A comparison of different Machine learning classifiers to predict

if a patient has diabetes or not.

Prepared by:

Ashraf Fashafsheh

Submitted in partial fulfillment of the requirements of the "HarvardX PH125.9x Data Science: Capstone"

22/2/2022

Contents

1 . Introduction:

Diabetes is a chronic condition in which the body develops a resistance to insulin, a hormone which converts food into glucose. Diabetes affects many people worldwide and is normally divided into Type 1 and Type 2 diabetes. Both have different characteristics .

This research intends to analyze and create a model on the PIMA Indian Diabetes dataset to predict if a infected or un infected.

Diabetes mellitus has become a major global public health problem in recent time. According to the International Diabetes Federation, there are currently 246 million diabetic people worldwide, and this number is expected to rise to 380 million by 2025. (1)

Data mining represents a significant advance in the type of analytical tools data mining into medical analysis are to increase diagnostic accuracy, to reduce costs and to save human resources. (2)

In recent years Data mining has been successfully applied to healthy databases to automate analysis of huge volumes of complex data, to predict and classification of disease. There are various major data mining techniques such as Support Vector Machine (SVM), Neural Network, Naïve Bayes and Design Tree that have been developed and used in data mining.

2. executive summary section

2 .1 Study population

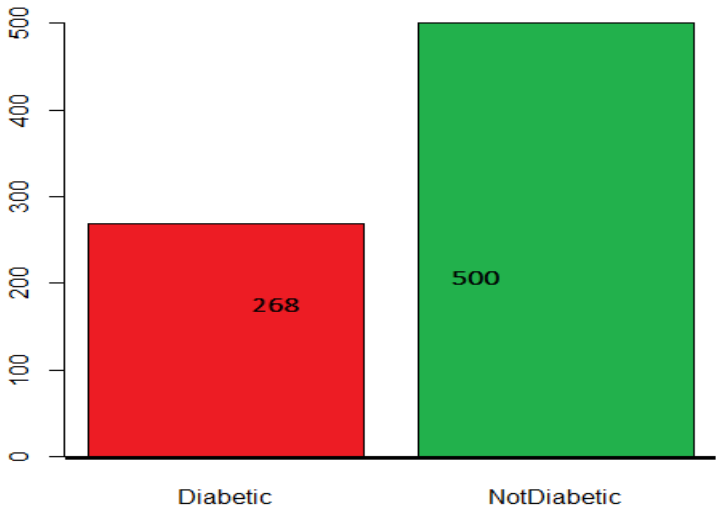
All patients are females at least 21 years old of Pima Indian.

2. 2 **Study Sample**

The study Sample consists of 768 females from Pima Indian heritage. This Sample taken from the National Institute of Diabetes and Digestive and Kidney Diseases.

2.3 **Dataset**

Pima Indians Diabetes dataset contains clinical tests and diagnoses of Pima Indian women of 21 years of age and above with diabetes. The dataset is made up of integer and real number data types. It has 768 instances eight predictive features and a class feature (one response), this Dataset consists of two groups as shown following figure



2.4 **Patient Dataset Variables**

Independent variables

- ✓ (Pregnancies)Number of times pregnant.
- ✓ (Glucose)Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
- ✓ (Blood pressure)Diastolic blood pressure (mm Hg).
- ✓ (Skin thickness)Triceps skin fold thickness (mm).
- ✓ (insulin)2-Hour serum insulin (mu U/ml).
- ✓ (BMI)Body mass index (weight in kg/(height in m)^2).
- ✓ Diabetes pedigree function.
- ✓ Age (years).

Dependent Variables (Class)

- ✓ Outcome (0: uninfected, 1: infected).

2. 5 **Goals of our study**

- Predict if a patient has diabetes or not by using Machine Learning (ML) such as Support Vector Machine (SVM), Naïve Bayes, Design Tree and Neural Network.
- Identify the best kernel used in support vector machines in terms of accuracy, sensitivity and privacy.
- Determine the best tool for Machine Learning in terms of accuracy in the prediction process.

3. Methodology

3.1 **Introduction**

In this study, we will use Support Vector Machines (SVM Kernel Functions), Naive Bayes, Artificial neural network and Decision tree to predict Diabetes; this algorithm will be executed on Diabetes dataset that we tacked from the National Institute of Diabetes and Digestive and Kidney Diseases.

3.2 The algorithm's used

Support Vector Machine (SVM) is a supervised machine learning algorithmic rule which might be used for each classification or regression challenges. However, it's principally utilized in classification issues. In this algorithmic rule, we plot each data item as a point in n-dimensional space where n is number of features one has with the value of each feature being the value of a particular coordinate [3]. Then, we perform classification by finding the hyper-plane that differentiates the two classes well shown in the figure below.

