to the Federal Bureau of Investigation, cybercrime extortionists are increasingly attacking corporate websites and networks, crippling their ability to operate and demanding payments to restore their service. More than 20 cases are reported each month to the FBI and many go unreported in order to keep the victim's name out of the public domain. Perpetrators typically use a distributed denial-of-service attack.However, other cyber extortion techniques exist such as doxing and bug poaching.

5.3.Cyber Warfare

Cyber warfare is usually defined as a cyber attack or series of attacks that target a country. It has the potential to wreak havoc on government and civilian infrastructure and disrupt critical systems, resulting in damage to the state and even loss of life.Cyberwarfare is the use of digital attacks against an enemy state, causing comparable harm to actual warfare and/or disrupting the vital computer systems. There is significant debate among experts regarding the definition of cyberwarfare, and even if such a thing exists.

CHAPTER 2

Literature survey

In a deeply connected world, like the one we are facing nowa-days, hackers continuously try to find new targets and develop new tools to break through cyberdefenses. Moreover, the lack

privacy and security of the new upcoming technologies and the users' lack of awareness pose a real threat to our personal life. In the following, we present some works that face cyber- security and discuss the countermeasures available today.

IT security includes cyber security as a subset. Cyber security protects the digital data on your networks, computers, and devices from unauthorised access, attack, and destruction. While IT security protects both physical and digital data, cyber security protects the digital data on your networks, computers, and devices from unauthorised access, attack, and destruction. In this section, we'll talk about how cyber security works. Brenner describes the first method for identifying measures for assessing crime that originates in cyberspace. Although she acknowledges that designing metrics and scales for cybercrime is extremely difficult, due to apprehension, scale, and evidence issues, she proposes a simple taxonomy of harms consisting of three main types, namely individual, systemic. Kshetri attempts to define a cost-benefit calculus using a similar methodology to Laube but he focuses on the attacker's point of view He

describes the characteristics of cybercriminals, cybercrime victims, and law enforcement officials, arguing that when these three types of entities interact, they create a vicious cycle of cybercrime. He develops a calculation that analyses an attacker's rewards and costs, as well as arguments for whether or not a cybercrime will occur. With the use of interruption detection, this paper uses machine learning and information digging approaches for digital inquiry. The crime triangle is sometimes used to define cybercrime, which states that for a cybercrime to occur, three variables must exist: a victim, a motive, and an opportunity. The victim is the person who will be attacked, the motive is what motivates the criminal to perform the crime, and the opportunity is when the crime will be committed (e.g., it can be an innate vulnerability in the system or an unprotected device). While today's attacks are more sophisticated and targeted to specific victims based on the attacker's goal, such as financial gain, espionage, coercion, or retribution, opportunistic untargeted attacks are still common.

"Opportunistic attacks" are defined as attacks that target victims based on their vulnerability to attack. Camellia is a 128-bit block cipher proposed in this publication. Camellia supports 128-bit block sizes and 128-, 192-, and 256-bit keys, i.e. the Advanced Encryption Standard's interface specifications (AES). Camellia is notable for its efficiency on both software and hardware platforms, in addition to its high level of security. Camellia has been proven to give good security against both differential and linear cryptanalysis. Camellia has at least comparable encryption speed in software and hardware to the AES finalists, namely MARS, RC6, Rijndael, Serpent, and Twofish.

The author of this utilised machine learning and sentiment analysis to cyber security in order to

establish a way for detecting cyber risks that were previously undetectable by traditional technologies. Greenfield provides a methodology for experimentally assessing harm that includes a number of processes. Functional integrity, material support and amenity, freedom from humiliation, privacy or autonomy, and reputation are the five fundamental dimensions where injury might appear. They also establish five levels of scale for various sorts of harm and investigate the cascading nature of harm by looking at real-world crimes that have generated significant societal impact. Grant et al. coined the term "cyberspace cartography" and applied the concept of "cyber-geography" to military operations. They also suggest that their ontology might be used in research to help solve the attribution problem of being unable to quickly identify hostile actors in cyberspace . Chertoff describes the state of Internet jurisdiction law and the problem of assigning legal authority to a particular forum when a suit traverses multiple states. They present four possible formulations for defining the controlling jurisdiction in situations

in a clear and equitable manner. These regulations are based on either the citizenship of the offending information, data, or system's subject, the location where the harm occurred, the citizenship of the data creator, or the citizenship of the data holder or custodian. A high-quality standalone literature review, according to Mathieu and Guy provides reliable information and insights into previous research, allowing other researchers to seek new

directions on similar issues of interest. Furthermore, the findings of this study can be utilised as references in related fields or as a basis for future research. Lin compares nuclear and cyber technology and regulation, outlining a slew of contrasts, as well as a few parallels, between the potential difficulties that these two technologies bring, which he categorises as strategy,
operations, acquisition, and arms control. The author of the paper claimed that online security attacks have been carried out by hacker-activist organisations with the goal of causing harm to web services in a specific context. On Twitter content, the author demonstrated a sentiment analysis method. The author's strategy was based on a daily collection of tweets from users who utilise the platform to share their opinions on pertinent subjects and to deliver content connected to web security assaults. The information was transformed into data that could be statistically examined to determine whether an attack was likely or not. The latter was accomplished by examining the aggregate sentiment of users and hacktivist groups in response to a worldwide incident. Edwards use a publicly available dataset of data breaches to uncover trends in
data breaches using a Bayesian Generalised Linear Model. They conclude that while the amount and frequency of data breaches have remained consistent in recent years, their impact is increasing as threat actors improve their ability to monetize personal information and the quantity of electronic financial transactions grows. A concentrated literature analysis of machine learning and data mining methods for cyber analytics in support of intrusion detection was reported in a survey study.

Van Slyke created a taxonomy of harms for white-collar crimes by focusing on the victimisation aspect of these crimes. They look at a number of white-collar offences and the costs associated with them. They combine desktop research with victim surveys, focusing on the long-term consequences of damages in specific persons.

Tounsi and Rais (2018) give an overview of open-source/free threat intelligence technologies and compare them to AlliaCERT TI.2 features. Their research discovered that sharing threat intelligence quickly, as any business would desire in order to collaborate, is insufficient to prevent targeted attacks. Furthermore, firms that share personal information must have a high level of trust. Another issue is determining how much data must be shared in order to prevent attacks and cooperate, as well as in what format. Tounsi et al. provide their analysis in order to determine whether the standard is superior. Finally, the research examines the best threat intelligence systems, dividing them into two categories: those that prioritise standardisation and autonomous analyses, and those that prioritise high-speed requirements.

Furthermore, although Tounsi focuses on the best approach to maintain trust among enterprises while also sharing information about cyber dangers, the authors of Toch (2018) focus on the type of data that is necessary from cybersecurity systems that are supposed to protect our privacy from prying eyes. The article's suggested taxonomy covers the hazards of several types of cybersecurity technologies that are linked to a certain hack. Almost all cyber-security technological categories, according to the taxonomy, require some level of access to personally sensitive data. This finding can help not only in deciding which technique to use, but also in building more privacy-aware cyber-security solutions with little or no compromise in terms of their ability to safeguard data from cyber-attacks.

The research mentioned above attempted to examine systems and best practises for reducing cyber hazards. We have a study by Chang et al. (2013) on the most up-to-date web-based malware threats and how to defend against them. The paper begins with a review of the assault model and the vulnerabilities that enable these attacks, then assesses the present state of the malware problem and looks into security techniques. As a result, the report identifies three types of techniques for analysing, identifying, and defending against web-based malware: (1) employing virtual computers to create honeypots; (2) using code analysis and testing approaches to uncover Web application vulnerabilities; and (3) creating reputation-based blacklists. Each category has its own set of benefits and drawbacks, as well as how different approaches complement one another and how they might be combined.

In Xu et al. (2013), the authors take a different technique than the previous ones, analysing network-layer traffic and application-layer website contents simultaneously in order to detect malicious web applications at run-time. There are two types of methodologies for detecting fraudulent websites that are currently available: static and dynamic approaches. The first method examines URLs and contents, while the second employs client honeypots to examine run-time behaviour. Experiments with this method revealed that cross-layer detection was capable of achieving the same detection efficiency as the dynamic method. However, it was far faster than the dynamic one.

Another useful resource for understanding the growing worry about cybersecurity is the United Nations Office on Substances and Crime's Guidance Note3, which is a global leader in the battle against illicit drugs and international crime. The goal of this guideline note is to provide a complete overview of today's most common cybersecurity risks. Cybercrime activities such as online radicalization and the illicit sale of pharmaceutical solutions are displayed and described in depth to illustrate how UNODC may provide technical help to solve cybercrime concerns at both the regional and national levels. While there is a significant increase in interest in cybersecurity risks in the scientific literature, there are several blogs and web pages warning about new future cybersecurity threats on the web side.

CHAPTER 3

Methodology

Static analysis (or code analysis) and dynamic analysis are the two most prevalent malware analysis approaches employed by malware analysts (or behaviour analysis). These two methods enable analysts to swiftly and thoroughly comprehend the threats and goals of a specific sample malware.

Static analysis necessitates a solid understanding of programming and the x86 assembly language paradigm. You don't have to run the malware during the static analysis. Malware samples' source code is rarely published. You must first disassemble and decompile, and then reverse engineer the low-level assembly code. Because static analysis is safer than dynamic analysis, most malware analysts undertake it first. The intricacy of modern malware makes static analysis difficult, and some malware uses anti-debugging techniques to prevent malware analysts from examining the code.

In malware analysis, dynamic analysis (behaviour analysis) is a technique that executes the malware and monitors its activity. It also keeps track of any changes that occur during the malware's execution. Infecting a computer with malware that has been downloaded from the internet can be extremely dangerous. Malware infestation on your system might result in file deletion, registry changes, file alteration, data theft, and other issues. You'll need a secure environment to conduct malware analysis, and the network shouldn't be connected to any production networks.

Machine learning algorithms employ computational approaches to "learn" information directly from data rather than depending on a model. As the number of samples available for learning grows, the algorithms adaptively enhance their performance.

Detecting Malicious URLs using Machine Learning

The malicious URLs can be detected using the lexical features along with tokenization of the URL strings. I aim to build a basic binary classifier that would help classify the URLs as malicious or benign. Steps followed in building the machine learning classifier:

- Data Preprocessing
- Data Visualisation
- Algorithm and Predication

Data Preprocessing

In Machine Learning, data preprocessing refers to the process of cleaning and organising raw data in order to make it appropriate for creating and training Machine Learning models.Preprocessing data is necessary to ensure high-quality results. Data cleansing, data integration, data reduction, and data transformation are the four stages of data preprocessing, which make the process easier.

Data Visualisation

Data visualisation in machine learning is essential for understanding how data is used in a machine learning model and for assessing it. Facets is a free, open-source Python tool that may be used to efficiently visualise and analyse data.

Algorithm and Predication

In machine learning, an "algorithm" is a technique that is conducted on data to build a "model." Pattern recognition is performed by machine learning algorithms. Algorithms can either "learn" from data or "fit" to a dataset. There are numerous machine learning algorithms to choose from.

- Logistic Regression
- Decision Trees
- Random Forest

❖ Logistic Regression

A statistical analysis approach for predicting a data value based on past observations of a data collection is known as logistic regression. By examining the connection between one or more existing independent variables, a logistic regression model may predict a dependent data variable. It's a type of statistical software that estimates probabilities using a logistic regression equation to analyse the relationship between a dependent variable and one or more independent variables. This form of analysis can assist you in predicting the probability of an event or a decision occurring.

Logistic regression uses an equation as the representation, very much like linear regression. Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modeled is a binary values (0 or 1) rather than a numeric value.

