



DIABETES DETECTION

Course : MIS 637-A (SPRING 2017)

MAJOR : INFORMATION SYSTEM

Prepared by : NIHARIKA GULATI

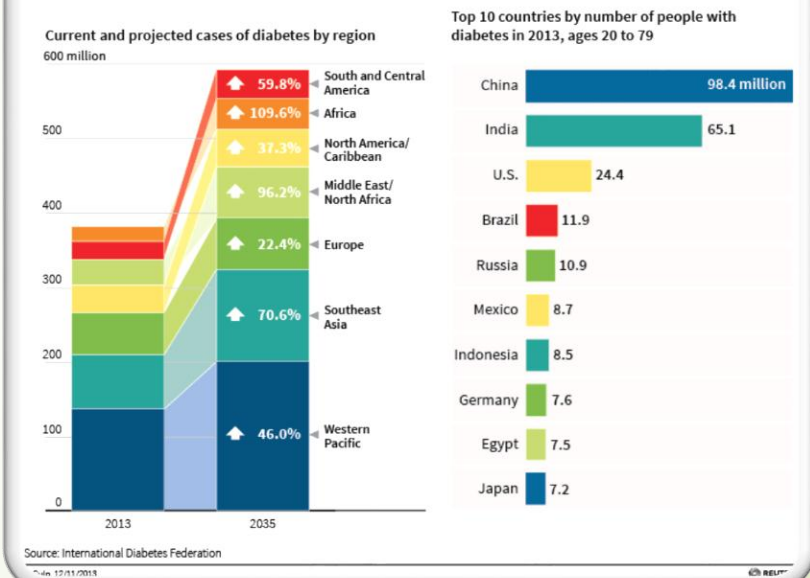
Guided by : Prof. MAHMOUD DANESHMAND

Diabetes

INTRODUCTION

- Diabetes is one of the deadliest diseases in the world. It is not only a disease but also creator of different kinds of diseases like heart attack, blindness etc. The normal identifying process is that patients need to visit a diagnostic center, consult their 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.

World diabetes cases expected to jump 55 percent by 2035





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

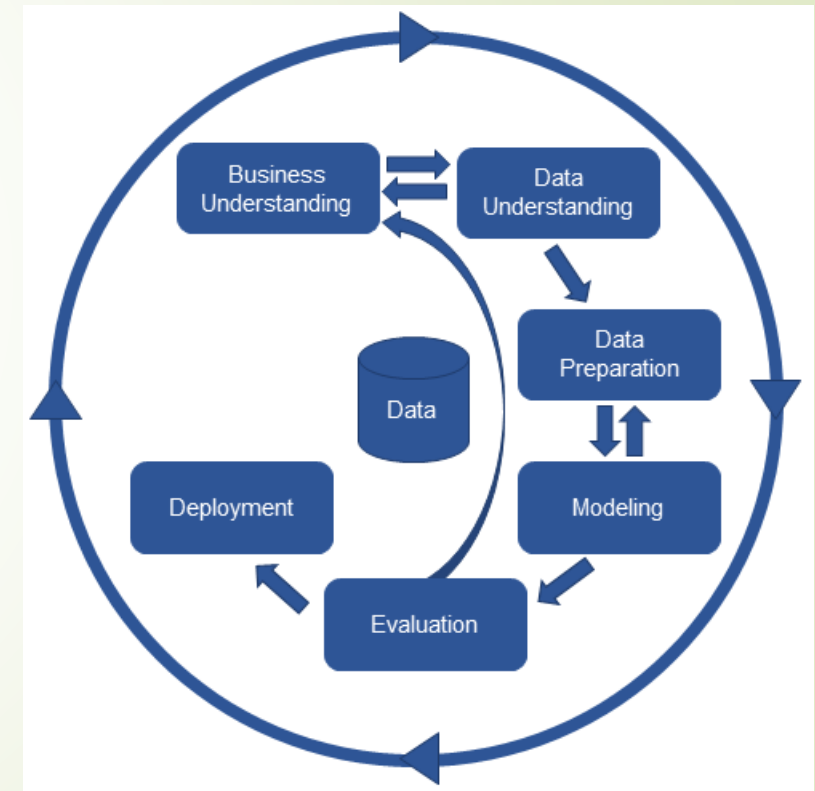
- Select appropriate data modeling technique

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 21 years 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 m²)

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.