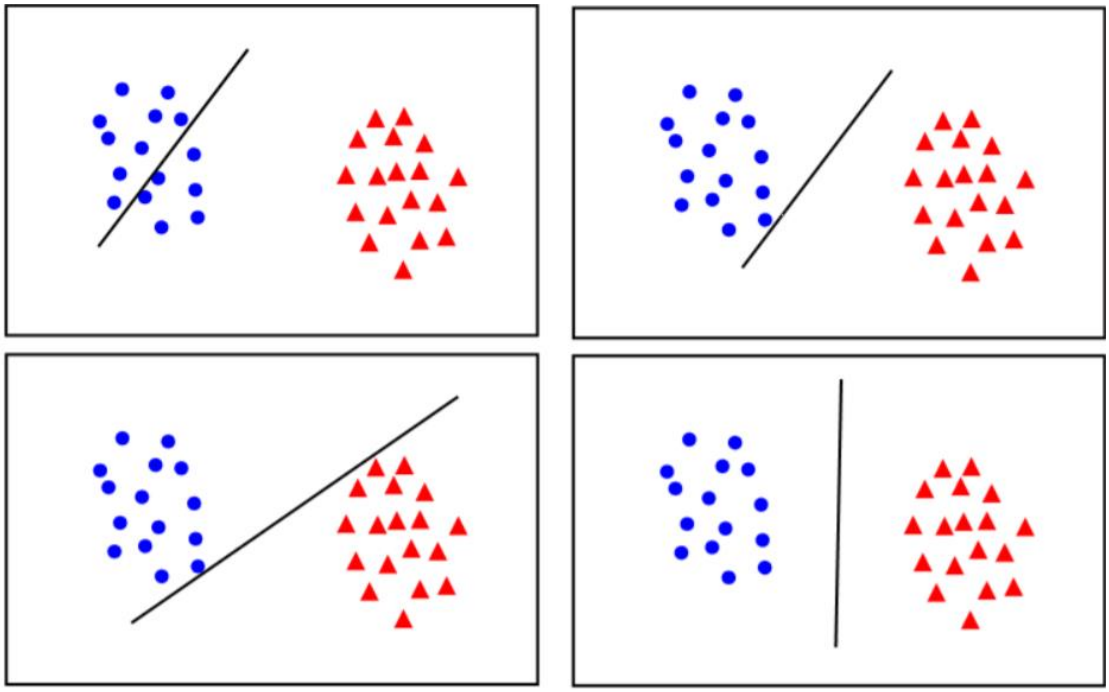


Figure 1 shows the perform classification by finding the hyper-plane that differentiates the two classes.

SVM is a one of the popular machine learning algorithm for regression, classification. It is a supervised learning algorithm that analyses data used for classification and regression. SVM modeling involves two steps, firstly to train a data set and to obtain a model & then, to use this model to predict information of a testing data set. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane where SVM model represents the training data points as points in space and then mapping is done so that the points which are of different classes are divided by a gap that is as wide as possible. Mapping is done in to the same space for new data points and then predicted on which side of the gap they fall.

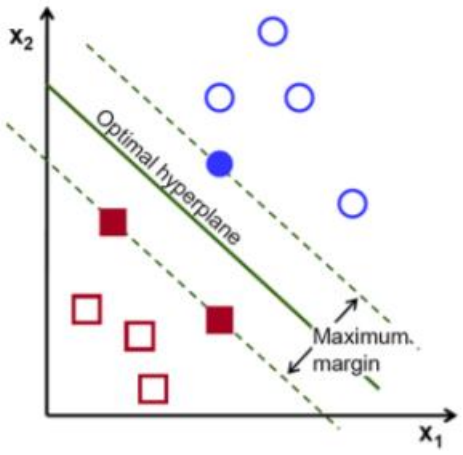


Figure 2 SVM model Graph

In SVM algorithm, plotting is done as each data item is taken as a point in n-dimensional space where n is number of features, with the value of each feature being the value of a particular coordinate. Then, classification is performed by locating the hyper-plane that separates the two classes very well.

3.2.1 **SVM Kernel Functions**

SVM algorithms use a set of mathematical functions that are defined as the kernel. The function of kernel is to take data as input and transform it into the required form. Different SVM algorithms use different types of kernel functions. These functions can be different types, for example *linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid.*

Introduce Kernel functions for sequence data, graphs, text, images, as well as vectors. The most used type of kernel function is RBF, because it has localized and finite response along the entire x-axis.

The kernel functions return the inner product between two points in a suitable feature space, thus by defining a notion of similarity, with little computational cost even in very high-dimensional spaces.

