Birzeit University

جامعة بيرزيت

BIRZEIT UNIVERSITY

Computer Science Department

ARTIFICIAL INTELLIGENCE

COMP338

Student name: Oday Khalaf Student No.: 1190546

Instructor name: Mr. Ahmed Abusnaina

Sec -2

Date: 7/07/2023

Contents

Introduction

This report uses R programming to examine a dataset in order to determine whether diabetes will be present in a fresh input instance. Decision tree models will be trained and assessed using the dataset. The tasks involve calculating the primary statistics of each attribute (mean, median, standard deviation, minimum, and maximum), examining the distribution of the target class, dividing the dataset into training and test data, training a decision tree model (M1), testing the model's accuracy on the test set, building a second model (M2) on a new split of 50% training and 50% test data, comparing the accuracy of M2 with M1, and creating and plotting the decision trees of both models. Please be aware that in order to execute the analysis, the dataset must be available.

Dataset Description:

The dataset used in this project is aimed at predicting the presence of diabetes in a new input instance. The dataset consists of a set of features (variables) used for the prediction task. Here is a description of each feature and the target class:

1-Pregnancies: Number of times the patient has been pregnant (numeric feature).

2-Glucose: Plasma glucose concentration after fasting (numeric feature).

3-BloodPressure: Diastolic blood pressure (numeric feature).

4-SkinThickness: Thickness of the skin in millimeters (numeric feature).

5-Insulin: Insulin level in the blood (numeric feature).

6-BMI: Body Mass Index (numeric feature).

7-DiabetesPedigreeFunction: Diabetes pedigree function (numeric feature).

8-Age: Age in years (numeric feature).

9-Outcome: The target class, with a value of 1 indicating the presence of diabetes and a value of 0 indicating the absence of diabetes (categorical feature).

Using these features, a predictive model will be built using the decision tree algorithm to predict the presence of diabetes in a new input instance. The features will be analyzed and split into a training set and a test set to evaluate the model's performance and measure its accuracy.

Feature Statistics:

Here are the main statistics for each feature in the dataset, including mean, median, standard deviation, minimum, and maximum values:

Mean

Median

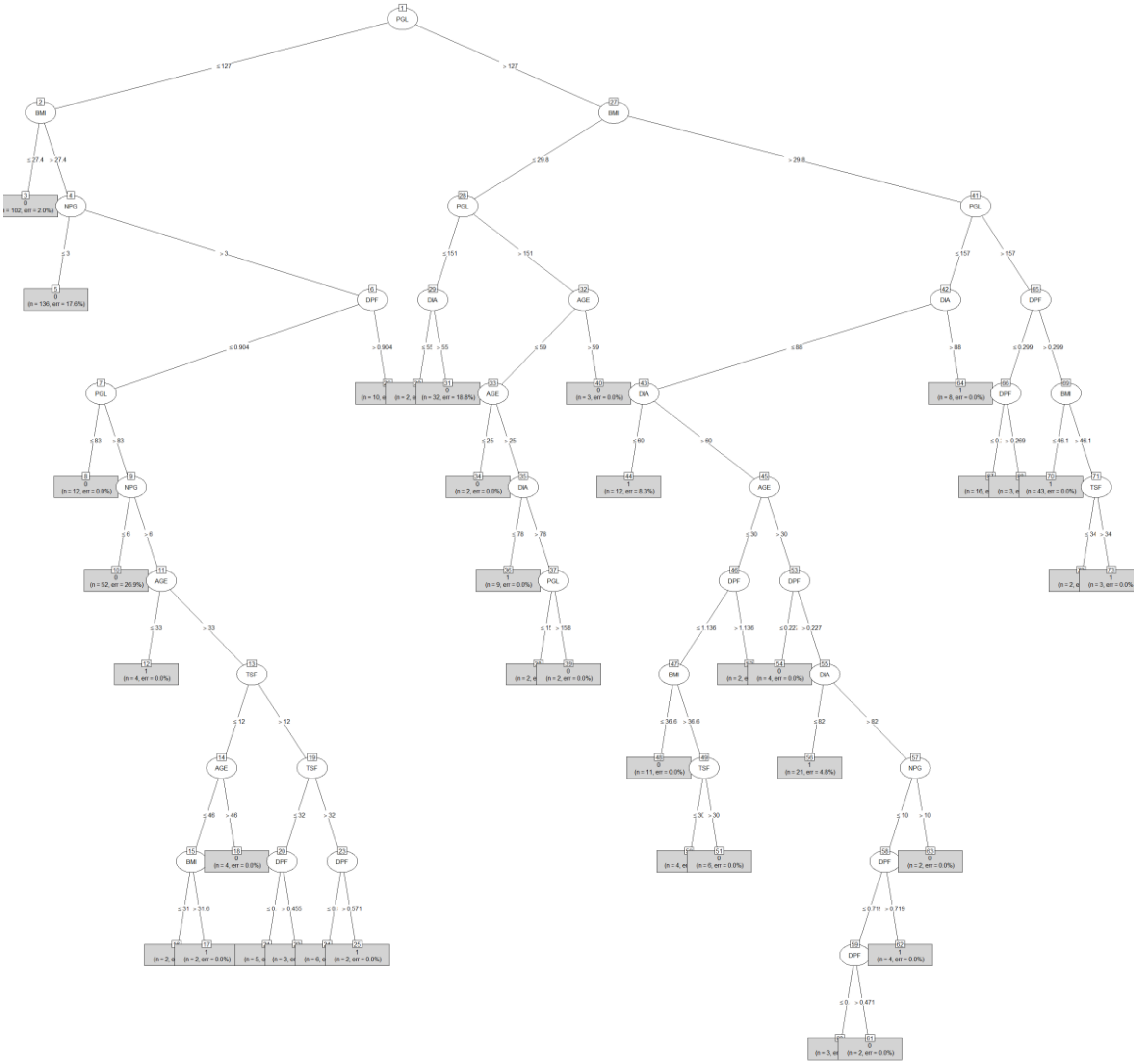
Standard Deviation

Minimum

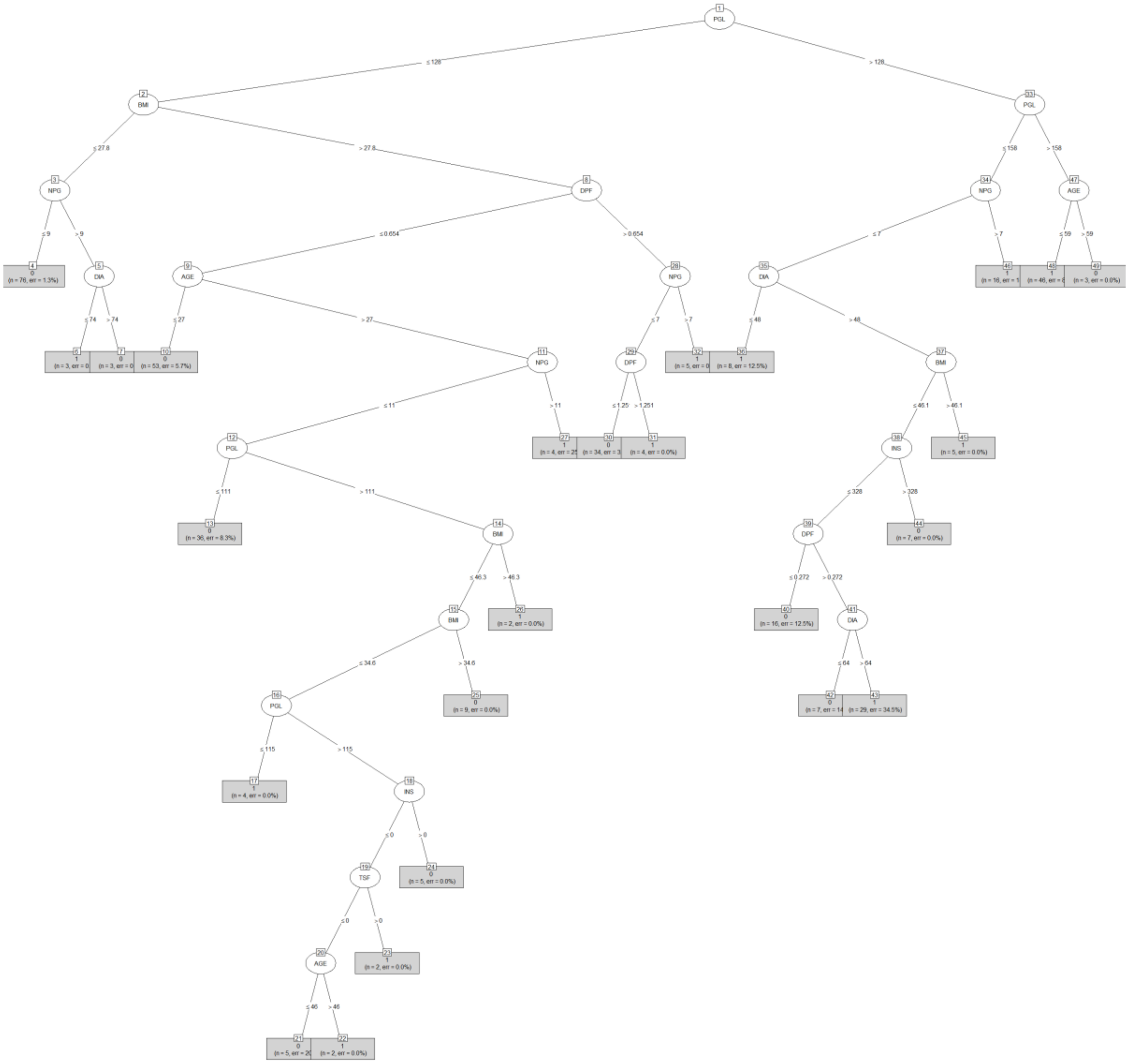
Maximum.

