**INTRODUCTION**

There are 3 types of diabetes mellitus,

Type 1 diabetes/juvenile diabetes/insulin dependent diabetes mellitus(IDDM) - It’s a condition whereby the body makes little or no insulin, its most prevalent among children, teenagers and young adults, the disease though can develop at any age.

Type 2 diabetes/Non-Insulin Dependent Diabetes Mellitus(NIDDM) - It’s the most common type of diabetes mellitus. It develops as of insulin resistance or when cell in the body are unable to utilize the available insulin adequately.

Gestational diabetes-it’s a type of diabetes that occurs during pregnancy, when you’re your body fails to make enough insulin. The condition may resolve after delivery but increases the individuals risk of developing type 2 diabetes later in life.

Diabetes type 2 is one of the top 10 diseases globally that contribute to adult mortality annually. As at 2019,1it was estimated that 463 million adults were living with diabetes, of these 9% were women and 9.6 were men. Its estimated that this figure could rise to 598 million by 2030 and 700 million by 2045(Saeedi et al 2019)

Type 2 Diabetes is more prevalent in adults age 20-74 years of age. The predisposing factors include genetics, obesity, Physical inactivity, insulin resistance, age, race, cardiovascular diseases, hypertension, gestational diabetes and polycystic Ovary Cyst.

**LITERATURE REVIEW**

According to (Thomas,2021), Lately, Machine learning has played a key role in many health related project. Some of the uses of machine learning in healthcare sector include predicting illness and treatment, forecasting health risks to various populations, assisting with health care record and workflow and identifying opportunities for clinical trials.

Artificial intelligence is bringing a paradigm shift into the healthcare sector. This is as result of increasing health care data and rapid progress of analytics techniques (Kumar, Das, Gupta, Gupta & Bidraa,2021). Researchers have developed some apps that aid in real-time diagnosis of diseases. Some of these apps even predict the risk of developing certain diseases, and encourage the individual to seek further help from health professional to ascertain their health conditions techniques (Kumar, Das, Gupta, Gupta & Bidraa,2021)

Beta bionics are developing an app that they call iLet. The app is meant to help manage blood sugar levels around the clock for those with type 1 diabetes. Some of the highly rated Diabetes apps include MySugr, Glucose Buddy, Diabetes M, Beat Diabetes, One Touch Reveal and Your Blood Sugar (Doyle, 2022)

Machine learning models have the ability to bring meaning to data and aid in early prediction and diagnosis of diseases. It is for this reason that they are used in diabetes problem to bring out solutions (Kumar, Das, Gupta, Gupta & Bidraa,2021)

**THE DATASET**

This dataset used in this machine learning project was derived from the data. world website. The dataset contains data from 15,000 women, all within the ages of (21-72). The dataset has 10 attributes. The attributes are as defined below.

Pregnancies – Number of pregnancies that a woman has had throughout her life

Plasma Glucose- - The level of fasting blood glucose/sugar in the blood of an individual.

Diastolic Blood Pressure- the pressure on the arteries, during the hearts resting period between heart beats

Triceps Thickness – this measures the skinfold at the middle of the less active arm to estimate the body fat of an individual

Serum Insulin- the amount of fasting insulin in the blood of an individual

BMI – measures how far an individual deviate from their normal weight compared to their height

Diabetes Pedigree – it’s a factor that score the likelihood of an individual suffering from diabetes according to genetic pre-disposition.

Age- The age of the participant as at the time of data collection

Diabetic – The Status that show weather an individual is Diabetic or not

Patients ID- A unique number used to identify a specific participant during data collection

Below is a section of the uploaded dataset, table 1.0 showing the dataset and table 1.1 showing the dataset information.

**The uploaded dataset**

|  | **Patient**  **ID** | **Pregnancies** | **Plasma**  **Glucose** | **Diastolic**  **Blood Pressure** | **Triceps**  **Thickness** | **Serum Insulin** | **BMI** | **Diabetes Pedigree** | **Age** | **Diabetic** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 1354778 | 0 | 171 | 80 | 34 | 23 | 43.509726 | 1.213191 | 21 | 0 |
| **1** | 1147438 | 8 | 92 | 93 | 47 | 36 | 21.240576 | 0.158365 | 23 | 0 |
| **2** | 1640031 | 7 | 115 | 47 | 52 | 35 | 41.511523 | 0.079019 | 23 | 0 |
| **3** | 1883350 | 9 | 103 | 78 | 25 | 304 | 29.582192 | 1.282870 | 43 | 1 |
| **4** | 1424119 | 1 | 85 | 59 | 27 | 35 | 42.604536 | 0.549542 | 22 | 0 |

