Chasing Phishing URLs

Final Project Report

Martinus Boom
uc2024154464@student.uc.pt

UNIVERSIDADE D
COIMBRA

Pattern Recognition
2024/2025

Prof. César Alexandre Domingues Teixeira
cteixei@dei.uc.pt

Department of Informatics Engineering
University of Coimbra

May 11, 2025

Contents

1 Introduction

The objective of this project is to develop machine learning classifiers for detecting phishing URLs. In Part 1, the focus was on data preprocessing, feature engineering, dimensionality reduction, and implementing Minimum Distance Classifiers (MDC) with Fisher Linear Discriminant Analysis (LDA). The PhiUSIIL Phishing URL Dataset was used for this purpose. In Part 2, the project was extended by implementing and evaluating additional classifiers, including Linear Bayes (LDA), Quadratic Bayes (QDA), k-Nearest Neighbors (k-NN), and Support Vector Machines (SVM). Each classifier was integrated with the same pipeline structure used in Part 1, ensuring consistency. These models were ranked during cross-validation and later tested on the held-out test set to assess their generalization and compare their performance in a consistent evaluation process.

2 Dataset Description

The PhiUSIIL Phishing URL Dataset consists of 235,795 labeled samples, comprising 134,850 legitimate and 100,945 phishing URLs. Each sample includes 54 numerical features extracted from URL strings and HTML content.

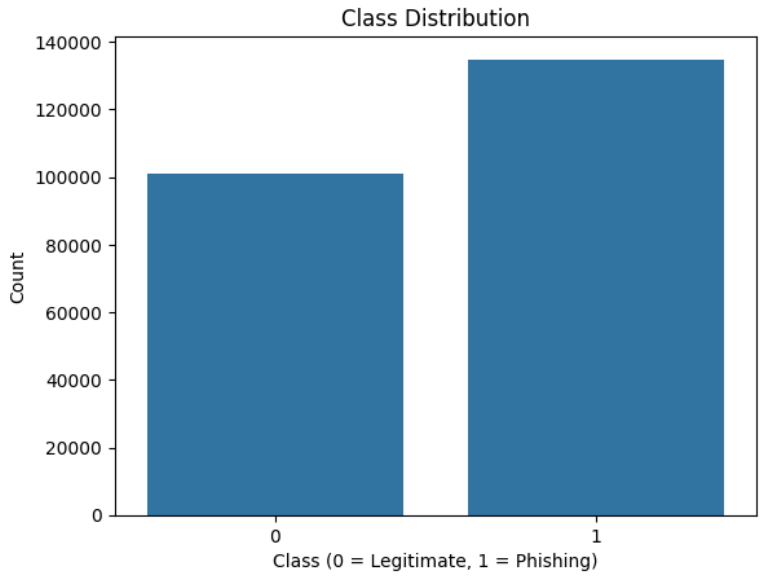

Figure 1: Class Distribution (0 = Legitimate, 1 = Phishing)

3 Data Preprocessing

The preprocessing phase included exploratory analysis, data splitting, scaling, feature selection, and dimensionality reduction techniques. An initial exploratory data analysis (EDA) confirmed the absence of missing values and duplicate records. It was also identified that the dataset was imbalanced, with phishing URLs representing approximately 42.8% of the total samples.

During the EDA, it was observed that the label vector `y` had an unexpected two-dimensional shape, which could interfere with the learning algorithms. To correct this, the `squeeze()` function was applied, converting it to a one-dimensional array.

After confirming the data structure, the dataset was split into 70% for training and 30% for testing. The training set was used in cross-validation for feature engineering and model selection, while the test set was held out for the final evaluation of the selected models.

3.1 Data Scaling

Feature scaling was applied to the non-categorical features to standardize their ranges before applying selection or reduction techniques. Two scaling methods were tested:

- **StandardScaler**: standardizes features by removing the mean and scaling to unit variance.
- **MinMaxScaler**: scales features to a range between 0 and 1.

3.2 Feature Selection and Dimensionality Reduction

Two distinct strategies were implemented for feature engineering:

1. **Scaling followed by Feature Selection**, where the most informative features were selected after scaling the data.
2. **Scaling followed by PCA Reduction**, where dimensionality was reduced using Principal Component Analysis (PCA) after scaling.

Feature selection and dimensionality reduction were applied separately. PCA was not applied after feature selection to avoid unnecessary transformations and to keep the feature selection results interpretable.

The feature selection techniques included:

- **Kruskal-Wallis Test**: non-parametric statistical test used to identify features significantly associated with the class labels (p-value < 0.05).
- **Fisher Score**: selecting features with high discriminative power between classes, based on the ratio of between-class variance to within-class variance (threshold > 0.01).
- **AUC per Feature**: computing the individual ROC-AUC for each feature and selecting those with AUC greater than 0.6.
- **RFE with SVM**: Recursive Feature Elimination using a linear SVM estimator (LinearSVC), selecting the top-ranked features based on a wrapper approach.
- **LASSO and Ridge (Logistic Regression)**: using L1 and L2-regularized logistic regression to select features with non-zero coefficients, controlling sparsity with the penalty term.
- **mRMR (Minimum Redundancy Maximum Relevance)**: selecting features with high relevance to the target and low redundancy among themselves, using a greedy approximation with mutual information.
- **Correlation Filtering**: removing features with Pearson correlation greater than 0.7.
- **Random Forest Importance**: selecting features with importance greater than 0.01.
- **Mutual Information**: selecting features with mutual information scores greater than 0.01.
- **RFECV** (Recursive Feature Elimination with Cross-Validation) was performed using LDA as the estimator because LDA was the core model used in the classification pipeline. Using it ensures the feature selection process is aligned with the model's behavior.

Dimensionality reduction was performed using Principal Component Analysis (**PCA**), retaining 95% of the explained variance in the dataset.

3.3 Handling Class Imbalance

Due to class imbalance, the Synthetic Minority Over-sampling Technique (**SMOTE**) was applied during cross-validation on the training folds. SMOTE creates new samples of the minority class by interpolating between existing instances. It was applied only to the training data in each fold to prevent data leakage. The validation and test sets were not modified and were used as they were for model evaluation.

4 Fisher LDA and Minimum Distance Classification

As stated in the introduction, the primary goal of this phase was to implement Minimum Distance Classifiers (**MDC**) combined with Fisher Linear Discriminant Analysis (**LDA**). LDA was applied as a dimensionality reduction technique, projecting the data into a one-dimensional space that maximizes the separability between the phishing and legitimate classes. This projection provided a simplified representation of the data while retaining the most relevant information for classification.

On the reduced LDA space, classification was performed using the **Euclidean** variant of the Minimum Distance Classifier. The decision was based on the distance between each sample and the class centroids in the LDA-transformed space, with each sample assigned to the class with the nearest centroid.

5 Linear and Quadratic Bayes Classification

The Linear Bayes classifier was implemented using the Linear Discriminant Analysis (**LDA**) model, interpreted under a Bayesian framework. It assumes that each class follows a multivariate normal distribution with a shared covariance matrix. Classification is performed by computing the posterior probabilities for each class and assigning the sample to the one with the highest probability. This results in linear decision boundaries in the feature space.

The Quadratic Bayes classifier was implemented using the Quadratic Discriminant Analysis (**QDA**) model. Unlike LDA, QDA uses a different covariance matrix for each class, which allows the decision boundaries to adapt to the shape of each class. However, this can cause problems when the selected features are highly correlated or have low variance, making the covariance matrices hard to invert. To prevent this, a small regularization value (`reg_param` $= 0.01$) was added. When invalid probability values still occurred, the evaluation metrics were set as `NaN` so that the feature sets would still be included in the overall comparison.

6 k-Nearest Neighbors (k-NN) Classification

The k-Nearest Neighbors (**k-NN**) classifier was implemented using a distance-based approach, assigning a class to a sample based on the majority label among its 25 nearest neighbors in the training set, using Euclidean distance.

A fixed value of $k = 25$ was used consistently across all experiments, instead of performing cross-validation for hyperparameter selection. This decision ensured uniform comparison between feature sets while keeping the evaluation time tractable.

Unlike probabilistic models such as Bayes classifiers, k-NN does not assume any data distribution. It makes decisions purely based on the spatial proximity of labeled samples in the feature space.

7 Support Vector Machine (SVM) Classification

The Support Vector Machine (**SVM**) classifier was implemented using the kernel-based formulation provided by the `scikit-learn` library. Classification was performed by identifying the optimal separating hyperplane that maximizes the margin between classes in the transformed feature space. The `rbf` (Radial Basis Function) kernel was used by default, allowing for non-linear decision boundaries.

SVM was trained independently for each feature set, relying on the same evaluation pipeline as the other classifiers. Class probabilities were estimated using Platt scaling, enabling ROC-AUC computation. Due to its ability to handle high-dimensional data and model complex class boundaries, SVM serves as a powerful baseline for comparison against both generative and distance-based classifiers.

8 Experimental Setup

8.1 Cross-Validation for Training and Validation

A 5-fold stratified cross-validation was performed on 70% of the dataset. Feature engineering, including scaling and either feature selection or dimensionality reduction, was applied within each fold to prevent data leakage. As mentioned previously, SMOTE was applied only to the training data in each fold to handle class imbalance. After SMOTE, each classifier was trained using the transformed data. For Minimum Distance Classifier (MDC), Fisher Linear Discriminant Analysis (LDA) was applied before classification. For all other classifiers (Linear Bayes, Quadratic Bayes, k-NN, and SVM), training was performed directly on the resampled data. Metrics were computed on the validation folds and averaged across all folds.

8.2 Evaluation Metrics

Model performance was evaluated using the following metrics, computed during cross-validation and on the test set:

- **Accuracy**: Overall proportion of correct predictions.
- **Sensitivity (Recall)**: Proportion of phishing URLs correctly identified.
- **Specificity**: Proportion of legitimate URLs correctly identified.
- **F1-score**: Harmonic mean of precision and recall.
- **ROC-AUC**: Area under the Receiver Operating Characteristic curve, indicating discriminative ability.

8.3 Model Ranking and Test Set Evaluation

All classifiers models were ranked according to their cross-validation metrics, using ROC-AUC as the primary criterion, followed by F1-score and Accuracy as tie-breakers. This ranking logic was used because ROC-AUC provides a reliable measure of the model's ability to distinguish between classes, while F1-score and Accuracy offer complementary insights into balanced performance and overall correctness. The same ranking criteria were applied after evaluation on the 30% test set. Differences in ranking between cross-validation and test set results were analyzed to assess generalization ability.

9 Results and Discussion

9.1 Results

9.1.1 Cross-Validation Results

The following tables present the cross-validation results for each classifier individually. Each classifier was evaluated across different feature sets, and performance metrics were averaged over the cross-validation folds.