3.2.2 **Naive Bayes**

It's one of the most popular machine learning methods, as it is characterized by speed in processing and efficiency in prediction operations It relies on the statistical concept Bayes 'theorem, which calculates the probability of a specific result by verifying what is available and known and is called Naive because it depends on the principle of Independence Assumptions.

3.2.3 **Artificial neural network**

Neural networks same as SVMs, is a supervised machine learning algorithm which can handle a variety of classification or pattern recognition problems.
They are trained to generate an output as a combination of the input variables.

3.2.4 **Decision tree**

Decision tree (DT) is supervised learning that's used for classification and regression. A decision tree helps decision-makers in knowing all possible alternatives and the possibilities of obtaining them and uses the best option among future cases.

3.3 **Procedure of Work-Study**

1- Collecting Diabetes patient's dataset from **https://www.kaggle.com/uciml/pima-indians-diabetes-database**.

2- Preprocessing: process missing data, outliers.

3- Split Data: We dividing our data set into two subset's.

- Training set (60%): a subset to train a model and applying Machine Learning tools Support Vector Machines (SVM), Neural Network, Naïve Bayes and Design Tree.
- Test set (20%): a subset to test is used to evaluate a given model.

4. **Data Analysis:**

- ❖ Data analysis is conduct using R (programming language)

4.1 Pre-Processing

4.1.1 Missing value

Our variables, Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI all had zero values in there columns as shown in table 1.

Table 1 shows Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI all had zero values.

```
 Pregnancies        Glucose      BloodPressure    SkinThickness      Insulin          BMI        DiabetesPedigreeFunctic
Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   :  0.00   Min.   :  0.0   Min.   : 0.00   Min.   :0.0780
1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.:  0.00   1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437
Median : 3.000   Median :117.0   Median : 72.00   Median : 23.00   Median : 30.5   Median :32.00   Median :0.3725
Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   : 20.54   Mean   : 79.8   Mean   :31.99   Mean   :0.4719
3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.: 32.00   3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262
Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   : 99.00   Max.   :846.0   Max.   :67.10   Max.   :2.4200
      Age             Outcome
Min.   :21.00   Min.   :0.000
1st Qu.:24.00   1st Qu.:0.000
Median :29.00   Median :0.000
Mean   :33.24   Mean   :0.349
3rd Qu.:41.00   3rd Qu.:1.000
Max.   :81.00   Max.   :1.000
```

Since this is not possible for a human to have 0 of any of these features, firstly, we used listwise estimation only complete cases only, then we performed imputation using multiple imputation method for those variables, the accuracy of our model when apply multiple imputation is more than listwise, I compared mean in three cases (before treatment, listwise and multiple imputation), as shown in table 2.

Table 2 shows the mean of independent variables for three cases.

cases	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BIM	Diabetes Pedigree Function	Age
Mean	3.845	120.9	69.11	20.54	79.8	31.99	0.4719	33.24
Mean_del	3.301	122.6	70.66	29.15	156.06	33.09	0.5230	30.86
Mean_imp	3.845	121.7	72.35	28.98	152.4	32.46	0.4719	33.24

Dataset contains 652 missing values, and we are interested in knowing the proportion of missing values by feature, and I used multiple imputations as shown table 3.

Table 3 shows the dataset after multiple imputations

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction
Min. : 0.000	Min. : 44.0	Min. : 24.00	Min. : 7.00	Min. : 14.0	Min. :18.20	Min. :0.0780
1st Qu.: 1.000	1st Qu.: 99.0	1st Qu.: 64.00	1st Qu.:21.00	1st Qu.: 72.0	1st Qu.:27.48	1st Qu.:0.2437
Median : 3.000	Median :117.0	Median : 72.00	Median :29.00	Median :120.0	Median :32.25	Median :0.3725
Mean : 3.845	Mean :121.6	Mean : 72.34	Mean :28.73	Mean :153.1	Mean :32.47	Mean :0.4719
3rd Qu.: 6.000	3rd Qu.:140.2	3rd Qu.: 80.00	3rd Qu.:36.00	3rd Qu.:182.0	3rd Qu.:36.60	3rd Qu.:0.6262
Max. :17.000	Max. :199.0	Max. :122.00	Max. :99.00	Max. :846.0	Max. :67.10	Max. :2.4200

Age	Outcome
Min. :21.00	Min. :0.000
1st Qu.:24.00	1st Qu.:0.000
Median :29.00	Median :0.000
Mean :33.24	Mean :0.349
3rd Qu.:41.00	3rd Qu.:1.000
Max. :81.00	Max. :1.000

