Diabetes Prediction

Final Report

Team members:

1) Akshay Rajendra Mutha - arm8@njit.edu

2) Hitarthi Shirish Mokashi - hm57@njit.edu

3) Nikhil Teja Vedagiri - nv39@njit.edu

Introduction

Diabetes is the world's fourth leading cause of death, as well as a cause of renal disease, blindness, and heart disease. Data mining techniques assist doctors in making accurate diagnoses and treating diseases, reducing the workload of specialists. The goal of this study was to use data mining techniques to predict diabetes. In the model we used various types of data modelling techniques which helped us compare to different models and achieve better results.

Dataset Description

The data was taken from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset's goal is to diagnose whether a patient has diabetes using diagnostic metrics included in the collection.The selection of these cases from a wider database was subjected to many limitations. All of the patients at this clinic are Pima Indian women who are at least 21 years old.There are various medical predictor factors in the dataset, as well as one goal variable, Outcome.The pregnancies number the patient has had, their BMI, insulin level, age, and other factors are all variables predictor .There are 768 records and 9 features columns in the our dataset.Each feature can be either be integer or float data type and Some features have zero values which represent missing values in data.

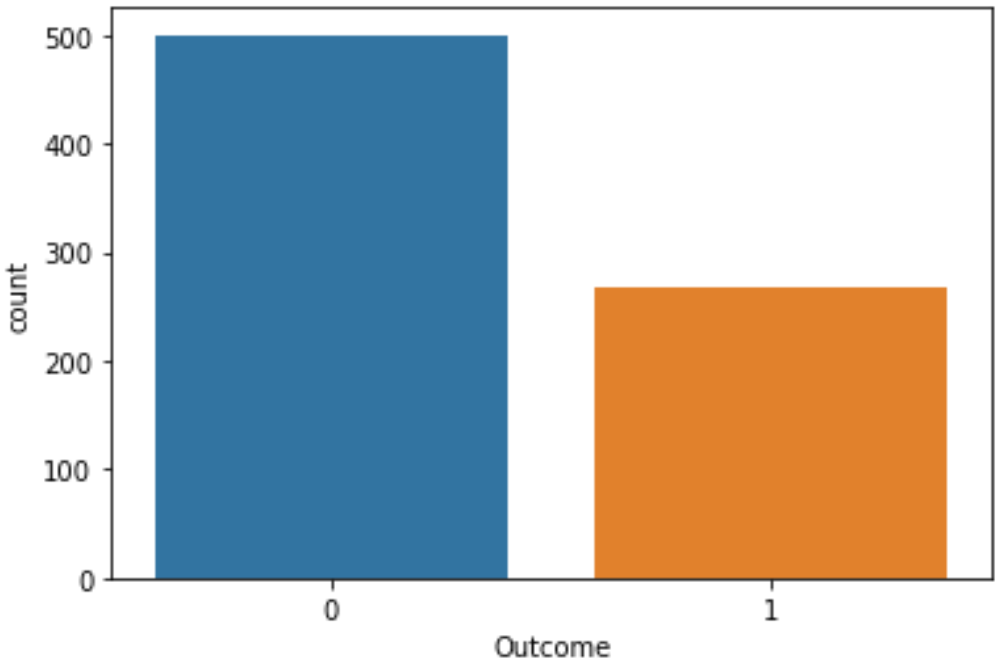
Dataset URL: https://www.kaggle.com/uciml/pima-indians-diabetes-database