Feature Set	**ROC-AUC**	**F1-score**	**Accuracy**	**Sensitivity**	**Specificity**
Selected by RFECV (StandardScaler)	0.5689	0.9992	0.9991	0.9999	0.9980
Selected by RFECV (MinMaxScaler)	0.5688	0.9992	0.9991	0.9999	0.9980
Kruskal (StandardScaler)	0.5666	0.9992	0.9991	0.9999	0.9980
LASSO_L1 (StandardScaler)	0.5666	0.9992	0.9991	0.9999	0.9980
LASSO_L2 (StandardScaler)	0.5666	0.9992	0.9991	0.9999	0.9980
Filtered by correlation (StandardScaler)	0.5666	0.9992	0.9991	0.9999	0.9980
Kruskal (MinMaxScaler)	0.5662	0.9992	0.9991	0.9999	0.9980
LASSO_L1 (MinMaxScaler)	0.5662	0.9992	0.9991	0.9999	0.9980
LASSO_L2 (MinMaxScaler)	0.5662	0.9992	0.9991	0.9999	0.9980
Filtered by correlation (MinMaxScaler)	0.5662	0.9992	0.9991	0.9999	0.9980
Selected by Mutual Information (StandardScaler)	0.5649	0.9992	0.9991	0.9999	0.9980
Selected by Mutual Information (MinMaxScaler)	0.5646	0.9992	0.9991	0.9999	0.9980
All scaled features (StandardScaler)	0.5645	0.9993	0.9992	1.0000	0.9981
Fisher (StandardScaler)	0.5643	0.9990	0.9989	0.9999	0.9979
All scaled features (MinMaxScaler)	0.5640	0.9993	0.9992	1.0000	0.9981
Fisher (MinMaxScaler)	0.5640	0.9990	0.9989	0.9999	0.9979
AUC (StandardScaler)	0.5502	0.9988	0.9986	0.9999	0.9967
AUC (MinMaxScaler)	0.5501	0.9988	0.9986	0.9999	0.9967
Selected by RF importance (MinMaxScaler)	0.5469	0.9987	0.9985	0.9999	0.9966
Selected by RF importance (StandardScaler)	0.5469	0.9987	0.9985	0.9999	0.9965
All scaled features (MinMaxScaler) + PCA	0.4589	0.9942	0.9934	0.9969	0.9852
mRMR (StandardScaler)	0.4563	0.9886	0.9870	0.9902	0.9808
mRMR (MinMaxScaler)	0.4562	0.9886	0.9870	0.9902	0.9808
All scaled features (StandardScaler) + PCA	0.4167	0.9965	0.9960	0.9987	0.9811
RFE_SVM (MinMaxScaler)	0.3381	0.9728	0.9694	0.9780	0.9485
RFE_SVM (StandardScaler)	0.3378	0.9726	0.9691	0.9776	0.9483

Table 1: Cross-Validation Results for Euclidean MDC Models

Feature Set	ROC-AUC	F1-score	Accuracy	Sensitivity	Specificity
AUC (StandardScaler)	1.0000	0.9988	0.9986	0.9999	0.9972
AUC (MinMaxScaler)	1.0000	0.9988	0.9986	0.9999	0.9972
Kruskal (StandardScaler)	1.0000	0.9992	0.9991	0.9999	0.9980
LASSO_L1 (StandardScaler)	1.0000	0.9992	0.9991	0.9999	0.9980
LASSO_L2 (StandardScaler)	1.0000	0.9992	0.9991	0.9999	0.9980
Filtered by correlation (StandardScaler)	1.0000	0.9992	0.9991	0.9999	0.9980
Selected by RFECV (StandardScaler)	1.0000	0.9992	0.9991	0.9999	0.9980
Kruskal (MinMaxScaler)	0.9999	0.9992	0.9991	0.9999	0.9980
LASSO_L1 (MinMaxScaler)	0.9999	0.9992	0.9991	0.9999	0.9980
LASSO_L2 (MinMaxScaler)	0.9999	0.9992	0.9991	0.9999	0.9980
Filtered by correlation (MinMaxScaler)	0.9999	0.9992	0.9991	0.9999	0.9980
Selected by RFECV (MinMaxScaler)	0.9999	0.9992	0.9991	0.9999	0.9980
All scaled features (StandardScaler)	0.9999	0.9993	0.9992	1.0000	0.9981
All scaled features (MinMaxScaler)	0.9999	0.9993	0.9992	1.0000	0.9981
All scaled features (StandardScaler) + PCA	0.9999	0.9965	0.9960	0.9987	0.9811
Selected by Mutual Information (StandardScaler)	0.9999	0.9992	0.9991	0.9999	0.9980
Selected by Mutual Information (MinMaxScaler)	0.9999	0.9992	0.9991	0.9999	0.9980
Fisher (StandardScaler)	0.9999	0.9990	0.9989	0.9999	0.9979
Selected by RF importance (StandardScaler)	0.9999	0.9987	0.9985	0.9999	0.9965
Selected by RF importance (MinMaxScaler)	0.9999	0.9987	0.9985	0.9999	0.9966
Fisher (MinMaxScaler)	0.9999	0.9990	0.9989	0.9999	0.9979
All scaled features (MinMaxScaler) + PCA	0.9998	0.9942	0.9934	0.9969	0.9852
mRMR (StandardScaler)	0.9996	0.9886	0.9870	0.9902	0.9808
mRMR (MinMaxScaler)	0.9996	0.9886	0.9870	0.9902	0.9808
RFE_SVM (MinMaxScaler)	0.9970	0.9728	0.9694	0.9780	0.9485
RFE_SVM (StandardScaler)	0.9969	0.9726	0.9691	0.9776	0.9483

Table 2: Cross-Validation Results for Linear Bayes Models

Feature Set	ROC-AUC	F1-score	Accuracy	Sensitivity	Specificity
Selected by RF importance (StandardScaler)	1.0000	0.9978	0.9974	0.9999	0.9948
Selected by RFECV (StandardScaler)	0.9999	0.9995	0.9994	0.9999	0.9988
All scaled features (StandardScaler)	0.9998	0.9991	0.9990	1.0000	0.9980
AUC (StandardScaler)	0.9998	0.9966	0.9962	0.9985	0.9879
Selected by Mutual Information (StandardScaler)	0.9998	0.9983	0.9980	0.9994	0.9960
mRMR (StandardScaler)	0.9997	0.9864	0.9847	0.9911	0.9680
Kruskal (StandardScaler)	0.9997	0.9989	0.9988	0.9999	0.9977
LASSO_L1 (StandardScaler)	0.9997	0.9989	0.9988	0.9999	0.9977
LASSO_L2 (StandardScaler)	0.9997	0.9989	0.9988	0.9999	0.9977
Filtered by correlation (StandardScaler)	0.9997	0.9989	0.9988	0.9999	0.9977
AUC (MinMaxScaler)	0.9997	0.9955	0.9948	0.9980	0.9834
Selected by RF importance (MinMaxScaler)	0.9997	0.9961	0.9955	0.9983	0.9859
Fisher (StandardScaler)	0.9996	0.9982	0.9980	0.9997	0.9959
Selected by Mutual Information (MinMaxScaler)	0.9996	0.9956	0.9950	0.9979	0.9832
All scaled features (MinMaxScaler)	0.9995	0.9953	0.9946	0.9978	0.9814
Selected by RFECV (MinMaxScaler)	0.9994	0.9952	0.9945	0.9977	0.9812
Kruskal (MinMaxScaler)	0.9994	0.9951	0.9944	0.9977	0.9810
LASSO_L1 (MinMaxScaler)	0.9994	0.9951	0.9944	0.9977	0.9810
LASSO_L2 (MinMaxScaler)	0.9994	0.9951	0.9944	0.9977	0.9810
Filtered by correlation (MinMaxScaler)	0.9994	0.9951	0.9944	0.9977	0.9810
Fisher (MinMaxScaler)	0.9993	0.9952	0.9945	0.9977	0.9812
All scaled features (MinMaxScaler) + PCA	0.9992	0.9915	0.9903	0.9963	0.9693
mRMR (MinMaxScaler)	0.9987	0.9840	0.9818	0.9912	0.9534
All scaled features (StandardScaler) + PCA	0.9980	0.9967	0.9963	0.9989	0.9822
RFE_SVM (StandardScaler)	0.9977	0.9784	0.9748	0.9881	0.9484
RFE_SVM (MinMaxScaler)	0.9962	0.9699	0.9647	0.9842	0.9306

Table 3: Cross-Validation Results for Quadratic Bayes Models

Feature Set	ROC-AUC	F1-score	Accuracy	Sensitivity	Specificity
Selected by RF importance (MinMaxScaler)	1.0000	0.9995	0.9994	1.0000	0.9988
Selected by RF importance (StandardScaler)	0.9999	0.9997	0.9996	1.0000	0.9991
Selected by Mutual Information (StandardScaler)	0.9998	0.9985	0.9982	0.9996	0.9958
All scaled features (StandardScaler)	0.9998	0.9983	0.9981	0.9996	0.9953
Kruskal (StandardScaler)	0.9998	0.9981	0.9978	0.9996	0.9949
LASSO_L1 (StandardScaler)	0.9998	0.9981	0.9978	0.9996	0.9949
LASSO_L2 (StandardScaler)	0.9998	0.9981	0.9978	0.9996	0.9949
Filtered by correlation (StandardScaler)	0.9998	0.9981	0.9978	0.9996	0.9949
mRMR (MinMaxScaler)	0.9998	0.9986	0.9984	0.9997	0.9970
Selected by RFECV (StandardScaler)	0.9998	0.9980	0.9977	0.9995	0.9948
All scaled features (StandardScaler) + PCA	0.9998	0.9977	0.9974	0.9994	0.9940
Fisher (StandardScaler)	0.9998	0.9977	0.9974	0.9994	0.9940
mRMR (StandardScaler)	0.9998	0.9990	0.9989	1.0000	0.9977
AUC (StandardScaler)	0.9997	0.9976	0.9972	0.9994	0.9932
Selected by Mutual Information (MinMaxScaler)	0.9994	0.9950	0.9943	0.9981	0.9823
Fisher (MinMaxScaler)	0.9994	0.9950	0.9943	0.9981	0.9823
Kruskal (MinMaxScaler)	0.9994	0.9950	0.9943	0.9981	0.9823
LASSO_L1 (MinMaxScaler)	0.9994	0.9950	0.9943	0.9981	0.9823
LASSO_L2 (MinMaxScaler)	0.9994	0.9950	0.9943	0.9981	0.9823
Filtered by correlation (MinMaxScaler)	0.9994	0.9950	0.9943	0.9981	0.9823
Selected by RFECV (MinMaxScaler)	0.9994	0.9951	0.9944	0.9981	0.9826
All scaled features (MinMaxScaler)	0.9994	0.9947	0.9940	0.9980	0.9807
AUC (MinMaxScaler)	0.9993	0.9945	0.9937	0.9979	0.9802
All scaled features (MinMaxScaler) + PCA	0.9992	0.9938	0.9929	0.9976	0.9783
RFE_SVM (StandardScaler)	0.9986	0.9920	0.9908	0.9969	0.9753
RFE_SVM (MinMaxScaler)	0.9980	0.9888	0.9871	0.9952	0.9663

