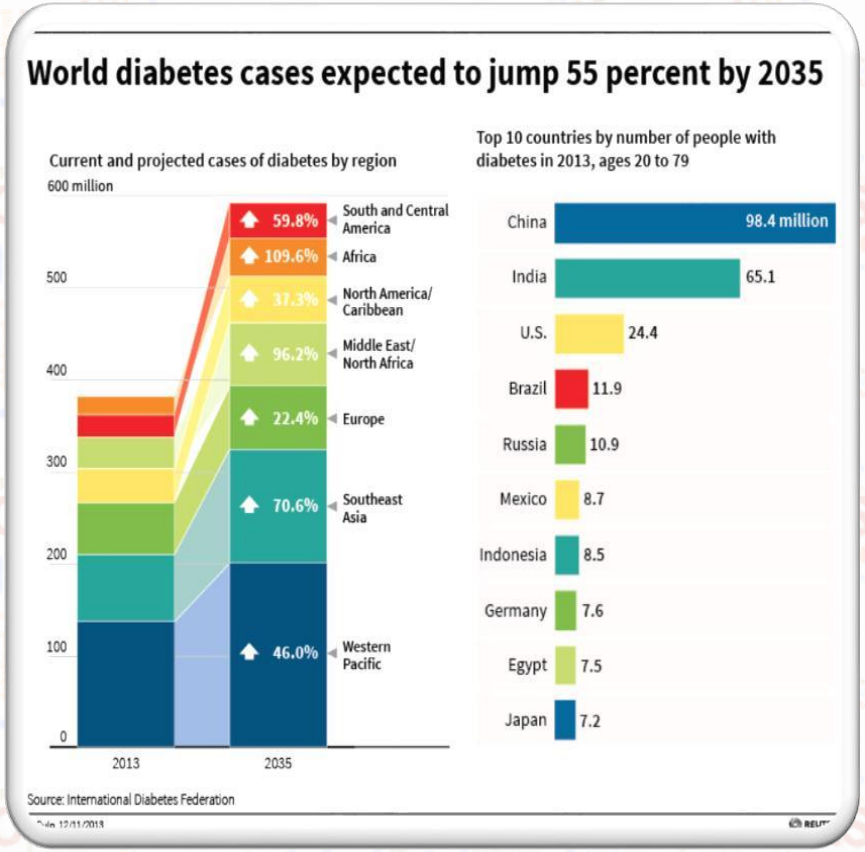
DIABETES DETECTION

ARAVINDA RAMAN J

- DATA SCIENCE AND ML INTERN
- EXPOSYS DATA LABS
- E-mail: aravindaraman04@gmail.com

INTRODUCTION

- Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Blood glucose is your main source of energy and comes from the food you eat.
- Insulin, a hormone made by the pancreas, helps glucose from food get into your cells to be used for energy. Sometimes your body doesn't make enough—or any—insulin or doesn't use insulin well. Glucose then stays in your blood and doesn't reach your cells.
- Over time, having too much glucose in your blood can cause health problems. Although diabetes has no cure, you can take steps to manage your diabetes and stay healthy.
- Diabetes is also creator of different kinds of diseases like heart attack, blindness etc.. The normal identifying process is that the patients need to visit a diagnostic center, consult a doctor and sit tight for a day or more to get their reports
- Cause of diabetes vary depending on the genetic makeup, family history, ethnicity, health etc..
- Diabetes & pre-diabetes is diagnosed by blood test.



STANDARD PROCESS: CRISP DM

- Crisp DM process is used to better understand the problem and give us better insight of whole process.
- CRISP DM(Cross Industry Standard Process For Data Mining) has six phases:

1. **Business/Research Understanding Phase**

- Determine the business objectives
- To assess the problem and determine data mining goals
- To come up with a strategy to meet goals and objectives

2. **Data Understanding Phase**

- Collect the data
- Assess and analyse the data

STANDARD PROCESS:CRISP DM

3. **Data Preparation Phase**

- Clean the data i.e. remove any missing values or outliers etc.
- Transform the data
- Select specific data for analysis