Data Visualisation

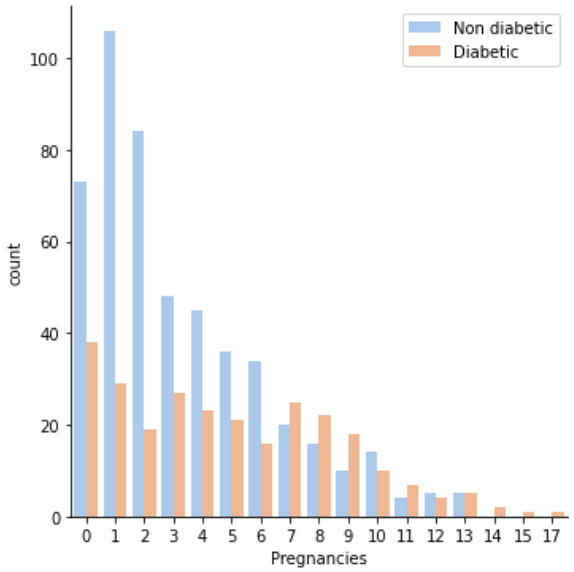
The graphical description of information and data is known as data visualization. Data visualization tools make it easy to examine and comprehend trends, outliers, and patterns in data by engaging visually like charts, graphs, and maps.

1. **Count plot**: A count plot is a histogram across a category variable rather than a quantitative variable. You can compare counts across nested variables using the same basic API and settings as for barplot().

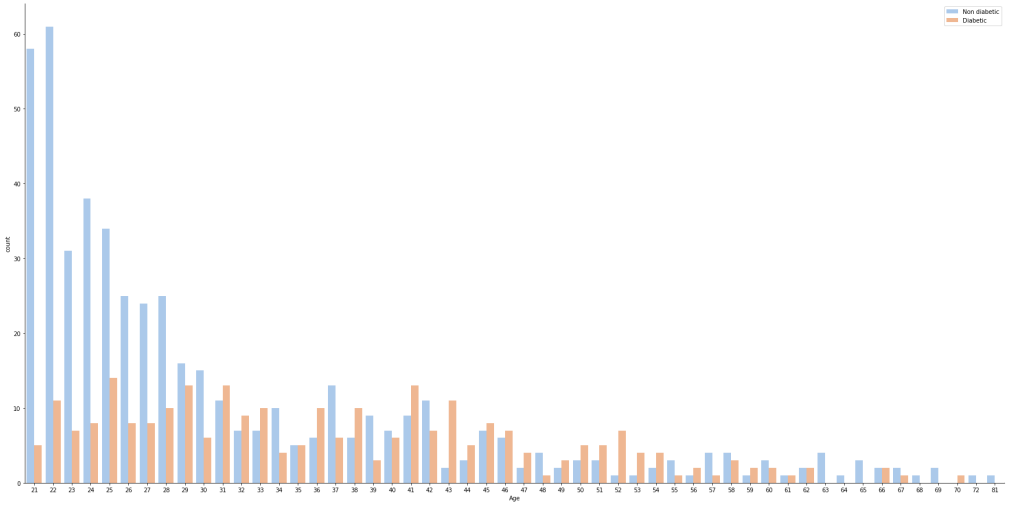
Here we made a countplot having outcomes 0 or 1 where 0 represents non diabetic and 1 represents diabetic.



2. **Catplot**: This function gives you access to numerous axes-level functions that use one of several visual representations to depict the relationship between a number and one or more category variables. Here we made a catplot to show how many pregnant women suffered diabetes.