Table 4: Cross-Validation Results for k-NN Models

Feature Set	ROC-AUC	F1-score	Accuracy	Sensitivity	Specificity
Selected by Mutual Information (MinMaxScaler)	1.0000	0.9999	0.9999	1.0000	0.9999
All scaled features (MinMaxScaler)	1.0000	0.9999	0.9998	1.0000	0.9996
Kruskal (MinMaxScaler)	1.0000	0.9998	0.9998	0.9999	0.9996
LASSO_L1 (MinMaxScaler)	1.0000	0.9998	0.9998	0.9999	0.9996
LASSO_L2 (MinMaxScaler)	1.0000	0.9998	0.9998	0.9999	0.9996
Filtered by correlation (MinMaxScaler)	1.0000	0.9998	0.9998	0.9999	0.9996
Fisher (MinMaxScaler)	1.0000	0.9998	0.9998	0.9999	0.9996
Selected by Mutual Information (StandardScaler)	1.0000	0.9999	0.9998	1.0000	0.9996
Selected by RFECV (StandardScaler)	1.0000	0.9998	0.9998	0.9999	0.9996
Selected by RF importance (StandardScaler)	1.0000	0.9999	0.9999	1.0000	0.9997
AUC (StandardScaler)	1.0000	0.9999	0.9999	1.0000	0.9997
Selected by RF importance (MinMaxScaler)	1.0000	0.9998	0.9998	0.9999	0.9996
AUC (MinMaxScaler)	1.0000	0.9999	0.9998	1.0000	0.9996
Selected by RFECV (MinMaxScaler)	1.0000	0.9999	0.9999	1.0000	0.9997
Fisher (StandardScaler)	1.0000	0.9998	0.9998	0.9999	0.9996
Kruskal (StandardScaler)	1.0000	0.9998	0.9998	0.9999	0.9996
LASSO_L1 (StandardScaler)	1.0000	0.9998	0.9998	0.9999	0.9996
LASSO_L2 (StandardScaler)	1.0000	0.9998	0.9998	0.9999	0.9996
Filtered by correlation (StandardScaler)	1.0000	0.9998	0.9998	0.9999	0.9996
All scaled features (StandardScaler)	1.0000	0.9998	0.9997	1.0000	0.9994
All scaled features (StandardScaler) + PCA	1.0000	0.9996	0.9995	0.9999	0.9990
mRMR (StandardScaler)	1.0000	0.9994	0.9993	0.9997	0.9986
mRMR (MinMaxScaler)	1.0000	0.9971	0.9967	0.9987	0.9944
All scaled features (MinMaxScaler) + PCA	1.0000	0.9986	0.9984	0.9995	0.9973
RFE_SVM (StandardScaler)	0.9994	0.9935	0.9925	0.9970	0.9880
RFE_SVM (MinMaxScaler)	0.9973	0.9829	0.9805	0.9911	0.9693

Table 5: Cross-Validation Results for SVM Models

9.1.2 Test Set Results

After testing, the ranking of the models shifted. The tables below present the performance of all models on the test set for each classifier, sorted by ROC-AUC, F1-score, and Accuracy.

Rank	Feature Set	Accuracy	Sensitivity	Specificity	F1-score	ROC-AUC
1	Selected by RFECV (MinMaxScaler)	0.9992	1.0000	0.9981	0.9993	0.5693
2	Selected by RF importance (StandardScaler)	0.7798	0.6151	0.9997	0.7616	0.9372
3	Selected by Mutual Information (StandardScaler)	0.6610	0.4618	0.9270	0.6091	0.8826
4	Filtered by correlation (StandardScaler)	0.6663	0.4703	0.9280	0.6171	0.8807
5	All scaled features (StandardScaler)	0.6663	0.4703	0.9280	0.6171	0.8807
6	Kruskal (MinMaxScaler)	0.9875	0.9815	0.9954	0.9890	0.4535
7	LASSO_L1 (MinMaxScaler)	0.9875	0.9815	0.9954	0.9890	0.4535
8	LASSO_L2 (MinMaxScaler)	0.9875	0.9815	0.9954	0.9890	0.4535
9	Filtered by correlation (MinMaxScaler)	0.9992	1.0000	0.9981	0.9993	0.5693
10	Selected by Mutual Information (MinMaxScaler)	0.9991	1.0000	0.9980	0.9992	0.5638
11	Selected by RF importance (MinMaxScaler)	0.9913	0.9849	0.9998	0.9924	0.5055
12	All scaled features (StandardScaler) + PCA	0.9751	0.9655	0.9879	0.9779	0.5993
13	Fisher (StandardScaler)	0.8096	0.6685	0.9980	0.8006	0.8346
14	AUC (StandardScaler)	0.8126	0.6738	0.9980	0.8044	0.8349
15	Fisher (MinMaxScaler)	0.9875	0.9815	0.9954	0.9890	0.4536
16	All scaled features (MinMaxScaler)	0.9992	1.0000	0.9981	0.9993	0.5693
17	AUC (MinMaxScaler)	0.9882	0.9812	0.9975	0.9896	0.4890
18	Selected by RFECV (StandardScaler)	0.6519	0.4481	0.9242	0.5955	0.8887
19	All scaled features (MinMaxScaler) + PCA	0.9932	0.9921	0.9948	0.9941	0.4562
20	mRMR (StandardScaler)	0.8096	0.6686	0.9980	0.8006	0.8345
21	All scaled features (StandardScaler) + PCA	0.9751	0.9655	0.9879	0.9779	0.5993
22	mRMR (MinMaxScaler)	0.9875	0.9815	0.9954	0.9890	0.4535
23	RFE_SVM (MinMaxScaler)	0.4281	0.0000	1.0000	0.0000	0.9764
24	RFE_SVM (StandardScaler)	0.4281	0.0000	1.0000	0.0000	0.9831

Table 6: Test Set Results for Euclidean MDC Models

Rank	Feature Set	Accuracy	Sensitivity	Specificity	F1-score	ROC-AUC
1	Selected by RF importance (MinMaxScaler)	0.9913	0.9849	0.9998	0.9924	1.0000
2	Filtered by correlation (MinMaxScaler)	0.9992	1.0000	0.9981	0.9993	0.9999
3	Selected by RFECV (MinMaxScaler)	0.9992	1.0000	0.9981	0.9993	0.9999
4	All scaled features (MinMaxScaler)	0.9992	1.0000	0.9981	0.9993	0.9999
5	Selected by Mutual Information (MinMaxScaler)	0.9991	1.0000	0.9980	0.9992	0.9999
6	All scaled features (MinMaxScaler) + PCA	0.9932	0.9921	0.9948	0.9941	0.9998
7	AUC (MinMaxScaler)	0.9882	0.9812	0.9975	0.9896	0.9997
8	Kruskal (MinMaxScaler)	0.9875	0.9815	0.9954	0.9890	0.9996
9	LASSO_L1 (MinMaxScaler)	0.9875	0.9815	0.9954	0.9890	0.9996
10	LASSO_L2 (MinMaxScaler)	0.9875	0.9815	0.9954	0.9890	0.9996
11	mRMR (MinMaxScaler)	0.9875	0.9815	0.9954	0.9890	0.9996
12	Fisher (MinMaxScaler)	0.9875	0.9815	0.9954	0.9890	0.9996
13	Selected by RF importance (StandardScaler)	0.7798	0.6151	0.9997	0.7616	0.9993
14	AUC (StandardScaler)	0.8126	0.6738	0.9980	0.8044	0.9983
15	Fisher (StandardScaler)	0.8096	0.6685	0.9980	0.8006	0.9982
16	Kruskal (StandardScaler)	0.8096	0.6686	0.9980	0.8006	0.9982
17	LASSO_L1 (StandardScaler)	0.8096	0.6686	0.9980	0.8006	0.9982
18	LASSO_L2 (StandardScaler)	0.8096	0.6686	0.9980	0.8006	0.9982
19	mRMR (StandardScaler)	0.8096	0.6686	0.9980	0.8006	0.9982
20	All scaled features (StandardScaler) + PCA	0.9751	0.9655	0.9879	0.9779	0.9949
21	RFE_SVM (MinMaxScaler)	0.4281	0.0000	1.0000	0.0000	0.9838
22	RFE_SVM (StandardScaler)	0.4281	0.0000	1.0000	0.0000	0.9838
23	Filtered by correlation (StandardScaler)	0.6663	0.4703	0.9280	0.6171	0.9172
24	All scaled features (StandardScaler)	0.6663	0.4703	0.9280	0.6171	0.9172
25	Selected by Mutual Information (StandardScaler)	0.6610	0.4618	0.9270	0.6091	0.9155
26	Selected by RFECV (StandardScaler)	0.6519	0.4481	0.9242	0.5955	0.9114