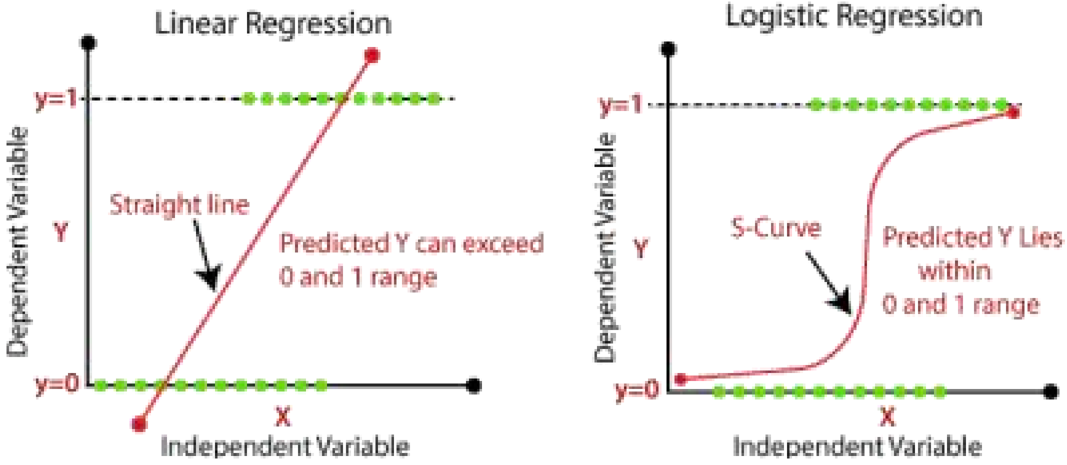


Fig: Linear Regression Variable Graph

Below is an example logistic regression equation:

y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

❖ Decision Trees

The Decision Tree is a supervised learning technique that may be used to solve both classification and regression issues, however it is most commonly employed to solve classification problems. Internal nodes represent dataset properties, branches represent decision rules, and each leaf node reflects the conclusion. The Decision Node and the Leaf Node are the two nodes of a Decision Tree. Decision nodes are used to make any decision and have several branches, whereas Leaf nodes are the results of such decisions and have no additional branches. The decisions or tests are based on the characteristics of the given dataset.

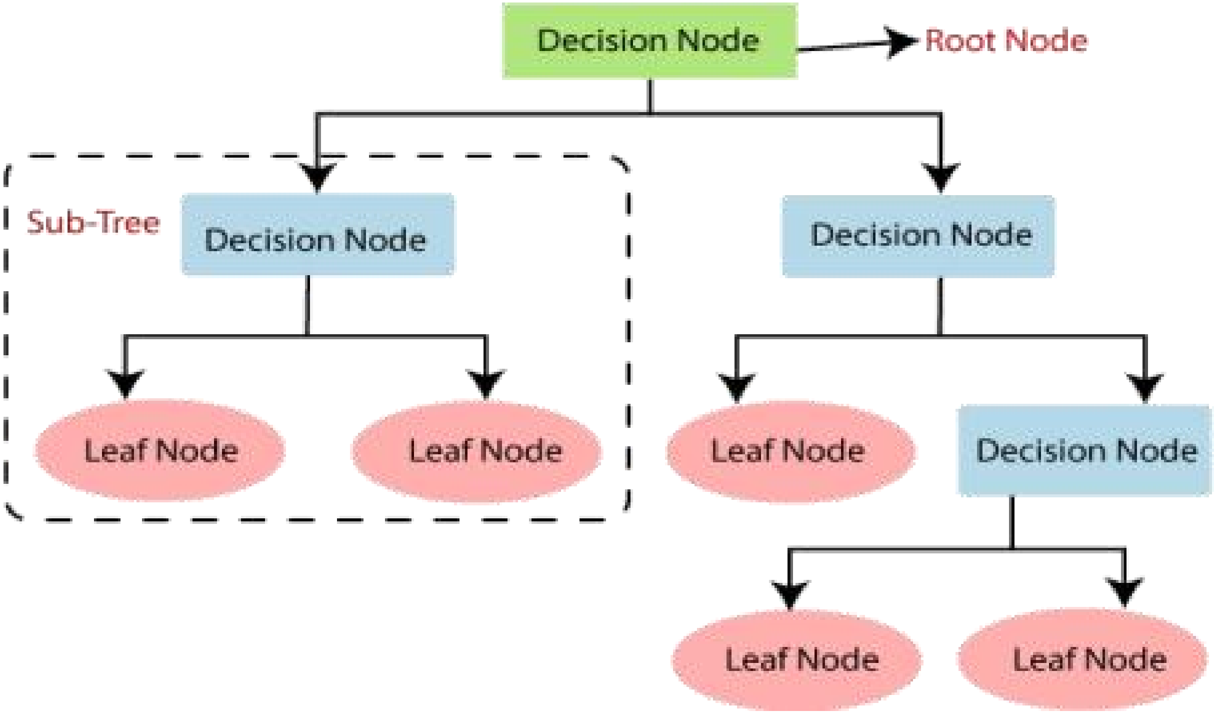
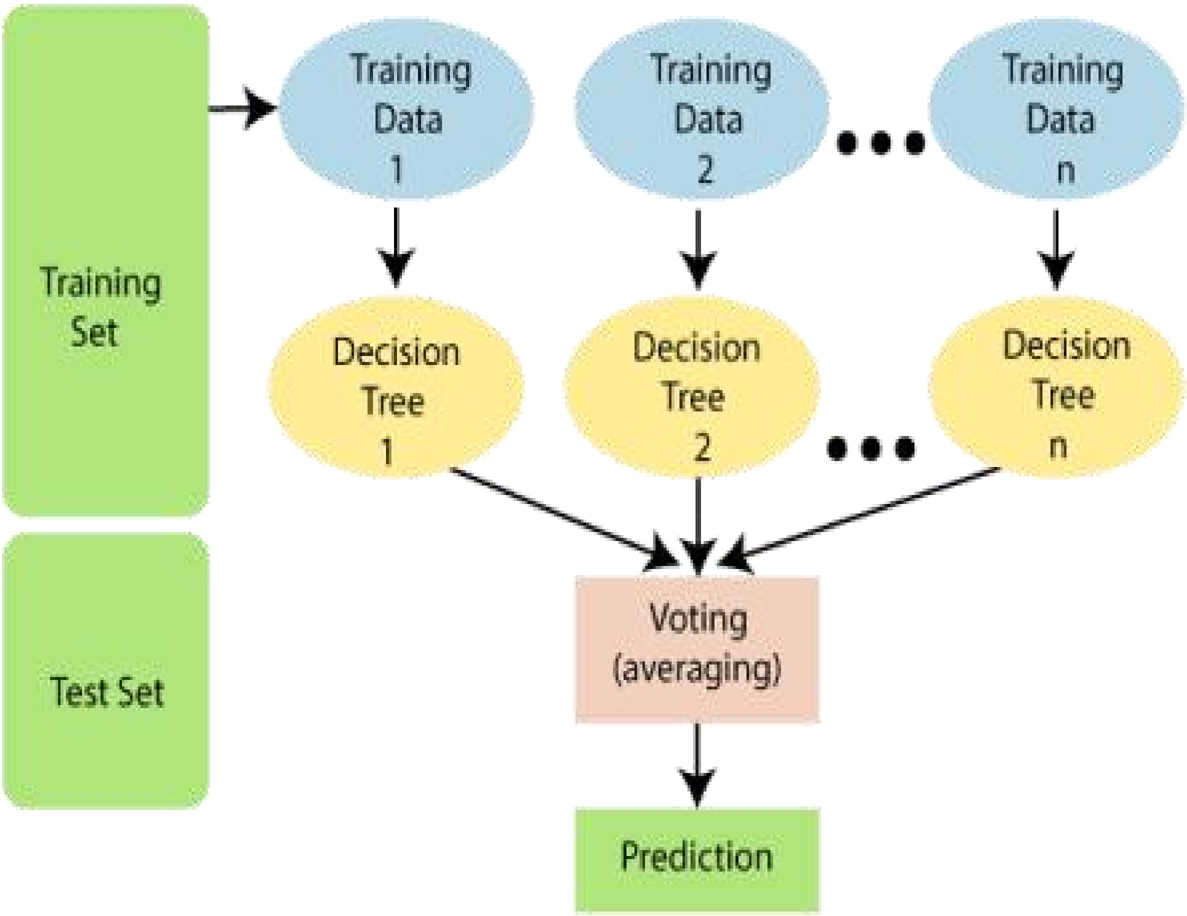


Fig: Decision Tree Flow Chat

The procedure for determining the class of a given dataset in a decision tree starts at the root node of the tree. This algorithm checks the values of the root attribute with the values of the record (actual dataset) attribute and then follows the branch and jumps to the next node based on the comparison.

❖ Random Forest

Random Forest is a well-known supervised machine learning algorithm. In machine learning, it can be used for both classification and regression. It is based on ensemble learning, which is the process of integrating numerous classifiers to solve a complex problem and improve the model's performance. "Random Forest is a classifier that contains a number of decision trees on various subsets of a given dataset and takes the average to enhance the predictive accuracy of that dataset," as the name suggests. Rather than depending on a single decision tree, the random forest collects predictions from each tree and ranks them according to the majority of votes.



Below are some points that explain why we should use the Random Forest algorithm:

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

CHAPTER4

Design Coding

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import seaborn as sns #visualisation
import matplotlib.pyplot as plt #visualisation
%matplotlib inline
import plotly.express as px
import plotly.graph_objects as go
import plotly.offline as py
import plotly.express as px

import os
import io

# Upload data set
from google.colab import files
uploaded = files.upload()

#Reads the data set
df = pd.read_csv(io.BytesIO(uploaded['datafiles.csv']))
df.head()
df.shape
df.columns
```

O/P

```
Index(['S. No', 'Category', 'State/UT', '2019', '2020', '2021',
 'Percentage Share of State/UT (2021)',
  'Mid-Year Projected Population (in Lakhs) (2021)+',
  'Rate of Total Cyber Crimes (2021)++'],
  dtype='object')
```

df.describe()

O/P

```
2019    2020    2021    Percentage Share of State/UT (2021) Mid-Year Projected
Population (in Lakhs) (2021)+          Rate of Total Cyber Crimes (2021)++
count   39.000000      39.000000      39.000000      39.000000      39.000000
        39.000000
mean    947.461538     1676.615385    2096.000000    7.689744       1017.987179
        1.689744
std     2724.974532    4832.658115    6065.161416    22.257391      2885.991893
        1.811193
min     0.000000       0.000000       0.000000       0.000000       0.700000
        0.000000
25%     9.500000       11.500000      24.500000      0.100000       18.300000
        0.500000
50%     102.000000     176.000000     239.000000     0.900000       284.000000
        1.000000
75%     439.500000     772.000000     886.500000     3.250000       663.850000
        2.200000
max     12317.000000 21796.000000 27248.000000 100.000000     13233.800000
        8.900000
```

df.info()

O/P

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 39 entries, 0 to 38
Data columns (total 9 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                            --------------  -----
 0   S. No                               39 non-null     object
 1   Category                                   39 non-null     object
 2   State/UT                                   39 non-null     object
 3   2019                                       39 non-null     int64
 4   2020                                       39 non-null     int64
 5   2021                                       39 non-null     int64
 6   Percentage Share of State/UT (2021)             39 non-null     float64
 7   Mid-Year Projected Population (in Lakhs) (2021)+  39 non-null     float64
 8   Rate of Total Cyber Crimes (2021)++             39 non-null     float64
dtypes: float64(3), int64(3), object(3)
memory usage: 2.9+ KB
```

```
df.isnull().sum()
```

O/P

```
S. No                                      0
Category                                   0
State/UT                                   0
2019                                       0
2020                                       0
2021                                       0
Percentage Share of State/UT (2021)              0
Mid-Year Projected Population (in Lakhs) (2021)+    0
Rate of Total Cyber Crimes (2021)++              0
dtype: int64
```

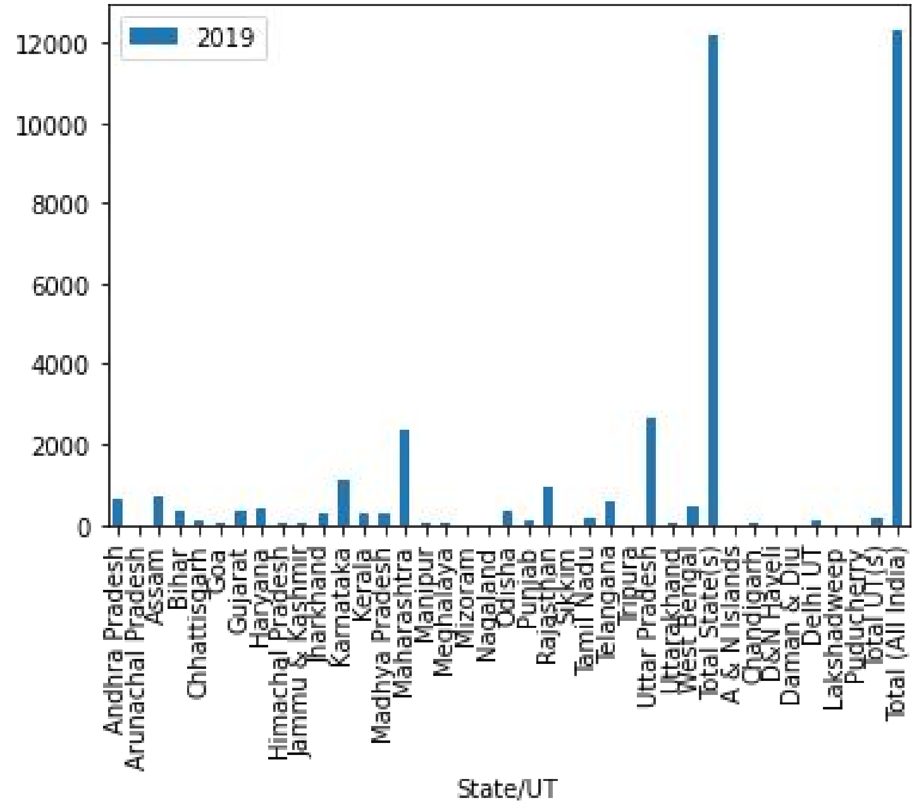
Visualisation

```
corr = df.corr()
corr.style.background_gradient(cmap = 'coolwarm')
```