4.1.2 Coding Response Feature

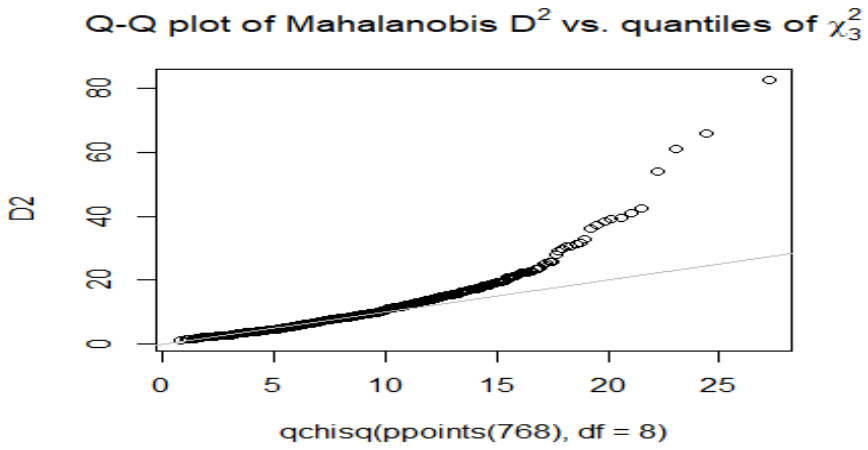
For easier analysis, we will encode the 0 and 1 value for the diagnosis to “Not Diabetic” and “Diabetic”, respectively, by converting the numerical response feature to a factor.

dataset <- **as.factor**(**ifelse**(diabetes**$**outcome == 0, "NotDiabetic", "Diabetic"))

4.1.3 **Outliers**

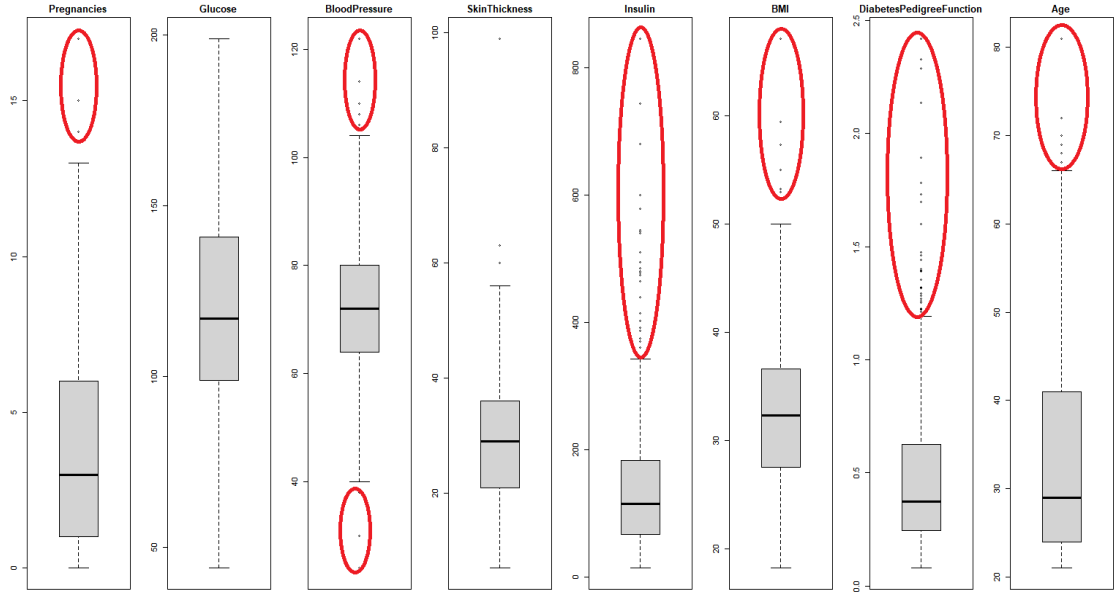
Notice, in the plot in figure 1; cases with extreme Mahalanobis distances (upper right) are likely to be true multivariate outliers. But because these outliers are fact value, so we will not remove it.

Figure 1 shows the outliers in our dataset



In figure 4, we show the outliers of independent variables except Skin Thickness and Glucose. And we note these outliers cannot remove it because outliers are not always bad data points. Sometimes they are the most important data of all.

Figure 2 shows the outliers for eight independent variables



If you analyze an ECG signal, all its peaks are outliers. And they convey the most important information.

4.1.4 Multicollinearity

In the table 4 shows all independent variable not have any of them multicollinearity since all VIF less than 5.