**Table 1.0**

**Dataset information**

|  |  |
| --- | --- |
| # Column Non | Null Count Dtype |
| 0 Patient ID 15000 non | null int64 |
| 1 Pregnancies 15000 non | null int64 |
| 2 Plasma Glucose 15000 non | null int64 |
| 3 Diastolic Blood Pressure 15000 non | null int64 |
| 4 Triceps Thickness 15000 non | null int64 |
| 5 Serum Insulin 15000 non | null int64 |
| 6 BMI 15000 non | null float64 |
| 7 Diabetes Pedigree 15000 non | null float64 |
| 8 Age 15000 non | null int64 |
| 9 Diabetic 15000 non | null int64 |
| Dtype count: float64(2), int64(8) | |

**Table 1.1**

**DATA PRE-PROCESSING**

**Data description**

|  | **Patient ID** | **Pregnancies** | **Plasma Glucose** | **Diastolic Blood Pressure** | **Triceps Thickness** | **Serum Insulin** | **BMI** | **Diabetes Pedigree** | **Age** | **Diabetic** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 1.500000e+04 | 15000.000000 | 15000.000000 | 15000.000000 | 15000.000000 | 15000.000000 | 15000.000000 | 15000.000000 | 15000.000000 | 15000.000000 |
| **mean** | 1.502922e+06 | 3.224533 | 107.856867 | 71.220667 | 28.814000 | 137.852133 | 31.007733 | 0.080867 | 30.137733 | 0.333333 |
| **std** | 2.892534e+05 | 3.391020 | 31.981975 | 16.758716 | 14.555716 | 133.068252 | 9.763696 | 0.286472 | 12.089703 | 0.471420 |
| **min** | 1.000038e+06 | 0.000000 | 44.000000 | 24.000000 | 7.000000 | 14.000000 | 18.000000 | 0.000000 | 21.000000 | 0.000000 |
| **25%** | 1.252866e+06 | 0.000000 | 84.000000 | 58.000000 | 15.000000 | 39.000000 | 21.000000 | 0.000000 | 22.000000 | 0.000000 |
| **50%** | 1.505508e+06 | 2.000000 | 104.000000 | 72.000000 | 31.000000 | 83.000000 | 31.000000 | 0.000000 | 24.000000 | 0.000000 |
| **75%** | 1.755205e+06 | 6.000000 | 129.000000 | 85.000000 | 41.000000 | 195.000000 | 39.000000 | 0.000000 | 35.000000 | 1.000000 |
| **max** | 1.999997e+06 | 14.000000 | 192.000000 | 117.000000 | 93.000000 | 799.000000 | 56.000000 | 2.000000 | 77.000000 | 1.000000 |

The dataset had no missing value. We applied the pandas function, data. describe () to outlay the descriptive statistics of the data set. The table 2.0 below describes the characteristics of the data frame that include, central tendency, dispersion and distribution.

