

Diagnostically predict diabetes on Pima Indians dataset

RAM SWAROOP HOLALKERE KRISHNA
NET ID – RH668/ RU-ID 191004219

Diagnostically predict diabetes on Pima Indians dataset

INDEX

1. Introduction.....	2-4
2. Data Cleaning and Pre-Processing	5-7
3. Exploratory Data Analysis	8-10
4. Data Modelling and Evaluation	11-19
5. Conclusion and Applications	19
6. References.....	20

Diagnostically predict diabetes on Pima Indians dataset

Introduction

Diabetes is a disease that occurs when the blood glucose, also called blood sugar, is too high. Blood glucose is your main source of energy and comes from the food you eat. Blood glucose is the main source of energy and comes from the food we eat. The insulin hormone made by the pancreas, helps glucose from food get into the cells to be used for energy. Sometimes the body doesn't produce enough insulin and glucose remains in the blood. There can also be a case when the body doesn't produce any insulin. As a result, there is high blood sugar level in the bloodstream which can damage kidney, nerves and other organs.

There are several types of diabetes. Most common types are Type1, Type2 and Gestational Diabetes.

In type1 diabetes, the body does not produce insulin, the immune system attacks the cells in pancreas that produce insulin. This type of diabetes is diagnosed in children and young adults most frequently. A patient with type1 diabetes needs to take insulin every day.

In type2 diabetes, the body does not use or make insulin well. type2 diabetes is commonly diagnosed in middle aged or old people, however it can be seen during childhood.

Gestational diabetes is diagnosed with women during their pregnancy and it goes away after the child is born. However, there are chances of developing type2 diabetes at a later stage.

Pima Indians Diabetes Dataset

Who are the Pima Indians?

The Pima Indians are a group of Native Americans living in an area consisting of what is now central and southern Arizona. They lived along the Gila and Salt rivers in Arizona. They called themselves River people. Most of them were sedentary farmers who used rivers for irrigation. The Pima Indians spoke Uto- Aztecan language and existed around the 1600s. Although farming was their main occupation, during drought years hunting and gathering were done to supplement their diet. During drought years jackrabbits and mesquite beans became their dietary staples.

About the Dataset

The dataset is from National Institute of Diabetes and Digestive and Kidney Diseases. The main goal of the Project is to predict whether a patient has diabetes or not. In this dataset there are several medical predictors like BMI, Glucose level, Insulin level, Blood Pressure, Age, Pregnancies etc. There is one outcome variable which is binary (0 and 1), this is an indicator variable whether a patient has diabetes or not. All the patients mentioned in the dataset are female of pima Indian heritage. There is a total of 768 patients in the dataset.

Diagnostically predict diabetes on Pima Indians dataset

Problem Statement

To build a machine learning model to predict whether a patient is suffering from diabetes or not. This is a classification problem, where the patients will be classified into two groups (with or without diabetes) based on the predictors we have in our dataset.

We are implementing a classification technique for this problem. Classification is a supervised machine learning technique where we develop a model to predict the class of the variable. This is called supervised learning because we train our model on the known datapoints and later test it on the testing set.

Variables can be two types Dependent and Independent.

Dependent variables in the Dataset

Pregnancies	Number of pregnancies that the patient has gone through
Glucose	Glucose level in the blood
Age	Age of the patient in years.
BloodPressure	Blood pressure of the patient.
SkinThickness	Skin thickness measured in mm
Insulin	Serum Insulin level in the body.
DiabetesPedigreeFunction	DPF scores likelihood of Diabetes based on Family history. Higher the DPF more the likelihood.
BMI	Body mass index (weight in kg/(height in mts)^2)

