This study was conducted using Machine Learning classifiers to predict if a particular observation is at a risk of developing diabetes.

PREDICTION OF DIABETES

PIMA INDIAN DATASET

Parvin Soleymani

Pima Indians Diabetes PID

Abstract

Data Mining is used to extract information from the raw data information that is expressed in a comprehensible form and can be used for a variety of purposes. Data Mining plays a crucial role in health concern businesses, mainly in the early prediction of diseases, such as Diabetes.

Currently, the incidence of diabetes has increased worldwide and is expected to keep growing, with the greatest increase seen in metabolic forms of diabetes, notably type 2. Diabetes is one of fatal, metabolic and costly disease that increases blood sugar level, affecting people of all ages. According to the Public Health Agency of Canada, between 2003–2004 and 2013–2014, there was a relative increase of 37.3% of diagnosed diabetes cases, from 5.6% to 7.8% in Canada.

The prime objective of this project is to predict diabetes based on diagnostic measurements available in the Pima Indian Diabetes (PID) dataset, by identifying which type of algorithm model works best for this prediction. This manuscript explains the steps followed in the process to achieve this objective. At first, the dataset was prepared, and consecutively different classification techniques were applied to compare and analyze the best model performance on PID. Three machine learning classification algorithms were used in this experiment Decision Tree (J48), Naive Bayes, and Logical Regression

The performances of all the three algorithms were evaluated on various measures like Precision, Accuracy, F-Measure, and Recall. Results obtained show Logistic Regression outperforms with the highest accuracy of 78% compared with other algorithms.

Introduction

- ## What is diabetes?

Diabetes is a chronic condition that occurs when the body loses its ability to produce or properly use insulin, a hormone that controls sugar levels in the blood. There are three main types of diabetes: type 1, the body does not make insulin, type 2 (common type), body does not make or use insulin well, and gestational diabetes. Other types are uncommon

- ## How many Canadians live with diabetes? (Prevalence)

According to the most recent data provided by the Public Health Agency of Canada (1), about 3.0 million Canadians (8.1%) were living with diagnosed diabetes in 2013–2014, representing 1 in 300 children and youth (1–19 years), and 1 in 10 adults (20 years and older). The prevalence of diagnosed diabetes generally increases with age and is higher among males (8.7%) than among females (7.6%), both overall and in most age groups.10% of people living with diabetes in Canada have type 1 diabetes. 90 per cent of Canadians with diabetes are living with type 2 diabetes.

- ## Dataset - Attributes Information

The dataset used in this study is Pima Indians Diabetes Database of the National Institute of Diabetes and Digestive and Kidney Diseases research center.

The dataset consists of eight attributes, the independent variables, and one outcome (class), independent variable, which could be either positive or negative in terms of 1 and 0. Attributes are defined as follow:

- Number of times pregnant
- Plasma glucose concentration 2 hours in an oral glucose tolerance test
- Blood pressure (mm Hg)
- Triceps skinfold thickness (mm)
- 2-Hour serum insulin (mu U/ml)
- Body mass index (weight in kg/(height in m)^2)
- Diabetes pedigree function
- Age (years)
- Outcome - Class variable (0 or 1)

1. DATA PREPARATION

- ## Summary of Attributes

All the attributes in this dataset are quantitative type, both numeric and integer. And then the test was applied to the dataset and released there are no duplicated records.

```
'data.frame':   768 obs. of  9 variables:
 $ no.of.pregnency    : int  6 1 8 1 0 5 3 10 2 8 ...
 $ glucose.conc.      : int  148 85 183 89 137 116 78 115 197 125 ...
 $ blood.pressure     : int  72 66 64 66 40 74 50 0 70 96 ...
 $ skin.fold.thickness: int  35 29 0 23 35 0 32 0 45 0 ...
 $ serum.insulin      : int  0 0 0 94 168 0 88 0 543 0 ...
 $ BMI                : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
 $ pedigree.function  : num  0.627 0.351 0.672 0.167 2.288 ...
 $ Age                : int  50 31 32 21 33 30 26 29 53 54 ...
 $ Class              : int  1 0 1 0 1 0 1 0 1 1 ...
```

Values of Max, min, mean and standard deviation of attributes are shown on the table below. (Table 1)

Table 1 Summary of Attributes

	No. of Pregnancy	Glucose Conc.	Blood Pressure	Skin Fold Thickness	Serum Insulin	BMI	Pedigree Function	Age	Class
Min	0	0	0	0	0	0	0.07	21	0
1st Q	1	99	62	0	0	27.30	0.24	24	0
Median	3	117	72	23	30.5	32	0.37	29	0
Mean	3.8	120.9	69	20.54	79.8	31.99	0.47	33.24	0.35
3rd Q	6	140.2	80	32	127.2	36.60	0.62	41	1
MAX	17	199	122	99	846	67.10	2.42	81	1
STD	3.4	32.0	19.4	16.0	115.2	7.9	0.3	11.8	

