CS 513

"KNOWLEDGE DISCOVERY AND DATA MINING"

COMPARATIVE ANALYSIS OF CLASSIFICATION MODELS ON PHISHING WEBSITE DATASET

03 December, 2024

OUR TEAM:

AMRUTHA KANAKATTE RAVISHANKAR

CWID: 20027346

OJAS YOGENDRA VAZE

CWID: 20034747

HARSHKUMAR DIPESHKUMAR BHIKADIYA

CWID: 20033108

BHUMITI VIPULBHAI GOHEL

CWID: 20025320

AGENDA OVERVIEW

01

INTRODUCTION

02

PROBLEM STATEMENT

03

EXPLORATORY DATA ANALYSIS

04

CLASSIFICATION ALGORITHMS

05

MODEL CONCLUSION

INTRODUCTION

Cybercrimes, of which phishing assaults are one of the most common risks, have increased in tandem with the growing use of the internet for a variety of purposes. Phishing websites are malicious websites that pose as trustworthy websites in order to get private information, including credit card numbers, login passwords, and other private information. To protect consumers from such attacks, it is essential to identify these phishing websites.

In this study, we use a publically available dataset to investigate a machine learning method for phishing website detection. Our goal is to create an effective model that can differentiate between phishing and legal websites by utilizing a variety of classification methods.

DATASET DESCRIPTION

https://www.kaggle.com/datasets/eswarchandt/phishing-website-detector

Index	UsingIP	LongURL	ShortURL	Symbol@	Redirecting//	PrefixSuffix-	SubDomains	HTTPS	DomainRegLen	...	UsingPopupWindow	IframeRedirection	AgeofDomain	DNSRecording	WebsiteTraffic	PageRank	GoogleIndex	LinksPointingToPage	StatsReport	class
0	1	1	1	1	1	-1	0	1	-1	...	1	1	-1	-1	0	-1	1	1	1	-1
1	1	0	1	1	1	-1	-1	-1	-1	...	1	1	1	-1	1	-1	1	0	-1	-1
2	1	0	1	1	1	-1	-1	-1	1	...	1	1	-1	-1	1	-1	1	-1	1	-1
3	1	0	-1	1	1	-1	1	1	-1	...	-1	1	-1	-1	0	-1	1	1	1	1
4	-1	0	-1	1	-1	-1	1	1	-1	...	1	1	1	1	1	-1	1	-1	-1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
11049	1	-1	1	-1	1	1	1	1	-1	...	-1	-1	1	1	-1	-1	1	1	1	1
11050	-1	1	1	-1	-1	-1	1	-1	-1	...	-1	1	1	1	1	1	1	-1	1	-1
11051	1	-1	1	1	1	-1	1	-1	-1	...	1	1	1	1	1	-1	1	0	1	-1
11052	-1	-1	1	1	1	-1	-1	-1	1	...	-1	1	1	1	1	-1	1	1	1	-1
11053	-1	-1	1	1	1	-1	-1	-1	1	...	1	1	-1	1	-1	-1	-1	1	-1	-1

```
Dataset Overview
Total samples: 11054
Total features (excluding target): 31
Target column: 'class'
```

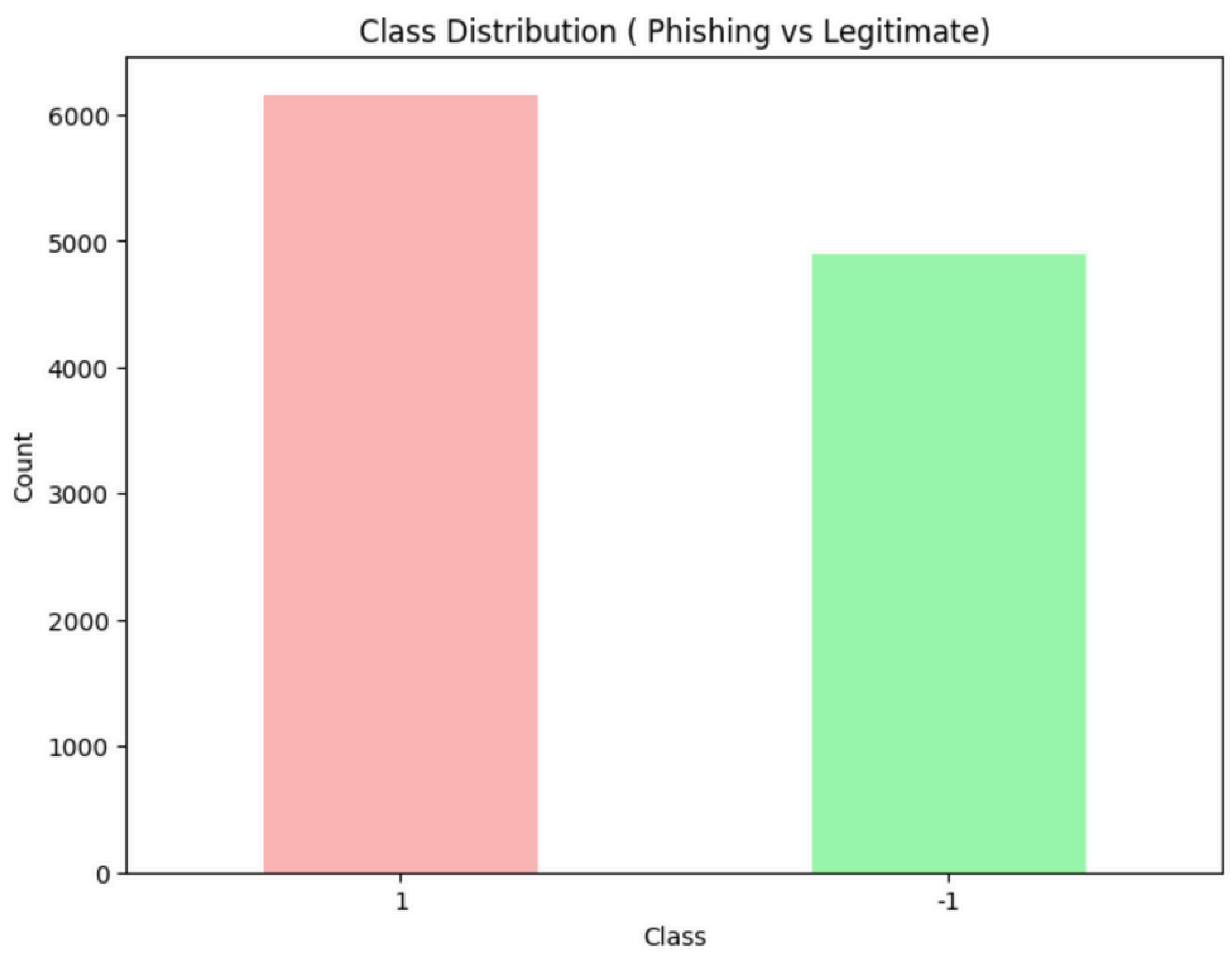
PROBLEM STATEMENT

Phishing attacks pose a serious risk to internet security because they can cause identity theft, financial loss, and other negative outcomes for both people and businesses. Phishing websites are constantly changing, making detection difficult even with improvements in security measures. This project's objective is to use machine learning techniques to create a trustworthy system for detecting phishing websites. Our goal is to accurately classify websites as either authentic or phishing by utilizing features that are retrieved from the website's domain, URL, and other aspects. The goal of this research is to aid in the creation of more potent online security technologies.

EXPLORATORY DATA ANALYSIS

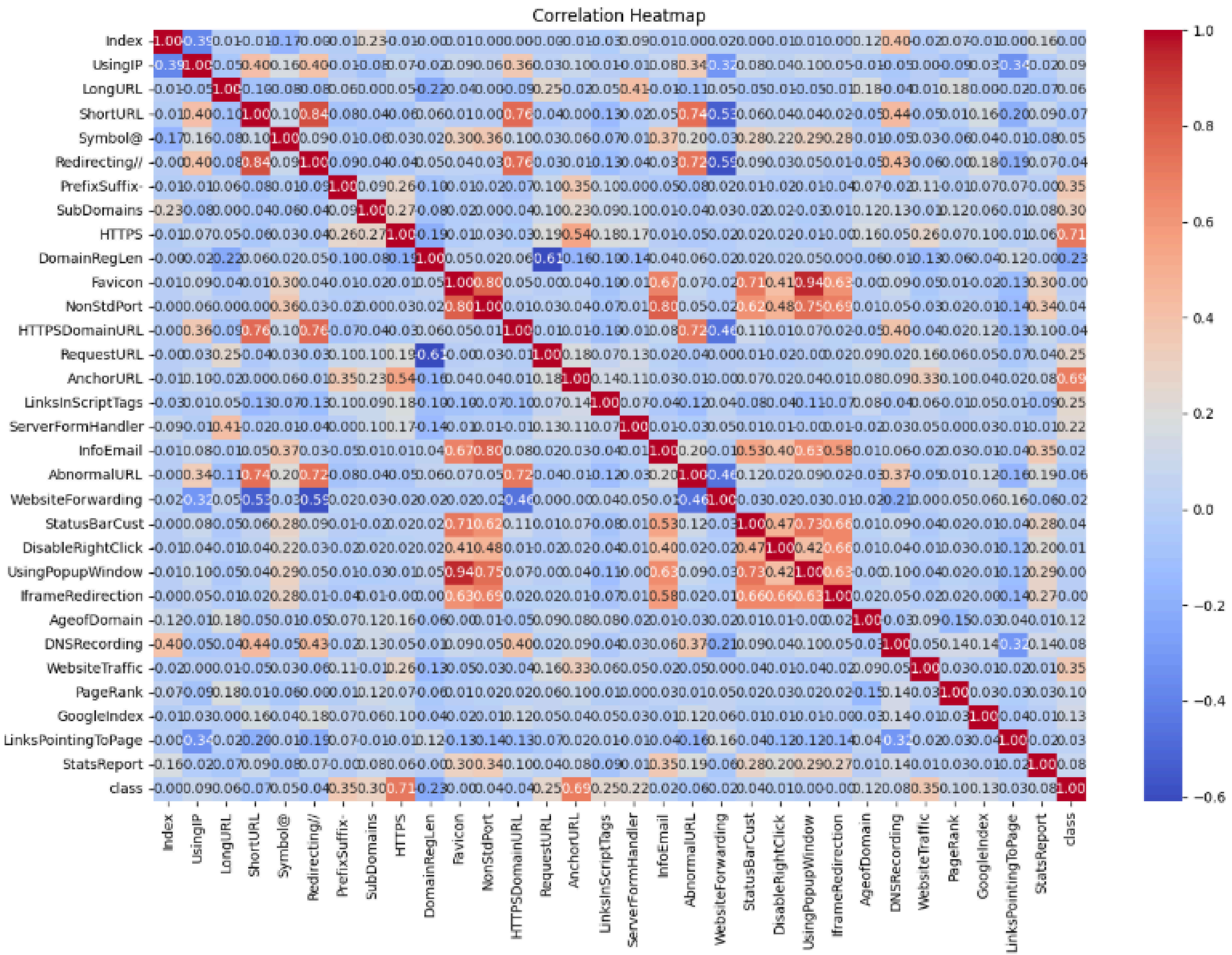
```
Missing values per column
Index                 0
UsingIP               0
LongURL               0
ShortURL              0
Symbol@               0
Redirecting//         0
PrefixSuffix-         0
SubDomains            0
HTTPS                 0
DomainRegLen          0
Favicon               0
NonStdPort            0
HTTPSDomainURL        0
RequestURL            0
AnchorURL             0
LinksInScriptTags     0
ServerFormHandler     0
InfoEmail             0
AbnormalURL           0
WebsiteForwarding     0
StatusBarCust         0
DisableRightClick     0
UsingPopupWindow      0
IframeRedirection     0
AgeofDomain           0
DNSRecording          0
WebsiteTraffic        0
PageRank              0
GoogleIndex           0
LinksPointingToPage   0
StatsReport           0
class                 0
dtype: int64
```

