DIABETES PREDICTION USING MACHINE LEARNING ALGORITHMS

C.Vijayalakshmi[1], Hema.B[2], Athira[1]

[1]Department of Statistics and Applied Mathematics
Central University of Tamil Nadu, Thiruvarur
[2]Department of Corporate Secretaryship
Kumararani Meena Muthiah College of Arts and Science, Chennai

ABSTRACT

Diabetes is a chronic disease. Diabetes rate is increasing day by day. Even the youngster of this generation face this disease. Diabetes occurs because of high sugar in blood. It destroys the life of human by this disease worldwide. One can study huge datasets by using big data analytics. It helps to find more informative knowledge to predict the data's outcome, hidden information and pattern regarding the datasets. When the sugar content in blood is high it produces many symptoms . Diabetes can also cause many other diseases like blindness, kidney failure, amputations, heart failure and stroke. Machine learning is a method of learning machine's language by emerging scientific data science fields in many ways. A diabetes prediction model is combined by overall result in this. It has few factors like Glucose, BMI, Age, Insulin etc. for this interpretation of the data. Comparative analysis also carried out to know about the accuracy rate.

KEYWORDS : Diabetes prediction; machine learning; networks;

INTRODUCTION

Diabetes is a common disease. It affects people of all age groups. First comprehend what occurs in the body without diabetes in order to comprehend diabetes and how it develops. It is caused by high sugar levels in blood. Carbohydrates that consume is sugar. Energy source of body is Carbohydrate meals and it must be consumed by all. Foods are broken down into glucose by the body and eat them. In the stream of the blood, circulation of glucose is happening in the body. Glucose may transport to brain, where it helps in performing and thinking process. Hyperglycemia is a typical complication of untreated diabetes that damages many organs, including neurons and blood vessels, over time. Diabetes is spread across the world, according to International Diabetes Federation a study was conducted and analyzed that millions of people suffer from diabetes. In 2014 and above 8.5 percent persons affected diabetes. Diabetes was the reason for 1.5 million deaths in 2019. All diabetes – related deaths were 48 percent and it was under the age of 70. By 2045, an estimated 700 million will have diabetes across the world, according to the International Diabetes Federation. With 1.5 million deaths projected in 2019, diabetes is the main causes of death worldwide. Due to a lack of effective treatment choices, diabetes more than doubles the risk of dying early in many parts of the world. The Western Pacific has the highest amount of deaths, with over 717,000 people succumbing to the disease in 2021. The most important indicator of diabetes progression is Obesity. The International Diabetes Federation believes that in certain wealthy nations, the number of newly diagnosed diabetes patients is actually decreasing. However, diabetes is rising as people's living situations improve. Because of the changing lifestyle, many people are suffering from this disease. The organization also sees the lack of diabetes diagnosis as a persistent issue. The cause of diabetes is unknown, but lifestyle and genetic factors have a significant impact. Diabetes patients are at risk of having many problems like heart disease and damage of nerve.

LITERATURE SURVEY

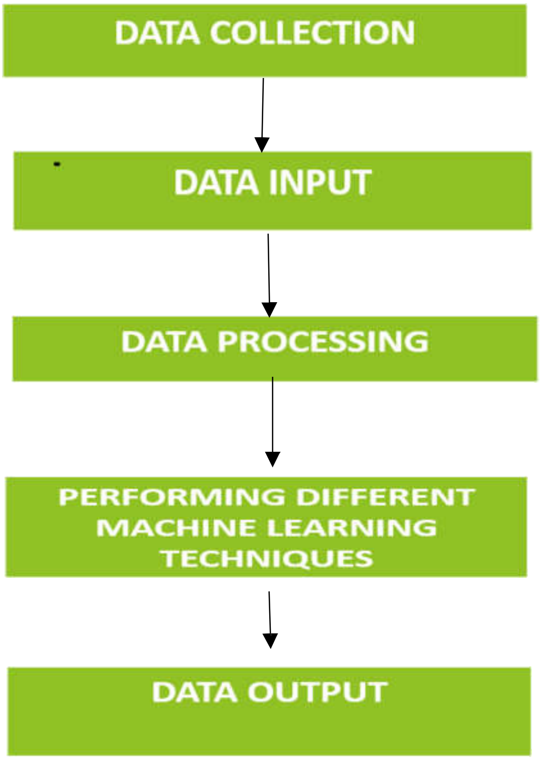
Name of author	Title Name	Journal Name
Someswar Pal, Navneet Mishra, Megha Bhusthan, Pankaj Singh Kholiya, Meghavi Rana,Arun Negi	*Deep Learning Techniques for Prediction and Diagnosis of Diabetes Mellitus*	*International Mobile and Embedded Technology Conference (MECON), March 2022*
Aastha Singh, Madhulika Bhatia	*Prediction of abnormal pregnancy in pregnant women with Advanced maternal age and Pregestational Diabetes using Machine learning models*	*12th International Conference on cloud Computing, Data Science and Engineering (Confluence) 27-28 Jan. 2022*
Umair Muneer Butt, Sukumar Letchmunan, Musbashir Ali, Fadratul Hafinaz Hassan, Anees Baqir and Hafiz Husnain Raza Sherazi	*Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications*	*Journal of Healthcare Engineering, vol. 2021, Article ID9930985, 17 pages,2021.*
Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadka	*Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop*	*International Conference On I-SMAC,978-1-5090-3243-3,2017*
Ayush Anand and Divya Shakti	*Prediction of Diabetes Based on Personal Lifestyle Indicators*	*1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015*
B. Nithya and Dr. V. Ilango	*Predictive Analytics in Health Care Using Machine Learning Tools and Techniques*	*International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017*
Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S	*Predictive Methodology for Diabetic Data Analysis in Big Data*	*2nd International Symposium on Big Data and Cloud Computing,2015.*

Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly	*Diagnosis of Diabetes Using Classification Mining Techniques*	*International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.*
P. Suresh Kumar and S. Pranavi	*Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics*	*International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017*
Mani Butwall and Shraddha Kumar	*A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier*	*International Journal of Computer Applications, Volume 120 - Number 8,2015*
K. Rajesh and V. Sangeetha	*Application of Data Mining Methods and Techniques for Diabetes Diagnosis*	*International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012*
Humar Kahramanli and Novruz Allahverdi	*Design of a Hybrid System for the Diabetes and Heart Disease*	*Expert Systems with Applications: An International Journal, Volume 35 Issue 1-2, July, 2008.*
B.M. Patil, R.C. Joshi and Durga Toshniwa	*Association Rule for Classification of Type-2 Diabetic Patients*	*ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing, February 09 - 11, 2010*
Dost Muhammad Khan1, Nawaz Mohamudally2	*An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm*	*Journal Of Computing, Volume 3, Issue 12, December 2011*
Mahmood, D.Y.; Hussein, M.A.	*Intrusion detection system based on K-star classifier and feature set reduction*	*IOSR J. Comput. Eng. 2013, 15, 107–112*
Sharma, R.; Kumar, S.; Maheshwari	*R. Comparative Analysis of Classification Techniques in Data Mining Using Different Datasets*	*Int. J. Comput. Sci. Mobile Comput. 2015, 44, 125–134*

METHODOLOGY

Various machine learning strategies used for predicting diabetes. Accurate outcome will be determined for each approach. Various strategies are employed in this instance. The accuracy measurements of the machine learning models are the output. As a result, the diabetes model can be forecasted.