**Table2.0**

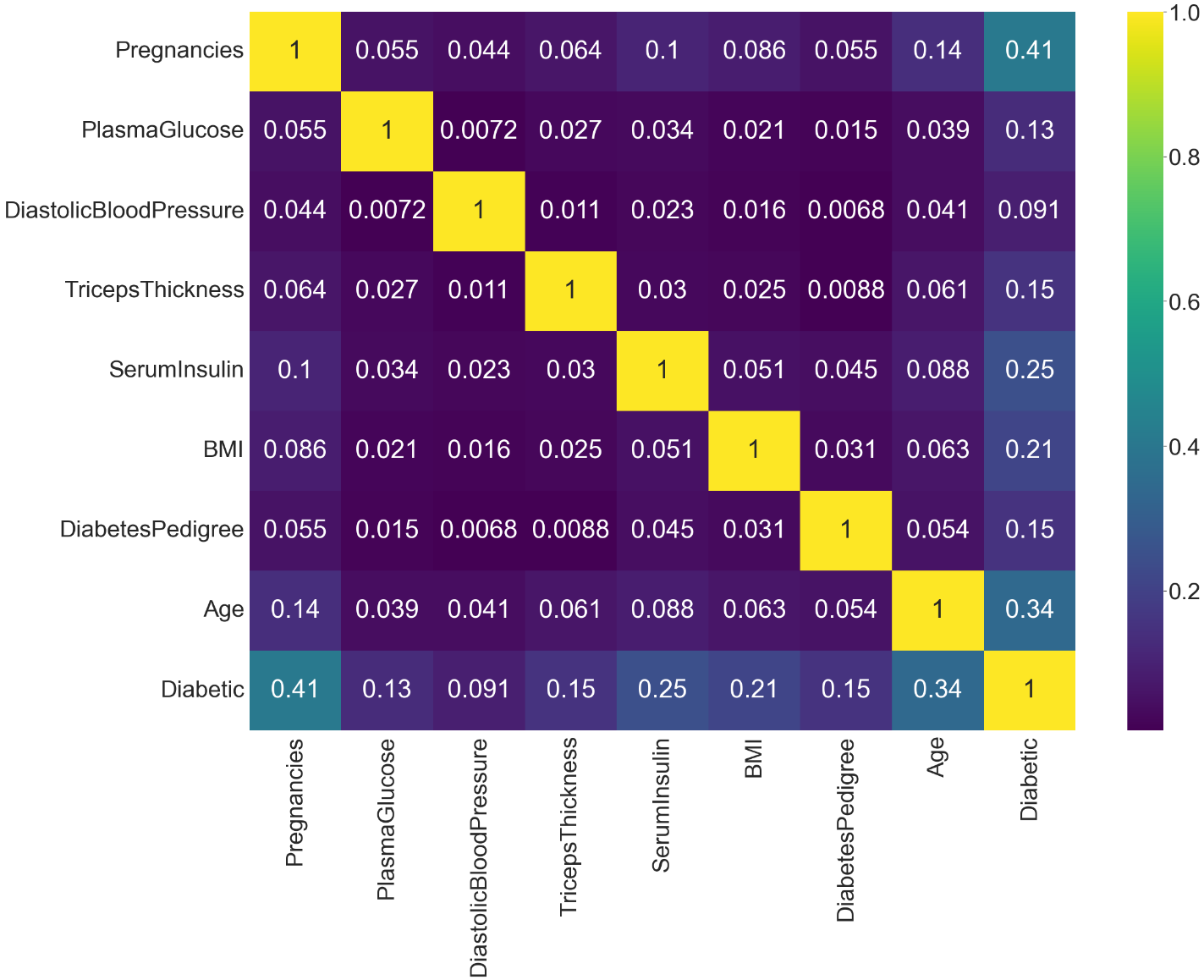
We then binned the data. This process provided us with new columns and type of data known as the categorical data. The categorical data would help us further understand our data frame by categorizing different subset of data into smaller groups. Below is a sample of the categorical data as shown by table 2.1

|  | **Diabetic** | **Age\_ group** | **BMI\_ chart** | **Plasma Glucose\_ cat** | **Pregnancies\_ grp** | **Diastolic Blood Pressure\_ chart** | **Serum Insulin\_ chart** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0 | young\_adults | 0BESE 3 | high | 0-5 | Normal | Lowbloodsugar |
| **1** | 0 | young\_adults | NORMAL | good | 6-10 | Highbloodpressure | Lowbloodsugar |
| **2** | 0 | young\_adults | 0BESE 3 | moderate | 6-10 | Normal | Lowbloodsugar |
| **3** | 1 | middle\_ age | OVERWEIGHT | moderate | 6-10 | Normal | Deadly |
| **4** | 0 | young\_adults | 0BESE 3 | good | 0-5 | Normal | Lowbloodsugar |
| **5** | 0 | young\_adults | NORMAL | good | 0-5 | Highbloodpressure | Deadly |
| **6** | 0 | young\_adults | NORMAL | high | 0-5 | Normal | Deadly |
| **7** | 0 | young\_adults | LOW | good | 0-5 | Elevated | Lowbloodsugar |

