ABSTRACT

In an generation wherein cybersecurity threats have end up more and more sophisticated, the want for sturdy prediction and detection structures to guard in opposition to cyber hacking breaches is paramount. This mission affords a singular technique to deal with this concern, using Machine Learning techniques, in particular the Random Forest Classifier, to are expecting and locate ability cyber hacking breaches. Implemented in Python, the proposed device makes use of a cautiously curated dataset of 5457 URLs, encompassing 87 extracted features. Crucially, the dataset maintains a balanced composition, precisely divided between 50% phishing and 50% legitimate URLs. The project's number one consciousness lies in correctly figuring out cyber threats at the same time as minimizing fake positives. Through rigorous schooling and evaluation, the carried out consequences display the system's fantastic performance. The Random Forest Classifier attains a commendable training accuracy of 99%, ensuring its ability to discern patterns and distinguish between legitimate and malicious URLs. The version additionally showcases a sturdy take a look at accuracy of 91%, similarly validating its reliability in real/international scenarios. In conclusion, this challenge stands as a pioneering attempt withinside the realm of cyber hacking breach prediction and detection, harnessing the strength of Machine Learning and the Random Forest Classifier to offer enhanced security measures. The notable accuracy done serves as a testimony to its effectiveness, empowering agencies to give a boost to their cybersecurity defenses in opposition to capability cyber threats and attacks.

Keywords: Cyber-Security, Machine Learning, Random Forest Classifier, And Phishing Detection.

I. INTRODUCTION

As the corporations have rapidly become the target of cyberattacks. Among the ransomware that causes such severe damage, the hackers may employ many attack types Today, maintaining system security, including the confidentiality of both corporate and personal data, is quite difficult. Millions of cyber attacks per day cause tremendous financial losses. Our study has three main objectives. Using actual cybercrime data, the initial phase is to predict a cybercrime strategy, and the accuracy results are then compared. The second is to examine if the information at hand can be used to predict cybercrime perpetrators. It is employed to hide a system's data. Information theft is caused by sensitive and highly confidential data as well as poor management. The hackers' techniques might be found in two different methods. One is to move through with legal action, get in touch with the victim, and let them know about the violations. The organizations should be aware of the sorts, trends, and patterns of assaults for the purpose of enabling them to monitor the system. We present a study on the consequences of these kinds of attacks in an effort for managing the prevention of occurring the breaches. We provide comprehensive study of the breaches that have occurred by the various organizations and financial effect Because of improvements in information technology, prices for memory and storage devices, and the expansion of the digital economy, businesses and governmental organizations now acquire more data every day. Businesses and organizations have the threat of data attacks because of the collecting of personal data on their computers. Computer networks are used in manufacturing, healthcare, research. This information is transferring every second through network. These attacks are used for profit and destroy the important information and use that for own need which rises the risk of data. Hybrid based detection is used to detect the high false positive rates and low false positive rates. Anomaly based detection analyse the behaviour of traffic, where signature based detection has the previous attacks records and able to detect the possibilities. It has high
likelihood of discovering undiscovered hazards. For finding breaches on massive amounts of data, typical
detection techniques are Ineffective. Machine learning techniques can increase the effectiveness of breach
detection. There are supervised and unsupervised methods of machine learning techniques. In supervised
method, the model will be trained by the labelled input. The model will distinguish between many classes
existing from the records of taken dataset, where as in another model system will be trained by the unlabelled
input in order to determine the content of exactly same existed in the dataset of input. Using Twitter data that
has been scraped and categorized, online social networks can serve as platforms and routes for exchanging
information. Methodology evaluates the relationship between user group sentiment and potential Cyberattacks.
We use a statistical modelling approach to address these problems and apply them for PRC data, which
demonstrates It shows that neither the number of breaches nor theirs frequency have grown over time.

II. METHODOLOGY

2.1 Methodology and Workflow :-

Data and Methodology:
The dataset used for prediction is taken from Kaggle as its data source. The dataset contains 300 instances with
organization, website and social network details and year, records, organization type and sources are the
attributes in the dataset.

B. Data Preprocessing:
In order to remove the null and duplicate values data preprocessing is performed. The adjustments made to the
data before we send for the particular algorithm is known as data preprocessing.

C. Feature Selection:
Finding the best feature from the features that are present in training data is the process of feature selection.
Correlation coefficient method is used for the feature selection.

D. Implementation:
We are using Decision Tree, Random Forest, K-Means and Multi-layer perceptron algorithms for predicting the
cyber breaches.

2.2 Requirements:-

Hardware requirements:
System : Pentium i3 Processor.
Hard Disk : 500 GB.
Monitor : 15” LED
Input Devices : Keyboard, Mouse
Ram : 4 GB

Software requirements:
Operating system : Windows 10.
Coding Language : Python 3.10.9.
Web Framework : Flask.

III. SYSTEM ARCHITECTURE

The complete system architecture can be described as the following components:
1. Data Collection and Preprocessing: It consists of data sources, data cleaning, and feature extraction.
2. Data Splitting: It divides the dataset into two sets, i.e., training and testing sets.
3. Machine Learning Model: Choosing a random forest classifier and then training the model using the training
datasets.
4. Evaluation: It does the testing of the model’s performance.
5. Real-time Monitoring and Detection: Real-time monitoring and detection of cyber threats.
Results

The output displays a pie-chart that shows two parts, i.e., legitimate and phishing.

V. CONCLUSION

The project, "Cyber Hacking Breaches Prediction and Detection Using Machine Learning," represents a full-size development within the subject of cybersecurity. By harnessing the power of Python programming language and employing the Random Forest Classifier algorithm, the proposed system has demonstrated exceptional accuracy, adaptability, and effectiveness in predicting and detecting cyber hacking breaches. The primary objective of the project was to address the limitations of earlier rule-based and signature-based
methods, which lacked the ability to adapt to evolving cyber threats and often produced high false positive rates. The proposed system overcomes these challenges by integrating advanced machine learning techniques, utilizing a balanced dataset, and focusing on feature-rich extraction to enhance its predictive capabilities. Throughout the project, meticulous attention was given to the dataset, which consisted of 5457 URLs with 87 extracted features. The dataset's balanced composition, comprising exactly 50% phishing and 50% legitimate URLs, played a pivotal role in minimizing bias and improving the system’s accuracy. The Random Forest Classifier, a powerful ensemble learning algorithm, was chosen for its ability to handle high-dimensional data, reduce overfitting, and adapt to diverse patterns in the dataset. The achieved results of the proposed system were truly remarkable. With a training accuracy of 99% and a test accuracy of 91%, the system demonstrated a high level of reliability in distinguishing between legitimate and malicious URLs. This level of accuracy significantly reduces false positives, ensuring minimal disruptions to legitimate URLs and enhancing user trust in the system's predictions. Moreover, the adaptability of the proposed system to emerging threats makes it a future-proof solution in the dynamic landscape of cybersecurity. Its ability to learn from diverse subsets of data and generalize to new attack vectors enables it to effectively detect novel threats and protect organizations from zero-day attacks. The project's success reaffirms the significance of integrating advanced machine learning techniques in cybersecurity endeavors. By leveraging the potential of Python and the Random Forest Classifier, the system presents a more sophisticated and scalable approach to address modern cybersecurity challenges. In conclusion, "Cyber Hacking Breaches Prediction and Detection Using Machine Learning” offers a pioneering solution to combat cyber threats.

VI. REFERENCES