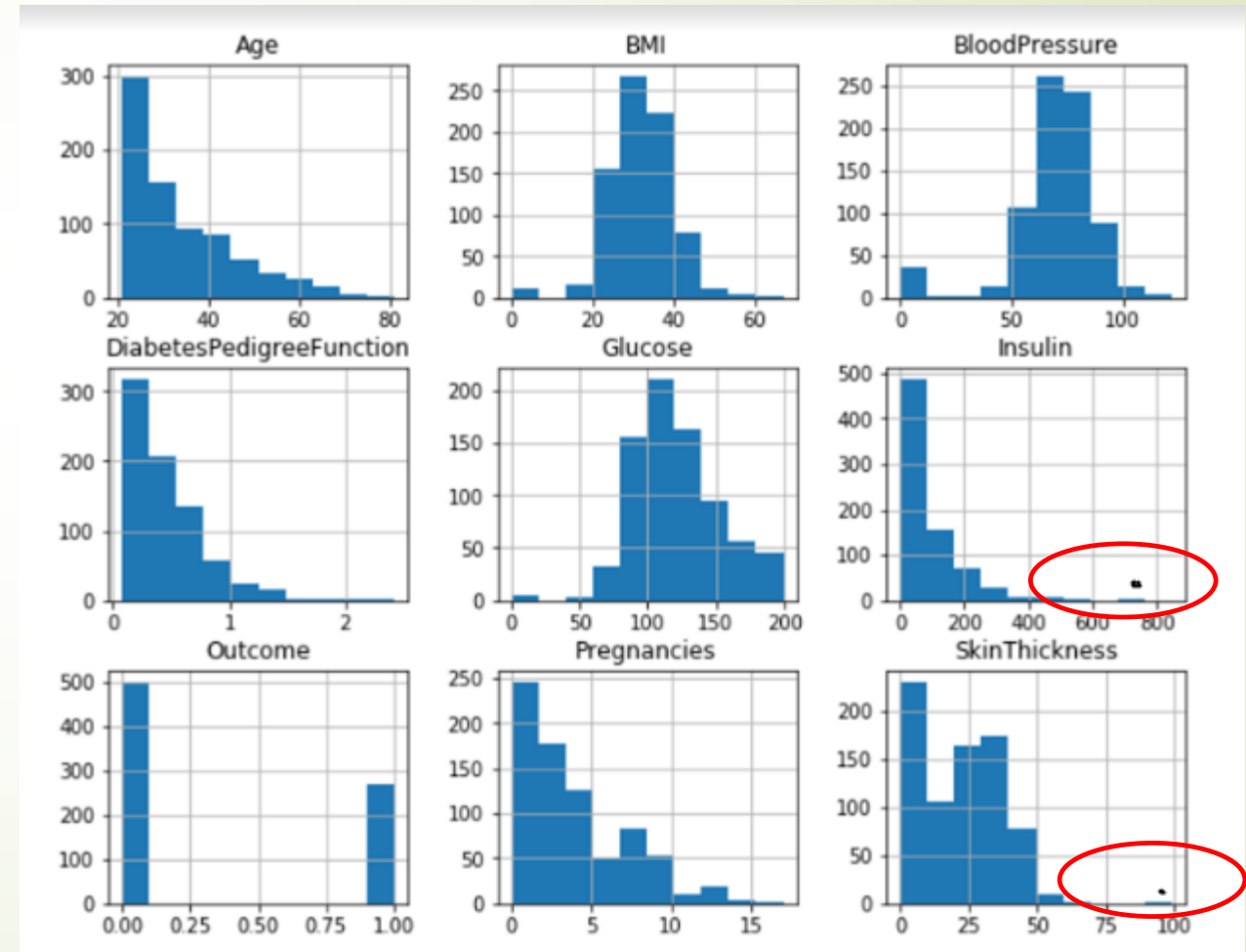
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	\
0	6	148.0	72.000000	35.00000	95.674746	33.600000	
1	1	85.0	66.000000	29.00000	95.674746	26.600000	
2	8	183.0	64.000000	25.09192	95.674746	23.300000	
3	1	89.0	66.000000	23.00000	94.000000	28.100000	
4	0	137.0	40.000000	35.00000	168.000000	43.100000	
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