**Table 2.1**

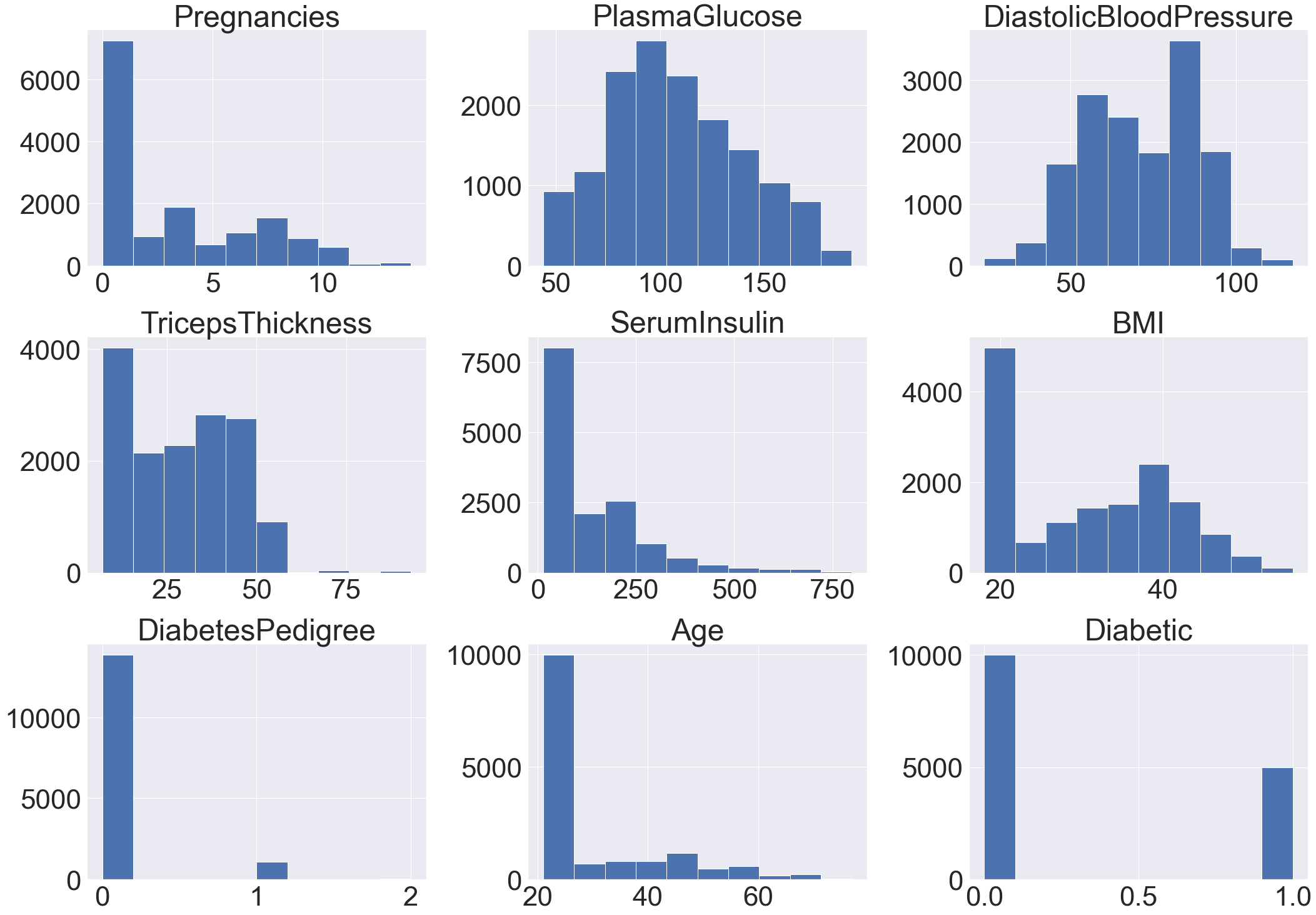
**Data visualization**

1. A heat map showing the correlation between attributes in the diabetic dataset.

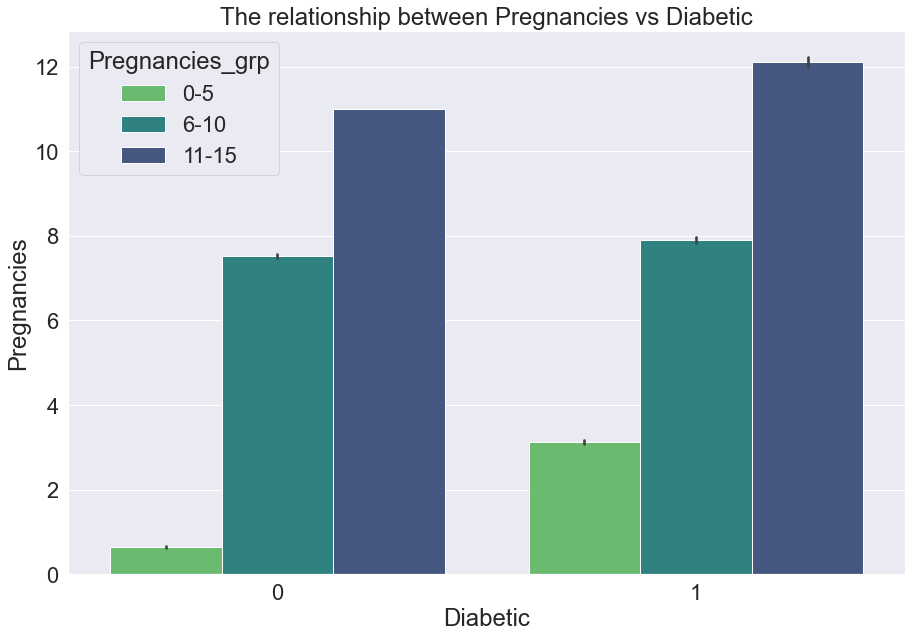


There is a moderate correlation between attributes Diabetic and Pregnancies and attributes Age and Diabetic. The rest of the attributes have a weak correlation with the target attribute and with each other.