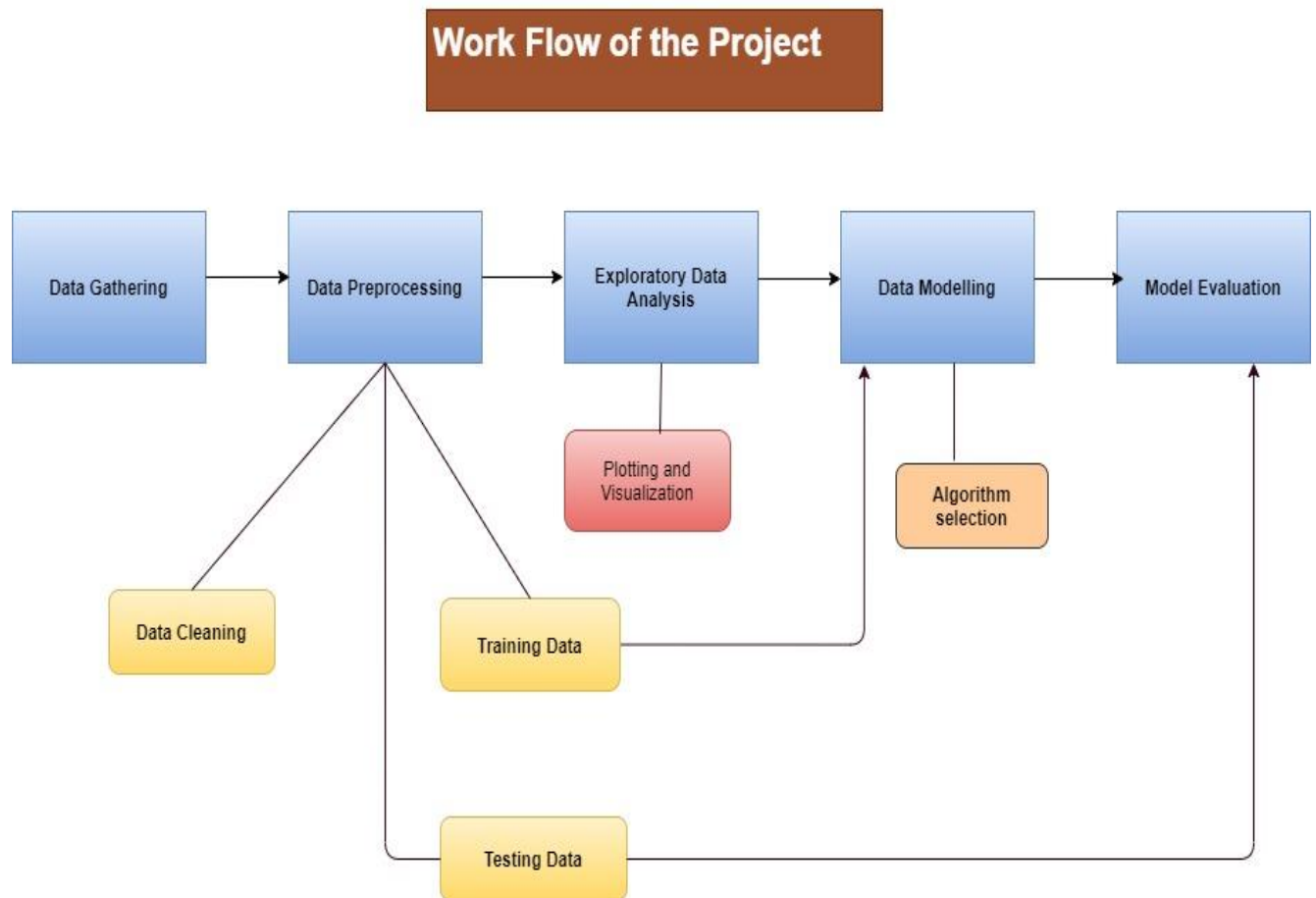
Independent variable in the Dataset

Outcome	Binary Class variable (0 and 1 value) 1 is people with Diabetes and 0 is without.
---------	--

Diagnostically predict diabetes on Pima Indians dataset

The main steps involved in the project are

- Data Gathering
- Data Preprocessing
- Exploratory Data Analysis (EDA)
- Data Modelling
- Model Evaluation



Diagnostically predict diabetes on Pima Indians dataset

Data Gathering – As mentioned before Pima Indians dataset was originally from National Institute of Diabetes and Digestive and Kidney Diseases. The dataset is available in Data.gov or the UCI repository.

Data Preprocessing

Data preprocessing also known as Data cleaning has the following steps:

- Checking for Null/empty values.
- Scaling the data.
- Dividing the data into train and test.

