# **Cybersecurity : Phishing Detection**

## **Dataset Overview**

* **Dataset:** Web Page Phishing Dataset
* **Total Rows:** 11,430
* **Total Columns:** 89
* **Target Variable:** status (labels: phishing, legitimate)
* **Missing Values:** None found
* **Balanced Classes:** 5,715 phishing and 5,715 legitimate samples

## **Exploratory Data Analysis (EDA)**

### **Dataset Structure**

* Categorical columns: url, status
* Most features are numerical indicators of URL characteristics.

### **Descriptive Statistics**

* Used .describe() to check distributions, outliers, and skewness.
* Visualized numeric features using histograms.
* Correlation matrix plotted with seaborn to assess feature relationships.

## **Feature Selection Using PCA**

1. **Non-relevant Columns Dropped:** url, status
2. **Standardized Data:** StandardScaler
3. **PCA Applied:** Extracted top 10 components.
4. **Top 10 features per PC** collected based on absolute weights.
5. **Final Selected Features (appeared ≥ 2 times):**

longest\_word\_path, nb\_slash, empty\_title, page\_rank, domain\_age, nb\_dots, domain\_in\_brand, prefix\_suffix, etc.

* **Reduced dataset shape:** (11,430 rows × 35 columns)

## **Model Building & Evaluation**

### **Models Trained:**

* Random Forest
* Logistic Regression
* SVM (RBF)
* KNN (k=5)
* Neural Network (Keras)
* XGBoost (optional)

### **Base Accuracy Comparison:**

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Random Forest | 95.5% |
| Logistic Regression | 91.4% |
| SVM (RBF Kernel) | 93.6% |
| KNN | 92.9% |
| Neural Network | 94.3% |

### **Metrics Used:**

* Accuracy
* Precision
* Recall
* F1-score
* Confusion Matrix

## **K-Fold Cross-Validation (5-Fold)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy ± SD** | **Precision ± SD** | **Recall ± SD** | **F1 ± SD** |
| Random Forest | 95.28 ± 0.25 | 95.17 ± 0.73 | 95.40 ± 0.33 | 95.28 ± 0.23 |
| Logistic Regression | 90.73 ± 0.38 | 91.32 ± 0.94 | 90.04 ± 0.74 | 90.67 ± 0.36 |
| SVM (RBF) | 93.70 ± 0.13 | 94.47 ± 0.42 | 92.84 ± 0.47 | 93.65 ± 0.14 |
| KNN (k=5) | 92.76 ± 0.37 | 93.98 ± 0.70 | 91.39 ± 0.50 | 92.66 ± 0.37 |

## **Neural Network (Keras)**

* **Architecture:**
  + Dense(64, ReLU) → Dropout(0.3)
  + Dense(32, ReLU) → Dropout(0.2)
  + Output: Dense(1, Sigmoid)
* **Loss:** Binary Crossentropy
* **Optimizer:** Adam
* **Best Accuracy:** ~94.6% validation accuracy
* **Test Accuracy:** 94.3%

## **Outlier Detection & Re-training**

### **IQR Method:**

* Reduced dataset to 2,970 records
* Model accuracy dropped slightly due to reduced support size

### **Z-score Method:**

* Final shape: 9,195 rows
* Models retrained; RF accuracy: 94.4%, SVM: 93.8%

## **Interactive Dashboard (Dash + Plotly)**

### **Dashboard Features:**

* Feature Histogram & Box Plot
* Feature Mean Comparison (by class)
* Correlation Heatmap
* Status Distribution Pie Chart
* Accuracy Comparison Bar Plot
* Confusion Matrix Viewer
* Neural Network Training Plots
* Performance Table

The dashboard makes this project highly presentable and deployment-ready for demo or real-world use.

## **Conclusion & Insights**

* Feature reduction using PCA effectively narrowed down crucial indicators.
* Random Forest and Neural Network models showed the best performance.
* Cross-validation and multiple model evaluations ensured robust validation.
* Dash-based dashboard created a user-friendly interface for exploring results.
* Outlier removal slightly reduced data size but validated model resilience.

**Author**

Ananya P S

20221CSD0106