4. **Modeling**

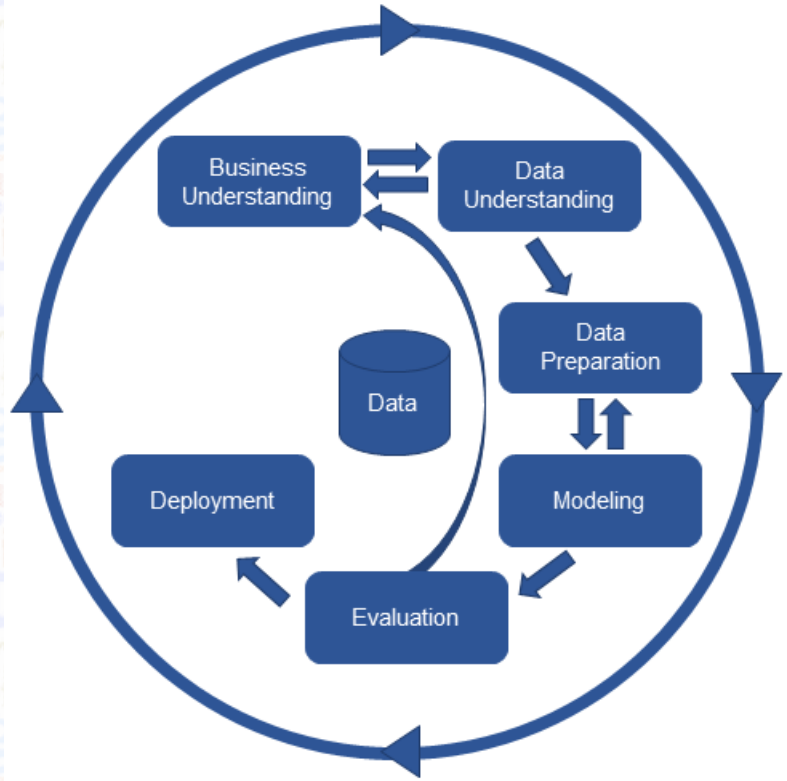
- Select appropriate data modelling technique

5. **Evaluation**

- Evaluate the model
- Calculate the accuracy and success rate of the model

6. **Deployment**

- Plan deployment
- Monitor Deployment
- Generate reports to test success of the model



BUSINESS UNDERSTANDING

What is the profound question?

To identify whether the patient is having diabetes or not?

Goal

Goal of this project is to identify the probability of diabetes in patients using data mining techniques.

Advantage of this project

The rules derived will be helpful for doctors to identify patients suffering from diabetes. Further predicting the disease early leads to treating the patient before it becomes critical.

DATA UNDERSTANDING

❑ Data Source

- This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes or not.
- This dataset has 769 samples of diabetic and healthy individuals.
- In particular, all patients here are **females** of at least 21 years of age.
- The diabetes dataset is credited to UCI machine learning database repository.

DATA UNDERSTANDING

- **Data Set Details:**

- The dataset consist of 769 samples, out of which 500 are non diabetic while 269 are diabetic people.
- All patients are **females** of at least 21years of age.
- The dataset has total 9 attributes out of which 8 are independent variables and one is the dependent variable i.e. target variable which determines whether patient is having diabetes or not.

- **Sample Data**

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1

DATA UNDERSTANDING

❖ Attributes Details:

- **Pregnancies**: No. of times pregnant
- **Glucose**: Plasma Glucose Concentration a 2 hour in an oral glucose tolerance test (mg/dl)

Plasma Glucose Test	Normal	Prediabetes	Diabetes
2 hour post-prandial	Below 140 mg/dl	140 to 199 mg/dl	200 mg/dl or more

A 2-hour value between 140 and 200 mg/dL is called impaired glucose tolerance. This is called "pre- diabetes." It means you are at increased risk of developing diabetes over time. A glucose level of 200 mg/dL or higher is used to diagnose diabetes.

- **Blood Pressure**: Diastolic Blood Pressure(mmHg)

If Diastolic B.P > 90 means High B.P (High Probability of Diabetes)

Diastolic B.P < 60 means low B.P (Less Probability of Diabetes)

DATA UNDERSTANDING

- **Skin Thickness**: Triceps Skin Fold Thickness (mm) –

  A value used to estimate body fat. Normal Triceps Skinfold Thickness in women is **23mm**. Higher thickness leads to obesity and chances of diabetes increases.

- **Insulin**: 2-Hour Serum Insulin (mu U/ml)

Feature	Normal Insulin Level
2 Hours After Glucose	**16-166 mIU/L**

Values above this range can be alarming.

- **BMI**: Body Mass Index (weight in kg/ height in m^2)

- Body Mass Index of **18.5 to 25** is within the normal range
- BMI between **25 and 30** then it falls within the overweight range.
- A BMI of **30 or over** falls within the obese range.