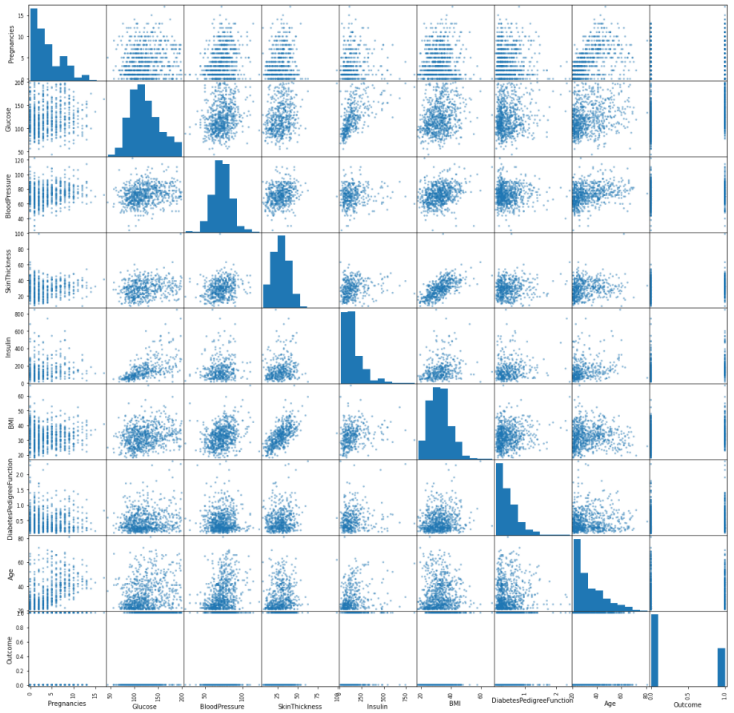


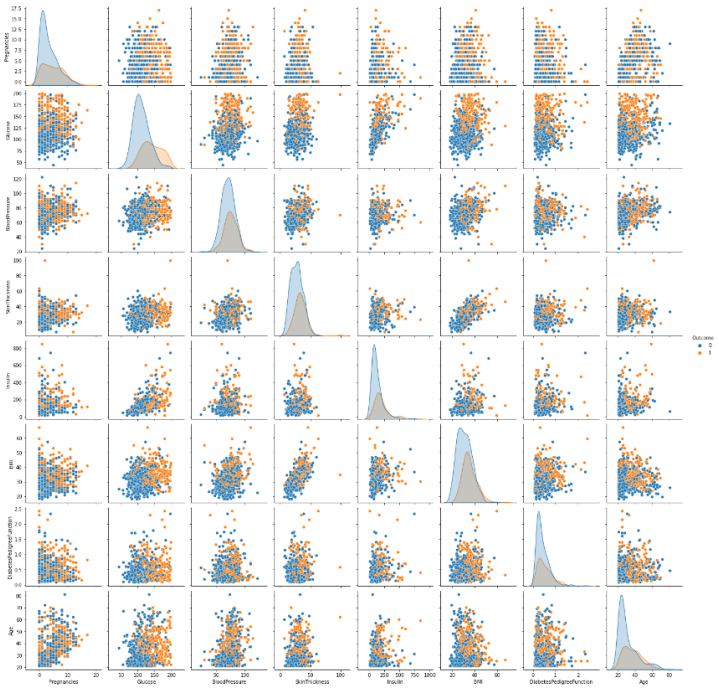
Here, we showed how many people tend to have diabetes compared to older people who tend to have diabetes at a young age.



3. Histogram

A histogram is a visual representation of data presented in the form of groupings. It is a precise approach for displaying numerical data distribution graphically. It's a type of bar plot. Here we used a scatterplot and pair plot matrix which showed us all the features and also to see the relationship between each other in a pair plots. This is illustrated as shown in the below graphs which is a scatter plot and a pairplot graph respectively.





4. Heatmap

A heatmap is a two-dimensional graphical representation of data that uses colors to represent the individual values in a matrix. From the heatmap we can observe that there is a greater correlation between Outcome and Glucose, BMI, Insulin. So we

select these features to accept the input from the user and pass it to our model which helps to predict the outcome.

Data preprocessing

Data Preprocessing is a data mining technique used in transformation of raw data into a usable and efficient format.It is an intrinsic step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; therefore, it is tremendously important that we preprocess our data before feeding it into our model. The following are the steps involved in data preprocessing:

1. **Data cleaning**: Data can contain a lot of useless and missing information. Data cleaning is carried out in order to handle this component. It entails dealing with missing data, noisy data, and so on.
    - Handling Null Values: In anydataset, there are always few null values. It doesn't really matter whether it is a regression, classification or any other kind of problem, no model can handle these NULL or NaN values on its own so we need to come about.

- Standardization: It is another intrinsic preprocessing step. we transform our values such that the mean of the values is replaced by NULL or NAN.

