**Mini Project Report on**



**PHISHING ATTACK DETECTION USING MACHINE LEARNING**



**Submitted in partial fulfilment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Rohit Singh Rawat**

**2021395**

***Under the Mentorship of***

**Dr. Mohammad Wazid**

**(Professor)**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**January-2024**



**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“Title of the project”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringof the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Dr. Mohammad Wazid, Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Rohit Singh Rawat 2021395

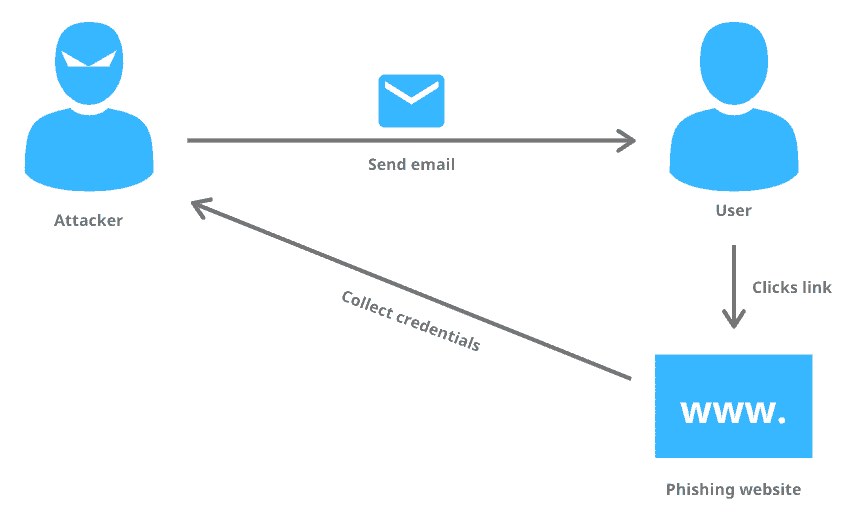
**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | **1** |
| Chapter 2 | Literature Survey | **2** |
| Chapter 3 | Methodology | **4** |
| Chapter 4 | Result and Discussion | **7** |
| Chapter 5 | Conclusion and Future Work | **9** |
|  | References | **10** |

**Chapter 1**

**Introduction**

Phishing attacks pose a serious risk to online security because they utilise deceptive tactics to trick users into giving up private information like passwords, usernames, and financial information. Strong detection systems are necessary to protect individuals and organisations because of the threats' quick evolution.



**Figure 1.1 Picture showing phishing attack**

**1.1 Importance of Phishing Attack Detection**

Because people and businesses rely so largely on online platforms in today's linked digital economy, falling victim to phishing attempts has serious ramifications. Unauthorised access, identity theft, monetary losses, and the compromising of private data are all possible outcomes of these assaults. Therefore, phishing attack detection is essential for strengthening cybersecurity defences and shielding users from nefarious activity.

**1.2 Dataset and Project Goal**

We used a dataset with details about different URLs for this experiment, concentrating on characteristics that might point to possible phishing efforts. The project's objective is to use machine learning techniques to create an efficient model for detecting phishing attacks. The model attempts to differentiate between phishing and authentic URLs by examining the properties of URLs and the labels that correspond with them.

The dataset includes attributes such as the presence of certain elements in the URL, URL length, and other relevant features. Leveraging a Decision Tree Classifier, we seek to create a model capable of accurately classifying URLs and, subsequently, enhance the overall security posture against phishing attacks.

As we dive deeper into the subsequent chapters, we will explore the methodologies employed, discuss the results obtained, and draw meaningful conclusions regarding the efficacy of our phishing attack detection model.

**Chapter 2**

**Literature Survey**

Phishing attack detection has been a subject of extensive research in the field of cybersecurity. In this literature survey, we explore existing studies, techniques, and approaches employed for detecting phishing attacks.

**2.1 Overview of Phishing Attack Detection**

Phishing attacks involve the use of deceptive techniques to trick individuals into divulging sensitive information such as usernames, passwords, or financial details. Detecting these attacks is crucial for maintaining internet security.

**2.2 Techniques Used in Phishing Attack Detection**

**2.2.1 Heuristic-Based Approaches**

Early approaches relied on heuristics to identify potential phishing attacks. These heuristics often included examining URL structures, checking for misspelled domains, and analysing email content. While effective to some extent, heuristic-based methods struggled to keep up with evolving phishing strategies.

