DETECTING PHISHING IN WEBSITES – A COMPARISON OF CLASSIFICATION TECHNIQUES

BY

R.BHARATHI

(17PCS002)

Project Report Submitted

In Partial fulfillment of the requirements for the award of

Master's Degree in Computer Science

Department of Computer Science,

Avinashilingam Institute for Home Science and Higher Education for Women,

(Deemed to be University),

Coimbatore-641043

April-2019

DETECTING PHISHING IN WEBSITES – A COMPARISON OF CLASSIFICATION TECHNIQUES

BY

R.BHARATHI

(17PCS002)

Project Report Submitted

In Partial fulfillment of the requirements for the award of

Master's Degree in Computer Science

Department of Computer Science,

Avinashilingam Institute for Home Science and Higher Education for Women,

(Deemed to be University),

Coimbatore-641043

April-2019

Signature of the Head of the Department **Signature of the Supervisor**

Viva Voce Examination Held on: _____

Signature of the Examiners

ACKNOWLEDGEMENT

ACKNOWLEDGEMENT

I would like to express my sincere thanks to God Almighty, for his constant love and grace that he showered upon me.

I would like to express my deep sense of reverential gratitude and sincere thanks to **Padma Shri, Dr. P. R. Krishnakumar, Chancellor,** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for his support and encouragement during the course of my study.

I owe my great deal of gratitude to **Dr. (Mrs.) V. Premavathy Vijayan, Vice Chancellor**, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for extending all resources that facilitated the conduct of the present work.

I express my gratitude to **Dr. (Mrs.) S. Kowsalya, Registrar**, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for providing all facilities necessary for the work.

I also thankful to **Dr. (Mrs.) K. Udaya Chandrika, Dean School of Physical Sciences & Computational of Sciences**, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for granting the facility required.

I wish to place my deep sense of gratitude to **Dr. (Mrs.) V. Radha, Professor and Head, Department of Computer Science,** for providing all the facilities required to complete the project.

I heartily thank my esteemed project guide **Dr. (Mrs.) B. Kalpana, Professor, Department of Computer Science**, for imparting the tremendous assistance and well-timed support for triumph of my project.

I express my honorable thanks to my project coordinator **Dr. (Mrs.) G. Padmavathi, Professor, Department of Computer Science,** for her kind advice and knowledgeable suggestions which helped me to complete my project successfully.

Finally, I take pride to thank my parents and those who helped me directly or indirectly for carrying out this work.

SYNOPSIS

SYNOPSIS

The project entitled as “Detecting phishing in websites – A comparison of classification techniques”. One of the online fraudulent act is website phishing. Phishing websites can cause a loss of revenue and leads to the damage of the brand image of organisation. The phishing website can be detected based on some important characteristics like URL, HTML and Java script, and Domain Identity. In order to detect phishing website, certain classification algorithms are used. The classification algorithm and techniques are implemented to extract the phishing websites characteristics to classify their legitimacy. In this project the performance of several algorithms are analyzed. The following algorithms are compared:

1. Support vector Machine
2. Naive Bayes
3. Random Forest

CONTENTS

TABLE OF CONTENT

INTRODUCTION

1. INTRODUCTION

Phishing refers to different types of online scams that 'phish' for personal and financial information (e.g., passwords, Social Security Number, bank account information, credit card numbers, or other personal information).These messages claim to come from a legitimate source: a well-known software company, online payment service, bank, or other reputed institution. Phishing messages can come from a growing number of sources, including: Email, phone calls, fraudulent software (e.g., anti-virus), Social Media messages (e.g., Facebook, Twitter), advertisements and text messages. Through the usage of data mining algorithm the phishing websites and the legitimate websites are identified, and the performance of different algorithms are compared.

1.1 PROBLEM DEFINITION

Compare the performance of Random Forest(RF), Naive Bayes(NB) and Support Vector Machine(SVM) in website phishing detection and to analyse the performance.

1.2 OVERVIEW OF THE PROJECT

Classification is the technique which is used to classify data records into one among set of predefined classes. A set of training data is developed that contains a certain set of the attributes as well as the likely outcomes. The job of the classification algorithm is to discover how that set of attributes helps in making the classification decision. In this project the performance measures of the Random Forest, Support Vector Machine and the Naive Bayes algorithms are analysed for detecting phishing websites.

SYSTEM SPECIFICATION

2. SYSTEM SPECIFICATION

This section describes the hardware and software specification needed for both development and implementation phases of the project.

2.1 HARDWARE SPECIFICATION

Monitor	: Generic PnP Monitor
RAM	: 2.00 GB
Processor	: Intel(R) Pentium(R)
Processor speed	: 2.10 GHZ
System Type	: 32-bit Operating System

2.2 SOFTWARE SPECIFICATION

Operating System	: Windows 8.1
Front End	: Spyder(Anaconda Distribution)
Back End	: Ms-Excel

2.3 ABOUT THE SOFTWARE

Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment.

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows users to launch applications and manage conda packages, environments and channels without using command-line commands.

The following applications are available by default in Navigator:

- JupyterLab
- Jupyter Notebook
- QtConsole
- **Spyder**
- Glueviz
- Orange
- Rstudio
- Visual Studio Code

Spyder is a powerful scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts. It offers a unique combination of the advanced editing, analysis, debugging, and profiling functionality of a comprehensive development tool with the data exploration, interactive execution, deep inspection, and beautiful visualization capabilities of a scientific package.

Top 5 Machine Learning Libraries in Python

- Numpy
- Pandas
- Matplotlib
- SciKit-Learn and
- NLTK

Numpy

Numpy is the fundamental package for scientific computing with Python. It is mostly used for solving **matrix** problems.

Pandas

Pandas is the most popular machine learning library written in python, for data manipulation and analysis.

Matplotlib

Matplotlib is a great library for **Data Visualization**.

SciKit-Learn

A library that provides a range of Supervised and Unsupervised Learning Algorithms. This library mainly focuses on **model** building.

NLTK

Natural Language Toolkit (NLTK**)** is a library for NLP **(Natural Langue Processing).**

Python Programming

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

- **Python is Interpreted** : Python is processed at runtime by the interpreter. Do not need to compile your program before executing it. This is similar to PERL and PHP.
- **Python is Interactive** : one can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
- **Python is Object-Oriented** :Python supports Object-Oriented style or technique of programming that encapsulates code within objects

.

PYTHON FEATURES

- **Easy-to-learn** − Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
- **Easy-to-read** − Python code is more clearly defined.
- **Easy-to-maintain** − Python's source code is fairly easy-to-maintain.
- **A broad standard library** − Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.
- **Interactive Mode** − Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
- **Portable** − Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
- **Extendable** – one can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
- **GUI Programming** − Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- **Scalable** − Python provides a better structure and support for large programs than shell scripting

USES OF PYTHON

Python is widely used for

- System programming
- GUI Programming
- Internet scripting
- Database Programming
- Gaming,Robotics
- Mozilla
- Government and non-profit organization

Excel

Microsoft Excel is a software program that allows users to organize, format and calculate data with formula using a spreadsheet system. This software is a part of the Microsoft office suite and compatible with other applications in the office suite.

Excel is a commercial spreadsheet application produced and distributed by Microsoft for Microsoft windows and Mac OS X. It features the ability to perform basic calculations, use graph tools, create pivot tables and create macro programming language.

Excel has the same basic features as every spreadsheet, which use a collection of cells arranged into rows and columns to organize data manipulation. They also display data as charts, histograms and line graphs.

The features of Microsoft Excel are

- Multi-threading recalculation(MTR) for commonly used functions
- Improved pivot tables
- More condition formatting options
- Additional image editing capabilities
- In-cell chart called spark lines
- Ability to preview before pasting
- Ability to customize the Ribbon

Uses of Excel

Microsoft Word is a word processing program used for **writing** letters, memos, reports and paper presentations. Microsoft Excel is a spreadsheet program used for calculations, **making** charts and recording data about all kinds of business processes.

2.4 About the Dataset:

Dataset of "**Phishing Websites Dataset**" were collected from the UCI Machine Learning Repository. This dataset was gathered from: Phish Tank archive, Miller Smiles archive, Google's searching operators.

It consists of 11055 websites samples. All samples are already categorized as 1 for "Legitimate", 0 for "Suspicious" and -1 for "Phishy".

Dataset phishing criteria is divided into 4 sections (Address Bar based Features, Abnormal Based Features, HTML and JavaScript based Features and Domain based Features) and it has 30 attributes.

Table: Categories of Features

Feature Group	Features Names
Address Bar based Features	Using the IP Address
	Long URL to Hide the Suspicious Part
	Using URL Shortening Services "TinyURL"
	URL's having "@" Symbol
	Redirecting using "//"
	Adding Prefix or Suffix Separated by (-) to the Domain
	Sub Domain and Multi Sub Domains
	HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)
	Domain Registration Length
	Favicon
	Using Non-Standard Port

Abnormal-based features	URL of Anchor
	Links in <Meta>
	<Script> and <Link> tags
	Server Form Handler (SFH)
	Submitting Information to Email and Abnormal URL

HTML and JavaScript-based features	Website Forwarding
	Status Bar Customization
	Disabling Right Click
	Using Pop-up Window
	IFrame Redirection
Domain-based features	Age of Domain
	Website Traffic
	Page Rank
	Google Index
	Number of Links Pointing to Page
	Statistical-Reports Based Feature
	DNS Record

SYSTEM DESIGN

3. SYSTEM DESIGN

System design is the art of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements. It evaluates the systems, actual functionality in relation to expected or intended functionality.

3.1 Input Design

Input design is a very essential part in the project and should be concentrated well as it is prone to error. The data are to be used, plays an important role. In order to get the meaningful output and to achieve good accuracy, the input should be acceptable and understandable.

In this project, the input information will be in the form of dataset. The dataset used here is a Phishing Websites dataset. The size of the dataset is 835KB.

