

“Cyber Crime: Analysis and Prediction Using Machine Learning Algorithms”

Dharaneesh J

Department of Computing
(Student- Data Science)

Coimbatore Institute of Technology
Coimbatore, India
71762132012@cit.edu.in

Dr. Rajarajeshwari.K

Department of Computing
(Assistant Professor Data Science)
Coimbatore Institute of Technology
Coimbatore India
rajarajeshwari@cit.edu.in

Sasika SN

Department of Computing
(Student- Data Science)

Coimbatore Institute of Technology
Coimbatore, India
71762132040@cit.edu.in

Ms. Deivarani S

Department of Computing
(Assistant Professor Data Science)
Coimbatore Institute of Technology
Coimbatore India
deivarani@cit.edu.in

Abinaya M

Department of Computing
(Student- Data Science)

Coimbatore Institute of Technology
Coimbatore, India
71762132001@cit.edu.in

ABSTRACT:

Cyber Crime is a threat to the security of any individual, a community or to nation as a whole. Cyber Crime analysis and prediction is an approach which is used to find the patterns or trends that occur. Nowadays the numbers of cyber crime are increasing at an alarming rate. In the current scenario of rapidly increasing crime rate, a nation needs to analyse and predict the cyber crime before it occurs. To achieve this, few Machine Learning Algorithms are suggested. This paper studies about the cyber crime dataset of NCRB 2021 and provides the insights on crimes that occur state wise. The data set contains various attributes related to cyber crime tampering computer source documents, computer related offences, Ransomware, Dishonestly receiving stolen computer resource or communication on device, etc. Support Vector Machine and Random Forest algorithms are employed for classification and prediction. The main aim of this analysis is to help the authorities to detect, prevent and solve the cyber crime at an accurate and faster rate.

Keywords: Cyber Crime, Ransomware, Support Vector Machine, Random Forest, Logistic regression

I.INTRODUCTION

The rise of cybercrime poses a threat to individuals, communities and entire countries. When faced with this major threat, advanced technology must be used for cybercrime analysis and prediction. This article uses machine learning algorithms to delve into key areas of cybercrime to identify patterns and trends in an evolving environment.

Today, cyber crimes are proliferating at an alarming rate and proactive measures need to be taken to identify and reduce these crimes. To address this issue, this study leverages the power of National Crime Registry Bureau (NCRB) 2021 data

to provide a better understanding of cybercrime occurring across the state. The data is rich in features related to cybercrime, including activities such as falsification of computer data, computer crimes, ransomware attacks, and indirect fair acceptance of unauthorized use of computer or telecommunications equipment. To solve various problems of cybercrime, this paper uses advanced machine learning, special support vector machines (SVM), logistic regression, and random forest for classification. The main purpose of this analysis is to provide authorities with the necessary tools to not only detect but also prevent and solve cybercrimes with a high level of accuracy and speed. By revealing the complexity of cybercrime trends and providing a foundation for predictive testing, this research aims to improve the security of individuals, communities and nations. This research contributes to ongoing efforts to protect cyberspace in an age where digital threats are becoming more significant.

II.CRIME ANALYSIS

A) DATASET

The dataset used in this study is "National Crime Records Bureau (NCRB) Cyber Crime Dataset" provided by Is India National Crime Records Bureau (NCRB). This data includes cybercrime incidents reported in India in 2021. It covers various types of cybercrime, including but not limited to hacking, online fraud, malware attacks, and various digital crimes. There are now three major crimes in this case. Category A is an offense under the IT Act and includes computer crime, cybercrime, sexual harassment in electronic form and other areas of the IT Act. Category B refers to offenses under the IPC, including cyberstalking, doxxing, forgery, forgery, fraud and other offences. Category C is an offense under SLL and includes offenses such as Gambling (Online), Lotteries Act (Online), Act (1957) and other sll offences.

The number of crimes for each state is listed and the state total for all three categories is also stated, and the total number of crimes for each state in 2021 is also listed. Before analysis, the data set went through a preliminary step that included the elimination of useful and negative values. Geographic information was created to enable identification within the state. Access to this information is obtained from the official NCRB website and is subject to the data usage terms and conditions set by the NCRB.

B) ANALYSIS

In this study, we leveraged the power of Microsoft Power BI to perform analysis of the National Crime Records Bureau (NCRB) 2021 Cyber Data set. The aim is to provide my valuable insights and patterns from this rich literature that reveals various aspects of cybercrime in India.