Table 7: Test Set Results for Bayes Linear Models

Rank	Feature Set	Accuracy	Sensitivity	Specificity	F1-score	ROC-AUC
1	Selected by RF importance (StandardScaler)	0.9955	0.9921	1.0000	0.9960	1.0000
2	AUC (StandardScaler)	0.9889	1.0000	0.9740	0.9904	0.9998
3	Selected by RF importance (MinMaxScaler)	0.9682	0.9450	0.9993	0.9715	0.9998
4	Selected by Mutual Information (MinMaxScaler)	0.9955	0.9985	0.9914	0.9961	0.9996
5	All scaled features (MinMaxScaler)	0.9948	0.9983	0.9901	0.9955	0.9994
6	Selected by RFECV (MinMaxScaler)	0.9948	0.9983	0.9901	0.9955	0.9994
7	Filtered by correlation (MinMaxScaler)	0.9948	0.9983	0.9901	0.9955	0.9994
8	Fisher (StandardScaler)	0.9870	0.9908	0.9820	0.9887	0.9992
9	mRMR (StandardScaler)	0.9870	0.9908	0.9820	0.9887	0.9992
10	Kruskal (StandardScaler)	0.9870	0.9908	0.9820	0.9887	0.9992
11	LASSO_L1 (StandardScaler)	0.9870	0.9908	0.9820	0.9887	0.9992
12	LASSO_L2 (StandardScaler)	0.9870	0.9908	0.9820	0.9887	0.9992
13	All scaled features (MinMaxScaler) + PCA	0.9903	0.9978	0.9802	0.9916	0.9991
14	Selected by RFECV (StandardScaler)	0.9639	0.9369	0.9999	0.9674	0.9990
15	Kruskal (MinMaxScaler)	0.9820	0.9813	0.9829	0.9842	0.9988
16	LASSO_L1 (MinMaxScaler)	0.9820	0.9813	0.9829	0.9842	0.9988
17	LASSO_L2 (MinMaxScaler)	0.9820	0.9813	0.9829	0.9842	0.9988
18	mRMR (MinMaxScaler)	0.9820	0.9813	0.9829	0.9842	0.9988
19	Fisher (MinMaxScaler)	0.9820	0.9813	0.9829	0.9842	0.9988
20	AUC (MinMaxScaler)	0.9822	0.9813	0.9833	0.9844	0.9987
21	Selected by Mutual Information (StandardScaler)	0.9728	0.9537	0.9984	0.9757	0.9986
22	All scaled features (StandardScaler)	0.9630	0.9406	0.9929	0.9667	0.9954
23	Filtered by correlation (StandardScaler)	0.9630	0.9406	0.9929	0.9667	0.9954
24	RFE_SVM (StandardScaler)	0.8583	0.7539	0.9979	0.8589	0.9835
25	RFE_SVM (MinMaxScaler)	0.4281	0.0000	1.0000	0.0000	0.9836
26	All scaled features (StandardScaler) + PCA	0.7987	0.8926	0.6733	0.8353	0.8487

Table 8: Test Set Results for Bayes Quadratic Models

Rank	Feature Set	Accuracy	Sensitivity	Specificity	F1-score	ROC-AUC
1	Selected by RF importance (MinMaxScaler)	0.9992	1.0000	0.9982	0.9993	0.9999
2	Selected by RF importance (StandardScaler)	0.9995	1.0000	0.9989	0.9996	0.9999
3	Selected by Mutual Information (StandardScaler)	0.9983	0.9992	0.9971	0.9985	0.9998
4	All scaled features (StandardScaler)	0.9982	0.9993	0.9968	0.9984	0.9998
5	Kruskal (StandardScaler)	0.9979	0.9990	0.9965	0.9982	0.9998
6	LASSO_L1 (StandardScaler)	0.9979	0.9990	0.9965	0.9982	0.9998
7	LASSO_L2 (StandardScaler)	0.9979	0.9990	0.9965	0.9982	0.9998
8	Filtered by correlation (StandardScaler)	0.9979	0.9990	0.9965	0.9982	0.9998
9	mRMR (MinMaxScaler)	0.9984	0.9989	0.9977	0.9986	0.9997
10	mRMR (StandardScaler)	0.9986	0.9988	0.9983	0.9988	0.9997
11	Selected by RFECV (StandardScaler)	0.9980	0.9989	0.9969	0.9983	0.9997
12	All scaled features (StandardScaler) + PCA	0.9974	0.9990	0.9954	0.9978	0.9997
13	Fisher (StandardScaler)	0.9976	0.9986	0.9963	0.9979	0.9996
14	AUC (StandardScaler)	0.9975	0.9982	0.9966	0.9978	0.9998
15	Selected by Mutual Information (MinMaxScaler)	0.9938	0.9957	0.9913	0.9946	0.9994
16	Fisher (MinMaxScaler)	0.9942	0.9960	0.9918	0.9949	0.9993
17	Kruskal (MinMaxScaler)	0.9941	0.9959	0.9917	0.9948	0.9993
18	LASSO_L1 (MinMaxScaler)	0.9941	0.9959	0.9917	0.9948	0.9993
19	LASSO_L2 (MinMaxScaler)	0.9941	0.9959	0.9917	0.9948	0.9993
20	Filtered by correlation (MinMaxScaler)	0.9941	0.9959	0.9917	0.9948	0.9993
21	Selected by RFECV (MinMaxScaler)	0.9939	0.9954	0.9918	0.9946	0.9993
22	All scaled features (MinMaxScaler)	0.9940	0.9961	0.9911	0.9947	0.9993
23	AUC (MinMaxScaler)	0.9938	0.9952	0.9918	0.9946	0.9992
24	All scaled features (MinMaxScaler) + PCA	0.9930	0.9950	0.9903	0.9939	0.9992
25	RFE_SVM (StandardScaler)	0.9900	0.9926	0.9866	0.9913	0.9984
26	RFE_SVM (MinMaxScaler)	0.9849	0.9890	0.9795	0.9868	0.9976

Table 9: Test Set Results for k-NN (k=25) Models

Rank	Feature Set	Accuracy	Sensitivity	Specificity	F1-score	ROC-AUC
1	All scaled features (MinMaxScaler)	0.9998	1.0000	0.9995	0.9998	1.0000
2	Filtered by correlation (MinMaxScaler)	0.9998	1.0000	0.9995	0.9998	1.0000
3	Selected by RFECV (MinMaxScaler)	0.9998	1.0000	0.9995	0.9998	1.0000
4	Selected by RF importance (MinMaxScaler)	0.9989	0.9982	0.9998	0.9990	1.0000
5	Selected by Mutual Information (MinMaxScaler)	0.9998	1.0000	0.9995	0.9998	1.0000
6	AUC (MinMaxScaler)	0.9963	1.0000	0.9914	0.9968	1.0000
7	All scaled features (MinMaxScaler) + PCA	0.9984	0.9990	0.9975	0.9986	1.0000
8	Kruskal (MinMaxScaler)	0.9964	1.0000	0.9916	0.9969	1.0000
9	LASSO_L1 (MinMaxScaler)	0.9964	1.0000	0.9916	0.9969	1.0000
10	LASSO_L2 (MinMaxScaler)	0.9964	1.0000	0.9916	0.9969	1.0000
11	mRMR (MinMaxScaler)	0.9964	1.0000	0.9916	0.9969	1.0000
12	Fisher (MinMaxScaler)	0.9964	1.0000	0.9916	0.9969	1.0000
13	All scaled features (StandardScaler) + PCA	0.6133	0.9771	0.1273	0.7430	0.9118
14	Selected by RF importance (StandardScaler)	0.4884	0.1120	0.9913	0.2003	0.8468
15	Filtered by correlation (StandardScaler)	0.5239	0.5459	0.4945	0.5674	0.5654
16	All scaled features (StandardScaler)	0.5239	0.5459	0.4945	0.5674	0.5654
17	Selected by RFECV (StandardScaler)	0.4980	0.3781	0.6583	0.4628	0.5628
18	Selected by Mutual Information (StandardScaler)	0.5140	0.4655	0.5788	0.5228	0.5466
19	Kruskal (StandardScaler)	0.5713	0.7567	0.3237	0.6688	0.5000
20	LASSO_L1 (StandardScaler)	0.5713	0.7567	0.3237	0.6688	0.5000
21	LASSO_L2 (StandardScaler)	0.5713	0.7567	0.3237	0.6688	0.5000
22	mRMR (StandardScaler)	0.5713	0.7567	0.3237	0.6688	0.5000
23	Fisher (StandardScaler)	0.5704	0.7548	0.3240	0.6677	0.5000
24	AUC (StandardScaler)	0.5151	0.5682	0.4443	0.5727	0.5000
25	RFE_SVM (StandardScaler)	0.4281	0.0000	1.0000	0.0000	0.5000
26	RFE_SVM (MinMaxScaler)	0.4281	0.0000	1.0000	0.0000	0.5000

Table 10: Test Set Results for SVM (rbf) Models

Due to the large number of evaluation plots generated, only the best model for each classifier is shown below. For a more detailed interpretation, additional visualizations and analysis are included in the Appendix..

9.1.3 Average Number of Selected Features per Method

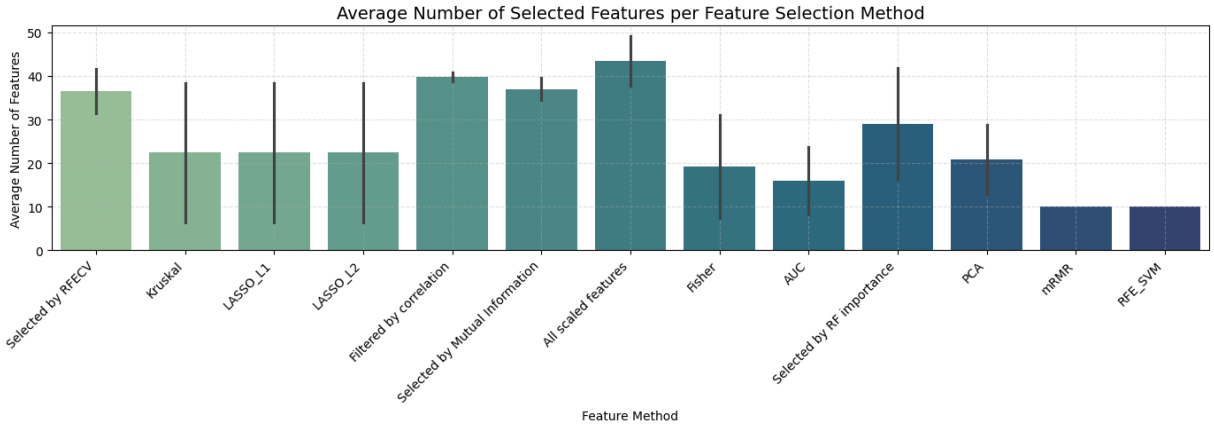

Figure 2: Average number of features selected by each method with standard deviation bars.

9.1.4 Mean Evaluation Metrics per Classifier

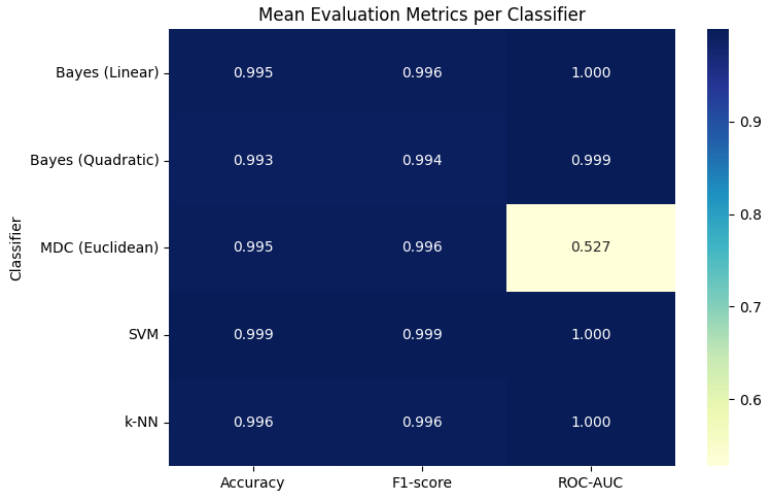

Figure 3: Mean Accuracy, F1-score, and ROC-AUC per classifier across all feature sets.