Inference: There are no missing values find in the dataset



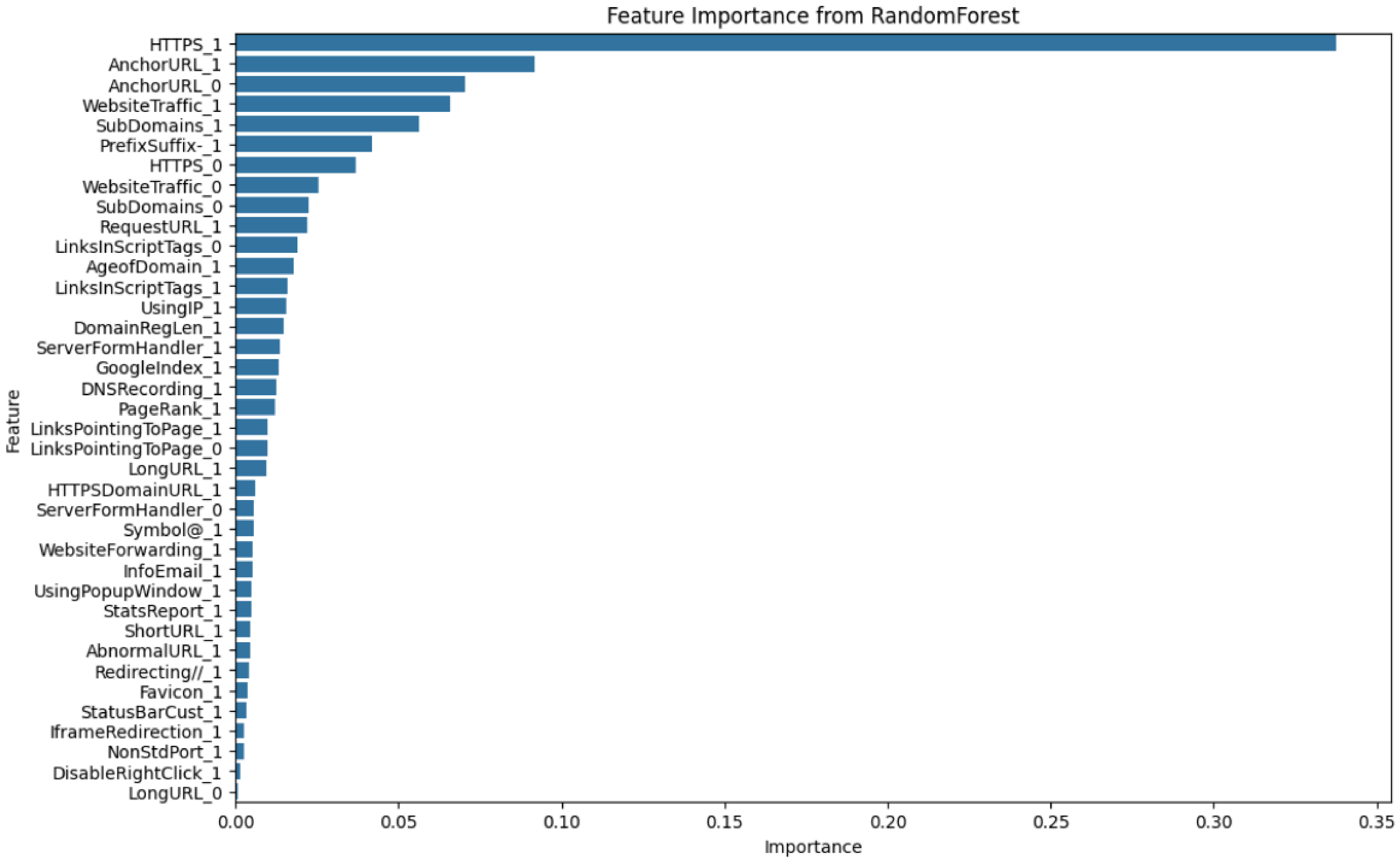
Inference: There are 6157 Phishing and 4897 Legitimate

EXPLORATORY DATA ANALYSIS



Inference: This shows the relationship between features

DATA PREPROCESSING TECHNIQUES



Inference: Guides decisions on which features to focus on for improving the model.

DATA PREPROCESSING TECHNIQUES

2. One Hot Encoding

```
Shape after One-Hot Encoding (11054, 38)
Index(['UsingIP_1', 'LongURL_0', 'LongURL_1', 'ShortURL_1', 'Symbol@_1',
       'Redirecting//_1', 'PrefixSuffix-_1', 'SubDomains_0', 'SubDomains_1',
       'HTTPS_0', 'HTTPS_1', 'DomainRegLen_1', 'Favicon_1', 'NonStdPort_1',
       'HTTPSDomainURL_1', 'RequestURL_1', 'AnchorURL_0', 'AnchorURL_1',
       'LinksInScriptTags_0', 'LinksInScriptTags_1', 'ServerFormHandler_0',
       'ServerFormHandler_1', 'InfoEmail_1', 'AbnormalURL_1',
       'WebsiteForwarding_1', 'StatusBarCust_1', 'DisableRightClick_1',
       'UsingPopupWindow_1', 'IframeRedirection_1', 'AgeofDomain_1',
       'DNSRecording_1', 'WebsiteTraffic_0', 'WebsiteTraffic_1', 'PageRank_1',
       'GoogleIndex_1', 'LinksPointingToPage_0', 'LinksPointingToPage_1',
       'StatsReport_1'],
      dtype='object')
```

Inference: Change in the size of data after One-Hot Encoding

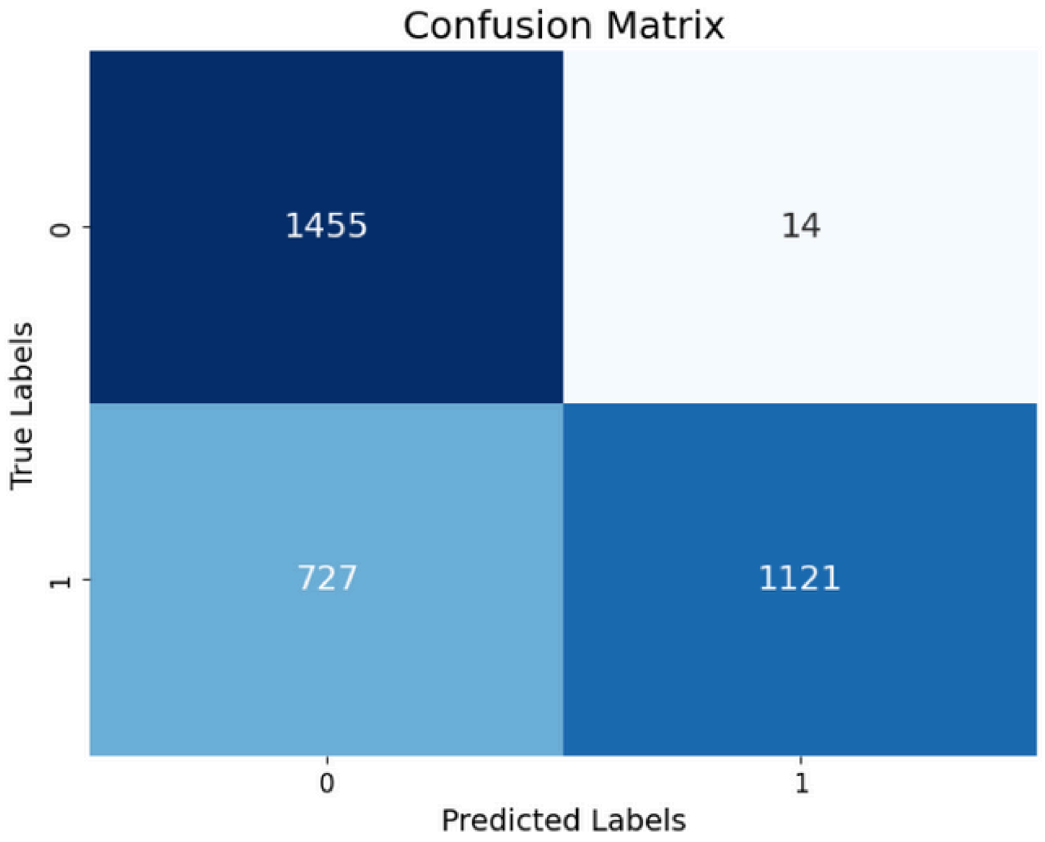
3. Oversampling using SMOTE

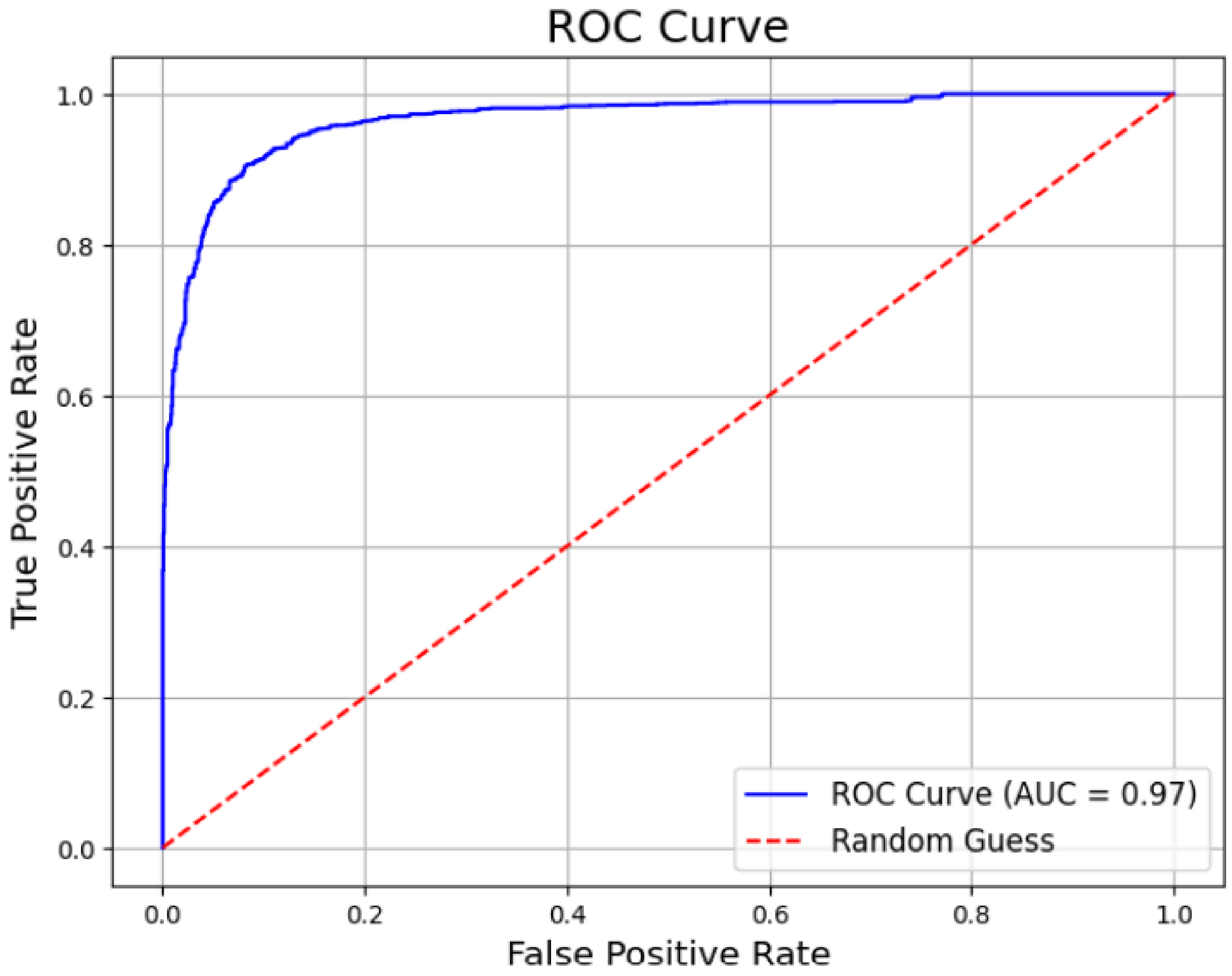
```
Training set shape before SMOTE: (7737, 38)
Testing set shape (unchanged): (3317, 38)
Training set shape after SMOTE: (8618, 38)
Testing set shape (unchanged): (3317, 38)
```

Inference: Oversampling balances class distribution by generating synthetic data for minority.