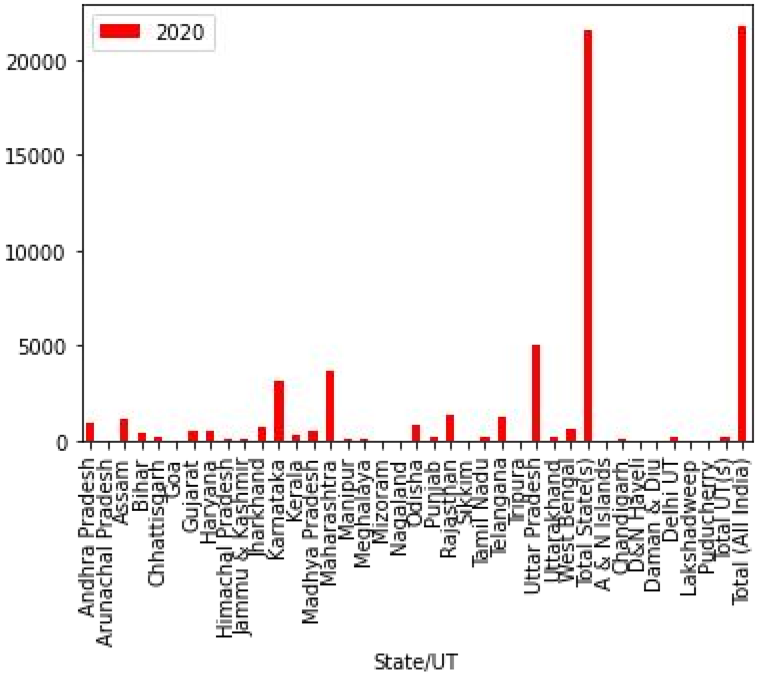
O/P

```
        2019    2020    2021    Percentage Share of State/UT (2021) Mid-Year Projected
Population (in Lakhs) (2021)+          Rate of Total Cyber Crimes (2021)++
2019    1.000000        0.998590        0.993830        0.993860        0.992970
        0.136820
2020    0.998590        1.000000        0.998014        0.998030        0.991394
        0.164416
2021    0.993830        0.998014        1.000000        0.999999        0.986735
        0.200750
Percentage Share of State/UT (2021) 0.993860        0.998030        0.999999
        1.000000        0.986789        0.200419
Mid-Year Projected Population (in Lakhs) (2021)+  0.992970        0.991394
        0.986735        0.986789        1.000000        0.077051
Rate of Total Cyber Crimes (2021)++          0.136820        0.164416        0.200750
        0.200419        0.077051        1.000000
```

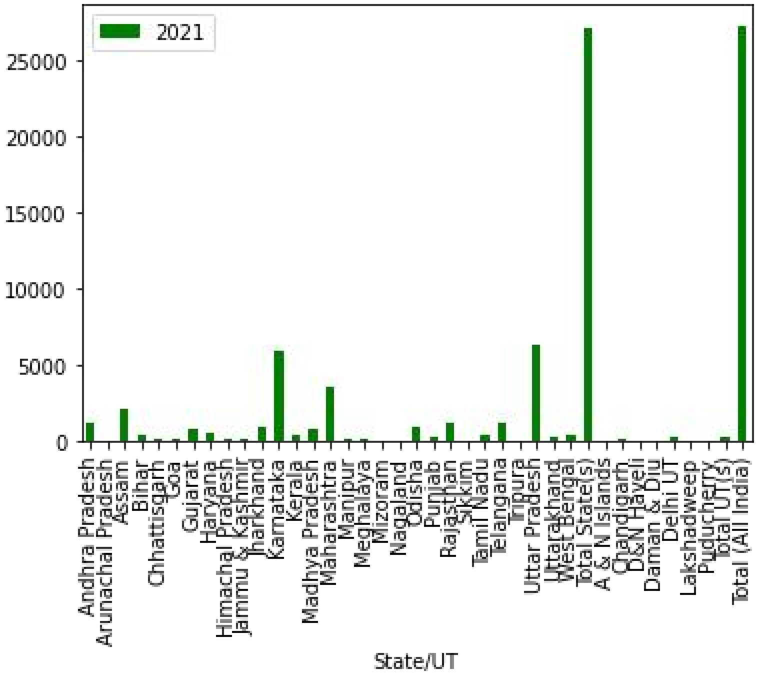
```
df.plot(kind='bar',x='State/UT',y='2019')
plt.show()
```



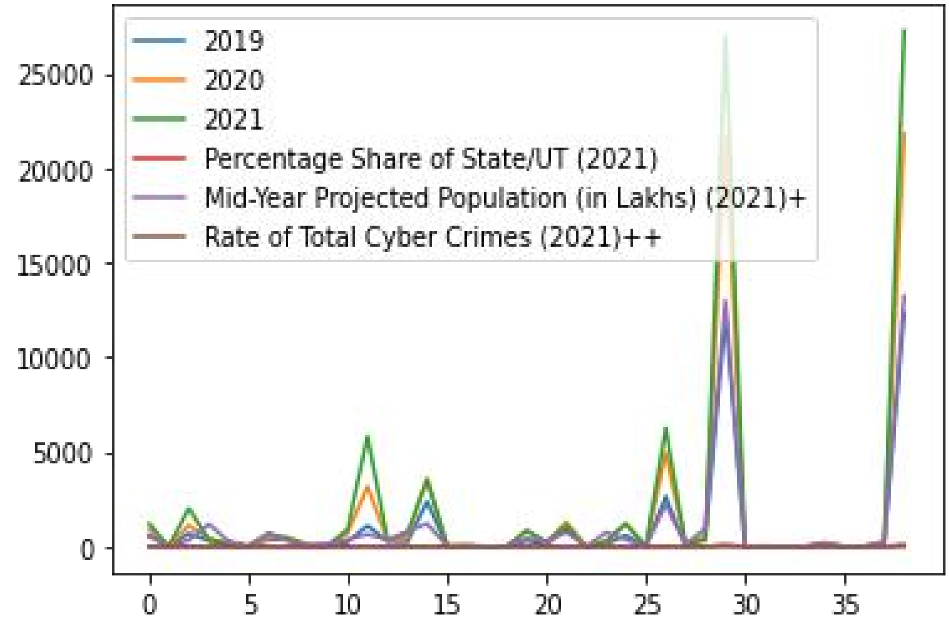
```
df.plot(kind='bar',x='State/UT',y='2020',color='r')
plt.show()
```



```
df.plot(kind='bar',x='State/UT',y='2021',color='g')
plt.show()
```



```
df.plot()
```

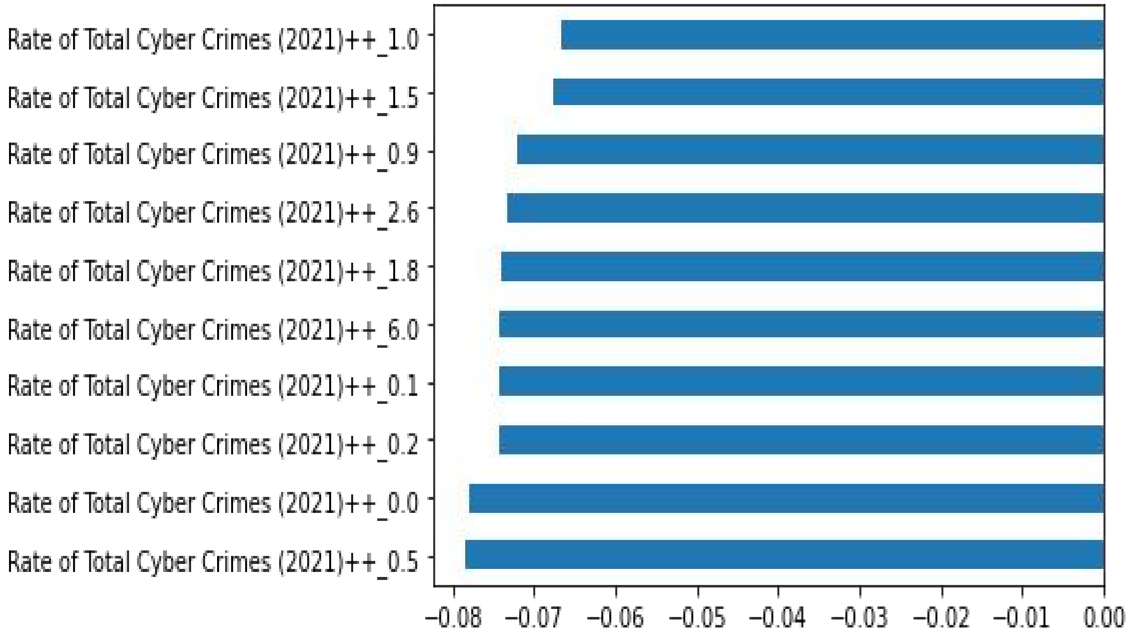


```
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import Normalizer
from sklearn.model_selection import train_test_split
from scipy.stats import skew

# categorical features
categorical_feat = [feature for feature in df.columns if df[feature].dtypes=='O']
print('Total categorical features: ', len(categorical_feat))
print('\n',categorical_feat)
O/P
Total categorical features:  3

 ['S. No', 'Category', 'State/UT']

for c in df.columns:
   if df[c].dtype=='float16' or  df[c].dtype=='float32' or  df[c].dtype=='float64':
        df[c].fillna(df[c].mean())

#fill in -999 for categoricals
df = df.fillna(-999)
# Label Encoding
for f in df.columns:
    if df[f].dtype=='object':
        lbl = LabelEncoder()
        lbl.fit(list(df[f].values))
        df[f] = lbl.transform(list(df[f].values))
from sklearn.model_selection import train_test_split
# Hot-Encode Categorical features
df = pd.get_dummies(df)

# Splitting dataset back into X and test data
X = df[:len(df)]
test = df[len(df):]

X.shape
df.columns.tolist()
```

