Project Based Expremental Learning Program

SB3001-NAAN MUDHALVAN REPORT

Project name: Ai based diabetes prediction using machine learning

Project id: Proj-212176 team-1

Done by..

P.Karthick

950621104037

INDEX

1.ABSTRACT

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes is a disease caused due to the increase level of blood glucose. This high blood glucose produces the symptoms of frequent urination, increased thirst, and increased hunger. Diabetes is a one of the leading cause of blindness, kidney failure, amputations, heart failure and stroke. When we eat, our body turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a key to open our cells, to allow the glucose to enter and allow us to use the glucose for energy. But with diabetes, this system does not work. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. The algorithms like K nearest neighbour, Logistic Regression, Random forest, Support vector machine and Decision tree are used. The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for predicting the diabetes. Diabetes is an illness caused because of high glucose level in a human body. Diabetes should not be ignored if it is untreated then Diabetes may cause some major issues in a person like: heart related problems, kidney problem, blood pressure, eye damage and it can also affects other organs of human body. Diabetes can be controlled if it is predicted earlier. To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying, Various Machine Learning Techniques. Machine learning techniques Provide better result for prediction by constructing models from datasets collected from patients. In this work we will use Machine Learning Classification and ensemble techniques on a dataset to predict diabetes. Which are K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB) and Random Forest (RF). The accuracy is different for every model when compared to other models. The Project work gives the accurate or higher accuracy model shows that the model is capable of predicting diabetes effectively. Our Result shows that Random Forest achieved higher accuracy compared to other machine learning techniques.

Keywords : Machine Learning, Diabetes, Decision tree, K nearest neighbour, Logistic Regression, Support vector Machine, Accuracy.

2.INTRODUCTION

Diabetes mellitus is one of the non-communicable diseases that pose a threat to human health. It has become a major global health problem. It is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin which it produces. It is found that diabetes causes blindness, amputation and kidney failure. Lack of awareness about diabetes, insufficient access to health services and essential medicines can lead to the above mentioned complications. According to a study by the World Health Organization (WHO), number of diabetic patients will raise to 552 million by 2030, which means that one in 10 adults will have diabetes by 2030. In 2014, the global prevalence of diabetes was estimated to be 9 % among adults aged 18+ years [1]. WHO insisted with an alarm that Diabetes is the 7th leading cause of death in the world. In 2012, an estimated 1.5 million deaths were directly caused by diabetes. Total deaths due to diabetes are projected to rise by more than 50 % in the next 10 years. Moreover, the International Diabetes Federation said that nearly 52 % of Indians are not aware that they are suffering from high blood sugar. More than 62 million diabetic individuals are currently diagnosed with the disease. It is predicted that, by 2030 diabetes mellitus may affect up to 79.4 million individuals in India. A nation-wide study, conducted by the Indian Council of Medical Research`s INDIAB (India Diabetes) has confirmed that one out of 10 people in Tamil Nadu is affected by diabetes, and every two persons with age group of 25 are in the pre-diabetic stage. It is stated that 14.8 per cent of urban population and 11 per cent of rural population of Tamil Nadu are affected by diabetes. Madras Diabetes Research Foundation suggested that about 42 lakh individuals have diabetes and 30 lakh people are in pre-diabetes stage. At least, 1,000 people avail treatment for diabetes out of the 12,000 outpatients who visit Rajiv Gandhi Government General Hospital (RGGGH), a leading Government hospital in Chennai.

3. Review of the Literature

- Raghupathi et al. [2] had done a review on big data analytics in healthcare. The promises and potentials of big data in healthcare were pointed out and an architectural framework and methodology for applying big data in healthcare were outlined. The advantages of using big data analytics in healthcare were elaborated and the available platforms and tools for applying big data analytics in healthcare were presented.

Aiswarya Iyer et al. [3] used Decision Tree and Naïve Bayes algorithms for the prediction of diabetes in pregnant women. Tenfold cross validation was done to prepare training and test data and the J48 algorithm was employed on the dataset using WEKA on the Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases.

- The authors concluded that both algorithms were efficient for the diagnosis of diabetic and Naïve Bayes technique resulted in least error rate. A.A. Aljumah et al. [4] suggested a predictive analysis of diabetic treatment using a regression based data mining technique.

- Oracle Data Miner (ODM) tool was deployed for predicting diabetics and support vector machine algorithm was employed for experimental analysis on Datasets of Non Communicable Diseases (NCD) risk factors in Saudi Arabia. Mohammed et al. [5] presented a review of existing applications of the Map Reduce programming framework and its implementation platform Hadoop in clinical big data and related medical health informatics. N.M. Saravana Kumar et al. [6] presented Predictive Analysis System Architecture with various stages of data mining. Prediction was carried out in Hadoop / Map Reduce environment. Predictive Pattern matching system was deployed to compare the analyzed threshold value with the obtained value. Saumya et al. [7] applied analytical techniques to reduce the hospital readmission of diabetic patients. In the proposed methodology, Hive was used as the preprocessing tool and R Hadoop as the analytics and predictive modeling tool.

- Classification was done using Logistic Regression, SVM, KNN and Decision Tree methods. Miss-classification error rates were also calculated. D. Peter Augustine [8] presented a concept paper on analyzing the data flowing from health monitoring devices. The present status of healthcare in India was presented. The application of Hadoop's map reduce in healthcare data was expounded. An interface HIPI (Hadoop Image Processing Interface) in Hadoop environment was also explicated. Sadhana et al. [9] analyzed the Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases data set using a proposed architecture which comprised of Hive and R. The raw data (CSV file) was given as input to the Hive for analyzing and partitioning. The output file was passed to the through R system for statistical computing. MuniKumar N et al. [10] has pointed out the massive shortage of proper healthcare facilities and addressed the concern to provide greater access to primary healthcare services in rural India. The ability of Big Data analytics in processing huge volumes of data in real-time situations to turn the dream of Swachh

Bharath (Clean &Healthy India) into reality was explicated. AditiBansal et al. [11] proposed an architecture consisting of Dynamic Hadoop Slot Allocation (DHSA) which used the slot based resource model. Two more alternatives for DHSA namely, Pool Independent DHSA (PIDHSA) and Pool dependent DHSA (PDDHSA) were also presented. It was found that DHSA focused on the maximum utilization of slots by allocating map (or reduce) slots to map and reduce tasks dynamically. K. Sharmila et al. [12] presented a survey paper on the advancement in the field of data mining, the latest adoption of Hadoop platform, deployment of big data algorithms and consequently the open challenges in the Indian medicinal data set.