DATA UNDERSTANDING

- **Diabetes Pedigree Function**:

It provides information about diabetes history in relatives and genetic relationship of those relatives with patients. Higher Pedigree Function means patient is more likely to have diabetes.

- **Age** (in years)

- **Outcome**:

Class Variable (0 or 1) where '**0**' denotes patient is not having diabetes and '**1**' denotes patient having diabetes.

The **dependent variable** is whether the patient is having diabetes or not.

DATA PREPARATION

- Data preparation stage includes data cleaning and transforming data if needed.
- Various things have to be taken into consideration for data cleaning like:
  - **Handling Zero/Null Values** – The zeroes shown in the table are not zeroes but null values . We have deduced this based upon our inference that certain attributes like skin thickness, insulin, BMI etc cannot be zero.

S.No	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
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0	0.232	54	1
10	4	110	92	0	0	37.6	0.191	30	0
11	10	168	74	0	0	38	0.537	34	1
12	10	139	80	0	0	27.1	1.441	57	0
13	1	189	60	23	846	30.1	0.398	59	1

The dataset had a lot of zero values.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	\
0	6	148.0	72.000000	35.00000	95.674746	33.600000	
1	1	85.0	66.000000	29.00000	95.674746	26.600000	
2	8	183.0	64.000000	25.09192	95.674746	23.300000	
3	1	89.0	66.000000	23.00000	94.000000	28.100000	
4	0	137.0	40.000000	35.00000	168.000000	43.100000	
5	5	116.0	74.000000	25.09192	95.674746	25.600000	
6	3	78.0	50.000000	32.00000	88.000000	31.000000	
7	10	115.0	72.533517	25.09192	95.674746	35.300000	
9	8	125.0	96.000000	25.09192	95.674746	32.056663	
10	4	110.0	92.000000	25.09192	95.674746	37.600000	

	DiabetesPedigreeFunction	Age	Outcome
0	0.627	50	1
1	0.351	31	0
2	0.672	32	1
3	0.167	21	0
4	2.288	33	1
5	0.201	30	0
6	0.248	26	1
7	0.134	29	0
9	0.232	54	1
10	0.191	30	0

The zero values have been replaced by the mean of that column.

DATA PREPARATION

- **Select appropriate attributes for analysis**

The dataset consist of 9 attributes i.e. **Pregnancies, Glucose, Blood Pressure, Diabetes Pedigree Function, Age, Skin Thickness, Insulin, BMI.** These 8 are independent attributes and one i.e. **Outcome** is the dependent attribute.

As all these attributes affect diabetes so we decided to keep all the independent variables for data mining process.

- **Data Splitting**:

- Data was divided into training and testing data into 66.6:33.3 ratio. 2/3rd was training data and 1/3rd was testing data.

DATA PREPARATION

- Different ranges were found out for each continuous variable in the data set. Based upon these ranges categorization was done.
- The features were categorized as per the below mentioned ranges and were denoted by 0,1, 2 & 3, in order to use them for classification.

Glucose

Plasma Glucose Test	Normal	Prediabetes	Diabetes
2 hour post-prandial	Below 140 mg/dl (0)	140 to 199 mg/dl (1)	200 mg/dl or more (2)

Blood Pressure(Diastolic)

Ranges	Low	Normal	High
	Below 60 (0)	60-90 (1)	90 or more (2)

Skin Thickness

Ranges	Low	Normal	High
	<23 (0)	23 (1)	>23 (2)

DATA PREPARATION

Insulin

Ranges	Low	Normal	High
2 Hours After Glucose	<16 mIU/L (0)	16-166 mIU/L (1)	>166 mIU/L (2)

BMI(weight in kg/ height in m^2)

Ranges	Under-weight	Normal	Over-weight	Obese
	<18.5 (0)	18.5-25 (1)	25-30 (2)	>30 (3)

Diabetes Pedigree Function

	Low	Medium	High
	0-0.78 (0)	0.79-1.561 (1)	>1.57 (2)