MODEL FORMULATION



DATASET DESCRIPTION

	Pregnancies	*Glucose*	*Insulin*	*BMI*	*Diabetes Pedigree Function*	*Age*	*Outcome*
1	*6*	*148*	*0*	*33.6*	*0.627*	*50*	*1*
2	*1*	*85*	*0*	*26.6*	*0.351*	*31*	*0*
3	*8*	*183*	*0*	*23.3*	*0.672*	*32*	*1*
4	*1*	*89*	*94*	*28.1*	*0.167*	*21*	*0*
5	*0*	*137*	*168*	*43.1*	*2.288*	*33*	*1*
6	*5*	*116*	*0*	*25.6*	*0.201*	*30*	*0*

The trait to predict with 0 and 1 is the outcome. It denotes the absence of diabetes for 0 and the presence of diabetes for 1.

INFERENCE :The dataset comprised 768 samples before outliers were removed. The dataset currently has 636 samples after outliers were removed.

STATISTICAL ANALYSIS

	Pregnancies	*Glucose*	*Insulin*	*BMI*	*Diabetes Pedigree Function*	*Age*	*Outcome*
Min.	*0.000*	*0.0*	*0.00*	*0.00*	*0.0780*	*21.00*	*0:500*
1^{st} Qu	*1.000*	*99.0*	*0.00*	*27.30*	*0.2437*	*24.00*	*1:268*
Median	*3.000*	*117.0*	*30.5*	*32.00*	*0.3725*	*29.00*	
Mean	*3.845*	*120.9*	*79.8*	*31.99*	*0.4719*	*33.24*	
3^{rd} Qu	*6.000*	*140.2*	*127.2*	*36.60*	*0.6262*	*41.00*	
Max	*17.000*	*199.0*	*846.0*	*67.10*	*2.4200*	*81.00*	

INFERENCE

The minimum value is the data's lowest value. The minimal value is for pregnancies, glucose, insulin, and BMI. The first quartile (Q1), which means the (n+1)/4th term. Insulin is at its lowest in the first quarter, while age is at its highest.

The data's midpoint is represented by the median. With the median, it represent a large number of data points with just one.

Median

n is odd,

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{observation}$$

n is even,

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th} \text{observation}}{2}$$

For glucose, the median is high, while for diabetes pedigree function, it is low.

The average of all the data is used in the mean formula.

$$\text{Mean} = \frac{\text{Sum of All Data Points}}{\text{Number of Data Points}}$$

The third quartile(Q3) is given as:

Third quartile (Q3) = (3(n+1)/4th term, also known as the upper quartile.

Third quarter is minimum for diabetes pedigree function and the maximum for glucose.

The maximum is the highest value of the data.

Age has the maximum value

	Minimum	*Maximum*	*Mean*	*Std Dev*	*Range*
Pregnancies	*0*	*17*	*3.845*	*3.37*	*17*
Glucose	*0*	*199*	*120.895*	*31.973*	*199*
Blood Pressure	*0*	*122*	*69.105*	*19.356*	*122*
Skin Thickness	*0*	*99*	*20.536*	*15.952*	*99*
Insulin	*0*	*846*	*79.799*	*115.244*	*846*
BMI	*0*	*67.1*	*31.993*	*7.884*	*67.1*
Diabetes Pedigree Function	*0.078*	*2.42*	*0.472*	*0.331*	*2.342*
Age	*21*	*81*	*33.241*	*11.76*	*60*
Outcome	*0*	*1*	*0.349*	*0.477*	*1*

INFERENCE

Dataset's maximum and minimum values are both relatively simple metrics. The difference between the greatest and least value of the data is the range of the data.

Max Value – Min Value = Range

Highest range is for the glucose.
Highest mean is for glucose.
Highest Standard deviation is for Insulin.

MACHINE LEARNING TECHNIQUES

Instances	*768*
Attributes	*9*
Test mode	*10-fold cross-validation*

In the context of the model, K is the specific value chosen. The value of K is ten in this case. As a result, 10-fold cross-validation is used.
Machine learning models employ cross validation to estimate previously unknown data.
Classifier model (full training set)

ZeroR predicts class value	*0.3489583333333333*

Zero R model for the dataset with an accuracy of 0.3489583333333333 .It is a positive predictive value.
Cross-validation

Correlation coefficient	*-0.0908*
Mean absolute error	*0.4548*
Root mean squared error	*0.4771*
Relative absolute error	*100 %*

Root relative squared error	*100 %*
Total Number of Instances	*768*

INFERENCE

The coefficient of correlation is negative. It describes how far two variables are moving in different directions.

In statistics, MAE is the amount of the deviation between a prediction and its true value.

$$\text{Mean absolute error} = \frac{\text{Sum of all absolute errors}}{\text{No of errors } (n)}$$

$$= \sum_{i=1}^{n} \frac{X_i - X}{n}$$

MAE = mean absolute error
Xi = prediction
X = true value
n = total number of data points

The model will try to reduce the inaccuracy as much as possible. The mean absolute error is minimal in this case.

The RMSE is calculated using the root of the square mean of squared deviation between actual and predicted outcomes.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

RMSE = root mean square error
Yi= prediction
Ŷı= true value
n = total number of data points

The root mean square error is abbreviated as RMSE. Because RMSE is low in this case, it is a better-fitting model. When compared to MSE, RMSE will always have more or equivalent mistakes. The greater the gap between them, the greater the variance in the sample's individual errors.