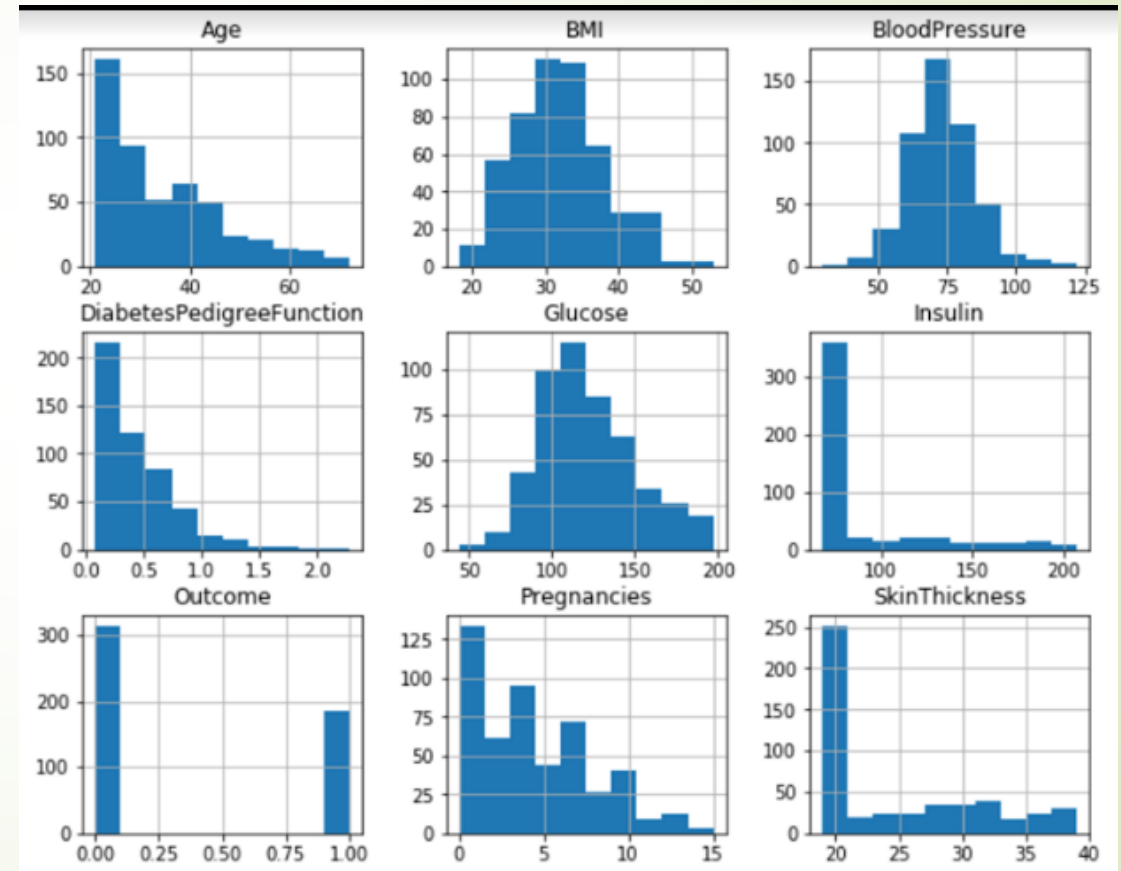
- **Check and remove outliers**

This graph shows the distribution of dataset of different attributes. By carefully studying the graph we figured out that Insulin and Skin thickness graphs have outliers. The black dots depicts the outliers.



DATA PREPARATION

- Outliers were found in **Insulin** and **Skin Thickness** attributes. They were removed using **IQR (Inter Quartile Range)** method.



Outliers Removed



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 60:40 ratio. Sixty percent was training data and forty percent was testing data.
- Out of 498 records after data cleaning, 298 records were used for trained and 200 records were used for testing.

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 m²)

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)


DATA PREPARATION

- The tables shows the data after categorization. Different ranges taken for categorization of data are denoted by 0,1, 2 & 3.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	1	1	1	1	1	3	0	1	1
1	0	0	1	1	1	2	0	0	0
2	1	1	1	0	1	1	0	0	1
3	0	0	1	0	1	2	0	0	0
4	0	0	0	1	2	3	2	0	1
5	0	0	1	0	1	2	0	0	0
6	0	0	0	1	1	3	0	0	1
7	1	0	1	0	1	3	0	0	0
9	1	0	2	0	1	3	0	1	1
10	0	0	2	0	1	3	0	0	0
11	1	1	1	0	1	3	0	0	1
12	1	0	1	0	1	2	1	1	0
14	0	1	1	0	2	2	0	1	1
15	1	0	1	0	1	3	0	0	1
17	1	0	1	0	1	3	0	0	1
18	0	0	0	1	1	3	0	0	0
19	0	0	1	1	1	3	0	0	1
21	1	0	1	0	1	3	0	1	0



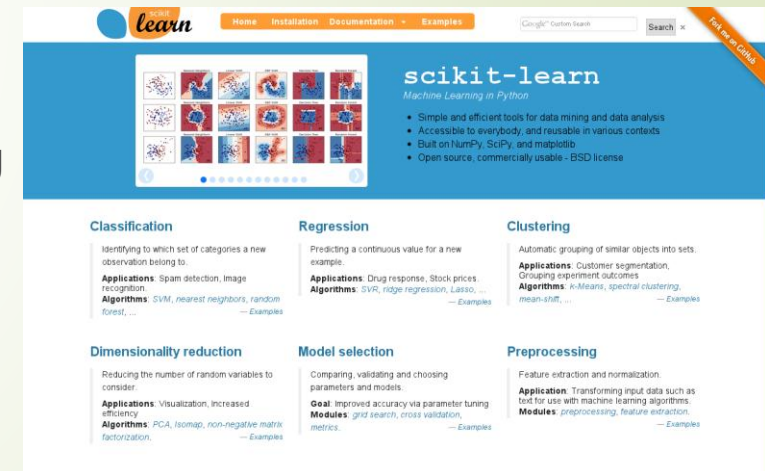
MODELING

- 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, we used Classification and Regression Tree Algorithm(CART) & K-Nearest Neighbor algorithms. Both of these algorithms are good for classifying dependent variables based upon categorized independent variables.
 - We compared both the algorithms to find the one which gives the best results based upon overall accuracy and precision.
- 