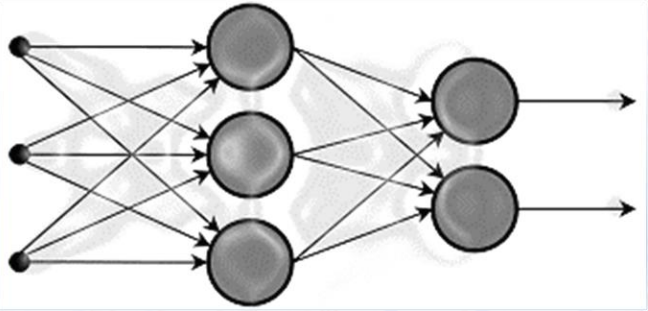
Age

Ranges	Young	Adult	Old
	20-44 (0)	44-64 (1)	64-100 (2)

No.of Pregnancies

Ranges	Normal	Above Normal	Highest
	<6 (0)	6-12 (1)	>12 (2)

MODELING



- This phase includes application of appropriate model to the data.
- Machine Learning Algorithms were used for modeling.
- As we have to classify the data into patients having diabetes or not, I have chosen **K-nearest neighbors (KNN)**, **Random decision forests** and **Artificial neural networks (ANN)** to see which one provides us with the best accuracy.

MODELING

- **Software Used**:
- Python-Scikit Learn
- Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.
- The library is built upon the SciPy (Scientific Python) that must be installed before we can use scikit-learn. This stack includes:
- **NumPy**: Base n-dimensional array package
- **SciPy**: Fundamental library for scientific computing
- **Matplotlib**: Comprehensive 2D/3D plotting
- **IPython**: Enhanced interactive console
- **Sympy**: Symbolic mathematics
- **Pandas**: Data structures and analysis

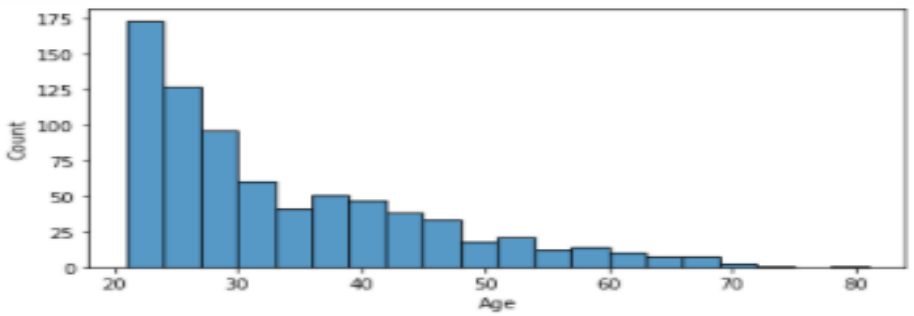
MODELING: USING PYTHON SCIKIT LEARN

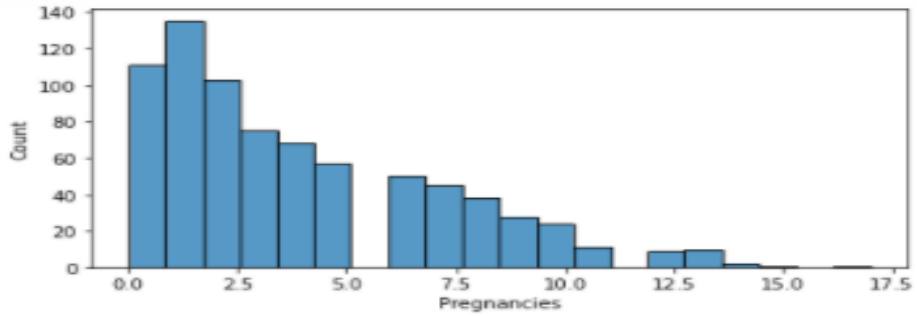
- The file containing data set is loaded in pandas.

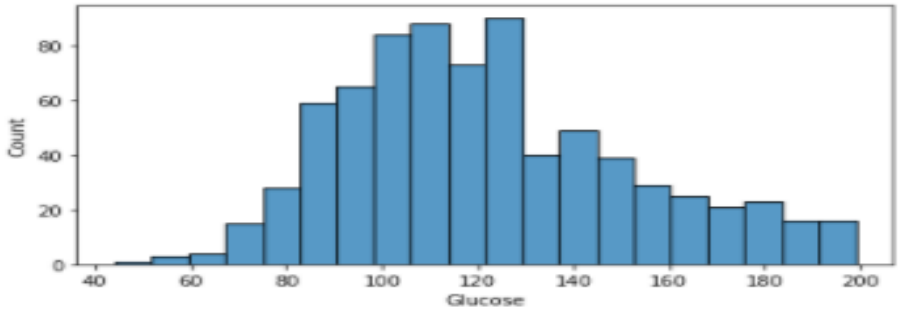
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	\
0	6	148.0	72.000000	35.000000	79.799479	33.600000	
1	1	85.0	66.000000	29.000000	79.799479	26.600000	
2	8	183.0	64.000000	20.536458	79.799479	23.300000	
3	1	89.0	66.000000	23.000000	94.000000	28.100000	
4	0	137.0	40.000000	35.000000	168.000000	43.100000	
5	5	116.0	74.000000	20.536458	79.799479	25.600000	
6	3	78.0	50.000000	32.000000	88.000000	31.000000	
7	10	115.0	69.105469	20.536458	79.799479	35.300000	
8	2	197.0	70.000000	45.000000	543.000000	30.500000	
9	8	125.0	96.000000	20.536458	79.799479	31.992578	

