

DETECTION OF PHISHING WEBSITES FROM URLs

Milestone 1: Project Initialization and Planning Phase

The "Project Initialization and Planning Phase" marks the project's outset, defining goals, scope, and stakeholders. This crucial phase establishes project parameters, identifies key team members, allocates resources, and outlines a realistic timeline. It also involves risk assessment and mitigation planning. Successful initiation sets the foundation for a well-organized and efficiently executed machine learning project, ensuring clarity, alignment, and proactive measures for potential challenges.

Activity 1: Define Problem Statement

Problem Statement: Phishing websites have become a significant threat to online security, compromising sensitive information and financial data of million of users worldwide. The rapid growth of phishing attacks has made it challenging for users to distinguish legitimate websites from fraudulent.

Ref. template: [Click Here](#)

Detection of phishing websites from urls Problem Statement Report: [Click Here](#)

Activity 2: Project Proposal (Proposed Solution)

The proposal for the detection of phishing websites from URLs project is to develop a system that can automatically identify and classify URLs as phishing or legitimate by collecting and preprocessing a large dataset of labeled URLs, developing a machine learning model to learn patterns and features from the URLs, evaluating the model's performance, and implementing a web application or API to predict phishing websites in real-time, with the goal of providing a robust and efficient solution to protect users from phishing attacks and improve online security.

Ref. template: [Click Here](#)

Detection of phishing websites from urls Project Proposal Report: [Click Here](#)

Activity 3: Initial Project Planning

It involves defining the project's objectives, scope and stakeholders, conducting a feasibility study, establishing a preliminary budget and resource allocation plan, developing a high-level project schedule and timeline, identify and assigning project roles and responsibilities, and defining communication and stakeholder management plans, with high accuracy and efficiency, and delivering a project charter, scope statement, and preliminary project plan outlining the objectives, scope, timeline, budget, resources, and roles.

Ref. template: [Click Here](#)

Detection of phishing websites from URLs Project Planning Report: [Click Here](#)

Milestone 2: Data Collection and Preprocessing Phase

The Data Collection and Preprocessing Phase involves executing a plan to gather relevant loan

application data from Kaggle, ensuring data quality through verification and addressing missing values. Preprocessing tasks include cleaning, encoding, and organizing the dataset for subsequent exploratory analysis and machine learning model development.

Activity 1: Data Collection Plan, Raw Data Sources Identified, Data Quality Report

The dataset for "detection of phishing websites from URLs " is sourced from Kaggle. It includes having IP address, URL length, domain, sub domain etc. Data quality is ensured through verification, addressing missing values, and maintaining adherence to ethical guidelines, establishing a reliable foundation for predictive modeling.

Ref. template: [Click Here](#)

Detection of phishing websites from URLs Data Collection Report: [Click Here](#)

Activity 2: Data Quality Report

The dataset for "Detection of phishing websites from URLs" is sourced from Kaggle. It includes URL strings, label, feature extracted from URLs. Data quality is ensured through thorough verification, addressing missing values, and maintaining adherence to ethical guidelines, establishing a reliable foundation for predictive modeling.

Ref. template: [Click Here](#)

Detection of phishing websites from URLs Data Quality Report: [Click Here](#)

Activity 3: Data Exploration and Preprocessing

Data Exploration involves analyzing the detection of phishing websites from URLs dataset to understand patterns, distributions, and outliers. Preprocessing includes handling missing values, scaling, and encoding categorical variables. These crucial steps enhance data quality, ensuring the reliability and effectiveness of subsequent analyses in the detection of phishing websites from URLs project.

Ref. template: [Click Here](#)

Detection of phishing websites from URLs Data Exploration and Preprocessing Report: [Click Here](#)

Milestone 3: Model Development Phase

The Model Development Phase entails crafting a predictive model for detecting phishing websites. It encompasses strategic feature selection, evaluating and selecting models (Random Forest, Logistic regression), initiating training with code, and rigorously validating and assessing model performance for informed decision-making in the detecting phishing websites.

Activity 1: Feature Selection Report

The Feature Selection Report outlines the rationale behind choosing specific features (e.g., having IP address, URL length, domain etc..) for detecting phishing websites model. It evaluates relevance, importance, and impact on predictive accuracy, ensuring the inclusion of key factors influencing the model's ability to detect phishing websites .

Ref. template: [Click Here](#)

Detection of phishing websites from URLs Feature Selection Report: [Click Here](#)

Activity 2: Model Selection Report

The model selection process for detecting phishing websites from URLs involved evaluating and comparing the performance of various machine learning algorithms, including Logistic regression and random forest. After thorough evaluation, random forest emerged as the top – performing model, achieving an accuracy in terms of precision, recall, and robustness. Random forest's ability to handle high-dimensional data, non -linear relationships, and feature interactions made it an idea choice for this problem.

Ref. template: [Click Here](#)

Detection of phishing websites from URLs Model Selection Report: [Click Here](#)

Activity 3: Initial Model Training Code, Model Validation and Evaluation Report

The initial model training, validation, and evaluation report for detecting phishing websites from URLs involved training a random forest classifier on a dataset of 10000 labeled URLs (80% training, 20% testing) using scikit learn library in python. The model was validated using cross-validation and evaluated using metrics such as accuracy, precision, recall, f1 score, demonstrating robust performance and effectiveness in detecting phishing websites. The code used for training and evaluation is available in the appendix, showcasing the implementation of the random forest classifier and evaluation metrics

Ref. template: [Click Here](#)

Detection of phishing websites from URLs Model Development Phase Template: [Click Here](#)

Milestone 4: Model Optimization and Tuning Phase

The Model Optimization and Tuning Phase involves refining machine learning models for peak performance. It includes optimized model code, fine-tuning hyperparameters, comparing performance metrics, and justifying the final model selection for enhanced predictive accuracy and efficiency.

Activity 1: Hyperparameter Tuning Documentation

The Gradient Boosting model was selected for its superior performance, exhibiting high accuracy during hyperparameter tuning. Its ability to handle complex relationships, minimize overfitting, and optimize predictive accuracy aligns with project objectives, justifying its selection as the final model.

Activity 2: Performance Metrics Comparison Report

The Performance Metrics Comparison Report contrasts the baseline and optimized metrics for various models, specifically highlighting the enhanced performance of the Gradient Boosting model. This assessment provides a clear understanding of the refined predictive capabilities achieved through hyperparameter tuning.

Activity 3: Final Model Selection Justification

The Final Model Selection Justification articulates the rationale for choosing Gradient Boosting as the ultimate model. Its exceptional accuracy, ability to handle complexity, and successful hyperparameter tuning align with project objectives, ensuring optimal loan approval predictions.

Ref. template: [Click Here](#)

Detection of phishing websites from URLs Model Optimization and Tuning Phase Report:

[Click Here](#)

Milestone 5: Project Files Submission and Documentation

For project file submission in Github, Kindly click the link and refer to the flow. [Click Here](#)

For the documentation, Kindly refer to the link. [Click Here](#)

Milestone 6: Project Demonstration

In the upcoming module called Project Demonstration, individuals will be required to record a video by sharing their screens. They will need to explain their project and demonstrate its execution during the presentation.