Checking for Null/empty values – There are no null values in the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
Pregnancies      768 non-null int64
Glucose          768 non-null float64
BloodPressure    768 non-null float64
SkinThickness    768 non-null float64
Insulin          768 non-null float64
BMI              768 non-null float64
DiabetesPedigreeFunction 768 non-null float64
Age              768 non-null int64
Outcome          768 non-null int64
dtypes: float64(6), int64(3)
```

However, there are many zeros in the dataset. These are medical predictors and these values cannot be zero. Pregnancies could be zero, but not BMI, insulin, glucose, Blood pressure etc. There are two ways to handle missing data or null values. First method is to remove the rows which has empty cells. Sometimes this method can be effective since there are no appropriate values to replace.

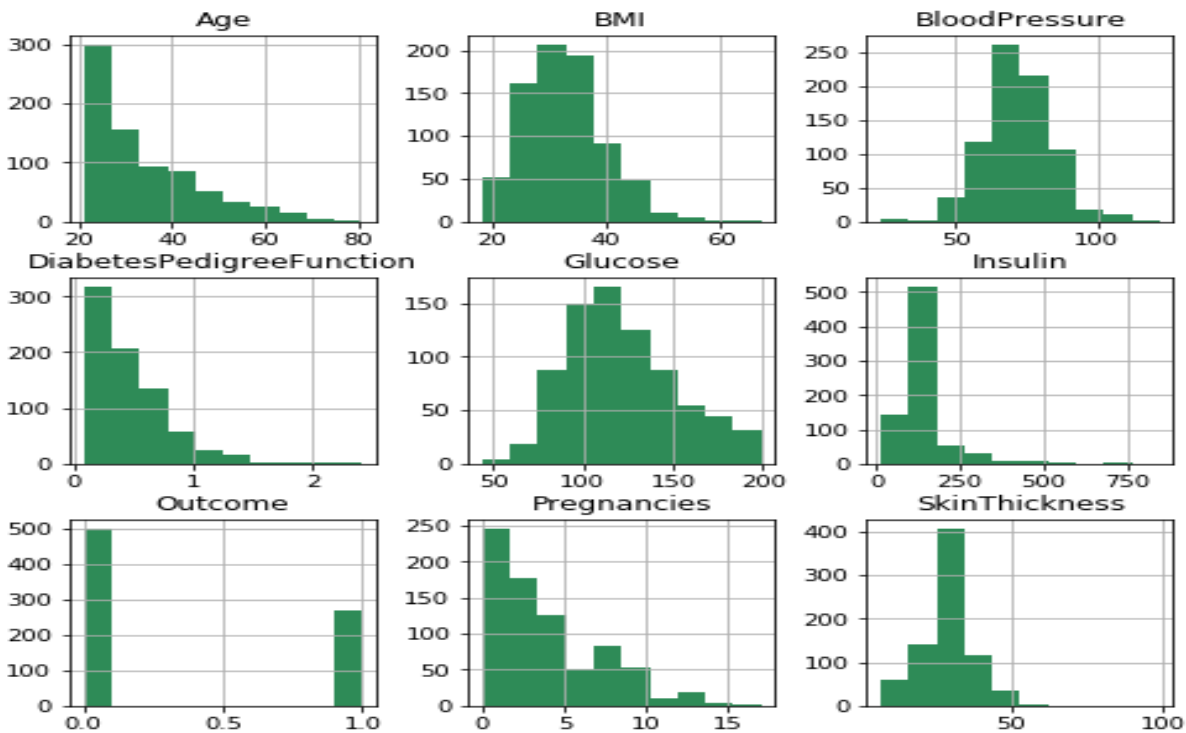
In our dataset, we have zero values for the following variables.

- Glucose
- Blood Pressure
- SkinThickness
- Insulin
- BMI

We need to find an appropriate statistic to replace with the zeros. The first step is to replace the zero values with Nan.

Diagnostically predict diabetes on Pima Indians dataset

Let's see the distribution of the variables.



Skewness of the data

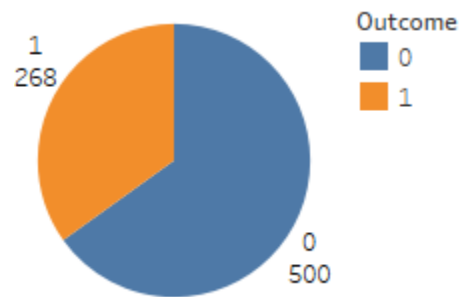
We observe a right skew distribution, in right skew distribution the mean is to the right of the peak. Also, our data has some outliers, although we cannot conclude this fact. For example, some patients have pregnancies more than 8-10 times. This seems highly unusual, but we cannot ignore its possibility. Hence, we fill the zero values with the median values rather than mean values.

Index	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	etesPedigreeFunc	Age	Outcome
count	768.00000	768.00000	768.00000	768.00000	768.00000	768.00000	768.00000	768.00000	768.00000
mean	3.84505	121.65625	72.38672	29.10807	140.67188	32.45521	0.47188	33.24089	0.34896
std	3.36958	30.43829	12.09664	8.79122	86.38306	6.87518	0.33133	11.76023	0.47695