9.2 Discussion

With the inclusion of four additional classifiers (k-NN, SVM, Bayes Linear and Bayes Quadratic), the test set evaluation provided a broader perspective on model performance. Although cross-validation rankings were based on ROC-AUC, F1-score, and Accuracy, several models that scored highly during validation did not retain their position on the test set. This confirms that certain models may have overfitted to the training data and did not generalize well.

Among the classifiers, SVM and k-NN consistently achieved top test performance across all metrics. In contrast, MDC (Euclidean) had strong Accuracy and F1-score, but very low ROC-AUC, which limits its suitability for imbalanced detection. Bayes models also showed competitive performance, with the linear version reaching perfect ROC-AUC in many setups. These results highlight the added value of comparing classifiers with different decision boundaries and sensitivities to feature scaling.

In terms of preprocessing, StandardScaler generally led to better results than MinMaxScaler, especially on the test set. This suggests that centering and normalizing feature distributions is more beneficial in this dataset than simple range compression. Feature selection methods based on filters (e.g., Fisher, Kruskal, Mutual Information) and embedded approaches (e.g., LASSO, RF importance) performed similarly. However, PCA-based reductions consistently showed a drop in ROC-AUC, confirming that while PCA can preserve variance, it may discard information relevant for class discrimination in this context.

Overall, the best models on the test set were not necessarily those with the highest cross-validation scores. This underlines the importance of combining validation metrics with final test evaluation when selecting models. Test performance must always guide the final choice, especially in scenarios prone to overfitting like phishing detection with high-dimensional data.

10 Conclusion

Adding k-NN, SVM, Bayes Linear, and Bayes Quadratic to the initial MDC baseline enabled a broader comparison of classifiers for phishing URL detection. SVM and k-NN achieved the best ROC-AUC scores on the test set, consistently outperforming MDC variants. Feature engineering had a clear impact: StandardScaler generally led to better results than MinMaxScaler, while PCA-based reductions often degraded performance. RFECV and RF importance remained strong feature selection strategies across models, but cross-validation alone was not always predictive of test performance.

Several classifiers showed signs of overfitting, performing well in training but poorly on the test set, reinforcing the need for careful test evaluation. Future work could focus on combining classifiers, analyzing decision boundaries, or testing robustness in more realistic phishing scenarios.

References

[1] Prasad, A., & Chandra, S. (2023). *PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning.* Computers & Security, 136, 103545.

[2] GeeksforGeeks. *Data Pre-processing with Scikit-Learn using Standard and MinMax Scaler.* Available at: https://www.geeksforgeeks.org/data-pre-processing-wit-sklearn-using-standard-and-minmax-scaler/

[3] TutorialsPoint. *Machine Learning - High Correlation Filter.* Available at: https://www.tutorialspoint.com/machine_learning/machine_learning_high_correlation_filter.htm

[4] GeeksforGeeks. *Feature Importance with Random Forests.* Available at: https://www.geeksforgeeks.org/feature-importance-with-random-forests/

[5] Mairam Nair. *Feature Selection: Mutual Information.* Available at: https://medium.com/@miramnair/feature-selection-mutual-information-a0def943e1ed

[6] GeeksforGeeks. *Recursive Feature Elimination with Cross-Validation in Scikit-Learn.* Available at: https://www.geeksforgeeks.org/recursive-feature-elimination-with-cross-validation-in-scikit-learn/

[7] Analytics Vidhya. *Recursive Feature Elimination.* Available at: https://www.analyticsvidhya.com/blog/2023/05/recursive-feature-elimination/

[8] GeeksforGeeks. *Principal Component Analysis with Python.* Available at: https://www.geeksforgeeks.org/principal-component-analysis-with-python/

[9] GeeksforGeeks. *ML — Linear Discriminant Analysis.* Available at: https://www.geeksforgeeks.org/ml-linear-discriminant-analysis/

[10] DataCamp. *k-Nearest Neighbors Classification Using Scikit-Learn.* Available at: https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn

[11] Scikit-Learn. *Support Vector Machines*. Available at:
https://scikit-learn.org/stable/modules/svm.html

[12] DataSklr. *Linear and Quadratic Discriminant Analysis*. Available at:
https://www.datasklr.com/select-classification-methods/linear-and-quadratic-discriminant-analysis

[13] GeeksforGeeks. *How to Perform a Kruskal-Wallis Test in Python*. Available at:
https://www.geeksforgeeks.org/how-to-perform-a-kruskal-wallis-test-in-python/

[14] Ranasingh. *Implementing Feature Selection Methods for Machine Learning*. Available at:
https://ranasinghiitkgp.medium.com/implementing-feature-selection-methods-for-machine-learning-bfa2e4b4e02

[15] Juan C. Olamendy. *A Comprehensive Guide to Stratified K-Fold Cross Validation for Unbalanced Data*. Available at:
https://medium.com/@juanc.olamendy/a-comprehensive-guide-to-stratified-k-fold-cross-validation-for-unbalanced-data-014691060f17

A Appendix

Correlation Analysis of Features

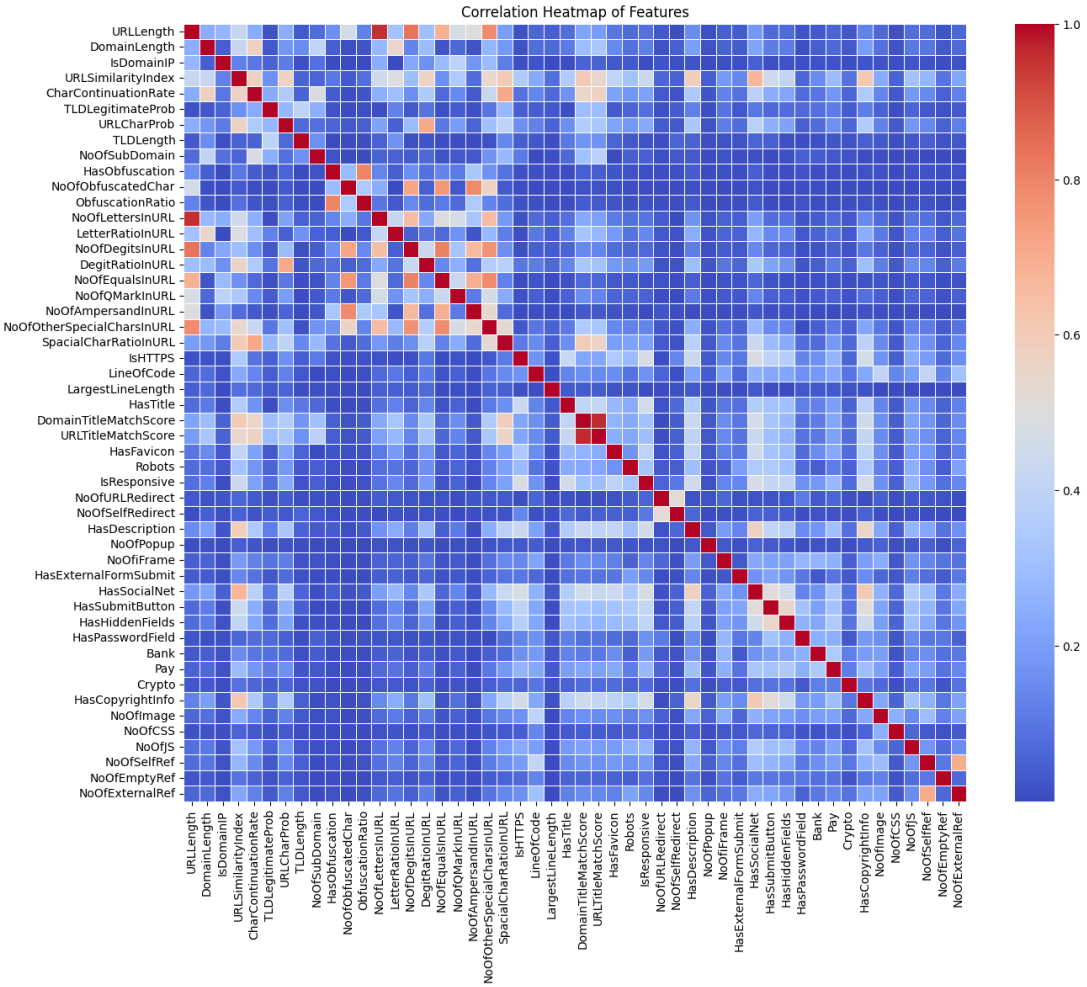

Figure 4: Correlation heatmap of the original features in the dataset

Evaluation on Test Set — MDC (Euclidean)