4.literature survey

KM Jyothi Rani Proposed a system for predicting diabetes based on Machine learning algorithms. In this paper they have used the dataset which contains 9 features and 2000 entries out of which outcome describes 0 means no diabetes, 1 means diabetes. They have used 5 machine learning algorithms in this paper out of these 5 algorithms Decision Tree algorithm provides training accuracy as 98% and testing accuracy as 99% Raja Krishnamoorthi proposed a diabetes healthcare disease prediction framework using machine learning techniques. The dataset contains 768 rows and 9 columns and 90% of the data is used for training and 10% used for the testing purpose and they performed hyper-parameter tuning to evaluate the Machine Learning models and used to increase the accuracy. Out of 5 algorithms best one is identified and hyper parameter tuning has been applied to provide better accuracy as a result of 86% Desmond Bala Bisandu proposed a system for diabetes prediction using data mining techniques. In this paper there are 5 parameters based on which diabetes is predicted and data is pre-processed to remove noise and to remove null values and classification and prediction was done using Naive Bayes Classifier and efficiency was around 95% B. Suvarnamukhi proposed a big data processing system which uses machine learning techniques for predicting diabetes. Due to rapid increase in technology the data is stored in the form of electronic records (EHR) and this data is processed using big data and for prediction of diabetes ELM is used and compared with other algorithms and diabetes which is predicted of 3 types Mitush Soni proposed machine learning algorithms for providing better accuracy in diabetes prediction. In this paper the dataset contains 500 negative outcomes means no diabetes and 268 positive outcomes means diabetes and For Predicting accurately they have used 6 machine learning algorithms and among these 6 algorithms random forest algorithm predicts with 77% accuracy The dataset consists of 2500 entries and 15 attributes and 768 items used for testing and they have used 5 algorithms out of which support vector machine provides 77% accuracy. Abdullah A. Aljumah and M.G Ahmad proposed a data mining application to predict diabetes in young and old patients using regression-based

mining technique. The dataset is used is a NCD risk factor report from Ministry of health report, Saudi Arabia and using data mining analysis on data set they have predicted the effectiveness in young and old group for different treatments. Salliah Shafia and Prof. Gufran Ahmad Ansari designed a model for Early Prediction of Diabetes Disease & Classification of Algorithms Using Machine Learning Approach. this research uses the WEKA tools to predict diabetes in patients from Pima India Diabetes Data Set consists of 7 attributes and 767 entries and in this paper, they have used 3 classification algorithms out of which Naïve bayes provides 74% accuracy. R M Anjana prepared a report on Prevalence of diabetes and prediabetes in urban and rural India. In this report they conducted a survey on urban and rural parts of india to estimate prevalence of diabetes and prediabetes and in the report, Chandigarh was found to be have highest diabetes percentage

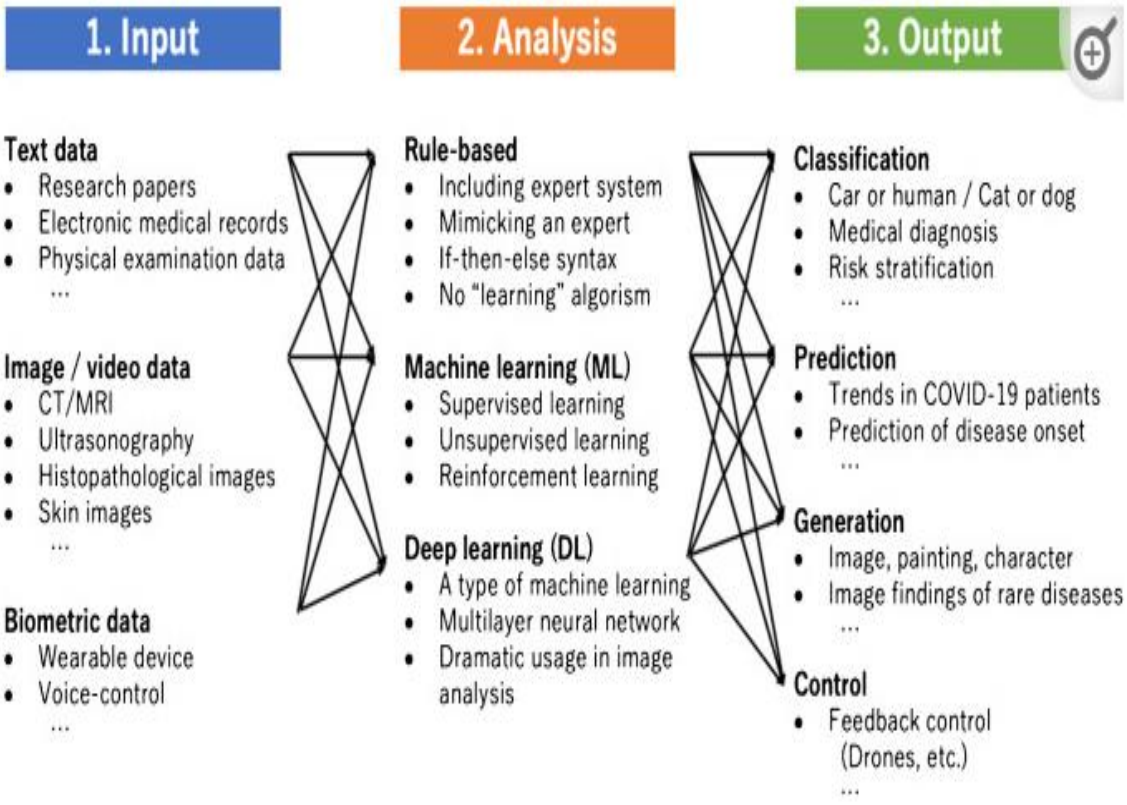
5.**Problem definition ,design thinking,innovation & problem solving**

Problem definition:

Customer problem statement template:

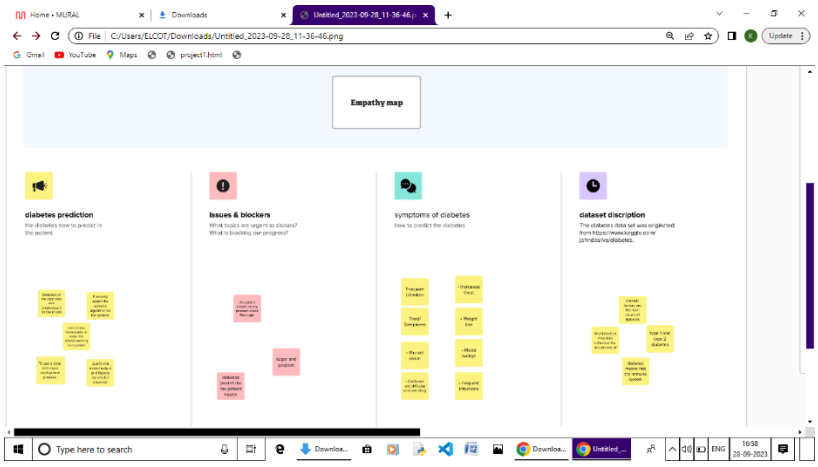
- We discuss the use of AI in medicine for diabetes, specifically in medical devices. The first AI-based medical device, BodyGuardian, was cleared by the US Food and Drug Administration (FDA) in 2012 when approval was given to a patch-like electrocardiogram equipped with an AI-based arrhythmia detection algorithm. Since then, the regulations on programmed medical devices, including AI, have advanced in various countries, including the USA, Europe, China, and Japan. Thanks to the outstanding development of deep learning technology and advancements in clinical applications these days, the number of approved AI-based medical devices has dramatically increased in both the USA and Europe in the past few years
- Currently, there are dozens of FDA-cleared AI-based medical devices using AI/machine learning technology. While most of these approvals are linked to radiology, cardiology, and oncology, three AI-based medical devices are related to diabetes management [5•]. In Japan, 12 types of AI-based medical devices have been approved as of 2020. However, all of them are for image analysis concerning radiology and diagnostic imaging, and there are no such medical devices approved for diabetes care.Efforts towards the clinical application of AI in the diagnosis and treatment of diabetes are mainly categorized into four areas: (1) automatic retinal