min	0.00000	44.00000	24.00000	7.00000	14.00000	18.20000	0.07800	21.00000	0.00000
25%	1.00000	99.75000	64.00000	25.00000	121.50000	27.50000	0.24375	24.00000	0.00000
50%	3.00000	117.00000	72.00000	29.00000	125.00000	32.30000	0.37250	29.00000	0.00000
75%	6.00000	140.25000	80.00000	32.00000	127.25000	36.60000	0.62625	41.00000	1.00000
max	17.00000	199.00000	122.00000	99.00000	846.00000	67.10000	2.42000	81.00000	1.00000

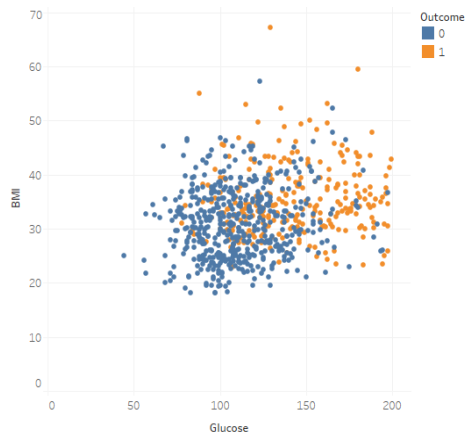
Diagnostically predict diabetes on Pima Indians dataset

Exploratory Data Analysis (EDA)

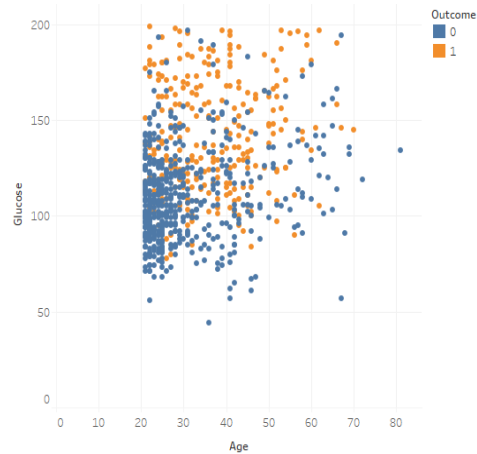
Outcome Variable – Out of 768 patients 268 have diabetes and 500 people are not affected.



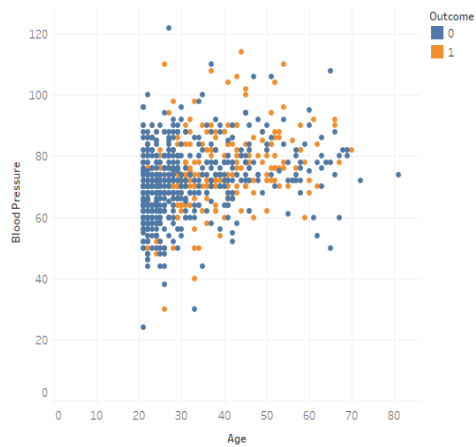
Glucose Vs BMI



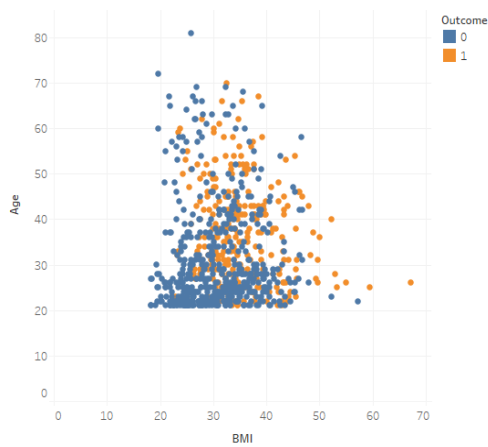
Glucose Vs Age



Age Vs Blood Pressure



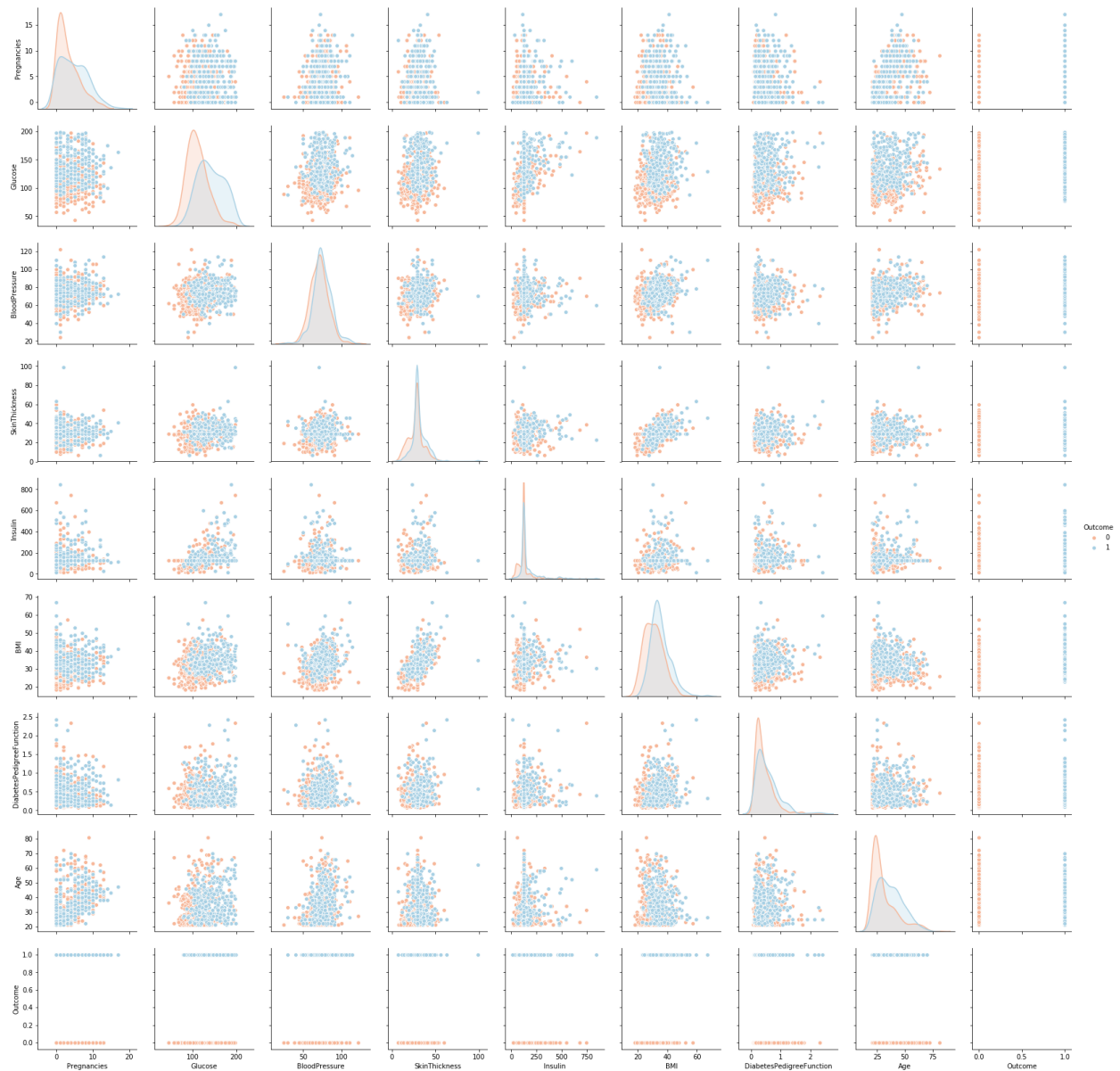
Age Vs BMI



Diagnostically predict diabetes on Pima Indians dataset

Scatter Plot of the variables – Before we saw scatter plots of few variables. The main factors for diabetes like glucose level, BMI, Age graphs which are highly correlated were shown before.

This is the scatter plot for each variable against all the variables.



Diagnostically predict diabetes on Pima Indians dataset

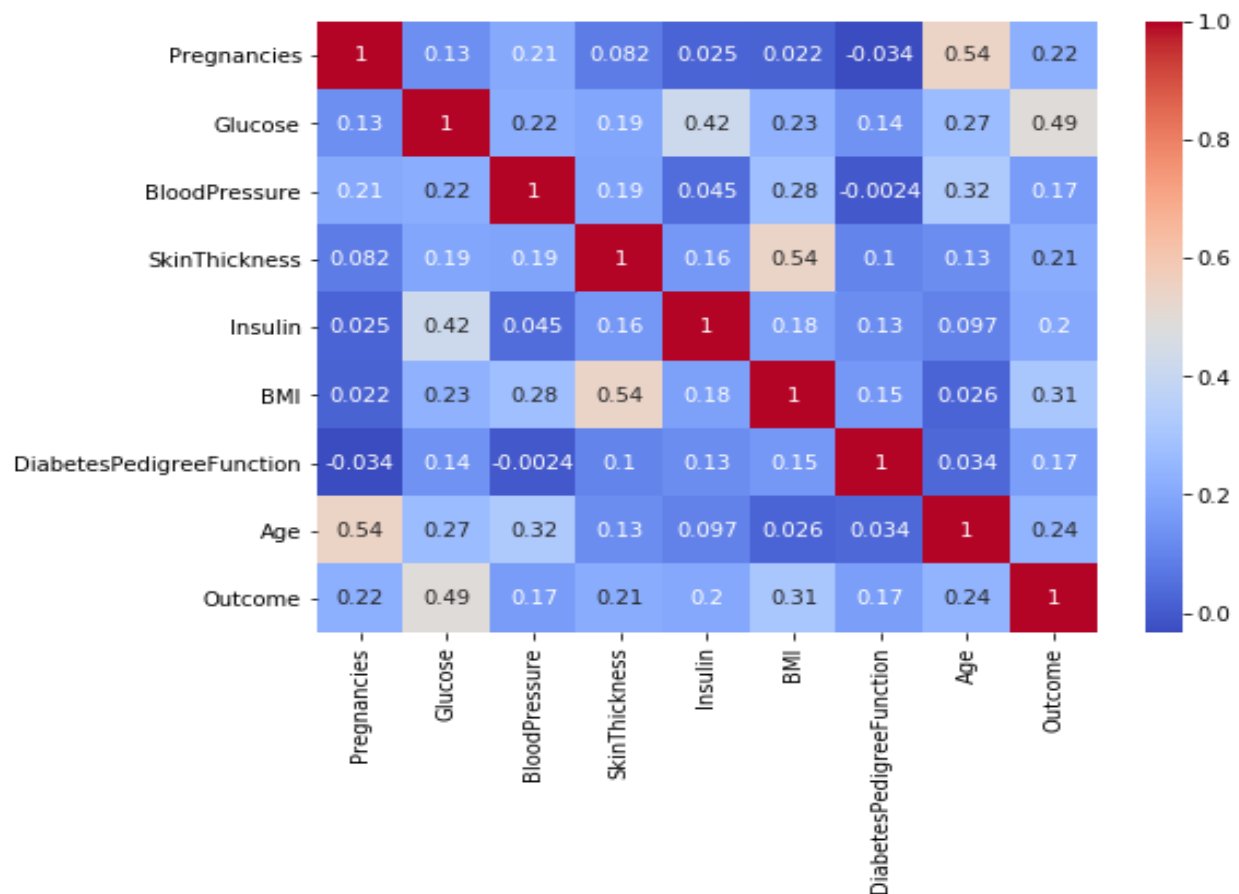
Correlation - In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data.

- 1 is a perfect positive correlation
- 0 is no correlation, the values don't seem linked at all.
- -1 is a perfect negative correlation

Formula to calculate correlation of two variables

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Heat map showing the correlation of all the variables.



Diagnostically predict diabetes on Pima Indians dataset

Data Scaling

Feature scaling is a method used to normalize the range of variables or features of data. In our data, there are many numerical variables and they are measured in different scales. For our modelling we need to scale those numbers into a normalized value.

The variables are normalized into a range from -1 to +1.

Scaling is done using Standard Scaler available in sklearn library.