3.2 Output Design

Output design generally refers to the results and information that are generated by the system. The dataset are taken as input, various techniques of data mining are being applied to generate variety of output. The output will feature the class lables that convey the presence or absence of phishing.

METHODOLOGY

4. METHODOLOGY

4.1 Proposed system

In this work implementation of the Support Vector Machine, Random Forest, and the Naive Bayes for classification is done**.** Classification is a data mining function that assigns items in a collection to target categories or classes. It is a two step process – training and testing. The goal of classification is to accurately predict the target class for each case in the data.

This project is designed with five modules. They are

- **Loading the dataset**
- **Data preprocessing**
- **Splitting the dataset - Training and Testing**
- **Classification Algorithm implementation**
- **Performance measures – Analysis**

4.2 Loading the dataset

Download the ’Phishing Website’ dataset from a CSV file in UCI repository. It consists of 11055 instances of websites samples and 30 attributes. To load the dataset into the python, pandas data frame can be created using the read_csv function. Prior to this pandas packages should be imported.

4.3 Data Preprocessing

Data preprocessing involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. In this project there is no null value or no noisy data. So data slicing method is used for separate the feature and labels from the dataset. It is helpful for splitting the training and test data.

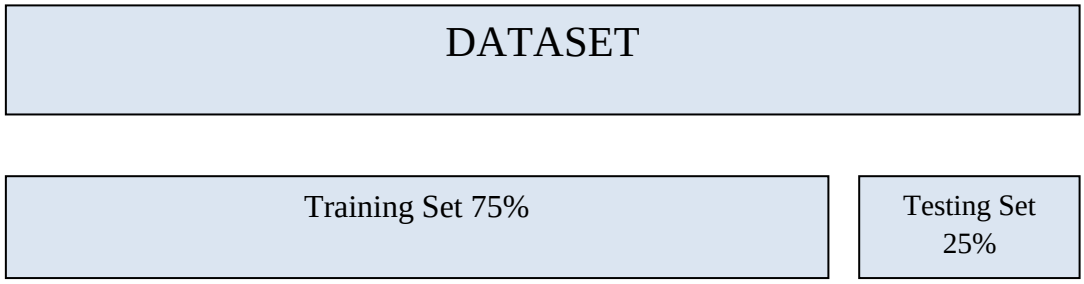
4.4 Splitting the dataset – Training and Testing

In order to assess the performance of the classifier. The data is split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later. We have the test dataset (or subset) in order to test our model's prediction on this subset.

4.4.1 Train Test Split:

It splits the data into a training set and a test set. The proportion of splitting the data should be mentioned, to use as a test set.

Figure: 4.1 **Train Test Split Format**



4.4.2 K Fold Cross Validation:

In the **k-fold cross validation** method, all the entries in the original training data set are **used** for both training as well as **validation**. Also, each entry is **used** for **validation** just once. The table below shows combination of training and testing data.

Table: 4.1 **K Fold Cross Validation Split format**

Split 1	**Test**	**Train**	**Train**	**Train**	**Train**
Split 2	**Train**	**Test**	**Train**	**Train**	**Train**
Split 3	**Train**	**Train**	**Test**	**Train**	**Train**
Split 4	**Train**	**Train**	**Train**	**Test**	**Train**
Split 5	**Train**	**Train**	**Train**	**Train**	**Test**

DATASET

Table: 4.2 **Example of K Fold Cross Validation**

Split 1	**0-2210**	Train	Train	Train	Train
Split 2	Train	**2211-4421**	Train	Train	Train
Split 3	Train	Train	**4422-6632**	Train	Train
Split 4	Train	Train	Train	**6633-8843**	Train
Split 5	Train	Train	Train	Train	**8834-11054**
	0-2210	**2211-4421**	**4422-6632**	**6633-8843**	**8834-11054**
	11055				

4.5 Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. A classification task begins with a data set in which the class assignments are known. The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values. In this project '-1 = phishy' and '1=legitimate' are the two classification labels. Classification models are tested by comparing the predicted values to known target values in a set of test data.

Apply three classification algorithm such as Random Forest, Support Vector Machine and Naive Bayesian for the Phishing Websites dataset.

4.5.1 Naive Bayesian classification

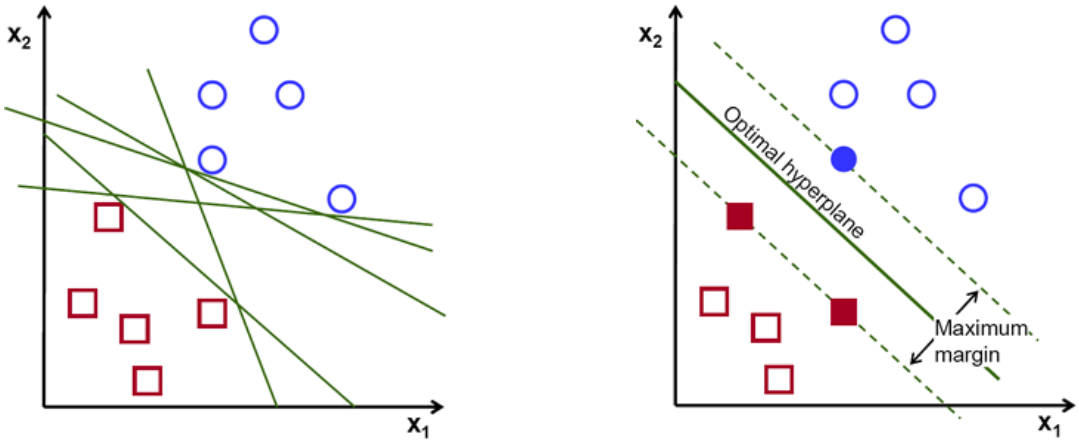
The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. It assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems.

$$Prob(B / A) = Prob(A / B)\ prob(B)/Prob(A) \quad (\textbf{Baye's therom}) \qquad (1)$$

4.5.2 Support Vector Machine Classification

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, perform classification by finding the hyper-plane that differentiate the two classes with a large margin .

Figure: 4.2 **Detect the optimal hyperplane**



The dots and square are the classes based on features. The hyper plane is the line that separates two classes. The different hyper plane will be drawn. The best hyper plane will be find from the margin width.

4.5.3 Random forest

Random forest algorithm is a supervised classification algorithm. This algorithm creates the forest with a number of trees. It builds multiple decision trees and merges them together to get a more accurate and stable prediction. . In the random forest classifier, the **higher the number** of trees, higher the accuracy. Random forest overcome the problem of over fitting of the decision trees.

Important terms:

Entropy - It is a measure of randomness and unpredictability in the dataset.

Information Gain - It is the measure of decrease in entropy after the dataset is split.

Root Node - The Top most decision node is known as Root Node.

Decision Node - Decision Node has two or more Branches.

Leaf Node - It has exactly one incoming Node and no outgoing Node

4.6 Performance measure

Accuracy:

The accuracy of a classifier refers to the ability of classifier to correctly predict the class label of new or previously unseen data . Accuracy can be estimated using one or more test sets that are independent of the training set. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. From the RF, SVM and NB classifier the accuracy percentage has been found. Accuracy refers to the percentage of correct predictions made by the model when compared with the actual classifications in the test data.

Accuracy=Number of correct predictions/Total of all cases to be predicted

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+FP+TN}} \qquad \textbf{(2)}$$

Confusion Matrix:

A confusion matrix displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. The matrix is n*n, where n is the number of classes. The Confusion matrix is in the form of

Table: 4.3 Confusion Matrix Format

Predicted

Classes	-1	1
-1	(True Negative)	(False Positive)
1	(False Negative)	(True Positive)

Actual

¶

-1=Phishy, 1=Legitimate

Table: 4.4 Example of Confusion Matrix

predicted classes

Classes	-1	1	Total
-1	1185 **(TN)**	64 **(FP)**	1249
1	26 **(FN)**	1489 **(TP)**	1515
Total	1211	1553	2764

Actual classes

Table: 4.5 Interpretation of Confusion matrix

TP	The number of legitimate website correctly classified as legitimate
TN	The number of websites classified correctly as phishing website
FP	The number of legitimate websites classified as phishing website
FN	The number of websites classified as legitimate websites when they were actually phishing websites.

Precision:

Precision is the number of correct positive result divided by the number of total predicted by the classifier positive.

True Positive + False Positive = Total Predicted Positive

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad \textbf{(3)}$$

Recall:

Recall is the number of correct Positive result divided by the number of all actual positive.

True Positive + False Negative = Actual Positiv

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad \textbf{(4)}$$

F1 Score:

The F1 score conveys the balance between the precision and the recall. It is also called the F Score or the F Measure.

$$\text{F1} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision+Recall}} \qquad \textbf{(5)}$$

EXPERIEMENTAL RESULTS AND ANALYSIS

5. EXPERIMENTAL RESULTS AND ANALYSIS

The experiment is performed in Spyder environment.

5.1 Evaluation Measures

In this experiments, 4 different measures are used for the evaluation of the classification quality: accuracy, precision, recall and F1 score.

Two types of splitting method are used for finding the classification quality:

Train Test Split – It splits the dataset randomly here the test data splitted as 25% among the total dataset.

K Fold Cross validation – It splits the data into 5 fold each fold is act as a training as well as testing data.

The performance evaluation of **Train Test Split(Test 25%)** is listed below:

Table: 5.1 Performance of the Algorithm

Metrics	RF	SVM	NB
Accuracy	0.9743	0.9674	0.8997
Precision	0.9721	0.9607	0.9142
Recall	0.9827	0.9828	0.9101
F1 Score	0.9803	0.9706	0.9122

The dataset has performed better in Random forest algorithm. The accuracy, precision, recall and f1 score the better result than the other algorithms.

