



Report

Assignment 3

Medical Data Analysis using Decision Trees

| | |
|--|-------------------|
| | علي بن دهيم |
| | عبدالرحمن الميمان |
| | خالد الدويسان |

Dataset Description :

In this code, we have analyzed the diabetes dataset using a decision tree classifier. The dataset contains information about patients and whether they have diabetes or not. There are eight input features and one output feature. The input features are Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age. The output feature is Outcome, which has two classes: 0 and 1.

Code :

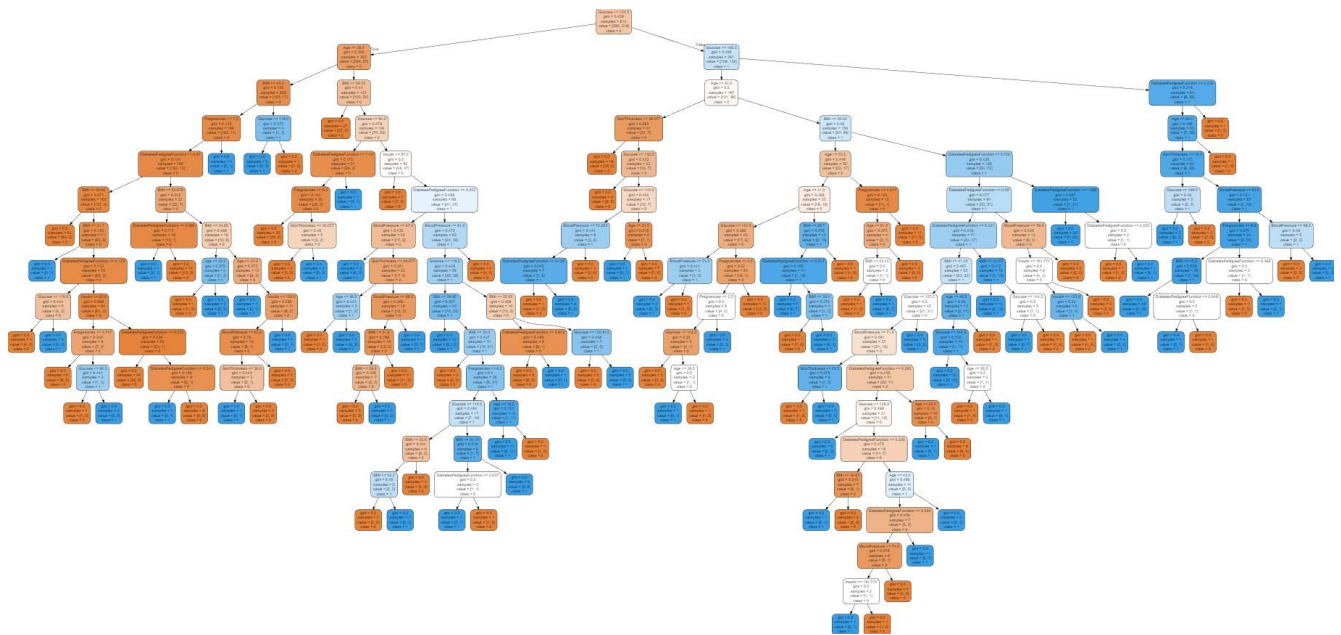
First, we read the dataset using pandas and replaced the missing values with the mean of the corresponding column. Then, we split the dataset into training and testing sets using three different ratios: 90/10, 80/20, and 70/30. For each split, we trained a decision tree classifier and evaluated its performance using accuracy, precision, recall, and F1 score metrics.



Assignment3.ipynb

Decision Tree Visualization :

We also visualized the decision tree using the graphviz library. The visualization shows the decision rules used by the classifier to predict the outcome of a patient.



Diabetes-Tree.dot

Furthermore, we analyzed the information gain of the split node features using the entropy criterion. The information gain represents the amount of information obtained by splitting the dataset based on a particular feature. The results showed that Glucose was the most informative feature, followed by BMI, Age, and DiabetesPedigreeFunction.

Finally, we compared the performance of the classifiers trained on different splits and selected the one with the highest accuracy. We then used the selected classifier to analyze the information gain of the split node features.

Insights and observations :

- The accuracy, precision, recall, and F1 score metrics were higher for the 80/20 split compared to the other splits .
- The Glucose feature was the most informative feature for predicting the outcome of a patient with diabetes. This confirms the importance of glucose monitoring for diabetes management.
- The BMI and Age features were also informative, which suggests that maintaining a healthy weight and aging gracefully are essential factors for diabetes prevention and management.
- The decision tree visualization provided a clear and intuitive representation of the decision rules used by the classifier. This can help healthcare professionals understand the factors that contribute to diabetes and develop personalized treatment plans for their patients.

```
1 90/10:
2 [0 0 1 0 1 0 1 0 1 1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 1 0 0 1
3 0 1 0 1 0 1 1 1 0 0 0 0 0 0 1 1 0 0 1 0 0 1 0 1 0 0 0 0 0 0 1 0 1 1 0 0 0 1
4 1 0 0]
5 Accuracy: 0.7272727272727273
6 Precision: 0.5769230769230769
7 Recall: 0.6
8 F1 score: 0.5882352941176471
9 -----
10 80/20:
11 [0 0 1 0 1 1 0 1 1 1 0 1 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 0 1 1 0 0 0
12 1 1 0 0 0 1 0 0 1 1 0 0 1 0 0 0 1 0 1 0 1 0 0 1 0 0 0 0 1 1 0 0 1 0 0 1 0
13 0 0 1 1 0 1 0 1 0 1 0 0 0 1 1 1 1 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 1
14 0 0 0 1 1 1 0 0 1 0 1 0 0 1 0 0 0 1 0 1 1 0 0 0 1 0 0 1 1 0 0 0 0 0 1 0 1
15 1 0 0 0 1 0]
16 Accuracy: 0.6753246753246753
17 Precision: 0.4642857142857143
18 Recall: 0.5652173913043478
19 F1 score: 0.5098039215686274
20 -----
21 70/30:
22 [1 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 1 1 1 0 1 1 0 1 0 0 0 0 0 1 1 0 0
23 0 0 0 1 1 0 0 0 0 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0 0 0 1 0 1 0 0 0 0 0 0 0 1
24 0 0 0 1 0 1 1 0 0 0 0 1 1 0 1 1 0 0 0 0 1 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 0
25 0 1 1 0 0 1 1 1 1 1 0 0 0 0 1 0 0 1 0 0 0 0 0 1 0 0 1 1 1 0 0 0 0 0 0 0 1
26 0 1 0 0 1 1 1 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1
27 0 0 1 1 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0 0 1 0 1 0 1 0 0 1 0 1 1 0 0 1 0
28 0 0 0 0 0 0 1 1 0]
29 Accuracy: 0.658008658008658
30 Precision: 0.547945205479452
31 Recall: 0.46511627906976744
32 F1 score: 0.5031446540880503
33 -----
34 Information Gain for Split Node Features:
35 Pregnancies: 0.0562
36 Glucose: 0.2807
37 BloodPressure: 0.0800
38 SkinThickness: 0.0502
39 Insulin: 0.0802
40 BMI: 0.1696
41 DiabetesPedigreeFunction: 0.1261
42 Age: 0.1570
43 |
```