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

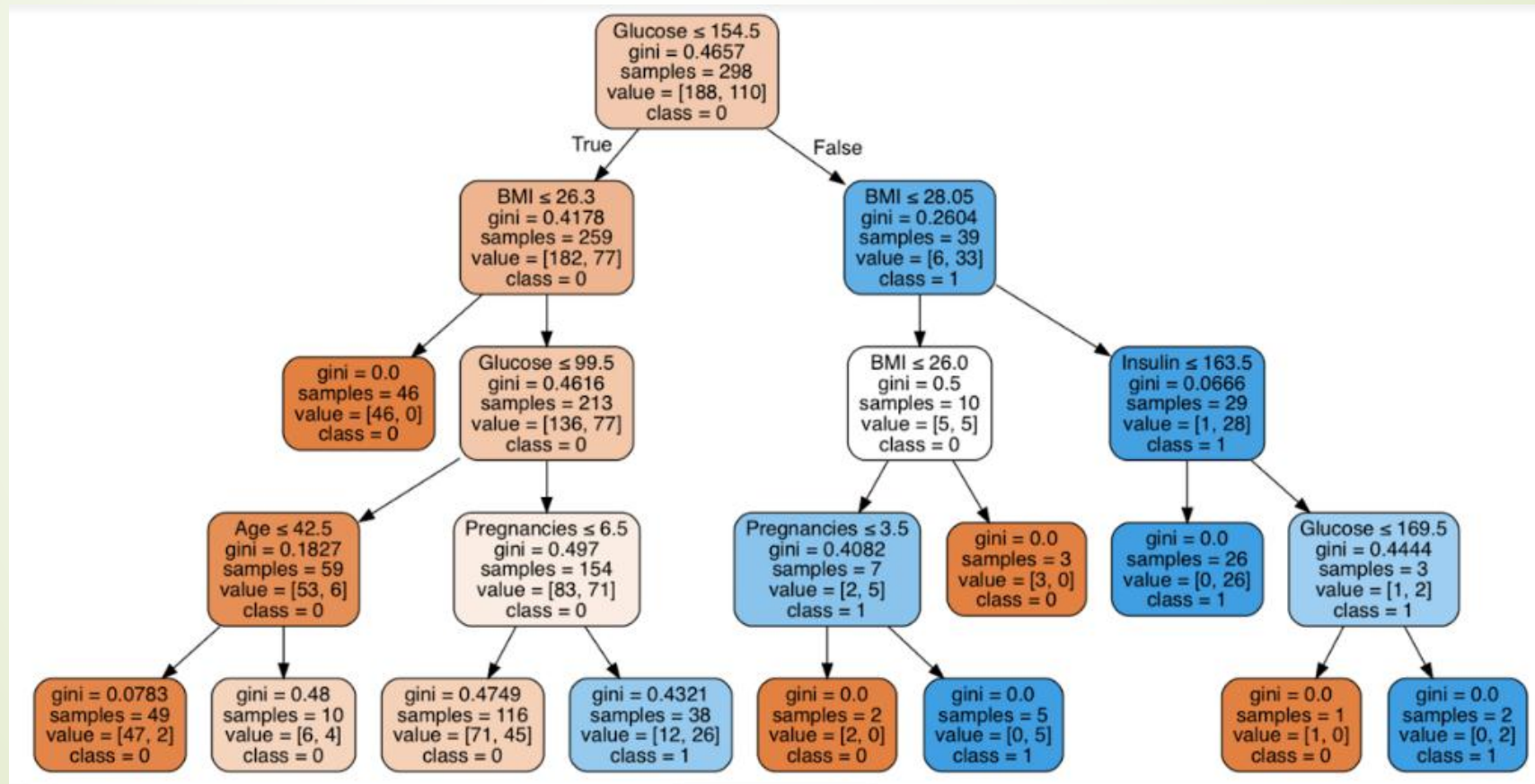
- The file containing data set is loaded in pandas.

	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

MODELING: USING PYTHON SCIKIT LEARN

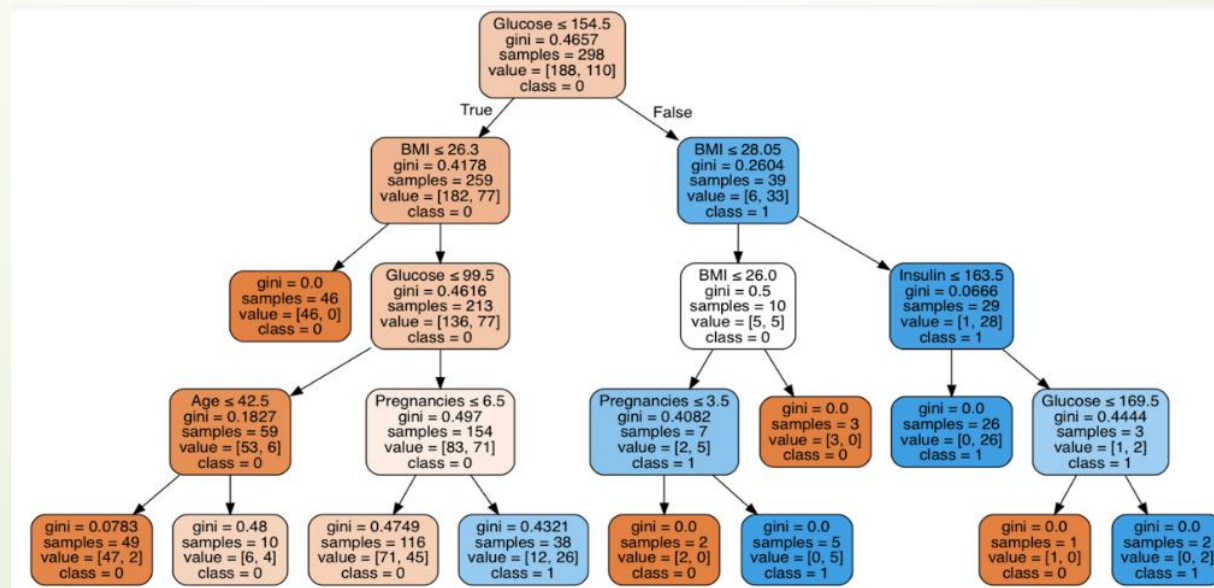
- The decision tree model i.e. CART is applied on the training dataset. The Decision Tree obtained gives the best result. The depth taken for this tree is **4** and total number of nodes are **21**.



MODELING: USING PYTHON SCIKIT LEARN

Node Description:

- The root node split in this tree started with Glucose attribute.
- No of samples – It is the count of those samples whose glucose value is less than 154.5
- Value – It gives the total no of samples for outcome '0' and '1'. For e.g. in root node the value 188 (represents class '0') is higher than 110 (represents class '1') , so this node is classified as class '0'
- Class – Diabetic – '1' or Non Diabetic – '0'
- Gini Impurity – This function measures the quality of a split. This factor measures how a randomly chosen element from the set would be incorrectly classified i.e. It is probability of misclassification of a record. It is used to minimize misclassifications.