- ## Outlier Values

Box plots below illustrate the outlier for each attribute

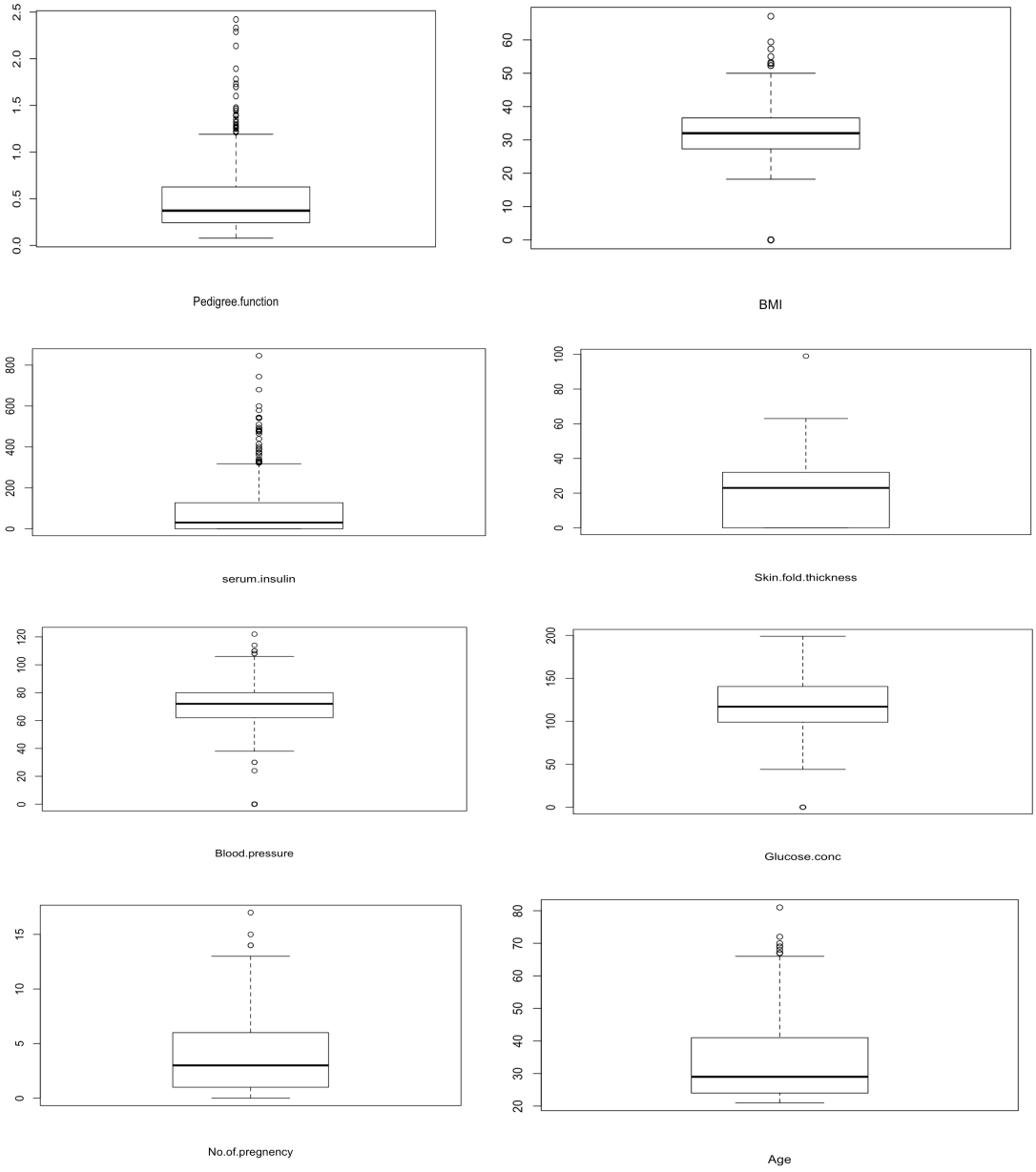


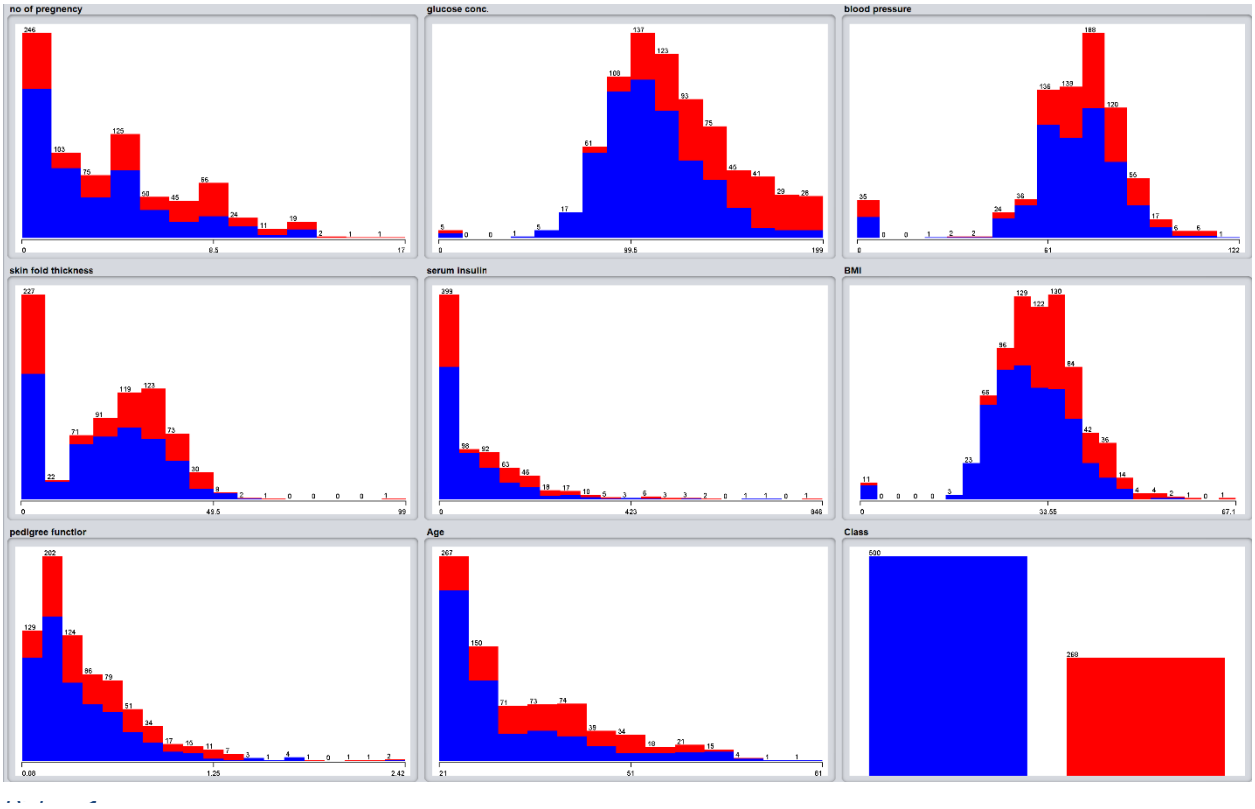
Figure 1 Box Plots

- **Predictor Attributes**

Histograms show the distribution of numeric attributes. According to Figure 2:

Attributes which look normally distributed: Glucose, Blood Pressure, Skin Thickness, and BMI.

Attributes look an exponential distribution: Pregnant, Insulin, Diabetes pedigree function, Age



- **Data Analysis**

  The dataset reflects an imbalanced class distribution since it possesses an unequal instances distribution.

Value 0 (Non-Diabetes): 500 Instances
Value 1 (Diabetes): 268 Instances

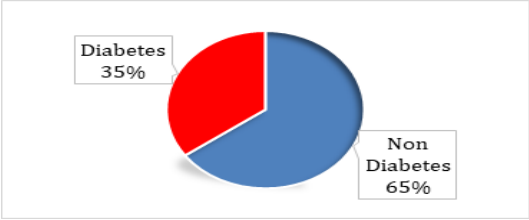


Figure 3 Pie Chart

This graph represents the correlation across all attributes.

