GLA UNIVERSITY, MATHURA

MINI PROJECT

(2020-2021)

DIABETES PREDICTION

MID-TERM REPORT

GLA University, Mathura

Institute of Engineering and Technology

Submitted By:-

1.Janvi Pangoriya(181500292)

2.Nidhi Gupta(181500422)

Under the Supervision

Of

Prof. Mohd. Amir Khan

Technical Trainer Department of Computer Engineering and Applications

CONTENTS

ABSTRACT

In this project we are creating a machine learning model for the prediction of diabetes in human beings and this model will be deployed on the cloud so as to provide a web based application to the user where the user can enter the parameters on which the prediction is being made and the model will return whether the person is diabetic or not. 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. Due to this, it is necessary to predict the disease and prevent further any health issues. Over time, high blood glucose leads to problems such as heart disease, stroke, kidney disease, eye problems, dental disease, nerve damage, foot problems etc. Although diabetes has no cure, you can take steps to manage your diabetes and stay healthy.

We are trying to predict the diabetes based on the following parameters in this project. There are 8 factors we have taken into consideration in this project. They are pregnancies (no of times the woman has been pregnant, the glucose concentration (Plasma glucose concentration a 2 hours in an oral glucose tolerance test), diastolic blood pressure (pressure in the arteries when the heart rests between beats. This is the time when the heart fills with blood and gets oxygen. A normal diastolic blood pressure is lower than 80. A reading of 90 or higher means you have high blood pressure), skin thickness (Triceps skin fold thickness), Insulin (2-Hour serum insulin), Body mass index (weight in kg / (height in m) ^ 2), Diabetes pedigree function, Age (years). The last column is the outcome which is the class variable which is a dependent variable results in 0 if the person is not diabetic and 1 if the person is diabetic.

INTRODUCTION

Diabetes is noxious diseases in the world. Diabetes caused because of obesity or high blood glucose level, and so forth. It affects the hormone insulin, resulting in abnormal metabolism of crabs and improves level of sugar in the blood. Diabetes occurs when body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffering from diabetes particularly from low or idle income countries. And this could be increased to 490 billion up to the year of 2030. However prevalence of diabetes is found among various Countries like Canada, China, and India etc. Population of India is now more than 100 million so the actual number of diabetics in India is 40 million. Diabetes is major cause of death in the world. Early prediction of disease like diabetes can be controlled and save the human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose we use the Pima Indian Diabetes Dataset, we apply various Machine Learning classification and ensemble Techniques to predict diabetes. Machine Learning is a method that is used to train computers or machines explicitly. Various Machine Learning Techniques provide efficient result to collect knowledge by building various classification and ensemble models from collected dataset. Such collected data can be useful to predict diabetes. Various techniques of Machine Learning can capable to do prediction, however it's tough to choose best technique. Thus for this purpose we apply popular classification and ensemble methods on dataset for prediction.

PROPOSED METHODOLOGY

Goal of the project is to investigate for model to predict diabetes with better accuracy.

A. Dataset Description- the data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset have many attributes of 768 patients. The 9th attribute is class variable of each data points. This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics. Distribution of Diabetic patient- We made a model to predict diabetes however the dataset was slightly imbalanced having around 500 classes labelled as 0 means negative means no diabetes and 268 labelled as 1 means positive means diabetic.

Table 1: Dataset Description

S No.	Attributes
1	Pregnancy
2	Glucose
3	Blood Pressure
4	Skin thickness
5	Insulin
6	BMI(Body Mass Index)
7	Diabetes Pedigree Function
8	Age

B. Data Preprocessing- Data pre-processing is most important process. Mostly healthcare related data contains missing vale and other impurities that can cause effectiveness of data. To improve quality and effectiveness obtained after mining process, Data pre-processing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate result and successful prediction. For Pima Indian diabetes dataset we need to perform pre-processing in two steps.

1). Missing Values removal- Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces dimensionality of data and help to work faster.

2). Splitting of data- After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train algorithm on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically aim of normalization is to bring all the attributes under same scale

C. Apply Machine Learning- When data has been ready we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes. The methods applied on Pima Indians diabetes dataset. Main objective to apply Machine Learning Techniques to analyze the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in

prediction. The following classification algorithm will be applied like Logistic Regression, K Nearest Neighbours, Random Forest, Decision Tree and many more to find the best accuracy.

D. MODEL BUILDING: This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms which are discussed above for diabetes prediction.

Procedure of Proposed Methodology

Step1: Import required libraries, Import diabetes dataset.

Step2: Pre-process data to remove missing data.

Step3: Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.

Step4: Select the machine learning algorithm i.e. K Nearest Neighbour, Support Vector Machine, Decision Tree, Logistic regression, Random Forest and Gradient boosting algorithm.

Step5: Build the classifier model for the mentioned machine learning algorithm based on training set.

Step6: Test the Classifier model for the mentioned machine learning algorithm based on test set.

Step7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.

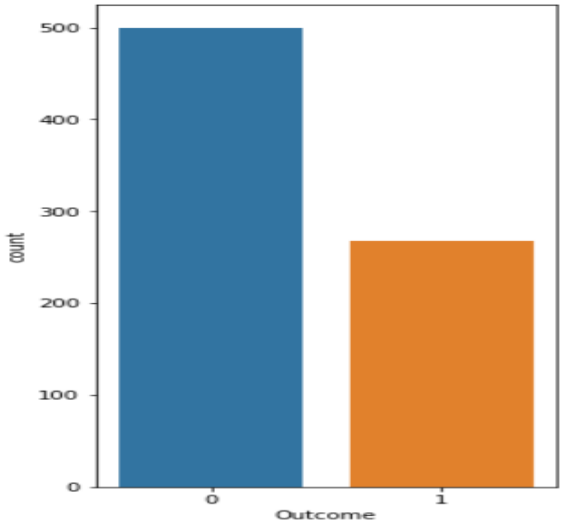
Step8: After analyzing based on various measures conclude the best performing algorithm.

E. Deploying the model on Cloud: After the best performing algorithm is selected ,the built model is deployed on cloud on the Heroku platform and all the calculations will be taking place on cloud and we will be getting a live link to the web based application where on entering the parameters the model will predict whether the patient is diabetic or not. The front-end part of the model is designed using HTML and CSS and the backend uses Flask .