```
from sklearn.preprocessing import StandardScaler sc = StandardScaler()
```

Dividing the Dataset into Train and Test

70 % of the data is taken for training the model and rest 30% is used for testing.

```
from sklearn.model_selection import train_test_split
```

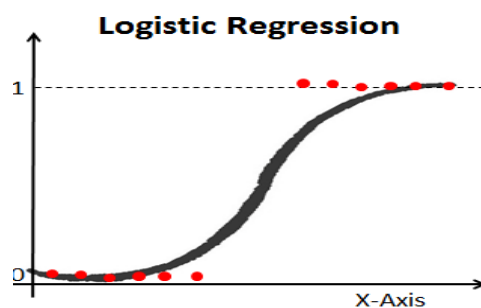
```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2)
```

Data Modelling

Algorithm used for modelling the data. We start with basic classification algorithm like Logistic Regression and Decision Trees, then implement ensemble methods like Random Forest, finally build an ANN model. We compare the model's efficiency and accuracy evaluating the best model for classification of diabetes patients.

- Logistic Regression
- Decision Trees
- Random Forest
- Naïve Bayes
- Support Vector Machines
- Artificial Neural Networks

1. Logistic Regression



Diagnostically predict diabetes on Pima Indians dataset

- The curve is the fit to the data to predict the probability that the patient has diabetes or not.
- The Y axis is the probability of the dependent variable from 0 to 1.
- The X axis are the independent variables.
- The Y axis in logistic regression is transformed from the probability of churn to the log (odds of churn).
- Now Y axis can go from -Infinity to +Infinity(Similar to Regression).
- This is done using the logit function. Given by $\log(p/1-p)$.
- Maximum likelihood is used to find the best fitting line.

Fitting the model

- The code for logistic regression model
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state = 0)
classifier.fit(x_train, y_train)

Evaluation of the model

Confusion Matrix - A **confusion matrix** is a table that is often used to describe the performance of a classification model.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

From the confusion matrix, we can calculate the following

Accuracy is given by: $(\text{True Positive} + \text{True Negative}) / \text{Total}$

Precision is given by $\text{True Positive} / (\text{True Positive} + \text{False Positive})$

Recall is given by $\text{True Positive} / (\text{True Positive} + \text{False Negative})$

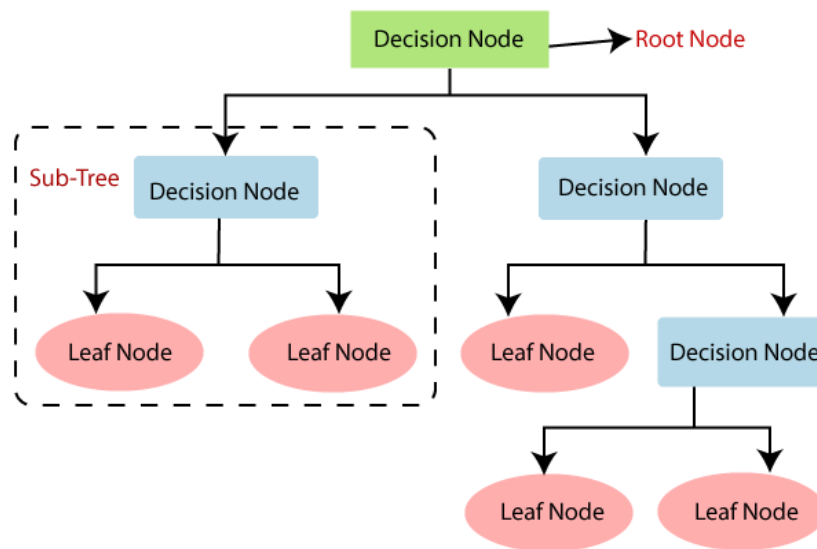
Diagnostically predict diabetes on Pima Indians dataset

Confusion Matrix for Logistic Regression

