# Project Design Phase-I

**Proposed Solution Template**

|  |  |
| --- | --- |
| Date | 30 September 2022 |
| Team ID | PNT2022TMID19048 |
| Project Name | Project – Web Phishing Detection |
| Maximum Marks | 2 Marks |

**Proposed Solution Template:**

Project team shall fill the following information in proposed solution template.

|  |  |  |
| --- | --- | --- |
| **S.No.** | **Parameter** | **Description** |
|  | Problem Statement (Problem to be solved) | Phishing detection techniques do suffer low detection accuracy and high false alarm especially when novel phishing approaches are introduced. Besides, the most common technique used, blacklist-based method is inefficient in responding to emanating [phishing attacks](https://www.sciencedirect.com/topics/computer-science/phishing-attack) since registering new domain has become easier, no comprehensive blacklist can ensure a perfect up-to-date database. Furthermore, page content inspection has been used by some strategies to overcome the false negative problems and complement the vulnerabilities of the stale lists. Moreover, page content inspection algorithms each have different approach to [phishing website detection](https://www.sciencedirect.com/topics/computer-science/website-phishing-detection) with varying degrees of accuracy. Therefore, ensemble can be seen to be a better solution as it can combine the similarity in accuracy and different error-detection rate properties in selected algorithms. |
|  | Idea / Solution description | A phishing website is a common social engineering method that mimics trustful uniform resource locators (URLs) and webpages. The objective of this project is to train machine learning models and deep neural nets on the dataset created to predict phishing websites. Both phishing and benign URLs of websites are gathered to form a dataset and from them required URL and website content-based features are extracted. The performance level of each model is measures and compared. |
|  | Novelty / Uniqueness | Before stating the ML model training, the data is split into 80-20 i.e., 8000 training samples & 2000 testing samples. From the dataset, it is clear that this is a supervised machine learning task. There are two major types of supervised machine learning problems, called classification and regression. |
|  | Social Impact / Customer Satisfaction | The website shows information regarding the services provided by us. It also contains information regarding ill- practices occurring in todays technological world. The website is created with an opinion such that people are not only able to distinguish between legitimate and fraudulent website, but also become aware of the mal-practices occrring in current world. They can stay away from the people trying to exploit ones personal information, like email address, password, debit card numbers, credit card details, CVV, bank account numbers, and the list goes on. |
|  | Business model | Although the use of URL lexical features alone has been shown to result in high accuracy (97%), phishers have learned how to make predicting a URL destination difficult by carefully manipulating the URL to evade detection. Therefore, combining these features with others, such as host, is the most effective approach . |
|  | Scalability of solution | Today's growing phishing websites pose significant threats due to their extremely undetectable risk. They anticipate internet users to mistake them as genuine ones in order to reveal user information and privacy, such as login ids, pass-words, credit card numbers, etc. without notice. |