Power BI is known for its data visualization capabilities and forms the basis of our analytics framework. We transform raw data into easy-to-understand explanations through diverse and interactive charts, graphs and dashboards. This information not only provides a deeper understanding of the data, but also enables stakeholders to gain quick and intuitive understanding.

The findings address various aspects of cybercrime, including time trends, geographic distribution, and the prevalence of certain types of cybercrime. A chart shows the evolution of cybercrime incidents over a six-year period, while a heat map shows regional hotspots of cyber activity. Pie charts and graphs provide a better understanding of the cyber threat landscape by clearly separating different categories of cybercrime by state or country.

These visual representations are not only research tools but also effective communication tools for discovery purposes. Using Power BI's drill-down capabilities, users can interact with data and zoom in to a specific time period, geographic area, or type of cybercrime for detailed analysis.

In summary, Power BI improves our understanding of cybercrime dynamics by playing a key role in transforming raw data into meaningful information. Visualizations created by this platform provide a visual explanation that supports informed decision-making and policy development in the field of cybersecurity. The combination of data exploration and visualization forms the basis for further analysis, including the use of machine learning algorithms for classification and prediction, as described in the next section of this study.

C) CLASSIFICATION

Classification is a data mining technique that classifies data to aid accurate prediction and analysis. It is one of the data mining methods designed to analyze large data sets. It is used to provide schemas that define the main data groups in the database. Classification involves predicting an outcome based on a given input. Classification algorithms attempt to find relationships between attributes so that results can be predicted. They analyze ideas and make predictions.

1) Support Vector Machines (SVMs) are a family of supervised learning methods renowned for their prowess in classification and regression tasks. At the heart of SVMs lies the concept of decision-making plans, which delineate boundaries between different class memberships within a dataset. These decision plans, often referred to as hyperplanes, efficiently segregate data points into distinct classes, making SVMs a quintessential example of hyperplane classifiers. SVMs, as hyperplane classifiers, operate in multidimensional feature spaces. They are primarily employed for classification tasks, where the objective is to classify data points into predefined categories.

2) Logistic regression serves as a fundamental statistical technique with widespread applicability in classification tasks. Differing from linear regression, which predicts continuous numeric values, logistic regression is specifically tailored for binary classification scenarios. In essence, it addresses questions with two possible outcomes, often expressed as 0 and 1, or "negative" and "positive." At its core, logistic regression operates by modelling the relationship between predictor variables and the probability of a binary outcome. It achieves this through the log-odds transformation, known as the logit function. The log-odds signify the natural logarithm of the odds of the positive outcome occurring. This mathematical transformation is pivotal in logistic regression, serving as the foundation for modelling probabilities.

3) Random Forest is a robust and versatile ensemble learning method that has gained prominence in various fields, including machine learning and data analysis. It stands as a powerful tool for both classification and regression tasks, known for its ability to deliver accurate and stable predictions. At its core, Random Forest is built upon the concept of an ensemble of decision trees. Decision trees are individual models that recursively split data into subsets based on the values of input features, ultimately leading to a decision or prediction.

III.METHODOLOGY

A) DATA COLLECTION

This study is based on obtaining general information about cyber crimes. The data selected for analysis is the National Crime Records Bureau (NCRB) Cyber Crime dataset, which includes cyber crime incidents in 2021. This dataset was obtained from the NCRB database in accordance with the Terms of Use and Conditions of Use. The columns in the database are as follows:

State/ut, Computer Crimes, Cyber Terrorism, Electronic Broadcasting/Obscene Conduct, Other Laws, Government Tag All (a), Cyber Stalking, Information Theft, Fraud, Fraud, Fraud, Online Extortion, Defamation, Others, All states (b), Gambling (Online), Lotteries Act (Online), Act 1957, Other Offenses, All States (c), All Crimes Year 2021, Risk

This column covers India in 2021 It provides an in-depth look at different types of cyber crimes reported across states and union territories.

In our study, we use three machine learning algorithms, namely support vector machine (SVM), logistic regression, and random forest, to complete the classification task.

B) PRELIMINARY DATA

Prior to analysis, rigorous preliminary data was conducted to ensure good information and communication was achieved. is necessary for the machine to operate. This phase includes handling missing values, handling implementation, and modeling geographic data to facilitate situation analysis.

C) DATA ANALYSIS

Data analysis means understanding all the features of the data set. This includes using data visualization tools, including Microsoft Power BI, to create visual explanations. Various graphs have been created to identify time trends, geographic distribution and expansion of various cybercrime groups.