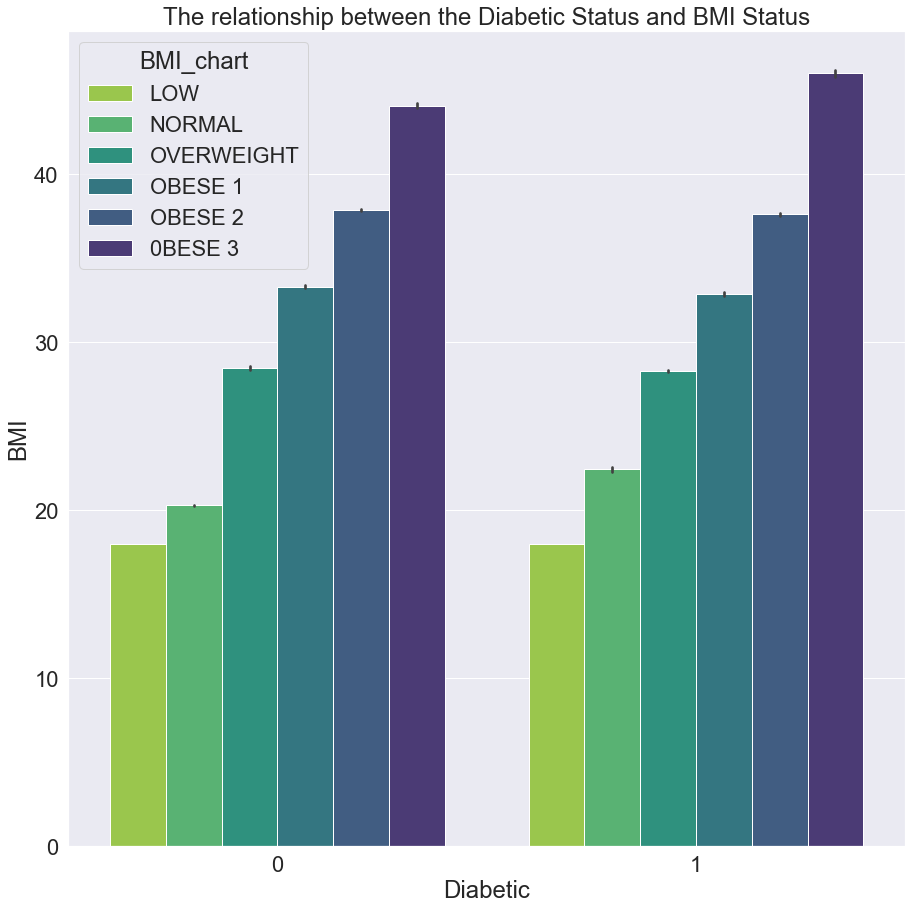
2.A histogram presentation of all the attributes of the diabetic dataset. It shows the distribution of data across each attribute



3.A bar graph showing the relation between being Diabetic and the number of pregnancies. It shows that having many pregnancies increases the risk of a woman suffering from diabetes.