- Handling Categorical Variables: It is another intrinsic aspect of Machine Learning.Categorical variables are basically the variables that are discrete and not continuous.

- One-Hot Encoding: So in One-Hot Encoding what we essentially do is that we create ’n’ columns where n is the number of unique values that the nominal variable can take.

- Multicollinearity: It occurs in our dataset because we have features that are firmly dependent on each other.

- Missing Data: This circumstance occurs when some data in the data is missing. It can be dealt with in a variety of ways.

  For example,we find the total number of missing values in the dataset and check if there is any dependency of missing values on other columns.There are several options for completing this challenge. You have the option of manually filling the missing values, using the attribute mean, or using the most likely value.

2. **Data Transformation**: This step is conducted to transform the data into a format that can be used in the mining process. We are using a function called MinMaxScaler and what it does is that it retains the shape of the original distribution of data. It doesn't change the information embedded in the original data. It doesn't reduce the importance of outliers. The default range for the feature MinMaxScaler is 0 to 1.

3. **Data Reduction**

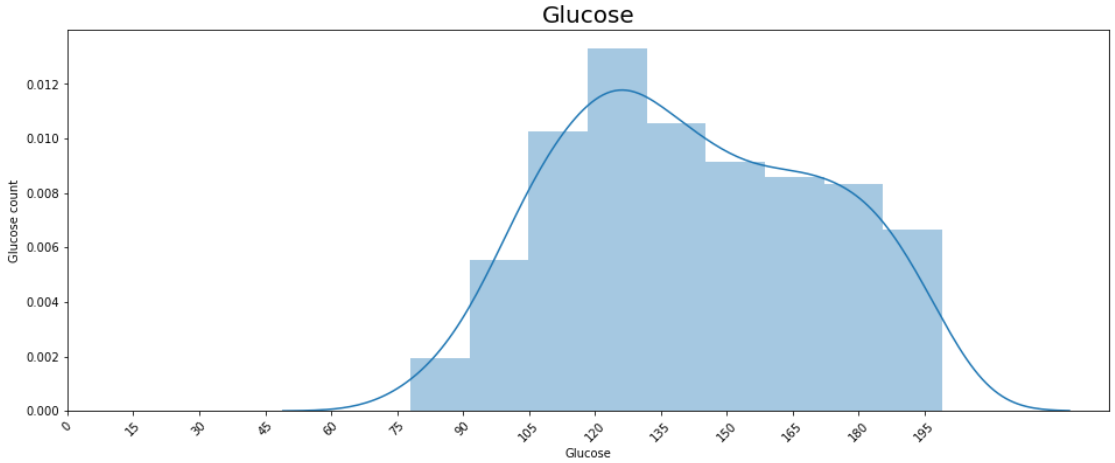
   In data mining we deal with large amounts of data. While dealing with large amounts of data, analysis becomes more heavy. So we use data reduction techniques to get rid

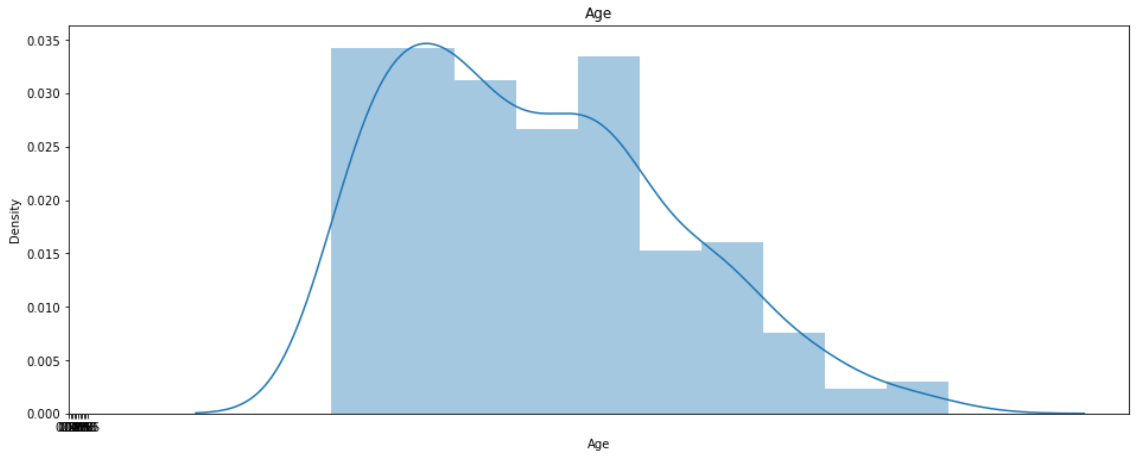
of this hurdle. Its goal is to simply to improve storage efficiency by lowering data storage and analysis expenses.