D) FEATURE SELECTION AND MODEL SELECTION

Feature selection is used to identify relevant features for classification while removing unimportant or redundant features. It is also thought that feature engineering techniques will create new features that can increase the predictive power of the model. Use StandardScaler to apply a scaling factor to scale the model, ensuring they contribute equally to the model.

For the task of classifying cybercrime incidents into different groups, we chose three learning systems, namely support vector machine (SVM), logistic regression, and random forest classifier. Our machine learning models were selected based on their robustness and suitability to the task at hand. By comparing the performance of these models, we aim to determine the best way to classify cybercrime incidents and ultimately help identify, prevent and mitigate digital crime.

E) DATA SPLITTING

The data set is divided into training and testing subsets, with 80% of the data split to train the model and 20% retained to evaluate performance. This method provides an unbiased assessment of the model's generalization ability.

F) MODEL TRAINING AND EVALUATION

1) Logistic Regression

Logistic regression model plays an important role in this study because it learns from the data how different features are related to the probability of falling into a particular risk. After training, the model is used to make predictions on test data, providing an estimate of the level of risk based on the learned model. Calculate and evaluate the model's accuracy, which is a measure of how well the model distributes events. Additionally, we dive deeper into the model's performance using the Classification report, which provides precision, recall, F1 score, and support metrics for each risk to provide a better understanding of its effectiveness in distributing cybercrime.

Model Accuracy: 0.875				
	precision	recall	f1-score	support
0.0	0.86	1.00	0.92	6
1.0	1.00	0.50	0.67	2
accuracy			0.88	8
macro avg	0.93	0.75	0.79	8
weighted avg	0.89	0.88	0.86	8

2) Support Vector Machine

The support vector machine initializes the inline SVM model, which is suitable for two classification tasks. This model fits the training data and allows it to learn patterns in the data. Predictions are made from the test data and the accuracy of the model is calculated to evaluate its performance. The final rule publishes the model's accuracy score and detailed information providing accuracy, return, F1 score, and support measures for each risk, thus providing an overview of SVM's performance in classifying cybercrime incidents.

Model Accuracy: 0.75				
	precision	recall	f1-score	support
0.0	0.83	0.83	0.83	6
1.0	0.50	0.50	0.50	2
accuracy			0.75	8
macro avg	0.67	0.67	0.67	8
weighted avg	0.75	0.75	0.75	8

3) Random Forest model

The Random Forest Classifier, with 100 decision trees and a specified random state for reproducibility, is initialized. The model is trained on the training data, allowing it to learn patterns within the dataset. Subsequently, the model is used to make predictions on the test data and the accuracy of the model is computed. The code snippet concludes by printing the model's accuracy score and generating a comprehensive classification report. This report includes precision, recall, F1-score, and support metrics for each risk category, providing an in-depth evaluation of the Random Forest Classifier's effectiveness in categorizing cybercrime incidents.

Model Accuracy: 0.96				
	precision	recall	f1-score	support
0.0	0.96	0.96	0.96	2
1.0	0.96	0.96	0.96	2
accuracy			0.96	4
macro avg	0.96	0.96	0.96	4
weighted avg	0.96	0.96	0.96	4

G) MODEL COMPARISON

Comparison of the accuracy achieved by three different machine learning models: Logistic Regression, Support Vector Machine (SVM), and Random Forest is done using a bar chart. Among the models, Random Forest achieved the highest accuracy, reaching a accuracy score of 0.96. In contrast, Logistic Regression and SVM exhibited lower accuracies, achieving accuracies of 0.875 and 0.75, respectively.

IV. EXPERIMENTAL RESULTS

A) DATASET OVERVIEW

The dataset under investigation, sourced from the National Crime Records Bureau (NCRB) for the year 2021, offers a comprehensive perspective on cybercrime incidents reported across various states and union territories in India. It encompasses a diverse array of cybercrime categories, shedding light on the evolving landscape of digital offenses.