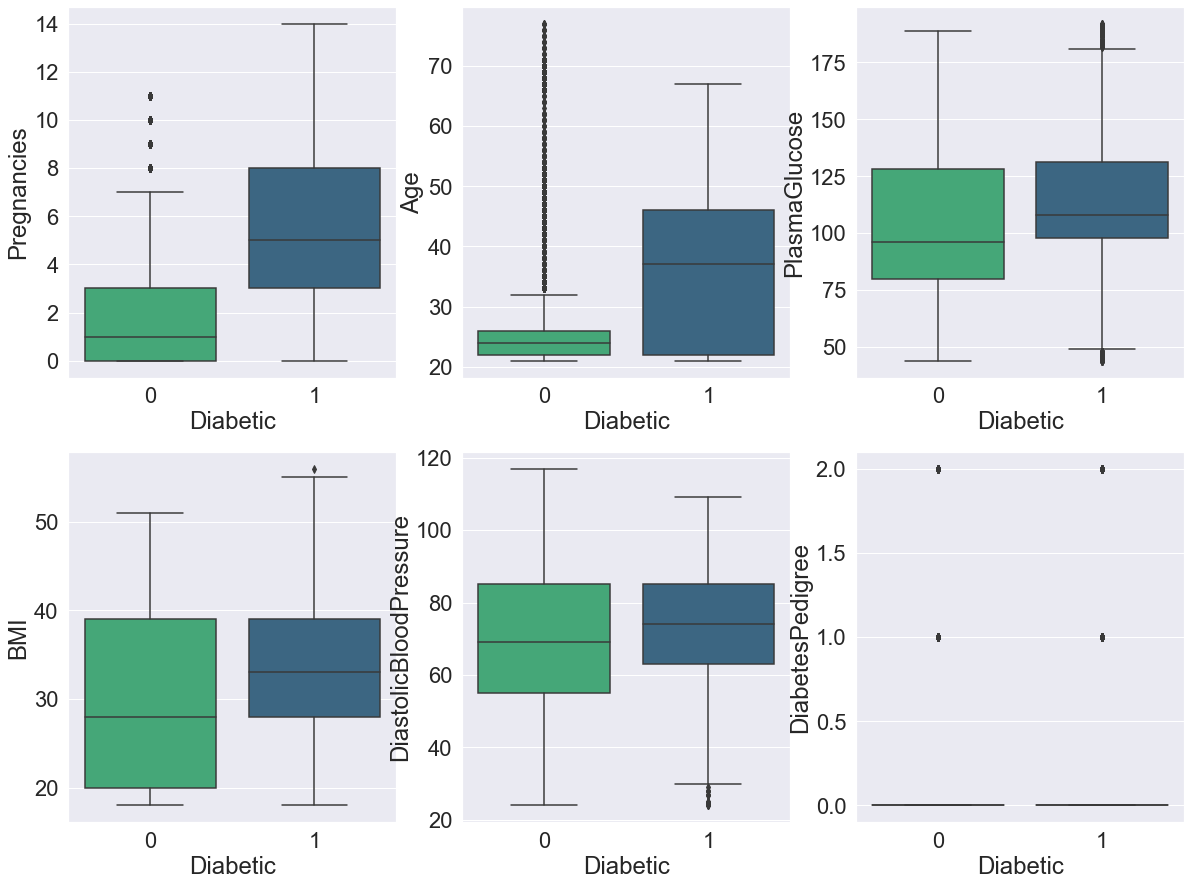


4.The box plot below shows the relationship between BMI status and being diabetic. There is a slight increase in the likelihood of suffering from diabetic in relation to being obese