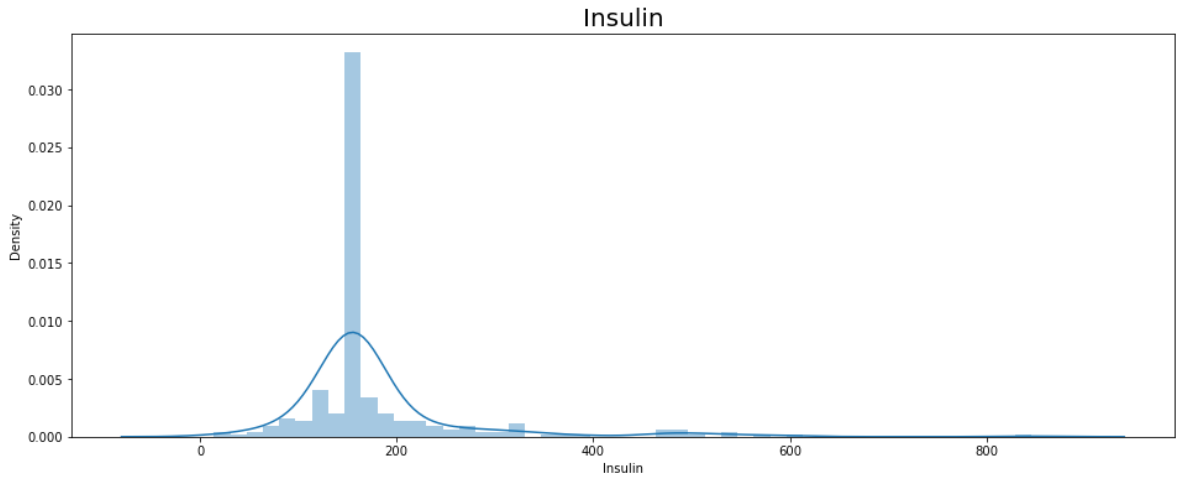
4. Exploratory data analysis

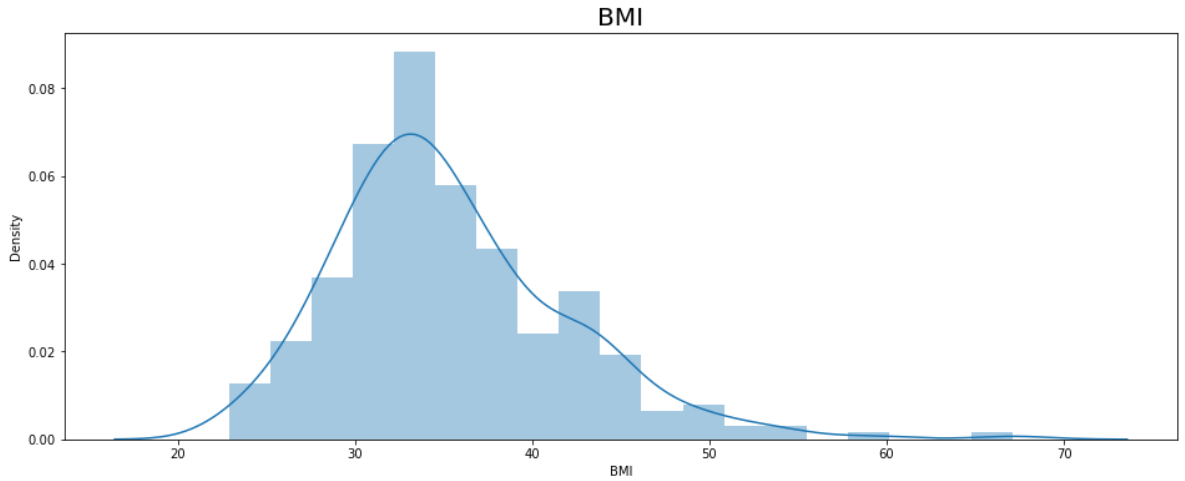
With the utilization of outline insights and graphical representations, exploratory information examination alludes to the vital preparation of doing to begin with examinations on information in order to reveal designs, spot irregularities, test speculations, and check assumptions.