PROBLEM STATEMENT

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. The accuracy will different for every model when compared to other models. The project work will give the accurate or higher accuracy model shows that the model is capable of predicting diabetes effectively.

To solve this problem efficiently we have collected a dataset from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset have many attributes of 768 patients. The 9th attribute is class variable of each data points. This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics. Distribution of Diabetic patient- We made a model to predict diabetes however the dataset was slightly imbalanced having around 500 classes labelled as 0 means negative means no diabetes and 268 labelled as 1 means positive means diabetic. The graph below shows the graphical representation of the dataset we have collected.



OBJECTIVE

We are creating a machine learning model for the prediction of diabetes in human beings and this model will be deployed on the cloud so as to provide a web based application to the user where the user can enter the parameters on which the prediction is being made and the model will return whether the person is diabetic or not. We are trying to predict the diabetes based on the following parameters in this project. There are 8 factors we have taken into consideration in this project. They are pregnancies (no of times the woman has been pregnant, the glucose concentration (Plasma glucose concentration a 2 hours in an oral glucose tolerance test), diastolic blood pressure (pressure in the arteries when the heart rests between beats. This is the time when the heart fills with blood and gets oxygen. A normal diastolic blood pressure is lower than 80. A reading of 90 or higher means you have high blood pressure), skin thickness (Triceps skin fold thickness), Insulin (2-Hour serum insulin), Body mass index (weight in kg / (height in m) ^ 2), Diabetes pedigree function, Age (years). The last column is the outcome which is the class variable which is a dependent variable results in 0 if the person is not diabetic and 1 if the person is diabetic.

IMPLEMENTATION DETAILS

Diabetes Prediction is a machine learning application which predicts whether the person is diabetic or not and this is implemented on cloud so that the user is provided with live link as a web based application and can enter various parameters on which the model predicts and gives the result. For implementing this project, we are using machine learning and cloud computing and the following flowchart represents our workflow.

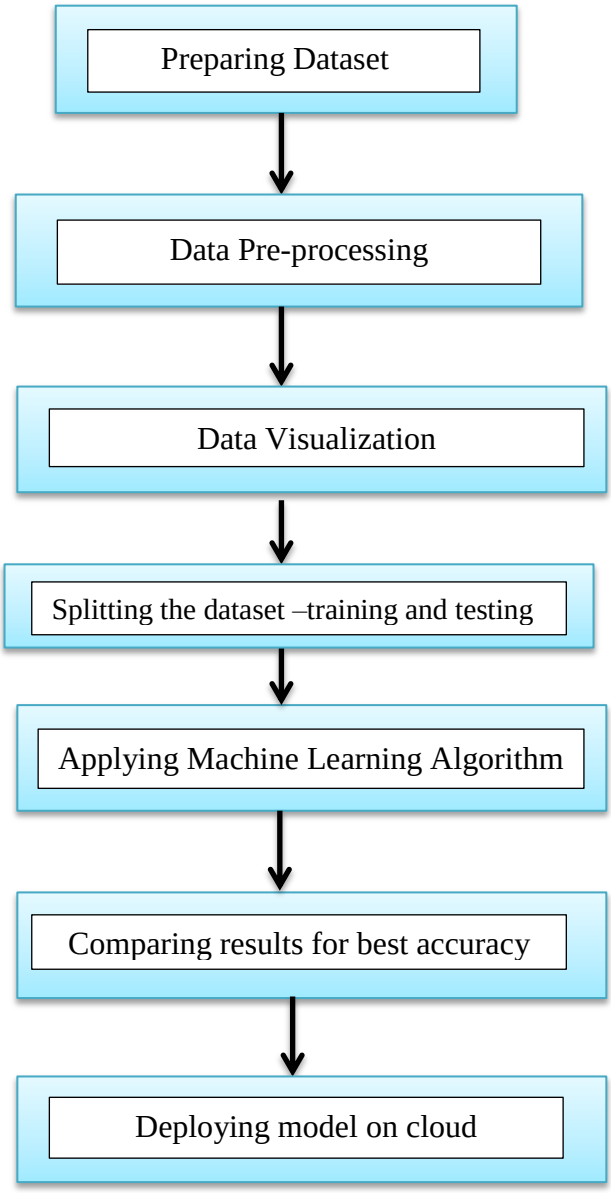


Fig 2: Flowchart showing the implementation details

Using Machine Learning for Diabetes Prediction

Diabetes occurs when body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffering from diabetes particularly from low or idle income countries. And this could be increased to 490 billion up to the year of 2030. However prevalence of diabetes is found among various Countries like Canada, China, and India etc. Population of India is now more than 100 million so the actual number of diabetics in India is 40 million. Diabetes is major cause of death in the world. Early prediction of disease like diabetes can be controlled and save the human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose we use the Pima Indian Diabetes Dataset, we apply various Machine Learning classification and ensemble Techniques to predict diabetes. Machine Learning is a method that is used to train computers or machines explicitly. Various Machine Learning Techniques provide efficient result to collect knowledge by building various classification and ensemble models from collected dataset. Such collected data can be useful to predict diabetes. Various techniques of Machine Learning can capable to do prediction, however it's tough to choose best technique. Thus for this purpose we apply popular classification and ensemble methods on dataset for prediction.

- **Collecting the data into the Dataset**

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variable includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

- **Data Pre-processing**

Basically it is the process of making the data ready to be fed to the machine learning algorithm. This includes the cleaning of the dataset like to check if there a null value in any of the attribute which is being considered for evaluation. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces dimensionality of data and help to work faster.

- **Data Visualization**

Data visualization gives us a clear idea of what the information means by giving it visual context through maps or graphs. This makes the data more natural for the human mind to comprehend and therefore makes it easier to identify trends, patterns, and outliers within large data sets. Data visualization takes the raw data, models it, and delivers the data so that conclusions can be reached. In advanced analytics, data scientists are creating machine learning algorithms to better compile essential data into visualizations that are easier to understand and interpret.

Specifically, data visualization uses visual data to communicate information in a manner that is universal, fast, and effective. This practice can help to identify which areas need to be improved, which factors affect customer satisfaction and dissatisfaction, and what to do with specific areas (outliners). Visualized data gives decision-makers a better prediction of future growth.