screening, (2) clinical diagnosis support, (3) patient self-management tools, and (4) risk stratification [6]. The first category is automatic retinal screening, an AI technology that automatically interprets the presence or absence of diabetic retinopathy—an important complication of diabetes—from fundus images. An example of this technology is the IDx-DR device manufactured by Digital Diagnostics Inc., approved by the FDA in 2018 for its high diagnostic performance by clinical trials [7]. Using this AI device, patients can be diagnosed with diabetic retinopathy or not without professional judgment from an ophthalmologist. Then, primary physicians can choose to have the patients with their fundus images see an ophthalmologist or re-examine the IDx-DR device 12 months later. This device facilitates the screening and diagnosis of diabetic retinopathy, especially in rural communities where patients have difficulties accessing an ophthalmologist .



Empathize&Discover

Empathy map canvas:



Dataset Description:

The diabetes data set was originated from https://www.kaggle.com/johndasilva/diabetes. Diabetes dataset containing 2000 cases. The objective is to predict based on the measures to predict if the patient is diabetic or not

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	62	35	0	33.6	0.127	47	1
1	0	84	82	31	125	38.2	0.233	23	0
2	0	145	0	0	0	44.2	0.630	31	1
3	0	135	68	42	250	42.3	0.365	24	1
4	1	139	62	41	480	40.7	0.536	21	0

- The diabetes data set consists of 2000 data points, with 9 features each.
- "Outcome" is the feature we are going to predict, 0 means No diabetes, 1 means diabetes.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               2000 non-null   int64
 1   Glucose                   2000 non-null   int64
 2   BloodPressure             2000 non-null   int64
 3   SkinThickness             2000 non-null   int64
 4   Insulin                   2000 non-null   int64
 5   BMI                       2000 non-null   float64
 6   DiabetesPedigreeFunction  2000 non-null   float64
 7   Age                       2000 non-null   int64
 8   Outcome                   2000 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 140.8 KB
```

METHODOLOGY :

In this section we shall learn about the various classifiers used in machine learning to predict diabetes. We shall also explain our proposed methodology to improve the accuracy. Five different methods were used in this paper. The different methods used are defined below. The output is the accuracy metrics of the machine learning models. Then, the model can be used in prediction.

Symptoms of Diabetes:

• Frequent Urination • Increased thirst
• Tired/Sleepiness • Weight loss
• Blurred vision• Mood swings
• Confusion and difficulty concentrating • frequent infections

6.Data cleaning and testing:

What is Diabetes?

Diabetes is a disease that occurs when the blood glucose level becomes high, which ultimately leads to other health problems such as heart diseases, kidney disease, etc. Diabetes is caused mainly due to the consumption of highly processed food, bad consumption habits, etc. Accordingto WHO, the number of people with

diabetes has been increased over the years.

Prerequisites

- Python 3.+
- Anaconda (Scikit Learn, Numpy, Pandas, Matplotlib, Seaborn)
- Jupyter Notebook.
- Basic understanding of supervised machine learning methods: specifically classification.

Data Preparation

As a Data Scientist, the most tedious task which we encounter is the acquiring and the preparationof a data set. Even though there is an abundance of data in this era, it is still hard to find a suitable data set that suits the problem you are trying to tackle. If there aren't any suitable data sets to be found, you might have to create your own.

Data Exploration

When encountered with a data set, first we should analyze and "**get to know**" the data set. This step is necessary to familiarize with the data, to gain some understanding of the potential featuresand to see if data cleaning is needed.

First, we will import the necessary libraries and import our data set to the Jupyter notebook. Wecan observe the mentioned columns in the data set.

Import the package:

First I have import the my dataset of the packages.

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.preprocessing import scale, StandardScaler

from sklearn.model_selection import train_test_split, GridSearchCV, cross_val_score

import warnings

warnings.simplefilter(action = "ignore")

- Import all packages using diabetes dataset.

First I have import numpy:

- Import Numpy as np is an incredibly powerful and versatile numerical computing library for Python and it is capable of numerous mathematical operations. When you import Numpy as 'np', you unlock the full potential of Numpy.
- The numpy used in the numerical value can be calculated and the functions can be used in the package was imported.
- The numpy is install in the command prompt using this command is **pip install numpy.**then you will import the package is **import numpy as np.**
- Numpy can be used in the numerical python the extension is np.
- The numpy using array can be performed in the numpy as np.the dataset will be numerical value can be inserted for the dataset the all dataset print the array the numpy was imported and all the functions can be performed.

Import the pandas:

The **import pandas** portion of the code tells Python to bring the pandas data analysis library into your current environment.

The **as pd** portion of the code then tells Python to give pandas the alias of **pd**. This allows you to use pandas functions by simply typing pd.function_name rather than pandas.function_name.

Once you've imported pandas, you can then use the functions built in it to create and analyze data.

The pandas was used in the dataframe and series all pandas functions can be used in the pandas as pd.

Import the matplotlib as plt:

importing the matplotlib module, defines x and y values for a plots, plots the data using the plot() function and it helps to display the plot by using the show() function . The plot() creates a line plot by connecting the points defined by x and y values.

Import the seaborn sns:

The **import seaborn** portion of the code tells Python to bring the Seaborn library into your current environment.

The **as sns** portion of the code then tells Python to give Seaborn the alias of **sns**. This allows you to use Seaborn functions by simply typing sns.function_name rather than seaborn.function_name.

Once you've imported Seaborn, you can then use the functions built in it to quickly visualize data.

Code:

```
diabetes = pd.read_csv('D:diabetes/diabetes.csv')
diabetes.columns
```

output:

```
Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
dtype='object')
```

Important: It should be noted that the above data set contains only limited features, where as in reality numerous features come into play.

We can examine the data set using the pandas' **head()** method.

- **diabetes.head(700)**

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	
0	6	148	72	35	0	33.6	0.627	50	1	
1	1	85	66	29	0	26.6	0.351	31	0	
2	8	183	64	0	0	23.3	0.672	32	1	
3	1	89	66	23	94	28.1	0.167	21	0	
4	0	137	40	35	168	43.1	2.288	33	1	
...	...	...	...	...	...	...	...	...	...	...
755	1	128	88	39	110	36.5	1.057	37	1	
756	7	137	90	41	0	32.0	0.391	39	0	

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
757		0	123	72	0	0	36.3	0.258	52	1
758		1	106	76	0	0	37.5	0.197	26	0
759		6	190	92	0	0	35.5	0.278	66	1

760 rows × 9 columns

We can find the dimensions of the data set using the panda Dataframes' 'shape' attribute.

- **diabetes.shape**

output

(768, 9)

We can observe that the data set contain 768 rows and 9 columns. '*Outcome*' is the column which we are going to predict, which says if the patient is diabetic or not. 1 means the person is diabetic and 0 means a person is not. We can identify that out of the 768 persons, 500 are labeled as 0 (non-diabetic) and 268 as 1 (diabetic)

diabetes ['Outcome'].value_counts()

Outcome
0 500
1 268
Name: count, dtype: int64

diabetes.groupby('Outcome').mean()

output:

Outcome	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	3.298000	109.980000	68.184000	19.664000	68.792000	30.304200	0.429734	31.190000
1	4.865672	141.257463	70.824627	22.164179	100.335821	35.142537	0.550500	37.067164

Missing or Null Data points

We can find any missing or null data points of the data set (if there is any) using the following pandas function.

diabetes.isnull().sum()
diabetes.isna().sum()

We can observe that there are no data points missing in the data set. If there were any, we should deal with them accordingly.

df.isnull().sum()

output

Pregnancies	0
Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0
DiabetesPedigreeFunction	0
Age	0
Outcome	0

dtype: int64

Unexpected Outliers

When analyzing the histogram we can identify that there are some outliers in some columns. We will further analyze those outliers and determine what we can do about them.

Blood pressure:

By observing the data we can see that there are 0 values for blood pressure. And it is evident that the readings of the data set seem wrong because a living person cannot have a diastolic blood pressure of zero. By observing the data we can see 35 counts where the value is 0.

```
print("Total : ", diabetes[diabetes.BloodPressure == 0].shape[0])Total:35 print(diabetes
[diabetes.BloodPressure == 0].groupby('Outcome')['Age'].count())

Outcome
0    19
1    16
Name: Age, dtype: int64
```

Plasma glucose levels:

Even after fasting glucose levels would not be as low as zero. Therefore zero is an invalid reading. By observing the data we can see 5 counts where the value is 0.

```
print("Total : ", diabetes[diabetes.Glucose == 0].shape[0])Total:5
print(diabetes[diabetes.Glucose == 0].groupby('Outcome')['Age'].count())Total:5

Outcome
0    3
1    2
Name: Age, dtype: int64
```

Skin Fold Thickness:

For normal people, skin fold thickness can't be less than 10 mm better yet zero. Total count where value is 0: 227.

```
print("Total : ", diabetes[diabetes.SkinThickness == 0].shape[0])Total :227
print(diabetes[diabetes.SkinThickness == 0].groupby('Outcome')['Age'].count())

Outcome
0    139
1     88
Name: Age, dtype: int64
```

BMI: Should not be 0 or close to zero unless the person is really underweight which could be life-threatening.

```
print("Total : ", diabetes[diabetes.BMI == 0].shape[0])Total :11  print(diabetes[diabetes.BMI
== 0].groupby('Outcome')['Age'].count())

Outcome
0    9
1    2
Name: Age, dtype: int64
```

Insulin: In a rare situation a person can have zero insulin but by observing the data, we can find that there is a total of 374 counts.

```
print("Total : ", diabetes[diabetes.Insulin == 0].shape[0])Total :374
print(diabetes[diabetes.Insulin == 0].groupby('Outcome')['Age'].count())

Outcome
0    236
1    138
Name: Age, dtype: int64
```

Here are several ways to handle invalid data values :

1. Ignore/remove these cases: This is not actually possible in most cases because that would mean losing valuable information. And in this case “skin thickness” and “insulin” columns mean to have a lot of invalid points. But it might work for “BMI”, “glucose ”and “blood pressure” data points.

2. Put average/mean values: This might work for some data sets, but in our case putting a mean value to the blood pressure column would send a wrong signal to the model.

3. Avoid using features: It is possible to not use the features with a lot of invalid values for the model. This may work for “skin thickness” but it's hard to predict that.

By the end of the data cleaning process, we have come to the conclusion that this given data set is incomplete. Since this is a demonstration for machine learning we will proceed with the given data with some minor adjustments.

We will remove the rows which the “BloodPressure”, “BMI” and “Glucose” are zero.

```
diabetes_mod = diabetes[(diabetes.BloodPressure != 0) & (diabetes.BMI != 0) &
(diabetes.Glucose != 0)]
print(diabetes_mod.shape)

output:
(724, 9)
```

```
diabetes_scores=lof.negative_outlier_factor_
np.sort(diabetes_scores)[0:30]
```

output

```
array([-3.30445978, -2.48884101, -2.28758733, -2.10500141, -2.05369597,
    -2.02885837, -2.01096252, -2.00720763, -1.98655427, -1.95338702,
    -1.91601291, -1.88815728, -1.8134966 , -1.80857804, -1.74187579,
    -1.73154315, -1.71639102, -1.71372358, -1.67587303, -1.64102097,
    -1.63498158, -1.62215678, -1.61146741, -1.59344933, -1.54582494,
    -1.54285259, -1.51413703, -1.49974262, -1.49619189, -1.48877158])
```

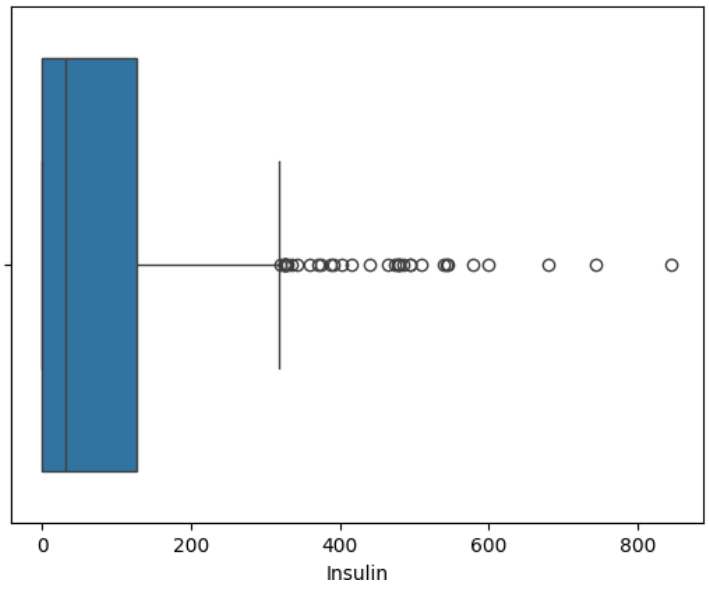
```
for feature in df:

    Q1 = df[feature].quantile(0.25)
    Q3 = df[feature].quantile(0.75)
    IQR = Q3-Q1
    lower = Q1- 1.5*IQR
    upper = Q3 + 1.5*IQR

    if df[(df[feature] > upper)].any(axis=None):
        print(feature,"yes")
    else:
print(feature, "no")
```

output
Pregnancies yes
Glucose no
BloodPressure yes
SkinThickness yes
Insulin yes
BMI yes
DiabetesPedigreeFunction yes
Age yes
Outcome no

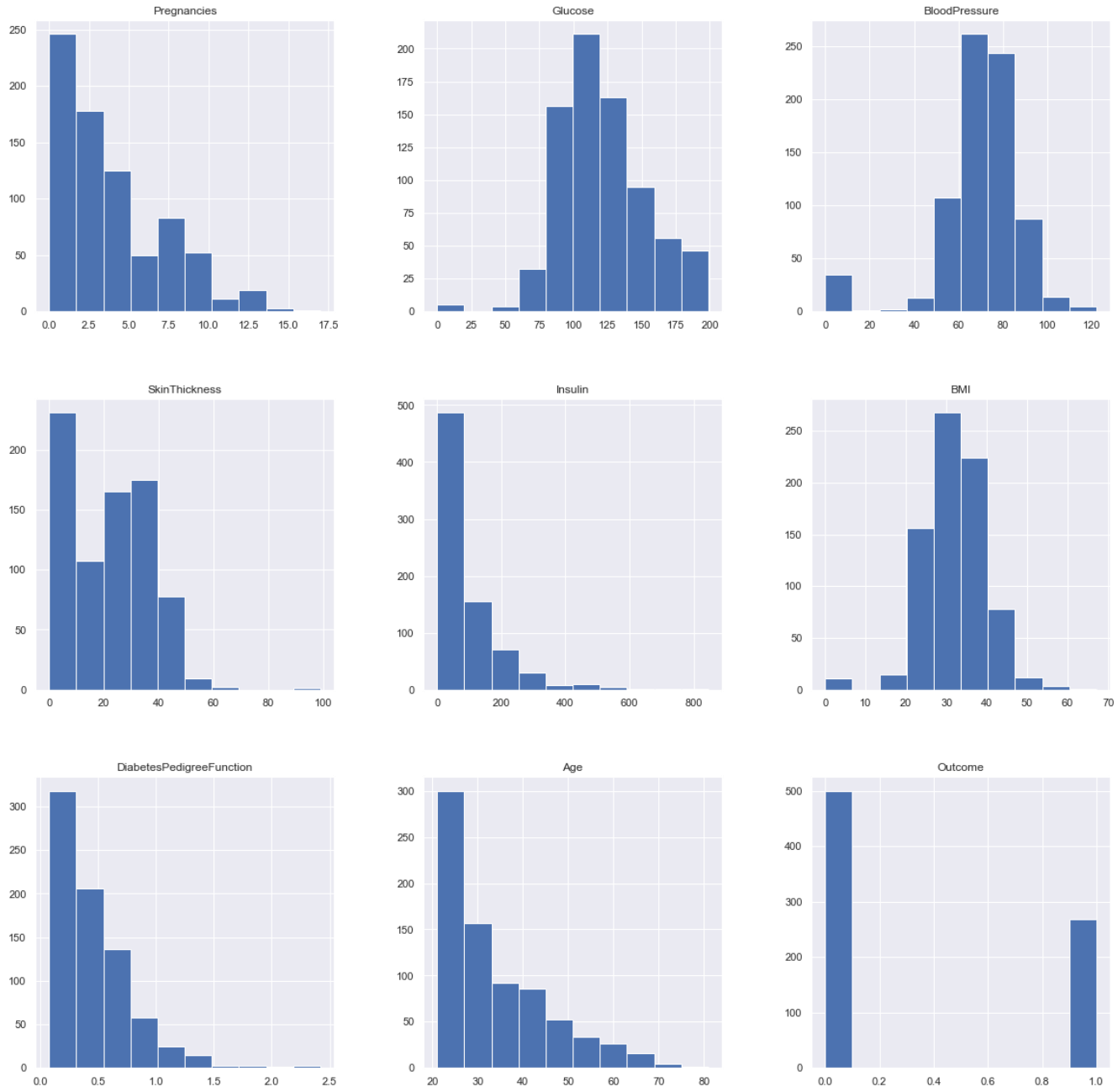
```
import seaborn as sns
sns.boxplot(x = df["Insulin"]);
```



7.Data Visualization

Plotting the data distribution plots before removing null values

```
p = diabetes_df.hist(figsize = (20,20))
```



Inference:So here we have seen the distribution of each features whether it is dependent data or independent data and one thing which could always strike that **why do we need to see the distribution of data?** So the answer is simple it is the best way to start the analysis of the dataset as **it shows the occurrence of every kind of value in the graphical structure which in turn lets us know the range of the data.**

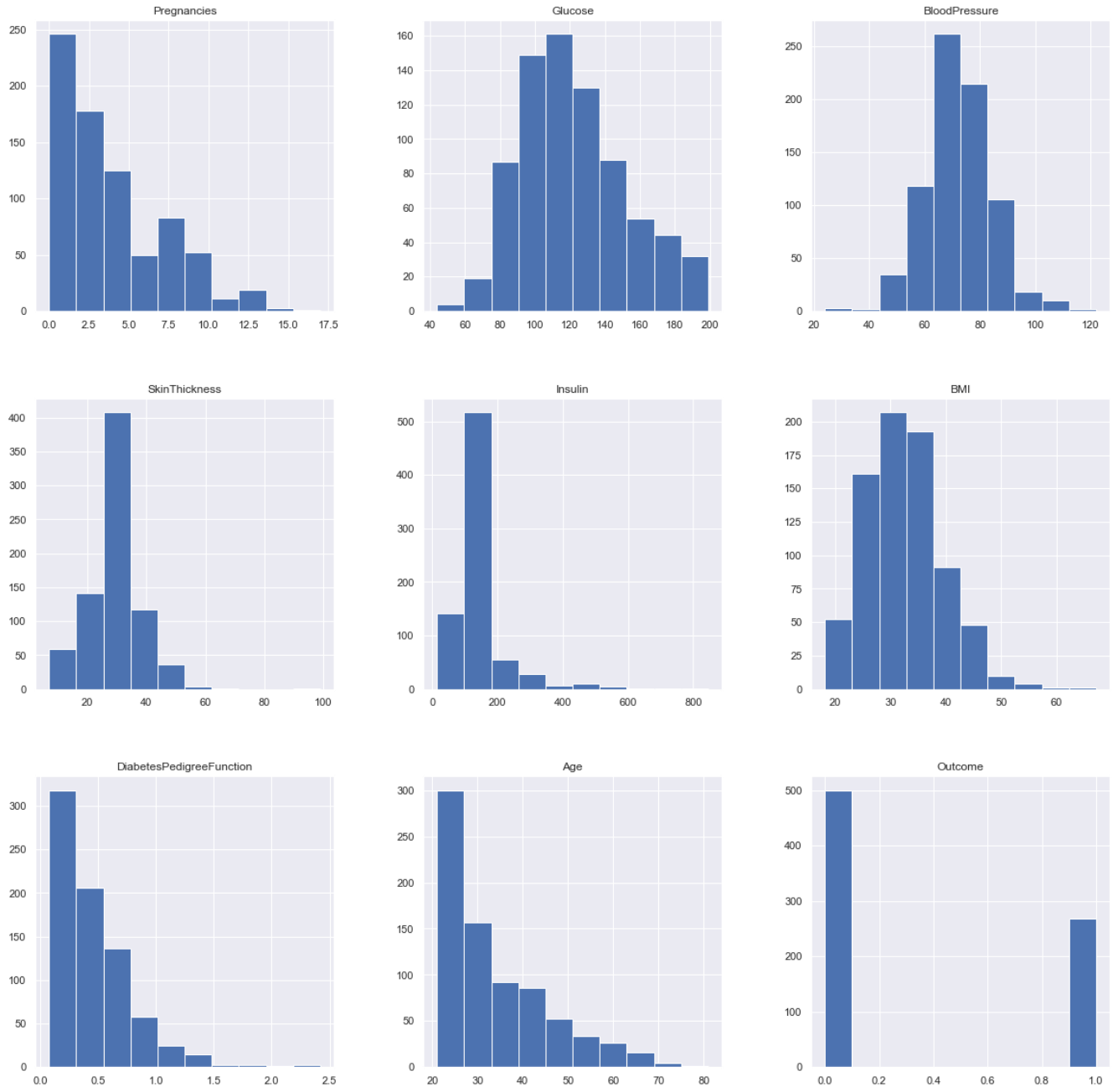
now we will be imputing the mean value of the column to each missing value of that particular column.