MODELING: USING PYTHON SCIKIT LEARN

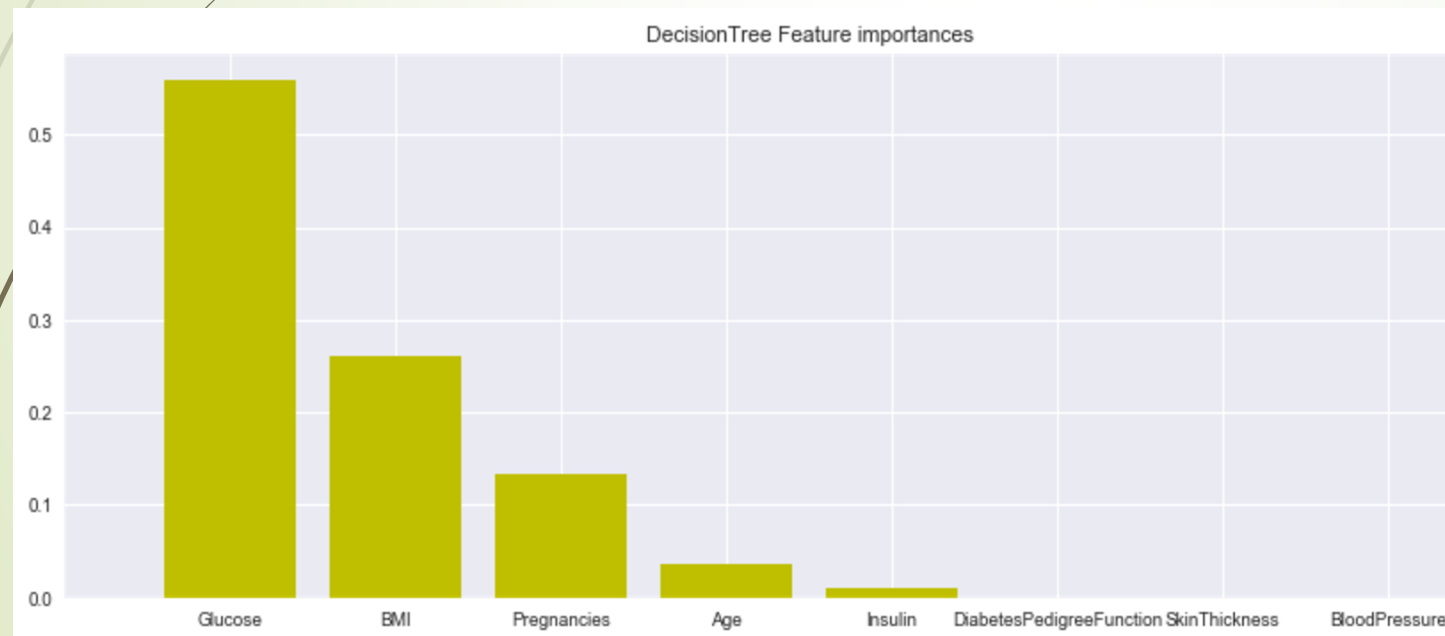
Decision Rules

S.No.	Antecedent	Consequent	Support	Confidence
1.	If Glucose<= 154.5 & BMI <= 26.3	No Diabetes	46/298	1
2.	If Glucose<= 154.5 & BMI <= 26.3 & Glucose <= 99.5 & Age <= 42.5	No Diabetes	47/298	0.95
3.	If Glucose<= 154.5 & BMI <= 26.3 & Glucose <= 99.5 & Age <= 42.5	No Diabetes	6/298	0.60
4.	If Glucose<= 154.5 & BMI <= 26.3 & Glucose <= 99.5 & Pregnancies<= 6.5	No Diabetes	71/298	0.61
5.	If Glucose<= 154.5 & BMI <= 26.3 & Glucose <= 99.5 & Pregnancies<= 6.5	Diabetes	26/298	0.68
6.	If Glucose<= 154.5 & BMI <= 28.05 & BMI <= 26.00	No Diabetes	3/298	1
7.	If Glucose<= 154.5 & BMI <= 28.05 & BMI <= 26.00 & Pregnancies <= 3.5	No Diabetes	2/298	1
8.	If Glucose<= 154.5 & BMI <= 28.05 & BMI <= 26.00 & Pregnancies <= 3.5	Diabetes	5/298	1
9.	If Glucose<= 154.5 & BMI <= 28.05 & Insulin <= 163.5	Diabetes	26/298	1
10.	If Glucose<= 154.5 & BMI <= 28.05 & Insulin <= 163.5 & Glucose <= 169.5	No Diabetes	1/298	1
11.	If Glucose<= 154.5 & BMI <= 28.05 & Insulin <= 163.5 & Glucose <= 169.5	Diabetes	2/298	1

EVALUATION

Feature Importance

- With the help of decision tree, we were able to figure out which features played an important role.
- This graph depicts the highest importance features.