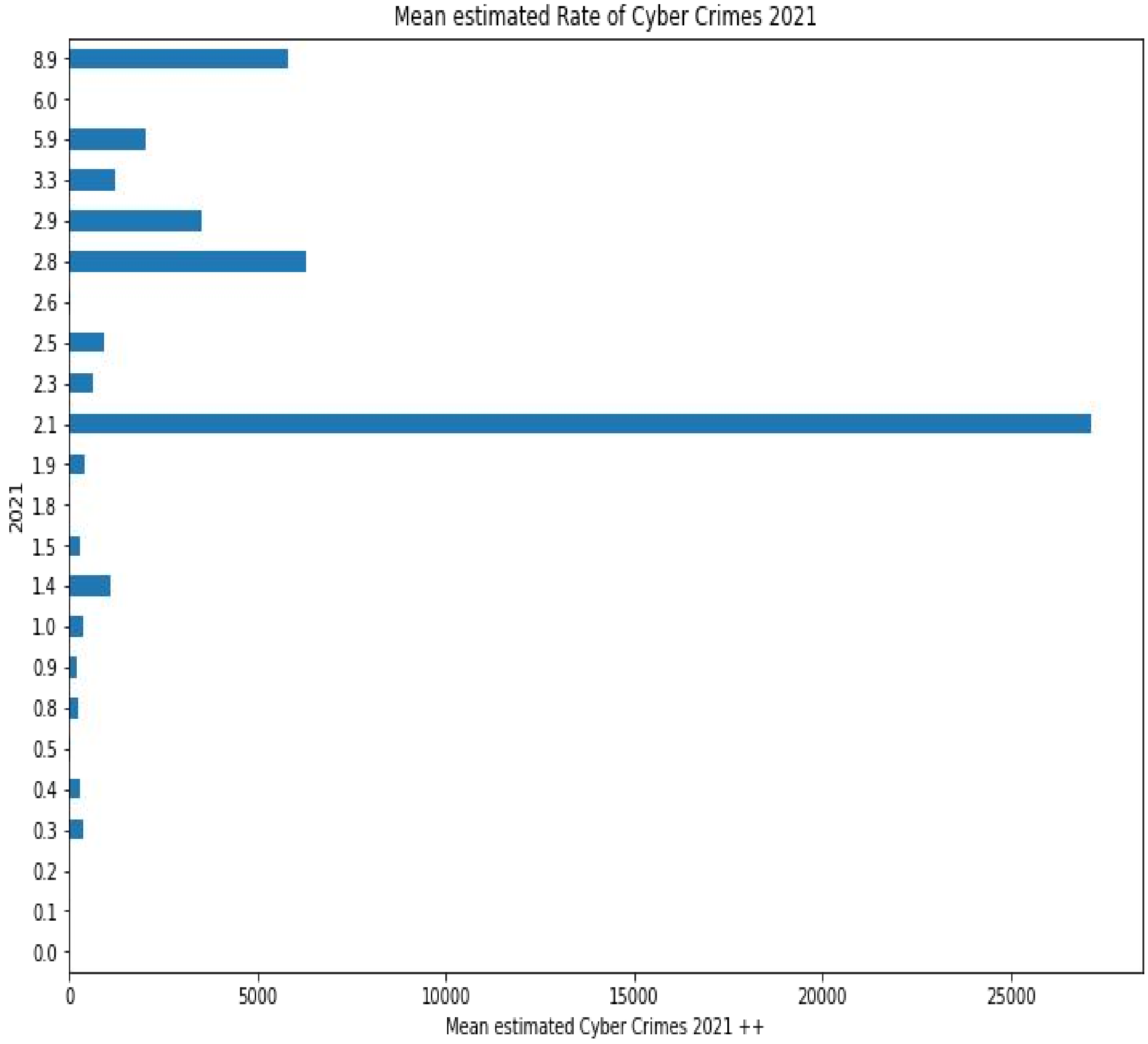
O/P

```
['S. No',
 'Category',
 'State/UT',
 '2019',
 '2020',
 '2021',
 'Percentage Share of State/UT (2021)',
 'Mid-Year Projected Population (in Lakhs) (2021)+',
 'Rate of Total Cyber Crimes (2021)++']
```

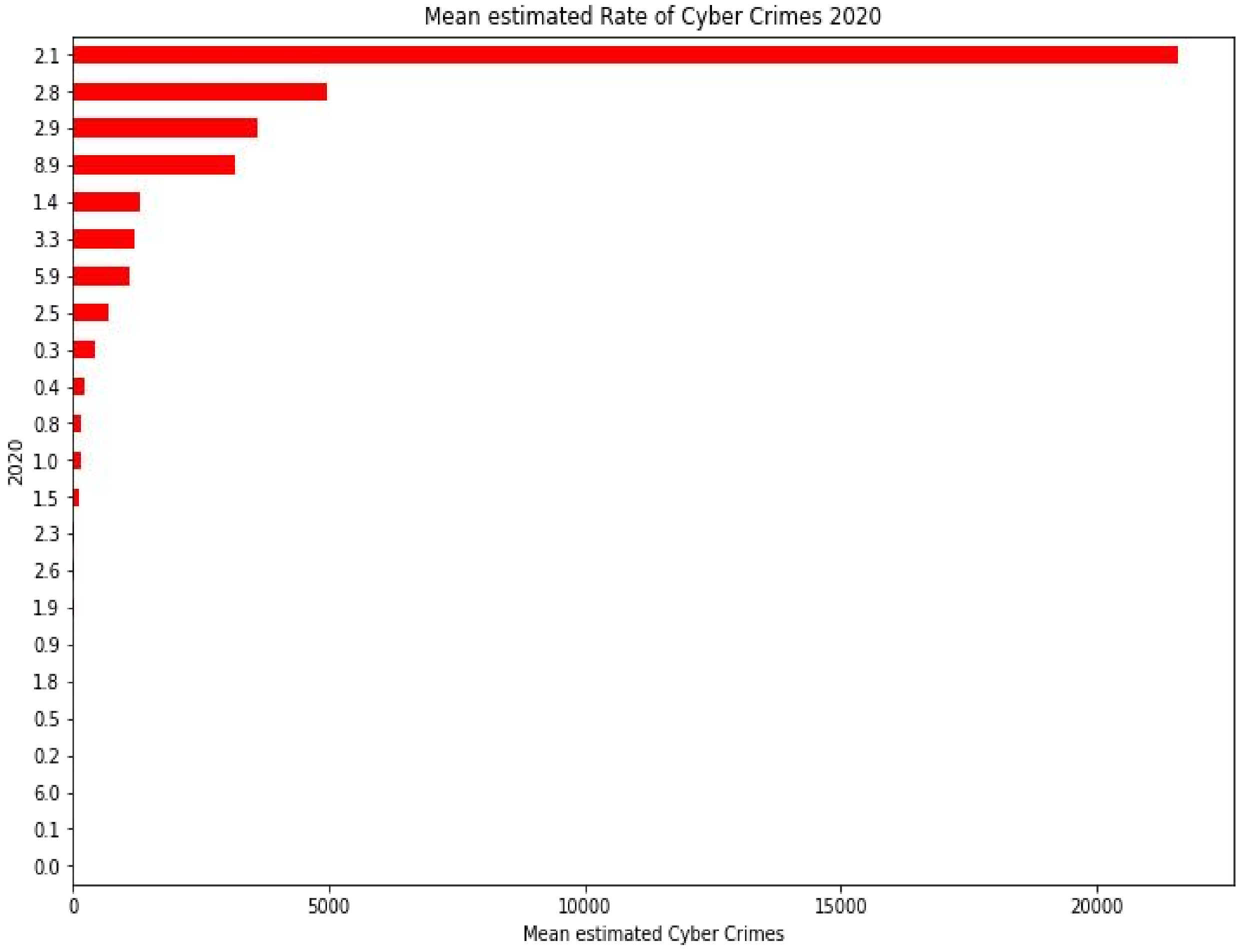
```
from category_encoders import OneHotEncoder
from sklearn.linear_model import LinearRegression, Ridge
from sklearn.preprocessing import StandardScaler, MinMaxScaler, MaxAbsScaler

cols_selected = ['Rate of Total Cyber Crimes (2021)++']
ohe = OneHotEncoder(cols=cols_selected, use_cat_names=True)
df_t = ohe.fit_transform(df[cols_selected+['2021']])

#scaler = MaxAbsScaler()
X = df_t.iloc[:,:-1]
y = df_t.iloc[:, -1].fillna(df_t.iloc[:, -1].mean()) / df_t.iloc[:, -1].max()

mdl = Ridge(alpha=0.1)
mdl.fit(X,y)

pd.Series(mdl.coef_, index=X.columns).sort_values().head(10).plot.barh()
```



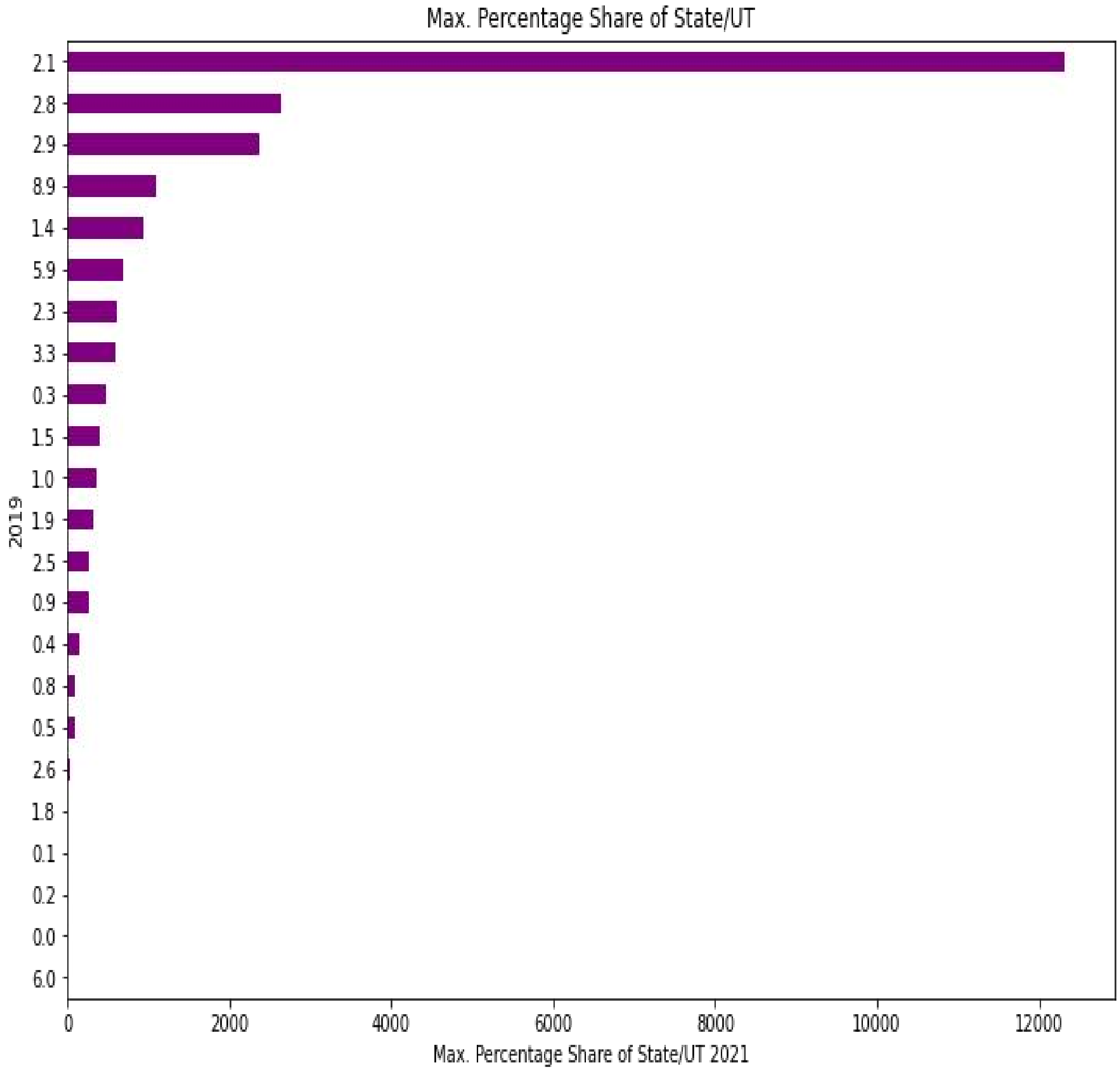
```
ax = df.groupby('Rate of Total Cyber Crimes (2021)++')['2021'].mean().plot(kind='barh',
figsize=(12,8),
title='Mean estimated Rate of Cyber Crimes 2021')
plt.xlabel('Mean estimated Cyber Crimes 2021 ++')
plt.ylabel('2021')
plt.show()
```



