Machine learning cybersecurity

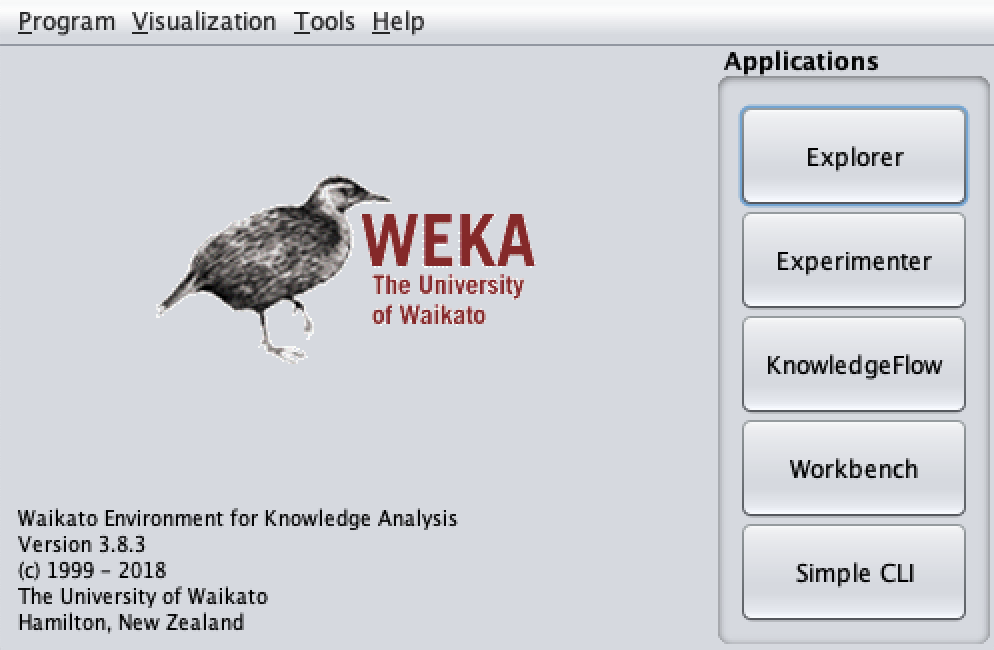
**Phishing detection**

# Lesson 2 lab Solutions

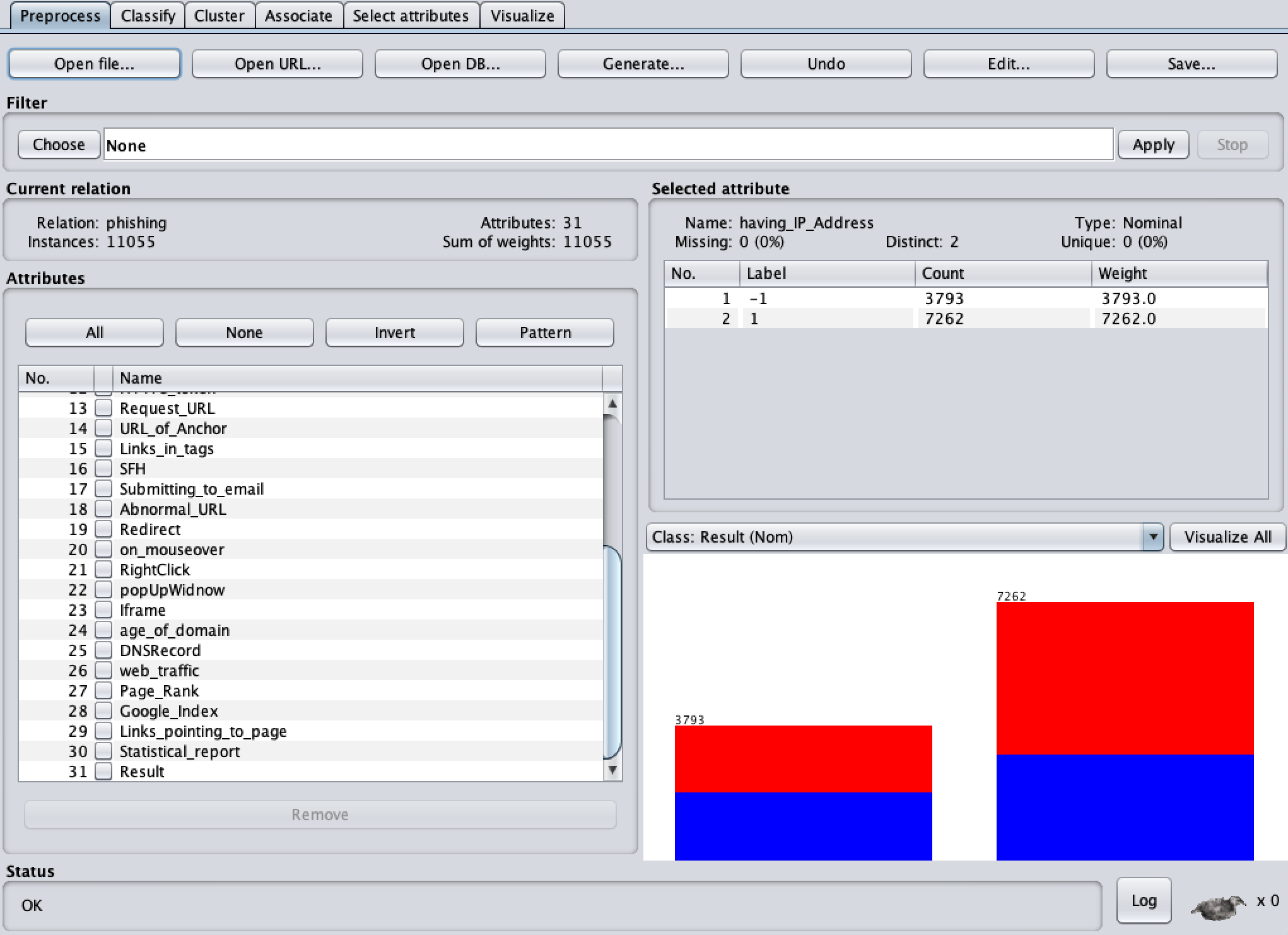
### **Lab Exercise 1**

In this exercise, WEKA will be used for the classification.