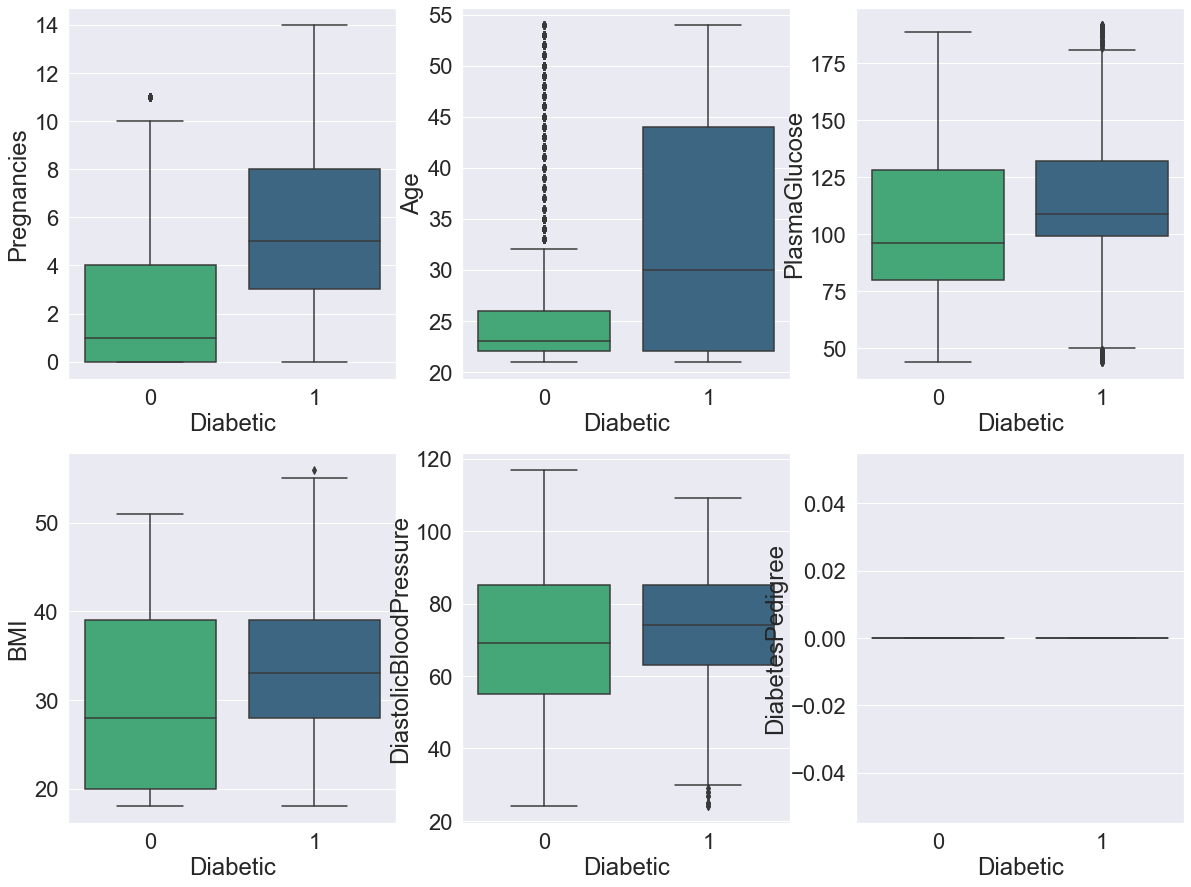
**Handling Outliers**

The outliers were detected and visualized using the boxplot technique as shown below.





We used Inter Quantile Range (IQR) technique to handle outliers. This method identifies the middle point of a frequency distribution within a dataset. This reduced our dataset from 15000 rows and 9 columns to 12358 rows and 9 columns. The boxplot visualization is a shown below.



**Separating features from the label**

We separated the features from the label. The features included Pregnancies, Plasma Glucose, Diastolic Blood Pressure, BMI, Diabetes Pedigree and Age. The attribute Diabetic became the Label.

**Feature scaling**

We scaled our data using the standardization method. Standardizing our data help the algorithms reach the minima of the cost much quicker. It also reduces the chances of algorithms being biased by bringing every feature to the same range.

**Splitting the data**

Both the feature and label data was split into training and test using a test size of 0.3. The data obtained was used to build classification algorithms in model building.

**CLASSIFICATION**

We used 4 machine learning classifiers that include Logistic Regression,

Decision Tree, Random Forest Classifier-Nearest Neighbors. The classification models performance is visualized below

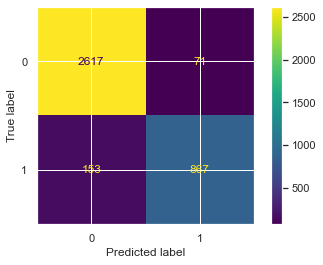
The best performing model as shown from the above chart was the Random Forest Classifier.