Clustering
Test mode: evaluate on training data

Clustering model (full training set)

Number of clusters selected by cross validation	*2*	
Number of iterations performed	*2*	
Attribute	1 (0.38)	0 (0.62)
Pregnancies		
mean	4.8483	3.2241
std. dev.	3.768	2.926
Glucose		
mean	142.7926	107.3415
std. dev.	31.308	23.7835
Blood Pressure		
mean	71.6206	67.5488
std. dev.	20.6869	18.2902
Skin Thickness		
mean	22.7605	19.1599
std. dev.	17.4576	14.7598
Insulin		
mean	118.3012	55.9702
std. dev.	155.4014	71.0018
BMI		
mean	35.3693	29.9027

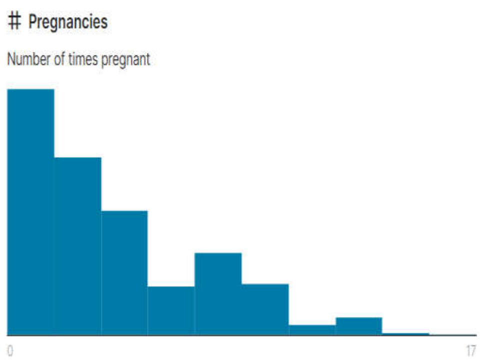
std. dev.	7.0256	7.6503
Diabetes Pedigree Function		
mean	0.579	0.4056
std. dev.	0.4037	0.2552
Age		
mean	37.2137	30.7821
std. dev.	11.3178	11.3365
Outcome		
mean	0.8872	0.0158
std. dev.	0.3163	0.1247

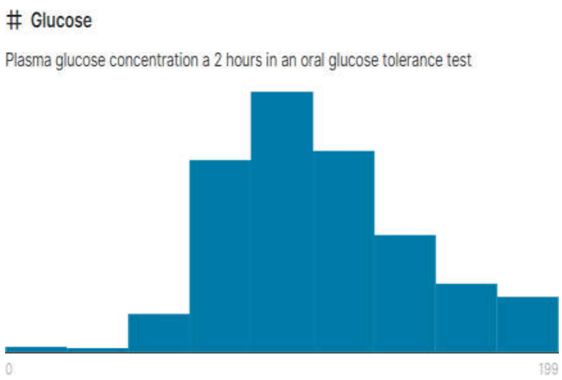
<u>Model and evaluation on training set</u>

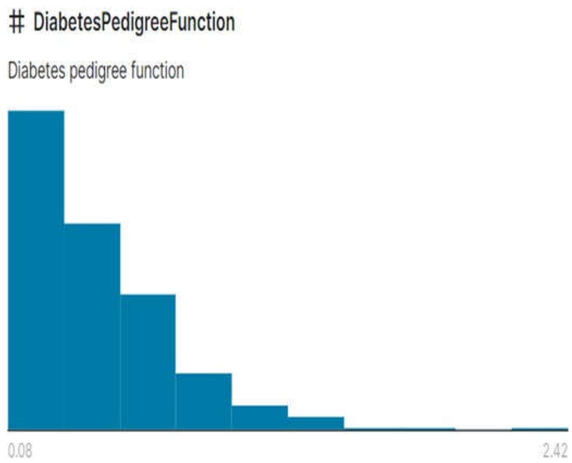
Clustered Instances

0	*289 (38%)*
1	*479 (62%)*
Log likelihood	*-29.34631*

GRAPHICAL REPRESENTATION

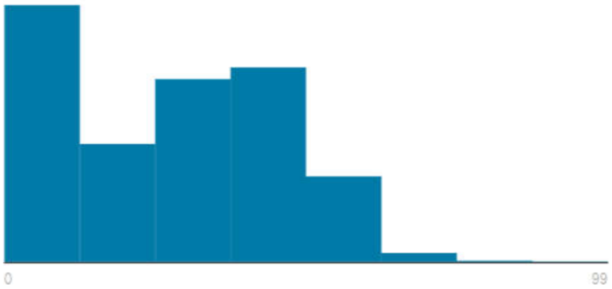






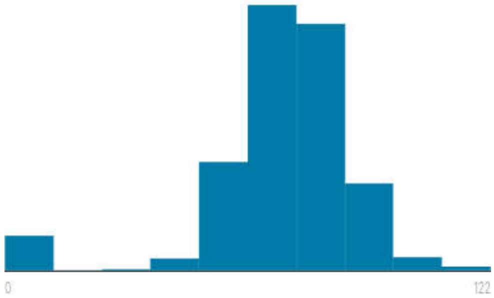
SkinThickness

Triceps skin fold thickness (mm)



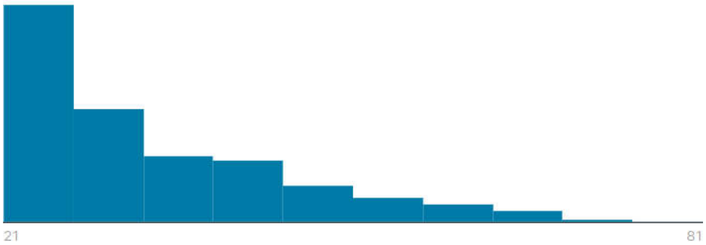
BloodPressure

Diastolic blood pressure (mm Hg)



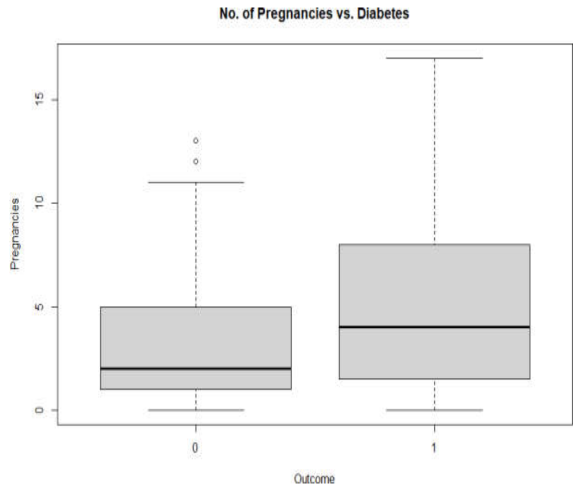
Age

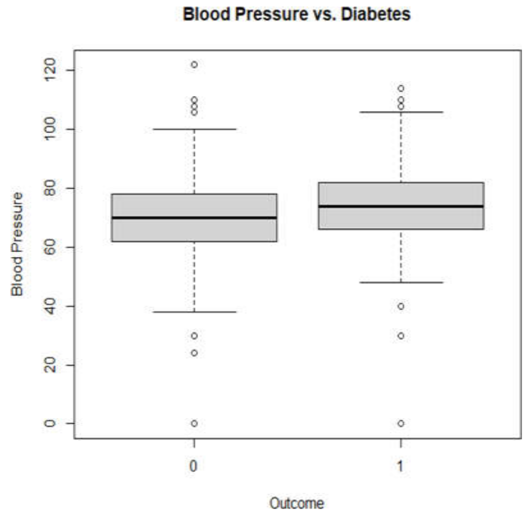
Age (years)

