**Digital Assessment 3**

**CBS3007 -** **Data Mining and Analytics**

Date: 6 October, 2024

Name: Anuj Parihar

Registration Number: 21BBS0162

**Link to Assessment Codebase and Dataset**:

<https://github.com/BearTS/data-mining-assignments/tree/main/Lab/DA%203>

**Question 1**

Collect the student attendance and performance dataset of your classroom to

identify students who are likely to drop out or fail early. Implement the KNN to

classify the above cases and display the list of students and their classes as per

classifications.

**Aim:** To classify the likelihood of a student dropping out using the KNN model and

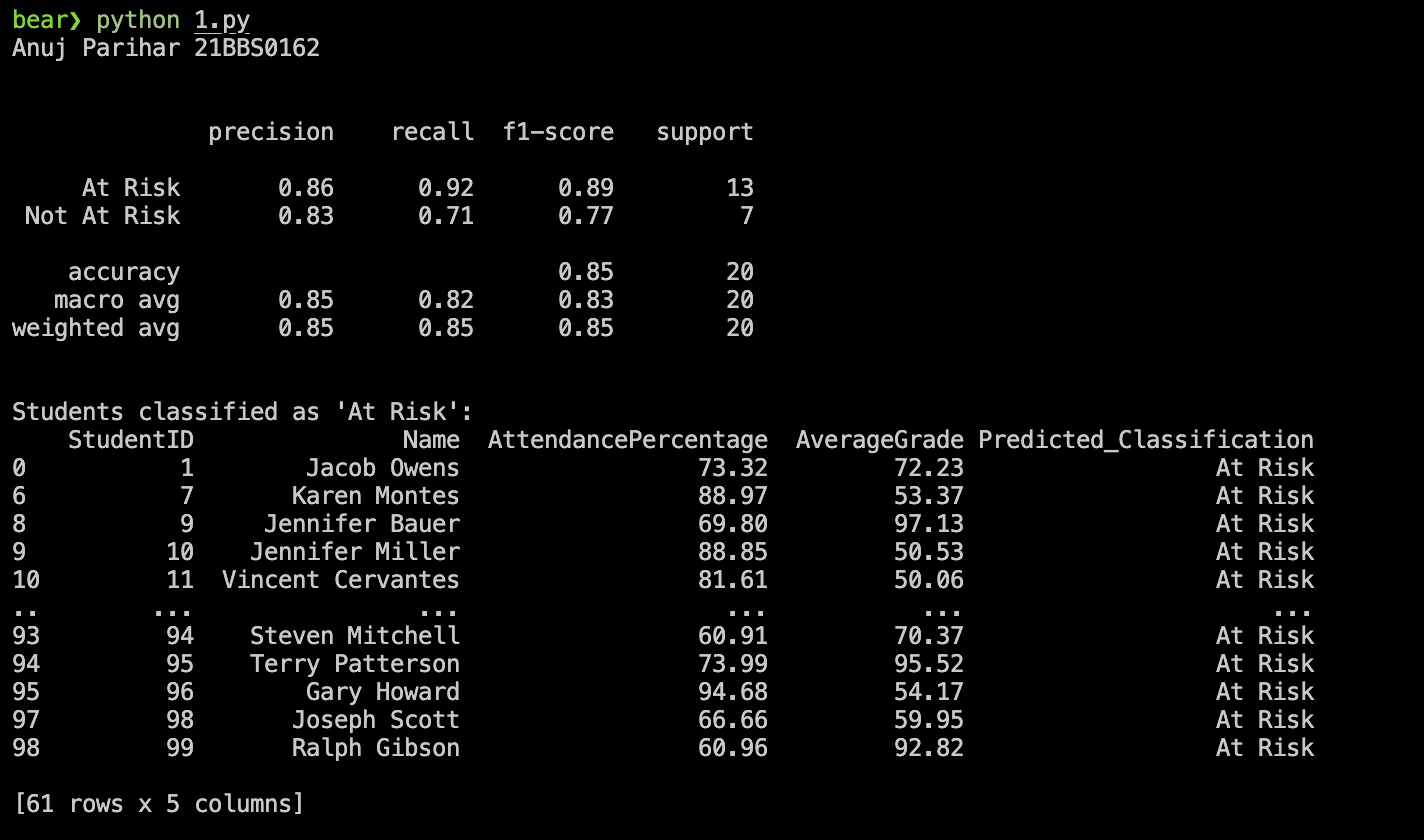
several criterions like attendance and performance in various exams,