The performance evaluation **of K fold Cross Validation** is listed below:

Table: 5.2 Accuracy - k Split = 5

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Accuracy	**RF**	0.9823	**0.9796**	**0.9769**	**0.9601**	**0.9402**	**0.9657**
	SVM	**0.9828**	0.9760	0.9706	0.9588	0.9344	0.9647
	NB	0.9014	0.9023	0.9122	0.8973	0.8968	0.9020

From Table 5.2 It is observed that on taking the average accuracy over the 5 folds(0.9657), Random Forest(RF) performs the best.

Table: 5.3 Precision - K Split = 5

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Precision	**RF**	0.9726	**0.9755**	**0.9767**	**0.9541**	**0.9427**	**0.9664**
	SVM	**0.9785**	0.9708	0.9649	0.9506	0.9509	0.9631
	NB	0.9141	0.9102	0.9276	0.8960	0.8995	0.9095

From Table 5.3 It is observed that on taking the average precision over the 5 folds(0.9664), Random Forest(RF) performs the best.

Table: 5.4 Recall - K Split = 5

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Recall	**RF**	0.9879	**0.9892**	0.9797	0.9717	**0.9512**	**0.9738**
	SVM	**0.9911**	0.9876	**0.9829**	**0.9774**	0.9300	**0.9738**
	NB	0.9096	0.9117	0.9141	0.9242	0.9170	0.9170

From Table 5.4 It is observed that on taking the average recall over the 5 folds(0.9738), Random Forest(RF) and Support Vector Machine(SVM) performs the best.

Table: 5.5 F1 Score - K Split = 5

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
F1 Score	**RF**	0.9835	**0.9823**	**0.9781**	0.9628	**0.9469**	**0.9701**
	SVM	**0.9847**	0.9791	0.9739	**0.9638**	0.9404	0.9684
	NB	0.9118	0.9109	0.9208	0.9099	0.9082	0.9123

From Table 5.5 It is observed that on taking the average f1 score over the 5 folds(0.9701), Random Forest(RF) performs the best.

Analysis:

In this phishing website dataset the two types of splitting method are used. In both the methods the Random Forest performs better in accuracy, precision, recall and f1 score. So it is concluded, the Random Forest algorithm gives best result for finding the phishing and legitimate websites.

CONCLUSION

6. CONCLUSION

Website Phishing is the most common online fraud all over the world. The performance has been analysed using the three classification algorithms namely Support Vector Machine(SVM), Random Forest(RF) and Naive Bayes(NB) classification algorithm. The Confusion Matrix is used depict to find the performance of the three algorithms. Two types of splitting methods are used - Train Test Split and K Fold Cross validation method. While using train test split, the random forest achieved 97% accuracy and in K Fold Cross Validation achieved 96% of accuracy. It may be concluded that RF gives an overall performance better than SVM and NB.

SCOPE OF FUTURE ENHANCEMENT

7. SCOPE FOR FUTURE ENHANCEMENT

Future work can includes proposing some new features, experimentally assigning new rules to some well known features and updating some other features. Also, experimental will be done based on other classification algorithms that can be applied on the data set and the best techniques can be identified.

BIBLIOGRAPHY

8. BIBLIOGRAPHY

- A. M. Shahiri and W. Hussain,"A Review on Predicting Student's Performance Using Data Mining Techniques", Proceeding Computer Science, vol. 72.
- Blum, A.I., and Langley. P. "Selection of relevant features and examples in machine learning", Artificial Intelligence, vol 97,1997.
- C.Romero, J. R. Romero and S. Ventura," A survey on pre-processing educational data", In Educational Data Mining. Springer International Publishing, 2014.
- C.Romero, S. Ventura, P. G. Espejo and C. Herv as," Data mining algorithms to classify students", in: Educational Data Mining, vol.2008.

Copyright © 2017 Aurelien Geron.

- Hands-On Machine Learning with Scikit-Learn and TensorFlow by Aurelien Geron

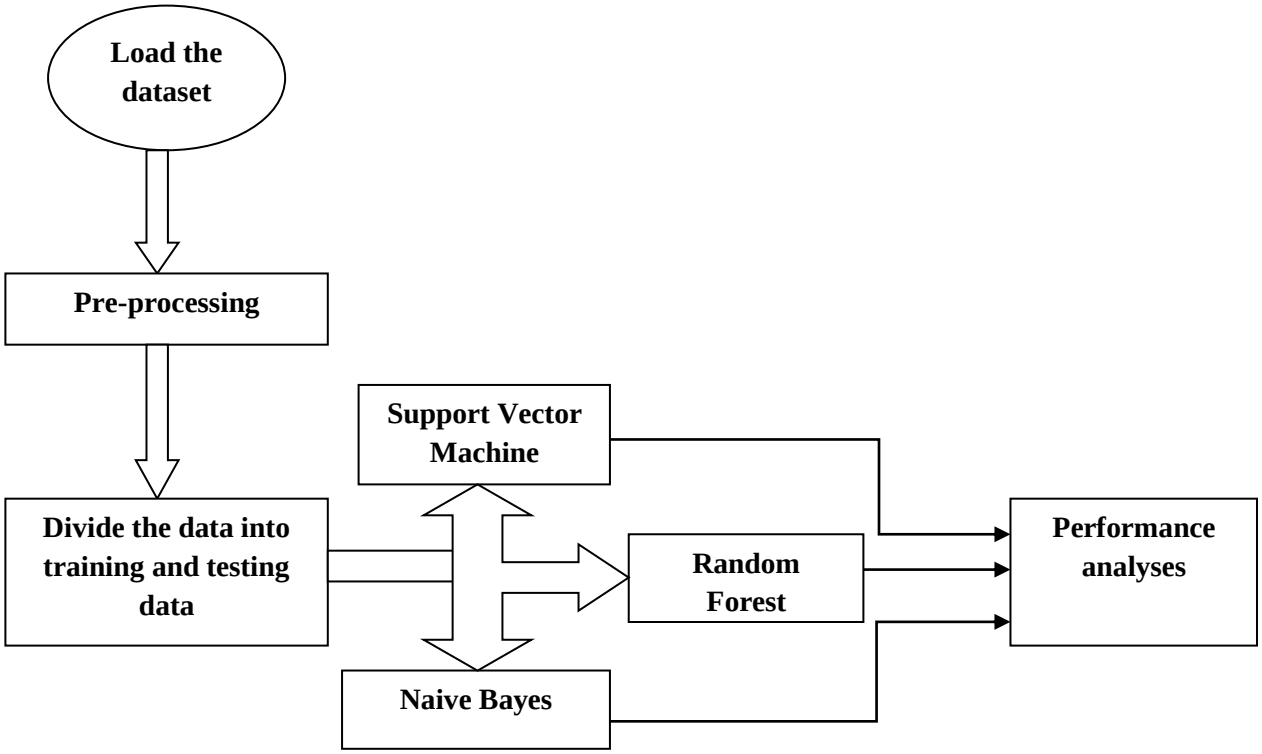
WEBSITES

- https://towardsdatascience.com
- https://www.analyticsvidhya.com
- https://www.datasciencecentral.com
- https://www.datacamp.com

APPENDIX

9. APPENDIX

9.1 WORK FLOW DIAGRAM



9.2 SCREEN SHOTS

IMPORT THE DATASET:

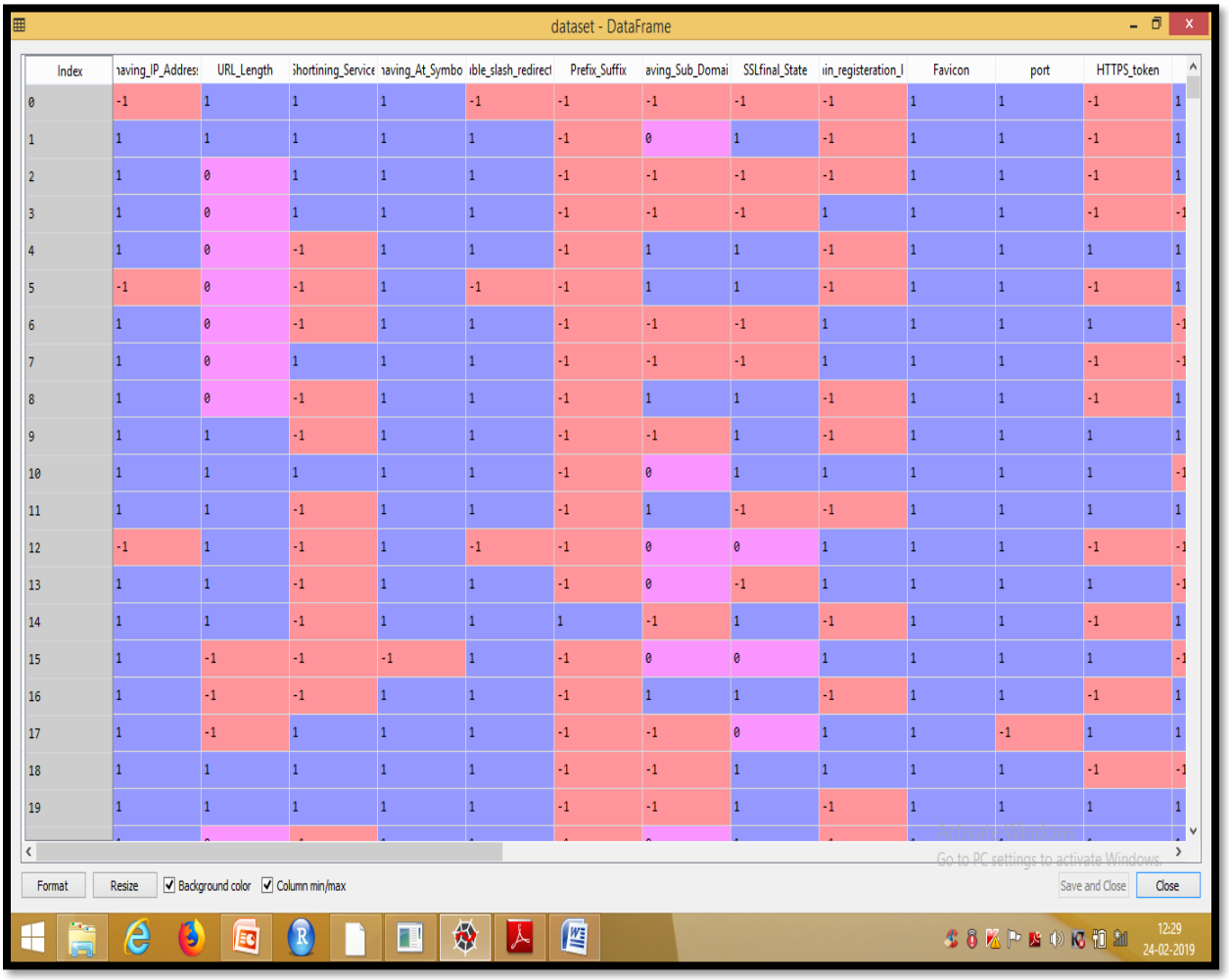


Figure: 9.1

The dataset will be loaded into the python using pandas data frame. It can be created using read_csv function. Before that the pandas packages should be imported.