- **Splitting the dataset into training and testing dataset**

The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset.

Train Dataset: Used to fit the machine learning model.

Test Dataset: Used to evaluate the fit machine learning model.

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. It is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modelling problem.

- **Applying Machine Learning Algorithm**

When data has been ready we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes. The methods applied on Pima

Indians diabetes dataset. Main objective to apply Machine Learning Techniques to analyze the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in prediction. The following classification algorithm will be applied like Logistic Regression, K Nearest Neighbours, Random Forest, Decision Tree and many more to find the best accuracy.

- **Comparing results for the Best Accuracy**

After the various machine learning algorithms have been applied for the prediction of whether the person is diabetic or not, we compare the results we get from the various algorithms and finds the one with the best accuracy in maximum number of cases so as to proceed forward with the algorithm and implement it in a real life application.

Using Cloud Computing for Diabetes Prediction

After we develop the machine learning model for prediction of diabetes by the evaluation of various parameters what we need to provide the user is the interface through which one can easily interact with the model and get the results. Moreover cloud provides fast computation power and hence interpreting whether the person whose details is entered is diabetic or not in a user friendly manner.

- Deployment of model

After the development of the machine learning model from the algorithm having the best accuracy, we will be deploying this model on cloud platform and for this we will be using Heroku as per our convenience. Heroku is a platform as a service (PaaS) that enables developers to build, run, and operate applications entirely in the cloud. This will provide user a user interface where the user can enter the parameters on which the model is predicting the result and hence can know the result.

PROGRESS CHECK

PROGRESS TILL DATE

We have started our project last week of Febuary,2021 and all of the work we have done till date is shared at the Github repository: https://github.com/JanviPangoriya/Diabetes-prediction . Following the flowchart as mentioned in the implementation details we have understood the problem statement briefly and studied the dataset that we have collected from kaggle and after which we have done the part of data processing and data visualisation and also applied two of the classification algorithm, the Naïve Bayes and the Logistic Regression to get the results of whether the patient is diabetic or not. Further we look forward to apply more of machine learning algorithms to get maximum accuracy as possible.

Understanding the Dataset

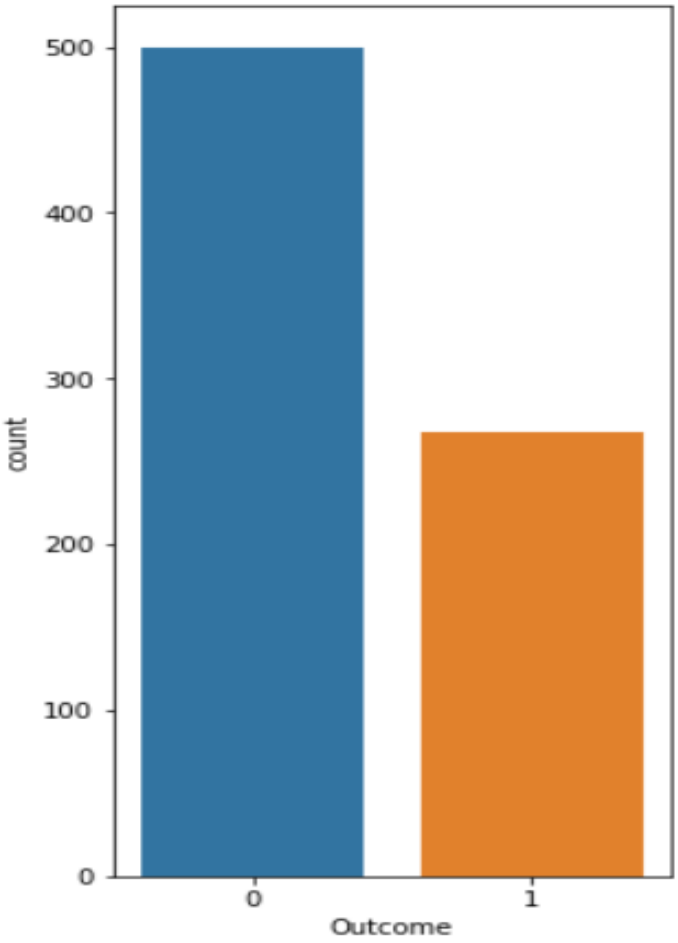
Exploratory data analysis is all about getting and overall understanding of data. It is mainly done to find its properties, patterns and visualizations. It helps us to assure that our data is correct and ready to use for machine learning algorithms. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variable includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

Data Visualisation

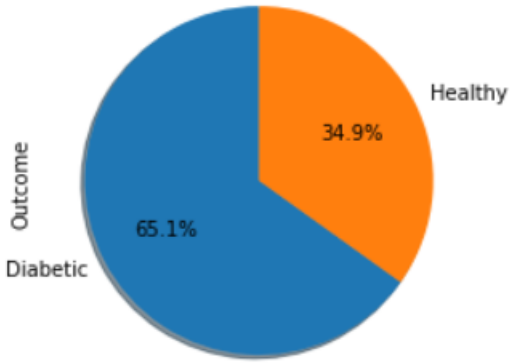
Visualizing data in different type of graphs will provide us with greater insights into our data. We will explore different options on visualizing our data and find out any patterns between within it.