```
array([[140, 17],  
       [ 38, 36]], dtype=int64)
```

Accuracy	0.7619
Precision	0.8917
Recall	0.7865

2. Decision Tree Classifier



- The approach of decision trees is divide and conquer.
- It splits the data into subsets, which are split repeatedly into smaller subsets and so on till the algorithm determines the data within the subsets are homogenous.
- Using the decision algorithm, we start at the tree root and split the data on the feature that results in the largest information gain. In an iterative process, we can then repeat this splitting procedure at each child node until the leaves are pure.
- This means that the samples at each leaf node all belong to the same class.

Diagnostically predict diabetes on Pima Indians dataset

Fitting the Model

- The code for Decision Tree classifier

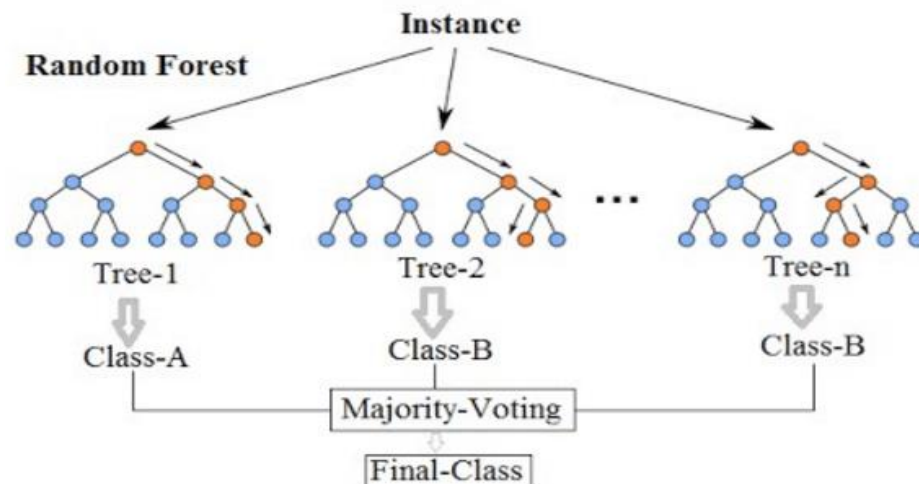
```
from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
classifier.fit(X_train, y_train)
```

Confusion Matrix for decision tree

```
array([[124, 33],
       [ 27, 47]], dtype=int64)
```

Accuracy	0.7402
Precision	0.7898
Recall	0.8211

3. Random Forest Classifier



- Random forest is an ensemble algorithm. Ensemble algorithms are which combine more than one algorithm of one kind or different kind.
- Random forest is a collection of decision trees.
- Random forest creates a subset of decision trees from randomly selected subset of training data. It aggregates the votes from different trees and decide the final class of the object.

Diagnostically predict diabetes on Pima Indians dataset

- Random forest assigns a weight for considering the impact of decision trees. Tree with high error rate are given low weight. The trees with low error rate have greater impact.

Fitting the model

- The code for Decision Tree classifier

```
from sklearn.ensemble import RandomForestClassifier  
  
classifier = RandomForestClassifier(n_estimators = 15, criterion = 'entropy',  
random_state = 0)  
  
classifier.fit(X_train, y_train)
```

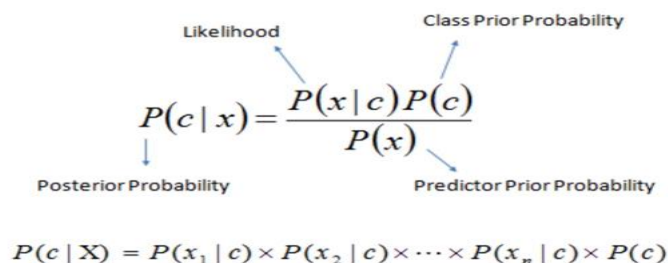
Confusion Matrix for Random Forest