CHECKING FOR NULL VALUES:

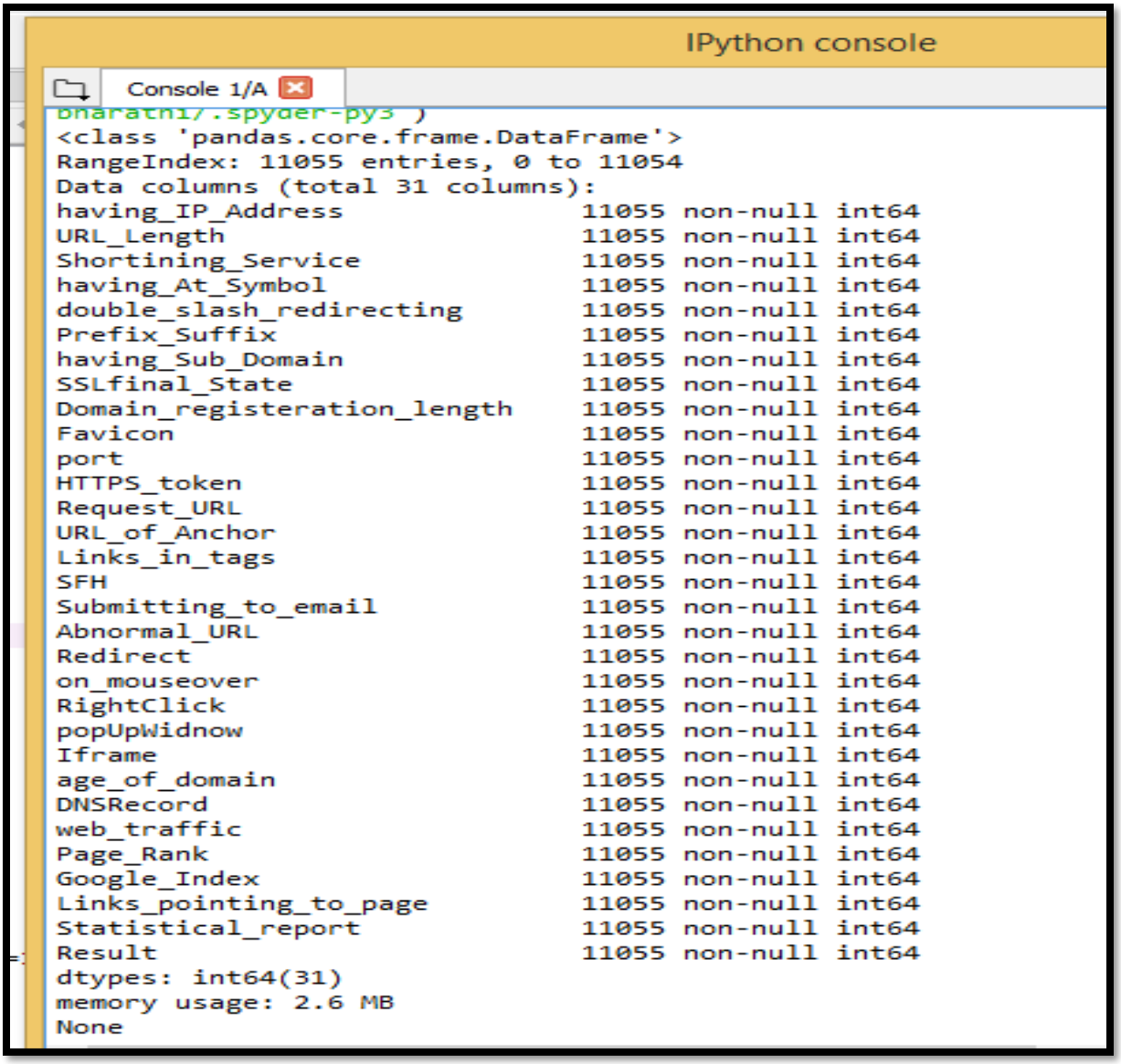


Figure: 9.2

The dataset contain 11055 instance and 30 attributes and it has no null values.

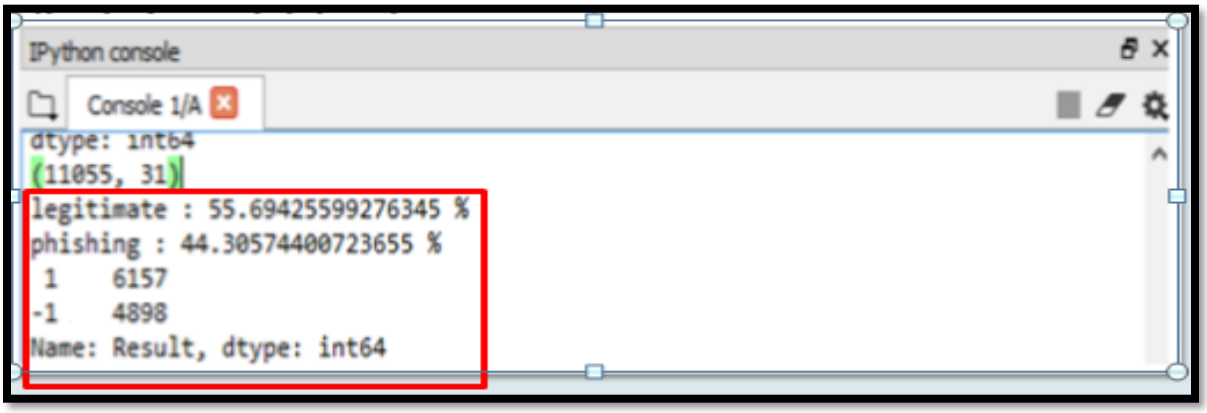


Figure: 9.3

Finding the number of legitimate and phishing website in the dataset .

Figure: 9.4

Displaying the number of phishing and legitimate websites in the dataset using the bar plot.

TRAIN TEST SPLIT:

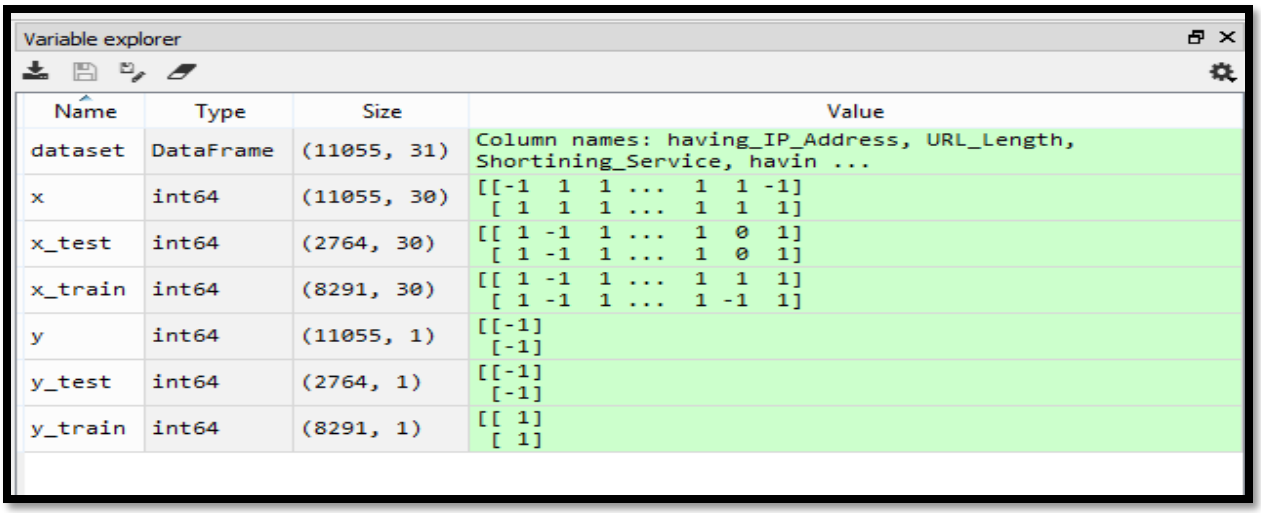


Figure: 9.5

In this project 25% of dataset is used as testing set and the rest of the 75% of dataset is used as training set. From the train test split method the dataset was spiltted.

