Machine Learning-Based Phishing Website Detection

High-Level Design (HLD)

System Overview

- **Data Layer**: Kaggle dataset, URL lexical and metadata features
- **Processing Layer**: Preprocessing, Feature Engineering, ML Model Training, Evaluation
- **Application Layer**: Streamlit/Web App for URL checking

Detailed Design (DLD)

Input Data

10,000 rows × 50 features (`Phishing_Legitimate_full.csv`), label column: **CLASS_LABEL** (1 = Phishing, 0 = Legitimate).

Preprocessing

- Drop ID columns
- Median imputation for missing values
- Feature scaling with StandardScaler (for LR)

Feature Set

- **Lexical**: NumDots, SubdomainLevel, UrlLength, NumDash, IpAddress
- **Metadata**: NoHttps, HostnameLength, PctExtHyperlinks, AbnormalFormAction
- **Behavioral**: PopUpWindow, RightClickDisabled, SubmitInfoToEmail

Model Pipeline

1. Load dataset & split (80/20)

2. Train models: Logistic Regression, Random Forest, Gradient Boosting
3. Evaluate with Accuracy, Precision, Recall, F1, ROC-AUC
4. Save best model (Random Forest)

Flowcharts / Diagrams (Text Format)

Workflow Diagram

```
User URL → Feature Extraction → Preprocessing → ML Model → Prediction → Result
```

Data Flow Diagram (DFD)

```
Level 0: User ⇔ Phishing Detection System ⇔ Dataset

Level 1:
Data Input → Feature Extractor → Preprocessor → Classifier → Output
```

Process Flow (Training Phase)

```
Start
  ↓
Load Dataset
  ↓
Preprocess Data
  ↓
Train Models (LR, RF, GB)
  ↓
Evaluate Models
  ↓
Save Best Model
  ↓
End
```

Process Flow (Prediction Phase)

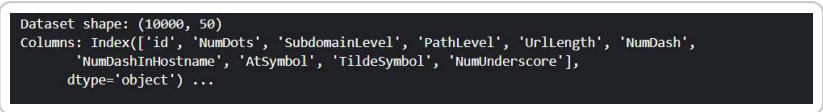
```
Start
  ↓
User Enters URL
  ↓
Extract Features
  ↓
Load Saved Model
  ↓
Predict Phishing/Legitimate
  ↓
Show Result
  ↓
End
```

Results

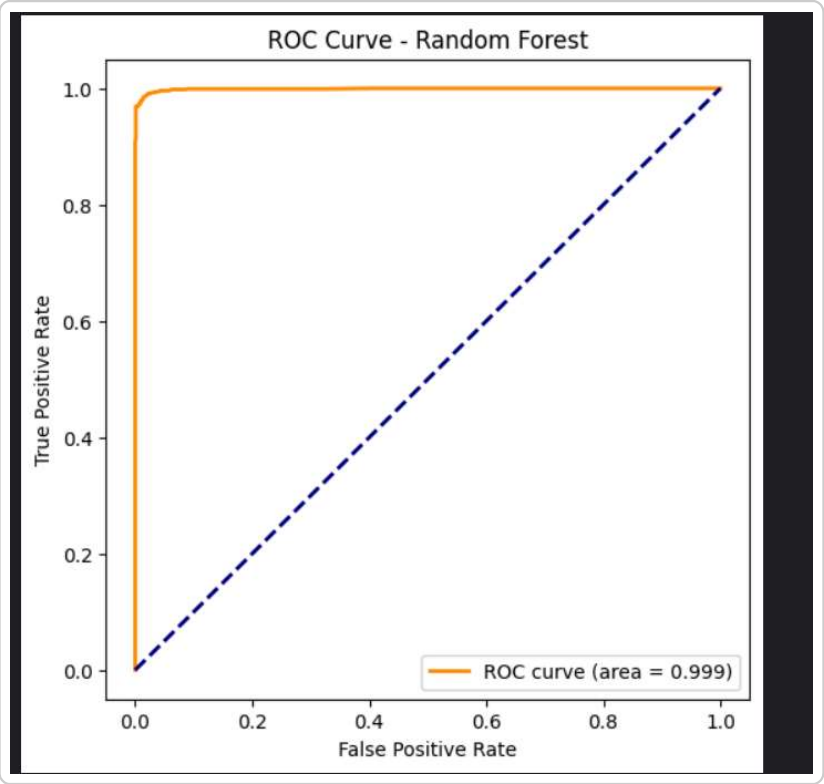
Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	96.7%	0.96	0.97	0.967	0.987
Random Forest	**99.4%**	0.99	0.99	0.994	0.999
Hist. Gradient Boost	98.9%	0.98	0.99	0.989	0.998