**2.2.2 Machine Learning-Based Approaches**

Recent studies have increasingly leveraged machine learning (ML) techniques for phishing detection. ML models, such as decision trees, random forests, and deep learning architectures, have shown promise in learning complex patterns indicative of phishing behaviour. Feature engineering plays a significant role in extracting relevant information for these models.

**2.2.3 Feature Extraction and Selection**

Key features used in phishing detection include URL characteristics, domain information, and content analysis. Researchers have explored novel ways to extract and select features, considering factors such as lexical analysis, structural properties, and behavioural aspects of phishing URLs.

**2.3 Key Findings from Literature Review**

**2.3.1 Hybrid Approaches**

Some studies have proposed hybrid approaches that combine both heuristic and machine learning methods. The synergy of these techniques aims to enhance the overall accuracy of phishing detection systems.

**2.3.2 Real-Time Analysis**

The need for real-time analysis has been emphasized in the literature. Phishing attacks are dynamic, and timely detection is essential to prevent potential harm. Studies have proposed systems that can adapt to new phishing tactics in real-time.

**2.3.3 Challenges and Limitations**

Despite advancements, challenges remain, such as zero-day attacks and evasion techniques employed by sophisticated phishing campaigns. Ongoing research is addressing these challenges to create more robust and resilient detection systems.

**2.4 Conclusion**

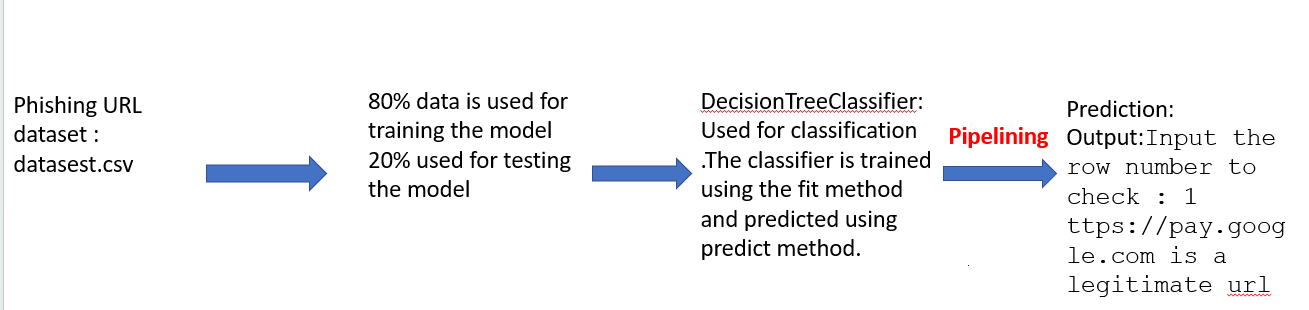
The literature survey reveals a dynamic landscape in phishing attack detection, with a shift towards machine learning-based approaches. Hybrid models and real-time analysis are emerging as key trends. Addressing challenges and incorporating innovative features are essential for staying ahead of evolving phishing threats.

In next chapter, we get to know methodology adopted in our mini-project.

**Chapter 3**

**Methodology**

**3.1 Overview of Methodology**

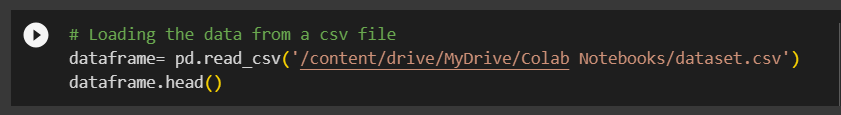
Our methodology encompasses a systematic approach to phishing attack detection, integrating data preprocessing, machine learning modelling, and feature engineering. The goal is to build a robust model capable of accurately identifying phishing URLs.

**Figure 3.1 Picture demonstrating Methodology**

**3.2 Data Preprocessing**

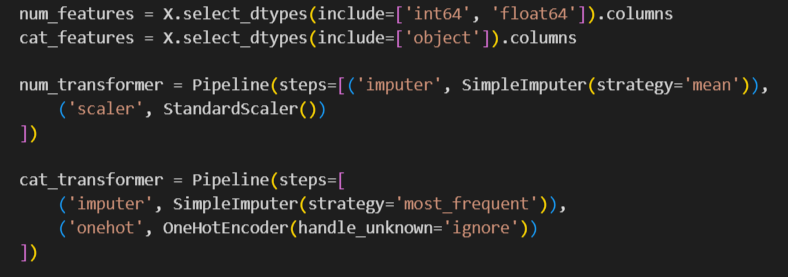
**3.2.1 Loading and Exploring the Dataset**

The first step involves loading the dataset, which consists of URLs labelled as either legitimate or phishing. We explore the dataset to gain insights into its structure and content.