NAIVE BAYES CLASSIFIER

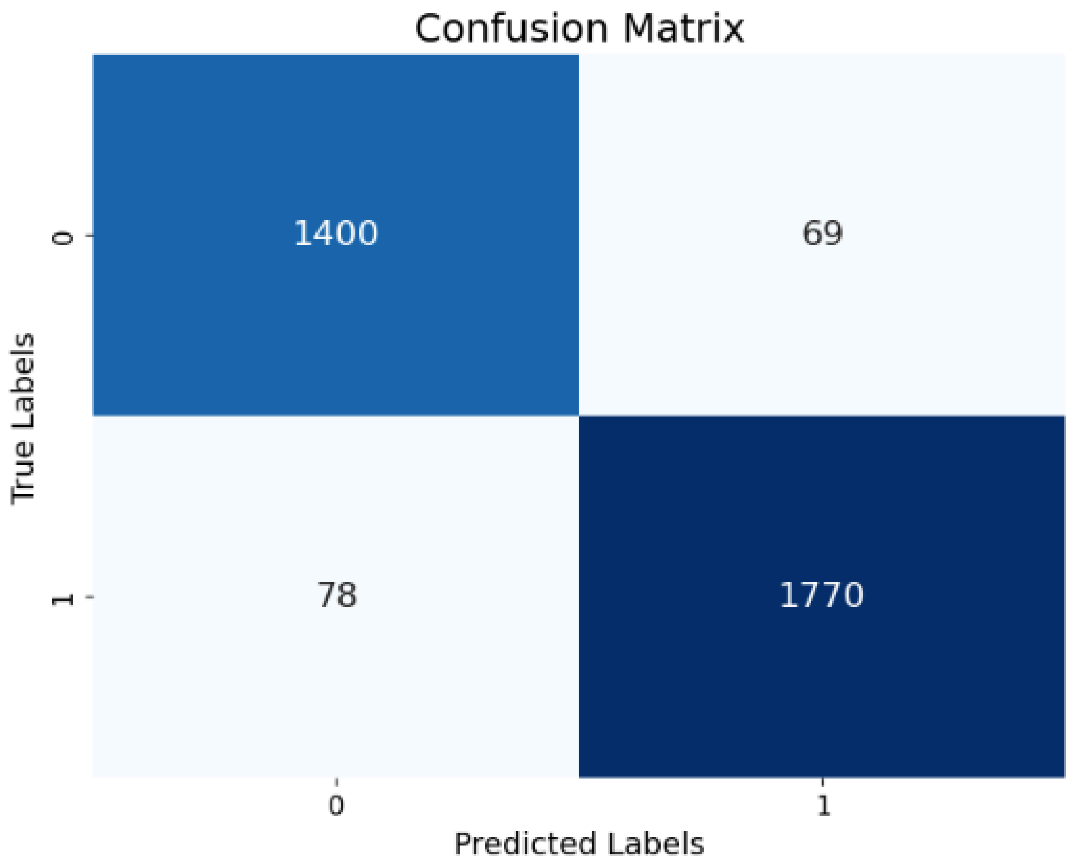
```
Accuracy score
0.776053662948448
Classification Report:
              precision    recall  f1-score   support

          -1       0.67      0.99      0.80      1469
           1       0.99      0.61      0.75      1848

    accuracy                           0.78      3317
   macro avg       0.83      0.80      0.77      3317
weighted avg       0.85      0.78      0.77      3317

Confusion Matrix:
[[1455   14]
 [ 727 1121]]
```

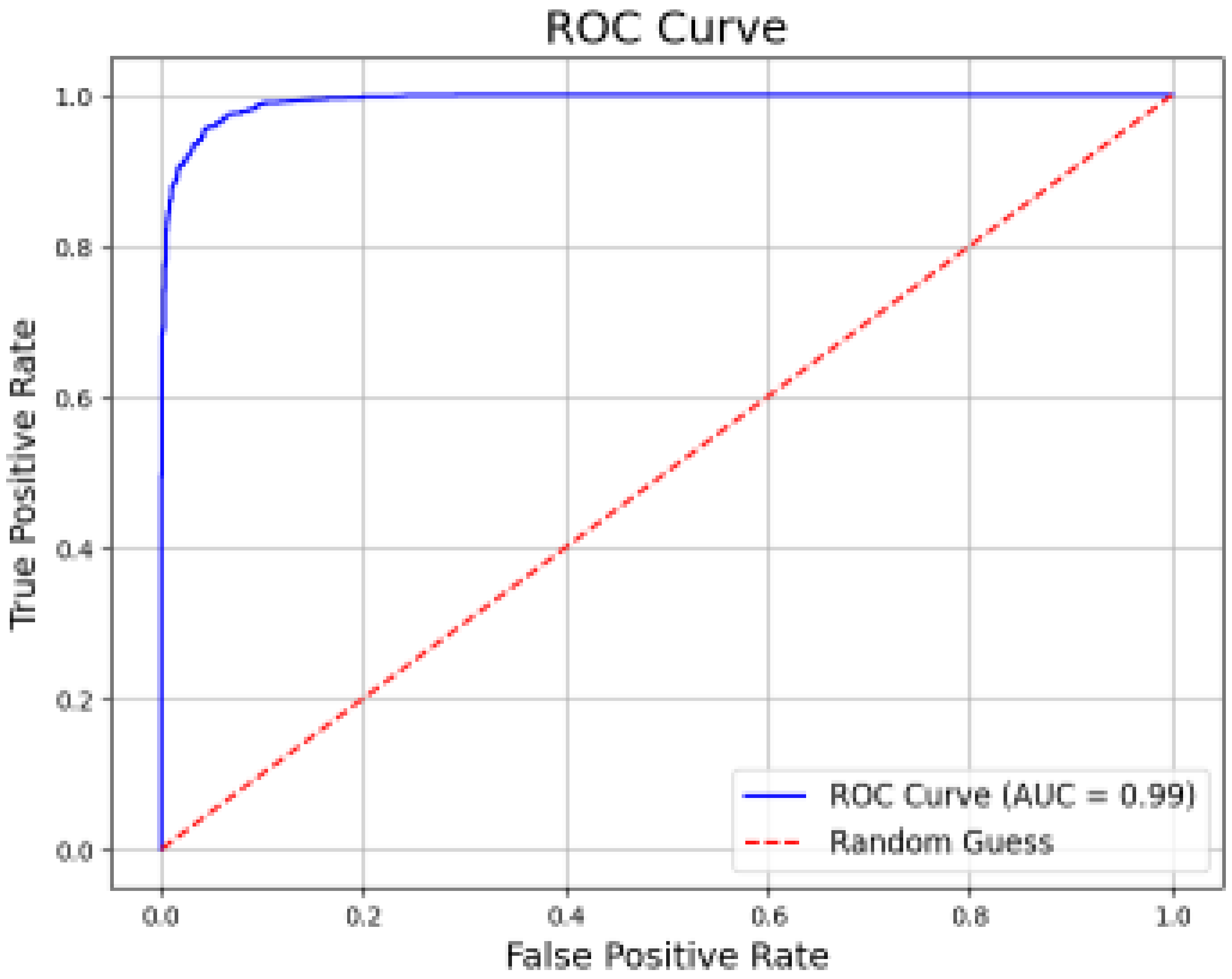




GRADIENT BOOSTING CLASSIFIER

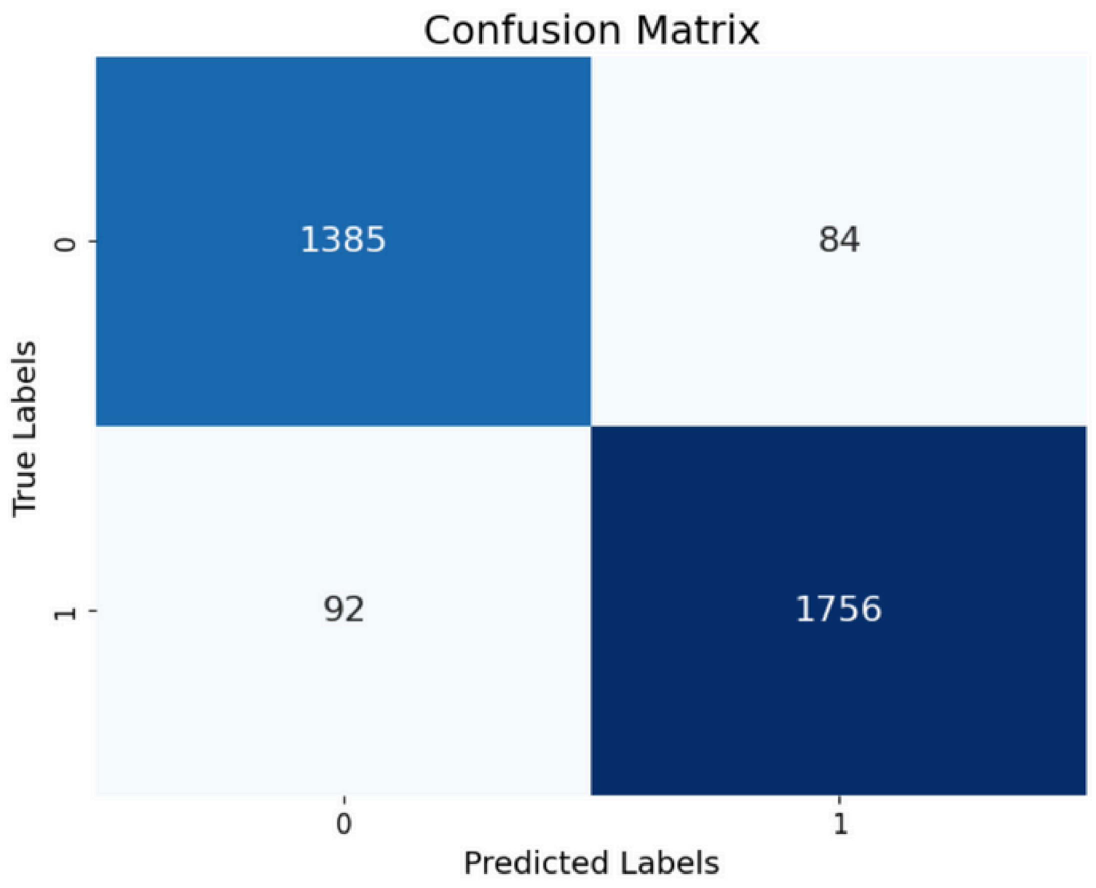
```
Accuracy score
0.9556828459451311
Classification Report:
              precision    recall  f1-score   support

          -1       0.95      0.95      0.95      1469
           1       0.96      0.96      0.96      1848

    accuracy                           0.96      3317
   macro avg       0.95      0.96      0.96      3317
weighted avg       0.96      0.96      0.96      3317

Confusion Matrix:
[[1400   69]
 [  78 1770]]
```