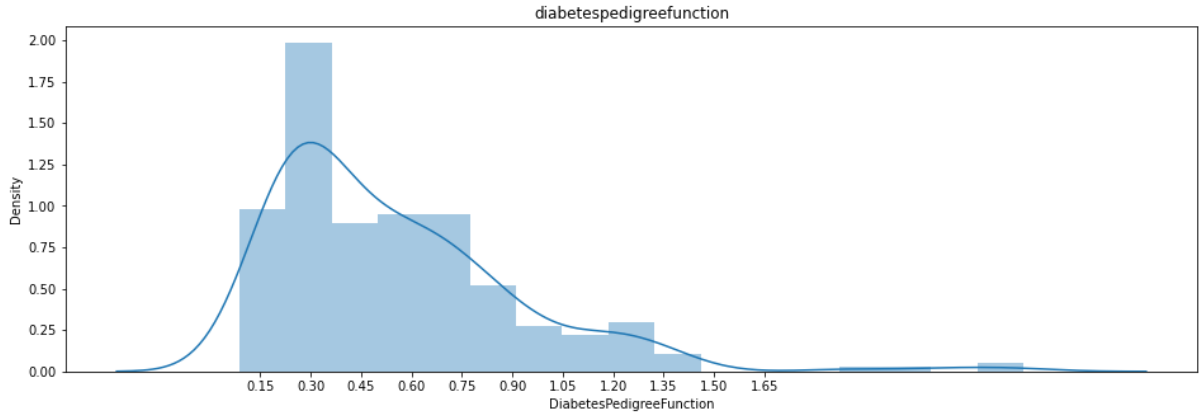
Here, we utilized dissemination plot to examine the highlights which appeared most elevated relationship with the Result i.e. Glucose, BMI, Affront, etc. A dispersion plot, too known as a Distplot, shows the variety in information dissemination. The whole dispersion of persistent information factors appears by the Seaborn Distplot.











There were various actions performed on the dataset like removing the unnecessary columns and splitting the data into train and test.The split we followed is a testing data 20% and training data 80%.Then we are finding the shapes of train and test split data.We obtain the results of the train and test shapes as follows.