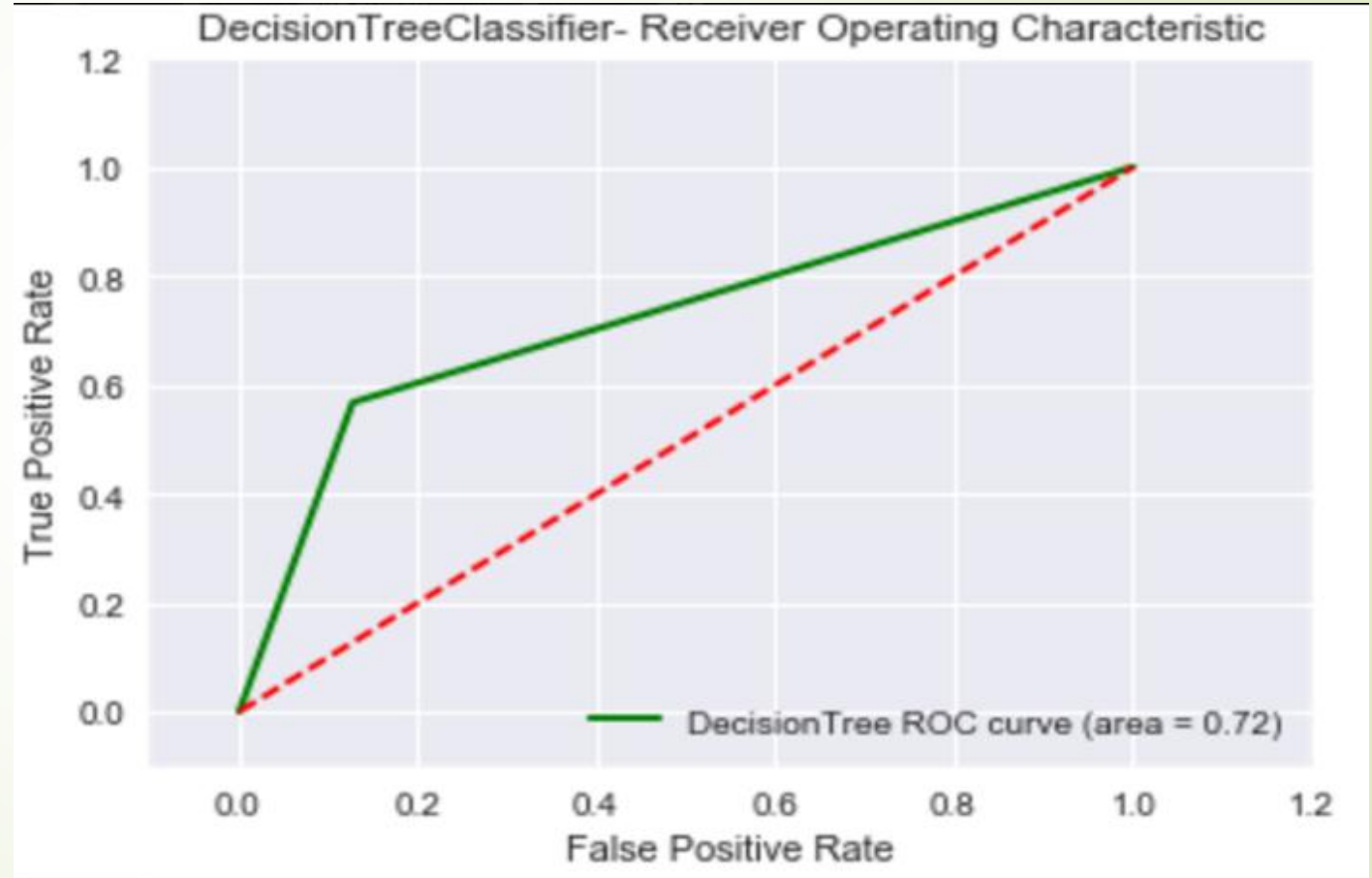


Decision Tree Feature Ranking		
1.	Glucose	0.558787
2.	BMI	0.260081
3.	Pregnancies	0.134401
4.	Age	0.036539
5.	Insulin	0.010191
6.	Diabetes Pedigree Function	0.000000
7.	Skin Thickness	0.000000
8.	Blood Pressure	0.000000

EVALUATION

ROC CURVE

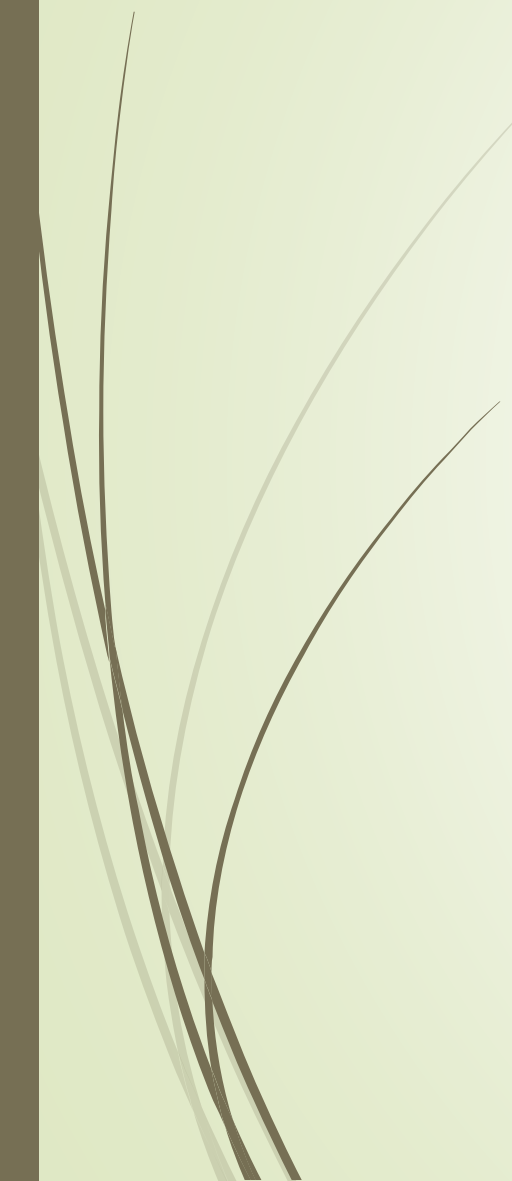
- The green line depicts Decision Tree curve area which is 0.72.
- The dashed line in the diagonal represents the ROC curve of a random predictor. The random predictor is commonly used as a baseline to see whether the model is useful or not.