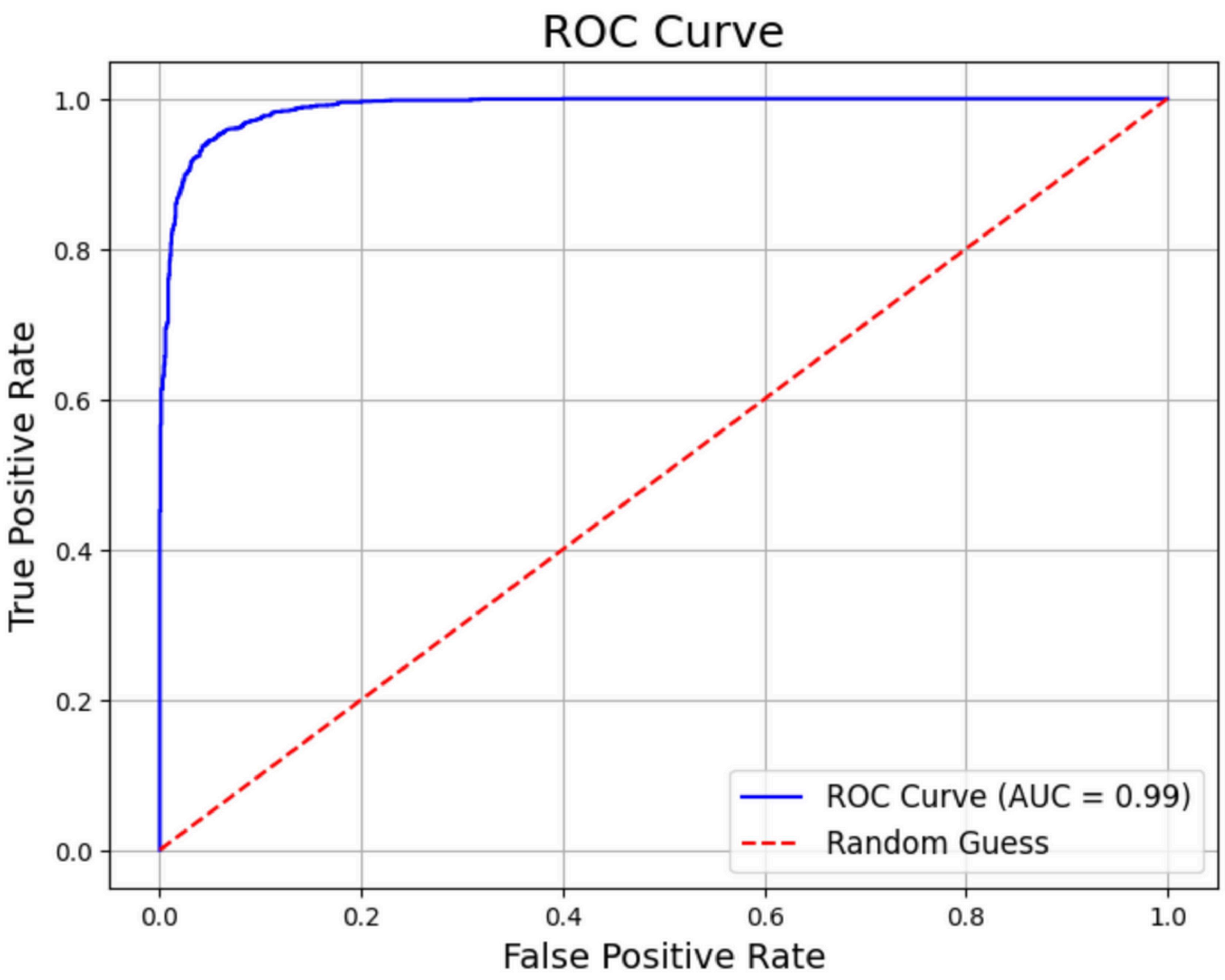




LOGISTIC REGRESSION CLASSIFIER

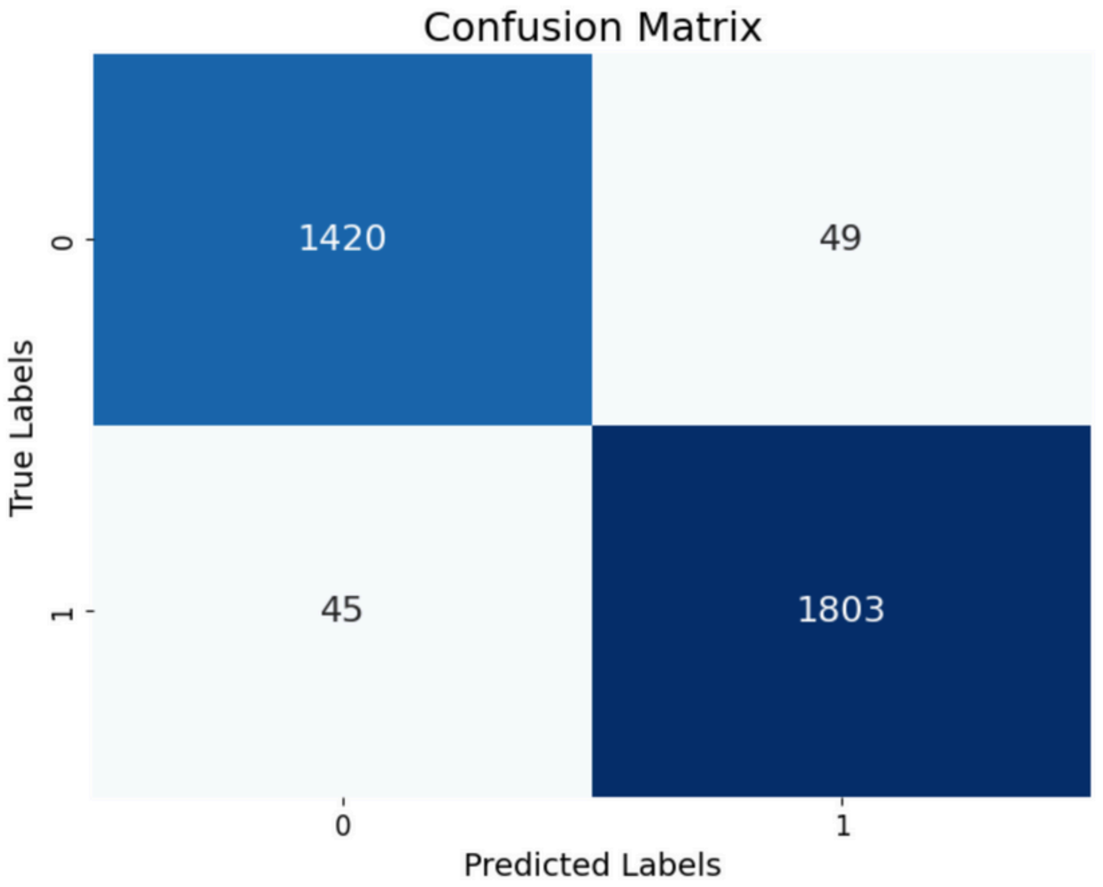
```
Accuracy score
0.9469400060295448
Classification Report:
              precision    recall  f1-score   support

          -1       0.94      0.94      0.94      1469
           1       0.95      0.95      0.95      1848

    accuracy                           0.95      3317
   macro avg       0.95      0.95      0.95      3317
weighted avg       0.95      0.95      0.95      3317

Confusion Matrix:
[[1385   84]
 [  92 1756]]
```