Key Dataset Statistics

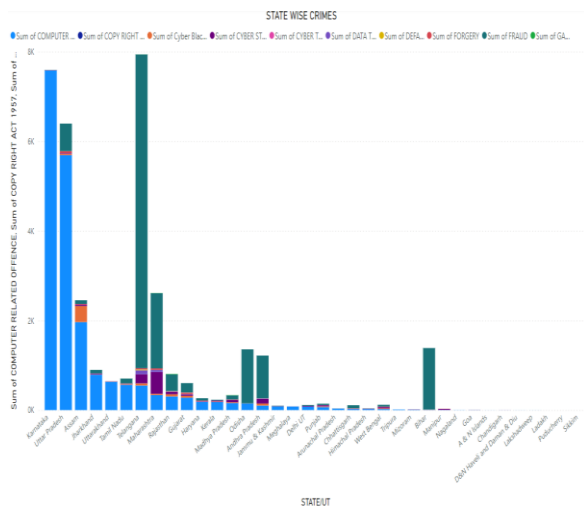
Number of States/Union Territories: 36

Total Reported Cybercrimes (2021): 52753

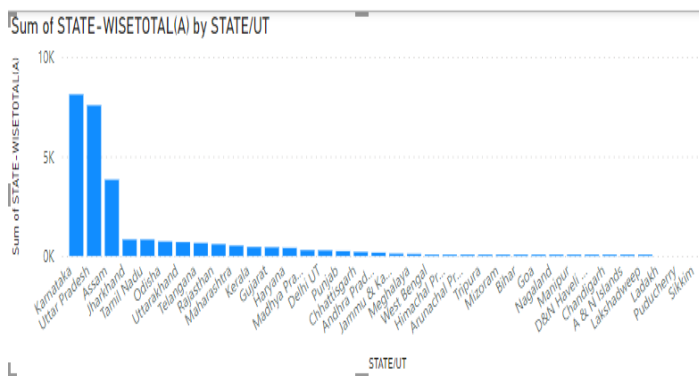
B) RESULTS

1) ANALYSIS

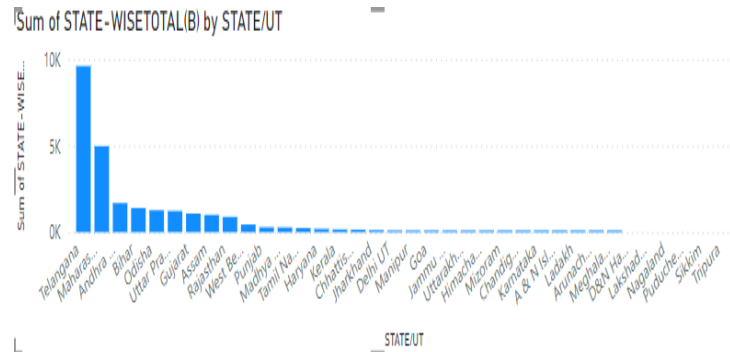
Firstly, we have compared all the crimes separately for each state through a bar chart. The representation is as follows



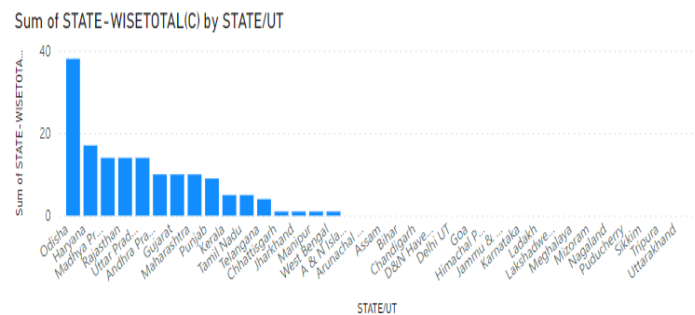
We have also compared the state wise total of each of the three categories of Crimes between the states and union territories using a bar chart which is the most common way of graphical representation.



In this figure we have compared the category OFFENCE UNDER THE IT ACT for all the states. Karnataka has the highest value of 8115.



In this figure we have compared the category OFFENCES UNDER IPC for all the states. Telangana has the highest value of 9605.



In this figure we have compared the category OFFENCES UNDER SLL for all the states. Odisha has the highest value of 38. From this we can infer that when compared to the other two categories of the crime this is less occurring.

Secondly, we have compared the total sum of all these three categories of crime for 2021 using both pie chart and a bar graph. The visualisation is as follows.