```
array([[140, 17],  
       [ 36, 38]], dtype=int64)
```

Accuracy	0.7705
Precision	0.8917
Recall	0.7954

4. Naïve Bayes Classifier

Naive Bayes Classifier



The diagram illustrates the components of Bayes' Theorem for a Naive Bayes Classifier. It shows the formula $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$ with arrows pointing from each term to its corresponding label: $P(c|x)$ is labeled 'Posterior Probability', $P(x|c)$ is labeled 'Likelihood', $P(c)$ is labeled 'Class Prior Probability', and $P(x)$ is labeled 'Predictor Prior Probability'. Below the main formula, the joint probability formula is given: $P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Posterior Probability Likelihood Class Prior Probability Predictor Prior Probability

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Naïve Bayes classifier is a probabilistic data model based on Bayes theorem.

Bayes Theorem is given by $P(A|B) = P(A) P(B|A) / P(B)$

- Using Bayes theorem, we can find the probability of **A** happening, given that **B** has occurred.
- Here, **B** is the evidence and **A** is the hypothesis.
- The assumption made here is that the predictors/features are independent.

Diagnostically predict diabetes on Pima Indians dataset

- That is presence of one feature does not affect the other.

Fitting the model

- The code for Naïve Bayes classifier
from sklearn.naive_bayes import GaussianNB
classifier = GaussianNB()
classifier.fit(X_train, y_train)

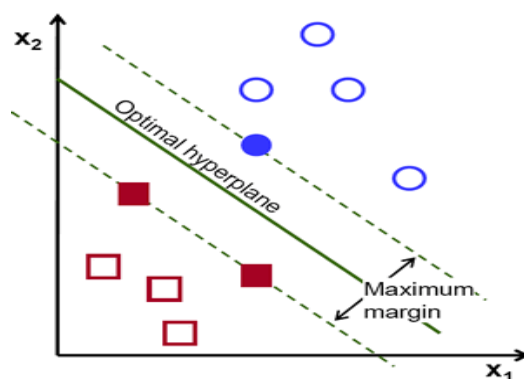
Confusion Matrix for Naïve Bayes Classifier

```
array([[137, 20],  
       [ 35, 39]], dtype=int64)
```

Accuracy	0.7619
Precision	0.8726
Recall	0.7965

There are different types of Naïve Bayes classifiers like Multinomial, Bernoulli, Gaussian Naïve Bayes. In our project we use the Gaussian Naïve Bayes classifier. This is because the predictors are continuous value. We assume that the values are sampled from Gaussian distribution. Bernoulli classifier is used when the predictors take up Boolean or discrete values.

5. Support Vector Machines



- In support vector machines, the objective is to find a hyperplane in an N-dimensional space that distinctly classifies the data points in our dataset.
- To separate the two classes of data points, there are many possible hyperplanes that could be chosen.

Diagnostically predict diabetes on Pima Indians dataset

- Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes.
- Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes.

Fitting the Model

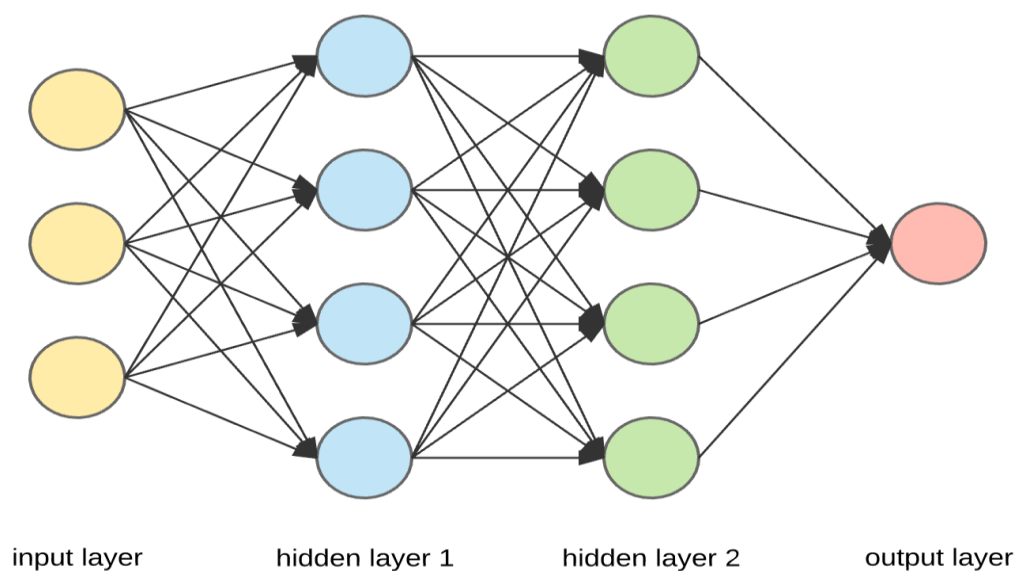
- The code for SVM classifier
from sklearn.svm import SVC
classifier = SVC(kernel = 'rbf', random_state = 0)
classifier.fit(X_train, y_train)

Confusion Matrix for SVM