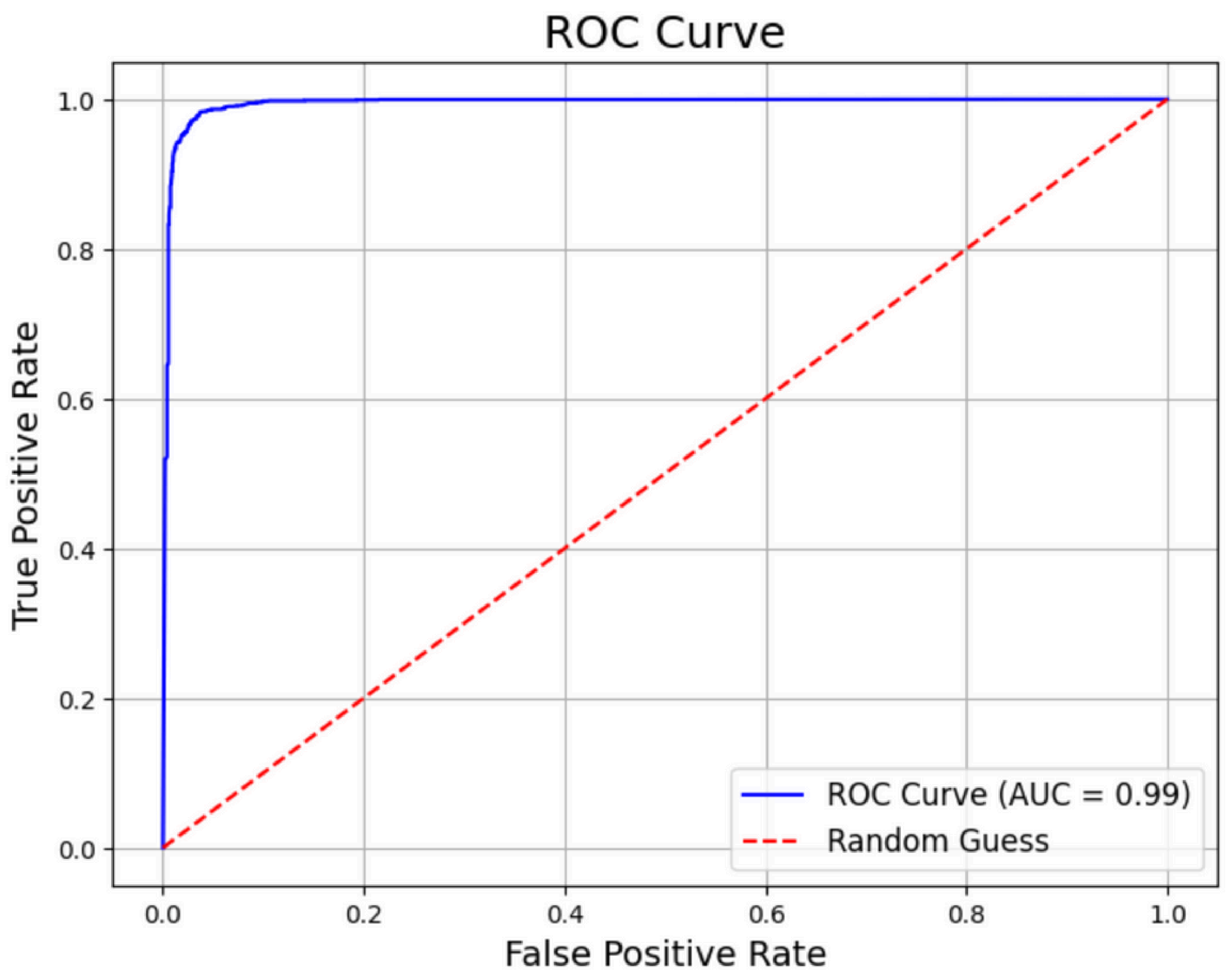




RANDOM FOREST CLASSIFIER

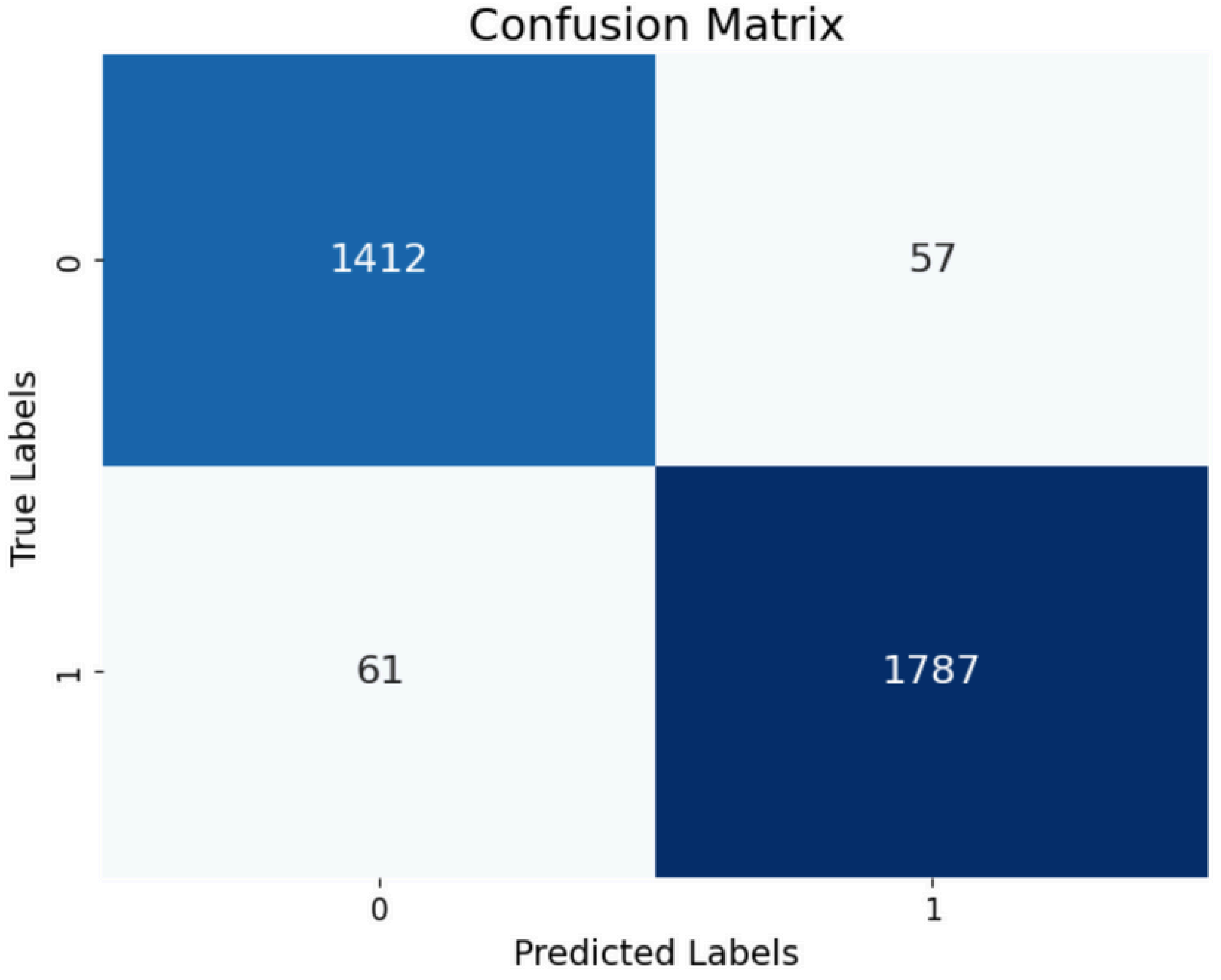
```
Accuracy score
0.9716611395839614
Classification Report:
              precision    recall  f1-score   support

          -1       0.97      0.97      0.97      1469
           1       0.97      0.98      0.97      1848

    accuracy                           0.97      3317
   macro avg       0.97      0.97      0.97      3317
weighted avg       0.97      0.97      0.97      3317

Confusion Matrix:
[[1420   49]
 [  45 1803]]
```

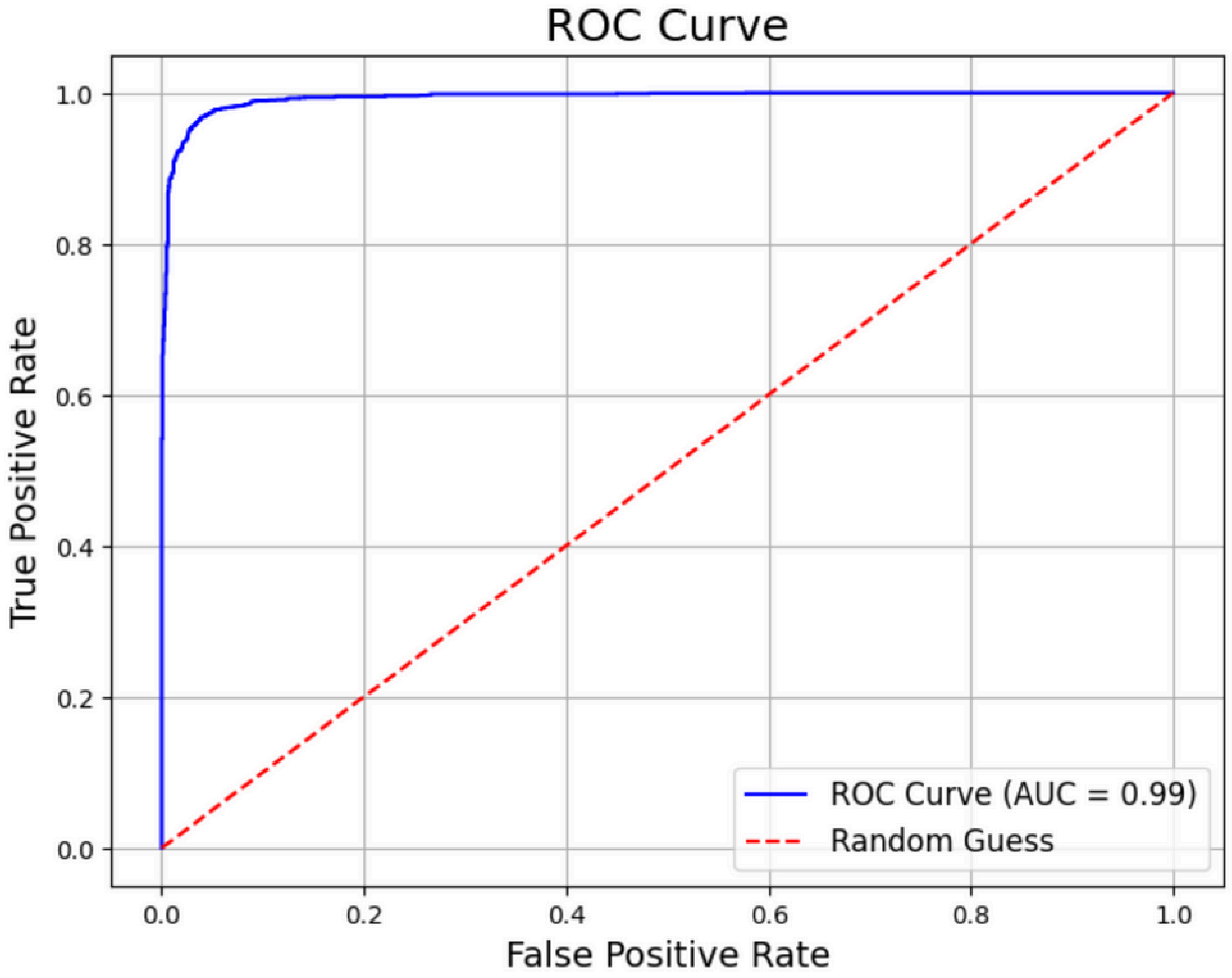




SUPPORT VECTOR MACHINE CLASSIFIER

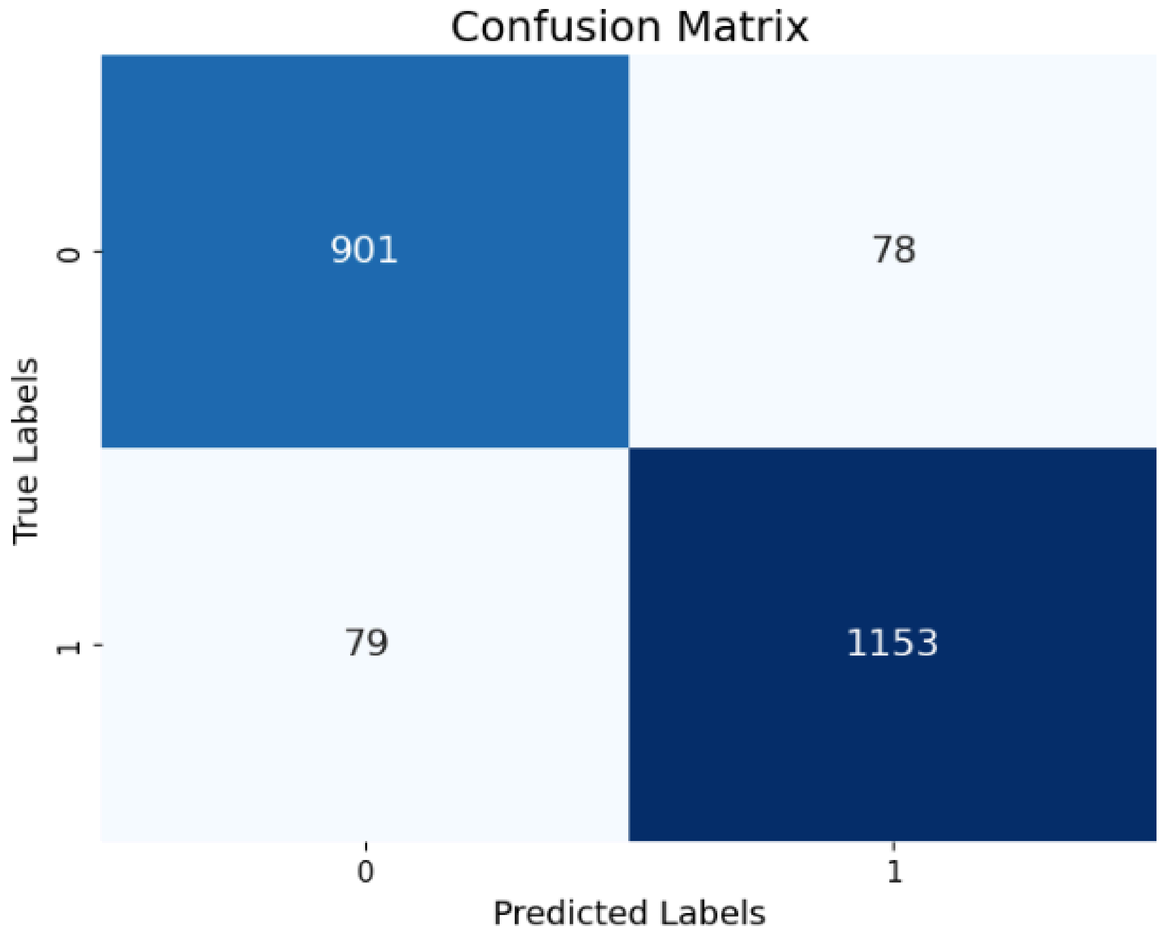
```
Accuracy score
0.9644256858607175
Classification Report:
              precision    recall  f1-score   support

          -1       0.96      0.96      0.96      1469
           1       0.97      0.97      0.97      1848

    accuracy                           0.96      3317
   macro avg       0.96      0.96      0.96      3317
weighted avg       0.96      0.96      0.96      3317

Confusion Matrix:
[[1412   57]
 [  61 1787]]
```