Visualizing the outcome column

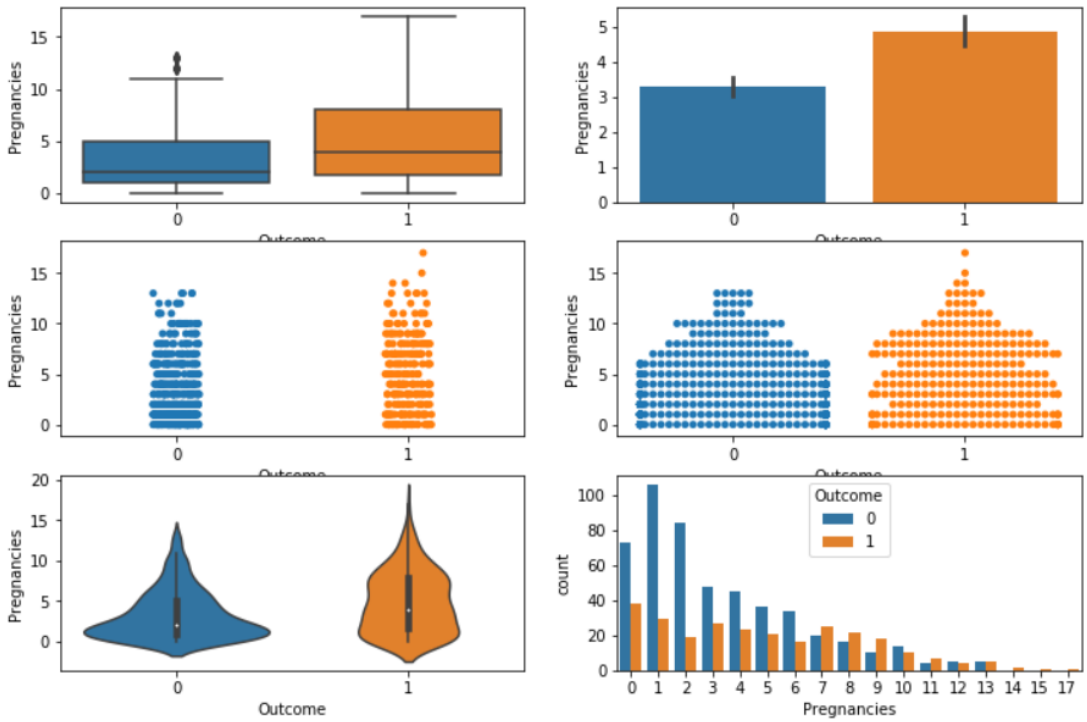
The data shows that these are records of 768 patients out of which 500 are non-diabetic and 268 are diabetic patients.



The dataset also shows that 65.1% are diabetic and 39.4% people are healthy on plotting a pie chart according to the given dataset.