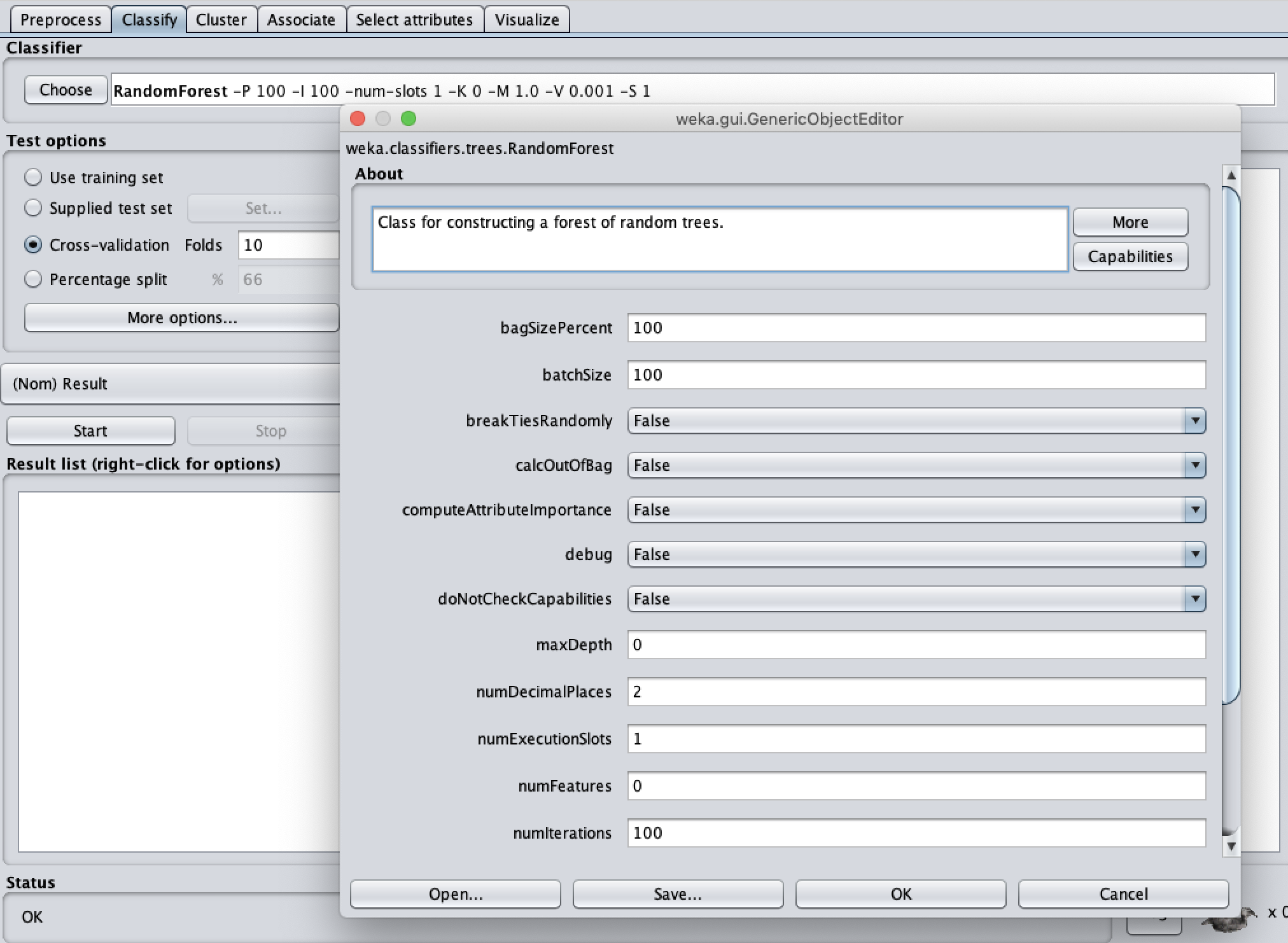
1. Open WEKA and select the Explorer in the applications section