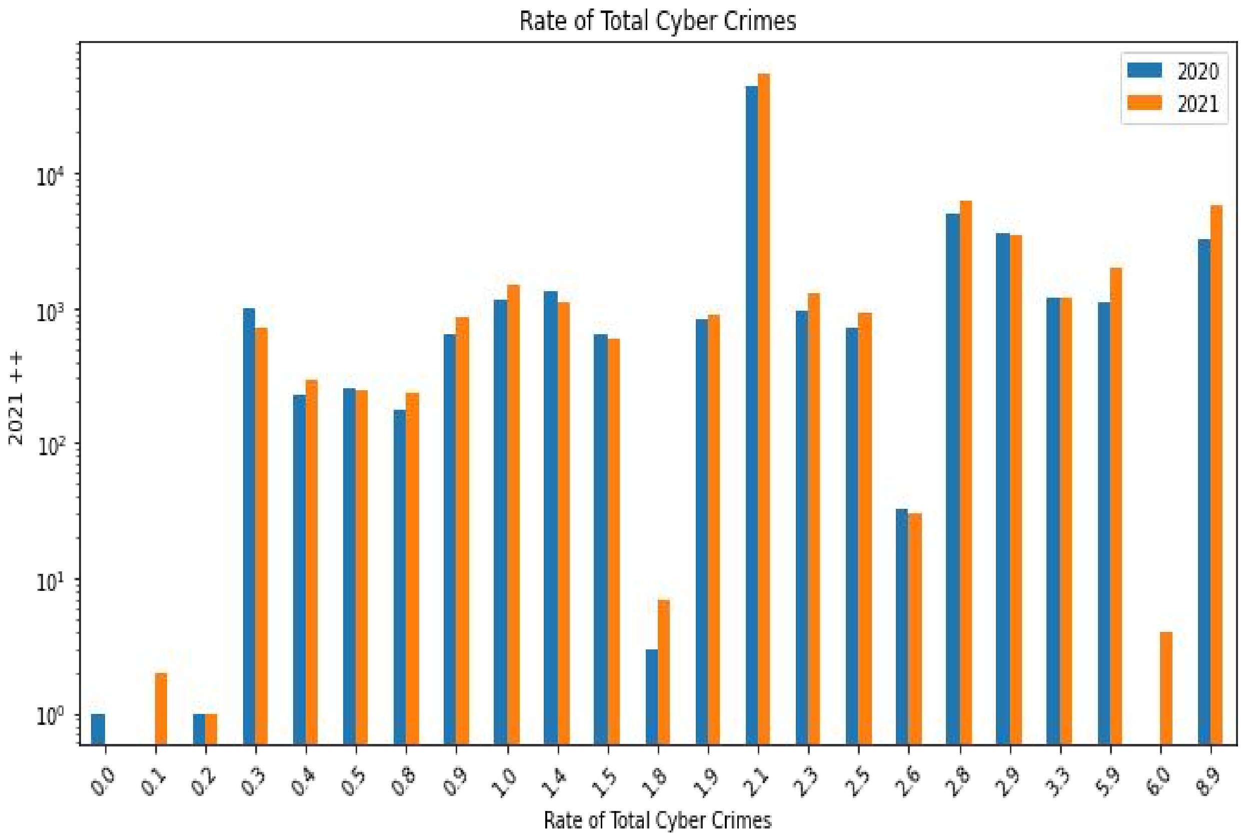
```
ax = df.groupby('Rate of Total Cyber Crimes
(2021)++')['2020'].min().sort_values(ascending=True).plot(kind='barh', figsize=(12,8),
color='r',
title='Mean estimated Rate of Cyber Crimes 2020')
plt.xlabel('Mean estimated Cyber Crimes')
plt.ylabel('2020')
plt.show()
```



```
ax = df.groupby('Rate of Total Cyber Crimes
(2021)++')['2019'].max().sort_values(ascending=True).plot(kind='barh', figsize=(12,8),
color='purple',
title='Max. Percentage Share of State/UT')
plt.xlabel('Max. Percentage Share of State/UT 2021')
plt.ylabel('2019')
plt.show()
```

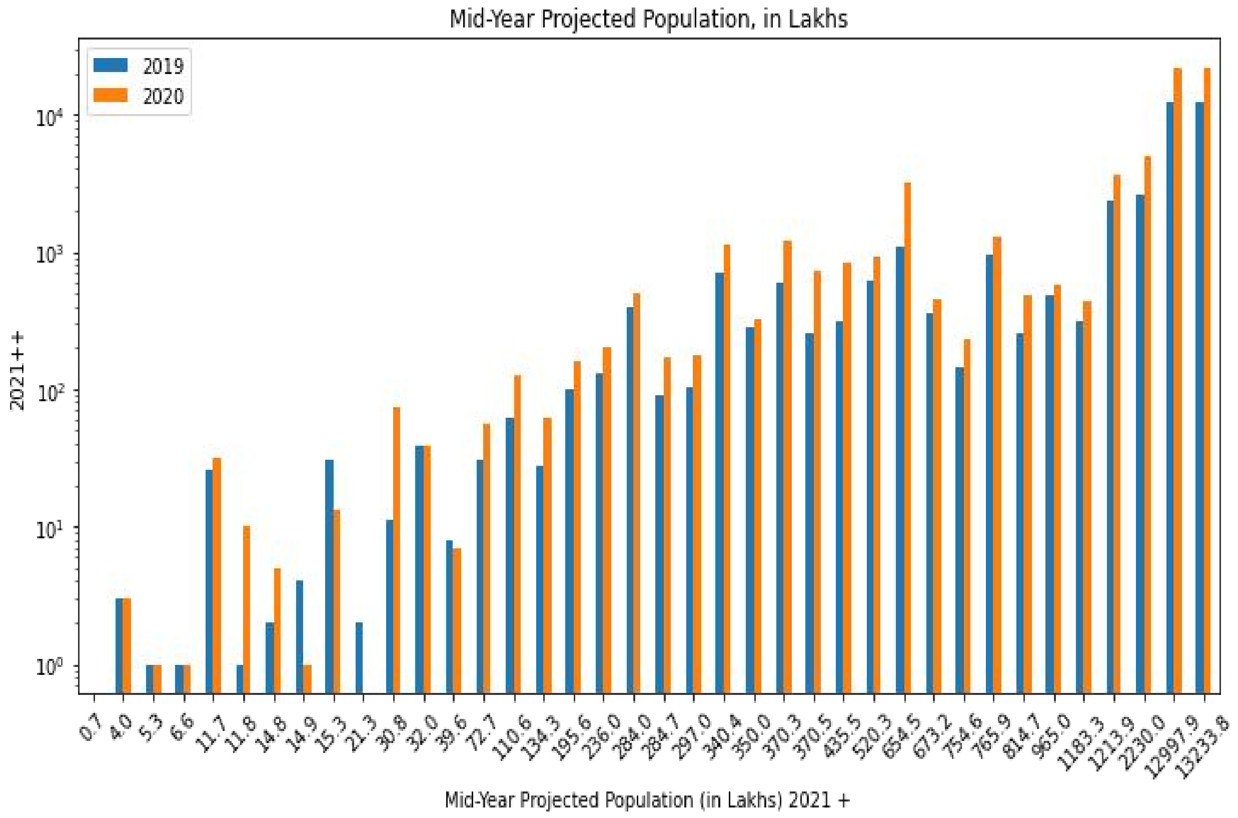


```
ax = df.groupby('Rate of Total Cyber Crimes
(2021)++')['2020','2021'].sum().plot(kind='bar', rot=45, figsize=(12,6),logy=True,
title='Rate of Total Cyber Crimes')
plt.xlabel('Rate of Total Cyber Crimes')
plt.ylabel('2021 ++')
plt.show()
```

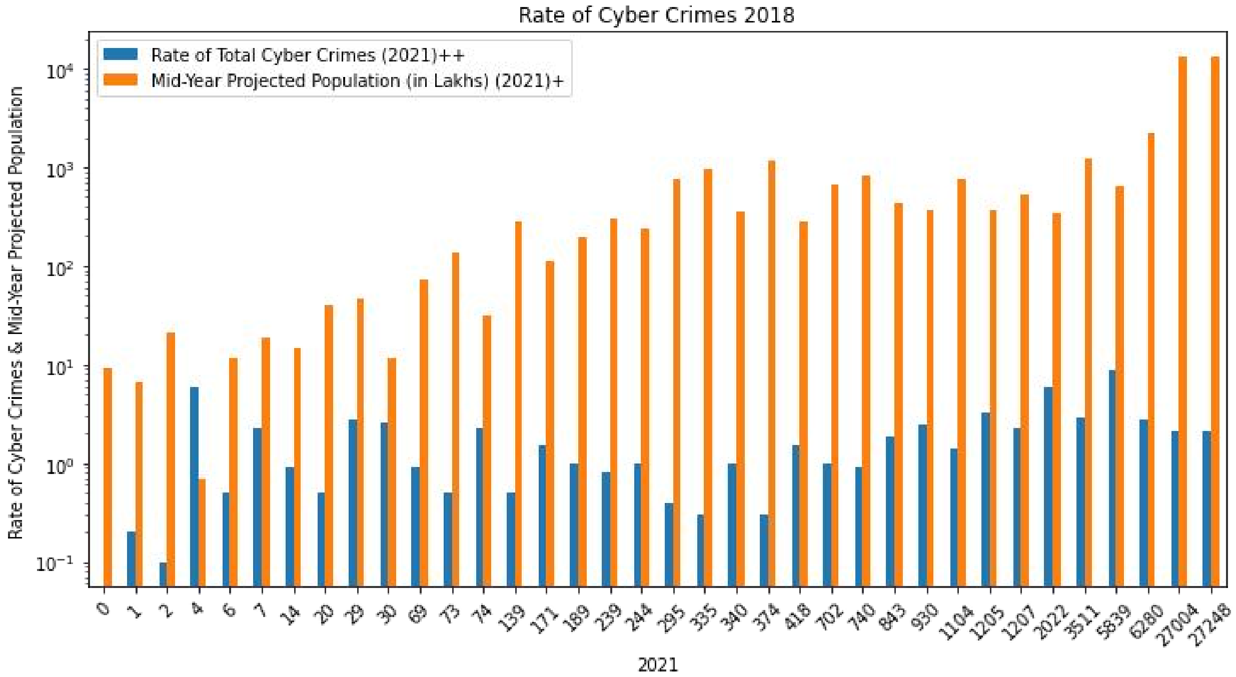


```
ax = df.groupby('Mid-Year Projected Population (in Lakhs)
(2021)+')['2019','2020'].sum().plot(kind='bar', rot=45, figsize=(12,6), logy=True,
title='Mid-Year Projected Population, in Lakhs')
plt.xlabel('Mid-Year Projected Population (in Lakhs) 2021 +')
plt.ylabel('2021++')
plt.show()
```

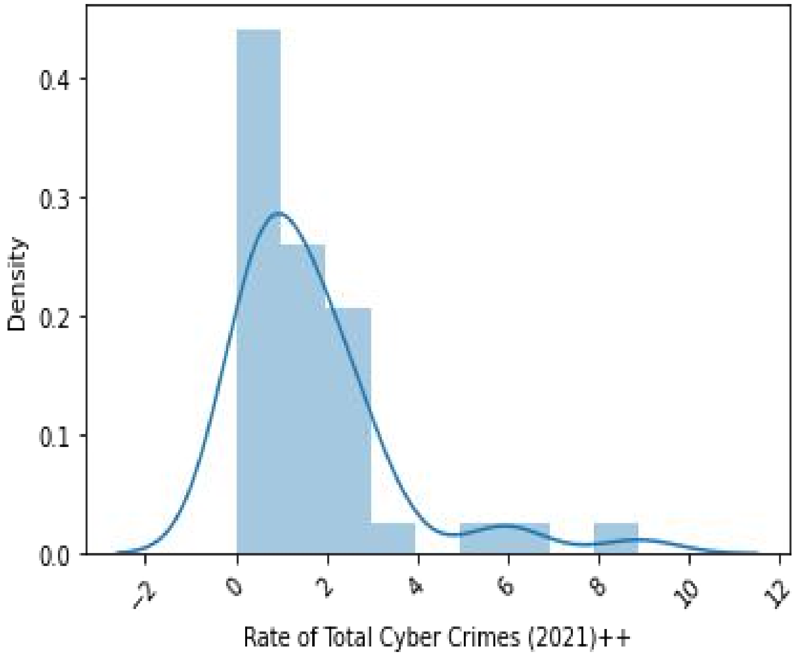
```
ax = df.groupby('Mid-Year Projected Population (in Lakhs)
(2021)+')['2019','2020'].sum().plot(kind='bar', rot=45, figsize=(12,6), logy=True,
title='Mid-Year Projected Population, in Lakhs')
plt.xlabel('Mid-Year Projected Population (in Lakhs) 2021 +')
plt.ylabel('2021++')
plt.show()
```



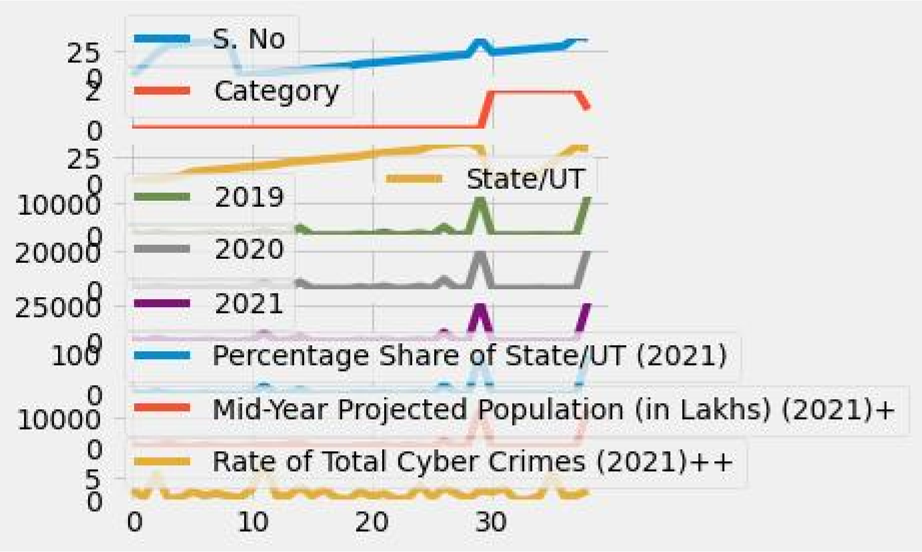
```
ax = df.groupby('2021')['Rate of Total Cyber Crimes (2021)++', 'Mid-Year Projected
Population (in Lakhs) (2021)+'].sum().plot(kind='bar', rot=45, figsize=(12,6), logy=True,
                                        title='Rate of Cyber Crimes 2018')
plt.xlabel('2021')
plt.ylabel('Rate of Cyber Crimes & Mid-Year Projected Population')

plt.show()

ax = df.groupby('Mid-Year Projected Population (in Lakhs)
(2021)+')['2019','2020'].sum().plot(kind='bar', rot=45, figsize=(12,6), logy=True,
title='Mid-Year Projected Population, in Lakhs')
plt.xlabel('Mid-Year Projected Population (in Lakhs) 2021 +')
plt.ylabel('2021++')
plt.show()
```