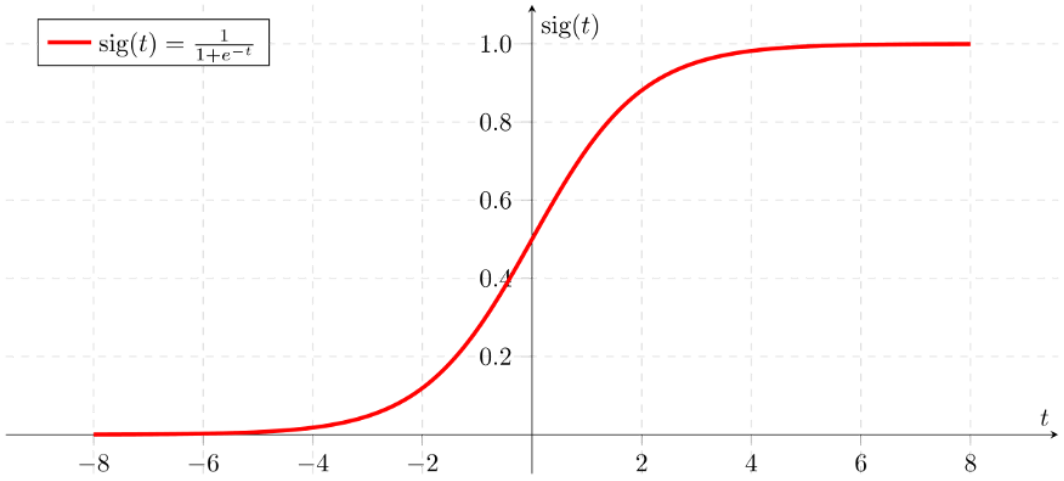
x_train shape: (614, 8) ,x_test shape: (154, 8), y_train shape: (614,), y_test shape: (154,)

Data Modelling

1. Logistic regression

Calculated Relapse could be a Machine Learning calculation that's utilized to solve classification issues. It could be a prescient explanatory approach that's based on the likelihood idea. The classification calculation calculated relapse is utilized to dole out perceptions to a discrete set of classes. The calculated sigmoid work deciphers the yield of calculated relapse into a likelihood esteem. A Calculated Relapse show is comparable to a Straight Relapse demonstration, but the Logistic Relapse employs a more complex task, which is known as the 'Sigmoid function' or the 'logistic function' rather than a direct work. The calculated relapse speculation proposes that the taken toll work be restricted to a esteem between 0 and 1.

This is the Logistic function or the Sigmoid function.



2. Support Vector Classifier

Back vectors are information focuses that are closer to the hyperplane and have an impact on the hyperplane's position and introduction. We maximize the classifier's edge by utilizing these back vectors. The hyperplane's position will be changed in the

event that the back vectors are erased. These are the focuses that will help us in building our SVM.

3. **Naive Bayes Algorithm**

   It's a classification strategy based on Bayes' Hypothesis and the presumption of indicator autonomy. A Gullible Bayes classifier, in straightforward terms, states that the presence of one include in a course is disconnected to the nearness of any other highlight.

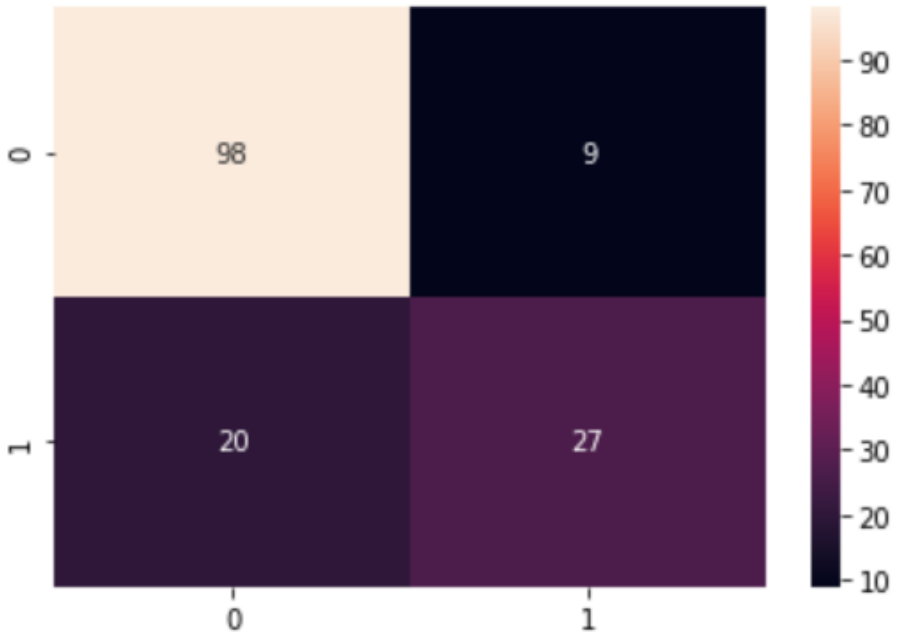
4. **Decision tree Algorithm**

   The foremost capable and broadly utilized instrument for categorization and expectation is the choice tree. A decision tree could be a flowchart-like tree structure in which each inside hub speaks to a trait test, each department reflects the test's conclusion, and each leaf hub (terminal hub) stores a lesson label.