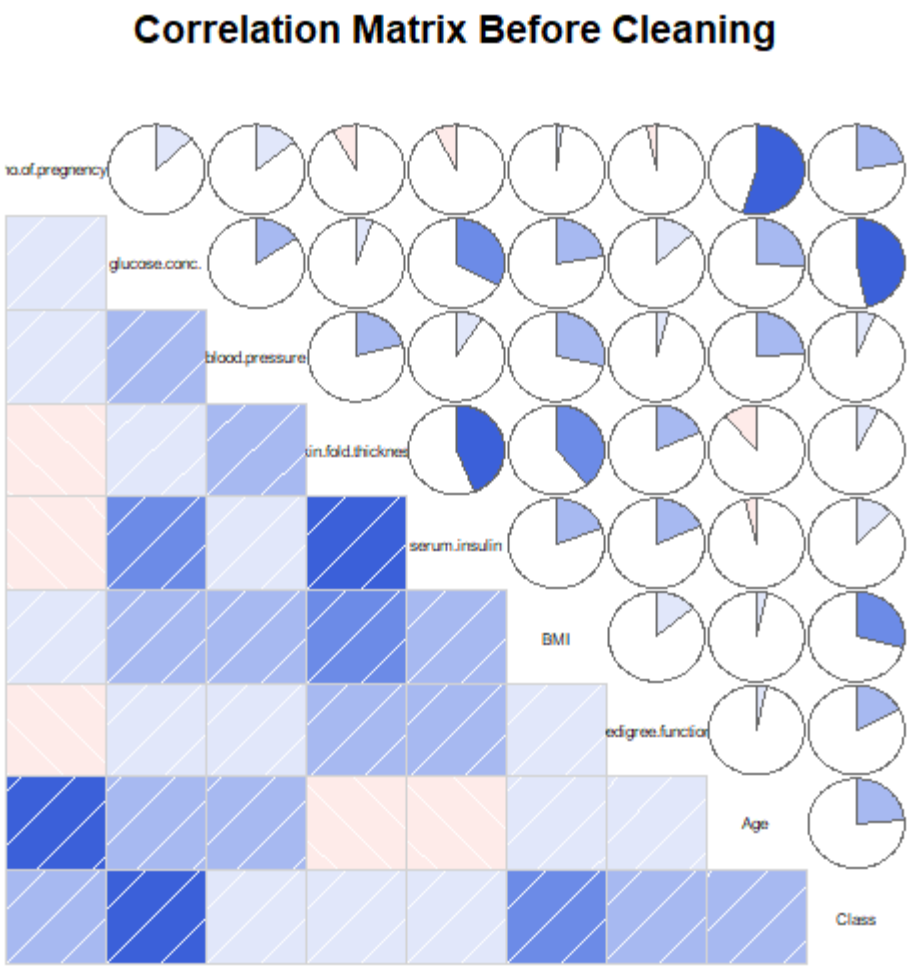


Figure 4 Correlation Matrix Before Cleaning Data

According to this graph Age and Number of Pregnancy, BMI and Skinfold thickness, Skinfold thickness and Serum Insulin have a strong correlation together.

Glucose and BMI both seem to have a higher correlation to the class.

- **Identify the influenced attributes:**

T-test was performed to identify if Glucose, Age, BMI and Pregnancies have influence in the Class attribute.

Ho = There is no difference in the average value of Pregnancies, Age, BMI, Glucose for all people with diabetes as compared to all those without

Results:

- T-test for Glucose: p-value - 2.2e-16
- T-test for Age: p-value - 1.302e-11
- T-test for BMI: p-value - 3.719e-16
- T-test for No of Pregnancies: p-value - 4.582e-09

Since p-values are considerably less than alpha level of 0.01 the null hypothesis is rejected, therefore, these attributes may have influence on the class attributes.

● Data Cleaning and Transformation

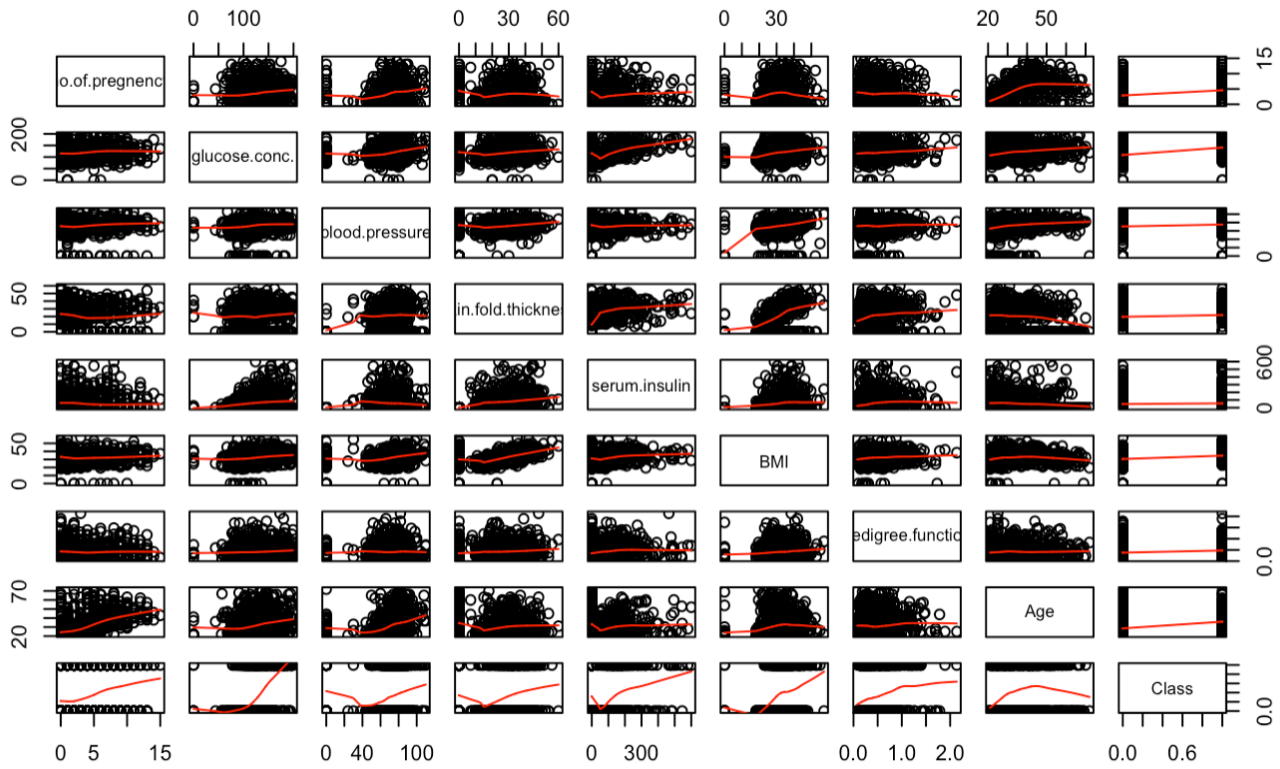
The Sapply() method was used to check the missing value (NAs) in columns. The result shows that there are no NAs in the dataset, however the summary displays zero values for attributes which are considered to be unappropriated.

```
no.of.pregnency       glucose.conc.      blood.pressure skin.fold.thickness       serum.insulin
              0                   0                   0                   0                   0
            BMI   pedigree.function                 Age               Class
              0                   0                   0                   0
```

The strategy applied to prepare the dataset was finding zero values of the attributes which could not be zero such as BMI, Glucose, Blood Pressure, Serum Insulin, and Skinfold thickness. Zero values were substituted with the median value of the attributes. Based on box plots the outliers are significant. In order to keep the dataset with a significant number of instances, only the outliers that seem to be far away from the dataset were eliminated.
Total number of instances after eliminating outliers

```
[1] 757   9
```

After handling zero values and removing outliers for Pedigree function, Number of pregnancies, Blood pressure, Skin fold thickness, Insulin, BMI and Age,
the dataset was cleaned and ready for building the different models.
The matrix of scatterplots below confirms that after cleaning the dataset there in less outliers in the attributes.



- **Feature selection**

Correlation Matrix after cleaning the data to find out the features which are more correlated to outcome.