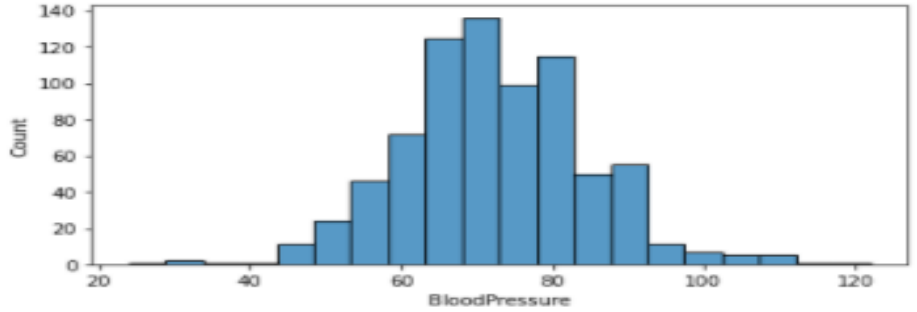
	DiabetesPedigreeFunction	Age	Outcome
0	0.627	50	1
1	0.351	31	0
2	0.672	32	1
3	0.167	21	0
4	2.288	33	1
5	0.201	30	0
6	0.248	26	1
7	0.134	29	0
8	0.158	53	1
9	0.232	54	1

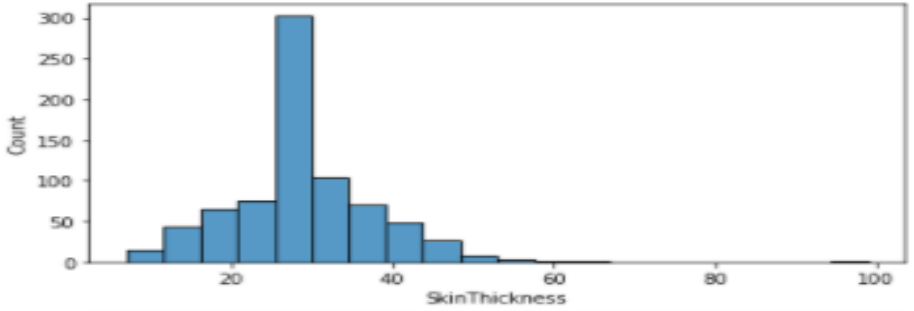
Histograms explaining the variation of count with different attributes mentioned before