Visual Results

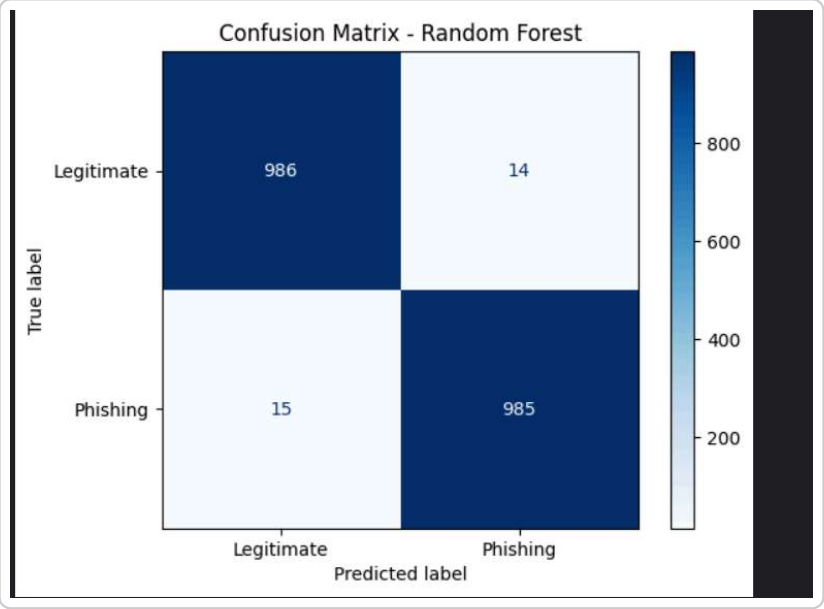
Dataset Description



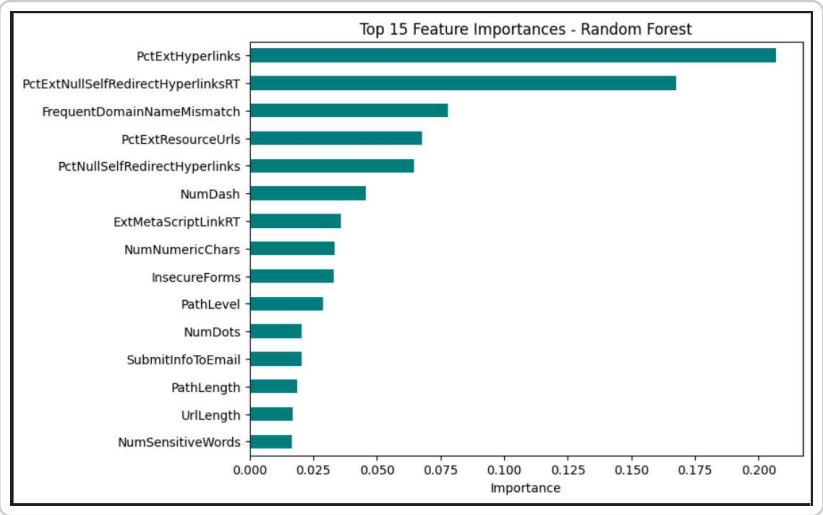
ROC Curve



Confusion Matrix



Feature Importances



References (IEEE Format)

1. N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection based associative classification data mining," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5948–5959, 2014.
2. Y. Zhang, J. Hong, and L. F. Cranor, "CANTINA: A content-based approach to detecting phishing websites," in *Proc. 16th Int. Conf. World Wide Web (WWW)*, 2007, pp. 639–648.
3. D. Miyamoto, et al., "An evaluation of machine learning-based methods for detection of phishing sites," in *Proc. APWG eCrime Researchers Summit*, 2008.
4. S. Marchal, G. Armano, et al., "PhishStorm: Detecting phishing with streaming analytics," *IEEE Trans. Comput.*, vol. 65, no. 5, pp. 1352–1365, 2016.
5. Kaggle, "Phishing Website Dataset," [Online]. Available: https://www.kaggle.com/dataset