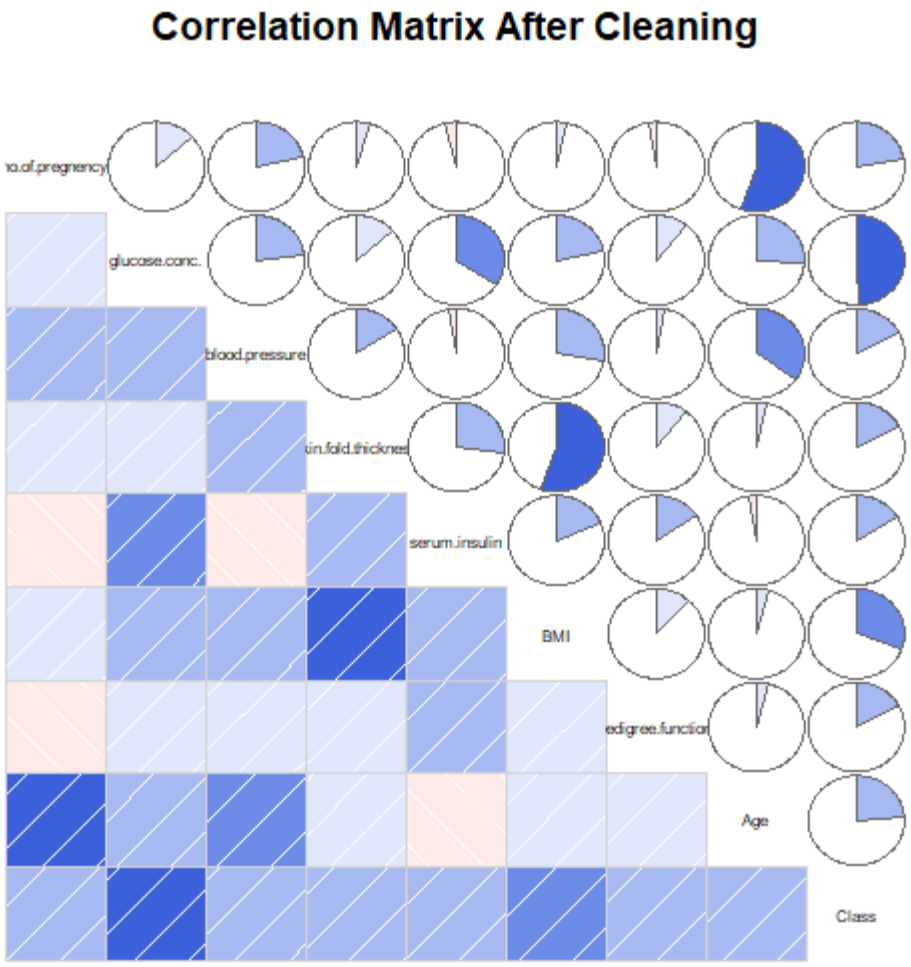


Figure 6 Correlation Matrix After Cleaning Data

After obtaining a clean dataset a correlation based on feature selection was calculated, to identify the attributes that have a moderate-to-high positive or negative correlation (close to -1 or 1). This with the intention to decide if dropping or not those attributes with a low correlation (value close to zero).

Weka supports correlation based on feature selection with the Correlation Attribute Eval technique that requires use of a Ranker search method.

As initially predicted, the result suggests that one attribute (glucose conc) has the highest correlation with the output class. It also suggests a host of attributes with some modest correlation (BMI, Age, No.of.Pregnancy). When using 0.2 as cut-off for relevant attributes, then the remaining attributes could possibly be removed (Skin.Fold Blood Pressure, Pedigree. Function, and Serum Insulin).

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 9 Class):
        Correlation Ranking Filter
Ranked attributes:
 0.492    2 glucose.conc.
 0.305    6 BMI
 0.238    8 Age
 0.225    1 no.of.pregnency
 0.176    4 skin.fold.thickness
 0.173    3 blood.pressure
 0.171    7 pedigree.function
 0.162    5 serum.insulin