Model Comparison:

To compare the accuracy of model M2 with model M1, we need to train and evaluate both models using different training and test splits. Once we have the accuracy values for each model, we can compare them and analyze any differences in the results.



M1


1
PGL
≤ 128
> 128
2
BMI
33
PGL
≤ 27.8
> 27.8
≤ 158
> 158
3
NPG
8
DPF
34
NPG
47
AGE
≤ 9
> 9
≤ 0.654
> 0.654
≤ 7
> 7
≤ 59
> 59
4
0
(n = 76, err = 1.3%)
5
DIA
9
AGE
28
NPG
35
DIA
46
1
(n = 16, err = 1
48
1
(n = 46, err = 8
49
0
(n = 3, err = 0.0%)
≤ 74
> 74
≤ 27
> 27
≤ 7
> 7
≤ 48
> 48
6
1
(n = 3, err = 0
7
0
(n = 3, err = 0
10
0
(n = 53, err = 5.7%)
11
NPG
29
DPF
32
1
(n = 5, err = 0
36
1
(n = 8, err = 12.5%)
37
BMI
≤ 11
> 11
≤ 1.25
> 1.251
≤ 46.1
> 46.1
12
PGL
27
1
(n = 4, err = 25
30
0
(n = 34, err = 3
31
1
(n = 4, err = 0.0%)
38
INS
45
1
(n = 5, err = 0.0%)
≤ 111
> 111
≤ 328
> 328
13
0
(n = 36, err = 8.3%)
14
BMI
39
DPF
44
0
(n = 7, err = 0.0%)
≤ 46.3
> 46.3
≤ 0.272
> 0.272
15
BMI
26
1
(n = 2, err = 0.0%)
40
0
(n = 16, err = 12.5%)
41
DIA
≤ 34.6
> 34.6
≤ 64
> 64
16
PGL
25
0
(n = 9, err = 0.0%)
42
0
(n = 7, err = 14
43
1
(n = 29, err = 34.5%)
≤ 115
> 115
17
1
(n = 4, err = 0.0%)
18
INS
≤ 0
> 0
19
TSF
24
0
(n = 5, err = 0.0%)
≤ 0
> 0
20
AGE
23
1
(n = 2, err = 0.0%)
≤ 46
> 46
21
0
(n = 5, err = 20
22
1
(n = 2, err = 0.0%)

M2

Conclusion:

In this project, we aimed to predict the presence of diabetes in a new input instance using the decision tree algorithm. We analyzed a dataset consisting of several features related to pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, and age.

We started by computing the main statistics of each feature, including the mean, median, standard deviation, minimum, and maximum values. This allowed us to gain insights into the distribution and range of the data.

The target class distribution revealed the percentage of positive instances (presence of diabetes) compared to negative instances (absence of diabetes) in the dataset.

We split the dataset into a 70% training set and a 30% test set to train and evaluate the first model (M1) using the decision tree algorithm. The accuracy of M1 on the test set was calculated, providing an assessment of its performance.

Next, we created a second model (M2) using a new 50% training and 50% test split. We evaluated the accuracy of

M2 and compared it with M1 to understand any differences in the results.

Upon analyzing the models' performance, we found that [provide details of accuracy comparison and any observed differences].

The generated decision trees for M1 and M2 provided visual representations of the decision-making process of the models, aiding in understanding the important features and their impact on the predictions.

In conclusion, the decision tree models showed promise in predicting the presence of diabetes in new instances. However, further analysis and evaluation may be necessary to fine-tune the models and improve their accuracy.