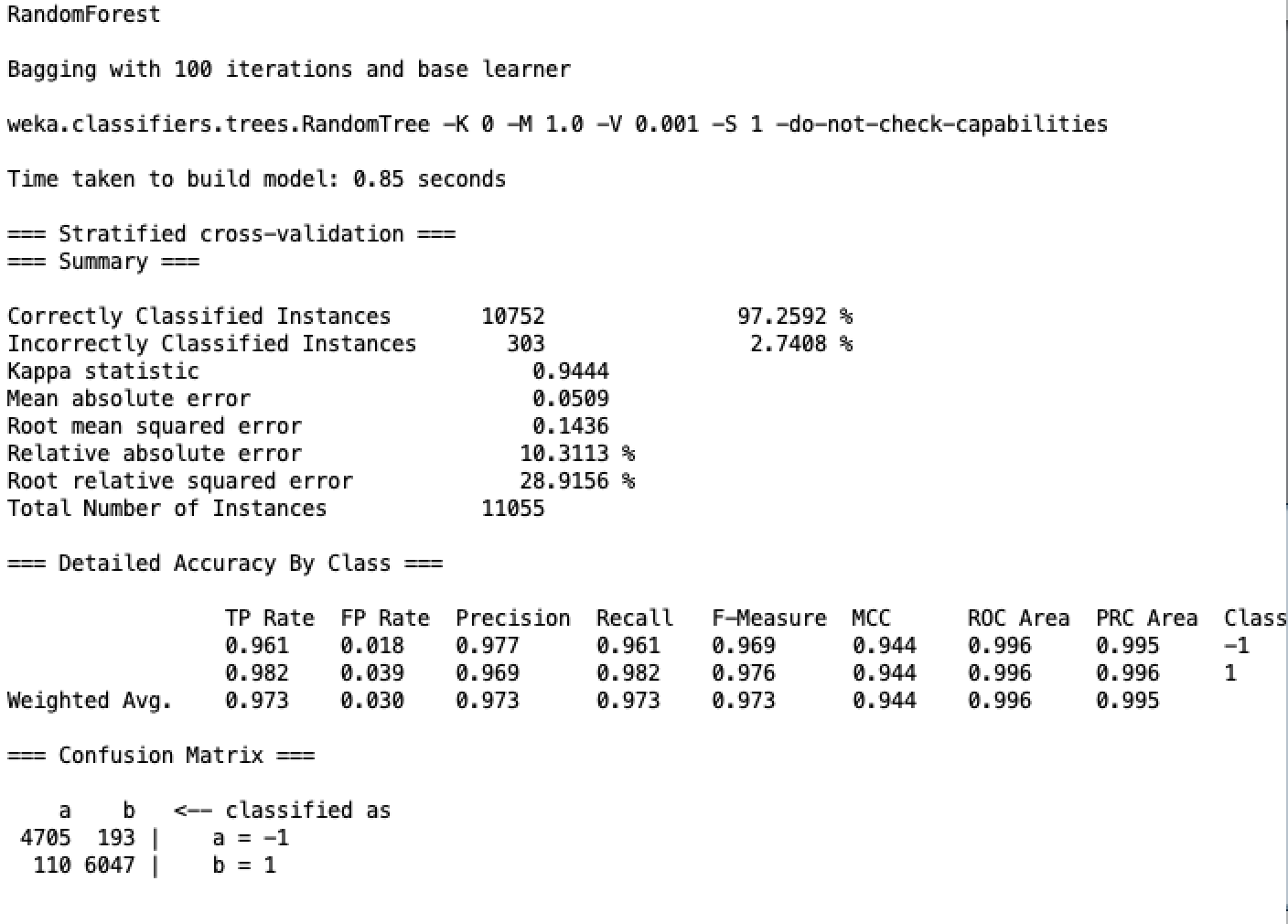
1. Open file and read the data. There are 11,055 records have been read from the phishing dataset.



1. In classify section, select RandomForest as the classifier and set the configuration



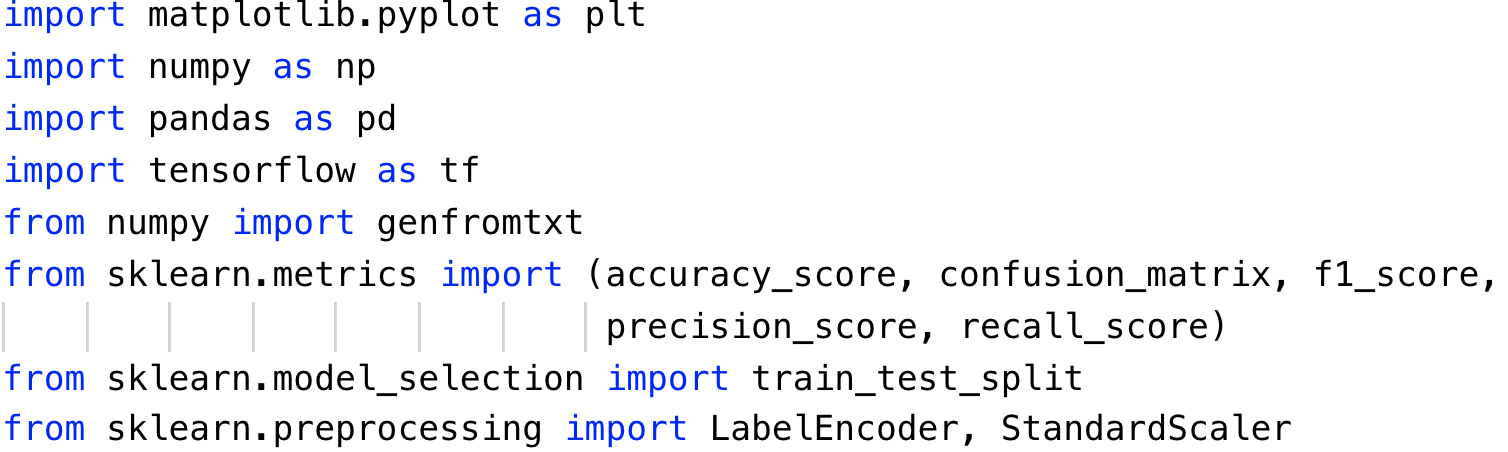
1. For the test option, 10-folds cross-validation has been chosen. Click the start button to perform the classification



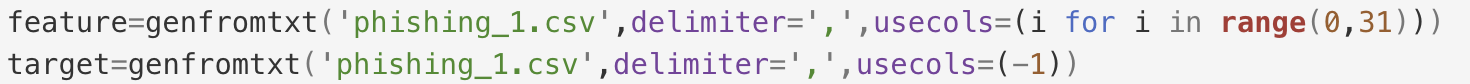
### **Lab Exercise 2**

The task of lab is to estimate whether a website is a phishing website. The dataset contains 102816 web hits and 30 features were recorded for each of the hit. Also, a class value has been given for each of the record.