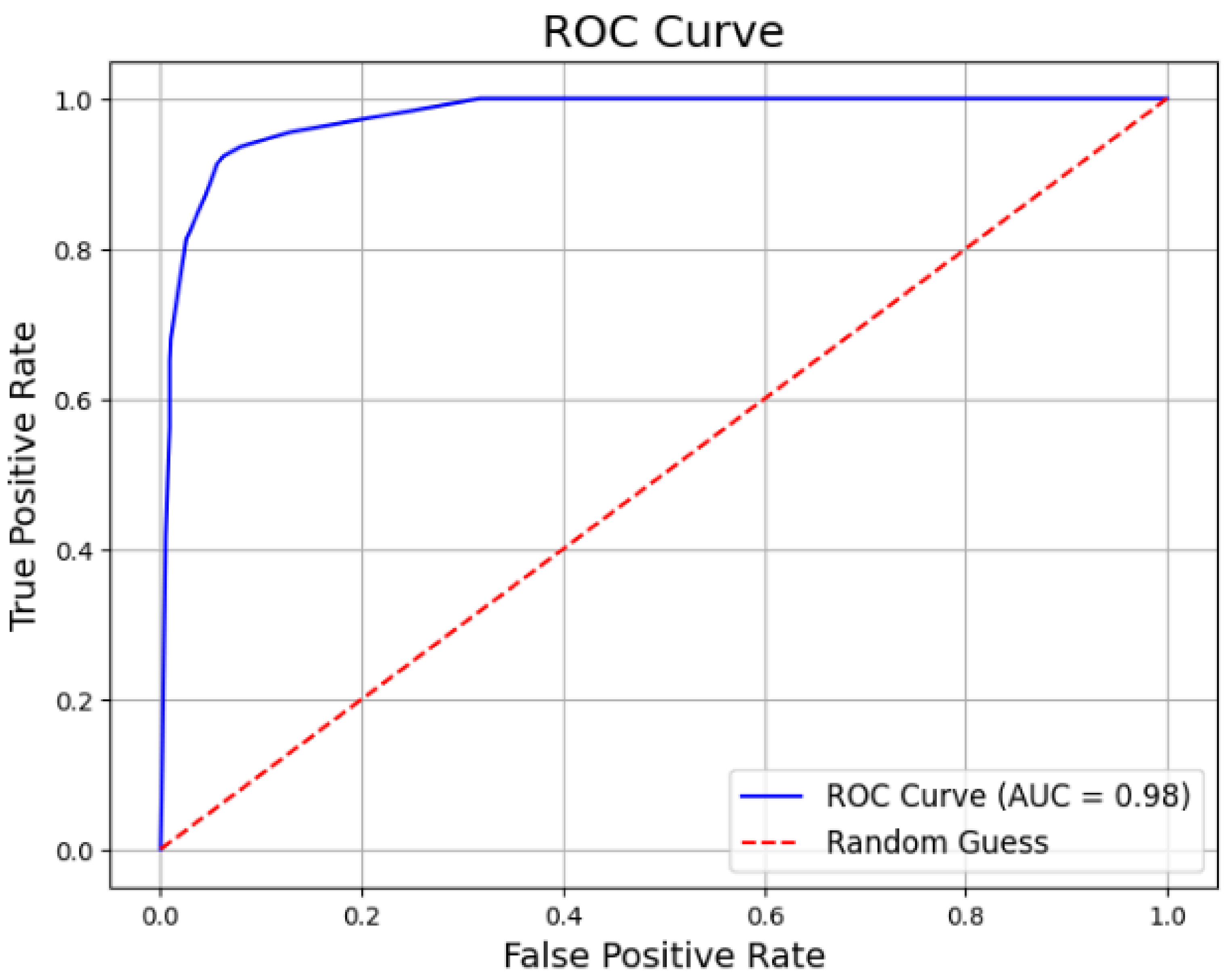


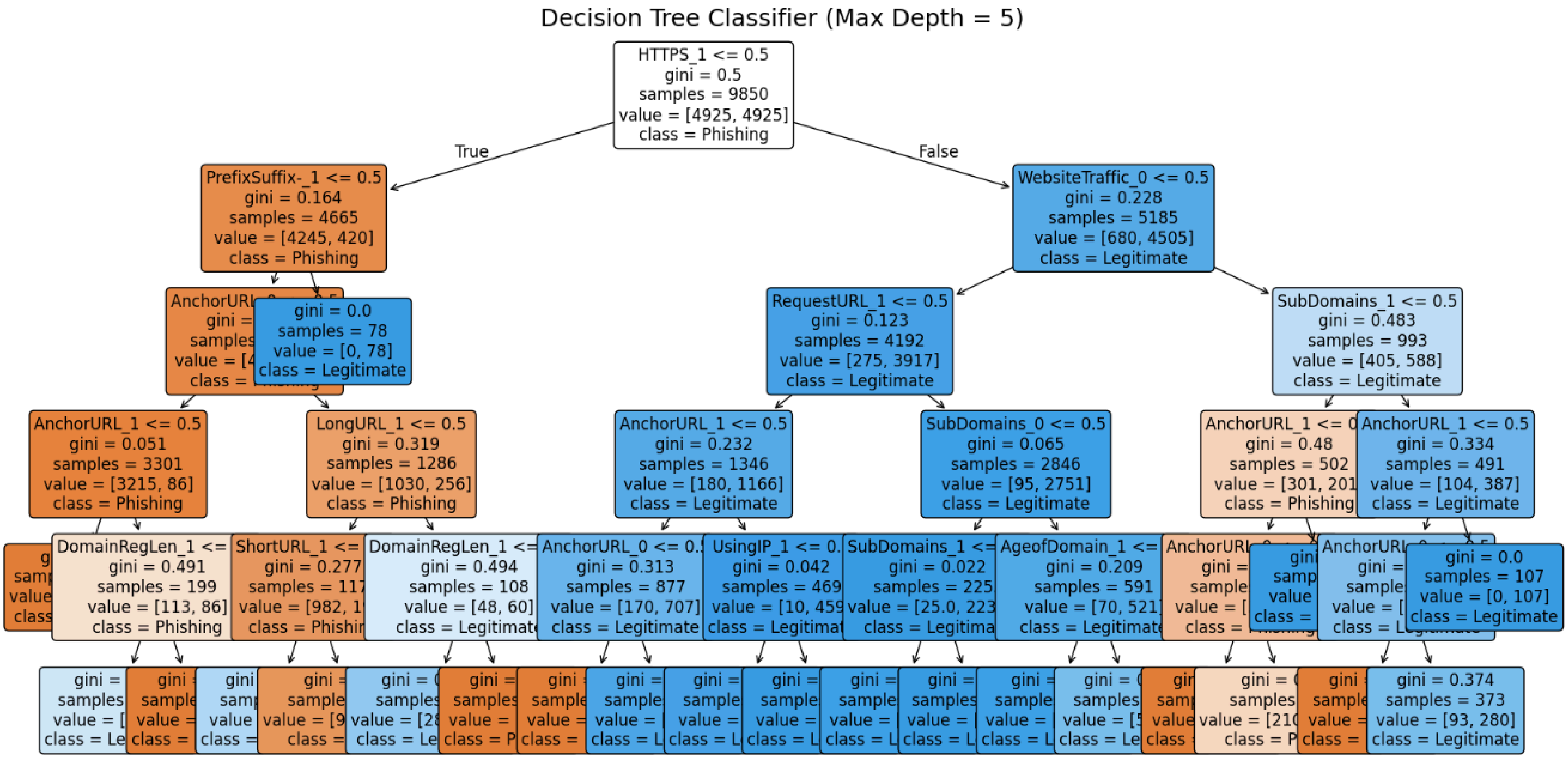


DECISION TREE CLASIFIER

```
Accuracy score
0.9289914066033469
Classification Report:
              precision    recall  f1-score   support

          -1       0.92      0.92      0.92       979
           1       0.94      0.94      0.94      1232

    accuracy                           0.93      2211
   macro avg       0.93      0.93      0.93      2211
weighted avg       0.93      0.93      0.93      2211
```