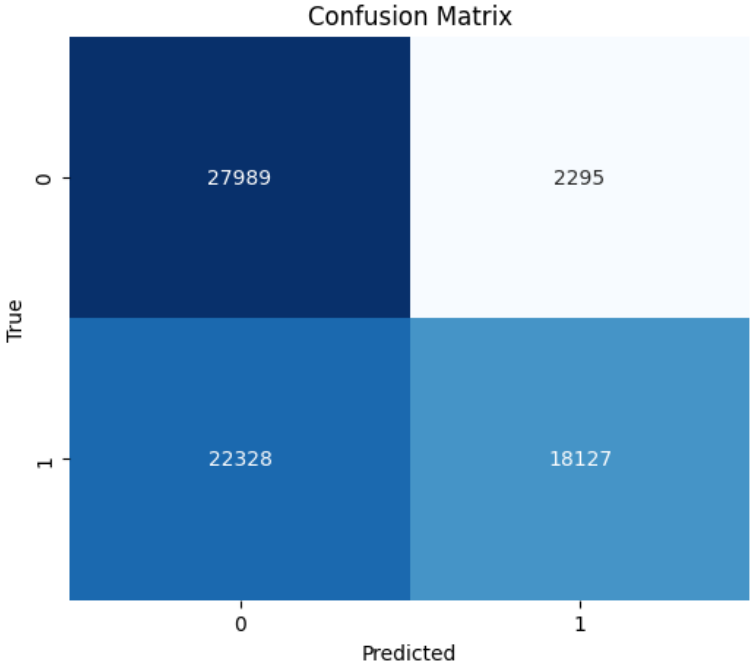

Figure 5: Confusion Matrix of the best MDC model on the test set

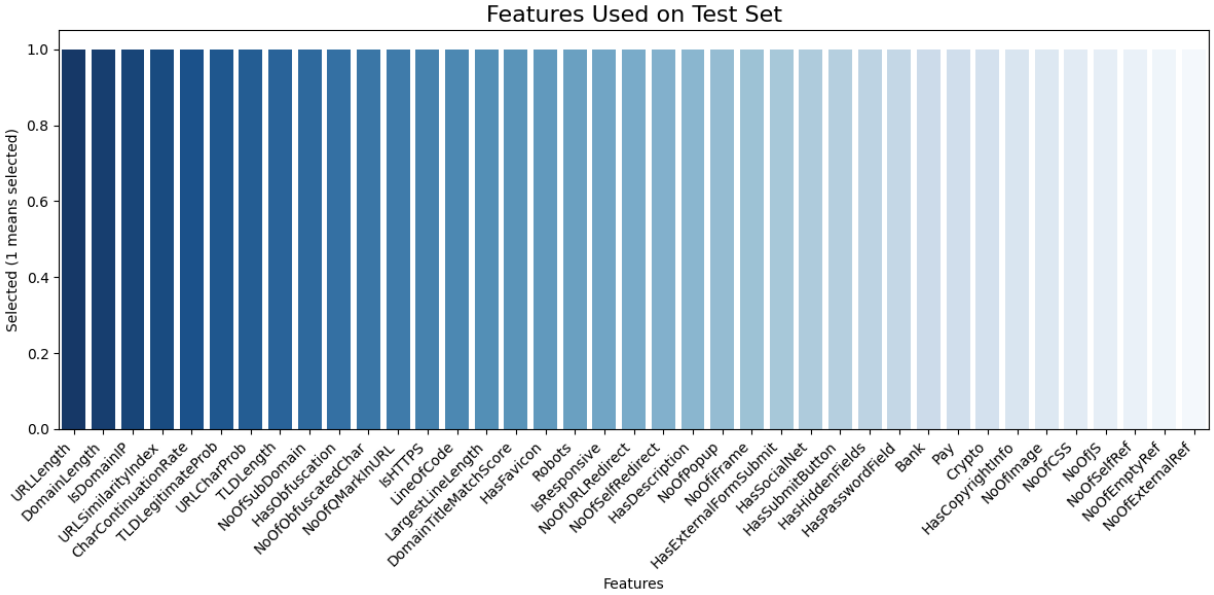

Figure 6: Features selected of the best MDC model

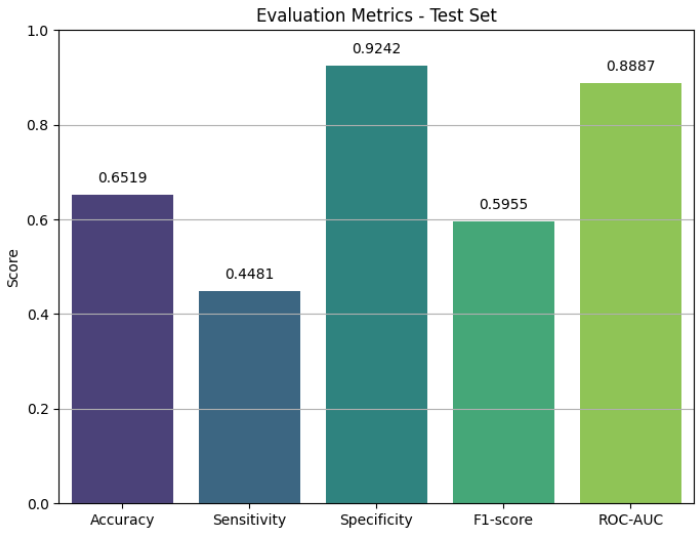

Figure 7: Evaluation metrics of the best MDC model on the test set

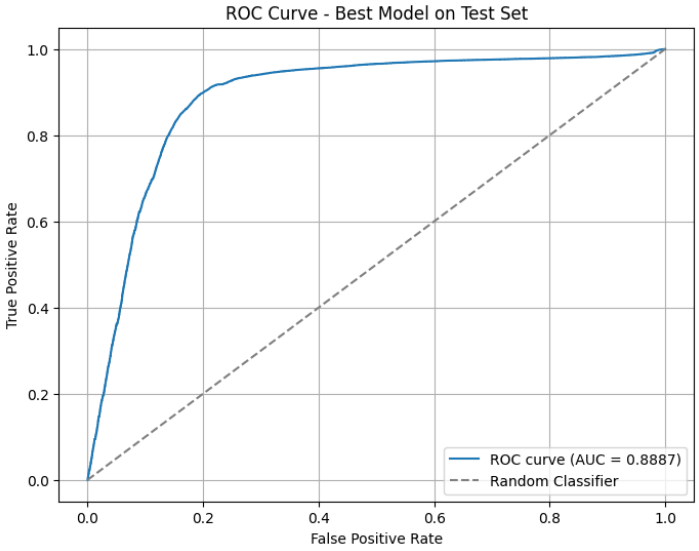

Figure 8: ROC curve of the best MDC model on the test set

Evaluation on Test Set — Bayes (Linear)

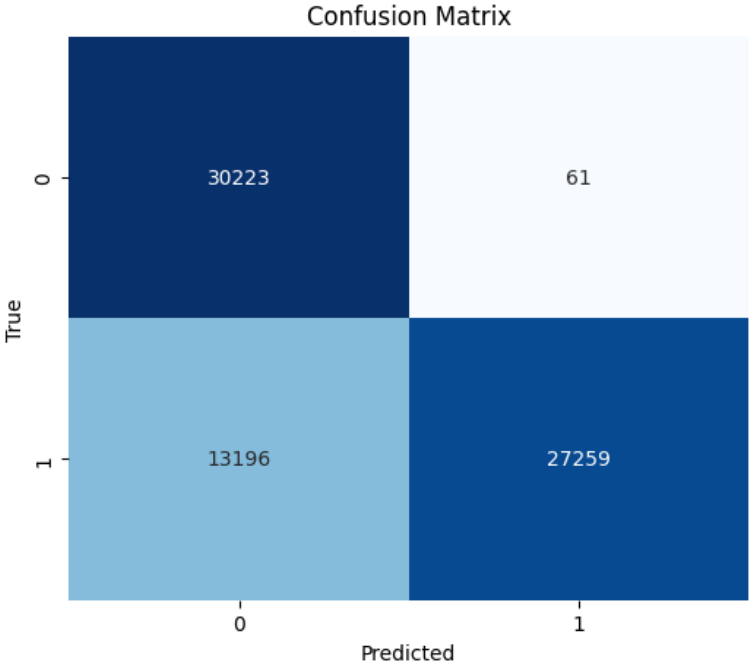

Figure 9: Confusion Matrix of the best Bayes Linear model on the test set

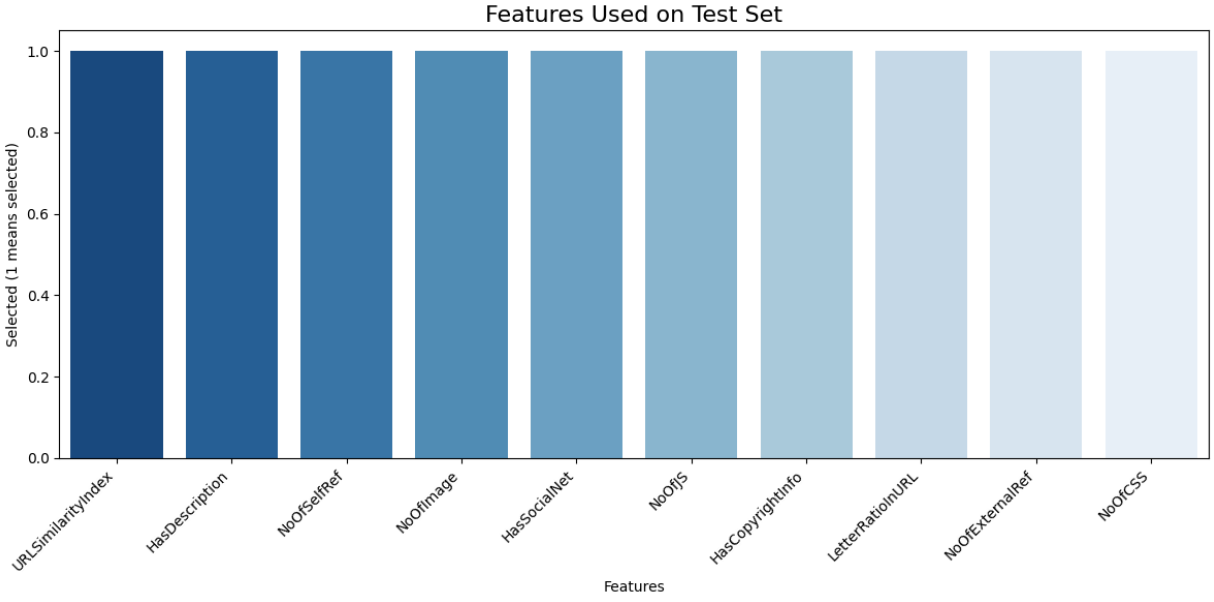

Figure 10: Features selected of the best Bayes Linear model

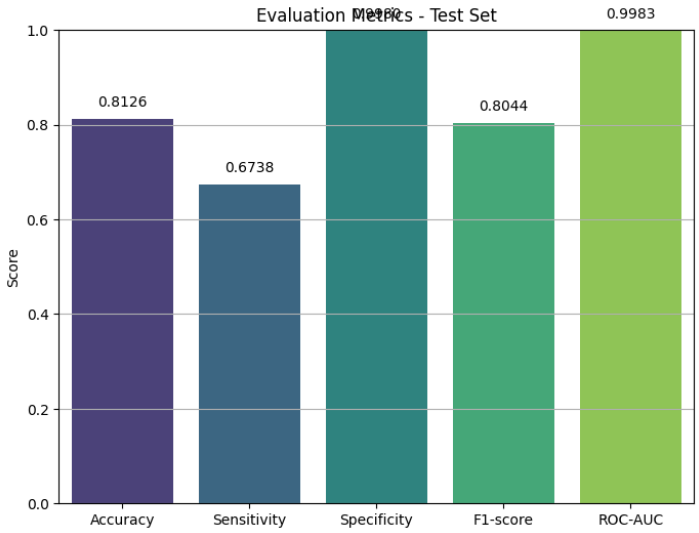

Figure 11: Evaluation metrics of the best Bayes Linear model on the test set

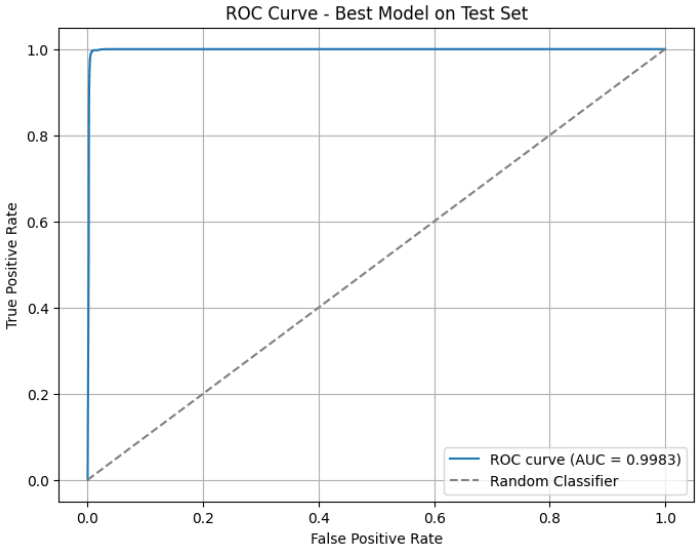

Figure 12: ROC curve of the best Bayes Linear model on the test set

Evaluation on Test Set — Bayes (Quadratic)