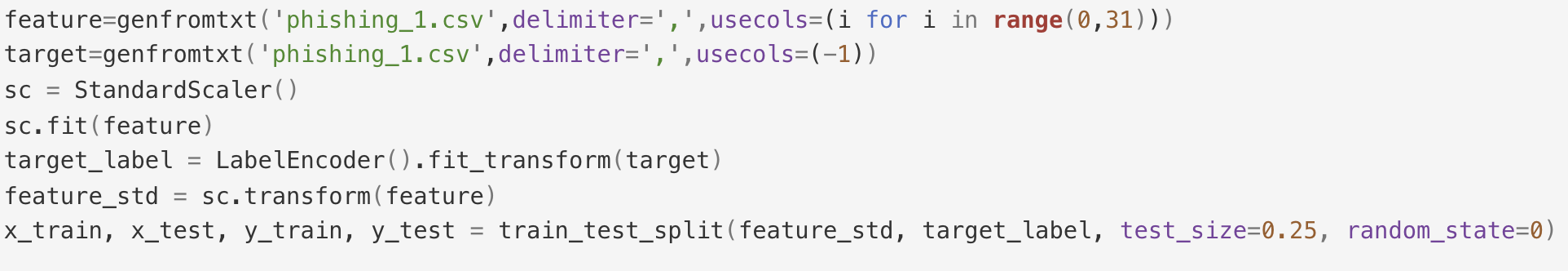
1. Import the required libraries



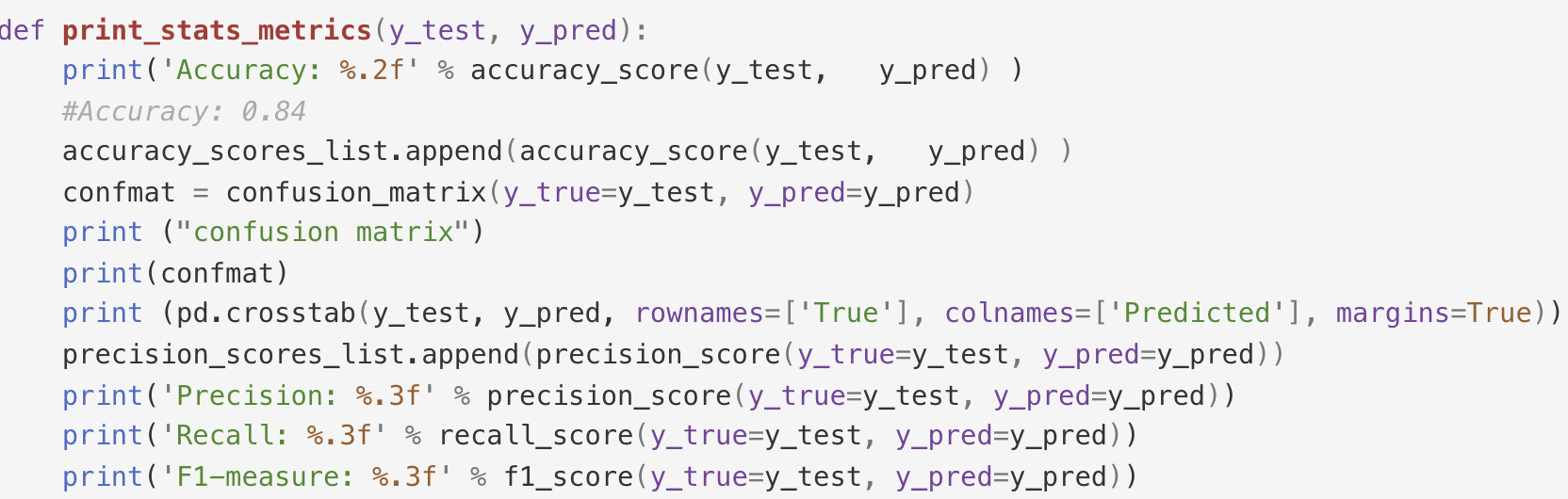
1. Read the features and class values from dataset