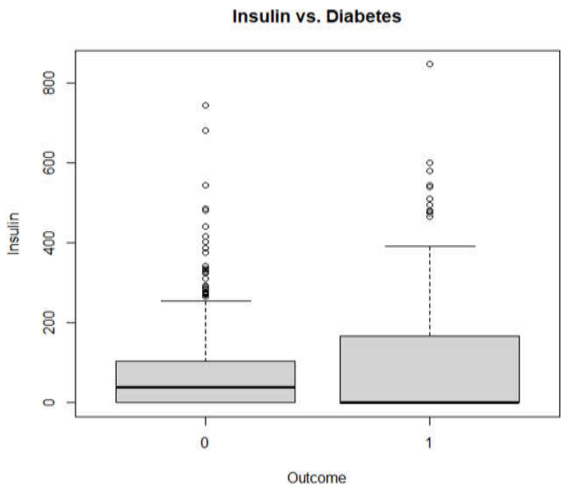


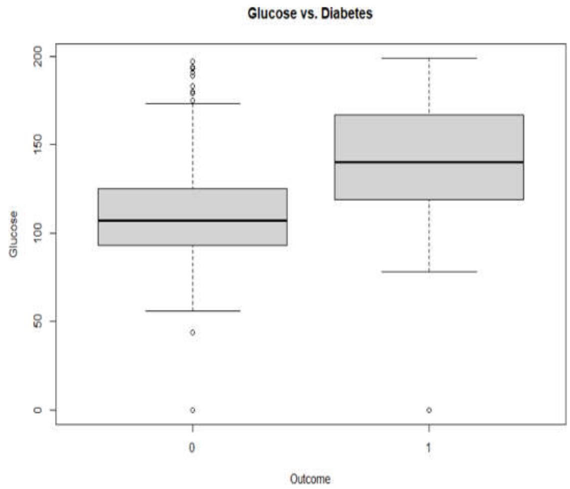
Outcome

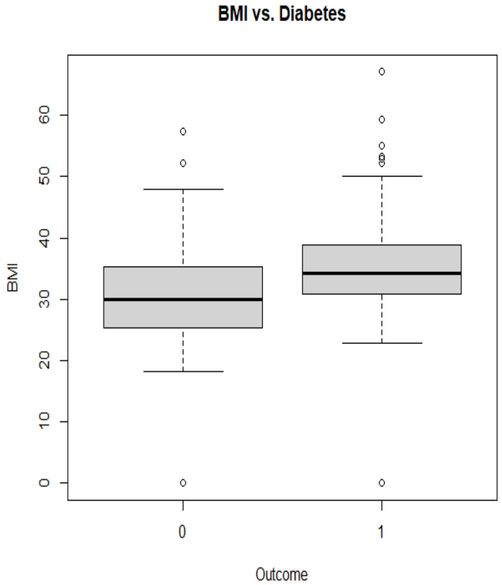
Class variable (0 or 1) 268 of 768 are 1, the others are 0

