Phishing Website Classification

(Data Mining Project)

Authors Himanshu Yadav, Swastika Tiwari

Faculty Advisor Prof Sharanjit Kaur

Affiliations Computer Science Department, Acharya Narendra Dev College, Delhi University

Introduction

Phishing is a deceptive online practice where attackers impersonate legitimate entities to steal sensitive information. It causes significant financial and reputational damage to individuals and organizations

Traditional methods of phishing detection, such as blacklisting and manual inspection, are often inadequate

This project explores the potential of machine learning to automate phishing website detection and improve accuracy

Objective

To identify the most influential features and build models with high accuracy to assist in automated phishing detection systems.

Methodology

Data Preprocessing

- Load ARFF file using liac-arff library.
- Convert data to pandas DataFrame.

Feature Selection

- Selected the most impactful features using domain specific knowledge.
- 15 attributes dropped and 14 attributes kept.

Model Building

- Training Testing dataset division(75%, 25%)
- Train classification models: Decision Tree, K-Nearest Neighbors.
- Optimize model parameters using cross-validation.

Model Evaluation

Evaluate performance using:
- Accuracy
- Confusion Matrix
- Classification Report

Dataset and Features

- **Source : UC Irvine Machine Learning Repo**
- **Size : dataset of 11,056 web URLs with lexical and security features**
- **Key Attributes : SSLfinal_State, URL_Length, annd web_traffic**
- **Lexical: URL Length, having_IP_Address, having_At_Symbol, double_slash_redirecting, HTTPS_token.**
- **Host-based: SSLfinal_State, Domain_registeration_length, age_of_domain, DNSRecord.**
- **Content-based: Request_URL, URL_of_Anchor, SFH, popUpWidnow, web_traffic**
- **Irrelevant features like Shortining_Service, Prefix_Suffix were dropped to prevent data leakage and overfitting, ensuring better model generalization.**