Table 4 shows the value VIF for independent variables

variable	VIF	Variable	VIF
Pregnancies	**1.47**	Insulin	**1.65**
Glucose	**1.68**	BMI	**2.10**
Blood Pressure	**1.28**	Diabetes Pedigree Function	**1.05**
Skin Thickness	**1.96**	Age	**1.70**

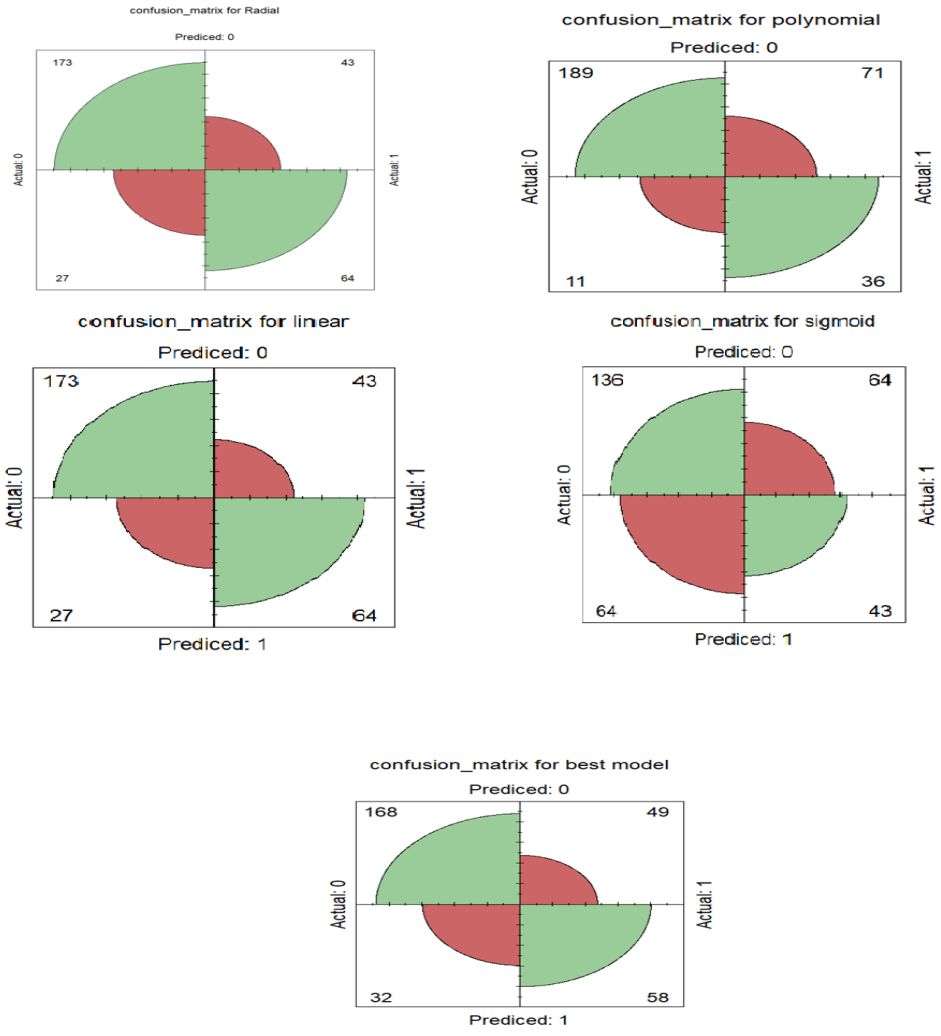
5. Analyze data and Results

In this project we will use Support Vector Machine (SVM), Neural Network, Naïve Bayes and Design Tree Algorithm to predict diabetes.

5.1 Fisrst Algorithm(SVM)

In figure 3 shows the confusion matrix displayed: The four-fold display shows the frequencies in a 2 x 2 table in a way confusion matrix for several kernel. But you can see all models performed similar, the best model is kernel=Redial. Because the best model given kernel =Radial.

Figure 5 shows the confusion matrix for kernel



SVMs have the parameters that impact its algorithm: Cost(C) and gamma.C is the trade-off between the classification of training examples and the margins and gamma is how far the influence of a single training example reaches. A small C maximizes the margin at the cost of accurately classifying training examples, and a large C ensures accurate classification of the training examples at the cost of

maximizing the margin. A small gamma indicates each training example has a far reach from the margins and a large gamma indicates each training example has a close reach from the margins.

Traditional $\epsilon\epsilon$-SVR works with the epsilon-insensitive hinge loss. The value of $\epsilon\epsilon$ defines a margin of tolerance where no penalty is given to errors.

Remember the support vectors are the instances across the margin, *i.e.* the samples being penalized, which slack variables are non-zero.

The larger $\epsilon\epsilon$ is, the larger errors you admit in your solution. By contrast, if $\epsilon \rightarrow 0+\epsilon \rightarrow 0+$, every error is penalized: you end with many (tending to the total number of instances) support vectors to sustain that.