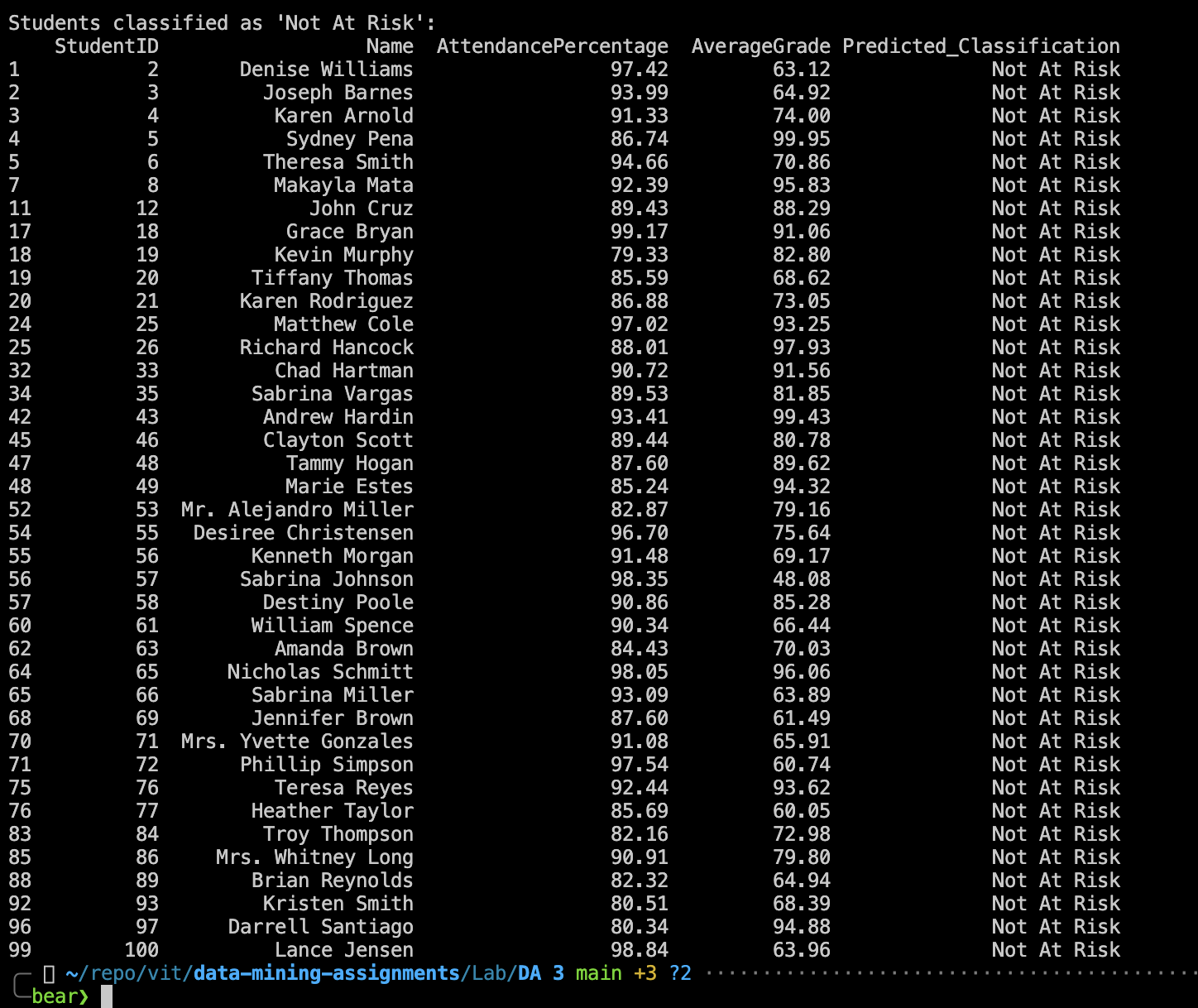
assignments and so on.

**Sample Input:** The entire input dataset is in the GitHub repository

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| StudentID | Name | AttendancePercentage | AverageGrade | ExtracurricularActivities | StudyHoursPerWeek |
| 1 | Jacob Owens | 73.32 | 72.23 | 1 | 16 |
| 2 | Denise Williams | 97.42 | 63.12 | 2 | 18 |
| 3 | Joseph Barnes | 93.99 | 64.92 | 1 | 1 |
| 4 | Karen Arnold | 91.33 | 74 | 2 | 3 |
| 5 | Sydney Pena | 86.74 | 99.95 | 4 | 0 |
| 6 | Theresa Smith | 94.66 | 70.86 | 1 | 31 |
| 7 | Karen Montes | 88.97 | 53.37 | 3 | 26 |
| 8 | Makayla Mata | 92.39 | 95.83 | 5 | 10 |
| 9 | Jennifer Bauer | 69.8 | 97.13 | 2 | 24 |
| 10 | Jennifer Miller | 88.85 | 50.53 | 4 | 23 |

**Output:**

****

****

**Results:**

The KNN to assess the likelihood of dropout of students has been implemented

successfully

**Question 2:**

Linear regression of 2 variables is to use one variable to forecast another variable value. Collect the DEMAT account counts of Indians for the past 60 months. Implement the Linear regression Technique to predict what will be count in JAN2025 in future. Collect the real time sample data from news sources to perform the algorithm