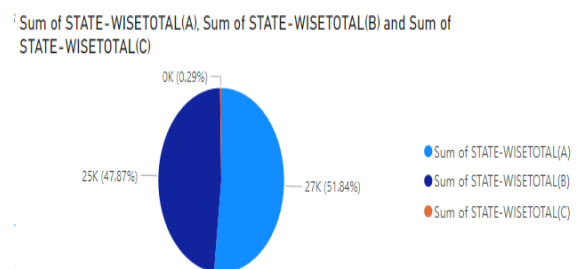


Figure a) Using a pie chart

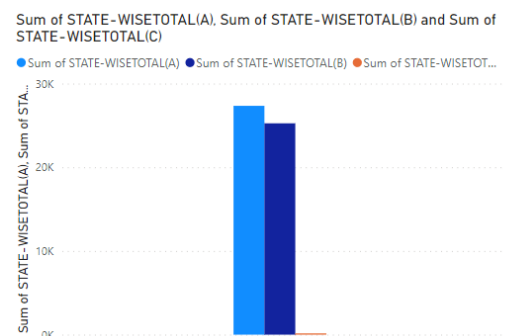
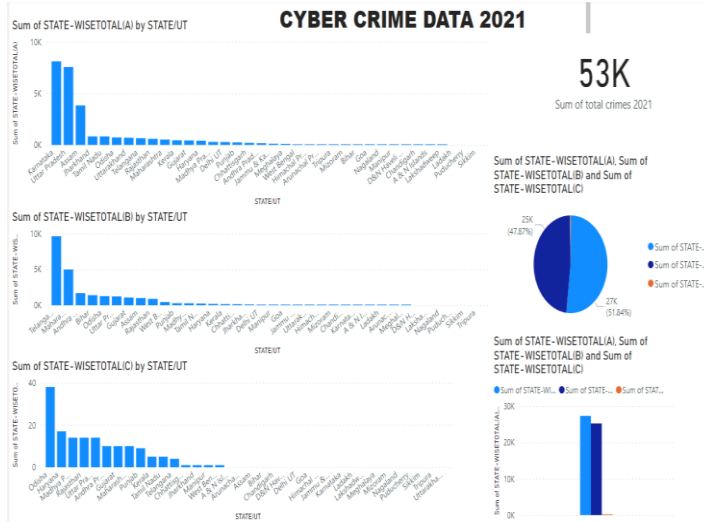


Figure b) using a bar chart

V.CONCLUSION

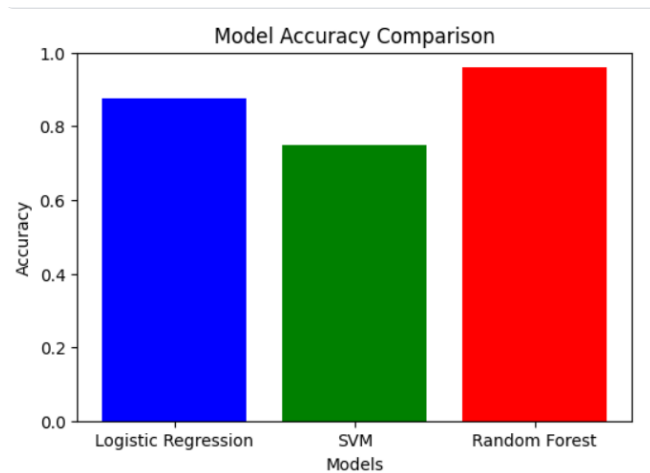
We infer that category A has a percentage of 51.84% which is the largest among all the three categories. Closely category B has a percentage of 47.87% and category C has the least percentage of 0.29%.

Finally, the total sum of all the categories of crimes in 2021 is about 53K.



DASHBOARD USING POWER BI

2) MODEL COMPARISON



The accuracy of Random Forest is found to be higher.

In an era characterized by the relentless proliferation of cyber threats, the utilization of advanced data analytics and machine learning techniques has become indispensable in understanding, mitigating, and combating cybercrimes. This research embarked on a journey to comprehend and address the intricate landscape of cybercrimes, leveraging the formidable capabilities of Support Vector Machine (SVM), Logistic regression and Random Forest model for classification. It is found that Telangana has the highest number of crimes (Approximately 10k) in the year 2021 followed by Karnataka. From this we can conclude that in future also Telangana will be the state with highest number of crimes as of our prediction.

The use of Power BI visualizations provided an intuitive visual narrative, shedding light on temporal trends, geographical distributions, and the prevalence of various cybercrime categories.

VI.REFERENCES

- [1] Crime Analysis and Prediction using Machine Learning, Olta Llaha, 2010.
- [2] Machine learning in crime prediction, Karabo Jenga, Cagatay Catal, Gorkem Kar, 2023.
- [3] Cybercrime: Identification and Prediction Using Machine Learning Techniques, K. Veena, K. Meena, Ramya Kuppusamy, Yuvaraja Teekaraman, Ravi V. Angadi, and Amruth Ramesh Thelkar, 2022.
- [4] Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention, Neil Shah, Nandish Bhagat, Manan Shah, 2021.
- [5] Dataset: https://ncrb.gov.in/sites/default/files/CII-2021/CII_2021Volume%201.pdf