K FOLD CROSSVALIDATION:

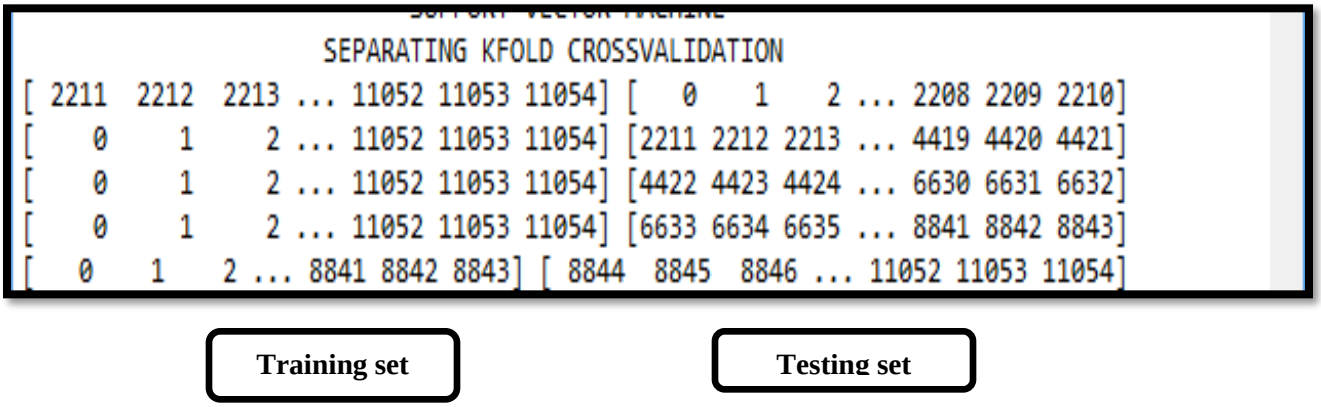


Figure: 9.6

In the K Fold Cross validation the dataset is separated into 5 Fold.

NAIVE BAYES PERFORMANCE METRICS

Train Test Split:

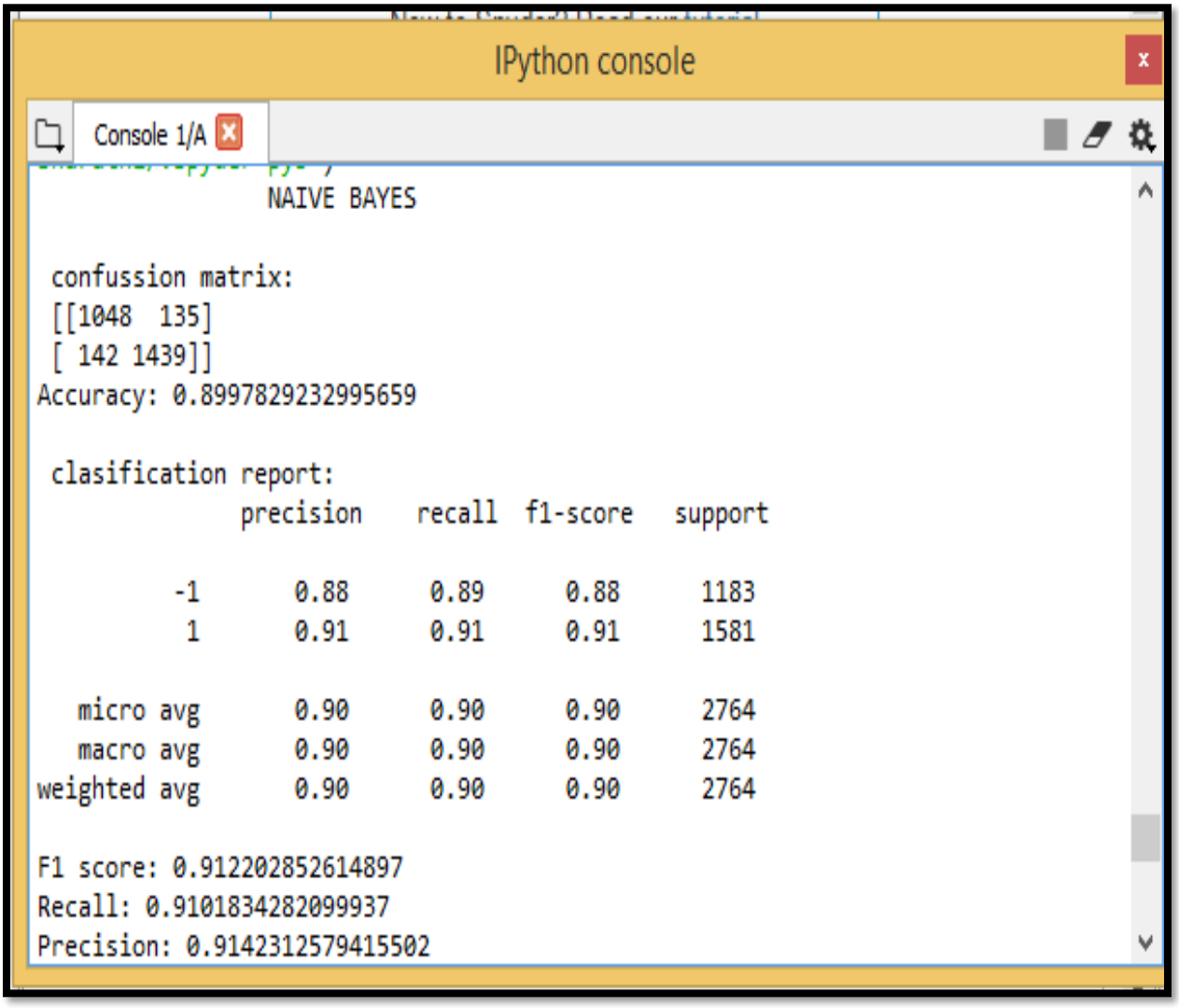


Figure: 9.7

K Fold Cross Validation:

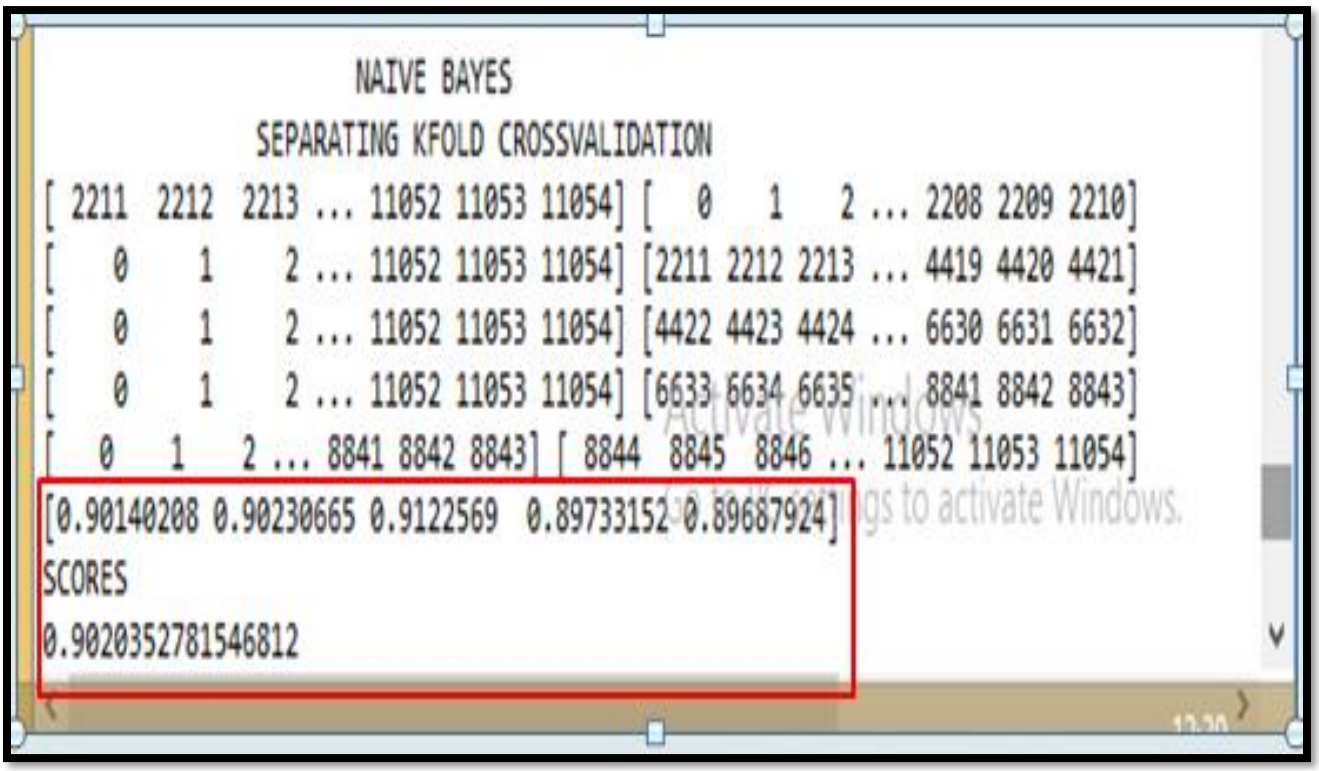


Figure: 9.8

The K Fold Cross Validation was separated as 5 split . The accuracy of the 5 split was found. The mean accuracy result is 0.9020

Visualization of Confusion Matrix:

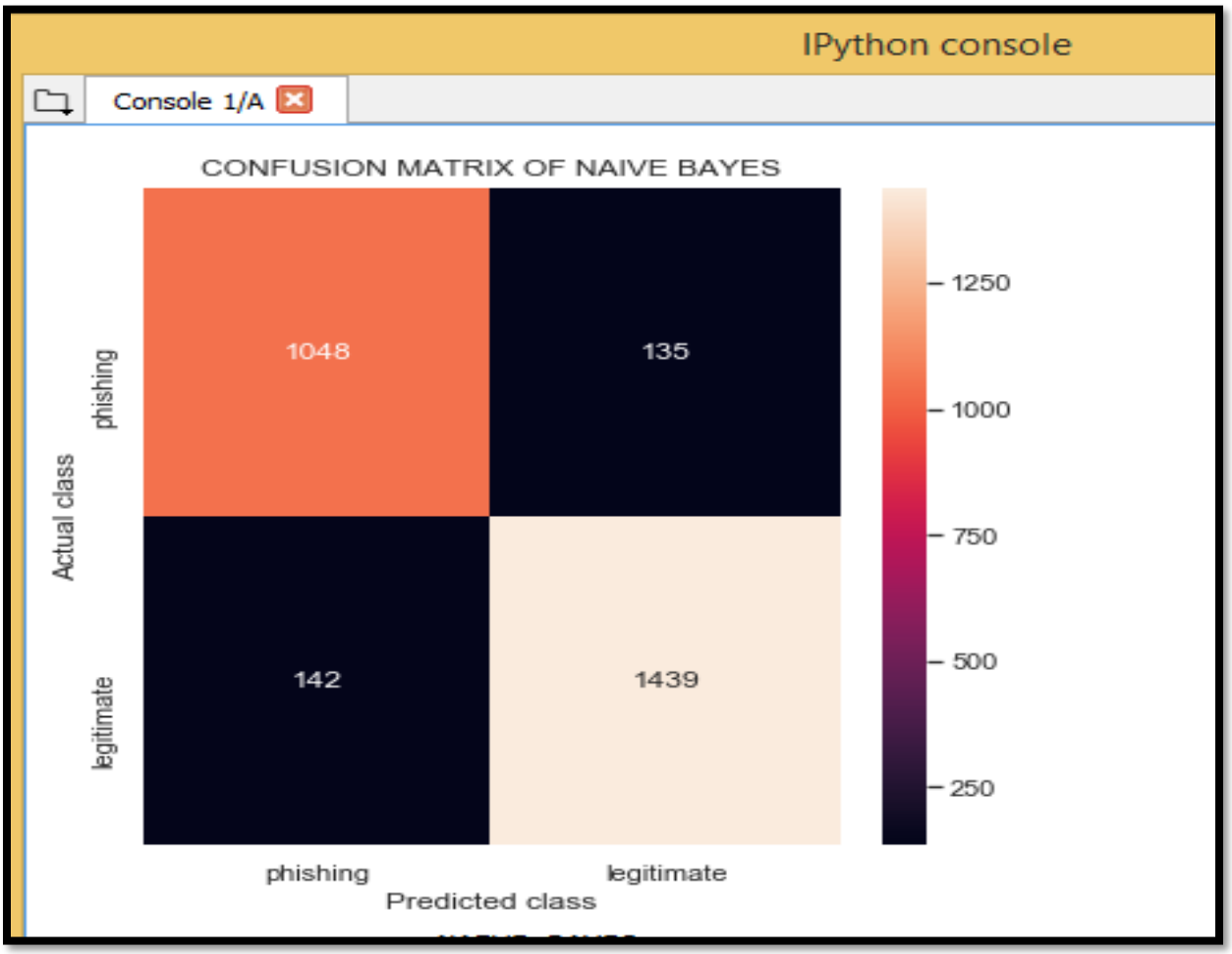


Figure: 9.9

1048 are correctly classified as phishing websites.
1439 are correctly classified as legitimate websites.
142 are wrongly classified as legitimate website.
135 are wrongly classified as phishing websites when they are legitimate.

- Out of those 2764 cases,
- The classifier predicted 1574 as legitimate websites, and phishing websites are 1190
- In actual classifier, 1581 websites in the sample are legitimate, and 1183 are phishing websites.

RANDOM FOREST PERFORMANCE METRICS

Train Test Split:

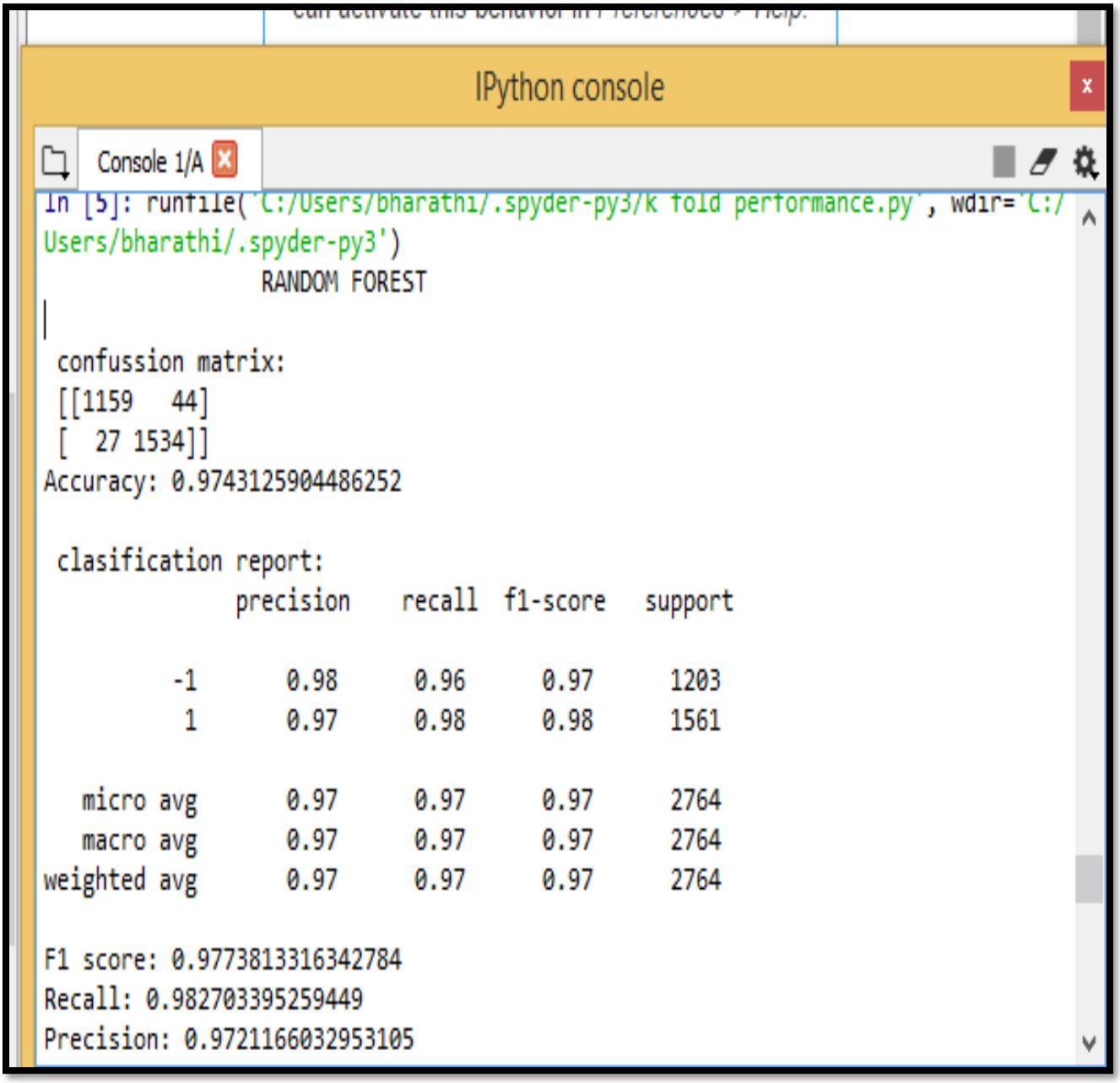


Figure: 9.10

K Fold Cross validation:

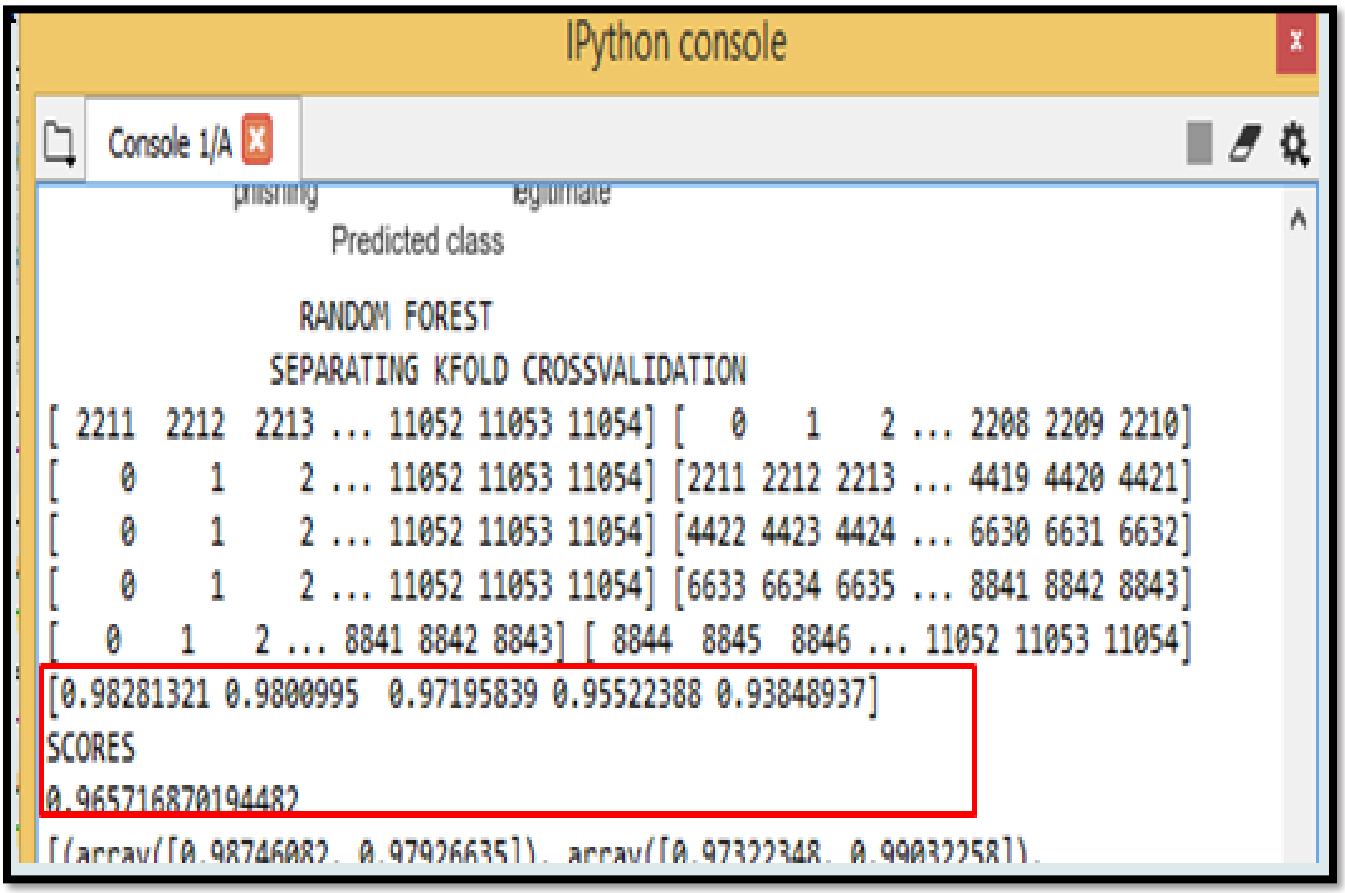


Figure: 9.11

The K Fold Cross Validation was separated as 5 split . The accuracy of the 5 split was found. The mean accuracy result is 0.9657.

Visualization of Confusion Matrix:

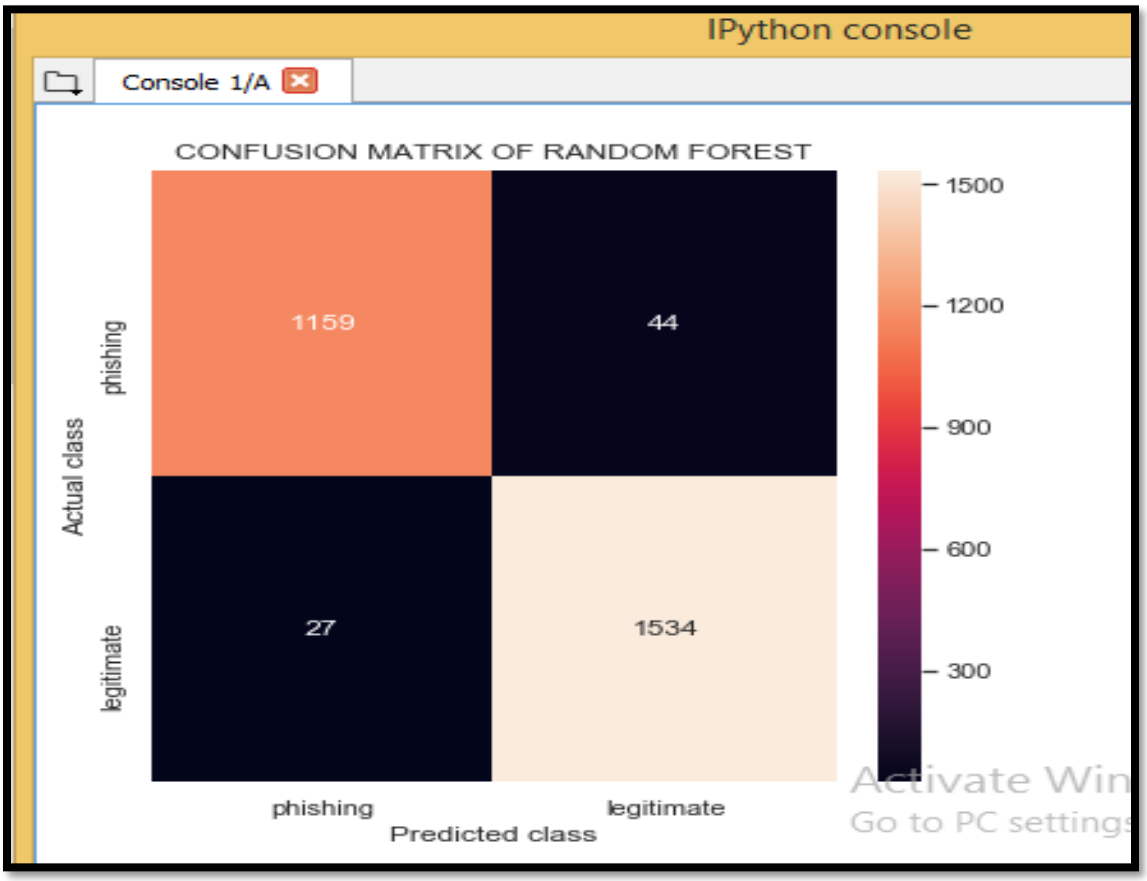


Figure: 9.12

1159 are correctly classified as phishing websites.
1534 are correctly classified as legitimate websites.
27 are wrongly classified as legitimate website when they are phishing.
44 are wrongly classified as phishing websites when they are legitimate.

- Out of those 2764 cases,
- The classifier predicted 1578 as legitimate websites, and phishing websites are 1186
- In actual classifier, 1561 websites in the sample are legitimate, and 1203 are phishing websites.

SUPPORT VECTOR MACHINE PERFORMANCE METRICS

Train Test Split:

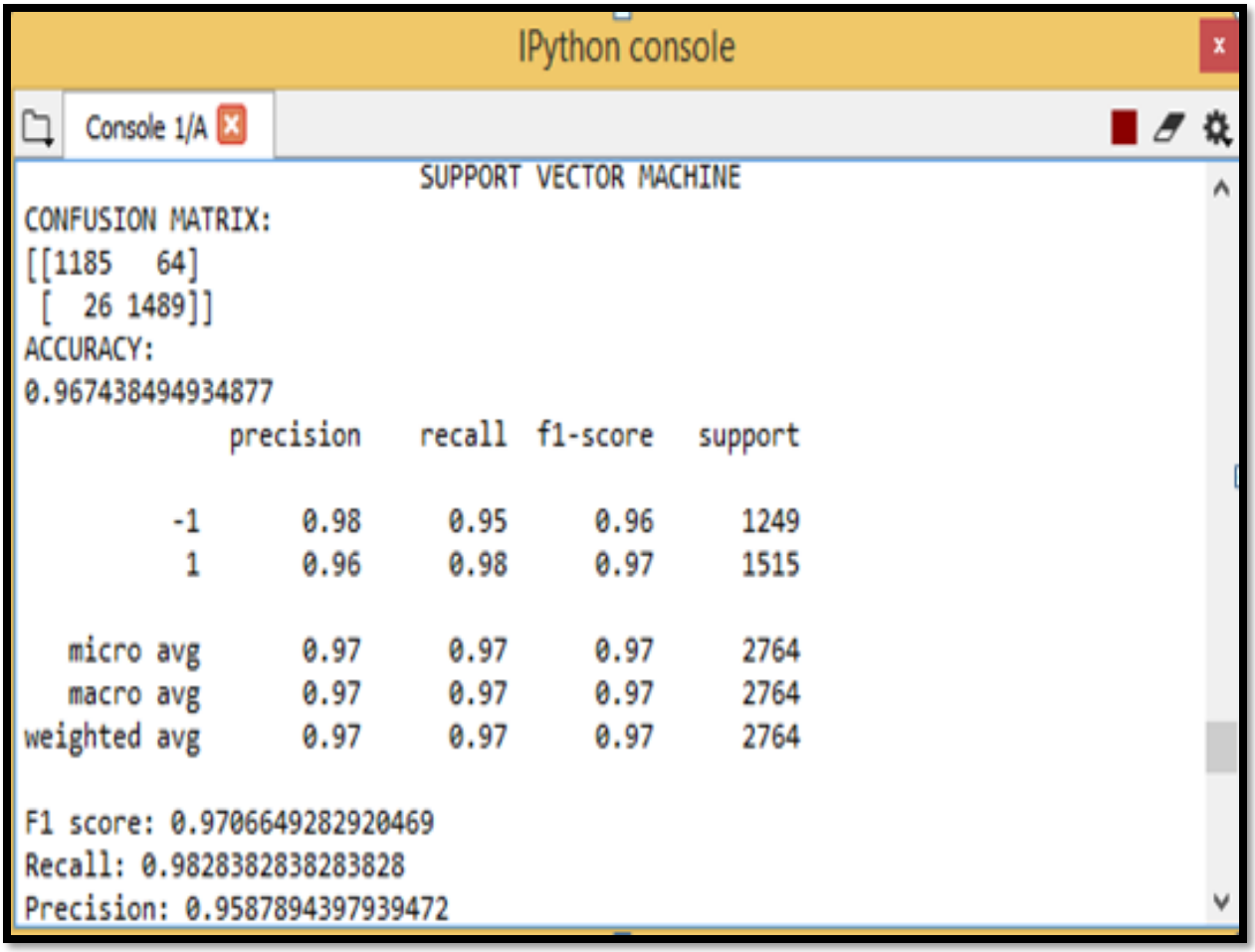


Figure: 9.13

K Fold Cross Validation:

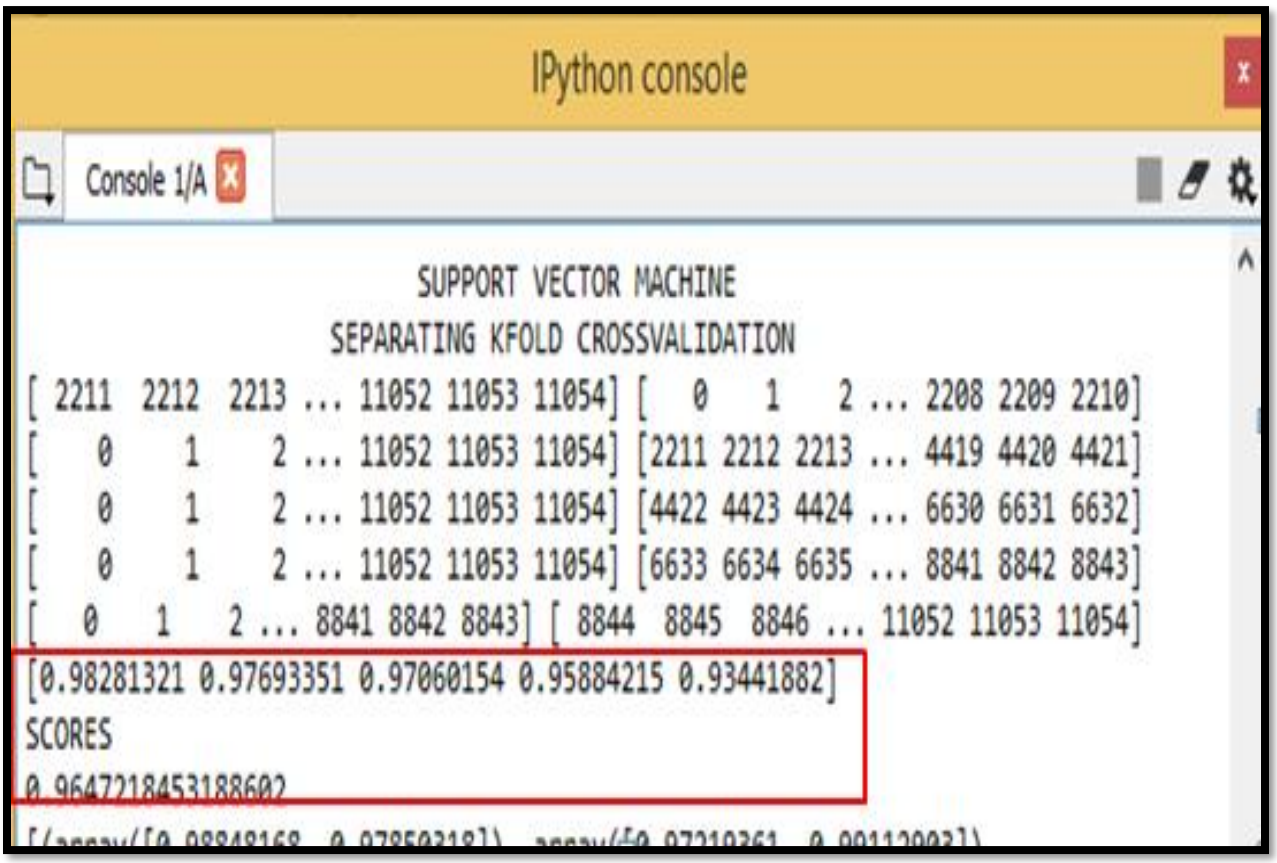


Figure: 9.14

The K Fold Cross Validation was separated as 5 split . The accuracy of the 5 split was found. The mean accuracy result is 0.9647.

Visualization of Confusion Matrix:

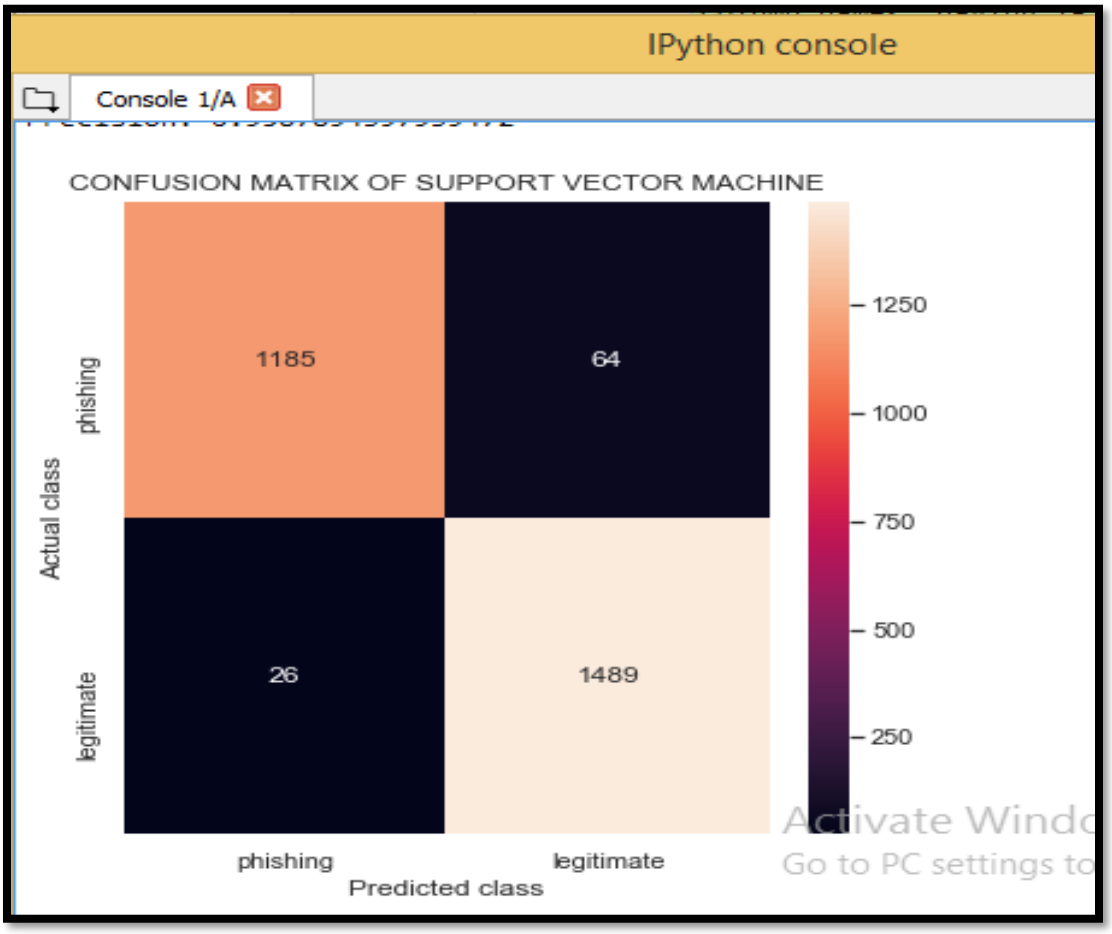


Figure: 9.15

1185 are correctly classified as phishing websites.
1489 are correctly classified as legitimate websites.
26 are wrongly classified as legitimate website when they are phishing.
64 are wrongly classified as phishing websites when they are legitimate.

- Out of those 2764 cases,
- The classifier predicted 1553 as legitimate websites, and phishing websites are 1211
- In actual classifier, 1515 websites in the sample are legitimate, and 1249 are phishing websites.

9.3 SAMPLE CODING

```
import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

import seaborn as sns; sns.set()

import warnings

warnings.simplefilter("ignore")


#importing the dataset

dataset = pd.read_csv("dataset.csv")

dataset = dataset.drop('index', 1) #removing unwanted column

x = dataset.iloc[: , :-1].values

y = dataset.iloc[:, -1:].values


from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x,y,test_size = 0.25, random_state =10 )


from sklearn.ensemble import RandomForestClassifier


#Create a Gaussian Classifier

clf=RandomForestClassifier(n_estimators=700,criterion="entropy",max_features='sqrt',random_state=10)


#Train the model using the training sets y_pred=clf.predict(X_test)

clf.fit(x_train,y_train)
```

```
# prediction on test set

y_pred=clf.predict(x_test)

from sklearn.metrics import
confusion_matrix,classification_report,accuracy_score,precision_score,recall_score,f1_score


print('                RANDOM FOREST')

print ('\n confussion matrix:\n',confusion_matrix(y_test, y_pred))

print ('Accuracy:', accuracy_score(y_test,y_pred))

print ('\n clasification report:\n', classification_report(y_test,y_pred))

print ('F1 score:', f1_score(y_test, y_pred))

print ('Recall:', recall_score(y_test, y_pred))

print ('Precision:', precision_score(y_test,y_pred))


from sklearn.metrics import confusion_matrix

mat = confusion_matrix(y_pred,y_test)

sns.heatmap(mat.T, square=False, annot=True, fmt='d', cbar=False)

plt.xlabel('true label')

plt.ylabel('predicted label')


from sklearn.metrics import roc_auc_score

roc=roc_auc_score(y_test, y_pred)


from sklearn.model_selection import KFold,cross_val_score

from sklearn.metrics import precision_recall_fscore_support

from sklearn.model_selection import GridSearchCV,cross_val_score
```

```
from sklearn.naive_bayes import BernoulliNB
print('                                RANDOM FOREST')
print('              SEPARATING KFOLD CROSSVALIDATION')


kf = KFold(n_splits=5)
for train, test in kf.split(x,y):
    print("%s %s" % (train, test))


model=RandomForestClassifier()
scores = cross_val_score(model, x, y, cv=kf)
print(scores,'                          SCORES')
scores
avg_score = np.mean(scores,axis=0)
print(avg_score)
kf = KFold(n_splits=5)
rf=RandomForestClassifier()
score_array =[]
accuracy=[]
for train_index, test_index in kf.split(x,y)
    x_train, x_test = x[train_index], x[test_index]
    y_train, y_test = y[train_index], y[test_index]
    krf=rf.fit(x_train,y_train)
    y_pred = krf.predict(x_test)
    score_array.append(precision_recall_fscore_support(y_test, y_pred, average=None))
 print(score_array)
```