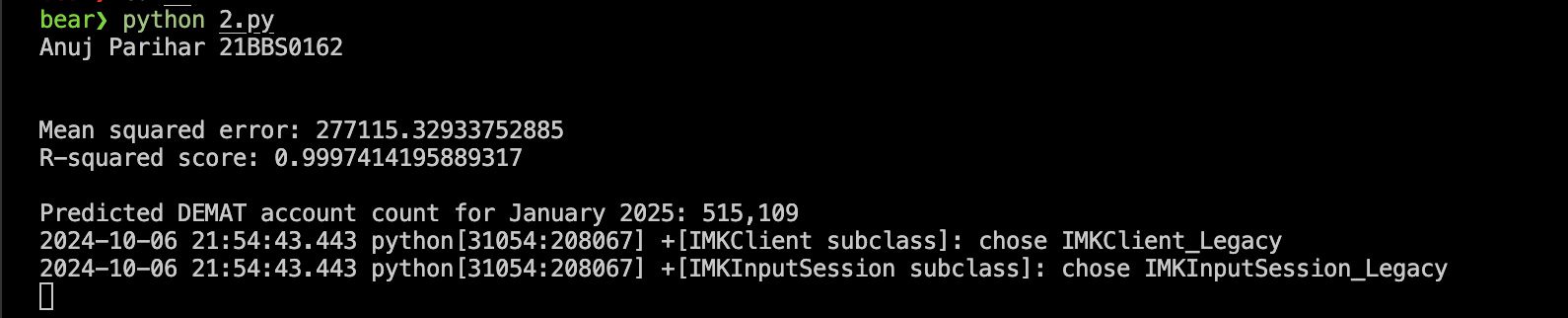
**Aim:** To predict the count of number of DEMAT accounts with the data for the past

60 months using linear regression of 2 variables

**Sample Input:** The entire input is in the GitHub Repository

|  |  |
| --- | --- |
| Month | DEMAT\_Count |
| 2019-01 | 400810 |
| 2019-02 | 402120 |
| 2019-03 | 403450 |
| 2019-04 | 404300 |
| 2019-05 | 405700 |
| 2019-06 | 407890 |
| 2019-07 | 409250 |
| 2019-08 | 410500 |
| 2019-09 | 412000 |

**Output:**

****

**A graph on a computer screen

Description automatically generated**

**Results:**

Linear Regression model to find the number of DEMAT accounts have been

implemented successfully

**Question 3**

Implement the Random Forest Supervised Machine Learning Algorithm that is

used widely in multi-Classifications in Fruits dataset. (Assume own dataset at

least 50 entries).

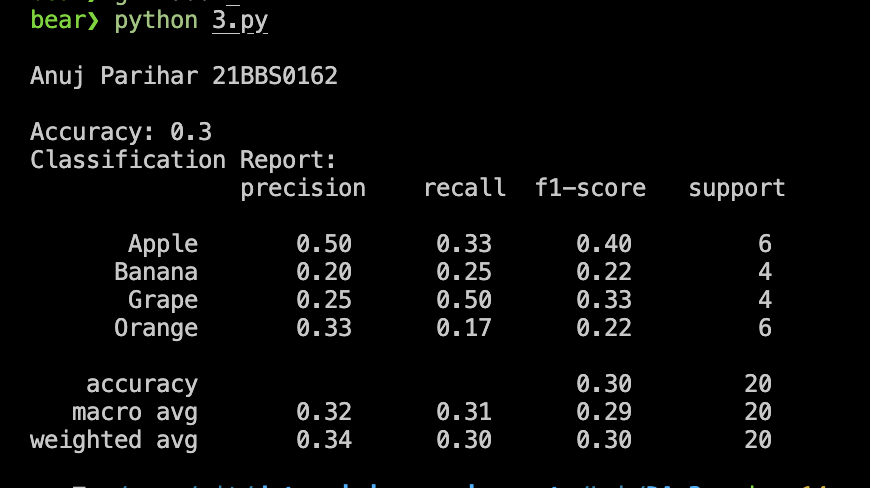
**Aim:** To classify the fruits using the Random Forest model and several criterions like

Colour, shape, volume, weight, density and so on

**Sample Input:** The input for the following question is in the GitHub Repository

|  |  |  |  |
| --- | --- | --- | --- |
| Weight | Color\_Score | Sugar\_Content | Fruit\_Type |
| 182 | 0.67 | 8 | Banana |
| 131 | 0.9 | 9 | Apple |
| 172 | 0.31 | 6 | Grape |
| 94 | 0.39 | 9 | Orange |
| 186 | 0.28 | 10 | Grape |
| 151 | 0.48 | 3 | Apple |
| 140 | 0.52 | 8 | Apple |
| 100 | 0.44 | 10 | Grape |
| 182 | 0.83 | 7 | Apple |

**Output:**

****

**Results:** The Random Forest Model to classify fruits has been implemented successfully