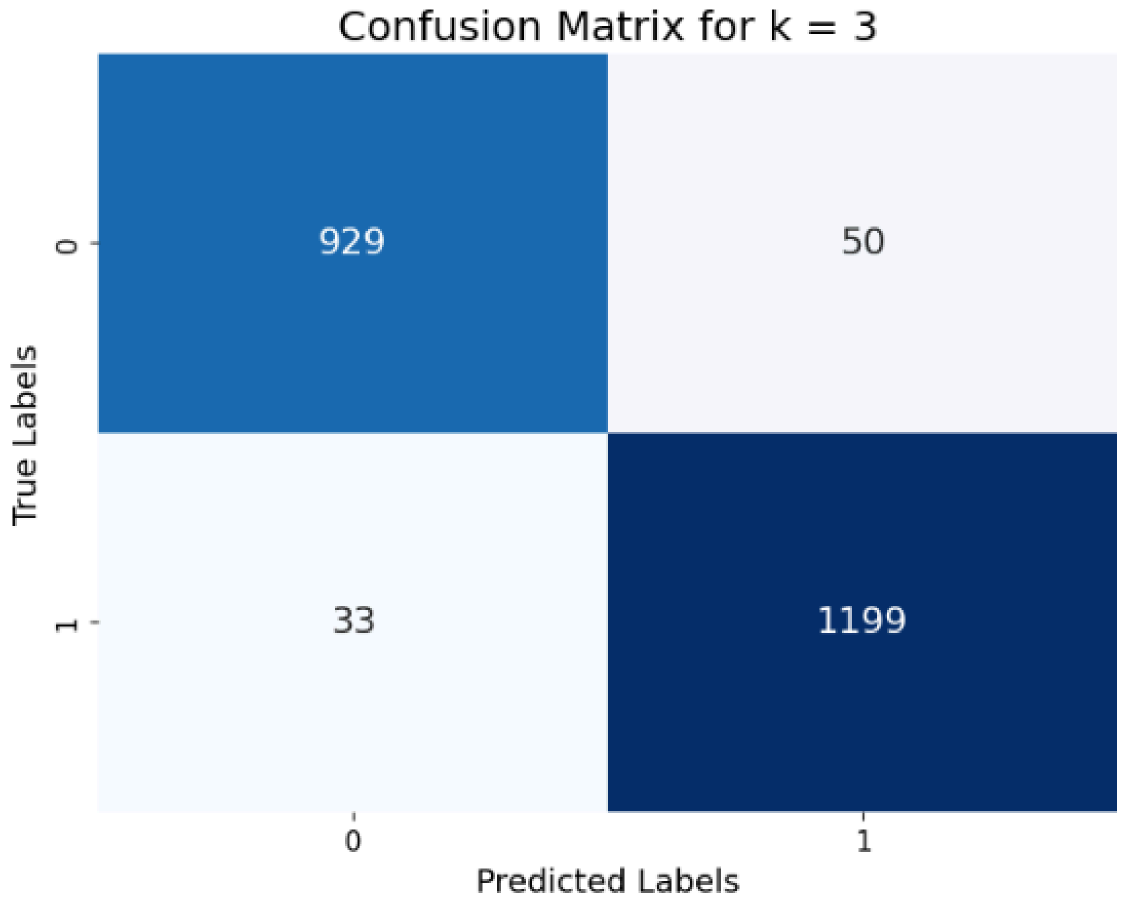


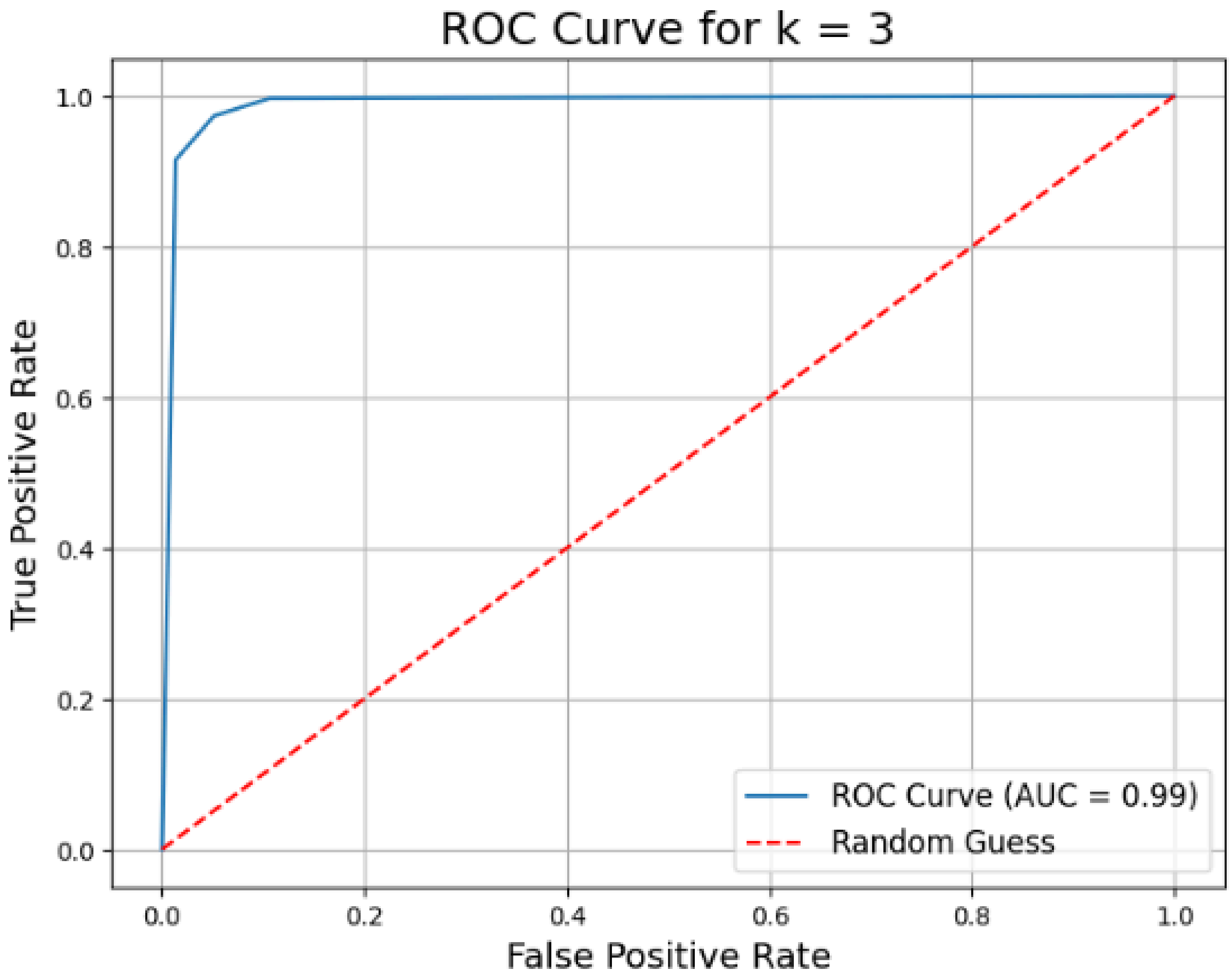

Decision Tree Classifier (Max Depth = 5)
HTTPS_1 <= 0.5
gini = 0.5
samples = 9850
value = [4925, 4925]
class = Phishing
True
False
PrefixSuffix-_1 <= 0.5
gini = 0.164
samples = 4665
value = [4245, 420]
class = Phishing
WebsiteTraffic_0 <= 0.5
gini = 0.228
samples = 5185
value = [680, 4505]
class = Legitimate
gini = 0.0
samples = 78
value = [0, 78]
class = Legitimate
RequestURL_1 <= 0.5
gini = 0.123
samples = 4192
value = [275, 3917]
class = Legitimate
SubDomains_1 <= 0.5
gini = 0.483
samples = 993
value = [405, 588]
class = Legitimate
AnchorURL_1 <= 0.5
gini = 0.051
samples = 3301
value = [3215, 86]
class = Phishing
LongURL_1 <= 0.5
gini = 0.319
samples = 1286
value = [1030, 256]
class = Phishing
AnchorURL_1 <= 0.5
gini = 0.232
samples = 1346
value = [180, 1166]
class = Legitimate
SubDomains_0 <= 0.5
gini = 0.065
samples = 2846
value = [95, 2751]
class = Legitimate
gini = 0.48
samples = 502
class = Phishing
AnchorURL_1 <= 0.5
gini = 0.334
samples = 491
value = [104, 387]
class = Legitimate
gini = 0.491
samples = 199
value = [113, 86]
class = Phishing
gini = 0.494
samples = 108
value = [48, 60]
gini = 0.313
samples = 877
value = [170, 707]
class = Legitimate
gini = 0.042
samples = 469
gini = 0.022
gini = 0.209
samples = 591
value = [70, 521]
gini = 0.0
samples = 107
value = [0, 107]
class = Legitimate
gini = 0.374
samples = 373
value = [93, 280]
class = Legitimate

K-NEAREST NEIGHBORS CLASSIFIER

- **For k=3**

```
Classification Report for k = 3:

              precision    recall  f1-score   support

          -1       0.97      0.95      0.96       979
           1       0.96      0.97      0.97      1232

    accuracy                           0.96      2211
   macro avg       0.96      0.96      0.96      2211
weighted avg       0.96      0.96      0.96      2211
```



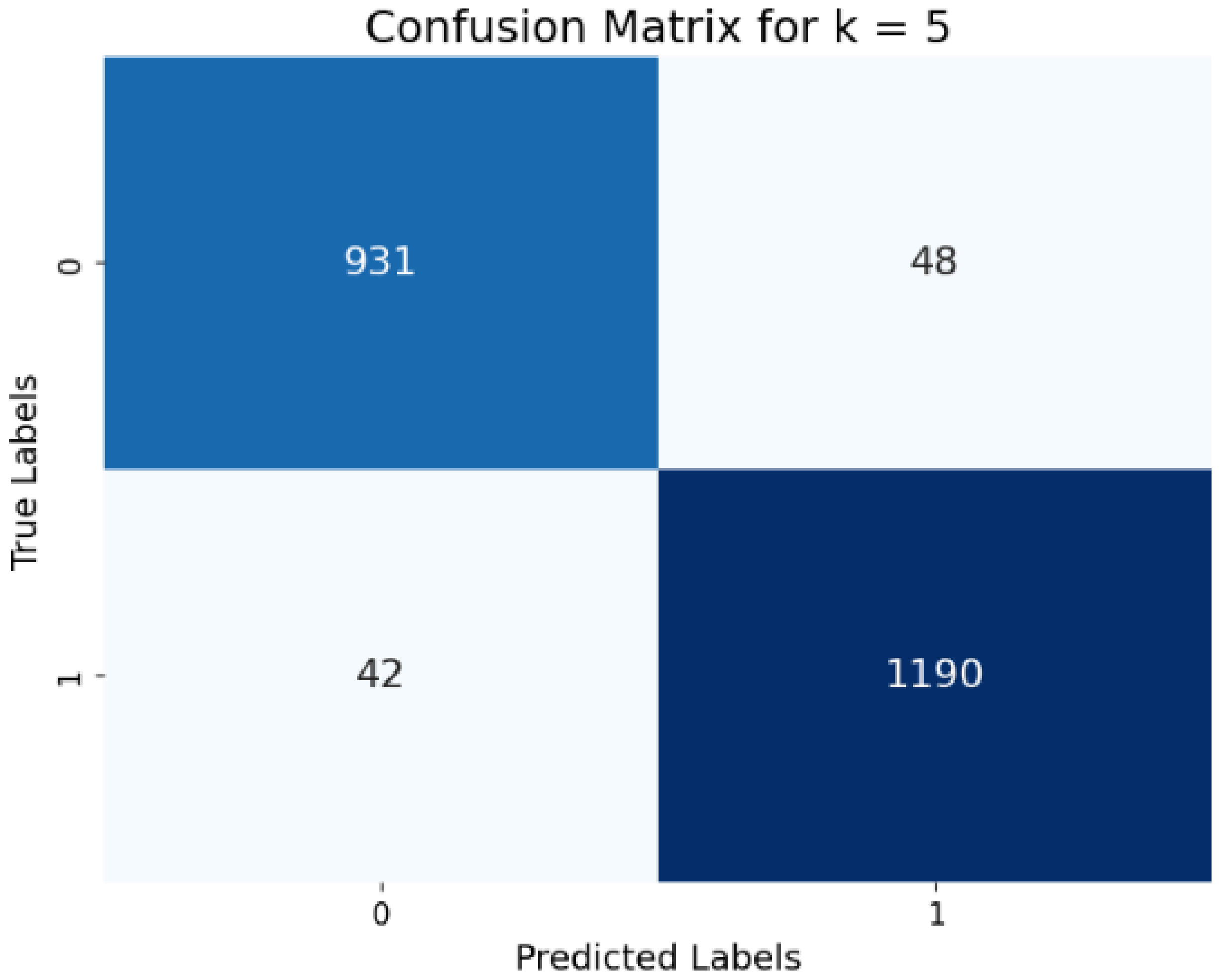


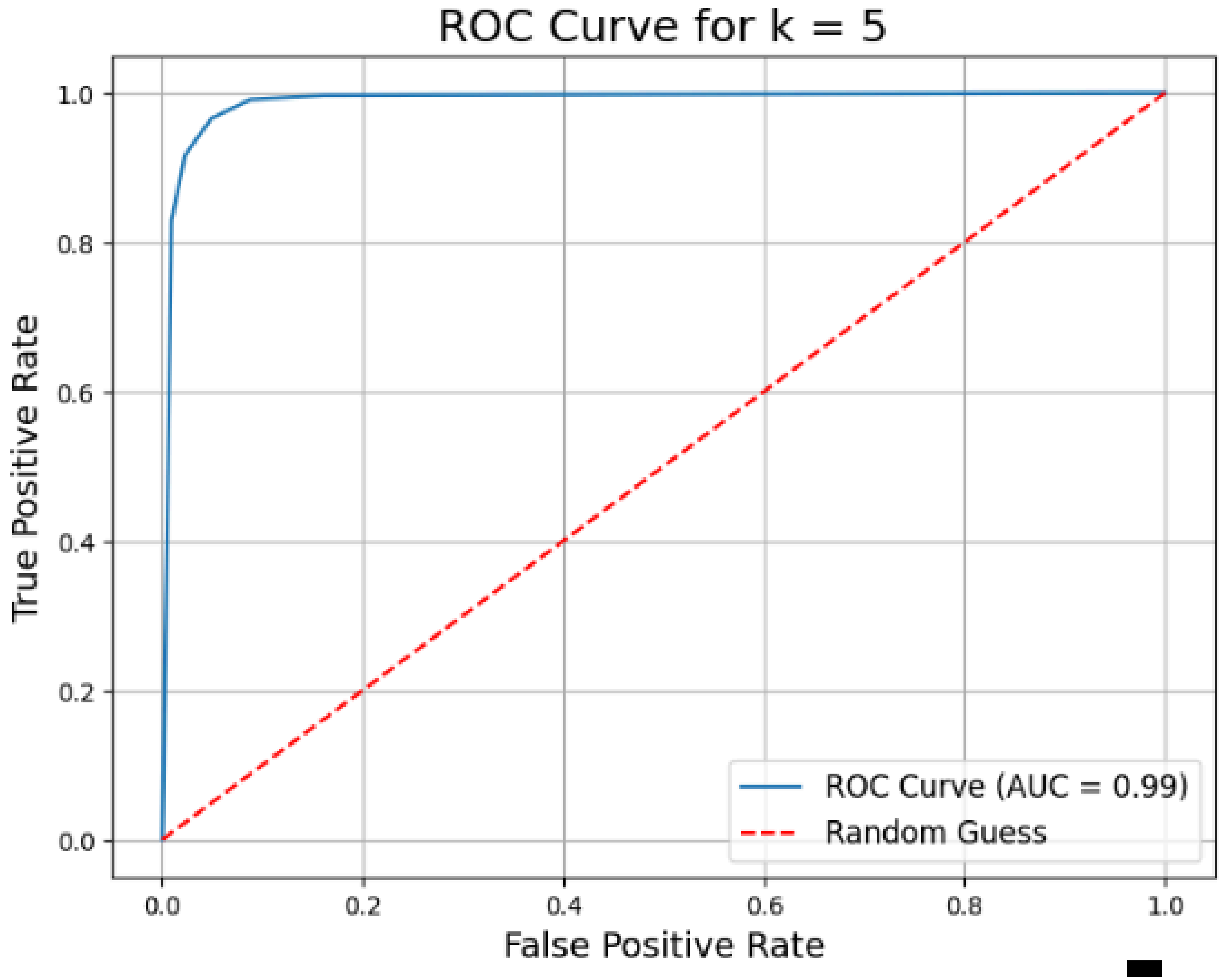
K-NEAREST NEIGHBORS CLASSIFIER

- **For k=5**

Classification Report for k = 5:

	precision	recall	f1-score	support
-1	0.96	0.95	0.95	979
1	0.96	0.97	0.96	1232
accuracy			0.96	2211
macro avg	0.96	0.96	0.96	2211
weighted avg	0.96	0.96	0.96	2211