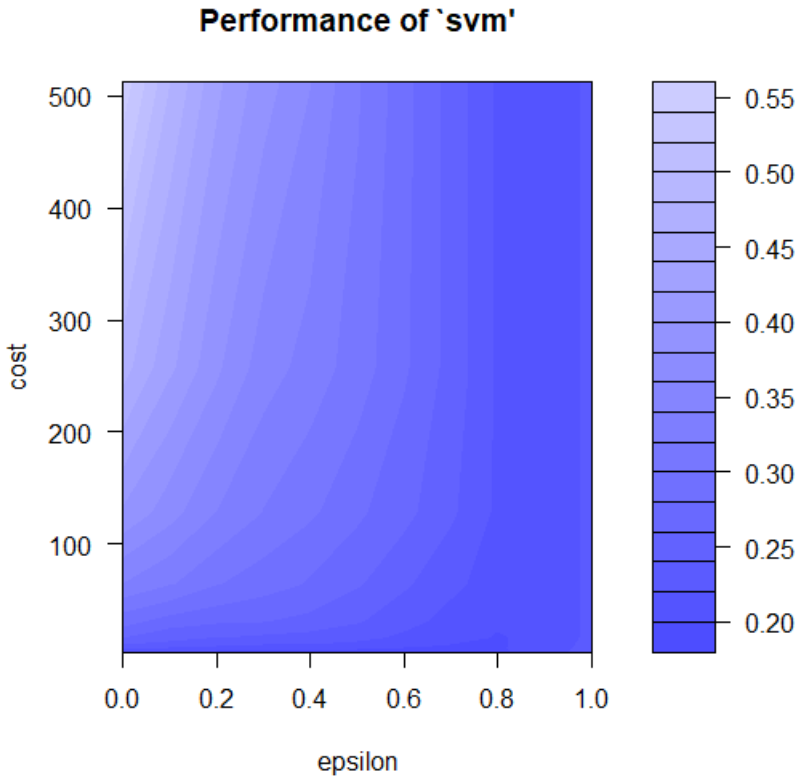
Confusion matrix-kernel=Redial

predicted	Actual		Total
	0	1	
0	173	43	216
1	27	64	91
total	200	107	307

kernel	Accuracy	Misclassification error	Sensitivity	Specificity
Redial	77.85%	22.15%	86.5%	59.8%
polynomial	73.3%	26.7%	94.5%	33.6%
linear	77.2%	22.8%	86.5%	59.8%
sigmoid	58.3%	41.7%	68%	40.2%

From the above table, we observed that the kernel = radial is the best nucleus in terms of accuracy, but in our case study (Predict whether the patient has diabetes or not) we interested sensitivity(is the ability of a test to correctly identify those with the disease (true positive rate)).Tuning the parameters' values for machine learning algorithms effectively improves model performance. Let's look at the list of parameters available with SVM. In the figure 4 we show the performance SVM.

Figure4 shows the performance of SVM



Further fine-tuning and our best values are around cost 500 and epsilon 1.0

svm(x, y = NULL, scale = TRUE, type = NULL, kernel = "radial", degree = 3, gamma = if (is.vector(x)) 1 else 1 / ncol(x), coef0 = 0, cost = 1, nu = 0.5, class.weights = NULL, cachesize = 40, tolerance = 0.001, epsilon = 0.1, shrinking = TRUE, cross = 0, probability = FALSE, fitted = TRUE,..., subset, na.action = na.omit)

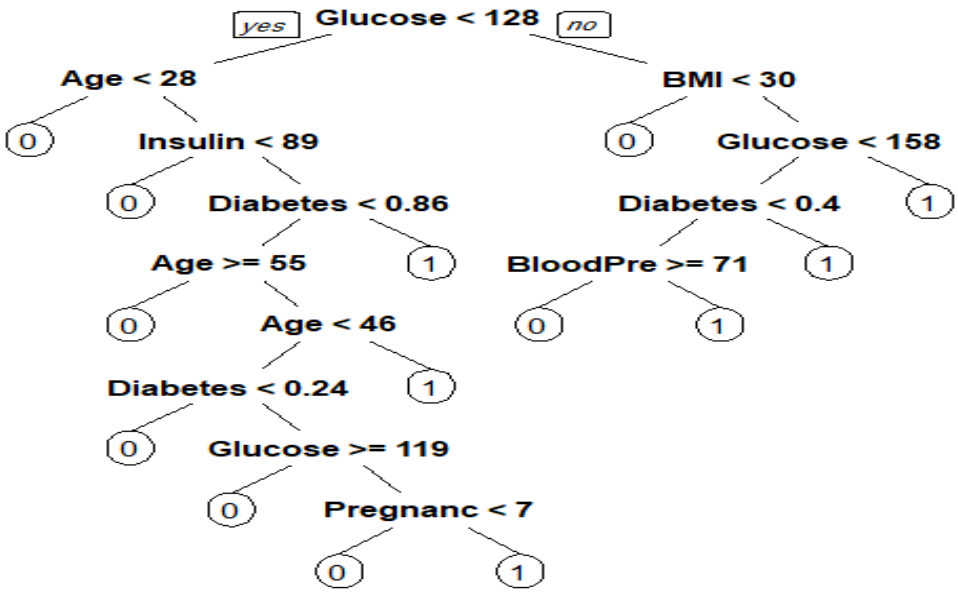
I am going to discuss about some important parameters having higher impact on model performance, “kernel”, “gamma” and “C”.