****

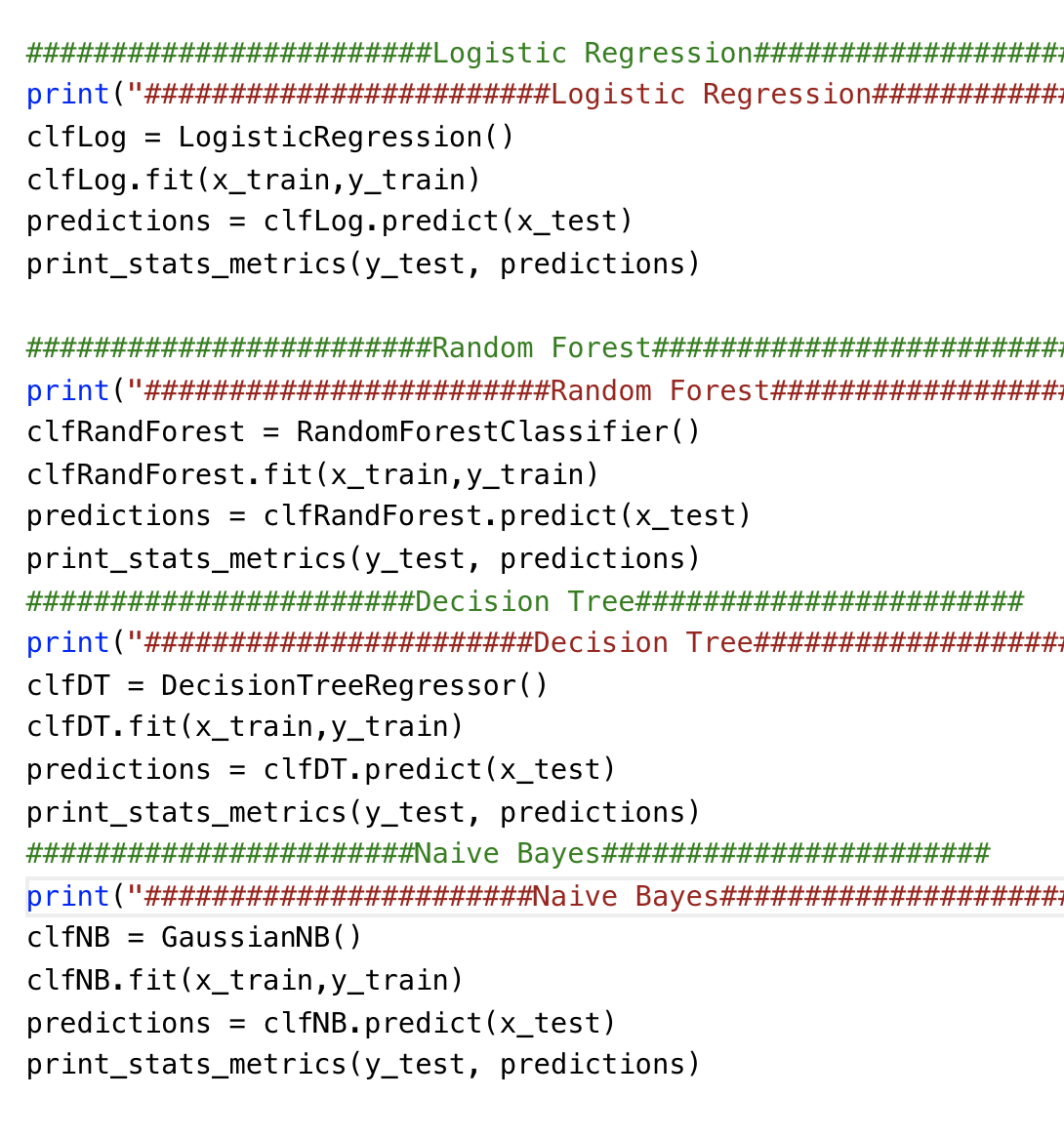
1. Encode the labels for class values, standard scale the feature values and split the



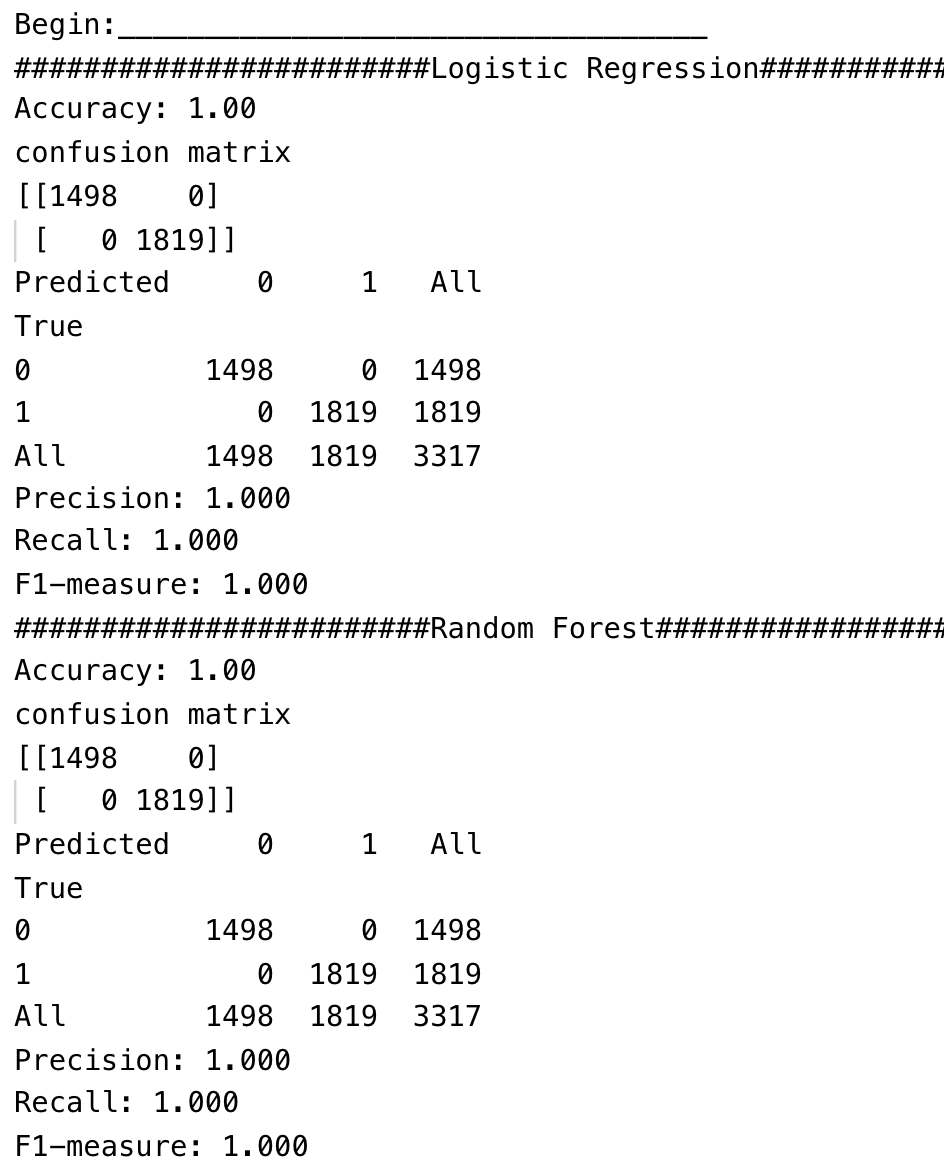
1. Define the function to print statistics metrics