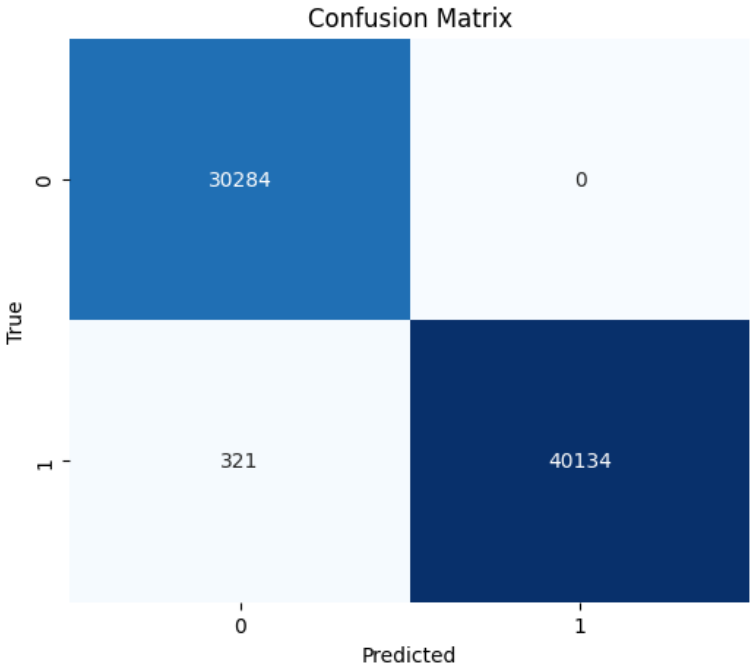

Figure 13: Confusion Matrix of the best Bayes Quadratic model on the test set

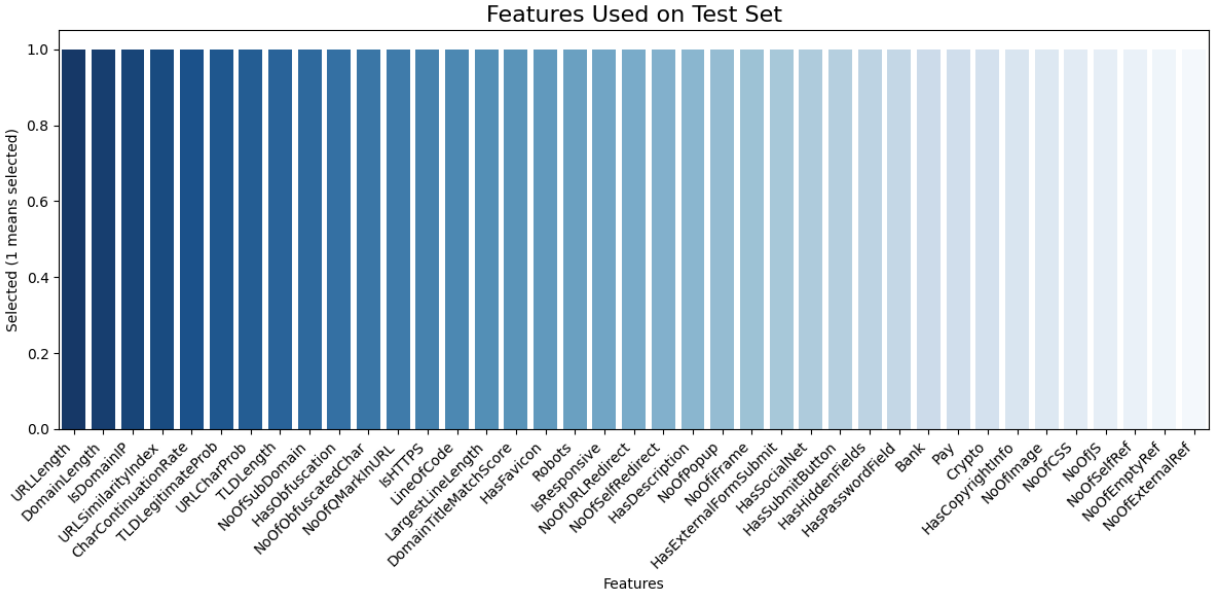

Figure 14: Features selected of the best Bayes Quadratic model

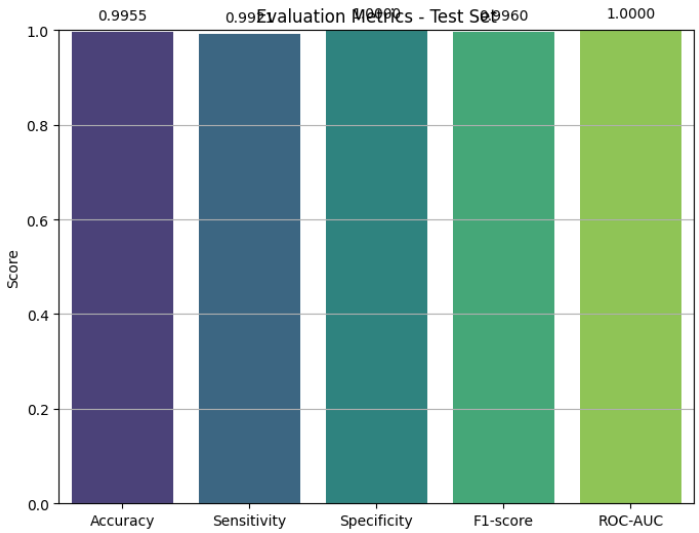

Figure 15: Evaluation metrics of the best Bayes Quadratic model on the test set

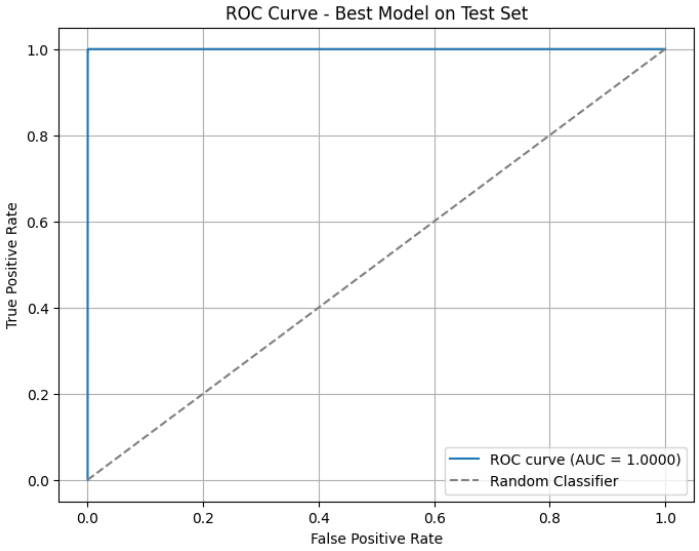

Figure 16: ROC curve of the best Bayes Quadratic model on the test set

Evaluation on Test Set — k-NN

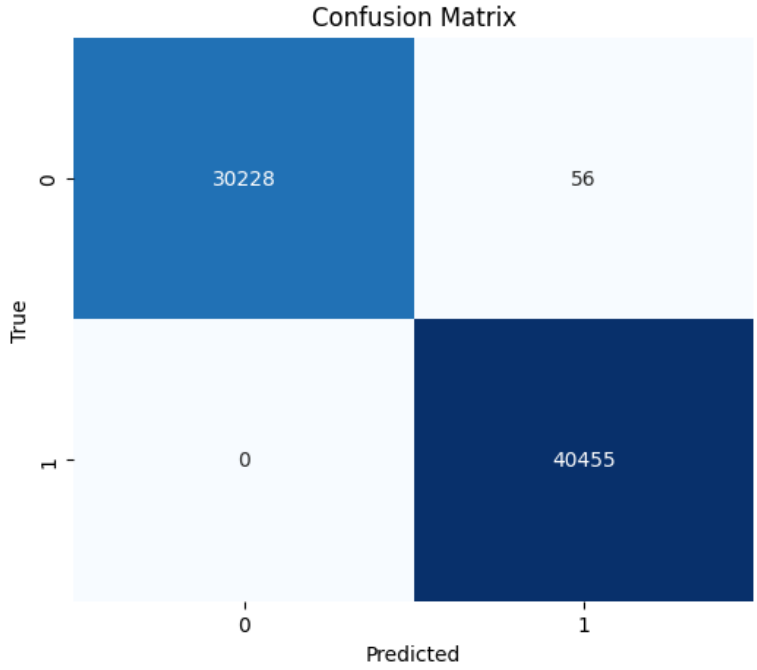

Figure 17: Confusion Matrix of the best k-NN model on the test set

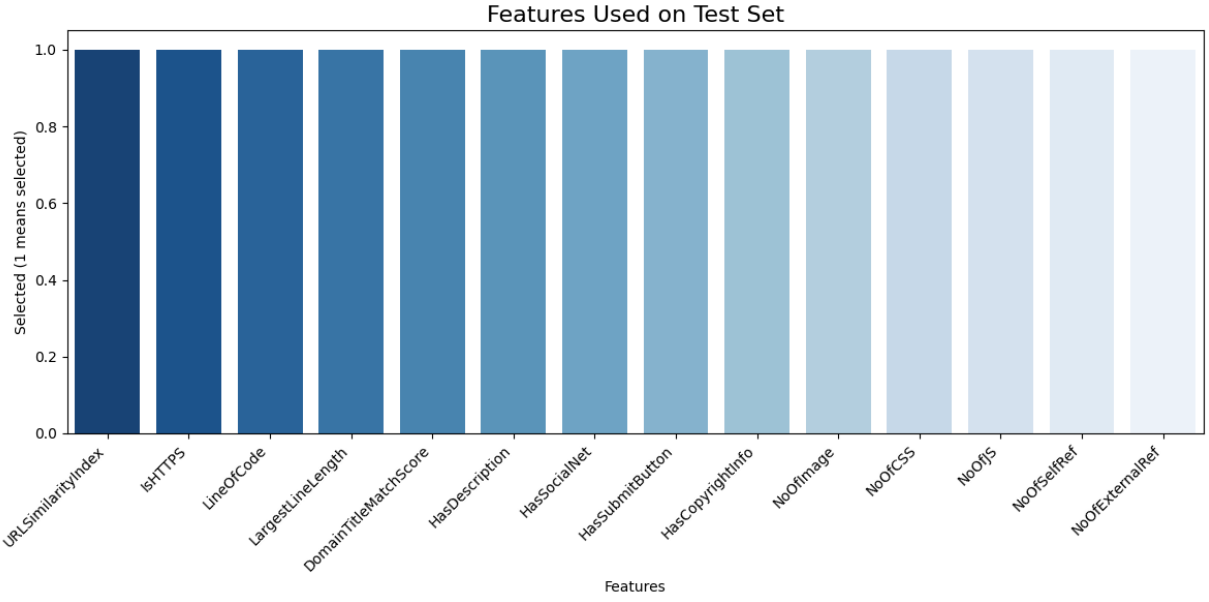

Figure 18: Features selected of the best k-NN model