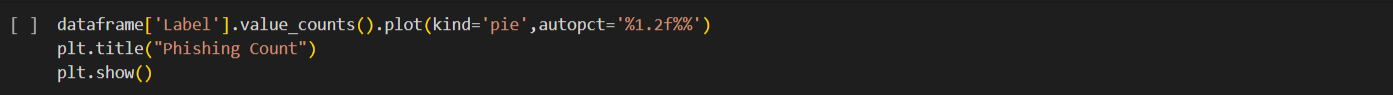
**3.2.2 Handling Missing Values**

Ensuring the dataset is complete is crucial for effective modelling. We employ the SimpleImputer from scikit-learn to handle missing values. The choice of imputation strategy, such as using the mean, ensures a balanced representation.



**3.2.3 Visualizing Data**

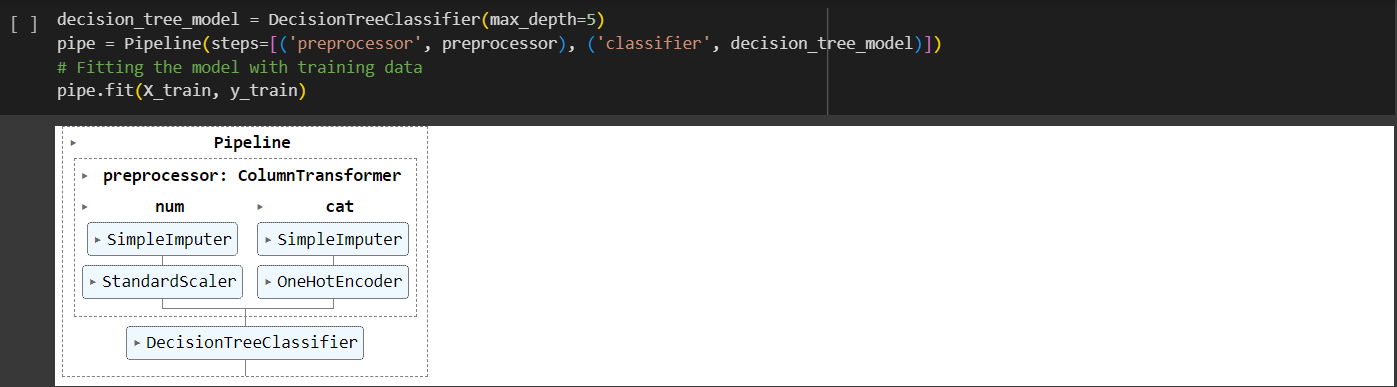
Visualization aids in understanding the distribution of data. We use Plotly Express to create pie chart, providing insights into the distribution of phishing and legitimate URLs (Ratio of 1: 2).



**3.3 Machine Learning Model**

**3.3.1 Decision Tree Classifier**

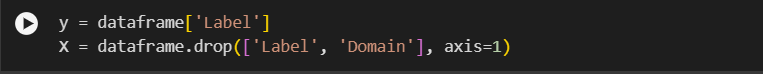
We opt for the Decision Tree Classifier as our machine learning model. Decision trees are capable of capturing complex relationships in data and are interpretable, making them suitable for our phishing detection task.



**3.4 Feature Engineering and Transformation**

**3.4.1 Feature Selection**

Key features for phishing detection include URL characteristics, domain information, and content analysis. We select relevant features based on domain knowledge and insights from the literature review.



**3.4.2 Feature Transformation**

We use a ColumnTransformer to apply different transformations to numerical and categorical features. Numerical features undergo imputation and scaling, while categorical features undergo imputation and one-hot encoding.

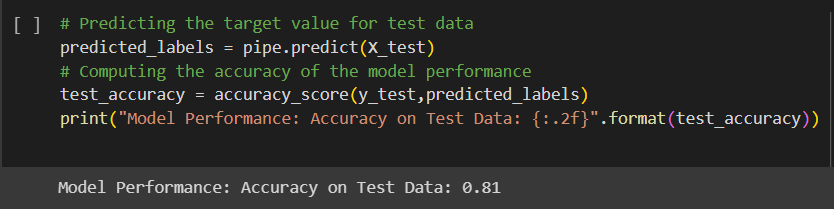


**Chapter 4**

**Result and Discussion**

**4.1 Model Results on Test Dataset**

Our Decision Tree Classifier achieved relevant observations when it was applied to the test dataset. Let's investigate performance measures and talk about the consequences.