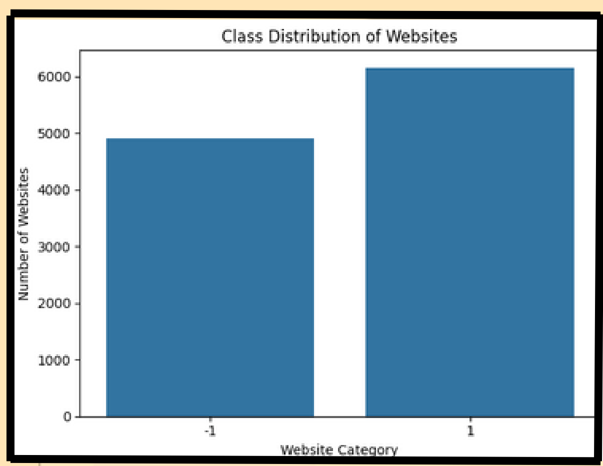

Figure 1. Count of Phishing vs Non Phishing Websites

here -1 indicates Phishing websites and 1 indicates Non-Phishing based websites

Results

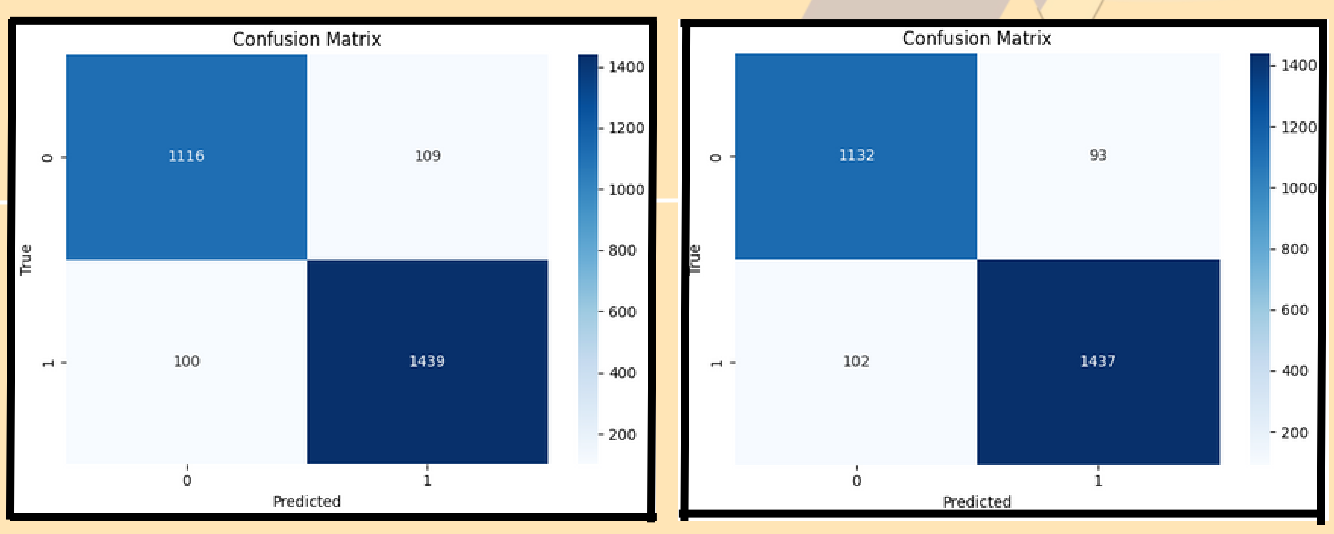

Figure 2. Confusion Matrix of (a) knn vs (b) decision tree model

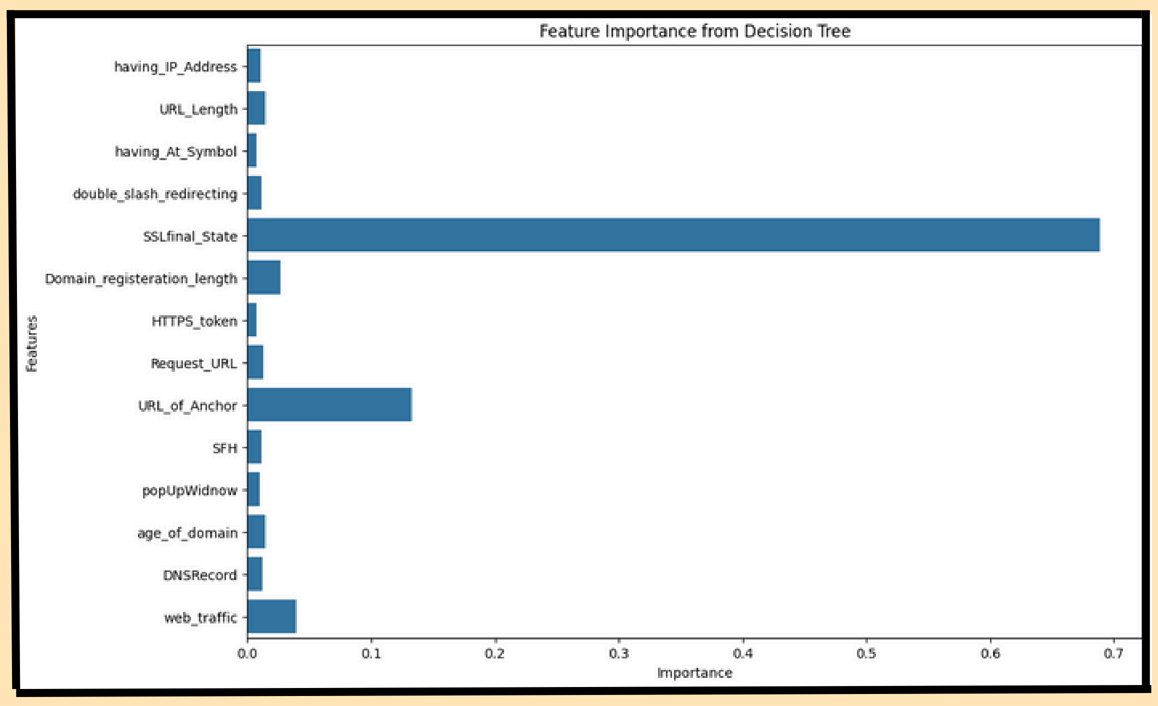

Figure 3.Feature importance using decision tree

- SSL certificate (Secure Sockets Layer) is one of the most critical factors .
- A secure and valid SSLfinal_State suggests that the website is likely to be legitimate.
- A missing, invalid, or suspicious SSLfinal_State strongly signals a phishing attempt.