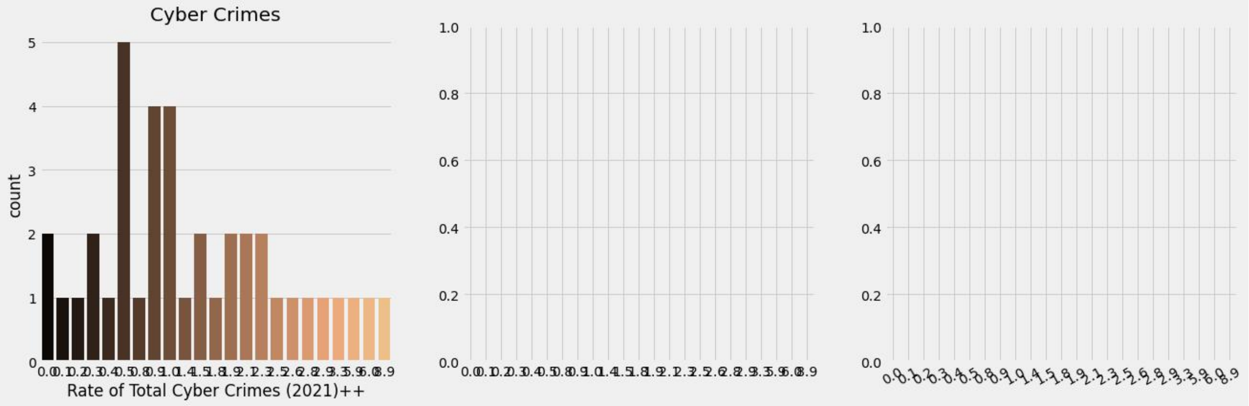
```
import matplotlib.ticker as ticker
ax = sns.distplot(df['Rate of Total Cyber Crimes (2021)++'])
plt.xticks(rotation=45)
ax.xaxis.set_major_locator(ticker.MultipleLocator(2))
figsize=(10, 4)
```



```
plt.style.use('fivethirtyeight')
df.plot(subplots=True, figsize=(4, 4), sharex=False, sharey=False)
plt.show()
```



```
df = df.rename(columns={'State/UT':'state'})
fig, ax = plt.subplots(1,3, figsize = (20,6), sharex=True)
sns.countplot(x='Rate of Total Cyber Crimes (2021)++',data=df, palette="copper",
ax=ax[0])

ax[0].title.set_text('Cyber Crimes')
plt.xticks(rotation=30)
plt.show()
```



```
import json
from urllib.request import urlopen
import pandas as pd
import matplotlib.pyplot as plt
import io
df = pd.read_csv(io.BytesIO(uploaded['monoxor.csv']))
df.head()
```

O/P

```
        req/baseUrl    req/body/note/title    req/body/note/desc    req/fresh
        req/headers/host       req/headers/user-agent        req/headers/content-type
        req/headers/org_id     req/headers/user_session_id req/headers/accept     ...
        req/hostname  req/ip   req/originalUrl req/path         req/protocol    req/secure
        req/stale         req/subdomains/0        req/xhr isSafe
0       /crm/note        Tina Johnson  Top recognize eat. Fact whom spend area thing ...
        False   example.com  insomnia/2020.4.2      application/json
        5f572820f65af8ac955b2e83   5fb27d3750b11901a35649fe */*        ...
        example.com  ::ffff:117.99.96.244      /crm/note        /note    http      False   True
        crm      False   True
1       /crm/note        Clayton Cooper          As possible American many prepare four
strong....       False   example.com  insomnia/2020.4.2      application/json
        5f572820f65af8ac955b2e83   5fb27d3750b11901a35649fe */*        ...
        example.com  ::ffff:117.99.96.244      /crm/note        /note    http      False   True
        crm      False   True
2       /crm/note        Curtis Wolfe    Tuesday Notes or 2 like 2 XSP Class           False
        example.com  insomnia/2020.4.2      application/json
        5f572820f65af8ac955b2e83   5fb27d3750b11901a35649fe */*        ...
        example.com  ::ffff:117.99.96.244      /crm/note        /note    http      False   True
        crm      False   False
3       /crm/note        Laura Fisher   State third represent energy campaign not forg...
        False   example.com  insomnia/2020.4.2      application/json
        5f572820f65af8ac955b2e83  5fb27d3750b11901a35649fe */*      ...
        example.com  ::ffff:117.99.96.244    /crm/note       /note   http    False   True
        crm     False   True
4       /crm/note       Tyler Santos    Us enjoy since. Time identify image position o...
        False   example.com  insomnia/2020.4.2      application/json
        5f572820f65af8ac955b2e83  5fb27d3750b11901a35649fe */*      ...
        example.com  ::ffff:117.99.96.244    /crm/note       /note   http    False   True
        crm     False   False
```

df.describe(include = 'all')

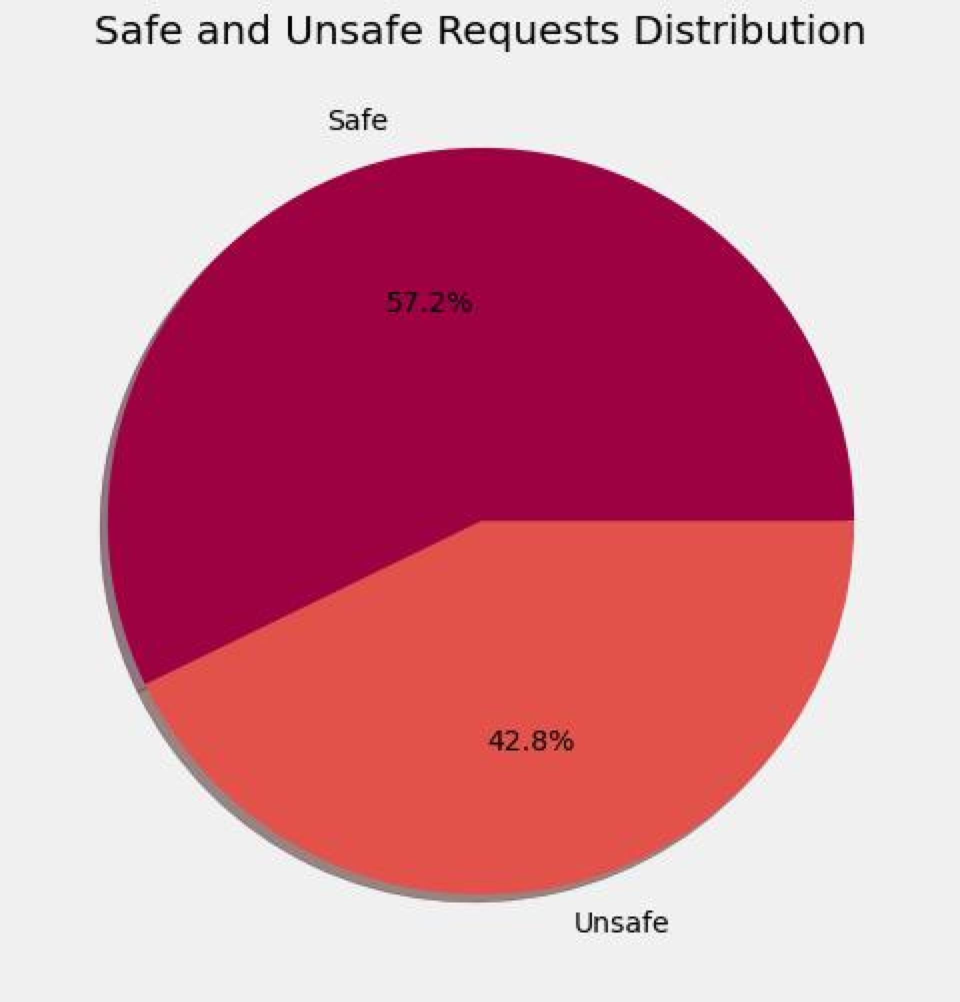
O/P

```
req/baseUrl     req/body/note/title     req/body/note/desc      req/fresh
        req/headers/host        req/headers/user-agent  req/headers/content-type
        req/headers/org_id      req/headers/user_session_id req/headers/accept      ...
        req/hostname    req/ip  req/originalUrl req/path        req/protocol    req/secure
        req/stale       req/subdomains/0        req/xhr isSafe
count   1000    1000    1000    1000    1000    1000    1000    1000    1000    1000    ...
        1000    1000    1000    1000    1000    1000    1000    1000    1000    1000
unique  1       994     750     1       1       1       1       1       1       1       ...
        1       1       1       1       1       1       1       1       1       2
top     /crm/note       Michael Smith note title <title onPropertyChange title onPro...
        False   example.com  insomnia/2020.4.2      application/json
        5f572820f65af8ac955b2e83  5fb27d3750b11901a35649fe */*      ...
        example.com  ::ffff:117.99.96.244    /crm/note       /note   http    False   True
        crm     False   True
freq    1000    3       36      1000    1000    1000    1000    1000    1000    1000    ...
        1000    1000    1000    1000    1000    1000    1000    1000    1000    572
mean    NaN     NaN     NaN     NaN     NaN     NaN     NaN     NaN     NaN     NaN     ...
        NaN     NaN     NaN     NaN     NaN     NaN     NaN     NaN     NaN     NaN
std     NaN     NaN     NaN     NaN     NaN     NaN     NaN     NaN     NaN     NaN     ...
        NaN     NaN     NaN     NaN     NaN     NaN     NaN     NaN     NaN     NaN
```

```
min     NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  ...
        NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN
25%     NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  ...
        NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN
50%     NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  ...
        NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN
75%     NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  ...
        NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN
max     NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  ...
        NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN  NaN
```

```
df_new = df[['req/body/note/title', 'req/body/note/desc', 'isSafe']]
df_new.head()
O/P
        req/body/note/title     req/body/note/desc     isSafe
0       Tina Johnson  Top recognize eat. Fact whom spend area thing ...  True
1       Clayton Cooper          As possible American many prepare four strong....  True
2       Curtis Wolfe  Tuesday Notes or 2 like 2 XSP Class          False
3       Laura Fisher  State third represent energy campaign not forg...    True
4       Tyler Santos  Us enjoy since. Time identify image position o...    False
```

```
mylabels='Safe','Unsafe'
cmap = plt.get_cmap('Spectral')
colors = [cmap(i) for i in np.linspace(0, 1, 8)]

plt.figure(figsize=(16,8))

plt.title("Safe and Unsafe Requests Distribution", size = 20)
plt.pie(df_new['isSafe'].value_counts(), labels=mylabels, autopct='%1.1f%%',
shadow=True, colors=colors)
plt.show()
```



```
not_safe = df_new['isSafe'].isin([False])
df_not_safe = df_new[not_safe].reset_index(drop=True)

df_not_safe = pd.DataFrame(df_not_safe)
df_not_safe
```

O/P

```
        req/body/note/title     req/body/note/desc    isSafe
0       Curtis Wolfe    Tuesday Notes or 2 like 2 XSP Class          False
1       Tyler Santos    Us enjoy since. Time identify image position o...      False
2       Tracy Smith     Area single occur chair opportunity art many. ...      False
3       Jacob Martin    note title <title onPropertyChange title onPro...      False
4       Colleen Riggs Meeting ")) or (("x"))=(("x Notes 12:30          False
...     ...     ...     ...
```

```
423     Gary Ruiz       <img src=1 href=1 onerror="javascript:alert(1)...       False
424     Jordan Brown Meeting ")) or (("x"))=(("x Notes 12:30           False
425     Larry Perez     ext1%3Cvideo+src%3D1+href%3D1+onerror%3D%22jav...
        False
426     Jason Tucker  ext1%3Cvideo+src%3D1+href%3D1+onerror%3D%22jav...
        False
427     James Rocha  <img src=1 href=1 onerror="javascript:alert(1)...       False
```

```
safe = df_new['isSafe'].isin([True])
df_safe = df_new[safe].reset_index(drop=True)

df_safe
```