EVALUATION

The graph is called a **Receiver Operating Characteristic curve** (or ROC curve.) It is a plot of the true positive rate against the false positive rate for the different points.

- It shows the tradeoff between True Positive Rate (TPR) and False Positive Rate (FPR) (any increase in (TPR) will be accompanied by a decrease in (FPR)).
 - The closer the curve towards TPR of the ROC space, the more accurate the test.
 - The closer the curve towards the FPR of the ROC space, the less accurate the test.
 - The Area Under the Curve (AUC) is a measure to determine the model effectiveness. The more AUC for a model comes to 1, the better it is. So models with higher AUCs are preferred over those with lower AUCs.
- 

EVALUATION

CONFUSION MATRIX

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

Confusion Matrix for the Decision Tree

	Predicted Diabetes	Predicted No Diabetes
Actual Diabetes	110	16
Actual No Diabetes	32	42

The Accuracy for the Decision Tree is 76%

- 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.")

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

$$\text{Recall} = \frac{tp}{tp + fn}$$

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

	precision	recall	f1-score	support
0	0.77	0.87	0.82	126
1	0.72	0.57	0.64	74
avg / total	0.76	0.76	0.75	200

```
[[110 16]
 [ 32 42]]
Decision Tree accuracy: 0.76
```

ALTERNATE MODEL COMPARISON

- We tried another machine learning algorithm to compare the accuracy of the model.
- We used K-Nearest Neighbor algorithm. The results obtained are as under:

Confusion Matrix KNN

	Predicted Diabetes	Predicted No Diabetes
Actual Diabetes	103	23
Actual No Diabetes	35	39

	precision	recall	f1-score	support
0	0.75	0.82	0.78	126
1	0.63	0.53	0.57	74
avg / total	0.70	0.71	0.70	200

```
[[103 23]
 [ 35 39]]
KNN accuracy: 0.71
```

The Accuracy of KNN Model is 71%

Confusion Matrix Decision Tree

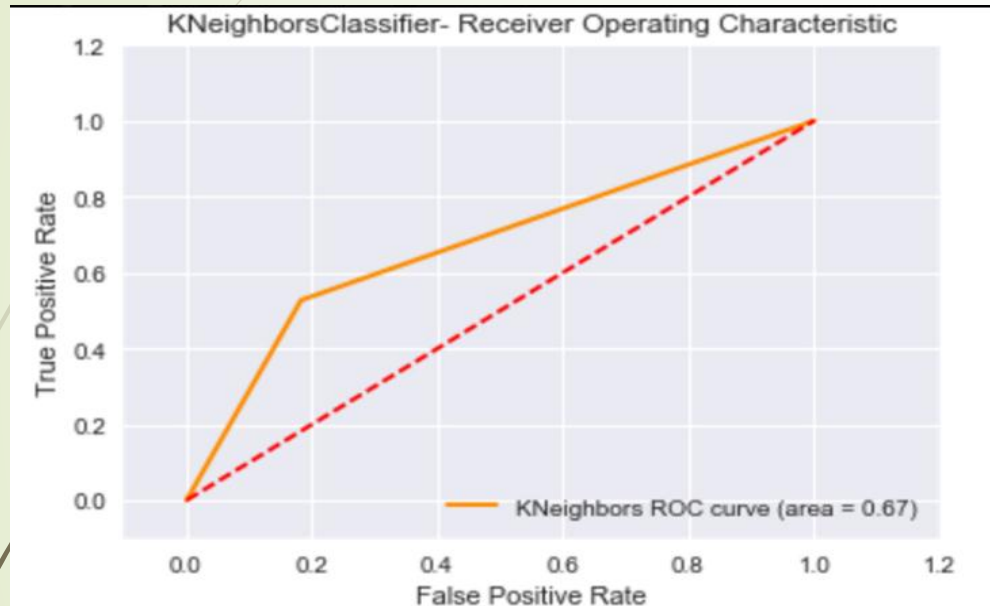
	Predicted Diabetes	Predicted No Diabetes
Actual Diabetes	110	16
Actual No Diabetes	32	42

	precision	recall	f1-score	support
0	0.77	0.87	0.82	126
1	0.72	0.57	0.64	74
avg / total	0.76	0.76	0.75	200