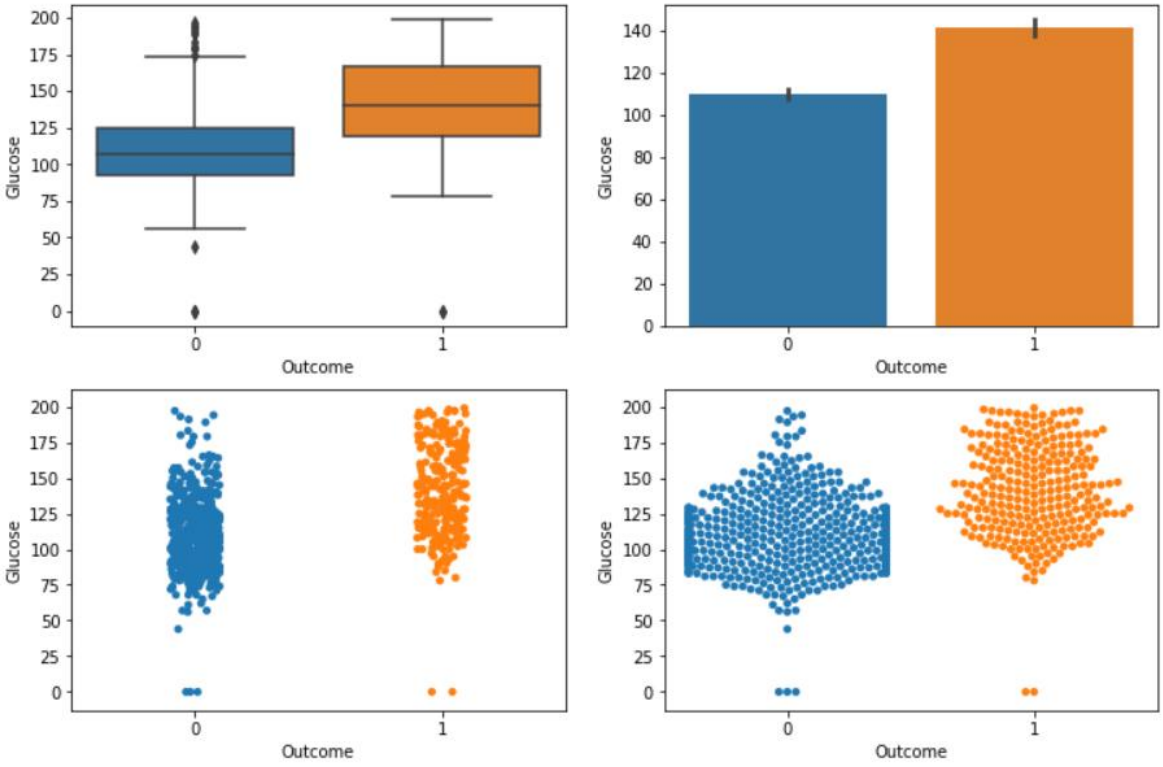


Analysis of "Pregnancies" Parameter with respect to the Outcome

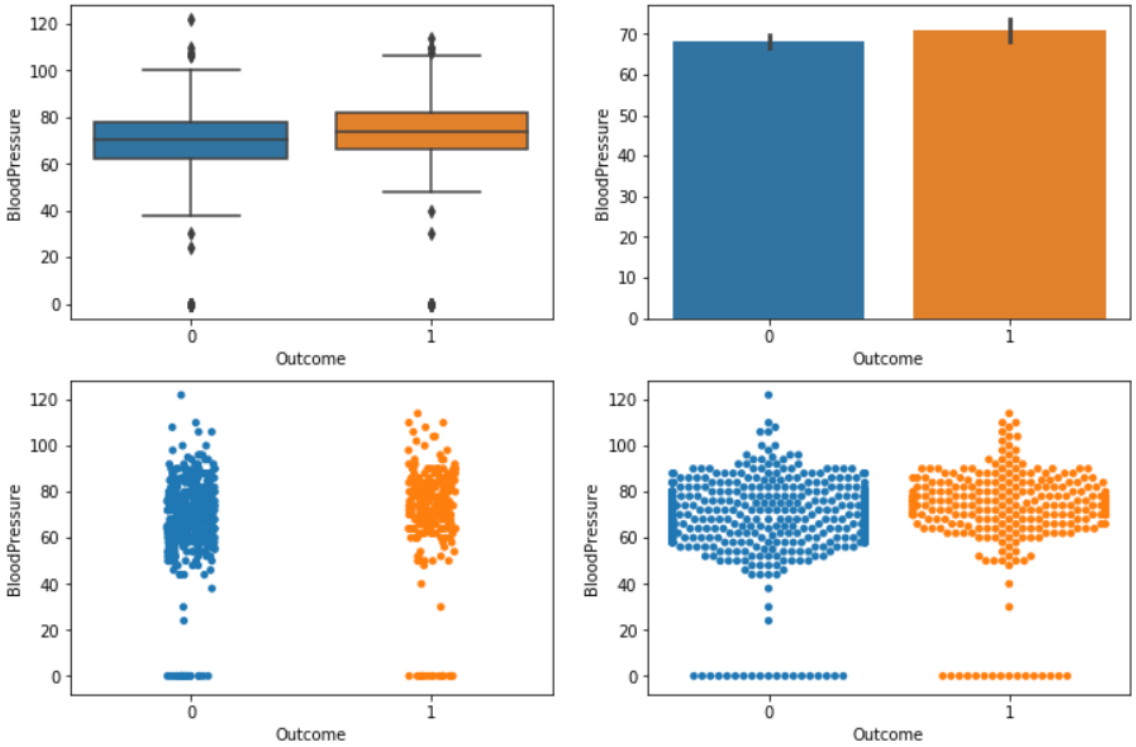


Outcome	0	1
Pregnancies		
0	73	38
1	106	29
2	84	19
3	48	27
4	45	23
5	36	21
6	34	16
7	20	25
8	16	22
9	10	18
10	14	10
11	4	7
12	5	4
13	5	5
14	0	2
15	0	1
17	0	1

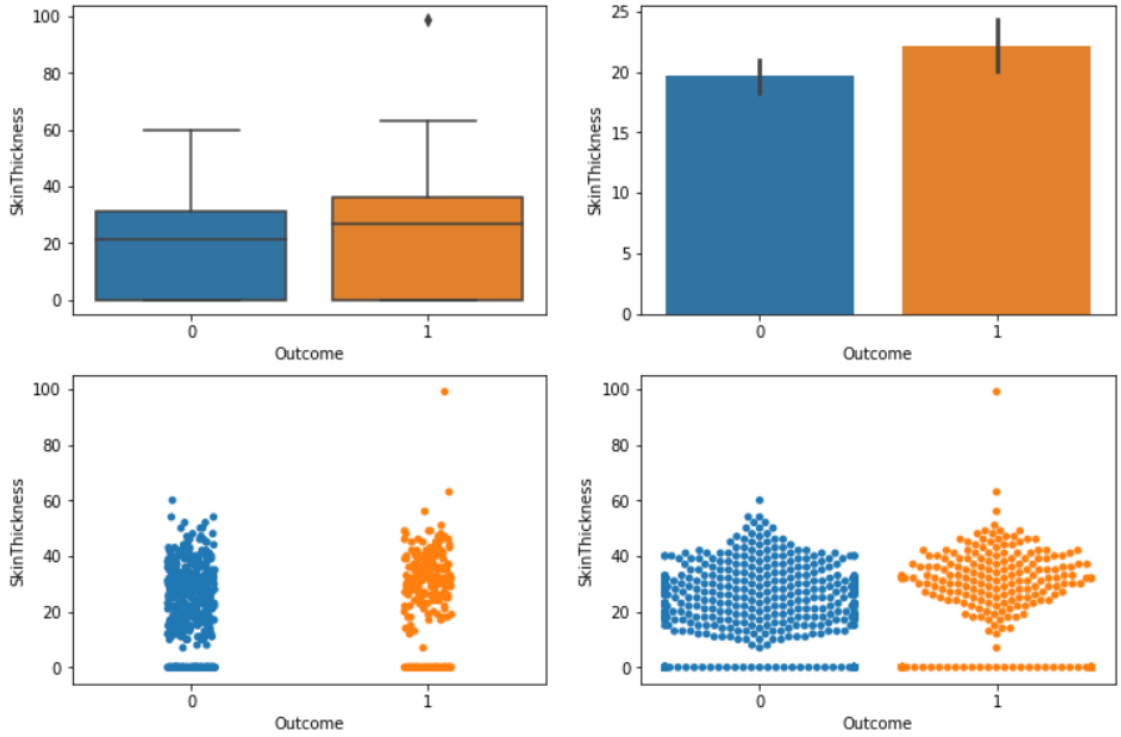
Analysis of “Glucose” parameter with respect to outcome



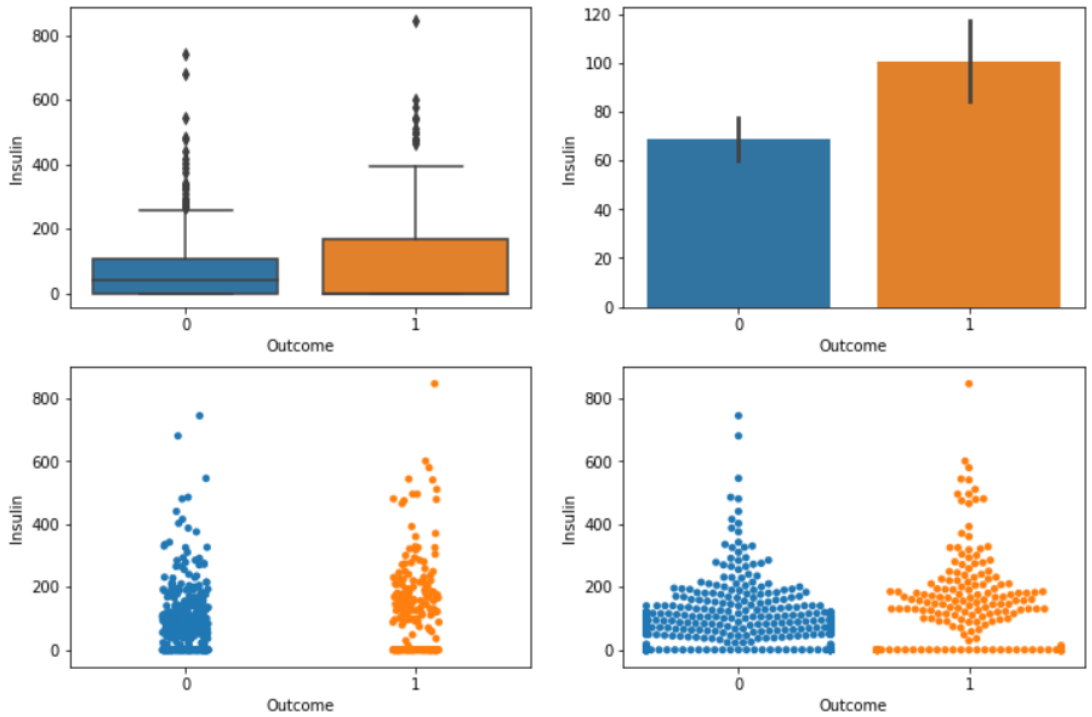
Analysis of “Blood Pressure” parameter with respect to Outcome