O/P

```
        req/body/note/title       req/body/note/desc     isSafe
0       Tina Johnson  Top recognize eat. Fact whom spend area thing ...  True
1       Clayton Cooper          As possible American many prepare four strong....  True
2       Laura Fisher   State third represent energy campaign not forg...      True
3       Kevin Gonzalez          Resource politics already close phone special ...     True
4       James Taylor  Truth care red give all own. Full better marke...        True
...     ...       ...       ...
567     James Smith    Mr support color history natural PM. Informati...        True
568     Bridget Elliott  Recently crime before five thought bit. Card f...        True
569     Jasmine Gibson          Set nature they then low resource truth. Edge ...      True
570     Mr. Antonio Valdez DDS         Push case them such face suffer. Letter middle...
        True
571     Brian Stephens          Family ready stay rule full than yet. Moment o...       True
```

Maliciuos Words in Safe and Unsafe requests.

```
import string
string.punctuation
def remove_punctuation(text):
   no_punct = [word for word in text if not word.isalpha()]
   no_punct = [word for word in no_punct if word!= '.']
```

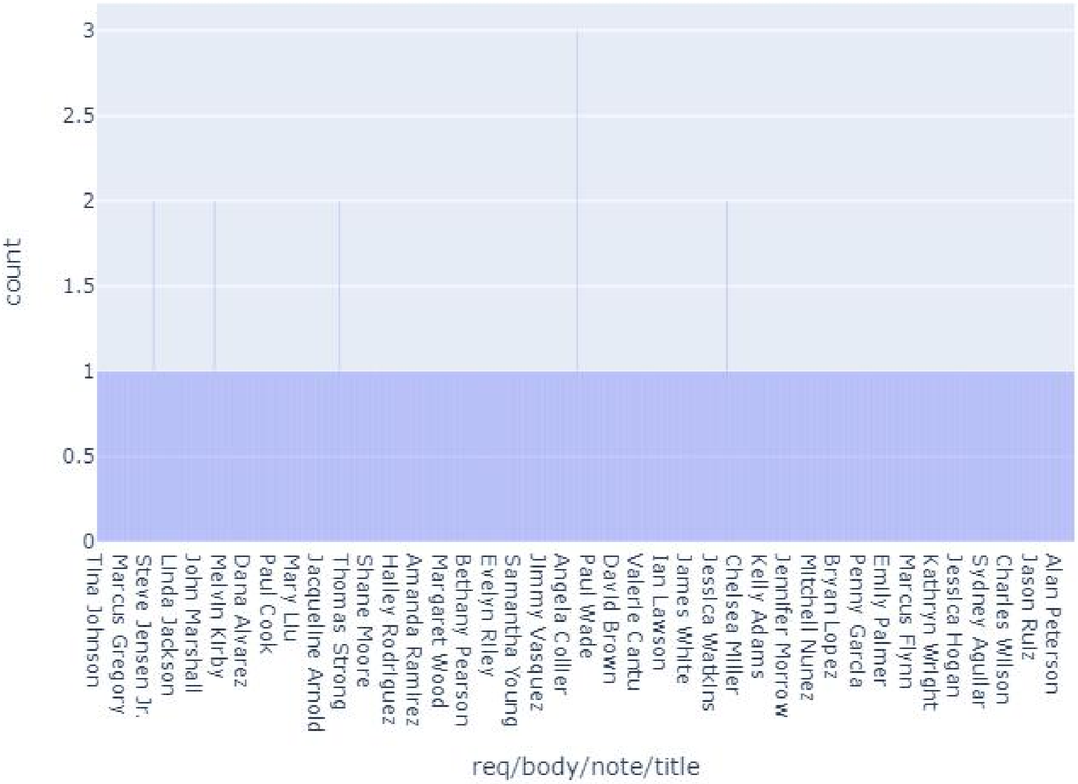
```
    words_wo_punct = ''.join(no_punct)

    return words_wo_punct
mal_words_ns = df_not_safe['req/body/note/desc'].apply(lambda x:
remove_punctuation(x))
print(mal_words_ns)

0                    2  2
1              \r\n
2                \r\n
3              <   ="::(1)"></ >
4           "))  ((""))=(("  12:30
            ...
423            < =1 =1 =":(1)"></>
424          "))  ((""))=(("  12:30
425    1%3+%31+%31+%3%22%3%281%29%22%3%3%2%3
426    1%3+%31+%31+%3%22%3%281%29%22%3%3%2%3
427            < =1 =1 =":(1)"></>
Name: req/body/note/desc, Length: 428, dtype: object

mal_words_ns.index = np.arange(len(mal_words_ns))

corpus_mal_words = []

for i in (range(len(mal_words_ns))):
  review = mal_words_ns[i]
  review = review.lower()
  review = review.split()

  if len(review) > 0:
      corpus_mal_words.append(review)
corpus_mal_words

print(len(corpus_mal_words))
```

```
string.punctuation

import string
string.punctuation
def remove_punctuation(text):
   no_punct = [word for word in text if not word.isalpha()]
   no_punct = [word for word in no_punct if word!= '.']
   words_wo_punct = ''.join(no_punct)

   return words_wo_punct
mal_words_s = df_safe['req/body/note/desc'].apply(lambda x: remove_punctuation(x))
print(mal_words_s)
```

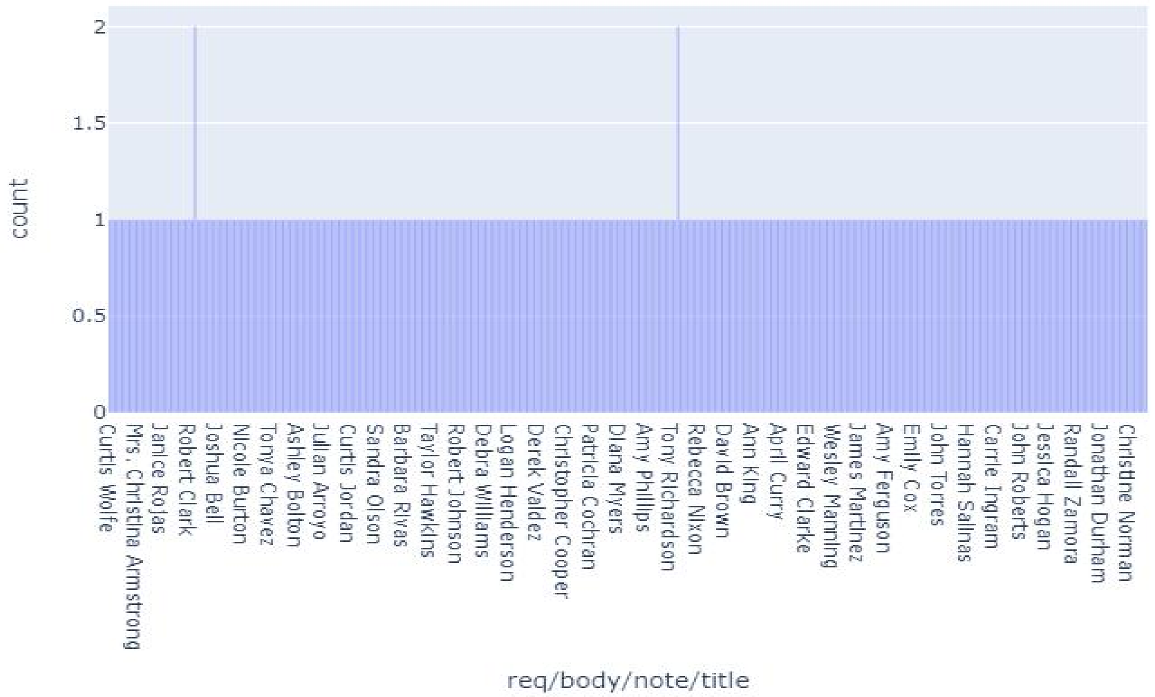
O/P

```
0           \r\n       \r\n
1
2                        \r\n
3
4                        \r\n
                 ...
567
568
569                      \r\n
570                    \r\n
571
Name: req/body/note/desc, Length: 572, dtype: object
```

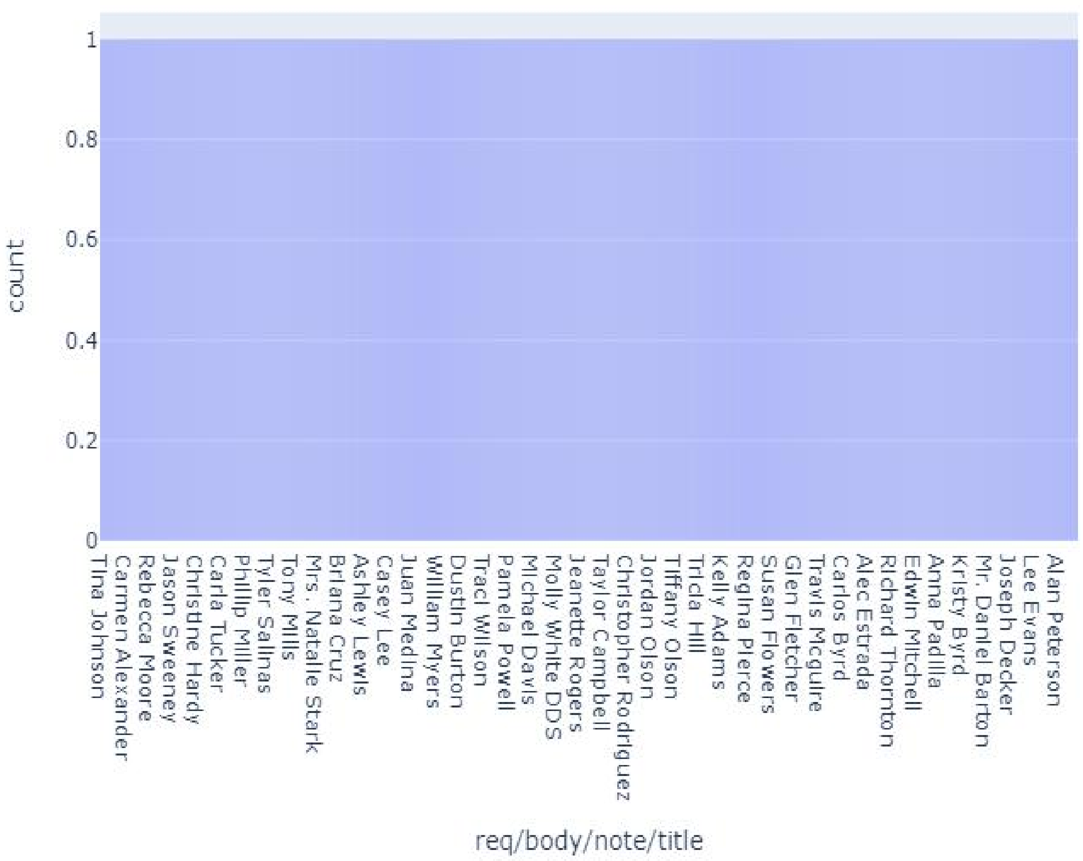
```
import plotly.express as px
px.histogram(df_new, x = 'req/body/note/title', opacity = 0.5)
```



```
import plotly.express as px
px.histogram(df_not_safe, x = 'req/body/note/title', opacity = 0.5)

ax = df.groupby('Mid-Year Projected Population (in Lakhs)
(2021)+')['2019','2020'].sum().plot(kind='bar', rot=45, figsize=(12,6), logy=True,
title='Mid-Year Projected Population, in Lakhs')
ax = df.groupby('Mid-Year Projected Population (in Lakhs)
(2021)+')['2019','2020'].sum().plot(kind='bar', rot=45, figsize=(12,6), logy=True,
title='Mid-Year Projected Population, in Lakhs')
plt.xlabel('Mid-Year Projected Population (in Lakhs) 2021 +')
plt.ylabel('2021++')
plt.show()
```



```
import plotly.express as px
px.histogram(df_safe, x = 'req/body/note/title', opacity = 0.5)

from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()
ax = df.groupby('Mid-Year Projected Population (in Lakhs)
(2021)+')['2019','2020'].sum().plot(kind='bar', rot=45, figsize=(12,6), logy=True,
title='Mid-Year Projected Population, in Lakhs')
df_new['isSafe'] = le.fit_transform(df_new['isSafe'])
df_new['desc length'] = df_new['req/body/note/desc'].astype(str).apply(len)
y = df_new['isSafe']
X = df_new.drop(columns = 'isSafe')
```



```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df_new['isSafe'] = le.fit_transform(df_new['isSafe'])
df_new['desc length'] = df_new['req/body/note/desc'].astype(str).apply(len)
y = df_new['isSafe']
X = df_new.drop(columns = 'isSafe')

import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
X.index = np.arange(len(X))
corpus = []
```

```
from tqdm import tqdm
for i in tqdm(range(len(X))):
  review = X['req/body/note/desc'][i]
  review = review.lower()
  review = review.split()
  ps = PorterStemmer()
  all_stopwords = stopwords.words('english')
  review = [ps.stem(word) for word in review if not word in set(all_stopwords)]
  review = ' '.join(review)
  corpus.append(review)

from sklearn.feature_extraction.text import CountVectorizer as CV
cv  = CV(max_features = 100,ngram_range=(1,1))

X_cv = cv.fit_transform(corpus).toarray()
y = y.values
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X_cv, y, test_size = 0.20, random_state
= 0)
from sklearn.naive_bayes import BernoulliNB
classifier = BernoulliNB()
classifier.fit(X_train, y_train)

y_pred = classifier.predict(X_test)
from sklearn.metrics import accuracy_score
from sklearn import metrics
acc = accuracy_score(y_test, y_pred)
print("Accuracy of the classifier: ",acc)
print("Confusion matrix is :\n",metrics.confusion_matrix(y_test,y_pred))
print("Classification report: \n" ,metrics.classification_report(y_test,y_pred))
```