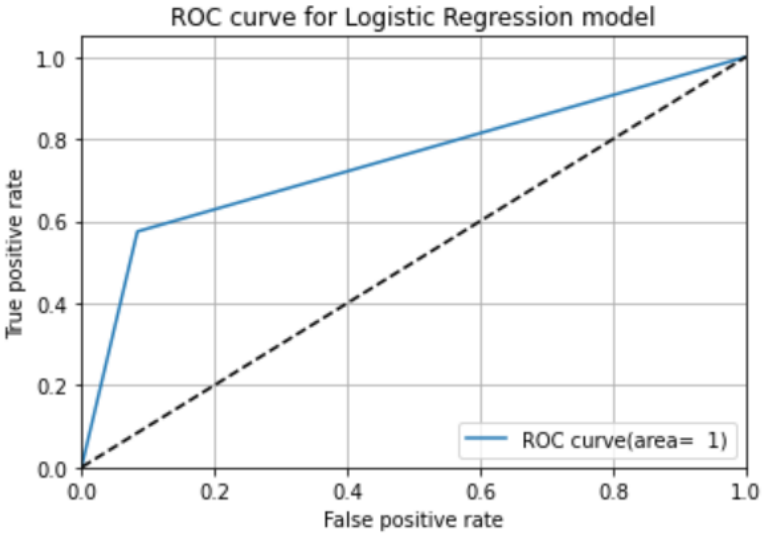
Model evaluation

We are progressing to utilize distinctive sorts of models to see what precision score we got by employing a work called accuracy_score.It computes the exactness, either the division (default) or the number (normalize=False) of adjusted forecasts. In multilabel classification, the work returns the subset accuracy. For all the models used, we gotten the taking after exactness for each of the models as recorded below(rounded to 2 decimal places)

- Logistic Regression: 81.16%
- Support Vector Classifier: 80.32%
- Naive Bayes: 79.87%
- Decision tree: 74.02%



The over chart may be a disarray framework heatmap for the calculated relapse calculation. We utilize this chart to provide coordinate comparisons of values like Genuine Positives, Wrong Positives, Genuine Negatives and Wrong Negatives. Where the over heatmap says 0,0 implies genuine negative and 1,1 implies genuine positive and 0,1 implies indeed individual is negative but appearing result positive and 1,0 implies individual is positive but shows negative so it's unsafe so we got to be more exact in our demonstration.



The over chart may be a ROC chart for a logistic relapse demonstrated. AUC-ROC bend may be an execution estimation for the classification issues at different limit settings. ROC may be a likelihood bend and AUC speaks to the degree or degree of distinguishableness. It tells how much the demonstrate is competent of recognizing between classes. Higher the AUC, the way better the show is at foreseeing classes as and 1 classes as 1. By relationship, the Higher the AUC, the better the model is at distinguishing between patients with the diabetic and not diabetic.An fabulous demonstration has AUC near to the 1 which suggests it features a great degree of distinguishableness. A destitute demonstrator has an AUC close which implies it has the most noticeably awful degree of distinctness. In reality, it implies it is responding to the result. It is anticipating 0s as 1s and 1s as 0s. And when AUC is 0.5, it implies the show has no course division capacity whatsoever.

Conclusion

Performing this project, we learned various Data modelling techniques. We learned the process of going through the raw data till the final data model evaluation step. Out of the 4 modelling techniques used, we found that the **Logistic Regression** modelling technique gave out the best results serving us with the accuracy of **81.16%**. Thus, we decided to keep Logistic regression as our final data modelling technique for the project.