Selected attributes: 2,6,8,1,4,3,7,5 : 8
```

2. Proposed Models

- **Assessing the Accuracy of the models**

This dataset is analyzed in Weka using 3 algorithms for the prediction of diabetics, Decision Tree (J48), Naive Bayes and Logic Regression. The accuracy of these models is compared together. For the testing model the 80% as the training set and 20% as the testing set was selected. According to Mean Absolute Error (MAE) of 0.316, this partitioning is valid. MAE measures the average magnitude of the errors in a set of predictions.

In 80% to 20% division R2 is higher, MAE and the error is lower compared to other division methods.

Table 2

	60%	70%	80%
MAE	0.3185	0.3260	0.3168
R2	0.3190	0.3152	0.3349
Error	1.182	1.100	1.110

Measures:

Table 3

Measures	Definitions	Formula
Accuracy(A)	Accuracy determines the accuracy of the algorithm in predicting instances	A=(TP+TN) / (Total no of samples)
Precision (P)	(also called positive predictive value) is the fraction of relevant instances among the retrieved instances	P = TP / (TP+ FP)
Recall (R)	(also known as sensitivity) is the fraction of the total amount of relevant instances that were actually retrieved.	R =TP / (TP+FN)
F-Measure	F-Measure is the weighted average of precision and recall.	F=2*(P*R) / (P+R)

Predictive Modeling/Classification

- ## Classification using Decision Tree:

Decision tree algorithm falls under the category of supervised learning. The main concept behind decision tree learning is to build a predictive model which is mapped to a tree structure, its goal is to achieve perfect classification with minimal number of decisions. (2)

This study has utilized the decision tree J48 (C4.5 algorithm) provided by Weka, which uses a measure called “information gain” to choose the attribute at each stage.

- ## Dataset Split Experiment

Experiments were performed and analyzed using different strategies for dataset split including training set, internal cross-validation 10-folds, 3 -folds and training split 80% to determine which of them provides the best performance.

The table below compares the performance of data split strategy based on various measures.

Table 4

Data Split Strategy	Accuracy	Precision	Recall	F-Measure	ROC	Mean Absolute Error
Training Set	78.20%	0.777	0.782	0.778	0.778	0.34
Cross- Validation / 10 folds	71.99%	0.722	0.720	0.721	0.775	0.34
Cross- Validation / 3 folds	73.71%	0.742	0.737	0.739	0.760	0.32
Percentage Split 80%	**74.17%**	**0.774**	**0.742**	**0.749**	**0.802**	**0.30**

- ## Result:

Training set has the highest accuracy in terms of dataset split strategy for all measures showing the maximum accuracy of 78.20% when compared to other strategies, however this is considered a normal symptom of over-fitting. **Percentage Split 80%** with 74.17% is the strategy recommended since it gives the highest ROC score with 0.802 and the lowest MAE score of 0.30

The table below shows the confusion Matrix for Split 80% based on 151 instances:

Table 5

	Actual Positives	Actual Negatives
Predicted Positives	(TP) 73	(FP) 28
Predicted Negatives	(FN) 11	(TN) 29

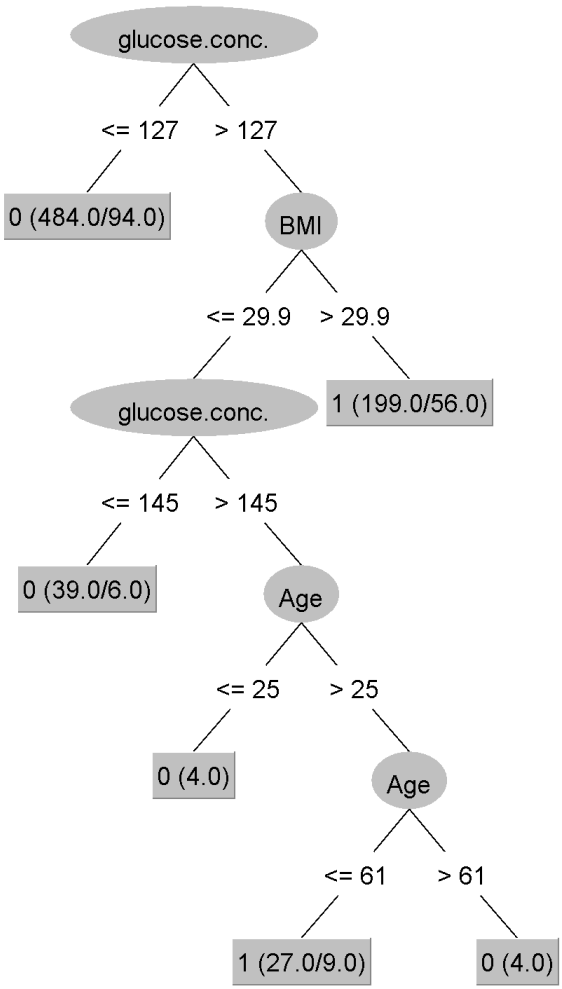


Figure 7 Decision tree

- ## Dataset Parameters Experiment

Experiments were performed and analyzed when altering 2 different types of parameters **minNumObj** and **unPruned**.

Table 6

Min Num Obj	Number of Leaves	Size of The Tree	Accuracy
1	8	15	74.17
2	6	11	74.17
5	3	5	74.83
10	3	5	76.16

Result: The number of leaves in the tree decreases as the size of each leaf grows. The tree size follows the same direction as the number of leaves: it decreases as the leaf size grows.

Pruning

J48 algorithm in Weka generates pruning by default, and it is specified in the parameters by the "unpruned" option here, which is "False". When running the experiment to switch the option to True, the following results are obtained:

- Default (pruned): 74.17% accuracy, tree has 6 leaves, 11 nodes
- Unpruned: 74.16% accuracy 10 leaves, 19 nodes

- **Classification using Naive Bayes:**

Naïve Bayes is a classification technique and it is used as a classifier. This classifies base on the probabilities, classifying an observation by evaluating its probability of belonging to one class vs. belonging on another. Since it is based on conditional probabilities it is considered as a powerful algorithm employed for classification purpose. The main assumption needed for the use of Naive Bayes is the independence of variables. (3)In this case, based on the correlation matrix, it is nearly impossible to believe that the variables in this dataset are not related in some way., but that does not mean that this algorithm will have poor predictive power in the case of this dataset.

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

The Variable y is a Class variable (1 or 0), which represents the patients diabetes test positive or negative. X represents other Variables in the dataset.

The table below represents the Naive Bayes result using four test options.

Table 7

Data Split Strategy	Accuracy %	Precision	Recall	F-Measure	ROC	MAE
Training Set	75.69	0.751	0.757	0.752	0.834	0.287
Cross- Validation / 10 folds	75.56	0.750	0.756	0.751	0.830	0.289
Cross- Validation / 3 folds	75.56	0.750	0.756	0.752	0.827	0.291
Percentage Split 80%	**76.82**	**0.765**	**0.768**	**0.766**	**0.813**	**0.291**

The table below shows the confusion Matrix for Split 80%:

Table 8

	Actual Positives	Actual Negatives
Predicted Positives	(TP)85	(FP)16
Predicted Negatives	(FN)19	(TN)31

What stands out from this test is that the 80-percentage split has a better result, based on the higher accuracy and almost the same mean absolute error.

- ## c. Logistic Regression:

Logistic regression is a supervised classification algorithm that can predict the probability of a dependent variable. In the logistic regression, the dependent variable can have only two possible values (0 or 1).

Strategy:

To perform the logistic regression a random sample of 80% of the observation is used for the training set and the remaining (20%) as the testing set.

Table 9 Statistics measurements using different splitting methods

Data Split Strategy	Accuracy	Precision	Recall	F-Measure	ROC	Mean Absolute Error
Training Set	76.00%	0.763	0.769	0.761	0.836	0.31
Cross- Validation / 10 folds	75.69%	0.750	0.750	0.740	0.840	0.31
Cross- Validation / 3 folds	75.56%	0.740	0.750	0.740	0.820	0.32
Percentage Split 80%	**78%**	**0.78**	**0.78**	**0.78**	**0.81**	**0.30**

According to the statistics shown on table 9, the model shows an accuracy percentage of 78% and error rate of 30% when the percentage split is 80%. The result indicates that the model works better with 80% split because the accuracy level is higher and Mean Absolute Error (MAE) is lower when compared to the other splitting methods.

Table 10 Confusion Matrix

	Actual Positives	Actual Negatives
Predicted Positives	90 (TP)	11 (FP)
Predicted Negatives	21 (FN)	29 (TN)

Confusion Matrix: Shows the actual and predicted labels from the classification problem using the percentage split of 80%

1. True Positive Rate: 78%

2. False Positive Rate: 31%
3. Precision value: The number of positive cases correctly classified from the total number of cases that the model identified as positive is 78%
4. Recall value: The number of positive cases correctly classified from the total number of real positive cases in the dataset is 78%
5. F-score: The model performance is 78%

Plotting the Model:

The plots below show residuals in four different ways. Checking residuals is a good way to understand the model and the data. It is important to assess this information and if needed redesign the model accordingly.

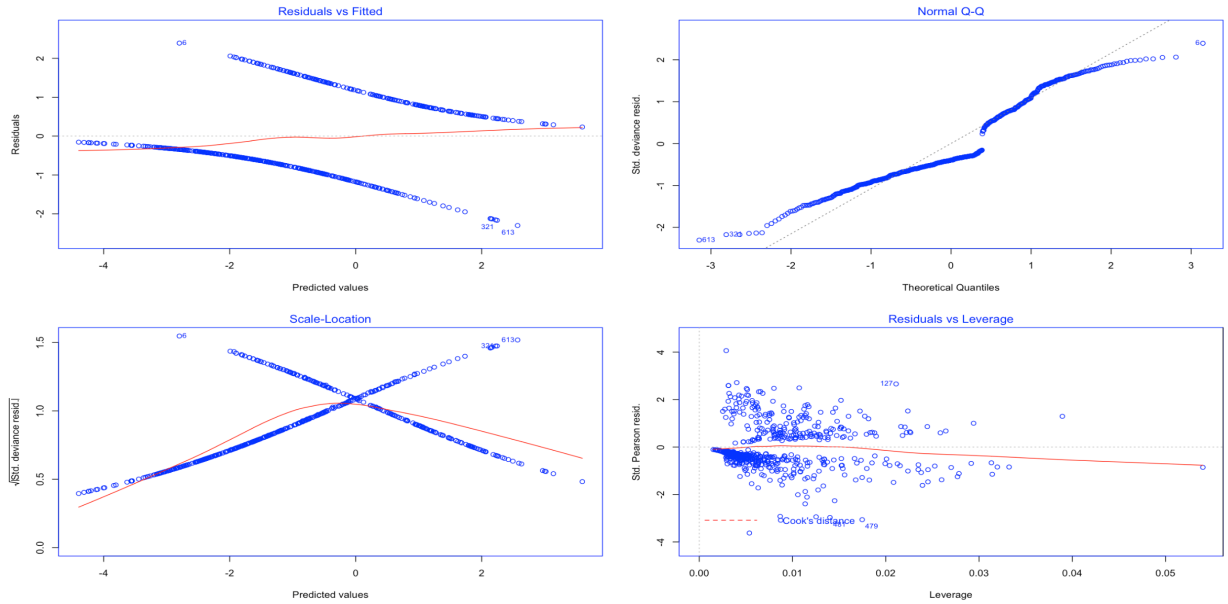


Figure 8 Regression Model Plots

Residuals vs Fitted: The dotted line in the graph shows the fitted line and points on this line have zero residuals. And, the red line shows a pattern of residual movement. The residuals have non-linear patterns therefore there is a non-linear relationship between observed value and the predictor variable. Two blue lines show the predistortion probability of 0 or 1 for the independent variables.

Normal Q-Q Plot: This plot shows if the residuals are normally distributed or not. Since the residuals are closed to the dotted line, they are normally distributed. As shown in the picture, the observation number 6 looks a little off.

Scale - Location Plot: This plot allows to check the assumption of equal variance across the range of predictors. The red line is close to an ideal horizontal line that indicates the homoscedasticity across the range. (4) .

Residuals vs Leverage Plot: Cook's distance tells which points have more influence on the regression model (5). The plot shows that all residuals are inside the Cook's distance. Usually, the points that are outside the cooked distance may have influence

on the model. In this case, the plot doesn't show any cases as all of the cases are within the Cook's distance line; however, point 127 may have influence on the model.

Summary:

Glucose level, BMI, pregnancies all have significant influence on the model. However, glucose and BMI seem to have a higher impact on diabetics compared to the other attributes. Overall, the accuracy of the model is 78%. This shows that using the logistic regression model is an accurate way to make diabetes predictions.

- ## d. Compare the results of the 3 techniques

ROC graphs are drawn using the true positive rates and true negative rates to summarize the confusion matrix in different thresholds. Area Under Curve (AUC) shows which method is better in predicting diabetics.

X= False Positive Rate **Y= True Positive Rate**

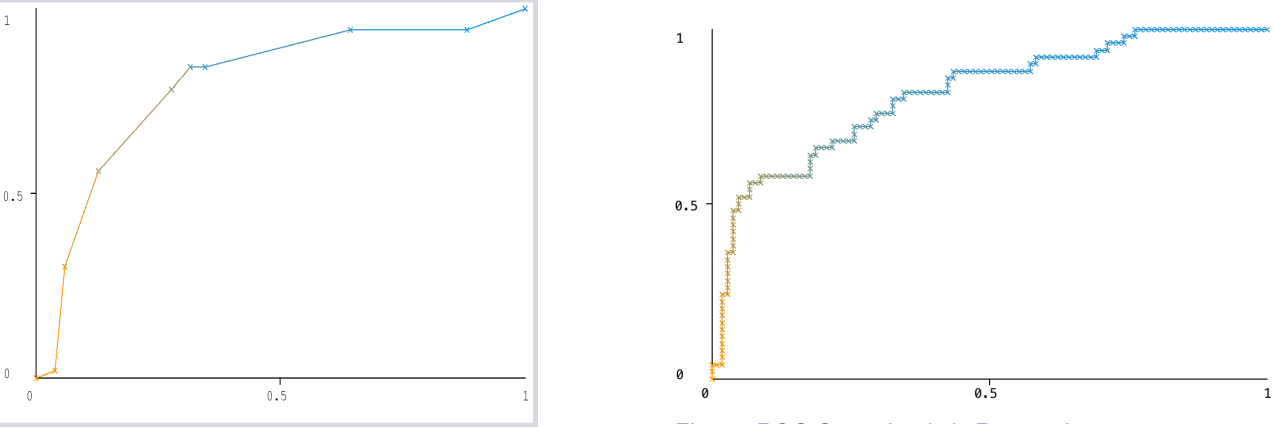


Figure 9 ROC Curve Decision Tree

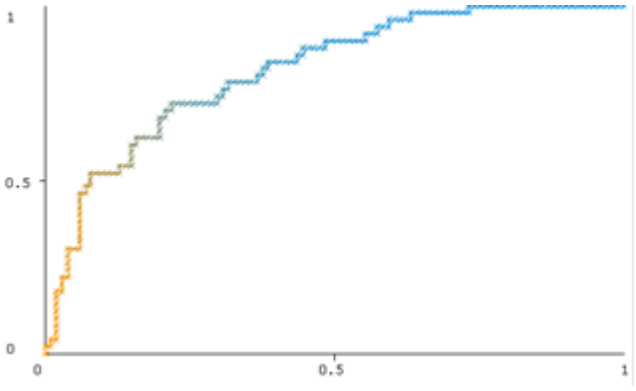


Figure 11 ROC Curve Naive Bayes

Based on the result of ROC Curve for all tests which performed the result shown that the logistic model did a better job in predicting false positive between threshold 0 and 0.5 intervals. At the threshold above the 0.5 the curves show that it is predicted correctly diabetics in logistic regression because the true positive rate is higher than false positive rate**.**

Table 11Comparative Performance of Classification Algorithms on Various Measures

Classification Algorithms	Accuracy	Precision	Recall	F-Measure	ROC	Mean Absolute Error
Decision Tree	74.17%	0.774	0.742	0.749	0.802	0.3
Naive Bayes	76.82%	0.765	0.768	0.766	0.813	0.3
Logistic Regression	**78%**	**0.78**	**0.78**	**0.78**	**0.81**	**0.3**

Conclusion: The Logic Regression model shows the best performance with the highest accuracy of 78% and with a ROC of 81% when compared to Decision Tree and Naive Bayes models.

3. Post-prediction Analysis

Cluster Analysis

Clustering is the process of grouping similar objects together based on their characteristics. It is an unsupervised learning technique, in which it determines the natural grouping of instances given for unlabeled data. (6) There are many types of clustering, but in this project, the k-Means clustering method was used.

Clustering was applied by initially setting the value of K = 2, (as in the dataset only two types of patients exist), one for patients with diabetes and the second for patients without diabetes.

The primary use of clustering algorithms is to discover the grouping structures inherent in data. The advantage of this approach is the structures of constructed data sets can be controlled.

