

Updated Report: Phishing Website Classification

Updated Report

Project discussion: Phishing Website Dataset

The dataset contains 1353 websites, each described using 10 attributes:

SFH, popUpWindow, SSLfinal_State, Request_URL, URL_of_Anchor, web_traffic, URL_Length, age_of_domain, having_IP_Address, Result

Data Preprocessing, Exploratory Analysis, and Feature Selection

Prepared by Yuliia Nykyporets

- Data extraction from .arff
- Type conversion and cleaning
- Correlation analysis and matrix
- PCA (Principal Component Analysis)
- Feature selection using SelectFromModel

Result: removal of web_traffic, PCA reduction to 5 components, and identification of top features via Decision Tree.

Main Analysis: ML Models and Classification

ML implementation by Olena Polishchuk

To classify websites into phishing / suspicious / legitimate categories, several ML models were implemented:

- Random Forest
- Gradient Boosting
- Support Vector Machine (SVM)
- Neural Networks

Updated Report: Phishing Website Classification

- Ensemble combinations

All models were implemented and validated using scikit-learn in Python. Cross-validation was applied.

Cross-validation

Conducted by Olena Polishchuk

Pipelines were developed to test models under consistent settings and ensure robustness.

Classification Tree

Designed and implemented by Yuliia Nykyporets

Two approaches tested:

- Recursive Binary Splitting
- Pruning with alpha via GridSearchCV

SVM

Implemented by Olena Polishchuk

Used SVC with RBF kernel, with and without feature selection.

Neural Networks & Ensembles

Implemented by Olena Polishchuk

- MLPClassifier used
- Voting/stacking ensemble approaches tested

Summary of Results

Model	Train Accuracy	Test Accuracy	Author
-------	----------------	---------------	--------

-----	-----	-----	-----
-------	-------	-------	-------

Updated Report: Phishing Website Classification

Decision Tree	-	-	Yuliia Nykyporets
SVM (RBF)	0.9536	0.8837	Olena Polishchuk
Random Forest	0.9691	0.8904	Olena Polishchuk
Gradient Boosting	0.9448	0.9172	Olena Polishchuk
Neural Networks	0.9294	0.9083	Olena Polishchuk
Ensemble	0.9636	0.8613	Olena Polishchuk

Conclusion

This project explored ML models for phishing detection. Responsibilities were divided between feature engineering (Yuliia Nykyporets) and model implementation (Olena Polishchuk).