```
diabetes_df_copy['Glucose'].fillna(diabetes_df_copy['Glucose'].mean(),
inplace = True)
diabetes_df_copy['BloodPressure'].fillna(diabetes_df_copy['BloodPressure'].
mean(), inplace = True)
diabetes_df_copy['SkinThickness'].fillna(diabetes_df_copy['SkinThickness'].
median(), inplace = True)
diabetes_df_copy['Insulin'].fillna(diabetes_df_copy['Insulin'].median(),
inplace = True)
diabetes_df_copy['BMI'].fillna(diabetes_df_copy['BMI'].median(),  inplace  =
True)
```

Plotting the distributions after removing the NAN values.

```
p = diabetes_df_copy.hist(figsize = (20,20))
```

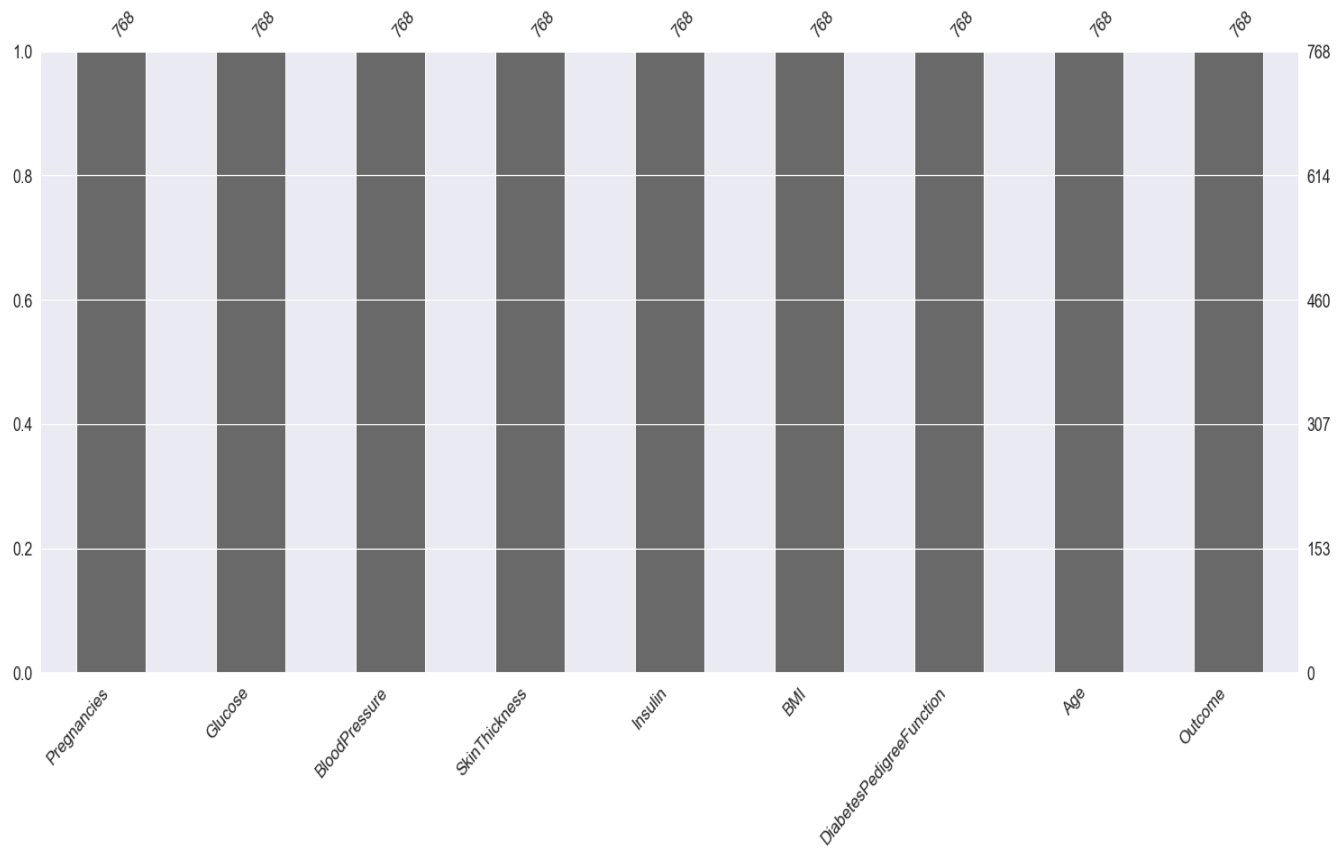
Output:



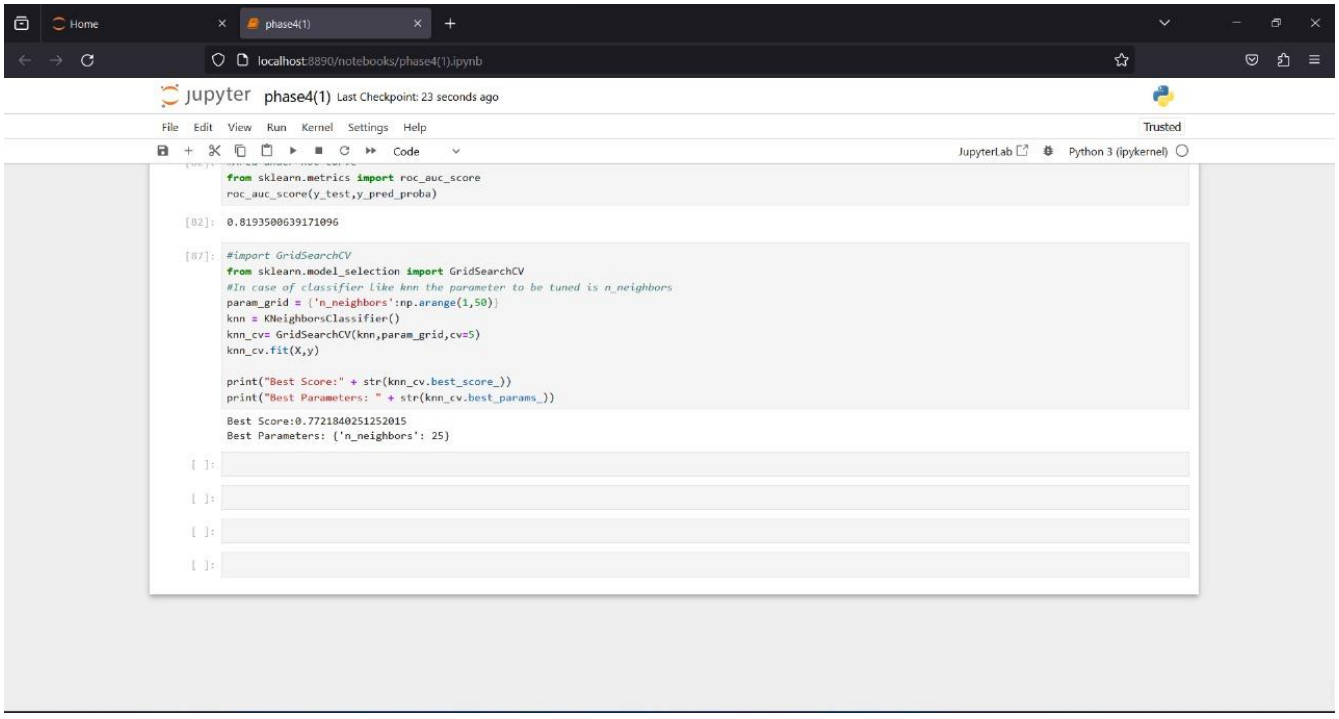
Inference:Here we are again using the hist plot to **see the distribution of the dataset** but this time we are using this visualization to see the changes that we can see after