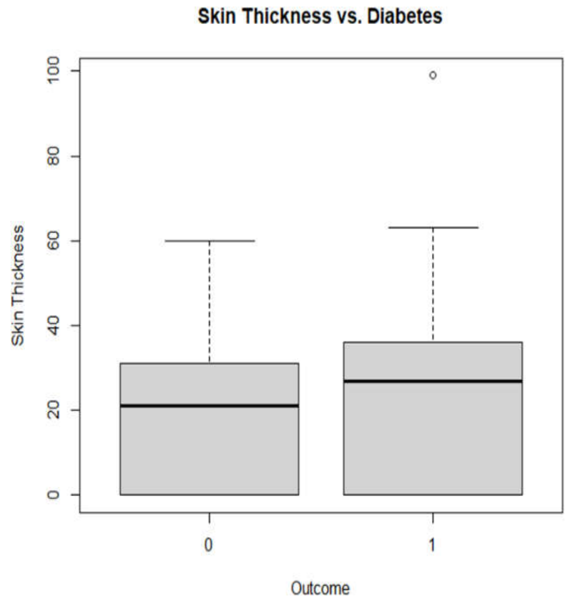

No. of Pregnancies vs. Diabetes
Pregnancies
0
5
10
15
0
1
Outcome

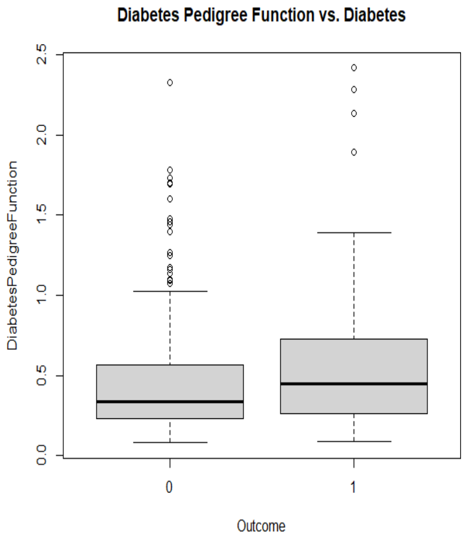

Blood Pressure vs. Diabetes
Blood Pressure
0
20
40
60
80
100
120
0
1
Outcome

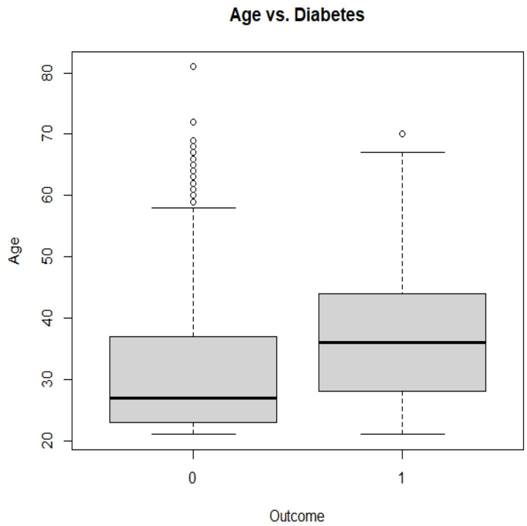

Insulin vs. Diabetes
Insulin
0
200
400
600
800
0
1
Outcome


Glucose vs. Diabetes
Glucose
0
50
100
150
200
0
1
Outcome



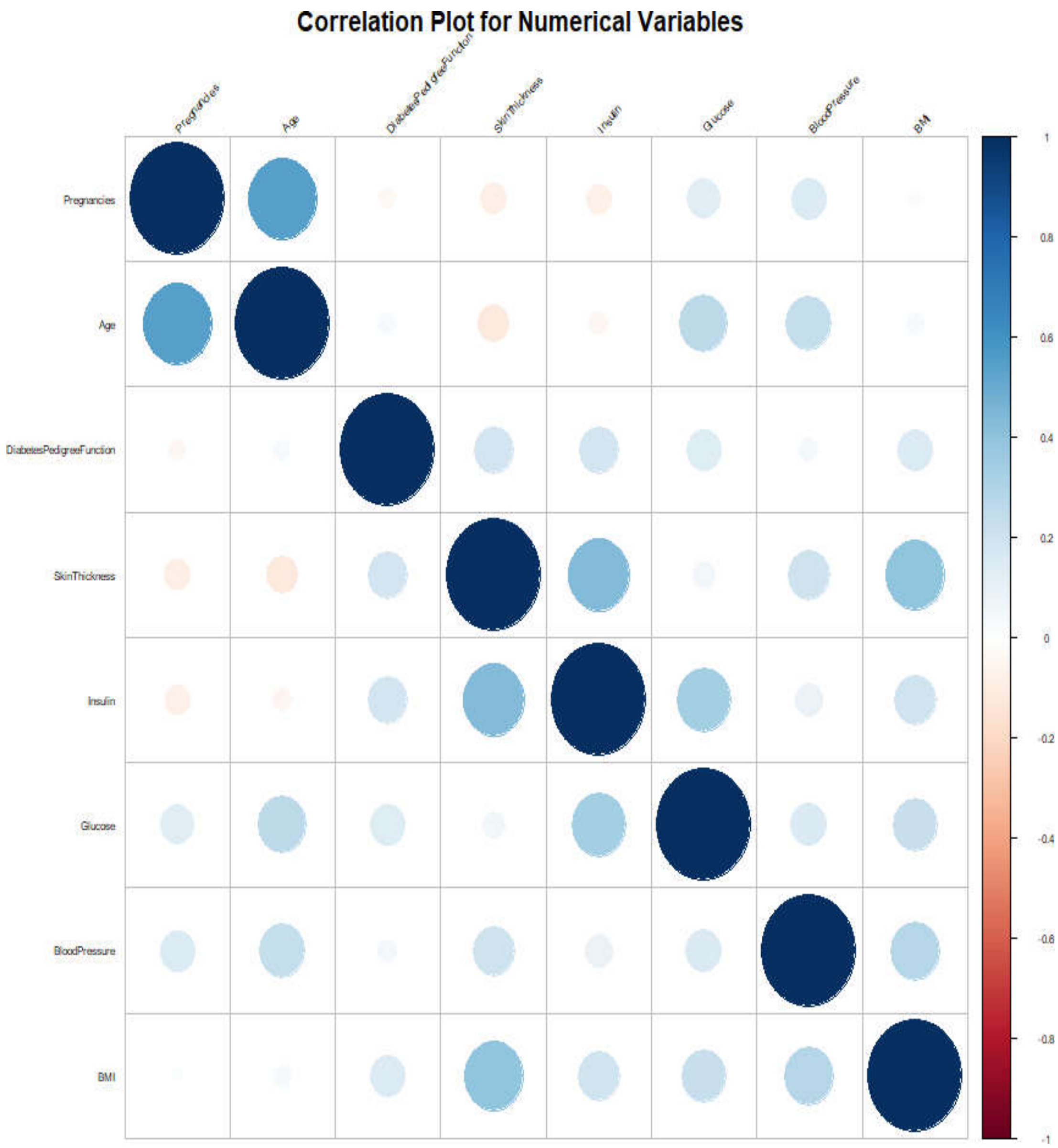






INFERENCE

Above graphs show that the data is biased towards data points having an outcome value of 0.The data is skewed towards data points with an outcome value of 0,which indicates that there is no diabetes present. No. of Pregnancies vs Diabetes graph is positively skewed. Glucose vs Diabetes, Blood Pressure vs Diabetes, BMI vs Diabetes are graphs in which both sides are nearly same. Insulin vs Diabetes, Skin Thickness vs Diabetes, Diabetes Pedigree Function vs Diabetes, Age vs Diabetes graphs are negatively skewed.



INFERENCE

In this data is represented using a dot-representation, positive correlation represents blue and negative correlation represents red. The greater the dot, the stronger the link. Because it shows the correlation of each variable the matrix is symmetrical and that the diagonal is perfectly positively correlated. Unfortunately, there is no correlation between any of the variables.

CORRELATION MATRIX

	Pregnancies	**Glucose**	**Blood Pressure**	**Skin Thickness**	**Insulin**	**BMI**	**Diabetes Pedigree Function**	**Age**
Pregnancies	1.00000000	0.12945867	0.14128198	-0.08167177	-0.07353461	0.01768309	-0.03352267	0.54434123
Glucose	0.12945867	1.00000000	0.15258959	0.05732789	0.33135711	0.22107107	0.13733730	0.26351432
Blood Pressure	0.14128198	0.15258959	1.00000000	0.20737054	0.08893338	0.28180529	0.04126495	0.23952795
Skin Thickness	-0.08167177	0.05732789	0.20737054	1.00000000	0.43678257	0.39257320	0.18392757	-0.11397026
Insulin	-0.07353461	0.33135711	0.08893338	0.43678257	1.00000000	0.19785	0.18507093	-0.04216295
BMI	0.01768309	0.22107107	0.28180529	0.39257320	0.19785906	1.00000000	0.14064695	0.03624187
Diabetes Pedigree Function	-0.03352267	0.13733730	0.04126495	0.18392757	0.18507093	0.14064695	1.00000000	0.03356131
Age	0.54434123	0.26351432	0.23952795	-0.11397026	-0.04216295	0.03624187	0.03356131	1.00000000

INFERENCE

The correlation matrix is entirely symmetrical in this case. There is no single attribute that has a strong relationship with the outcome value. Some traits have a negative association with the outcome value, whereas others have a positive link. The correlation between Pregnancies and Age is highly correlated. The correlation between Pregnancies and Diabetes Pedigree Function is least correlated.

Logistic Regression

Model	Estimate	Std.Error	Z Value	Pr(>|Z|)
Intercept	*-8.4169989*	*0.7106686*	*-11.844*	*2*10^16*
Glucose	*0.0345879*	*0.0036653*	*9.436*	*2*10^16*
Blood Pressure	*-0.0122052*	*0.0052309*	*-2.333*	*0.0196*
Skin Thickness	*0.0008423*	*0.0069023*	*0.122*	*0.9029*
Insulin	*-0.0013691*	*0.0009118*	*-1.502*	*0.1332*
BMI	*0.0878755*	*0.0148520*	*5.917*	*3.28*10^09*
Diabetes Pedigree Function	*0.8904459*	*0.2957255*	*3.011*	*0.0026*
Age	*0.0326696*	*0.0081094*	*4.029*	*5.16*10^05*

INFERENCE

The logistic regression coefficient is 0.734375.

The p-value (Pr (>|Z|)) column represents z-value column. It is statistically significant for the predictor variable in connection to the response variable in the model if the value of p is less than 0.05. Here, the p-values for *glucose*, *blood pressure*, *BMI*, *diabetes pedigree function*, and *age* are less than 0.05. As a result, the model's response variable has a statistical link with all of them.