```
Number of iterations: 7
Within cluster sum of squared errors: 79.45262213116335

Initial starting points (random):

Cluster 0: 3,106,30.9,24
Cluster 1: 5,97,35.6,52

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute          Full Data          0          1
                     (757.0)    (500.0)    (257.0)
=====================================================
no.of.pregnency       3.8388       2.05     7.3191
glucose.conc.       121.1189    115.636    131.786
BMI                  32.3215    32.0498    32.8502
Age                  33.1189     26.568    45.8638



Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      500 ( 66%)
1      257 ( 34%)


Class attribute: Class
Classes to Clusters:

   0   1  <-- assigned to cluster
 376 120 | 0
 124 137 | 1

Cluster 0 <-- 0
Cluster 1 <-- 1

Incorrectly clustered instances :       244.0    32.2325 %
```

The glucose is the most important factor to identify the diabetics for the person. Because the amount of calculated distance for Glucose is high in cluster 1.

The majority attributes in cluster one based on their distance from the mean of cluster 1 are No. of Pregnancy, BMI and Age.

The visualization provides how each attribute is useful in classification.

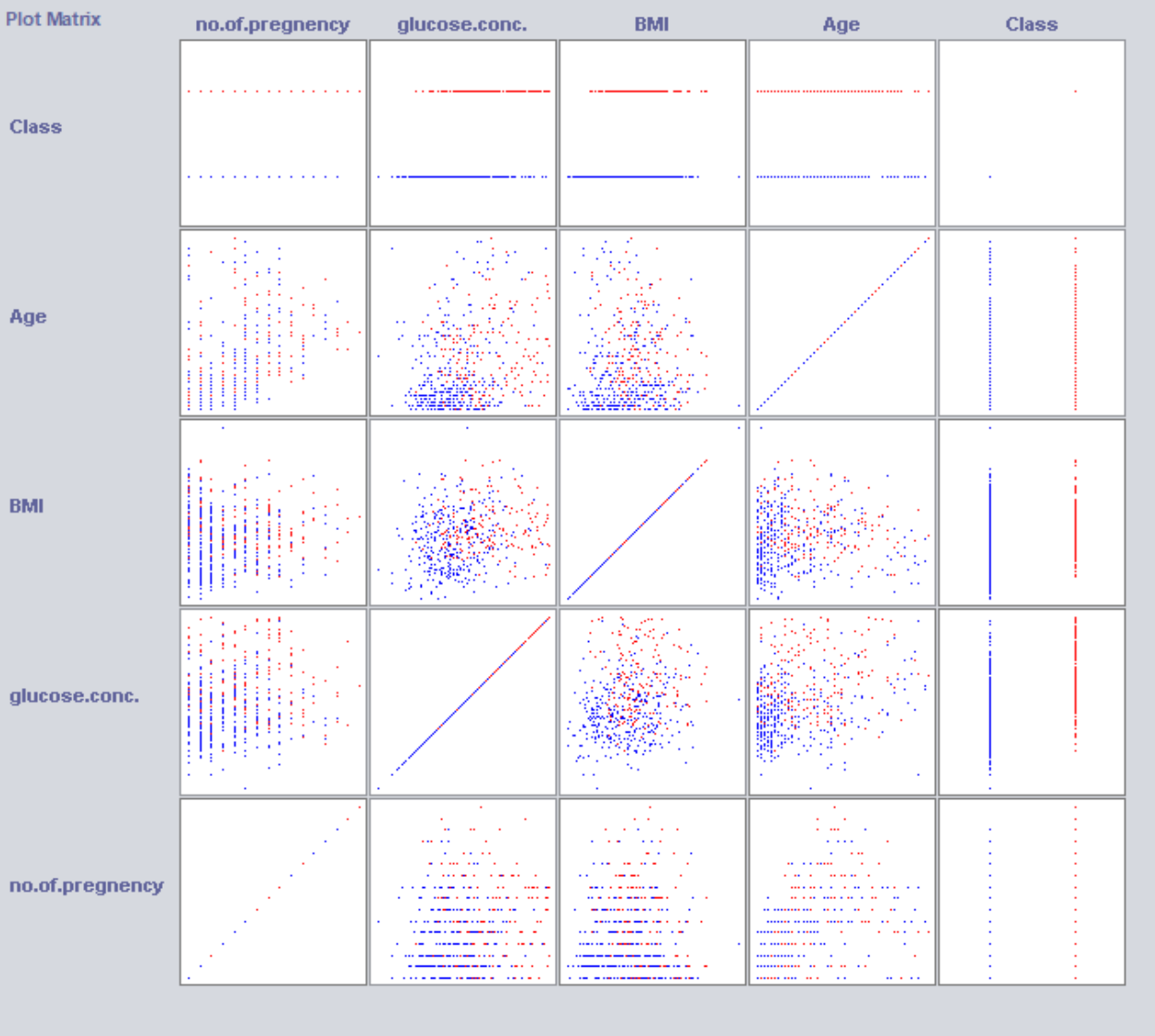


Figure 12

The result of clustering has been shown in blow picture

No.	1: Instance_number	2: no.of.pregnency	3: glucose.conc.	4: BMI	5: Age	6: Class	7: **Cluster**
	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal	Nominal
1	0.0	6.0	148.0	33.6	50.0	1	cluster1
2	1.0	1.0	85.0	26.6	31.0	0	cluster0
3	2.0	8.0	183.0	23.3	32.0	1	cluster1
4	3.0	1.0	89.0	28.1	21.0	0	cluster0
5	4.0	5.0	116.0	25.6	30.0	0	cluster0
6	5.0	3.0	78.0	31.0	26.0	1	cluster0
7	6.0	10.0	115.0	35.3	29.0	0	cluster1
8	7.0	2.0	197.0	30.5	53.0	1	cluster1
9	8.0	8.0	125.0	32.0	54.0	1	cluster1
10	9.0	4.0	110.0	37.6	30.0	0	cluster0
11	10.0	10.0	168.0	38.0	34.0	1	cluster1
12	11.0	10.0	139.0	27.1	57.0	0	cluster1
13	12.0	5.0	166.0	25.8	51.0	1	cluster1
14	13.0	7.0	100.0	30.0	32.0	1	cluster1
15	14.0	0.0	118.0	45.8	31.0	1	cluster0
16	15.0	7.0	107.0	29.6	31.0	1	cluster1
17	16.0	1.0	103.0	43.3	33.0	0	cluster0
18	17.0	1.0	115.0	34.6	32.0	1	cluster0
19	18.0	3.0	126.0	39.3	27.0	0	cluster0
20	19.0	8.0	99.0	35.4	50.0	0	cluster1
21	20.0	7.0	196.0	39.8	41.0	1	cluster1
22	21.0	9.0	119.0	29.0	29.0	1	cluster1
23	22.0	11.0	143.0	36.6	51.0	1	cluster1
24	23.0	10.0	125.0	31.1	41.0	1	cluster1
25	24.0	7.0	147.0	39.4	43.0	1	cluster1
26	25.0	1.0	97.0	23.2	22.0	0	cluster0
27	26.0	13.0	145.0	22.2	57.0	0	cluster1
28	27.0	5.0	117.0	34.1	38.0	0	cluster1
29	28.0	5.0	109.0	36.0	60.0	0	cluster1

- **Clustering performance:**

The primary use of clustering algorithms is to discover the grouping structures inherent in data. The advantage of this approach is the structures of constructed data sets can be controlled.

Table 12

	Cluster 1	Cluster 2
Tested positive	124	137
Tested negative	376	120

Incorrectly classified instances were 32.23% which show that the accuracy of K-means clustering method was 67.77%.

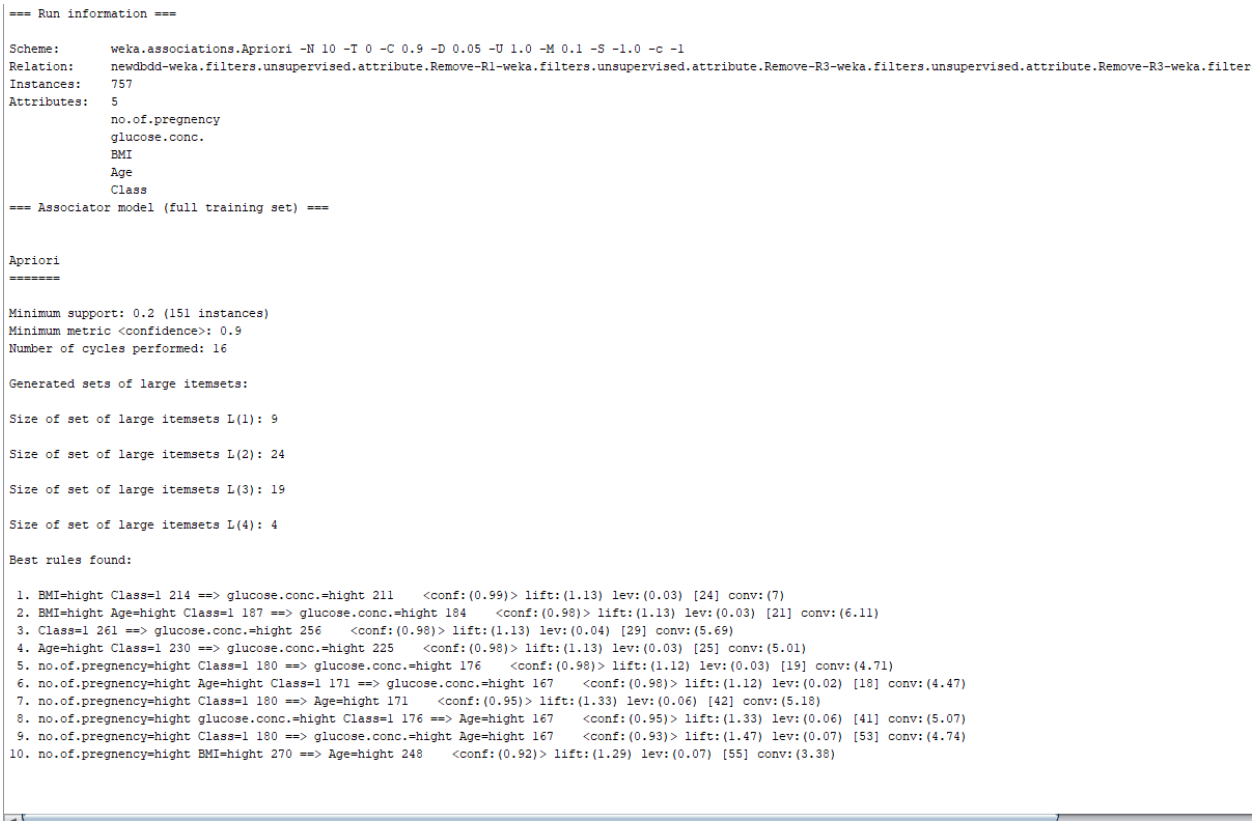
Result:

The most attractive property of the k-means algorithm in data mining is its efficiency in clustering large data sets. Classification is a data mining technique used to predict group membership for data instances (7). The classification is done using this algorithm and successfully classified the data set into two class labels namely tested positive and tested negative. K-mean might not be a good method for clustering diabetics and non-diabetics, because it does not guarantee clustering, since it generates different cluster centroid each time.

Pattern Mining:

Data mining techniques are also used to extract useful information to generate rules. Association rule mining is an important branch to determine the patterns and frequent items used in the dataset. It contains two parts: 1) Determines the frequent item set, 2) Generate rules.

There are several methods to generate rules from data using association rule mining algorithms. This test has been done by Apriori algorithm. Apriori works as an iterative method to identify the frequent item set in a given dataset, and to generate important rules from it. To determine the association between two item sets X and Y, there is a need to set the minimum support of that fraction of transactions which contains both X and Y called minsupp. The other important task is to set the minimum confidence that measures how often items in Y appear in transactions that contain X, known as minconf, to determine frequent item sets .There were only 268 patients with diabetes in the dataset, so only those instances were used to generate rules among them. To develop rules from a given dataset, set minimum support as 0.25 and minimum confidence as 0.9 to generate the following three different rules.