those null values are removed from the dataset and we can clearly see the difference **for example –** In age column after removal of the null values, **we can see that there is a spike at the range of 50 to 100 which is quite logical as well.**

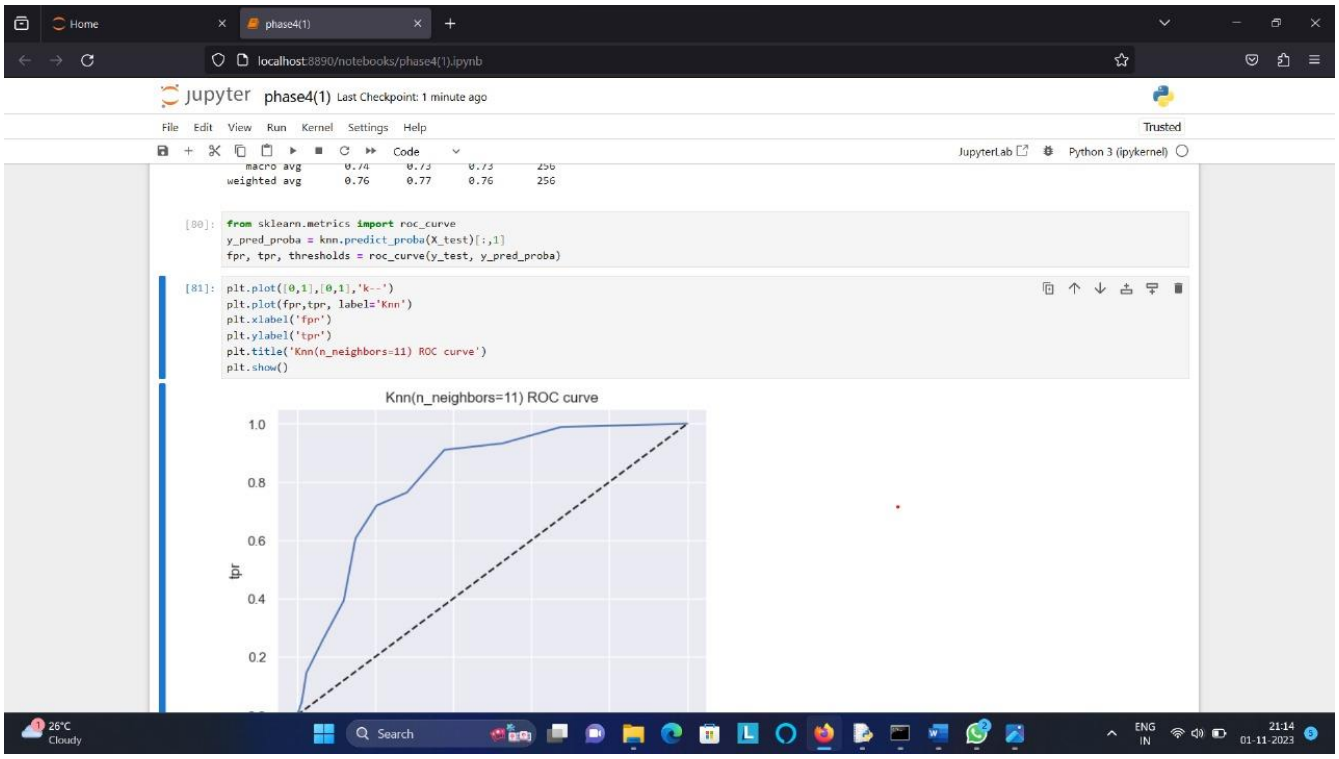
Plotting Null Count Analysis Plot

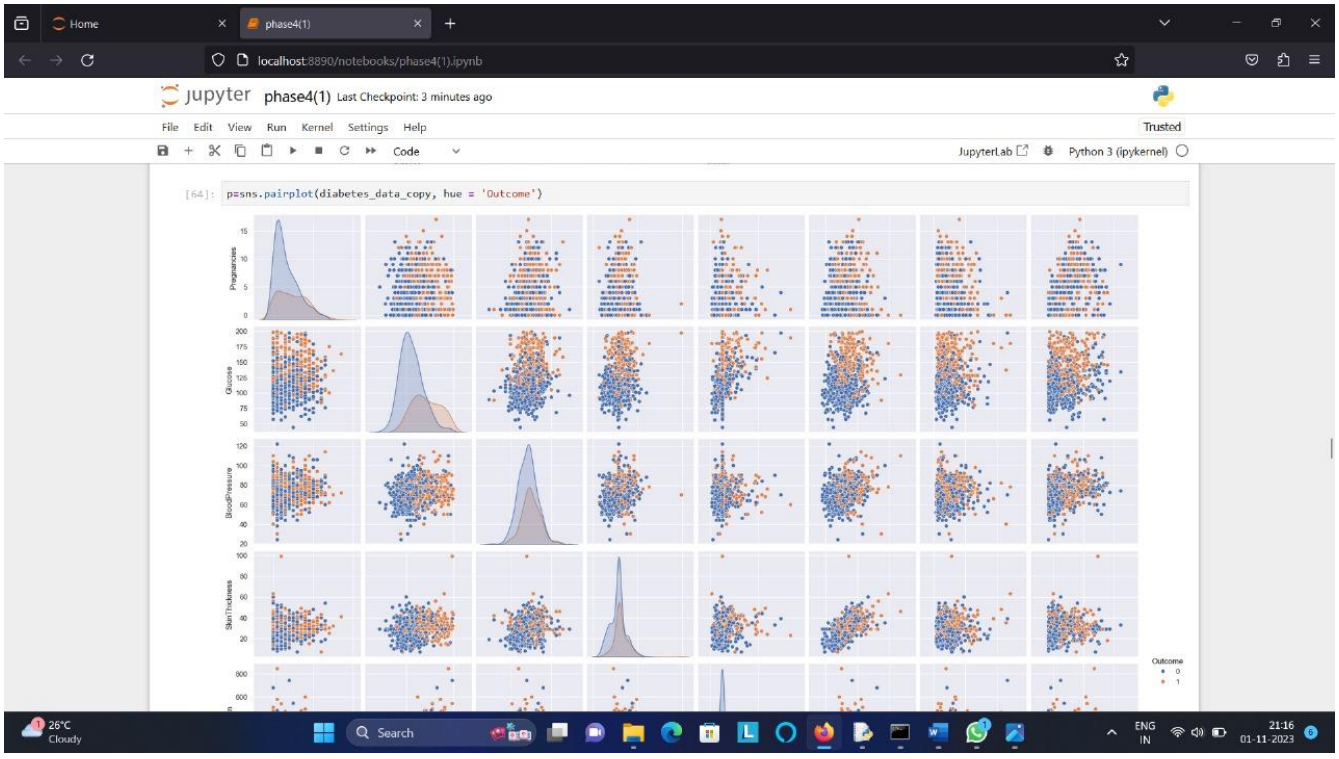
p = msno.bar(diabetes_df)

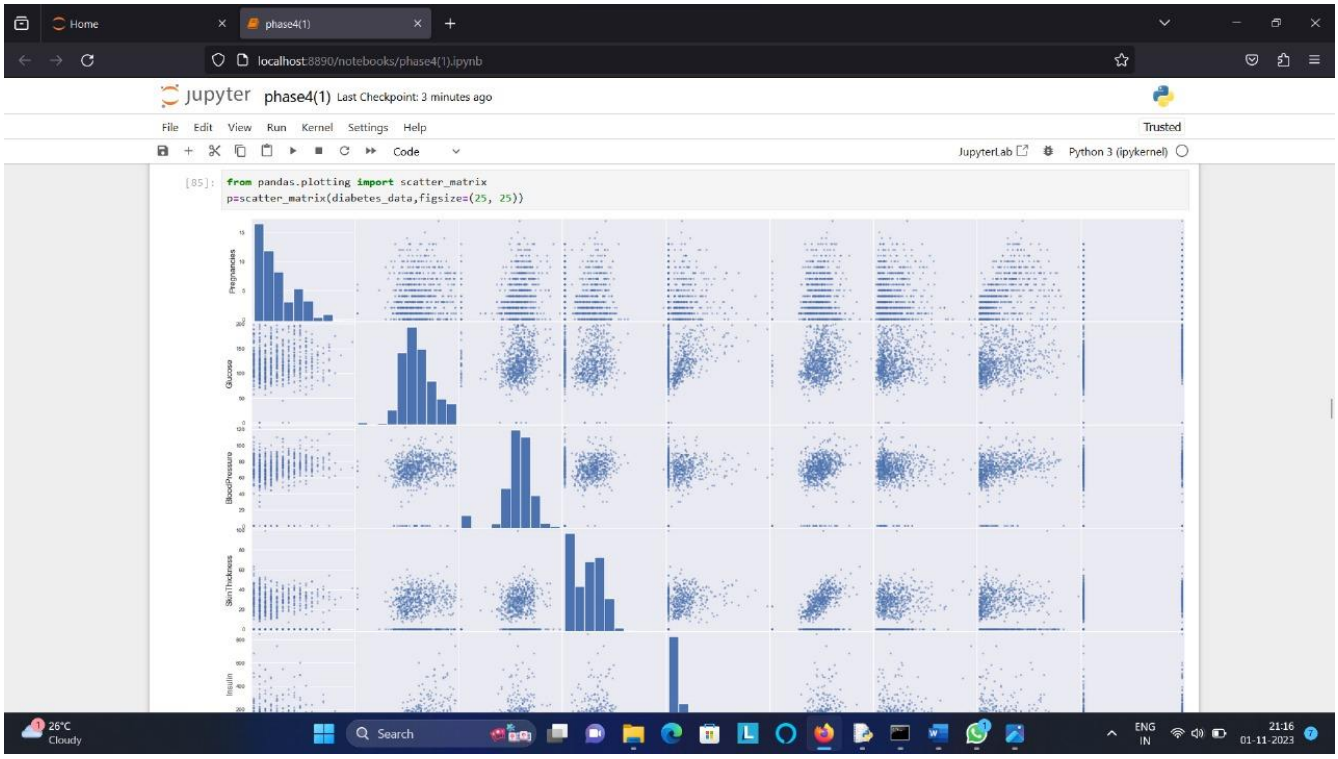


8.Code and output screenshot:




Home
phase4(1)
localhost:8890/notebooks/phase4(1).ipynb
jupyter phase4(1) Last Checkpoint: 1 minute ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
macro avg 0.74 0.73 0.73 256
weighted avg 0.76 0.77 0.76 256
[80]: from sklearn.metrics import roc_curve
y_pred_proba = knn.predict_proba(X_test)[:,1]
fpr, tpr, thresholds = roc_curve(y_test, y_pred_proba)
[81]: plt.plot([0,1],[0,1],'k--')
plt.plot(fpr,tpr, label='Knn')
plt.xlabel('fpr')
plt.ylabel('tpr')
plt.title('Knn(n_neighbors=11) ROC curve')
plt.show()
Knn(n_neighbors=11) ROC curve
tpr
1.0
0.8
0.6
0.4
0.2
26°C Cloudy
Search
ENG IN
21:14
01-11-2023


Home
phase4(1)
localhost:8890/notebooks/phase4(1).ipynb
jupyter phase4(1) Last Checkpoint: 3 minutes ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
[64]: p=sns.pairplot(diabetes_data_copy, hue = 'Outcome')
Pregnancies
Glucose
BloodPressure
SkinThickness
Outcome
0
1
26°C Cloudy
Search
ENG IN
21:16
01-11-2023



9.Conclusion:

After using all these patient records, we are able to build a machine learning model (random forest – best one) to accurately predict whether or not the patients in the dataset have diabetes or not along with that we were able to draw some insights from the data via data analysis and visualization.