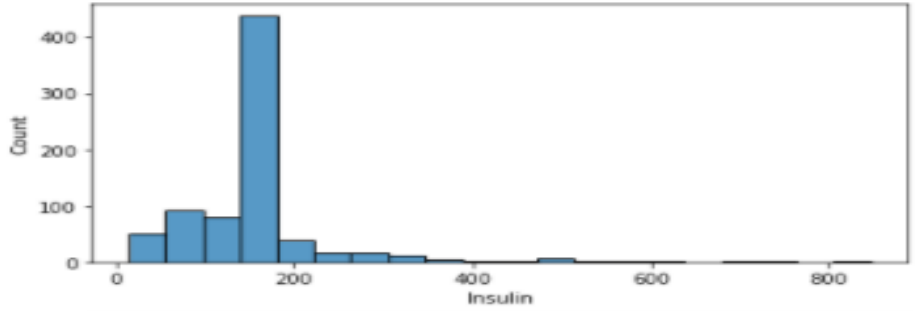


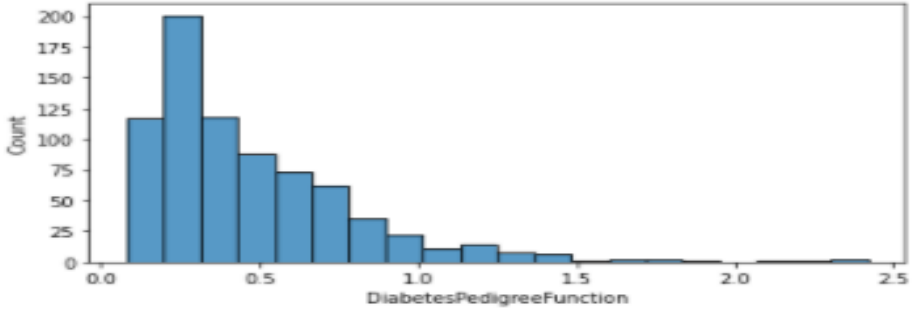


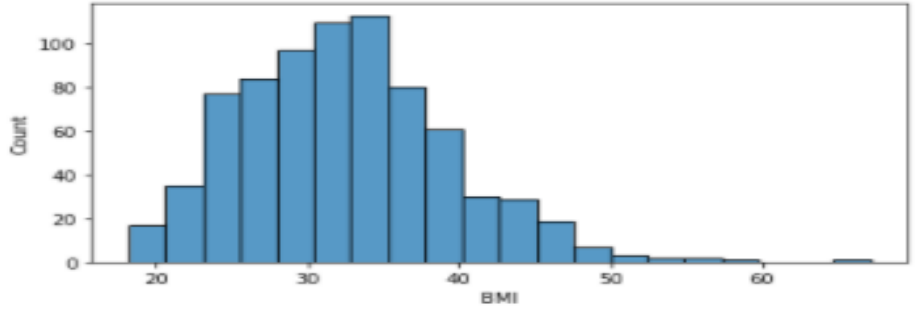




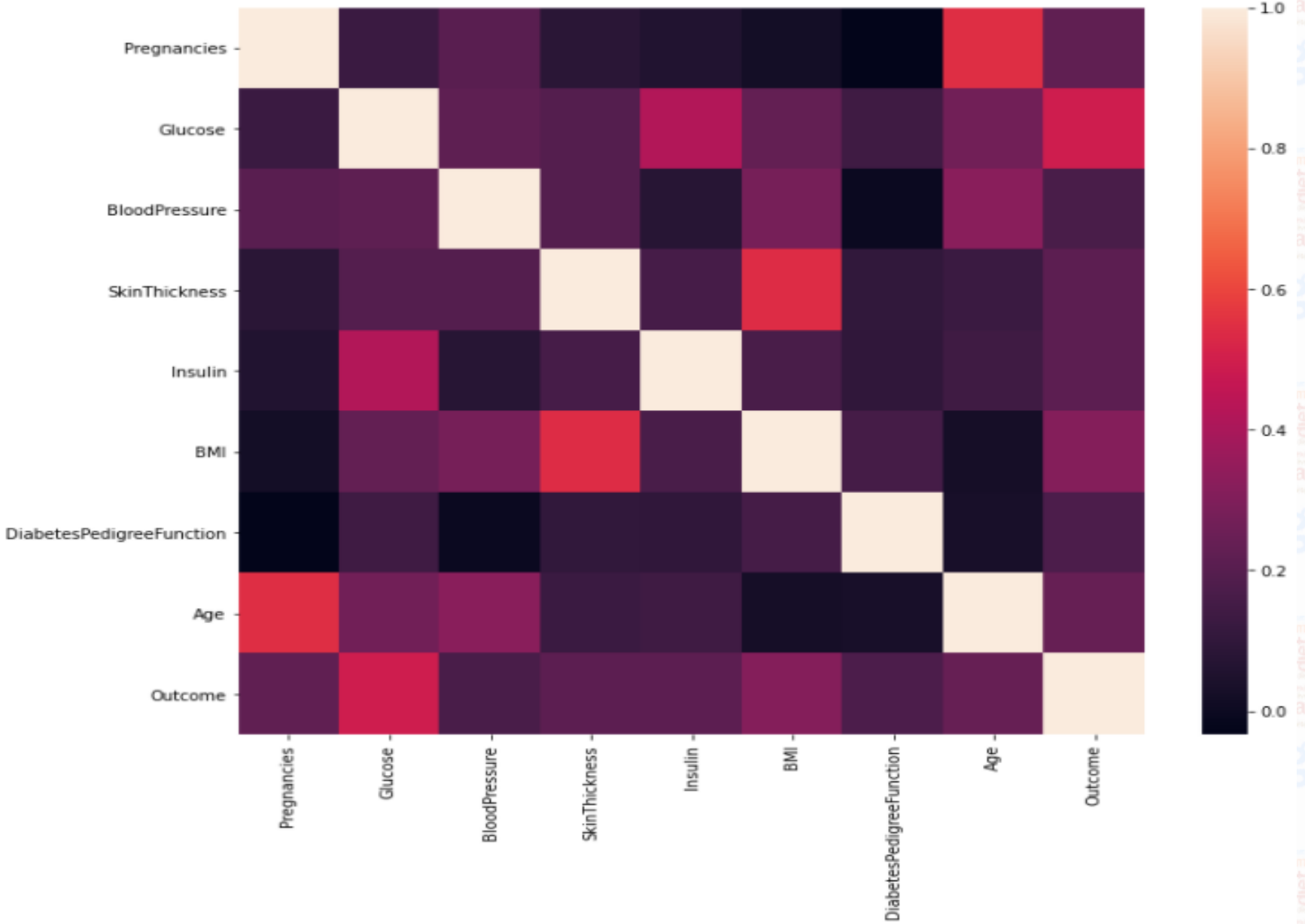




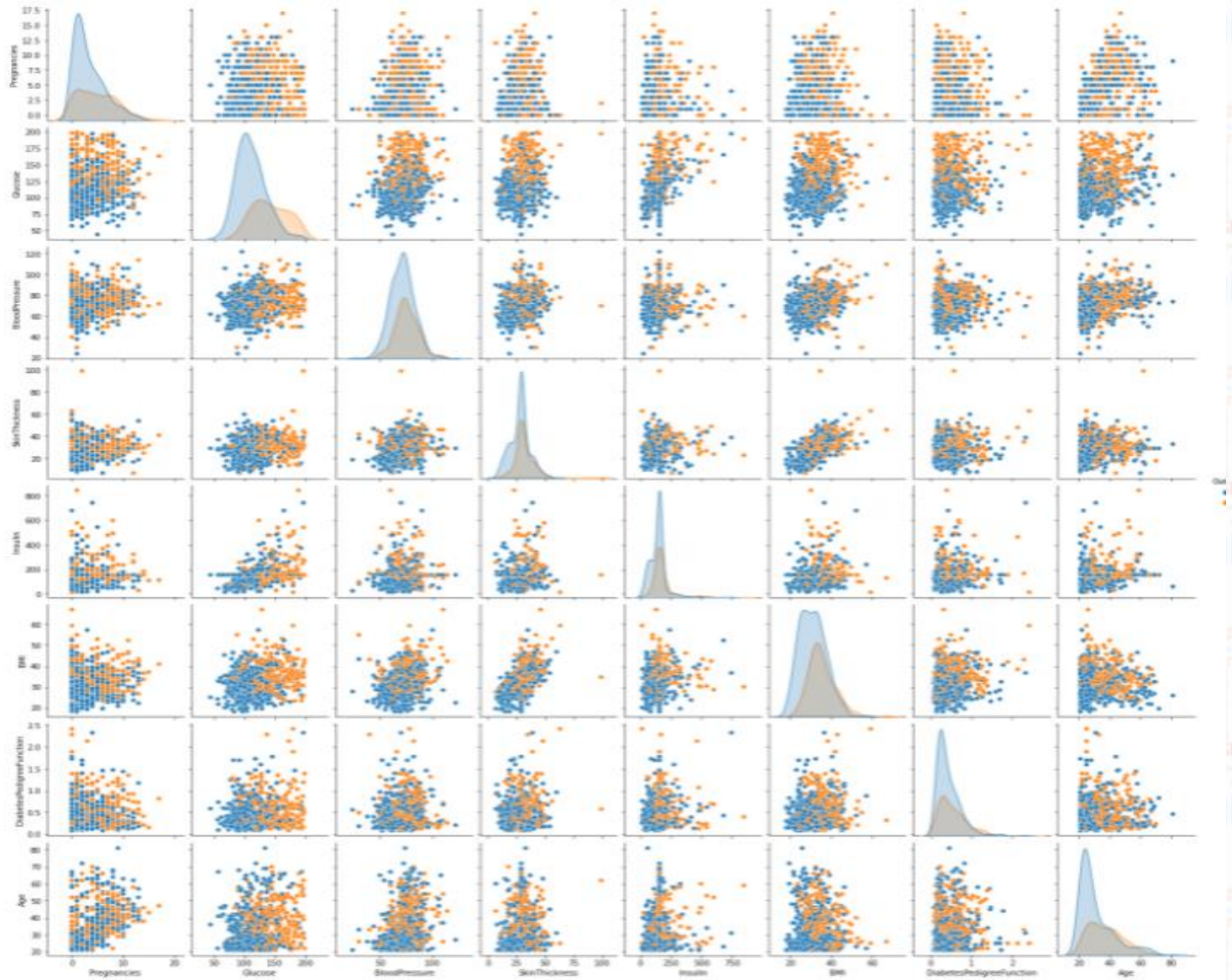




Heatmap (using seaborn) explaining how each attribute is related to the other



Graphs explaining the relationship between each of the parameters



EVALUATION

CONFUSION MATRIX

	Predicted Yes	Predicted No
Actual Yes	tp (true positive)	fp (false positive)
Actual No	fn (false negative)	tn (true negative)

- **true positives (TP):** These are cases in which we correctly predicted diabetes as result.
- **true negatives (TN):** We correctly predicted no diabetes and they don't have the disease.
- **false positives (FP):** We correctly predicted no diabetes, but they actually had the disease. (Also known as a "Type I error.")
- **false negatives (FN):** We correctly predicted diabetes, but they actually had no disease. (Also known as a "Type II error.")

Confusion Matrix for KNN

	Predicted Diabetes	Predicted No Diabetes
Actual Diabetes	**141**	**29**
Actual No Diabetes	**39**	**47**

$$\text{Accuracy} = \frac{tp + tn}{tp+tn+fp+fn} \qquad \text{Recall} = \frac{tp}{tp+fn}$$

