Big Data

****

|  |  |
| --- | --- |
| Study Program: | **BSSE/BSCS** |
| **Semester:** | **Spring 2022** |
| **Course Code/Title:** | **Big Data Programming** |
| **Resource Person:** | **Mr. Mazhar Javed**  **Mazhar.awan@umt.edu.pk** |
| **Weight age** | **20 %** |
| **Due Date** | **13th July 2022** |
| **Student Name** | **Muhammad Adnan Asad** |
| **Roll No** | **F2018065185** |

Graphical user interface

Description automatically generated with medium confidence

Command used to install pyspark

Background pattern

Description automatically generated

Importing pyspark dependency and checking version which is 3.3.0

Text

Description automatically generated

Importing more dependencies, then we create an app named ml-diabetes and in the third line loading the dataset from google drive. At the end we print the dataframe.

A screenshot of a computer

Description automatically generated with medium confidence

This command imports pandas as pd and with the help of pandas it shows the first 5 records of the dataset.

Graphical user interface, application

Description automatically generated

Data is grouped by outcome, we can see that 268 have positive outcome while 500 have negative outcome.

A screenshot of a computer

Description automatically generated with medium confidence

Another function to simply display the dataset by the help of topandas() function.

A screenshot of a computer

Description automatically generated with medium confidence

We can calculate the count, mean, stand deviation, min and max for all the numbers in the dataset. The command to achieve this function is called transpose()

Text

Description automatically generated

Shows all the columns

Text

Description automatically generated with low confidence

Plotting a scatter matrix for the dataset.

Text

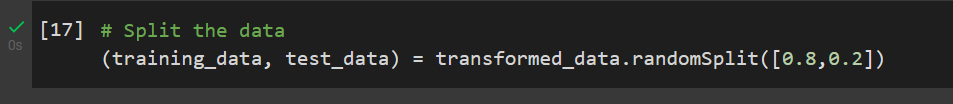
Description automatically generated

This checks for any and all missing rows

Text

Description automatically generated

This code will remove all unnecessary columns of data and show the remaining data.



This code will split the data into training and testing data with a 80% and 20% ratio respectivally.

Text

Description automatically generated

These are the models, the first one is the random forest classifier, in the first part the model is defined and then fitted with the training data. In the third part we test the model and at the end show the results and check its accuracy. The model shows a 76% accuracy.

Text

Description automatically generated

This is a Decision Tree Classifier model, here we fit it and test it in the first part. We then show our predictions and in the second part we check those predictions to reveal a 77% accuracy.

A screenshot of a computer

Description automatically generated with medium confidence

This one is a Logistic Regression Model, same as before we import dependencies, define model then fit it with the training data, then test it with testing data and alas compare predictions with the real data to check accuracy and we get 77%.

A screenshot of a computer

Description automatically generated with medium confidence

This model is a gradient boosted tree classifier model, here we import dependencies, define it, train it test it and check our predictions revealing 80% accuracy.