Kernel: We have already discussed about it. Here, we have various options available with kernel like, “linear”, “radial", "polynomial” and others (default value is “Radial”). Here “radial” and “polynomial” are useful for non-linear hyper-plane.

- ✓ Figure 5 shows the Design Tree output consist of three hidden layer, input layer and output layer. And showing all possible alternatives and the possibilities of obtaining them and uses the best option among future cases.

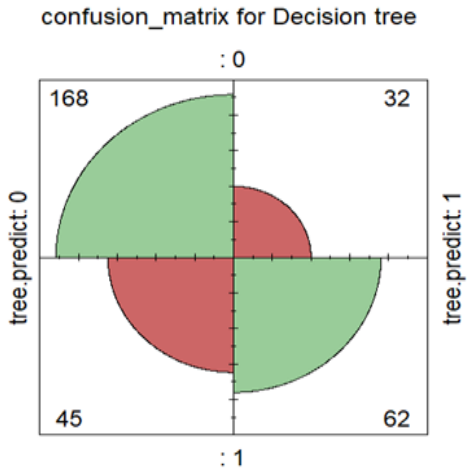
5.2 **Second Algorithm(Design Tree)**

Figure 5 shows the output of design tree



But the figure 6 shows confusion matrix.

Figure 6 shows the confusion matrix of Design Tree.



Algorithm	Accuracy	Misclassification error	Sensitivity	Specificity
Decision Tree	74.9%	25.1%	84%	57.9%

5.3 Third Algorithm(Neural Network)

I n figure 7 and 8 show the output of neural network that consist input variables (independent variables), Hidden layers and output layer, and confusion matrix.

Figure 7 shows the output of neural network

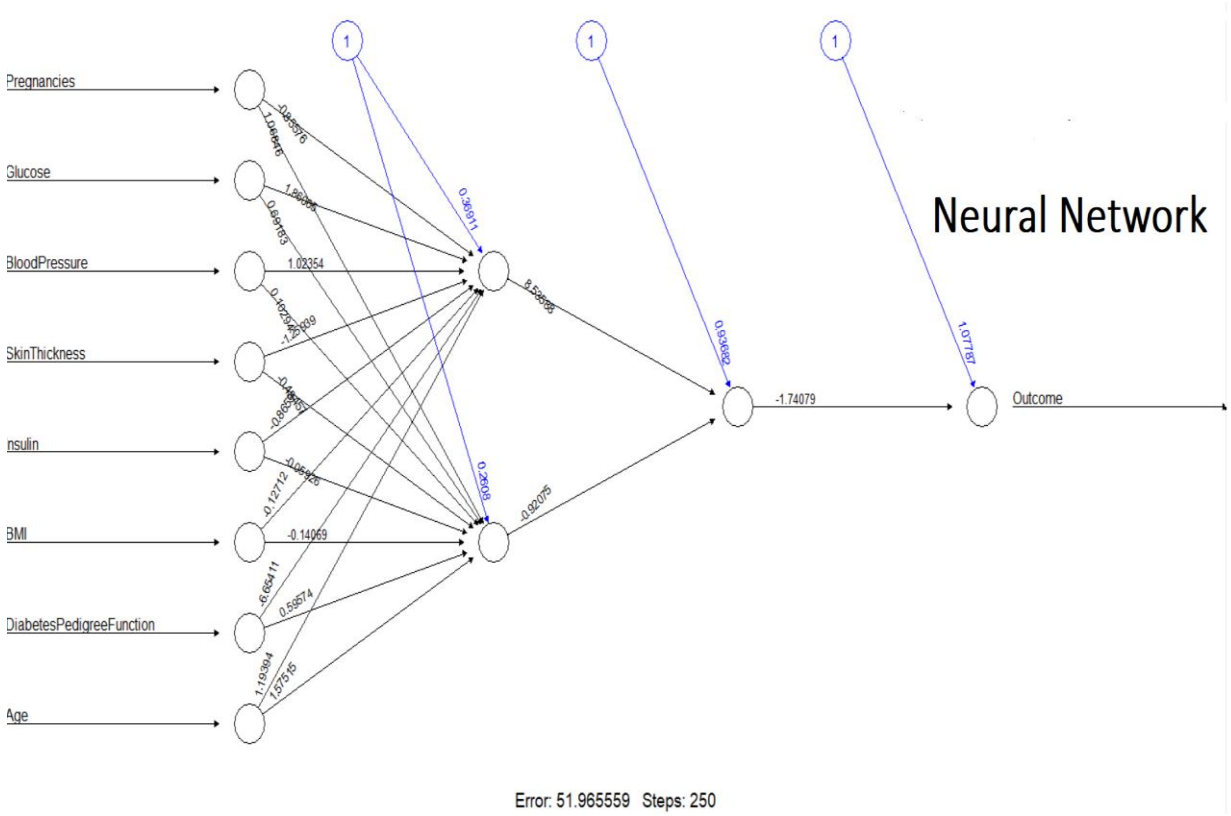
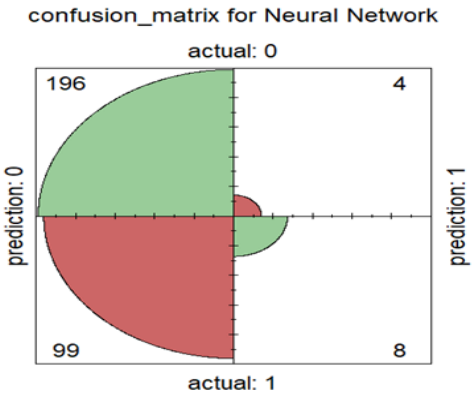


