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

Student ID	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:

```
bear❯ python 1.py
Anuj Parihar 21BBS0162


              precision    recall  f1-score   support

     At Risk       0.86      0.92      0.89        13
 Not At Risk       0.83      0.71      0.77         7

    accuracy                           0.85        20
   macro avg       0.85      0.82      0.83        20
weighted avg       0.85      0.85      0.85        20


Students classified as 'At Risk':
    StudentID              Name  AttendancePercentage  AverageGrade Predicted_Classification
0           1       Jacob Owens                 73.32         72.23                  At Risk
6           7      Karen Montes                 88.97         53.37                  At Risk
8           9     Jennifer Bauer                69.80         97.13                  At Risk
9          10    Jennifer Miller                88.85         50.53                  At Risk
10         11  Vincent Cervantes                81.61         50.06                  At Risk
..        ...               ...                  ...           ...                     ...
93         94    Steven Mitchell                60.91         70.37                  At Risk
94         95    Terry Patterson                73.99         95.52                  At Risk
95         96        Gary Howard                94.68         54.17                  At Risk
97         98       Joseph Scott                66.66         59.95                  At Risk
98         99       Ralph Gibson                60.96         92.82                  At Risk

[61 rows x 5 columns]

Students classified as 'Not At Risk':
    StudentID                 Name  AttendancePercentage  AverageGrade Predicted_Classification
1           2      Denise Williams                 97.42         63.12              Not At Risk
2           3        Joseph Barnes                 93.99         64.92              Not At Risk
3           4         Karen Arnold                 91.33         74.00              Not At Risk
4           5          Sydney Pena                 86.74         99.95              Not At Risk
5           6        Theresa Smith                 94.66         70.86              Not At Risk
7           8         Makayla Mata                 92.39         95.83              Not At Risk
11         12            John Cruz                 89.43         88.29              Not At Risk
17         18          Grace Bryan                 99.17         91.06              Not At Risk
18         19         Kevin Murphy                 79.33         82.80              Not At Risk
19         20       Tiffany Thomas                 85.59         68.62              Not At Risk
20         21      Karen Rodriguez                 86.88         73.05              Not At Risk
24         25         Matthew Cole                 97.02         93.25              Not At Risk
25         26      Richard Hancock                 88.01         97.93              Not At Risk
32         33         Chad Hartman                 90.72         91.56              Not At Risk
34         35       Sabrina Vargas                 89.53         81.85              Not At Risk
42         43        Andrew Hardin                 93.41         99.43              Not At Risk
45         46        Clayton Scott                 89.44         80.78              Not At Risk
47         48          Tammy Hogan                 87.60         89.62              Not At Risk
48         49         Marie Estes                  85.24         94.32              Not At Risk
52         53  Mr. Alejandro Miller                82.87         79.16              Not At Risk
54         55   Desiree Christensen                96.70         75.64              Not At Risk
55         56       Kenneth Morgan                 91.48         69.17              Not At Risk
56         57      Sabrina Johnson                 98.35         48.08              Not At Risk
57         58        Destiny Poole                 90.86         85.28              Not At Risk
60         61       William Spence                 90.34         66.44              Not At Risk
62         63         Amanda Brown                 84.43         70.03              Not At Risk
64         65     Nicholas Schmitt                 98.05         96.06              Not At Risk
65         66       Sabrina Miller                 93.09         63.89              Not At Risk
68         69        Jennifer Brown                87.60         61.49              Not At Risk
70         71  Mrs. Yvette Gonzales                91.08         65.91              Not At Risk
71         72      Phillip Simpson                 97.54         60.74              Not At Risk
75         76         Teresa Reyes                 92.44         93.62              Not At Risk
76         77       Heather Taylor                 85.69         60.05              Not At Risk
83         84        Troy Thompson                 82.16         72.98              Not At Risk
85         86    Mrs. Whitney Long                 90.91         79.80              Not At Risk
88         89       Brian Reynolds                 82.32         64.94              Not At Risk
92         93        Kristen Smith                 80.51         68.39              Not At Risk
96         97      Darrell Santiago                80.34         94.88              Not At Risk
99        100         Lance Jensen                 98.84         63.96              Not At Risk
┌─  ~/repo/vit/data-mining-assignments/Lab/DA 3 main +3 ?2 ·····
└─bear❯
```

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

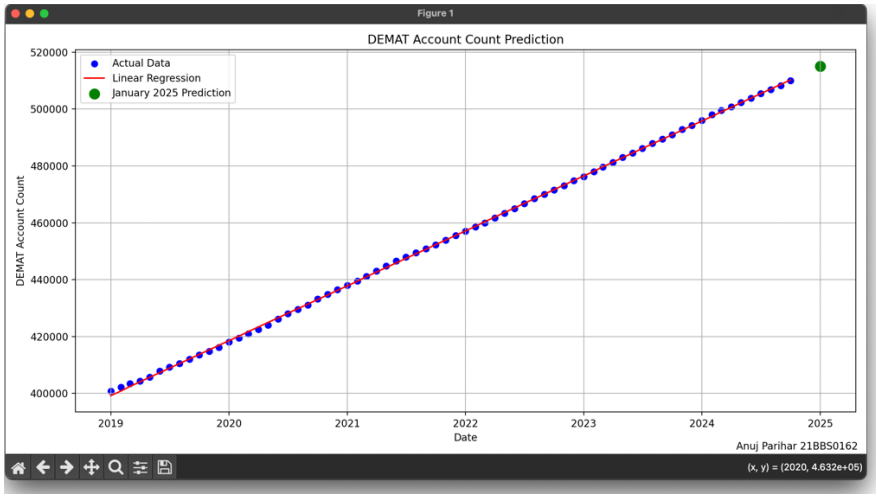
Month	DEMAT_Count
2019-01	400810
2019-02	402120
2019-03	403450
2019-04	404300
2019-05	405700
2019-06	407890
2019-07	409250
2019-08	410500
2019-09	412000

Output:

```
bear› python 2.py
Anuj Parihar 21BBS0162


Mean squared error: 277115.32933752885
R-squared score: 0.9997414195889317

Predicted DEMAT account count for January 2025: 515,109
2024-10-06 21:54:43.443 python[31054:208067] +[IMKClient subclass]: chose IMKClient_Legacy
2024-10-06 21:54:43.443 python[31054:208067] +[IMKInputSession subclass]: chose IMKInputSession_Legacy
```



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:

```
bear❯ python 3.py

Anuj Parihar 21BBS0162

Accuracy: 0.3
Classification Report:
              precision    recall  f1-score   support

       Apple       0.50      0.33      0.40         6
      Banana       0.20      0.25      0.22         4
       Grape       0.25      0.50      0.33         4
      Orange       0.33      0.17      0.22         6

    accuracy                           0.30        20
   macro avg       0.32      0.31      0.29        20
weighted avg       0.34      0.30      0.30        20
```

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