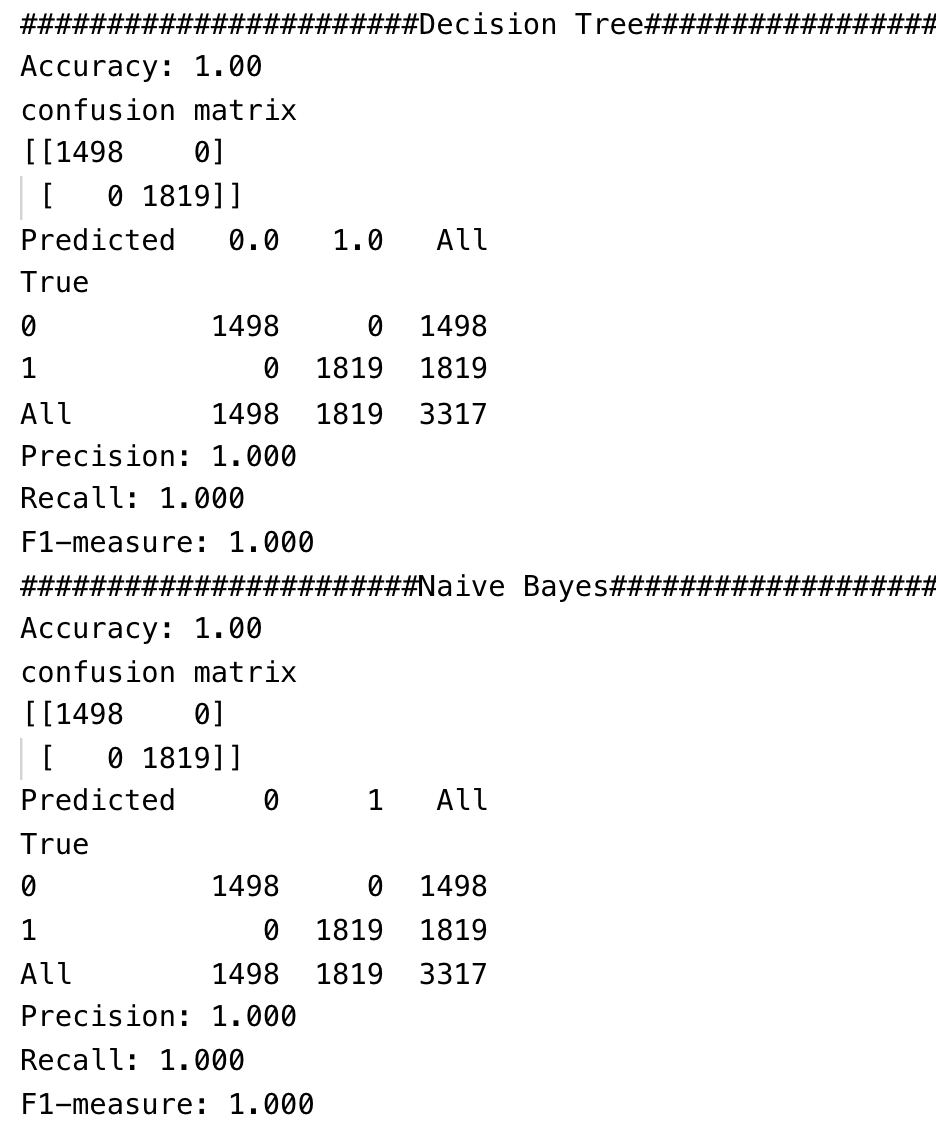
****

1. Specify the type of classifier, then fit it with the training set and perform the prediction with testing set. Finally, print out the statistics.



1. Results



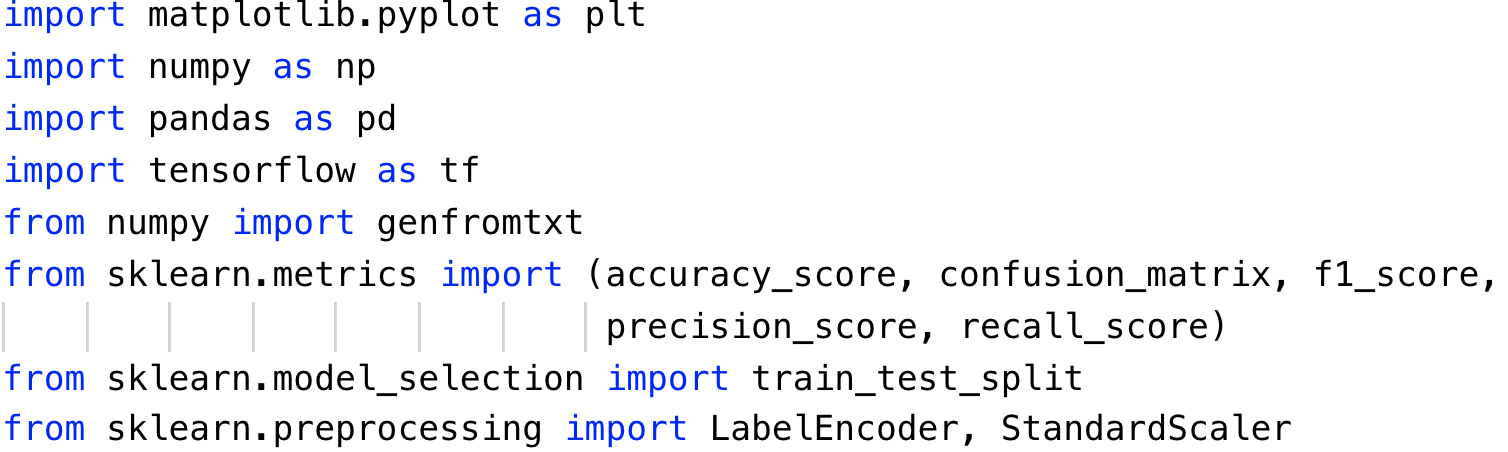


As we can see, the classifier has achieved excellent statistics metrics.