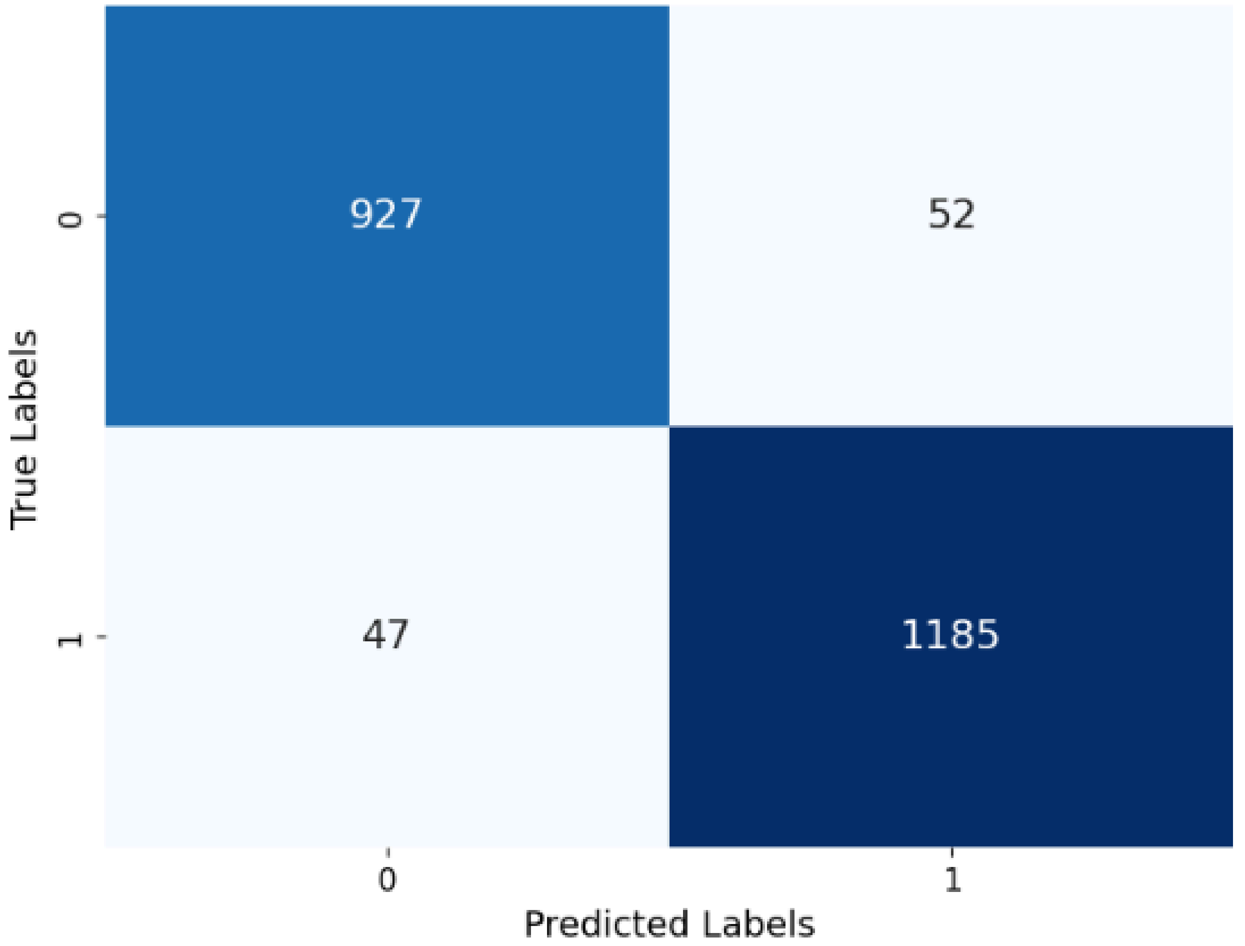


K-NEAREST NEIGHBORS CLASSIFIER

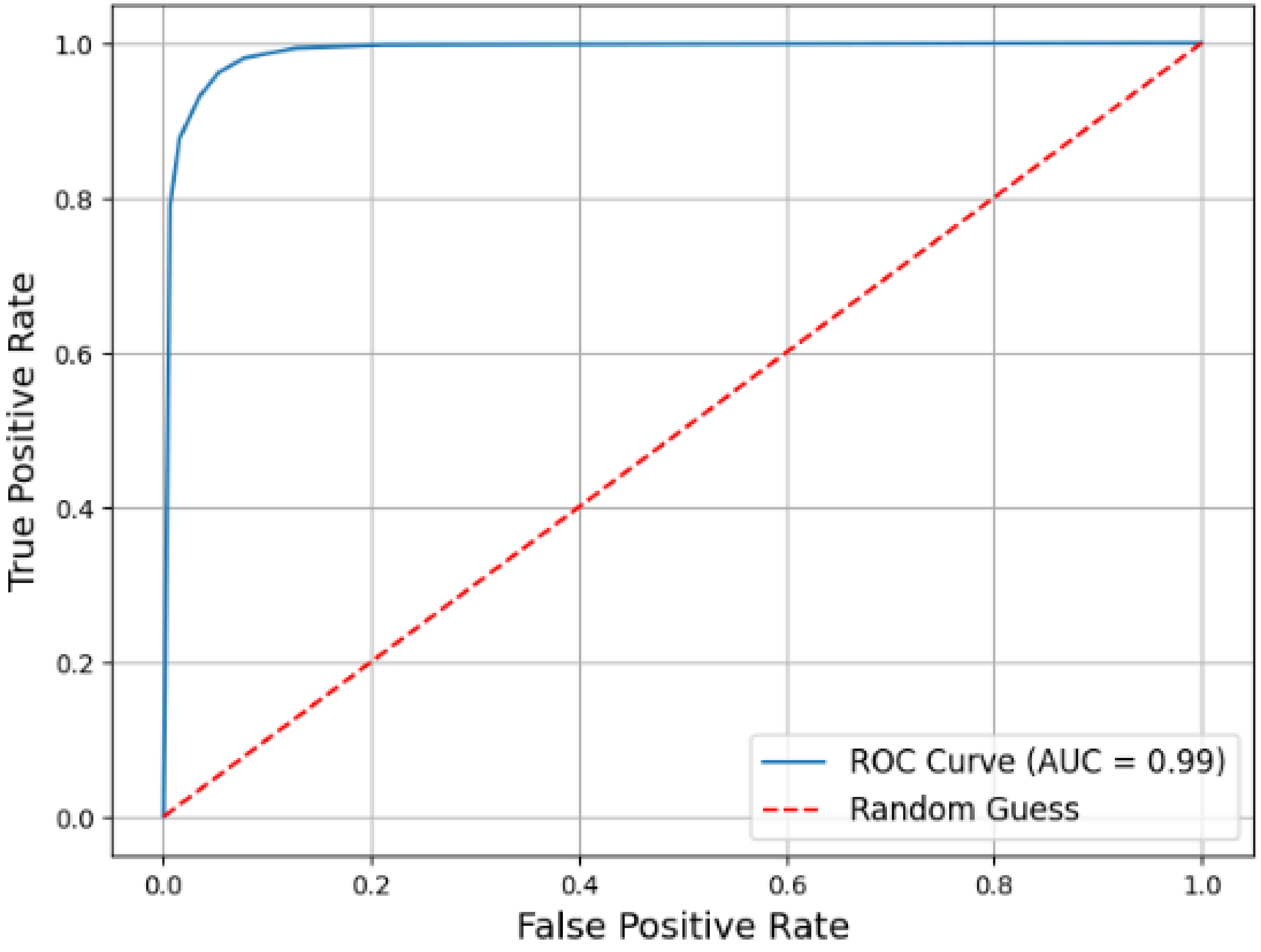
- **For k=7**

```
Classification Report for k = 7:

              precision    recall  f1-score   support

          -1       0.95      0.95      0.95       979
           1       0.96      0.96      0.96      1232

    accuracy                           0.96      2211
   macro avg       0.95      0.95      0.95      2211
weighted avg       0.96      0.96      0.96      2211
```

Confusion Matrix for k = 7

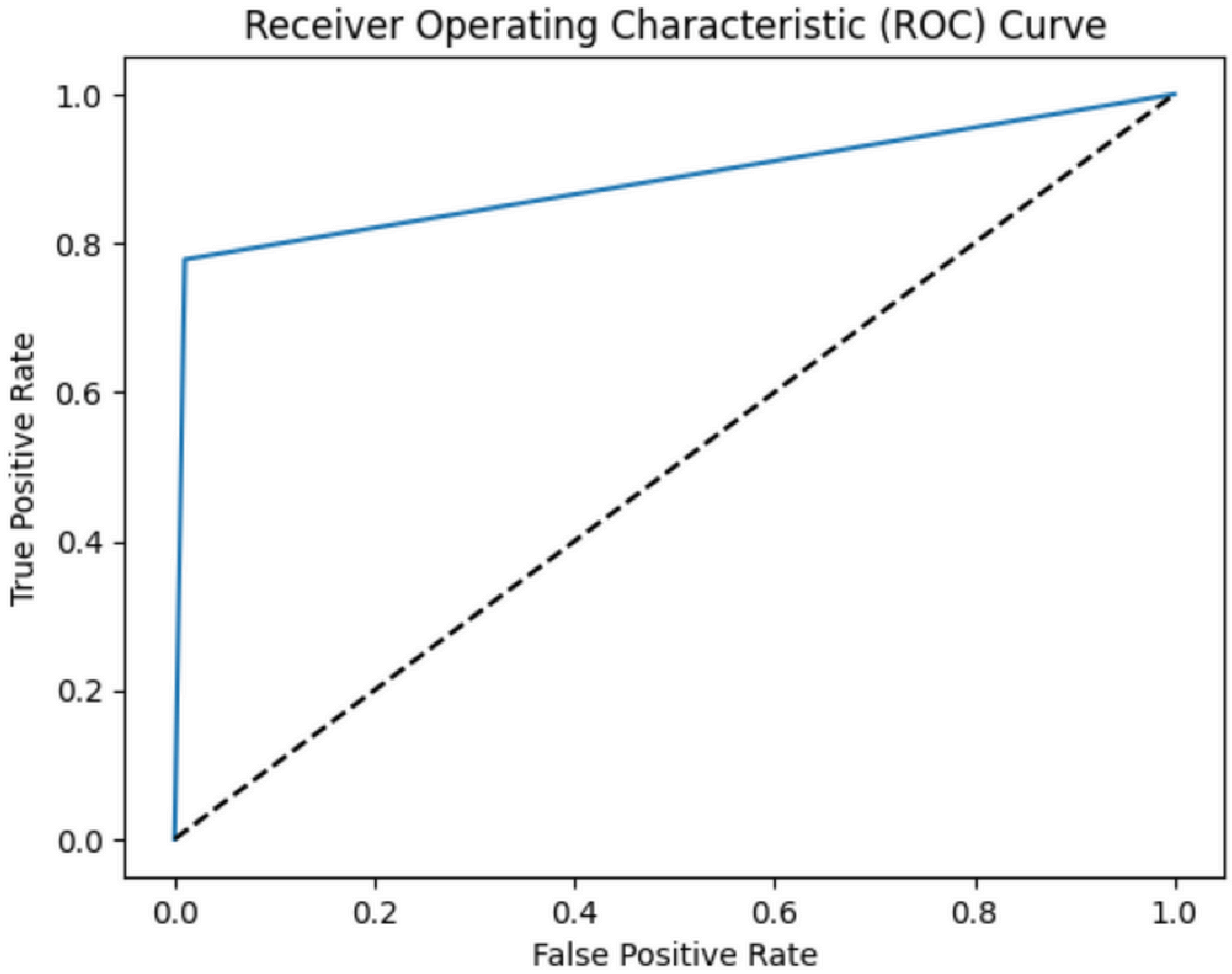


ROC Curve for k = 7



ARTIFICIAL NERUAL NETWORK

```
Accuracy: 0.8718721736508893
Classification Report:
              precision    recall  f1-score   support

          -1       0.78      0.99      0.87      1469
           1       0.99      0.78      0.87      1848

    accuracy                           0.87      3317
   macro avg       0.88      0.88      0.87      3317
weighted avg       0.90      0.87      0.87      3317

Confusion Matrix:
 [[1454   15]
 [ 410 1438]]
```



MODEL COMPARISON & CONCLUSION

Following is the conclusion using the various performance metrics for all the applied binary classification algorithms on our dataset:

Accuracy:**Random Forest Classifier** showed the highest accuracy of **97.1661%**
Precision:**Random Forest Classifier** has the highest precision of **97%**
Recall:**Random Forest Classifier** has the highest recall of **98%**
F1-score for **Random Forest Classifier** is **97**
Confusion matrix: False positive and False negative values are the least for **Random Forest Classifier**

So, concluding from the all the above-mentioned metrics "**Random Forest Classifier**" model is the best binary classification for our dataset.

Group 17

THANK YOU

Questions?