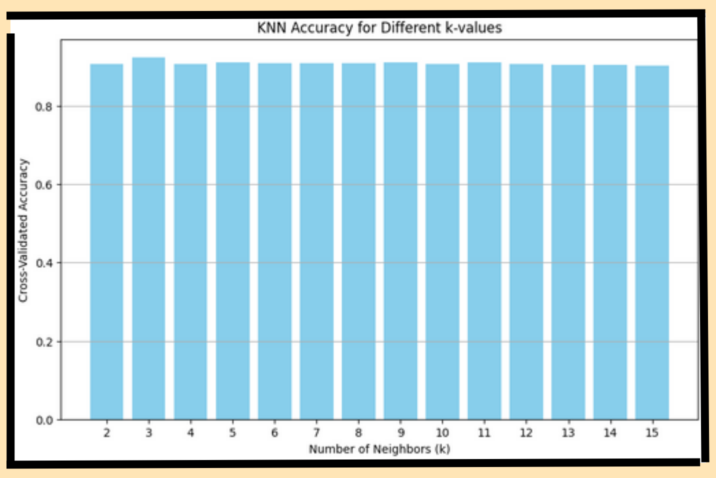

Figure 4: knn accuracy for k values

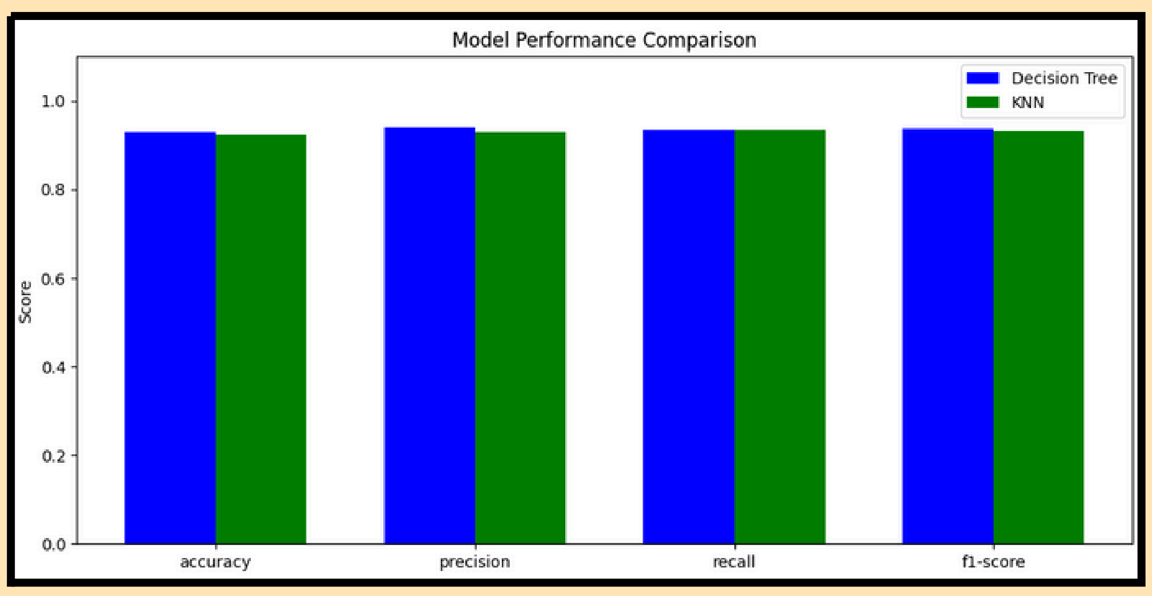

Figure 5. Model performance comparison

Conclusion

- Manually identified influential features for phishing website detection.
- Both the Decision Tree and KNN models perform well, achieving comparable accuracies of 92.94% and 92.43%, respectively.
- Demonstrated the potential of automated systems for phishing detection.

References

1. Tan P.N., Steinbach M, Karpatne A. and Kumar V. Introduction to Data Mining, Second edition, Sixth Impression, Pearson, 2023.