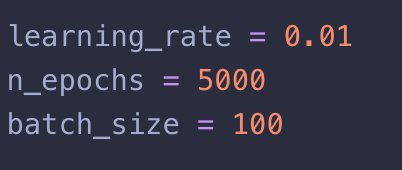
### **Lab Exercise 3**

In Exercise 3, tensorflow is applied for the deep-learning-based classifier.

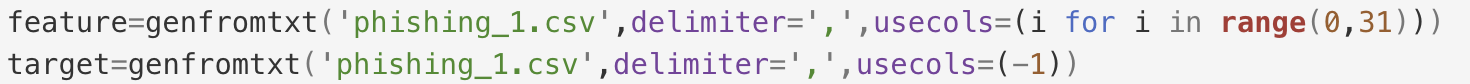
1. Import the required libraries



1. Define the parameters

****

1. Read the features and class values from dataset

****

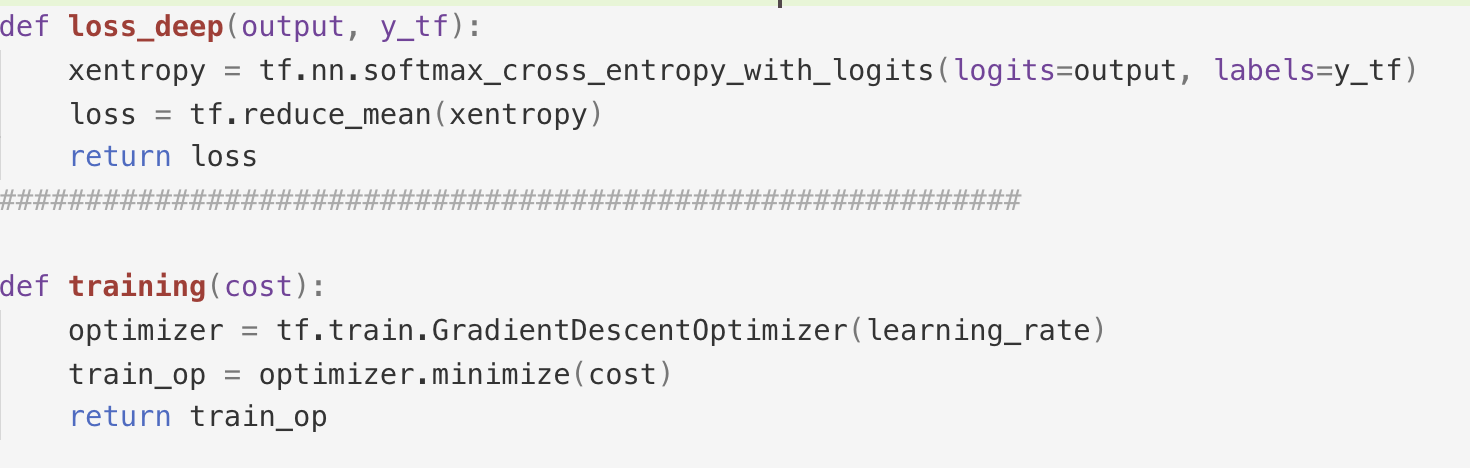
1. Encode the labels for class values, standard scale the feature values and split the dataset as well as the one hot encoding for class values



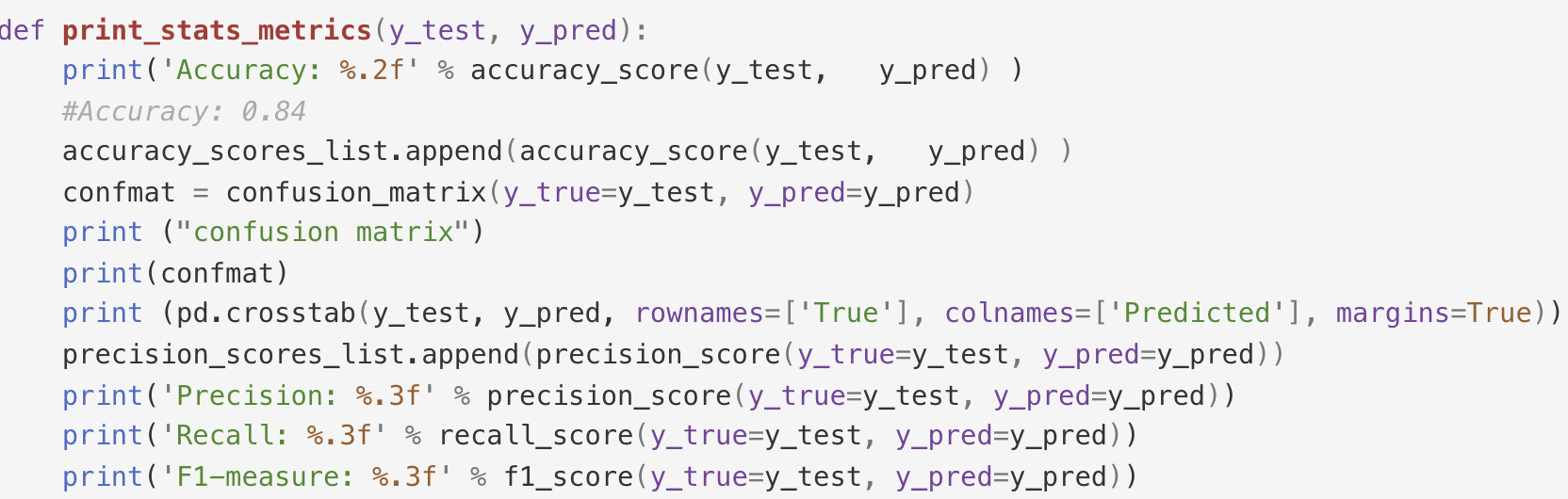
1. Define the architecture for deep neural network

Since the number of input features is 30, so the number of neurons in each layer should no more than 30 or less than 30.

