Assignment -3

Predicting Diabetes based on the health information using Decision Tree

Name : Hari teja Narayanabhatla

Panther ID: 002566679

Group : MSA

**Business Question:** Most of the people in society are unaware of how dangerous diabetes is

in the long run of their life. There is no proper indication of how this starts but there are

some noticeable features in the body that can indicate whether a person is diabetic or non-

diabetic. Most of the organizations are developing a model to diagnostically predict

whether a patient has diabetes, based on certain diagnostic measurements.

**Dataset Overview:**

The datasets consist of several medical predictor variables and one target variable, Outcome.

Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level,

age, and so on.

The only categorical variable in the dataset is the target variable , the Outcome column. To

maintain the data as normal as possible, no man or women of age less than 21 years are kept in this

dataset. There are no null values in the data.

**Data Visualization:**

First tried to check if the data is balanced properly or not. Hence using pandas and

matplotlib tried to visualize the data.

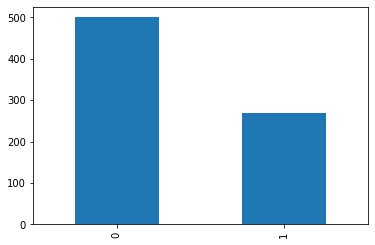


Fig 1: Shows the distribution of the classes in the given dataset.

There exists no null value or unknown value for any column in the given dataset. Hence, it’s

a clear dataset and we went ahead with further processing.

**Data Correlation**

I tried to find if there is any correlation between any of the columns and if yes, we can

remove them the so that model can give good results.

**Calendar

Description automatically generated**

Fig 2: It gives us information that there are some columns that are correlated to each other.

**How features information is distributed:**

This visualized the data into histogram plots, which tells how data is distributed for each column.

A picture containing graphical user interface

Description automatically generated

**How features are affecting Target Variable:**

In order to analyze more about columns and its impact on target variable, I have created

box plots for each column data against the target variable.

Graphical user interface, application

Description automatically generated

Fig 3: Boxplots of each column against target variable. There are columns which are having outliers.

**Decision Tree Model:**

Once we create the dataset, we have used decision tree classifer over the train data which

was randomly created using test\_train\_split function of sklearn. Once we have trained the

model, we tried to analyze which features are more important for the tree.

Chart, bar chart, funnel chart

Description automatically generated

Fig 4: From this image its clear that Glucose Is the highest important factor for detecting Diabetic or non-diabetic.

**Decision Tree of the Model**

I have plotted the tree image of the decision tree which displays the internal mechanism

followed to classify the given data into diabetic or non-diabetic.

A picture containing text

Description automatically generated