**MODEL EVALUATION**

**CLASSIFICATION REPORT AND CONFUSION MATRIX**

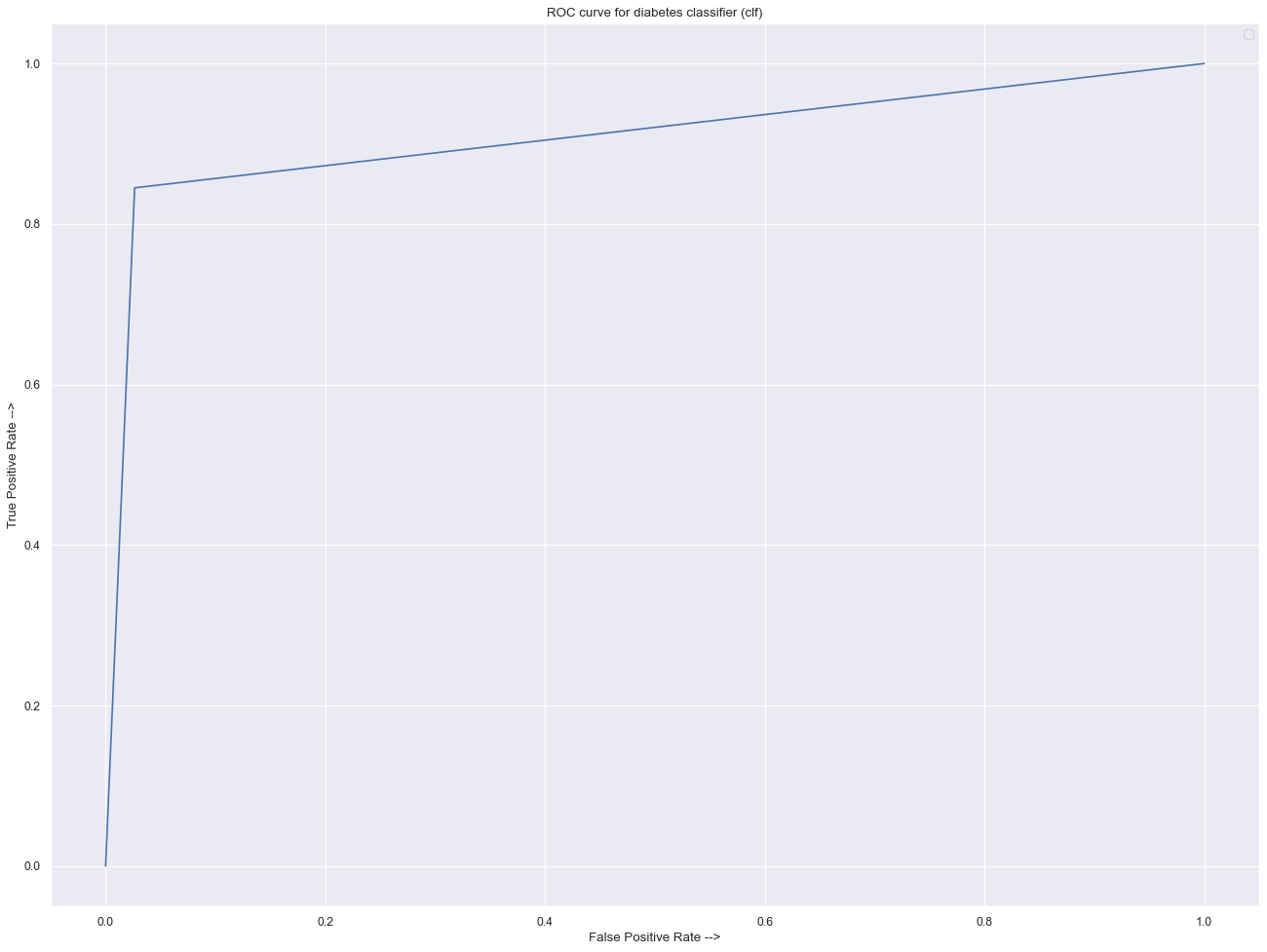
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Description** | **Precision** | **recall** | **fi-score** | **SUPPORT** |
| 0 | 0.94 | 0.97 | 0.96 | 2688 |
| 1 | 0.92 | 0.85 | 0.89 | 1020 |
| accuracy |  |  | 0.94 | 3708 |
| macro avg | 0.93 | 0.91 | 0.92 | 3708 |
| weighted avg | 0.94 | 0.94 | 0.94 | 3708 |

Below is a distribution plot showing the true label and the predicted label

****

**ROC CURVE**

We prepared a ROC CURVE using the confusion matrix True label and predicted label. The result is as shown below. It ranges between values (0.8 – 0.9). This means that our classifier (Random forest Classifier) has high chances of distinguishing between the True positives and the False positives than false negatives and false positives.



**REFERENCES**

1. Soni, M. and Varma, D., 2020. Diabetes Prediction using Machine Learning Techniques. [online] Ijert.org. Available at: <https://www.ijert.org/diabetes-prediction-using-machine-learning-techniques> [Accessed 23 February 2022].
2. Diabetes, T. and Health, N., 2017. Type 2 Diabetes | NIDDK. [online] National Institute of Diabetes and Digestive and Kidney Diseases. Available at: <https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/type-2-diabetes> [Accessed 23 February 2022].
3. Saeedi, P., Petersohna, a., Salpeaa, P., Malanda, B., Karurangaa, S., Unwinb, N., Colagiuric, S., Guariguatad, L., Motalae, A., Ogurtsovaf, K., E. Shawg, J., Bright, D. and William, R., 2019. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. [online] science direct. Available at: <https://www.sciencedirect.com/science/article/pii/S0168822719312306#!> [Accessed 23 February 2022].
4. Kumar, N., Das, N., Gupta, D., Gupta, K. and Bindra, J., 2021. *Efficient Automated Disease Diagnosis Using Machine Learning Models*. [online] The Jounal of Healthcare Engineering. Available at: <https://www.hindawi.com/journals/jhe/> [Accessed 11 March 2022].
5. Butt, Umair Muneer, E. D., Ali, Mubashir, L. S., Anees Sherazi, H. F. H. B., & Hafiz Husnain Raza. (2021). *Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications*. Publishing Open Access research journals & papers | Hindawi. <https://www.hindawi.com/journals/jhe/2021/9930985/>