Null deviance	*993.48 on 767 degrees of freedom*
Residual deviance	*738.68 on 760 degrees of freedom*
AIC	*754.68*

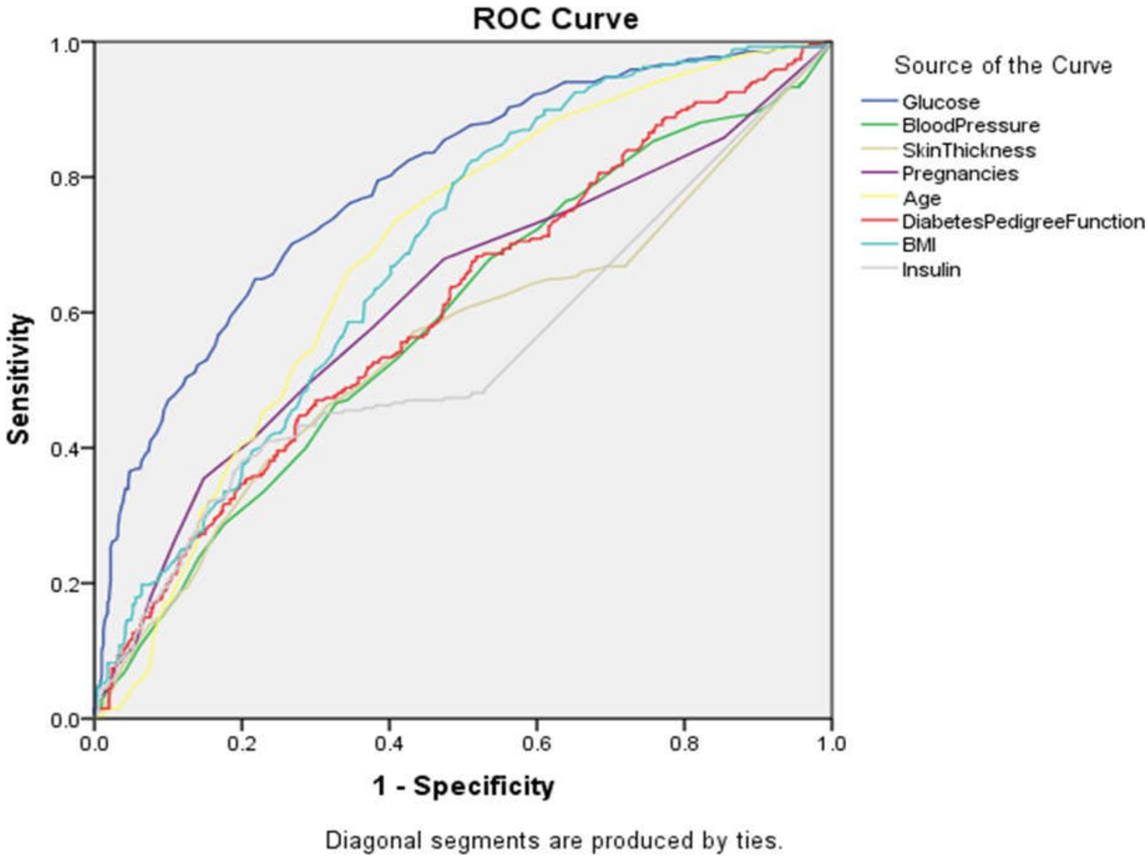
A good match for the model is shown by the deviance between Null Deviance and Residual Deviance. Larger disparity, the more accurate the model is. A null deviance occurs when an equation has just one intercept and has no variables, but a residual deviance occurs when all variables are considered. If the difference is significant enough, the model should be considered good. Because the difference between null and residual deviation isn't large in this case, the model doesn't fit well.

Because the AIC score is so high in this case, the model will not be a better-fit model.

Case Processing Summary

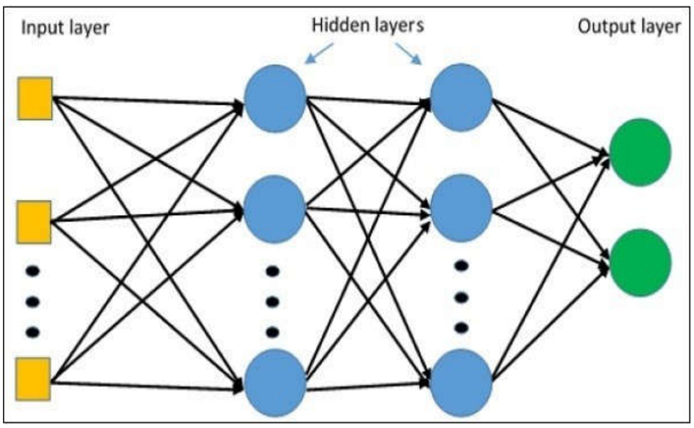
Outcome	Valid N (listwise)
Positive	268
Negative	500

The test result variable(s) with higher values suggest more evidence of a positive actual state. Positive-The actual positive state is 1.