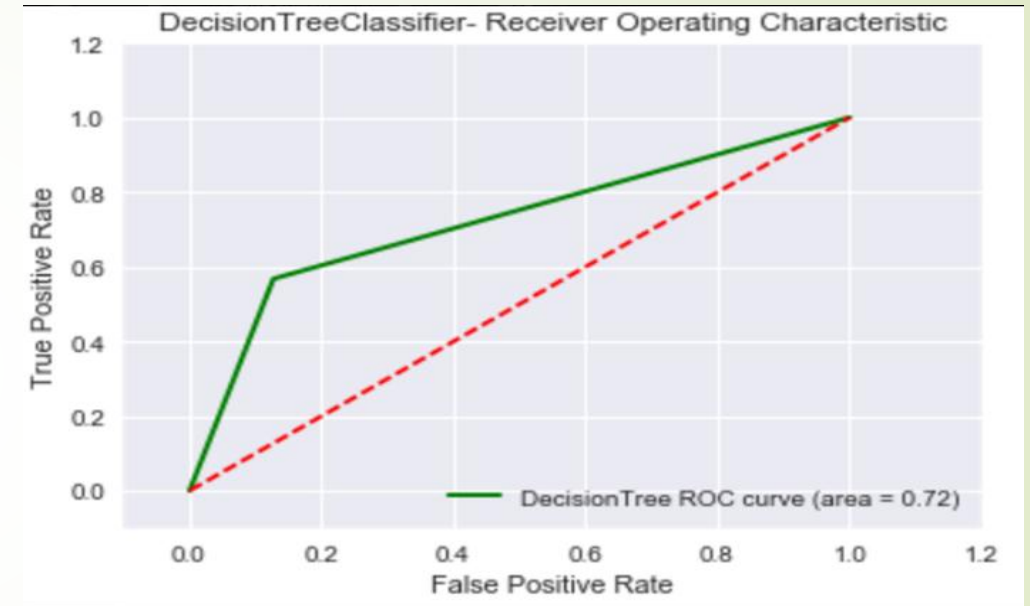
```
[[110 16]
 [ 32 42]]
Decision Tree accuracy: 0.76
```

The Accuracy of Decision Tree is 76%.

ALTERNATE MODEL COMPARISON



The Area Under Curve(AUC) for KNN is 0.67.



The Area Under Curve(AUC) for Decision Tree is 0.72.

So by comparing both the results we infer that Decision Tree Model is showing better results in comparison to K-Nearest Neighbor Model.



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.

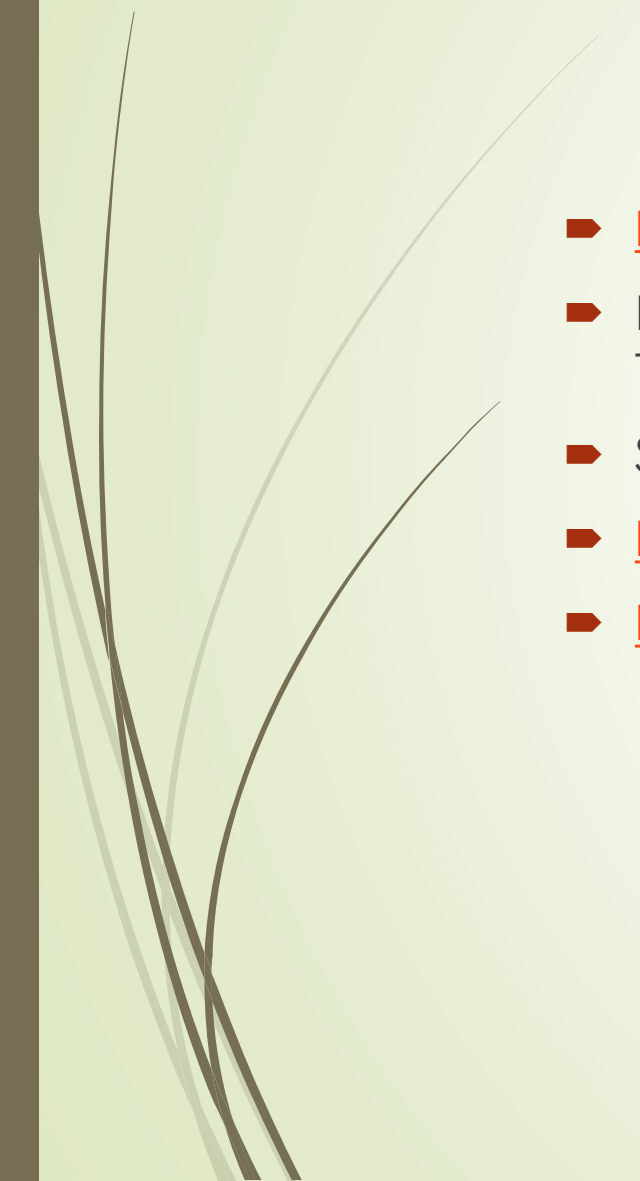


CONCLUSION

- The Decision Tree Model achieved 76% accuracy.
- Different options were taken into consideration to improve the accuracy. So finally by removing outliers, categorizing data, keeping tree depth to 4 we were able to achieve desired accuracy.
- During this process we figured out few attributes that played an important role . Out of eight attributes Glucose, BMI, Pregnancies, Age and Insulin were the important ones. As per our results and data the other factors like Diabetes Pedigree Function, Skin Thickness and Blood Pressure had negligible effect in determining diabetes.
- We even compared our CART model with K-Nearest Model and inferred that CART is the best amongst the two.



REFERENCE

- <http://archive.ics.uci.edu/ml/>
 - Discovering Knowledge in Data: An Introduction to Data Mining, By Daniel T. Larose
 - Slides & Lecture Notes
 - <http://scikit-learn.org/stable/>
 - <http://pandas.pydata.org/>
- 



THANK YOU