O/P

```
Accuracy of the classifier:  0.84
Confusion matrix is :
 [[ 49  32]
 [  0 119]]
Classification report:
              precision    recall  f1-score   support

           0       1.00      0.60      0.75        81
           1       0.79      1.00      0.88       119

    accuracy                           0.84       200
   macro avg       0.89      0.80      0.82       200
weighted avg       0.87      0.84      0.83       200

acc
0.84
```

An accuracy score of 84%.

Term Frequency - Inverse Document Frequency

```
from sklearn.feature_extraction.text import TfidfVectorizer as TV
tv  = TV(ngram_range =(1,1),max_features = 3000)
X_tv = tv.fit_transform(corpus).toarray()
X_train, X_test, y_train, y_test = train_test_split(X_tv, y, test_size = 0.20, random_state
= 0)
from sklearn.naive_bayes import MultinomialNB
classifier = MultinomialNB()
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
acc = accuracy_score(y_test, y_pred)
acc
0.84
```

Same accuracy as that of Bag of Words Technique, 84%.

Deep Learning Model

```
import tensorflow as tf
```

```
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from tensorflow.keras.layers import Dropout

tokenizer = Tokenizer(num_words = 100)
tokenizer.fit_on_texts(corpus)
sequences = tokenizer.texts_to_sequences(corpus)
padded = pad_sequences(sequences, padding='post')
word_index = tokenizer.word_index
vocab_size = len(tokenizer.word_index) + 1
embedding_dim = 64
model = tf.keras.Sequential([
    tf.keras.layers.Embedding(100, embedding_dim),
    tf.keras.layers.GlobalAveragePooling1D(),
    tf.keras.layers.Dropout(0.2),
    tf.keras.layers.Dense(6, activation='relu'),
    tf.keras.layers.Dropout(0.3),
    tf.keras.layers.Dense(1, activation='sigmoid')
])

model.summary()
```

O/p

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, None, 64)          6400

 global_average_pooling1d (G  (None, 64)               0
 lobalAveragePooling1D)

 dropout (Dropout)           (None, 64)                0

 dense (Dense)               (None, 6)                 390
 dropout_1 (Dropout)        (None, 6)              0


 dense_1 (Dense)            (None, 1)              7

=================================================================
Total params: 6,797
Trainable params: 6,797
Non-trainable params: 0
_________________________________________________________________
```

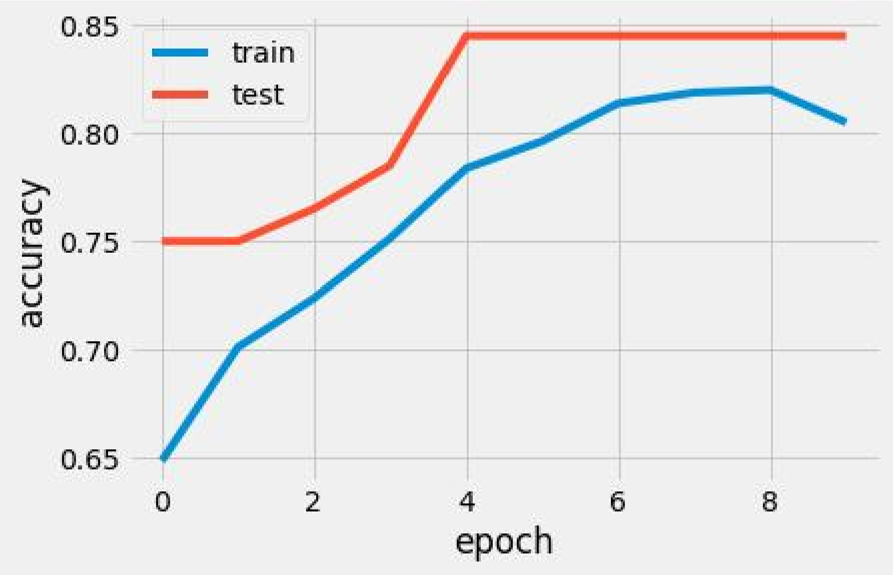
```
num_epochs = 10

model.compile(loss='binary_crossentropy',optimizer='adam',metrics=['accuracy'])
m = model.fit(padded,y,epochs= num_epochs,validation_split=0.2)
```

O/P

```
Epoch 1/10
25/25 [==============================] - 2s 20ms/step - loss: 0.6775 - accuracy:
0.6488 - val_loss: 0.6558 - val_accuracy: 0.7500
Epoch 2/10
25/25 [==============================] - 0s 3ms/step - loss: 0.6497 - accuracy:
0.7013 - val_loss: 0.6160 - val_accuracy: 0.7500
Epoch 3/10
25/25 [==============================] - 0s 4ms/step - loss: 0.6158 - accuracy:
0.7237 - val_loss: 0.5790 - val_accuracy: 0.7650
Epoch 4/10
25/25 [==============================] - 0s 3ms/step - loss: 0.5757 - accuracy:
0.7513 - val_loss: 0.5355 - val_accuracy: 0.7850
Epoch 5/10
25/25 [==============================] - 0s 3ms/step - loss: 0.5370 - accuracy:
0.7837 - val_loss: 0.5017 - val_accuracy: 0.8450
Epoch 6/10
25/25 [==============================] - 0s 3ms/step - loss: 0.5133 - accuracy:
0.7962 - val_loss: 0.4715 - val_accuracy: 0.8450
```

```
Epoch 7/10
25/25 [==============================] - 0s 3ms/step - loss: 0.4838 - accuracy:
0.8138 - val_loss: 0.4548 - val_accuracy: 0.8450
Epoch 8/10
25/25 [==============================] - 0s 3ms/step - loss: 0.4642 - accuracy:
0.8188 - val_loss: 0.4384 - val_accuracy: 0.8450
Epoch 9/10
25/25 [==============================] - 0s 3ms/step - loss: 0.4436 - accuracy:
0.8200 - val_loss: 0.4281 - val_accuracy: 0.8450
Epoch 10/10
25/25 [==============================] - 0s 4ms/step - loss: 0.4414 - accuracy:
0.8050 - val_loss: 0.4199 - val_accuracy: 0.8450
```

```
plt.plot(m.history['accuracy'])
plt.plot(m.history['val_accuracy'])
plt.ylabel('accuracy')
plt.xlabel('epoch')
plt.legend(['train', 'test'], loc='upper left')
plt.show()
```



Conclusion

The best accuracy achieved is 84.5%.

The main feature over which the safety is measured is the description feature. The concept of Natural Language Processing has been applied in order to analyse this feature, and create the learning model.

All the three models, Bag of Words model, TF-IDF model have a test accuracy of 84% and Deep Learning model has test accuracy of 84.5% .

CHAPTER 5

Results and discussions

Cybercrime is a broad phrase that refers to any criminal behaviour that involves the use of a computer. Hacking of consumer records and theft of intellectual property are examples of cyber crime on an organisational level. Many users believe that anti-spyware and antivirus software alone will safeguard them, their accounts, and their PCs. Consumers, as well as governmental and private businesses, are being targeted by cyber thieves who are growing more skilled.

This section outlines study results from descriptive statistics and later elaborates on the taxonomy of concepts found for surface-web cybercrime threat intelligence and later elaborates on taxonomy of concepts found for deep- and dark-web cybercrime threat intelligence. For the deep- and dark-web taxonomy, all concepts reported in the surface-web taxonomy were also found but were omitted for the sake of completeness and to highlight more specific results.

The expansion of the economy is directly proportional to the advancement of technology as it becomes more efficient. Because cybercriminals typically target chances that arise on systems, they take advantage of technological advancements to explore how they may infiltrate the systems, resulting in an increase in cybercrime activity.

CHAPTER 6

Conclusions

Recent studies published on the evolution of principal cyber threats in the security landscape. They present concerning scenarios, characterised by the constant growth of cybercrimes activities. Even though the level of awareness of cyber threats has increased, and law enforcement acts globally to combat them, illegal profits have reached amazing figures. The impact on society has become unsustainable, considering the global economic crisis. It’s necessary to work together to avoid the costs the global community suffers, which we can no longer sustain. The risk of business collapse is concrete, due to the high cost for enterprises in mitigating counter measures, and the damage caused by countless attacks. Nowadays customers have come to expect that organisations have a presence on the Internet, including a website and email capabilities. Use of the Internet is a risk that most companies have to take. The problem is to minimise the risks associated by doing so. If there is no technology, hopefully the cybercrimes would not be found anywhere. As it has been discussed in the paper, the preventive measures should be taken to prevent the society as well as the organisations from cybercrimes instead of avoiding the uses of the technology.

References

[1] Dawson, J. and Thomson, R., “The future cybersecurity workforce: Going beyond technical skills for successful cyber performance”, Frontiers in Psychology, 9(JUN), pp. 1–12, 2018, doi: 10.3389/fpsyg.2018.0074

[2] Praveen Paliwal, "Cyber Crime", Nations Congress on the Prevention of Crime and Treatment of Offenders, March 2016

[3] Le Compte, D. Elizondo and T. Watson, "A renewed approach to serious games for cyber security", 2015 7th International Conference on Cyber Conflict: Architectures in Cyberspace, pp. 203-216,2015

[4] Kshetri N. The simple economics of cybercrimes, IEEE Secur Priv, 4, pp. 33–39, 2006

[5] Hernández, A., Sanchez, V., Sánchez, G., Pérez, H., Olivares, J., Toscano, K., & Martinez, V. (2016,March). Security attack prediction based on user sentiment analysis of Twitter data. In 2016 IEEE international conference on industrial technology (ICIT) (pp. 610-617). IEEE.

[6] Buczak, A. L., & Guven, E ,”A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), pp. 1153-1176, 2016

[7] Apruzzese, G., Colajanni, M., Ferretti, L., Guido, A., & Marchetti, M. “On the effectiveness of machine and deep learning for cyber security”, 10th International Conference on Cyber Conflict (CyCon) pp. 371-390. IEEE, 2018

[8] Grace Odette Boussi," A Proposed Framework for Controlling Cyber- Crime", 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO),IEEE, India, 2020

[9] Priyanka Datta at. al.," A Technical Review Report on Cyber Crimes in India", International Conference on Emerging Smart Computing and Informatics (ESCI),IEEE, India, 2020

[10] F. Pasqualetti, F. Dorfler and F. Bullo, "Attack detection and identification in cyber-physical
systems", IEEE Transactions on Automatic Control, vol. 58, no. 11, pp. 2715-2729, 2013

[11] Adnan Amin at. Al. ," Classification of cyber-attacks based on rough set theory", First International Conference on Anti-Cybercrime (ICACC), IEEE, Saudi Arabia 2015

[12] R. Sabillon, J. Cano, V. Cavaller and J. Serra, "Cybercrime and Cybercriminals: A Comprehensive Study", International Journal of Computer Networks and Comm. Security, vol. 4, no. 6, pp. 165-176, 2016

[13] Z. Trabelsi, K. Hayawi, A. Braiki and S. Mathew, Network Attacks and Defenses: A Hands-on Approach, Boca Raton, Florida:CRC Press, 2013