$$\text{Precision} = \frac{tp}{tp+fp}$$

	precision	recall	f1-score	support
0	0.78	0.83	0.81	170
1	0.62	0.55	0.58	86
accuracy			0.73	256
macro avg	0.70	0.69	0.69	256
weighted avg	0.73	0.73	0.73	256

The Accuracy for KNN is 73%

ALTERNATE MODEL COMPARISON

- We tried alternate algorithms to compare the accuracy of the model.
- We used Random Forest Classifier and Artificial Neural Network. The results obtained are as under:

Confusion Matrix Random Forest Classifier

	Predicted Diabetes	Predicted No Diabetes
Actual Diabetes	145	25
Actual No Diabetes	42	44

	precision	recall	f1-score	support
0	0.77	0.85	0.81	170
1	0.63	0.51	0.56	86
accuracy			0.73	256
macro avg	0.70	0.68	0.69	256
weighted avg	0.73	0.73	0.73	256

The Accuracy of Random Forest Classifier Model is 73%

Confusion Matrix ANN

	Predicted Diabetes	Predicted No Diabetes
Actual Diabetes	145	25
Actual No Diabetes	34	52

	precision	recall	f1-score	support
0	0.86	0.76	0.81	170
1	0.61	0.76	0.68	86
accuracy			0.76	256
macro avg	0.74	0.76	0.74	256
weighted avg	0.78	0.76	0.76	256

The Accuracy of ANN is 76%.

DEPLOYMENT

- This is the last and the final phase of CRISP DM process. Deployment includes three important task :
- **Plan Deployment** – Planning basically includes the strategy to be formulated for implementing the model in real world. This model can now be used in medical organizations for easy and early detection of diabetes in patients.
- **Monitor Deployment** – In this, continuous monitoring of model takes place. Regular check is done to ensure model is working fine and if any error occurs can be easily detected.
- **Generate Reports** – Final statistical reports are generated which summarizes the overall performance of the model.

CONCLUSION

- The ANN model achieved 76% accuracy after various hyperparameter tunings which is the highest accuracy compared to KNN and Random Forest Classifier's 73%
- For the other 2 models different options were taken into consideration to improve the accuracy.
- In KNN we chose the number of neighbors as 17 which provided the best accuracy

and similarly in Random Forest Classifier we chose the number of estimators = 10

REFERENCE

- https://www.kaggle.com/uciml/pima-indians-diabetes-database
- Slides & Lecture Notes
- http://scikit-learn.org/stable/
- http://pandas.pydata.org/

THANK YOU

My Github : Please Click Here

My LinkedIn : Please Click Here