كلية الحوسبة والمعلوماتية |
College of Computing and Informatics
الجامعة السعودية الإلكترونية
SAUDI ELECTRONIC UNIVERSITY
2011-1432

Semester 1 – 2021/2022

Course Code	DS550
Course Name	Machine Learning
Assignment type	Critical Thinking
Module	10

Student ID	G200007615
Student Name	Abdulaziz Alqumayzi
CRN	15062

Solutions:

Critical Thinking Assignment 4

Introduction

In this activity, we will define and explain Random Forest algorithm. Provide a python programming example on how to apply the algorithm on a dataset. The dataset used is **Heart Failure Prediction** from Kaggle.

Data Set Information: Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worlwide. Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure. Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies. People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

Attribute information in Table 1.

Table 1

Attribute Information

Column	Data type
age	double
anaemia (decrease of red blood cells or hemoglobin)	boolean
creatinine_phosphokinase (level of the CPK enzyme in the blood (mcg/L))	Integer
diabetes (if the patient has diabetes)	boolean
ejection_fraction (percentage of blood leaving the heart at each contraction)	Integer
high_blood_pressure (If the patient has hypertension)	boolean
platelets (platelets in the blood (kiloplatelets/mL))	double
serum_creatinine (level of serum creatinine in the blood (mg/dL))	double
serum_sodium (level of serum sodium in the blood (mEq/L))	Integer
sex (woman or man)	binary
smoking (if the patient smokes or not)	boolean
time (follow-up period per days)	Integer
DEATH_EVENT (if the patient deceased during the follow-up period)	Integer

Random Forest Algorithm

Random Forest is a well-known machine learning algorithm that uses the supervised learning method. In machine learning, it may be utilized for both classification and regression issues. It is based on ensemble learning, which is a method of integrating several classifiers to solve a complicated issue and increase the model's performance. Random Forest is a classifier that combines a number of decision trees on different subsets of a dataset and averages the results to increase the dataset's predicted accuracy.

Benefits of Random Forest Algorithm

When compared to other algorithms, it takes less time to train. , predicts output with excellent accuracy, and it runs quickly even with a big dataset. It can keep its accuracy even when a major chunk of the data is absent.

Applications of Random Forest Algorithm

Banking: This algorithm is mostly used in the banking industry to identify loan risk. Medicine: This method may be used to identify illness trends as well as disease risks.

Scikit-learn and Random Forest Algorithm

Scikit-learn has two random forest algorithms, Random Forest Classifier and Random Forest Regressor.

The random forest classifier is a meta estimator that employs averaging to increase predicted accuracy and control over-fitting by fitting several decision tree classifiers on different sub-samples of the dataset.

The random forest regressor is a meta estimator that employs averaging to increase predicted accuracy and control over-fitting by fitting a number of classification decision trees on various sub-samples of the dataset.

Python Code

```
# importing needed packages
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix

#importing datasets
df = pd.read_csv('heart_failure_clinical_records_dataset.csv')

# check if null values exists
df.info()

# extracting independent and dependent variables
x= df.iloc[:,:-1].values
y= df.iloc[:,12].values # dependent variable is DEATH_EVENT
```

```
# spilt the dataset into test and train sets. 80% test 20% train
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.8,
random_state=9)

# fitting Random Forest to the dataset
clf = RandomForestClassifier(random_state= 9)
clf = clf.fit(X_train, y_train)

# validate the model

# predict class for X_test
y_pred= clf.predict(X_test)
print(y_pred)

# predict class log-probabilities for X_test
print(clf.predict_log_proba(X_test))

# predict class probabilities for X_test
print(clf.predict_proba(X_test))

# return the mean accuracy on the given test data and labels
print(clf.score(X_test,y_test))

# confusion matrix to determine the correct and incorrect predictions
print(confusion_matrix(y_test, y_pred))
# I got 19+22= 41 incorrect predictions and 144+55= 199 correct predictions
```

References

Bonaccorso, G. (2018). *Machine learning algorithms: Popular algorithms for Data Science and Machine Learning*. Packt

Maranhão, A. (2020, June 20). *Heart failure prediction*. Kaggle. Retrieved November 19, 2021, from https://www.kaggle.com/andrewmvd/heart-failure-clinical-data.

Sklearn.ensemble.randomforestclassifier. Scikit-learn . (2021). Retrieved November 19, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier.

Sklearn.ensemble.randomforestregressor. scikit-learn. (2021). Retrieved November 19, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html#sklearn.ensemble.RandomForestRegressor.

Machine learning random forest algorithm - javatpoint. www.javatpoint.com. (2021). Retrieved November 19, 2021, from https://www.javatpoint.com/machine-learning-random-forest-algorithm.