|  |  |
| --- | --- |
|  | **Sri Lanka Institute of Information Technology**  ISP- (IE 3092) |

PROJECT CHARTER

|  |
| --- |
|  |

|  |  |  |
| --- | --- | --- |
| GROUP NUMBER |  | (will be assigned by the lecturer-in-charge) |

PROJECT GROUP MEMBER DETAILS:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STUDENT NAME | STUDENT ID. | CONTACT NO. | EMAIL ADDRESS  (SLIIT mail address) | SIGNATURE |
| W.A.C.S Weerasinghe | IT20609580 | 0763612980 | It20609580@my.sliit.lk | chanodya |

|  |  |
| --- | --- |
| Project Topic | Identifying phishing websites using machine learning |

PROJECT DETAILS

|  |
| --- |
| Brief Description of the proposed project: |
| Introduction:  Phishing is a form of cybercrime where attackers use fraudulent techniques to deceive individuals into providing sensitive information such as login credentials, personal data, and financial information. It is a growing problem that has become increasingly sophisticated over the years, with attackers using techniques such as social engineering and impersonation to trick users into thinking they are interacting with a legitimate website.  Project Objective:  The objective of this project is to develop a machine-learning model that can accurately identify phishing websites. The model will be trained on a dataset of known phishing websites and legitimate websites, with the aim of creating a classifier that can accurately distinguish between the two. The goal is to develop a model that is accurate, reliable, and efficient, and can be used to help protect users from falling victim to phishing attacks.  Methodology:  The project will involve the following steps:   1. Data Collection: The first step will be to collect a large dataset of known phishing websites and legitimate websites. This will involve using a variety of sources, such as public datasets, web crawling, and manual data collection. 2. Data Pre-processing: The collected dataset will be cleaned and pre-processed to remove any irrelevant or duplicated data. The data will then be divided into training, validation, and testing sets. 3. Feature Extraction: The next step will be to extract relevant features from the data. This will involve using techniques such as URL analysis, HTML analysis, and content analysis to extract features such as domain age, SSL certificate information, URL length, and content similarity. 4. Machine Learning Model Training: The extracted features will be used to train a machine learning model. The project will explore different machine learning algorithms such as decision trees, random forests, Multilayer Perceptrons, XGBoost, Support Vector Machines, and neural networks to identify the most accurate and efficient model. 5. Model Evaluation: The trained model will be evaluated using various metrics such as accuracy, precision, and recall. The model will be compared with other existing methods to identify its effectiveness and efficiency. 6. Model Deployment: Once the model is developed, it will be deployed as a browser extension that can be used by individuals or organizations to identify phishing websites.   Conclusion:  The proposed project aims to address the growing problem of phishing attacks by developing a machine-learning model that can accurately identify phishing websites. The project will involve collecting a large dataset of known phishing and legitimate websites, extracting relevant features, training, and evaluating a machine learning model, and deploying the model as a web-based application or browser extension. The project is expected to contribute to the development of effective solutions for combating phishing attacks, and ultimately help to protect individuals and organizations from falling victim to these attacks. |

|  |
| --- |
| Identified Problem |
| While the project of Identifying Phishing Websites using Machine Learning is a promising one, there are a few potential problems that may arise during the implementation of the project. Some of these problems are:   1. Data Availability: One of the major challenges in developing a machine learning model for phishing detection is the availability of a large dataset. It can be challenging to obtain a significant amount of high-quality data that includes both legitimate and phishing websites. Additionally, the dataset may be biased, which can lead to poor performance of the machine learning model. 2. Feature Extraction: The features extracted from the websites play a crucial role in the accuracy of the machine learning model. However, feature extraction can be a challenging task, and identifying the most relevant features for phishing detection may require domain expertise. Additionally, features that are useful for detecting phishing today may not be relevant in the future as phishing techniques evolve. 3. Class Imbalance: The number of legitimate websites is significantly larger than the number of phishing websites. This class imbalance can make it difficult for the machine-learning model to learn from the data and may result in poor performance. 4. Dynamic Nature of Phishing Attacks: Phishing attacks are constantly evolving, and attackers are using new techniques to deceive users. This dynamic nature of phishing attacks means that the machine learning model needs to be regularly updated to stay effective. 5. Ethical Considerations: While the project aims to combat phishing attacks, it is essential to consider ethical considerations such as user privacy and data security. The data collected for the project must be handled appropriately, and the application must be transparent in its operation.   In conclusion, while Identifying Phishing Websites using Machine Learning is a promising project, there are several challenges that need to be addressed to ensure its success. These challenges include data availability, feature extraction, class imbalance, the dynamic nature of phishing attacks, and ethical considerations. By addressing these challenges, the project can achieve its objective of developing an effective solution for combating phishing attacks. |

|  |
| --- |
| Proposed Solution |
| To address the identified problems in the project of Identifying Phishing Websites using Machine Learning, the following proposed solutions can be implemented:   1. Data Availability: To overcome the challenge of data availability, we can use a combination of public datasets and web crawling to obtain a large dataset of legitimate and phishing websites. We can also incorporate user feedback to ensure that the dataset remains up to date. 2. Feature Extraction: We can use a combination of domain knowledge and automated techniques to extract relevant features from the websites. We can also use feature selection techniques to identify the most important features for detecting phishing attacks. 3. Class Imbalance: To overcome the problem of class imbalance, we can use techniques such as oversampling, undersampling, or a combination of these to balance the classes. 4. Dynamic Nature of Phishing Attacks: To ensure that the machine learning model stays effective, we can use a continuous learning approach where the model is regularly updated with new data. We can also incorporate techniques such as ensemble learning, where multiple models are combined to improve the overall performance. 5. Ethical Considerations: To address ethical considerations, we can ensure that the data collected is anonymized and handled appropriately. We can also make the application transparent in its operation and provide users with control over their data.   In conclusion, by implementing the proposed solutions, we can overcome the challenges identified in the project of Identifying Phishing Websites using Machine Learning. This will enable us to develop an effective solution for combating phishing attacks that is accurate, reliable, and efficient. |

Timeline (Please provide a brief description about the time line e.g. project charter submission, proposal submission, concept paper submission, progress presentations and final thesis submission, etc)

|  |  |
| --- | --- |
|  | ………………………………………………………………………………………………………………………………………………………… |
| **Chart, bar chart  Description automatically generated** | |

REFERENCE: (IEEE Format)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | [1] | A. Alswailem, B. Alabdullah. N. Alrumayh, A. Alsedrani, "Detecting Phishing Websites Using Machine Learning," in *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, Riyadh, 2019. | | [2] | R. Mahajan, I. Siddavatam, "Phishing Website Detection using Machine Learning Algorithms," in *International Journal of Computer Applications*, Mumbai, 2018. | | [3] | S. Jain, Dr. N. Choudhary, K. Jain, "Phishing Websites Detection Using Machine Learning," 2022. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=4121102. | | [4] | A. K. Dutta, "Detecting phishing websites using machine learning technique," 2021. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0258361. | |

…............................................................................................................................................................

|  |  |
| --- | --- |
| EVALUVATOR COMMENTS | |
|  |  |