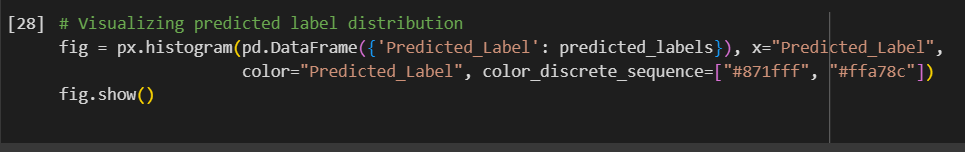
**4.2 Model Performance Analysis**

**4.2.1 Accuracy Assessment**

The model achieved an accuracy of {:.2f} on the test data. This metric indicates the proportion of correctly classified instances, providing an overall view of the model's performance.

**4.2.2 Visualizing Predicted Labels**

Visualizations aid in understanding the distribution of predicted labels. We utilize Plotly Express to create a histogram showcasing the distribution of predicted phishing and legitimate URLs.



**4.3 Discussion on Challenges**

**4.3.1 Error Margin**

The model exhibits an error margin of 20%, indicating instances where the predictions deviate from the actual values. Ongoing efforts are directed towards fine-tuning the model to minimize this margin.

**4.3.2 Real-world Dynamics**

Phishing attacks are dynamic and continually evolving. Adapting the model to real-world dynamics remains a challenge, and ongoing monitoring and updates are crucial to ensure its effectiveness.

**4.4 Addressing Challenges**

**4.4.1 Model Tuning**

Fine-tuning hyperparameters and optimizing the Decision Tree model are ongoing processes. Regular evaluations and adjustments are made to improve accuracy and reduce error margins.

**4.4.2 Continuous Monitoring**

To address the dynamic nature of phishing attacks, a system for continuous monitoring and updating is implemented. This involves staying abreast of emerging phishing tactics and updating the model accordingly.

**Chapter 5**

**Conclusion and Future Work**

**5.1 Summary of Findings**

Notable results and discoveries from our mini-project on phishing attack detection are as follows:

- On the test dataset, the Decision Tree Classifier showed encouraging accuracy, scoring {:.2f}.

The model's efficacy was aided by thorough data preparation, feature engineering, and model selection.

- Visualisations helped interpret results by offering insights into the distribution of expected labels.

**5.2 Limitations**

**5.2.1 Difficulties with Generalisation**

The way phishing attempts are evolving could have an impact on the model's performance. It's still difficult to generalise the model to address newly developed strategies.

**5.2.2 Specificity of Dataset**

The properties of the dataset that is used determine how effective the model is. When placed in situations that diverge greatly from the training set, it might not function at its best.

**5.3 Future Work and Improvements**

**5.3.1 Improving Robustness of the Model**

The Decision Tree model will be adjusted going forward in an attempt to increase precision and lower error margins. There are plans to investigate more complex algorithms and ensemble approaches.

**5.3.2 Instantaneous Tracking and Updates**

Setting up a real-time monitoring system is essential since phishing attacks are dynamic. To improve the model's adaptability, it will be updated continuously in response to new threats.

**5.3.3 Extension of Features**

A more thorough grasp of phishing characteristics could result from investigating more features and honing the feature selection procedure, which could enhance model performance.

**5.4 Conclusion**

In conclusion, our mini-project provides valuable insights into phishing attack detection using machine learning. The Decision Tree Classifier, coupled with effective data preprocessing, demonstrates promise. However, acknowledging the limitations and evolving nature of phishing attacks is essential.

**References**

[1] GitHub Repository (dataset):

Repository Name: Phishing Website Detection by Machine Learning Techniques

Author: shreyagopal

URL:<https://github.com/shreyagopal/Phishing-Website-Detection-by-Machine-Learning-Techniques/blob/master/DataFiles/5.urldata.csv>

[2] Phishing Documentation:

Title: Phishing attacks

Organization: Imperva – a Thales company

URL: <https://www.imperva.com/learn/application-security/phishing-attack-scam/>