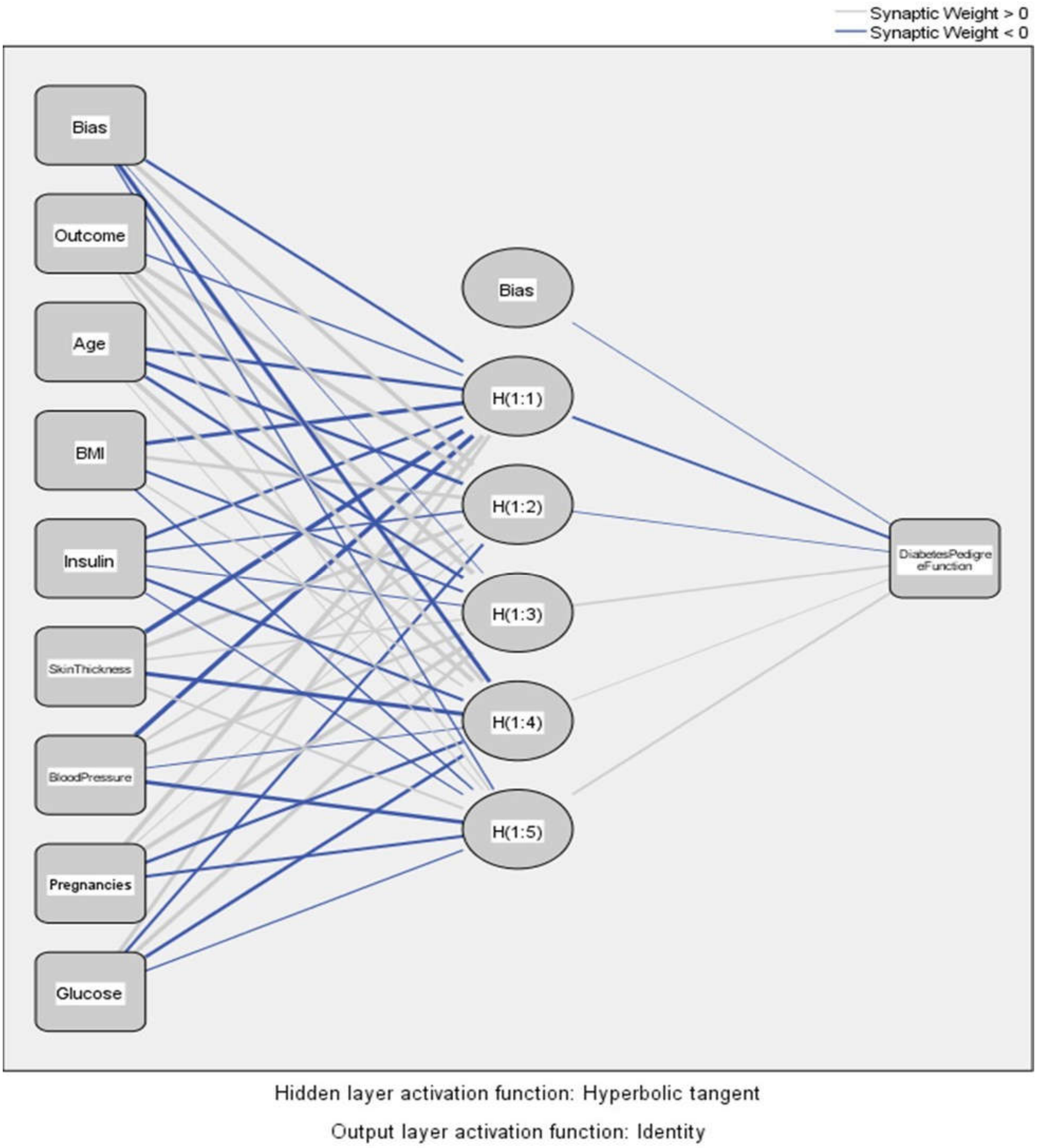
MultilayerPerceptron

- MLP is a type of feed forward ANN that is fully connected ANN.
- It consists of: input layer, hidden layer and output layer.



ANN

ANNs or SNNs underpins deep learning. It enables trainees to improve their accuracy over time by learning and improving their accuracy.



<u>Classification</u>

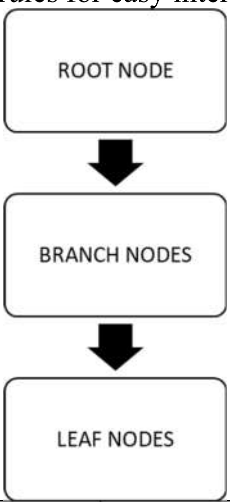
Sampled Observation	Predicted		
	0	1	Percent correct
Training			
0	290	55	84.1%
1	64	117	646%
Overall Percent	67.3%	32.7%	77.4%
Testing			
0	133	22	85.8%
1	30	57	65.5%
Overall Percent	67.4%	32.6%	78.5%

Dependent Variable : outcome

<u>Decision Tree</u>

• A popular way of representing and visualizing prediction models and algorithms.

• It visually defines the rules for easy interpretation and comprehension.



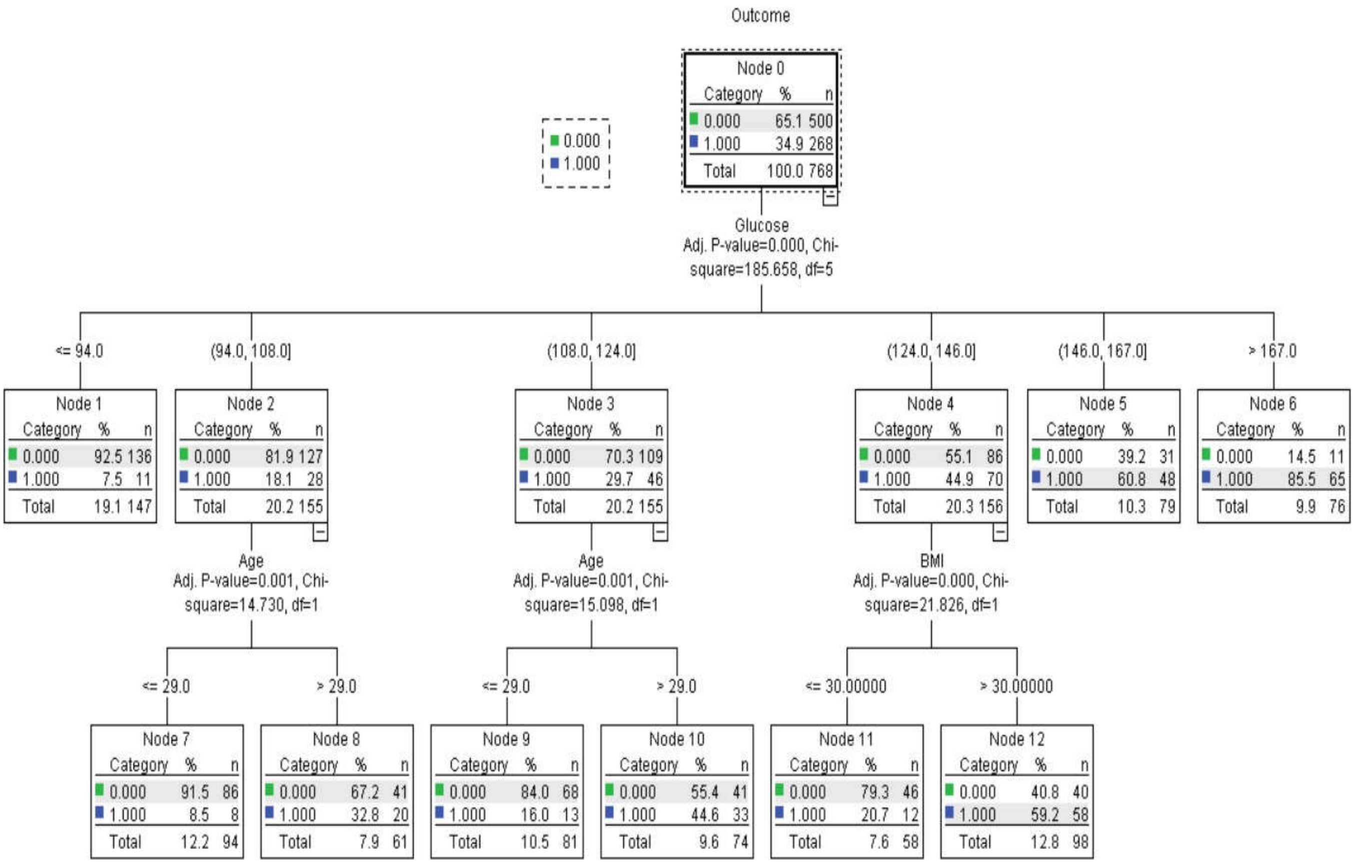
Tree pred	0	1
0	121	29
1	29	49

Tree prediction	0.745614

<u>Classification Tree :</u>

Observed	*Predicted*		
	0	*1*	*Percent Correct*
0	*418*	*82*	*83.6%*
1	*97*	*171*	*63.8%*
Overall Percentage	*67.1%*	*32.9%*	*76.7%*