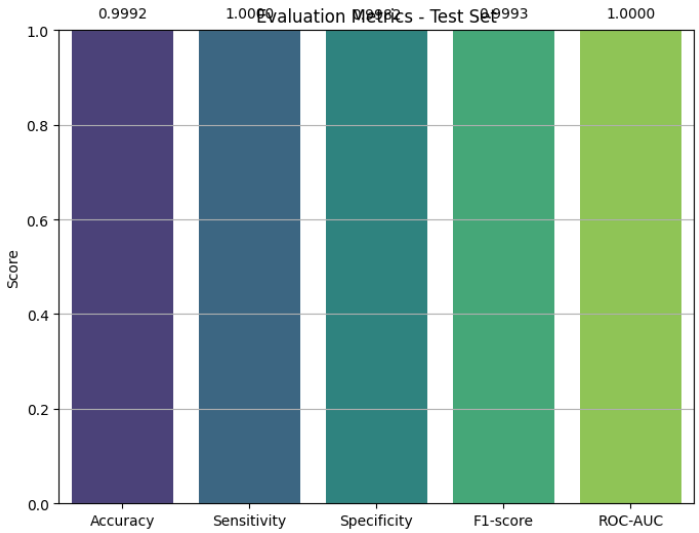

Figure 19: Evaluation metrics of the best k-NN model on the test set

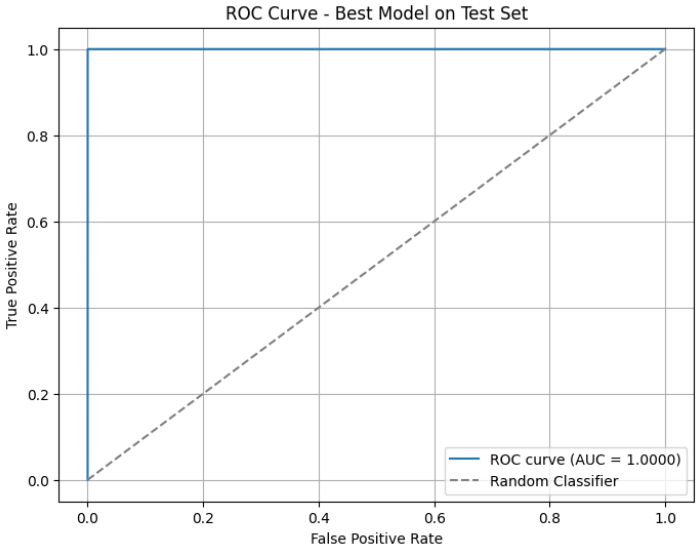

Figure 20: ROC curve of the best k-NN model on the test set

Evaluation on Test Set — SVM (RBF)

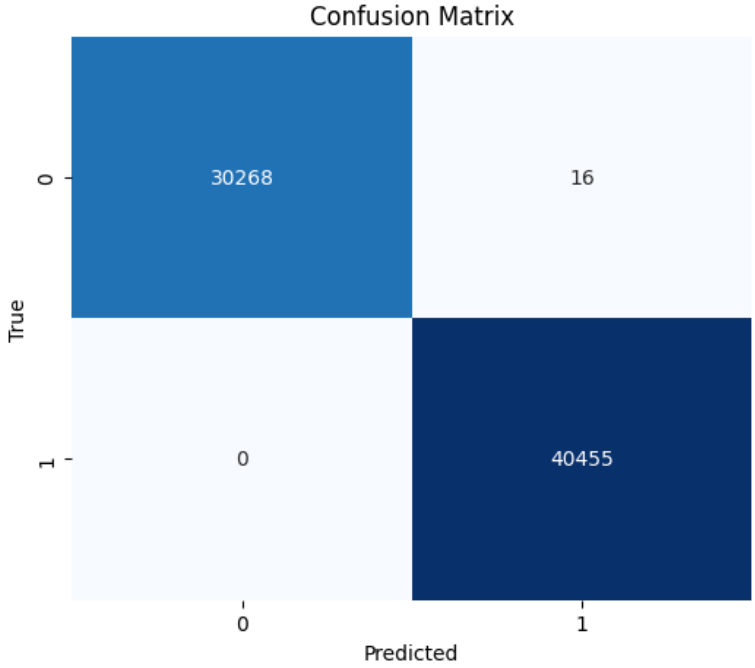

Figure 21: Confusion Matrix of the best SVM model on the test set

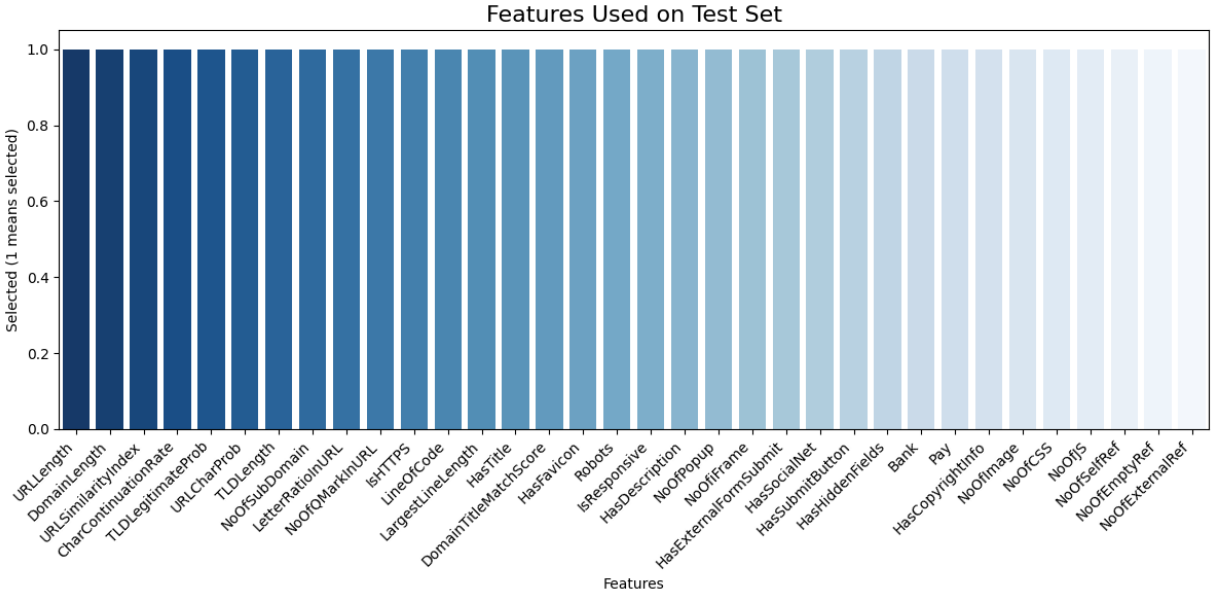

Figure 22: Features selected of the best SVM model

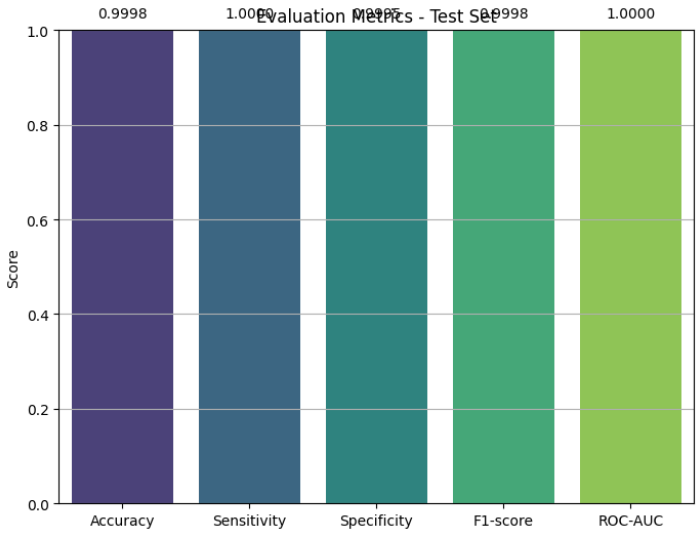

Figure 23: Evaluation metrics of the best SVM model on the test set

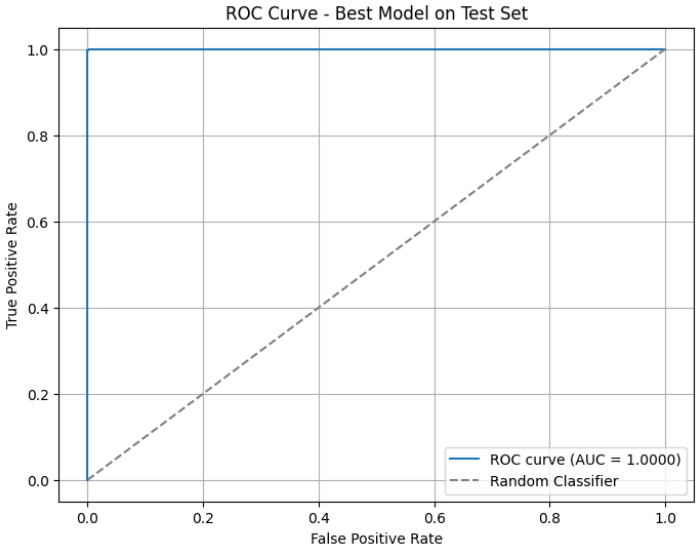

Figure 24: ROC curve of the best SVM model on the test set