```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     newdbdd-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R3-weka.filters.unsupervised.attribute.Remove-R3-weka.filter
Instances:    757
Attributes:   5
              no.of.pregnency
              glucose.conc.
              BMI
              Age
              Class
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.2 (151 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 9

Size of set of large itemsets L(2): 24

Size of set of large itemsets L(3): 19

Size of set of large itemsets L(4): 4

Best rules found:

 1. BMI=hight Class=1 214 ==> glucose.conc.=hight 211    <conf:(0.99)> lift:(1.13) lev:(0.03) [24] conv:(7)
 2. BMI=hight Age=hight Class=1 187 ==> glucose.conc.=hight 184    <conf:(0.98)> lift:(1.13) lev:(0.03) [21] conv:(6.11)
 3. Class=1 261 ==> glucose.conc.=hight 256    <conf:(0.98)> lift:(1.13) lev:(0.04) [29] conv:(5.69)
 4. Age=hight Class=1 230 ==> glucose.conc.=hight 225    <conf:(0.98)> lift:(1.13) lev:(0.03) [25] conv:(5.01)
 5. no.of.pregnency=hight Class=1 180 ==> glucose.conc.=hight 176    <conf:(0.98)> lift:(1.12) lev:(0.03) [19] conv:(4.71)
 6. no.of.pregnency=hight Age=hight Class=1 171 ==> glucose.conc.=hight 167    <conf:(0.98)> lift:(1.12) lev:(0.02) [18] conv:(4.47)
 7. no.of.pregnency=hight Class=1 180 ==> Age=hight 171    <conf:(0.95)> lift:(1.33) lev:(0.06) [42] conv:(5.18)
 8. no.of.pregnency=hight glucose.conc.=hight Class=1 176 ==> Age=hight 167    <conf:(0.95)> lift:(1.33) lev:(0.06) [41] conv:(5.07)
 9. no.of.pregnency=hight Class=1 180 ==> glucose.conc.=hight Age=hight 167    <conf:(0.93)> lift:(1.47) lev:(0.07) [53] conv:(4.74)
10. no.of.pregnency=hight BMI=hight 270 ==> Age=hight 248    <conf:(0.92)> lift:(1.29) lev:(0.07) [55] conv:(3.38)
```

- **Best rules are shown in:**

Rule1: If (BMI= high ∩ glucose.conc.= high) → Class = Yes

Rule2: If (Age = high ∩ glucose.conc.= high) → Class = Yes

Rule3: If (no.of.pregnancy = high ∩ age= high) → Class = Yes

Rule4: If (BMI = high ∩ glucose.conc.= high) → Class = Yes

4. Conclusions and Recommendations

Conclusion:

- The previous experimental study was conducted using 3 Machine Learning classifiers to predict the likeliness of diabetes. These models were used to compare their performance in terms of accuracy, precision, recall, accuracy and ROC Score. The final result of this experiment shows that the Logic Regression classifier plays the best performance in this prediction of diabetes with a highest accuracy of 78% in comparison to the other models.

- According to the Pima Indian dataset study results, Glucose is the stronger contributor for predicting diabetes.

- The methods utilized in this study could be used to train a model to predict a person's diabetes by taking inputs features like Glucose, BMI, Age and No. of Pregnancy.

Recommendation:

- The study for this dataset has some limitations such as: the target is defined as females older than 21 years which could be extended to male patients as well. By including both genders in the dataset, the prediction of diabetes would work in general and not for a specific community.

- For future and more extensive study if the company would like to predict the type of diabetes (1 and 2), they could provide more medical information on the dataset.

References

1. Public Health Agency of Canada. [Online] 2017. https://www.canada.ca/en/public-health/services/chronic-diseases/diabetes.html.
2. Introduction to Decision Trees . [Online] https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb.
3. Naive Bayes classifier. [Online] https://en.wikipedia.org/wiki/Naive_Bayes_classifier.
4. Linear Models in R: Diagnosing Our Regression Model. [Online] https://www.theanalysisfactor.com/linear-models-r-diagnosing-regression-model/.
5. Residual Plots Part 3— Scale-Location Plot. [Online] https://medium.com/data-distilled/residual-plots-part-3-scale-location-plot-113e469b99c.
6. Cluster analysis. [Online] https://en.wikipedia.org/wiki/Cluster_analysis.
7. Ioannis Kavakiotis, a,b,∗ Olga Tsave,c Athanasios Salifoglou,c Nicos Maglaveras,b,d Ioannis Vlahavas,a and Ioanna Chouvardab,d. Machine Learning and Data Mining Methods in Diabetes Research.