Classification



K-Nearest Neighbor Analysis

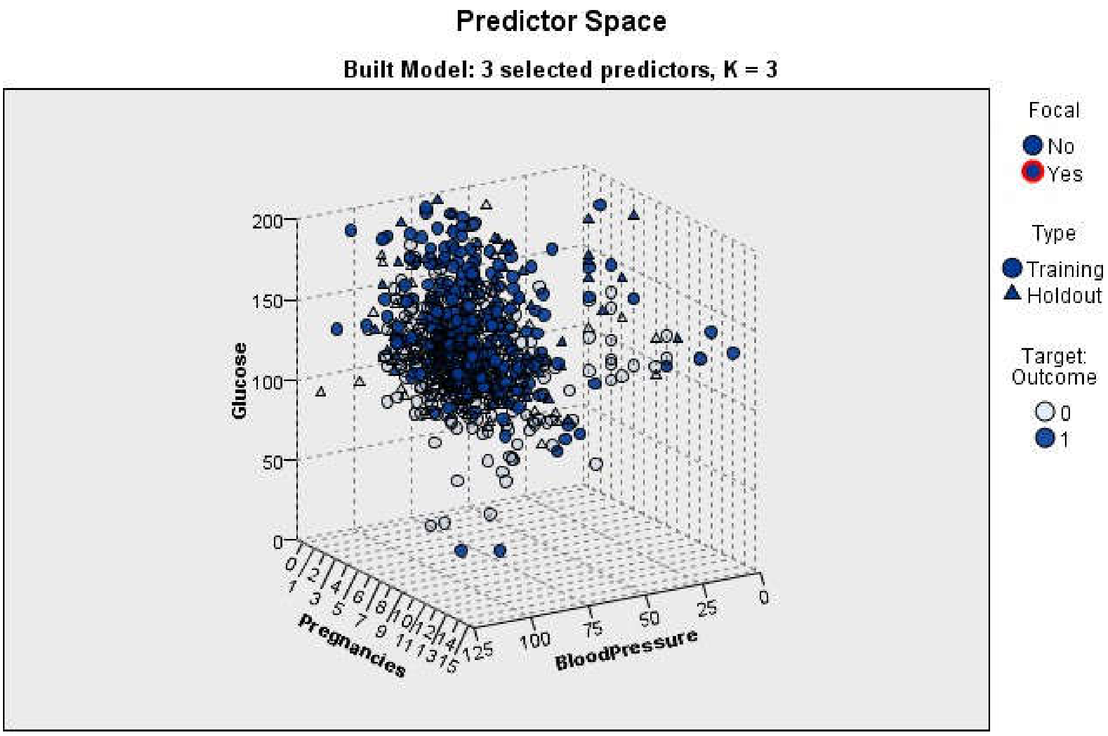
- K-Nearest Neighbor predicts the classification of a new sample point by identifying data points that are segregated into numerous groups.
- It is known as the LAZY algorithm.
- It uses a similarity metric to classify new data points.

Case Processing Summary

		N	Percent
Sample	Training	541	70.5%
	Holdout	226	29.5%
Valid		767	100.0%
Excluded		1	
Total		768	

K- Nearest Neighbor Analysis - Algorithm

1) Import data from .xls or.csv file.
2) Set K to zero (nearest neighbors).
3) Calculate the distance between query and current points for each sample in the training set. To an ordered collection, add distance and an example of index.
4) From small to large, sort ordered collection of distances and indexes.
5) Choose first K values from sorted collection.
6) Have labels for K values you selected.
7) Return mode of K labels.



Select points to use as focal records

This chart is a lower-dimensional projection of the predictor space, which contains a total of 8 predictors.

Results

NON-DAIBETES PATIENTS	500
DIABETES PATIENTS	268

Out of 768, we concluded that 500 patients do not have diabetes.

ALRITHM	ACCURACY
Logistic Regression	73%
Decision Tree	74%
Perceptron	67%
KNN	75%

COMPARATIVE ANALYSIS

Comparing the dataset with 'Diabetes Prediction Using Pipeline'.

Algorithms	Accuracy (Dataset used in this paper)	Accuracy (Dataset of Diabetes Prediction Using Pipeline)
Logistic Regression	73%	69%
Decision Tree	74%	62%
Perceptron	67%	54%
KNN	75%	68%

CONCLUSION

After analyzing the data by different machine learning techniques, it leads to the result that accuracy is obtained based on decision tree algorithm. In this various machine learning techniques are performed like logistic regression, decision tree, perceptron and KNN. This method helps the patients to find out the blood sugar content in body and predict whether the patient has diabetes or not at the earliest stage.

REFERENCES

- Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing,2015.
- Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.
- P. Suresh Kumar and S. Pranavi "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics", International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017.
- Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017.
- Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 2014, 15, 1929–1958.

- Sharma, R.; Kumar, S.; Maheshwari, R. Comparative Analysis of Classification Techniques in Data Mining Using Different Datasets. Int. J. Comput. Sci. Mobile Comput. 2015, 44, 125–134.
- Vidyasagar, M. Learning and Generalization: With Applications to Neural Networks; Springer Science & Business Media: London, UK, 2013.
- Mamuda, M.; Sathasivam, S. Predicting the survival of diabetes using neural network. In Proceedings of the AIP Conference Proceedings, Bydgoszcz, Poland, 9–11 May 2017; Volume 1870, pp. 40–46.
- Durairaj, M.; Kalaiselvi, G. Prediction of Diabetes using Back propagation Algorithm. Int. J. Innov. Technol. 2015, 1, 21–25.
- Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications, Volume 120 - Number 8,2015.
- K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.
- Humar Kahramanli and Novruz Allahverdi," Design of a Hybrid System for the Diabetes and Heart Disease", Expert Systems with Applications: An International Journal, Volume 35 Issue 1-2, July, 2008
- Ayush Anand and Divya Shakti," Prediction of Diabetes Based on Personal Lifestyle Indicators", 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015.
- B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.
- Somnath, R.; Suvojit, M.; Sanket, B.; Riyanka, K.; Priti, G.; Sayantan, M.; Subhas, B. Prediction of Diabetes Type-II Using a Two-Class Neural Network. In Proceedings of the 2017 International Conference on Computational Intelligence, Communications, and Business Analytics, Kolkata, India, 24–25 March 2017; pp. 65–71.