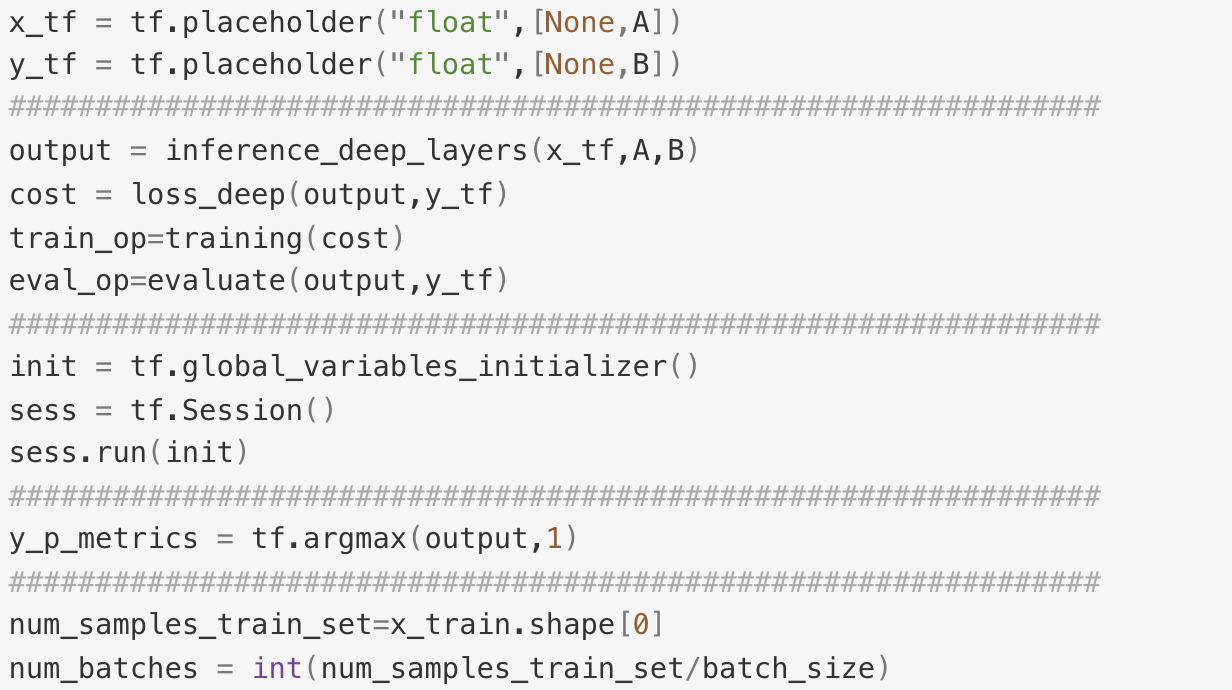
****



1. Define the function to print statistics metrics

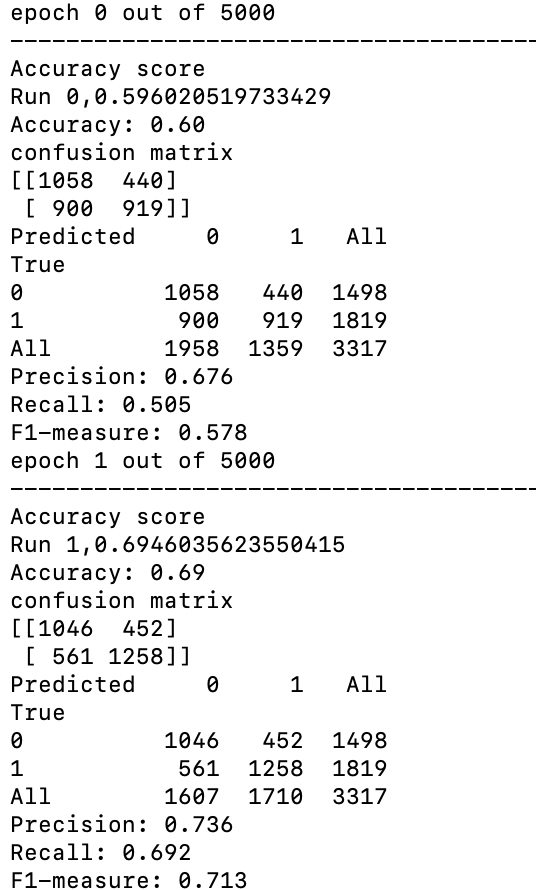
****

1. Define the placeholder and perform the deep learning for phishing dataset

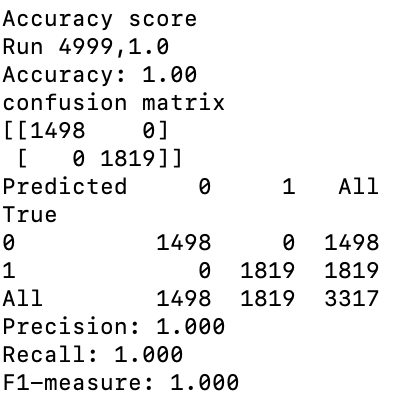


1. Results

The first two epochs:



The last epoch:



As we can see, all the statistics metrics have been improved after 5000 epochs.