Digital Assessment 4
CBS3007 - Data Mining and Analytics

Date: 3 November, 2024
Name: Anuj Parihar
Registration Number: 21BBS0162

Link to Assessment Codebase and Dataset:
https://github.com/BearTS/data-mining-assignments/tree/main/Lab/DA%204

Question 1

Consider a dataset of 50 user records with the attributes “Name”, “location” ,“Height”, “Weight”, ”Age”. Do the following tasks.
i) Create the dataset for the attributes given.
ii) Implement the Demo on Classification Technique using KNN.

Aim: The aim of this project is to implement a K-Nearest Neighbors (KNN) classification technique on a synthetic dataset of user information. The dataset includes the attributes Name, Location, Height, Weight, and Age for 50 individuals.

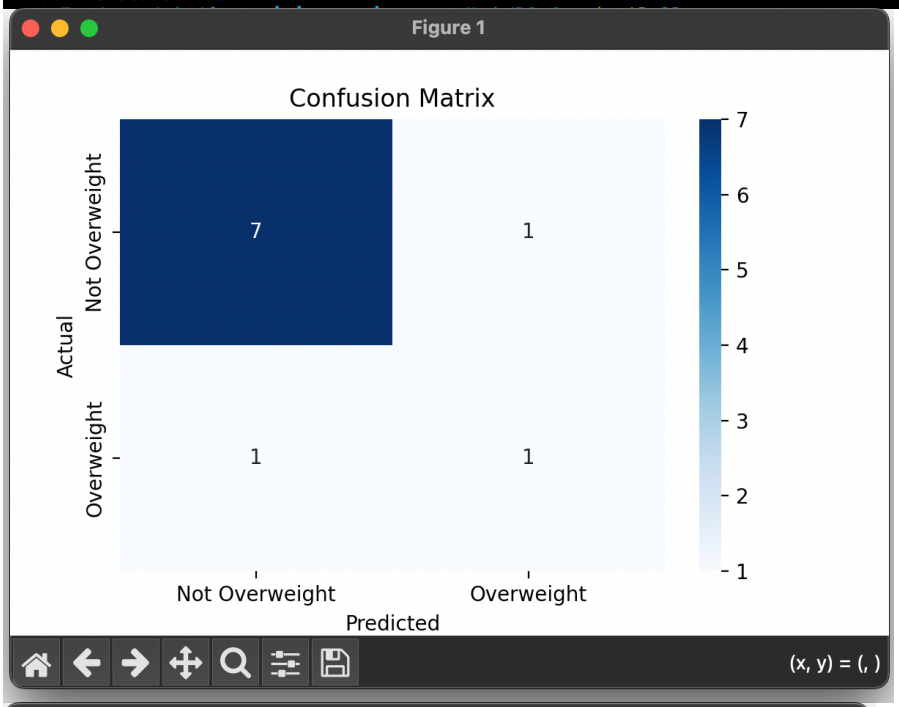
Sample Input: The entire input dataset is in the GitHub repository

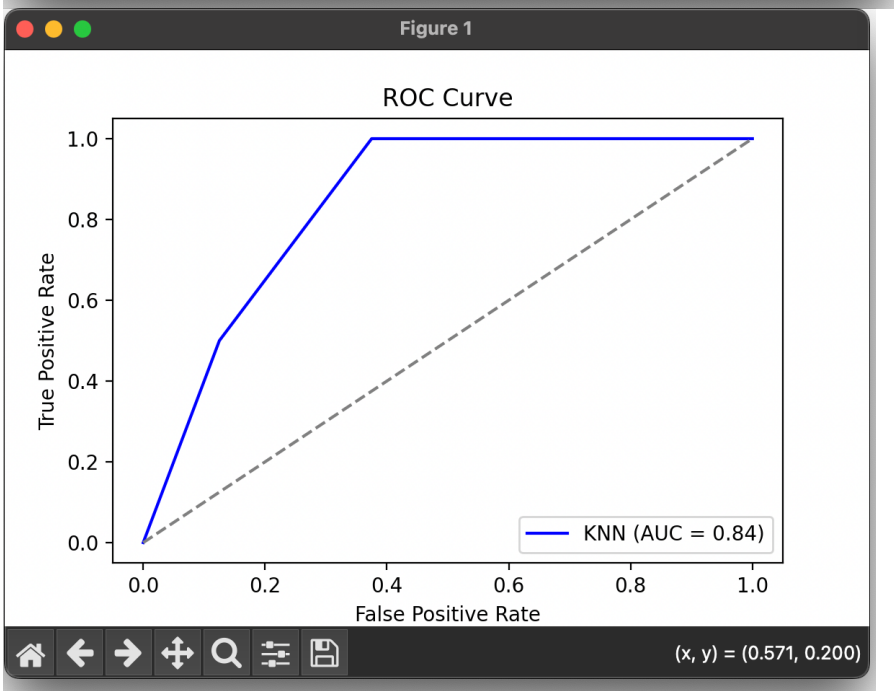
Name	Location	Height	Weight	Age
Julie King	Bengaluru	159	55	40
Brenda Martinez	Delhi	172	92	20
Carlos Miller	Indore	167	88	32
Mary Barnett	Delhi	164	74	45
Caitlin Doyle	Indore	168	51	25
Chad Wolfe	Mumbai	182	82	28
Brian Herring	Delhi	174	63	22
Jennifer Smith	Vellore	170	59	35
Edward Lane	Delhi	165	68	24
Elizabeth Ramirez	Delhi	172	73	18

Output:

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


KNN Classification Accuracy: 0.80
Confusion Matrix:
 [[7 1]
 [1 1]]
2024-11-03 22:01:39.405 python[92417:2848338] +[IMKClient subclass]: chose IMKClient_Modern
2024-11-03 22:01:39.405 python[92417:2848338] +[IMKInputSession subclass]: chose IMKInputSession_Modern
Precision: 0.50
Recall: 0.50
F1 Score: 0.50
ROC AUC Score: 0.84
```





Results:
The KNN classifier successfully classified individuals as "Overweight" or "Not Overweight" based on BMI with an accuracy of 0.80. Key metrics, including precision, recall, F1 score, and ROC AUC, indicate reliable performance, demonstrating the model's effectiveness in distinguishing between the two classes.

Question 2:

A probabilistic based learning algorithm used for classifying the following data that depicts the people choice of buying the phone. Apply the same to identify the probability of getting loan approval for the case age 30-70, has criminal record and More than 5 year exp.

Aim:
The aim of this implementation is to use a probabilistic-based learning algorithm (Naive Bayes classifier) to predict the probability of loan approval based on certain demographic and behavioural features. This analysis is crucial for financial institutions to assess risks associated with loan approvals and to make informed lending decisions.

Input:

S.No	Income	Criminal Record	EXP Load	Approved?
1	<30	No	1-5	Yes
2	30-70	Yes	1	No
3	30-70	No	1	No
4	30-70	Yes	1-5	Yes
5	30-70	No	>5	Yes
6	30-70	Yes	1-5	No
7	>70	Yes	>5	Yes
8	>70	No	>5	No
9	<30	Yes	1-5	No
10	30-70	No	1-5	Yes
11	30-70	No	1-5	No
12	30-70	Yes	>5	Yes

Output:

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


Probability of loan approval (No, Yes): [0.22111666 0.77888334]
Probability of loan approval (Yes): 0.78
```

Results:
The results of the Naive Bayes classification provide insights into the likelihood of loan approval for specific profiles. By understanding these probabilities, lenders can better evaluate applicants and potentially adjust their lending strategies based on the risk associated with various demographics. This model can be further refined with additional data and features to improve its predictive capabilities