```
array([[141, 16],  
       [ 42, 32]], dtype=int64)
```

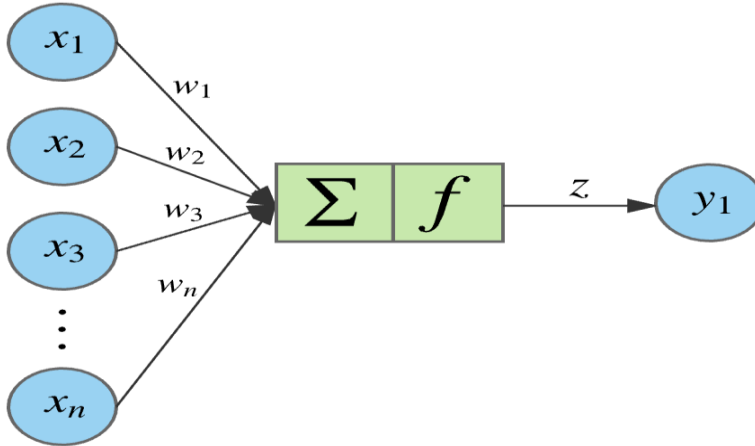
Accuracy	0.7489
Precision	0.8980
Recall	0.7704

6. Artificial Neural Networks (ANN)



Diagnostically predict diabetes on Pima Indians dataset

- ANN consist of an input layer, multiple hidden layers, and an output layer.
- Every node in one layer is connected to every other node in the next layer.



- This is a single node of the Artificial Neural Network.
- A given node takes the weighted sum of its inputs and passes it through a non-linear activation function. This is the output of the node, which then becomes the input of another node in the next layer.
- The signal flows from left to right, and the final output is calculated by performing this procedure for all the nodes.
- Training this deep neural network means learning the weights associated with all the edges.

The equation of the node

$$z = f(x \cdot w) = f \left(\sum_{i=1}^n x_i w_i \right)$$

$$x \in d_{1 \times n}, w \in d_{n \times 1}, z \in d_{1 \times 1}$$

- The weighted sum of its inputs passed through an activation function. It can be represented as a vector dot product. N is the number of inputs to the nodes.
- The ANN randomly initializes the weights. Then it performs a forward propagation using the weights calculates the output of each node.
- The error rate is measured by comparing actual and predicted values using a loss function.

Diagnostically predict diabetes on Pima Indians dataset

- Then after this calculation it performs backward propagation and calculates each weights contribution to the error.
- The goal is to minimize the error rate.
- Backpropagation is mainly done using Gradient Descent.

Fitting the Model

- The Dense function in Keras constructs a fully connected ANN.
- We can use different activation functions, the most simple is the Sigmoid function, that is used for Logistic Regression. We used the tanh function in our project.
- We have a optimizer which is used to minimize the loss function. We use the adam optimizer in our project.
- We use a loss function to minimize the loss. We use Binary Cross Entropy.

The code for ANN

Initializing the ANN

```
classifier = Sequential()
```

Adding the input layer and the first hidden layer

```
classifier.add(Dense(units = 6, kernel_initializer = 'uniform', activation = 'relu', input_dim = 8))
```

Adding the output layer

```
classifier.add(Dense(units = 1, kernel_initializer = 'uniform', activation = 'sigmoid'))
```

Compiling the ANN

```
classifier.compile(optimizer = 'adam', loss = 'binary_crossentropy', metrics = ['accuracy'])
```

Fitting the ANN to the Training set

```
classifier.fit(X_train, y_train, batch_size = 10, nb_epoch = 100)
```

Confusion Matrix for ANN

```
array([[136, 21],  
       [ 32, 42]], dtype=int64)
```

Accuracy	0.7706
Precision	0.8662
Recall	0.8095

Diagnostically predict diabetes on Pima Indians dataset

Model Evaluation

Model	Accuracy	Precision	Recall
Logistic Regression	0.7619	0.8917	0.7865
Decision Tree	0.7402	0.7898	0.8211
Random Forest	0.7705	0.8917	0.7954
Naïve Bayes	0.7619	0.8726	0.7965
SVM	0.7489	0.8980	0.704
ANN	0.7706	0.8662	0.8095

Conclusion

Here we observe that ANN has the best Accuracy compared to all the other models. These values should be close to one for the model to be efficient. Followed by Random Forest which is an ensemble algorithm. However, these models are efficient, but they are complicated and would need hardware and memory resources when dealing with large scale datasets.

We successfully predicted whether a patient could have diabetes, based on their body parameters on pima Indians dataset. In our project we classified between patients who has diabetes or not. Six different models were developed for classification and comparison was done based on performance parameters. These data models are extremely effective in conducting Predictive analysis. This prediction would be extremely useful as necessary measures can be taken by individuals to be healthy. Prediction of this disease in advance can help people to adapt healthy lifestyle.

Diagnostically predict diabetes on Pima Indians dataset

References

Dataset Link - <https://data.world/data-society/pima-indians-diabetes-database>

<https://diabetes.diabetesjournals.org/content/53/5/1181>

<https://care.diabetesjournals.org/content/29/8/1866>

<https://towardsdatascience.com/model-evaluation-techniques-for-classification-models-eac30092c38b>

<https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35>

<https://machinelearningmastery.com/loss-and-loss-functions-for-training-deep-learning-neural-networks/>

<https://www.mathsisfun.com/data/correlation.html>

<https://missinglink.ai/guides/neural-network-concepts/7-types-neural-network-activation-functions-right/>

<https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes>

<https://www.statisticshowto.com/probability-and-statistics/skewed-distribution/>

https://en.wikipedia.org/wiki/Feature_scaling