Logo

Description automatically generated

Diabetes project

Prepared for: Data analysis CSCI 322

|  |  |
| --- | --- |
| Abdelrahman Ahmed Ameen | 202001064 |
| Mohamed Ali Nassar | 221000965 |
| Ahmed shrife zahran | 202000905 |
| Mohamed Ibrahim salama | 211001298 |
| Mouhab Samer Sayed | 202000412 |

**Contents**

[**Introduction** 3](#_Toc135442082)

[**Data Description** 3](#_Toc135442083)

[**Data source:** 3](#_Toc135442084)

[**Size and format:** 3](#_Toc135442085)

[**Data type:** 3](#_Toc135442086)

[**Missing values:** 3](#_Toc135442087)

[**Outliers:** 3](#_Toc135442088)

[**Data quality:** 3](#_Toc135442089)

[**Data Cleaning** 4](#_Toc135442090)

[**Exploratory Data Analysis (EDA):** 5](#_Toc135442091)

[**Modelling:** 5](#_Toc135442092)

[**Results:** 6](#_Toc135442093)

[**Conclusion:** 6](#_Toc135442094)

# **Introduction**

Diabetes is a chronic disease that affects how the body turns food into energy. This project will use data analysis to better understand diabetes, identify risk factors, and look at the different attributes that is in the data to get a prediction for this disease. The data we used is a set of different attributes like gender, age, glucose level in blood, has diabetes. And from this data set we get set a decision tree classifier to get prediction about weather a patient has diabetes with accuracy of 95%

# **Data Description**

The data I used (diabetes\_prediction\_dataset-Copy1.edit.csv) and includes the following:

## **Data source:**

Data was obtained from Kaggle data sets

## **Size and format:**

The data consist of 100000 rows and 9 columns gender', 'age', 'hypertension', 'heart\_disease', 'smoking\_history', 'bmi', 'HbA1c\_level', 'blood\_glucose\_level', 'diabetes'

## **Data type:**

The data contains both numerical and categorical types, but we do make it all numerical when modelling it

## **Missing values:**

The data didn’t consist of a lot of missing values the number of missing values sum of 32 missing values in age, HbA1c\_level, etc.

## **Data quality:**

The data was clean and didn’t have a issues and no duplication.

# **Data Cleaning**

The steps we took for cleaning and handling the data was quite simple since the data wasn’t full of null values and outliers so after checking the data, we did the following to clean it:

* **Handling missing values:** the missing values in the data were handled in different ways depending on weather they were numerical or categorical. For example, numerical data such as Age we handled by filling in the missing values with the mean of the column, and the categorical were handled differently since we can’t use the mean on it, it was handled by using “bfill” which fills in the null values with the previous.
* **Handling outliers:** there were some outliers in the data as in the age column there were values that were Zero which didn’t make sense, so we filled them with the mean value of the age.
* **Data validation:** by checking the values of the data and their data types in the Checking the data process and see the unique values and info about the data and seeing that all Dtypes did match and there were no more outliers and null values in the data

# **Exploratory Data Analysis (EDA):**

For the Exploratory Data Analysis (EDA) section from my data set we check the relationship between different sets of data like in the following

* **Summary statistics:** For the Age the amount of patient is at its highest from 40 to 55. And there is a higher number of Female in the data set 58.6%, and 41.4% male.
* **Visualizations:**

A picture containing text, screenshot, diagram, plot

Description automatically generatedA picture containing text, screenshot, diagram, circle

Description automatically generated A picture containing text, screenshot, rectangle, pattern

Description automatically generated

* **Correlations:** From the correlation heat map we can get a couple of things for example there is a positive relation between Blood\_glucose\_level and diabetes which is a symptom of diabetes so it makes sense, and the positive relation between body mass and diabetes.

A picture containing text, screenshot, diagram

Description automatically generated

# **Modelling:**

* For the Modelling of the data we used the Decision Tree Classifier for the data but first was pre-processing of the data were we create 3 new columns in the data DataFrame: age\_group, glucose\_level\_range, and bmi\_range. These columns contain categorical values that represent the ranges of the age, blood\_glucose\_level, and bmi columns, respectively. Were the purpose of was to create column which will make it easier to use these columns in machine learning models. During the modelling phase of the code, creating X that contains all the features except for the diabetes column, Y that contains the [‘diabetes’]. Then Train the X and Y sets and creating a list with numerical columns and another with the categorical columns. Then creating the decision tree classifier and then checking the accuracy, F1 score, precision of the model.

# **Results:**

For the results of the project, we discovered the and the correlation and the predication model has a 95% Accuracy.

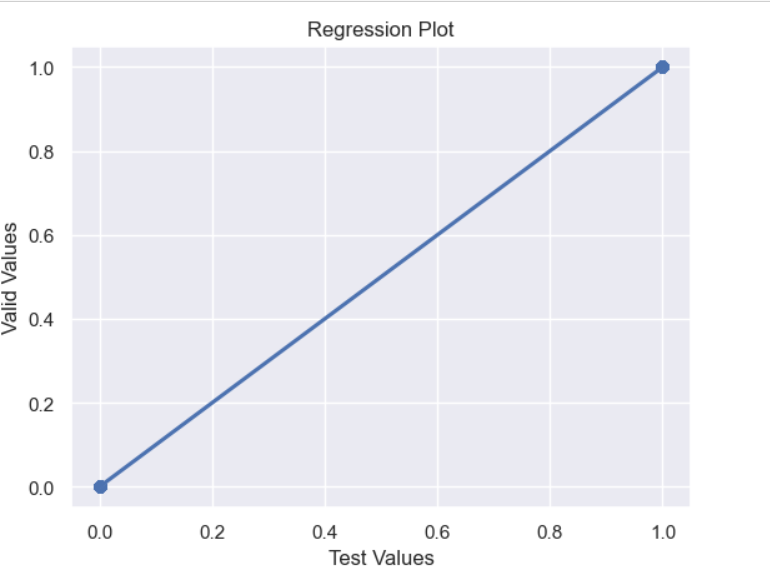
For the results of the model:

* **Model results:** in the model the results were as following.

DTC Accuracy: 95.295

DTC Precision: 95.295

DTC F1 score: 95.295

* **Key results:** decision tree classifier that can accurately predict whether or not a patient has diabetes. The model has an accuracy of 95%, a precision of 95%, and an F1 score of 95%.
* **Conclusion\_visualization:** 

# **Conclusion:**

For this project the question was whether we can predict if the patient has or doesn’t have diabetes from a set of data using decision tree classifier and in conclusion we managed to build a decision tree classifier to predict the probability of a patient having diabetes or not with an accuracy of 90%