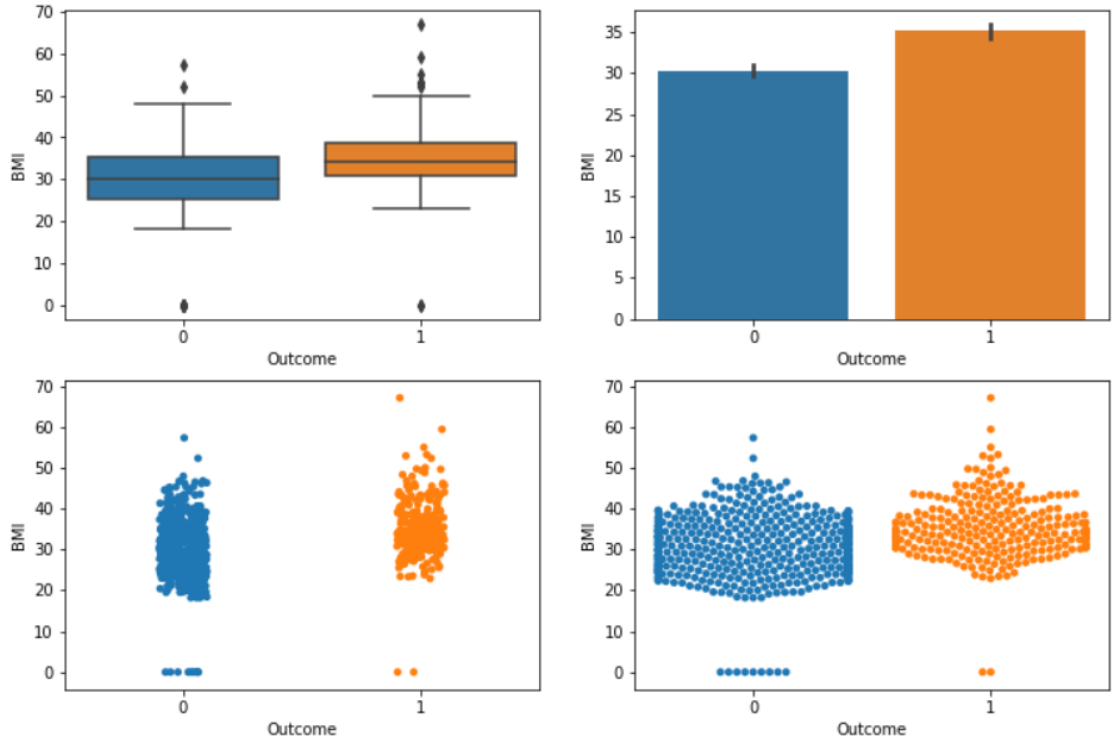
Analysis of “Skin Thickness” parameter with Outcome



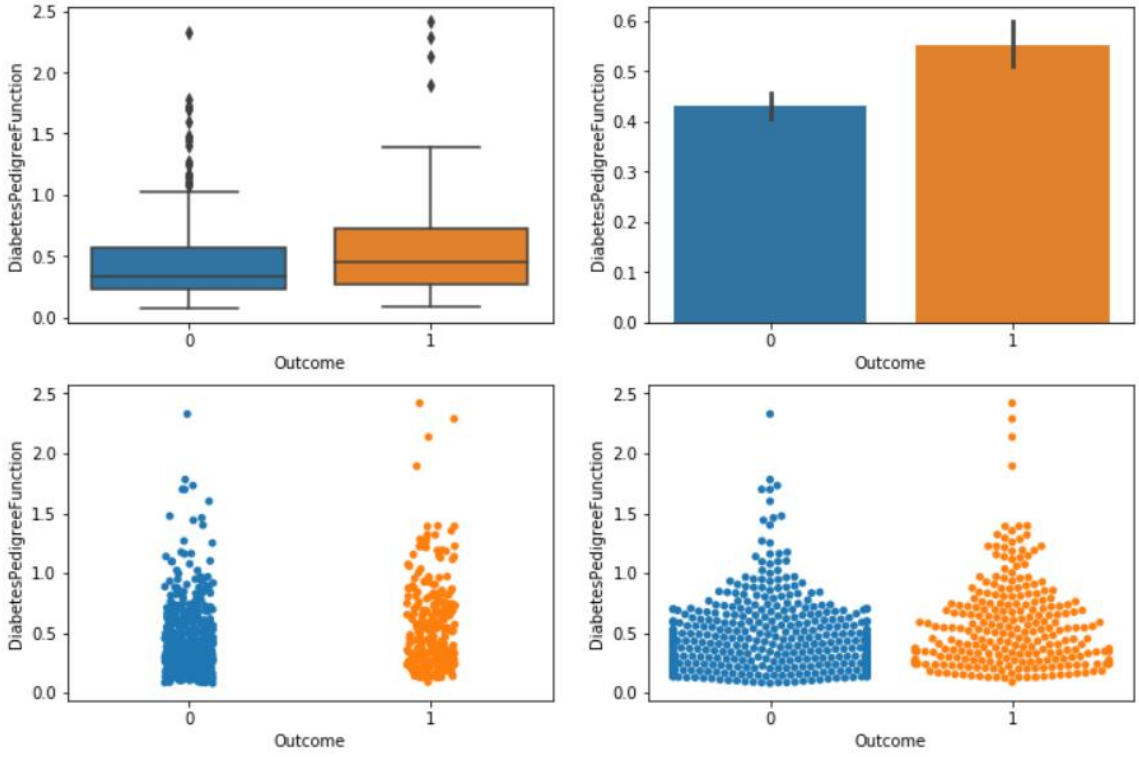
Analysis of “Insulin” parameter with Outcome