Figure 8 shows the confusion matrix and accuracy of Neural Network

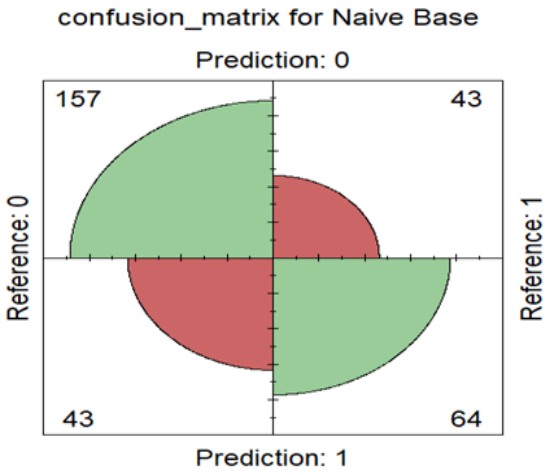


Algorithm	Accuracy	Misclassification error	Sensitivity	Specificity
Decision Tree	64.5%	35.5%	66.4%	66.7%

5.4 Fourth Algorithm(Naïve Bayse)

Figure9 shows the confusion matrix for Naïve Bayes

Figure 9 shows the confusion matrix and accuracy of Naïve Bayes.



Algorithm	Accuracy	Misclassification error	Sensitivity	Specificity
Decision Tree	72%	28%	78.5%	59.8%

Now we will compare between algorithms that used in this project to determine the best tool for Machine Learning in terms of accuracy in the prediction process. As shown in the following table

Algorthim	Accuracy	Sensitivity	Specificity
Support Vector Machine(SVM)	72.2%	59.8%	86.5%
Design Tree	74.9%	57.9%	84%
Artificial Neural Network	64.5%	66.7%	66.4%
Naïve Bayes	72%	59.8%	78.5%

From this table we noted that SVM algorithm is more accuracy than other.

6. **Future Work**

1. Unfortunately day after day, the diabetes is increasing. Therefore, more

Patient's data will be collected to make bigger training data set for further testing and evaluation to increase the prediction and improve the proposed model accuracy with more patients' data.

2- Increase the number of study variables, especially the family record, place of residence, and other variables, with the assistance of experts in the field.

7. Conclusion

In terms of accuracy, Support Vector Machine(SVM) and Design Tree (DT), have scored high 77.2% and 74.9% respectively. But in our project we interested in specificity that is the ability of a test to correctly identify those with the disease (true positive rate).

- Radial is the best kernel used in support vector machines in terms of accuracy, sensitivity and privacy.
- The missing value data was processed in two ways (listwise, and multiple imputation) and the accuracy results were close to 0.74 to 0.772

References

1. International Diabetes Federation, Diabetes Atlas, 3rd ed. Brussels, Belgium: International Diabetes Federation, 2007.

2. Marjan Khajehei, Faried Etemady, "Data Mining and Medical Research Studies," cimsim, pp.119-122, 2010 Second International Conference on Computational Intelligence, Modelling and Simulation, 2010.

3. Puneet Yadav, Rajat Varshney, Vishan Kumar Gupta. Diagnosis of Breast Cancer using Decision Tree Models and SVM (2016