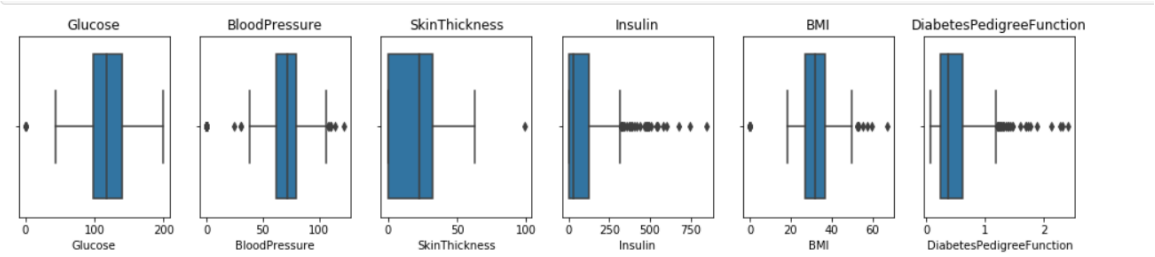
Analysis of “BMI” parameter with respect to Outcome



Analysis of “DiabetesPedigreeFunction Parameter” with Outcome



Visualizing the Outliners



Splitting the Dataset

The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset.

This has been implemented in our project**.** We have considered 30% our dataset as the test dataset and 70% of our dataset as training dataset.

```
In [28]: from sklearn.model_selection import train_test_split
         x = df.drop('Outcome',axis =1)
         y = df['Outcome']
         xtrain ,xtest , ytrain , ytest = train_test_split(x,y,test_size= 0.3)

In [29]: xtrain.shape
Out[29]: (537, 8)

In [30]: xtest.shape
Out[30]: (231, 8)

In [31]: ytrain.shape
Out[31]: (537,)

In [32]: ytest.shape
Out[32]: (231,)
```

DATA PREPROCESSING In the data pre-processing ,we have checked for the null values in each of the attribute and by using a method called simple Imputer we have replaced all the missing values by the mean of the coloumn.

```
print("total number of rows : {0}".format(len(df)))
print("number of rows missing glucose: {0}".format(len(df.loc[df['Glucose'] == 0])))
print("number of rows missing BloodPressure: {0}".format(len(df.loc[df['BloodPressure'] == 0])))
print("number of rows missing SkinThickness: {0}".format(len(df.loc[df['SkinThickness'] == 0])))
print("number of rows missing Insulin: {0}".format(len(df.loc[df['Insulin'] == 0])))
print("number of rows missing age: {0}".format(len(df.loc[df['Age'] == 0])))
print("number of rows missing BMI: {0}".format(len(df.loc[df['BMI'] == 0])))
print("number of rows missing DiabetesPedigreeFunction: {0}".format(len(df.loc[df['DiabetesPedigreeFunction'] == False])))
```

```
total number of rows : 768
number of rows missing glucose: 5
number of rows missing BloodPressure: 35
number of rows missing SkinThickness: 227
number of rows missing Insulin: 374
number of rows missing age: 0
number of rows missing BMI: 11
number of rows missing DiabetesPedigreeFunction: 0
```

```
from sklearn.impute import SimpleImputer
si = SimpleImputer(missing_values=0,strategy="mean")
xtrain = si.fit_transform(xtrain)
```

```
xtest = si.fit_transform(xtest)
```

Applying Machine Learning Algorithm

We have planned to use various machine learning algorithms that support classification as we a dependent variable ‘Outcome’ based on all the parameters and gives output as 0 for a non-diabetic person and 1 for a diabetic person. Up till now we have applied the Naive Bayes Algorithm over our dataset which gives the result as follows.

Naïve Bayes Algorithm Applied

A Naive Bayes classifier is a probabilistic machine learning model that’s used for classification task. The crux of the classifier is based on the Bayes theorem.

Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

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. That is presence of one particular feature does not affect the other. Hence it is called naive.

Multinomial Naive Bayes:

This is mostly used for document classification problem, i.e whether a document belongs to the category of sports, politics, technology etc. The features/predictors used by the classifier are the frequency of the words present in the document.

Naive Bayes Algorithm

```
In [44]: from sklearn.naive_bayes import MultinomialNB
         naive = MultinomialNB()
         naive.fit(xtrain,ytrain)

Out[44]: MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)

In [45]: ytest_pred = naive.predict(xtest)

In [46]: from sklearn.metrics import accuracy_score,classification_report,confusion_matrix,mean_absolute_error
         print(accuracy_score(ytest,ytest_pred)*100)

         58.44155844155844

In [48]: print(confusion_matrix(ytest,ytest_pred))

         [[86 58]
          [38 49]]
```

CLASSIFICATION REPORT FOR NAÏVE BAYES ALGORITHM

```
In [49]: print(classification_report(ytest,ytest_pred))
                       precision    recall  f1-score   support

                    0       0.69      0.60      0.64       144
                    1       0.46      0.56      0.51        87

             accuracy                           0.58       231
            macro avg       0.58      0.58      0.57       231
         weighted avg       0.60      0.58      0.59       231
```

WORK TO BE DONE

We have planned to apply various other classification machine learning algorithms for achieving better accuracy over this dataset and hence to achieve the accurate prediction and implement it in real life application to predict before- hand whether the person is diabetic or not. Some of the algorithms that we have planned to apply are:-

1) **Support Vector Machine-**

Support Vector Machine also known as SVM is a supervised machine learning algorithm. SVM is most popular classification technique. SVM creates a hyper plane that separate two classes. It can create a hyper plane or set of hyper plane in high dimensional space. This hyper plane can be used for classification or regression also. SVM differentiates instances in specific classes and can also classify the entities which are not supported by data. Separation is done by through hyper plane performs the separation to the closest training point of any class. Algorithm-

• Select the hyper plane which divides the class better.

• To find the better hyper plane you have to calculate the distance between the planes and the data which is called Margin.

• If the distance between the classes is low then the chance of miss conception is high and vice versa.

So we need to

• Select the class which has the high margin. Margin = distance to positive point + Distance to negative point.

2) **K-Nearest Neighbour –**

KNN is also a supervised machine learning algorithm. KNN helps to solve both the classification and regression problems. KNN is lazy prediction technique. KNN assumes that similar things are near to each other. Many times data points which are similar are very near to each other. KNN helps to group new work based on similarity measure. KNN algorithm record all the records and classify them according to their similarity measure. For finding the distance between the points uses tree like structure. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set — it's nearest neighbours. Here K= Number of nearby neighbours, it's always a positive integer. Neighbour's value is chosen from set of class. Closeness is mainly de

fined in terms of Euclidean distance. The Euclidean distance between two points P and Q i.e. P (p1,p2, …. Pn) and Q (q1, q2,..qn) is defined by the following equation:-

$$d(P,Q) = \sum_{i=1}^{n} (P_i - Q_i)^2$$

Algorithm-

• Take a sample dataset of columns and rows named as Pima Indian Diabetes data set.

• Take a test dataset of attributes and rows.

• Find the Euclidean distance by the help of formula-

$$EculideanDistance = \sqrt{\sum_{i=1}^{y} \sum_{j=1}^{m} \sum_{l=1}^{n-1} (R_{(j,l)} - P_{(i,l)})^2}$$

• Then, Decide a random value of K. is the no. of nearest neighbours

• Then with the help of these minimum distance and Euclidean distance find out the nth column of each.

• Find out the same output values. If the values are same, then the patient is diabetic, otherwise not.

3. **Decision Tree-**

Decision tree is a basic classification method. It is supervised learning method. Decision tree used when response variable is categorical. Decision tree has tree like structure based model which describes classification process based on input feature. Input variables are any types like graph, text, discrete, continuous etc.

Steps for Decision Tree Algorithm-

• Construct tree with nodes as input feature.

• Select feature to predict the output from input feature whose information gain is highest.

• The highest information gain is calculated for each attribute in each node of tree.

• Repeat step 2 to form a sub tree using the feature which is not used in above node.

4**) Logistic Regression-**

Logistic regression is also a supervised learning classification algorithm. It is used to estimate the probability of a binary response based on one or more predictors. They can be continuous or discrete. Logistic regression used when we want to classify or distinguish some data items into categories. It classify the data in binary form means only in 0 and 1 which refer case to classify patient that is positive or negative for diabetes. Main aim of logistic regression is to best fit which is responsible for describing the relationship between target and predictor variable. Logistic regression is a based on Linear regression model. Logistic regression model uses sigmoid function to predict probability of positive and negative class.

Sigmoid function P = 1/1+e - (a+bx) Here P = probability, a and b = parameter of Model. Ensembling- Ensembling is a machine learning technique Ensemble means using multiple learning algorithms together for some task. It provides better prediction than any other individual model that's why it is used. The main cause of error is noise bias and variance, ensemble methods help to reduce or minimize these errors. There are two popular ensemble methods such as – Bagging, Boosting, ada-boosting, Gradient boosting, voting, averaging etc. Here In these work we have used Bagging (Random forest) and Gradient boosting ensemble methods for predicting diabetes.

5.) Random Forest –

It is type of ensemble learning method and also used for classification and regression tasks. The accuracy it gives is grater then compared to other models. This method can easily handle large datasets. Random Forest is developed by Leo Bremen. It is popular ensemble Learning Method. Random Forest Improve Performance of Decision Tree by reducing variance. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.

Algorithm-

• The first step is to select the "R" features from the total features "m" where R<M

• Among the "R" features, the node using the best split point

. • Split the node into sub nodes using the best split.

• Repeat a to c steps until ”l” number of nodes has been reached.

• Built forest by repeating steps a to d for “a” number of times to create “n” number of trees. The random forest finds the best split using the Gin-Index

DEPLOYEMENT OF MODEL ON CLOUD

After we develop the machine learning model for prediction of diabetes by the evaluation of various parameters what we need to provide the user is the interface through which one can easily interact with the model and get the results. Moreover cloud provides fast computation power and hence interpreting whether the person whose details is entered is diabetic or not in a user friendly manner.

- Deployment of model

After the development of the machine learning model from the algorithm having the best accuracy, we will be deploying this model on cloud platform and for this we will be using Heroku as per our convenience. Heroku is a platform as a service (PaaS) that enables developers to build, run, and operate applications entirely in the cloud. This will provide user a user interface where the user can enter the parameters on which the model is predicting the result and hence can know the result.

CONCLUSION

1. The dataset have nine attributes (parameters) in which there are eight independent variables(Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age) and one dependent variable (Outcome).
2. BMI and Diabetes Pedigree Function are a float data type and other parameters are integer data type.
3. The parameters do not contain any null values (missing values). However, this cannot be true. As Insulin, Skin Thickness, Blood Pressure, BMI, Glucose have zero values.
4. The Outcome parameter shows that there are 500 healthy people and 268 Diabetic people. It means that 65% people are diabetic and 34.9% people are healthy.
5. The parameters Glucose, Blood Pressure, BMI are normally distributed. Pregnancies, Insulin, Age, Diabetes Pedigree Function are rightly skewed.
6. The missing values '0' is replaced by the mean of the parameter to explore the dataset.
7. Blood Pressure, Skin Thickness, Insulin, BMI have outliers.
8. There is no convincing relationship between the parameters. Pregnancies and age have some kind of a linear line. Blood Pressure and age have little relation. Most of the aged people have Blood Pressure. Insulin and Glucose have some relation.
9. Glucose, Age BMI and Pregnancies are the most Correlated features with the Outcome. Insulin and Diabetes Pedigree Function have little correlation with the outcome. Blood Pressure and Skin Thickness have tiny correlation with the outcome.
10. Age and Pregnancies, Insulin and Skin Thickness, BMI and Skin Thickness, Insulin and Glucose are little correlated.
11. The middle aged women are most likely to be diabetic than the young women. As the percentage of diabetic women are 48% and 59% in the age group of 31-40 and 41-55.
12. After Pregnancy people have more chance of diabetics.
13. People with high Glucose level are more likely to have diabetics.
14. People with high Blood Pressure have more chance of diabetics.
15. People with high Insulin level are more likely to have Diabetes.

REFERENCES

1. https://www.ijert.org/diabetes-prediction-using-machine-learning-techniques
2. https://www.kaggle.com/uciml/pima-indians-diabetes-database
3. https://www.medicalnewstoday.com/articles/high-diastolic-pressure