Diabetes prediction system using ml & dl techniques

Nandini Gupta, Shubhangi Malik, Hardik Chawla, Surinder Kaur *

Bharati Vidyapeeth's College of Engineering, GGSIPU, Delhi, INDIA
Emails: guptanandini12345@gmail.com; shubhangimalik28@gmail.com; hardikchawla111@gmail.com; kaur.surinder@bharatividyapeeth.edu

* Correspondence: kaur.surinder@bharatividyapeeth.edu

Abstract

Diabetes nowadays is a familiar and long-term disease. If a prediction is made early, better treatment can be provided. The preprocessing data approach is extremely useful in predicting the disease at an early stage. "Many tools are used in determining significant characteristics such as selection, Prediction, and association rule mining for diabetes. The principal component analysis method was used to select significant attributes. Our judgments denote a strong association of diabetes with body mass indicator (BMI) and glucose degree. The study implemented logistic regression, decision trees, and ANN techniques to process Pima Indian diabetes datasets and predict whether people at risk have diabetes. It was analyzed that random forest had the best accuracy of **80.52 %.** Out of 500 negative records & 268 positive records, our model correctly analyzed 403 records & 216 records, respectively.

Keywords: Body Mass Indicator; Artificial Neural Network; Logistic Regression; Random Forest

1. Introduction

A continuous sickness or condition or whose impact can be seen in the long run is termed a persistent condition or state. "These diseases affect the quality of life, thus deteriorating it. Diabetes is one of the diseases whose presence today is worldwide"[2]. One of the major reasons for death worldwide is the chronic disease of diabetes. "Diseases like these are also cost concerns. Governments and individuals spend a major portion of the budget on chronic diseases. The worldwide statistics for diabetes within the year 2013 revealed around 382 million individuals had this disease around the world. It was the fifth major reason for death in women and the eighth-most reason for death for both sexes in 2012. It has been noted that developed nations have a high probability of diabetes. In 2017, around 451 million grown-ups were treated with diabetes. It is estimated that in 2045, around 693 million patients with diabetes will exist around the globe, and a large portion of the populace will be undiscovered. Likewise, in 2017, 850 million USD was spent on patients with diabetes. Research on biological data is restricted, but with time, computational and statistical models are being used for analysis. A reasonable amount of knowledge is being gathered by healthcare organizations"[18]. "This can be made a reality when new models are developed to find out from the observed data using the data processing techniques. Data mining is the process of drawing out data and can also be utilized to create the choice-making process efficiently in the medical domain"[2]. Several information handling methods are used in disease prediction from biomedical information. "Diagnosis of diabetes is itself a challenge for quantitative research. A few boundaries like A1c, fructosamine, white blood corpuscle count, fibrinogen, and hematological indices were displayed as insufficient because of certain limitations. Various examinations tried to involve these boundaries for the determination of diabetes. Some of the treatments have been considered to boost A1c, including chronic ingestion of liquor, salicylates, and narcotics. Ingestion of vitamin C

DOI: https://doi.org/10.54216/FPA.010201
Received: January 02, 2020 Accepted: March 19, 2020

might raise A1c when assessed by electrophoresis, but levels might seem to lessen when assessed by chromatography. Most studies have suggested a better white blood corpuscle count, thanks to chronic inflammation during hypertension. A case history of diabetes has not been related to BMI and insulin. However, an increased BMI isn't always related to abdominal obesity"[5]. Only one boundary isn't powerful enough to precisely analyze diabetes and should be deceiving inside the dynamic interaction. "Thus, different parameters are to be mixed to efficiently predict diabetes at an early stage. A few existing strategies had not given powerful outcomes when various boundaries were utilized for predicting diabetes. In our review, diabetes is anticipated with the help of genuine traits and, in this way, the relationship of the contrasting credits. We examined the diagnosis of diabetes.

1.1 Diabetes categories

"Diabetes is a plague disease that happens whose major reason is a decrease of insulin within the body. Different types of diabetes are distinguished at diagnosis, so determining the type of diabetes depends on the disease's conditions. The old division was of two sorts of diabetes, that is, insulin-reliant and non-insulin reliant. The new grouping of diabetes was developed by America Diabetes.
Association: Type I diabetes, type II, gestational diabetes, and"[13] different sorts.

1.1.1. Type I diabetes

"Type I diabetes (insulin-dependent diabetes mellitus) may be a chronic disease that occurs when the pancreas releases a small amount of insulin (a hormone that's required for importing sugar). Some elements incorporating hereditary qualities and disease with certain infections can cause type I diabetes. Although type 1 diabetes usually occurs in childhood and adolescence, adults also are vulnerable to this disease"[17].

1.1.2. Type 2 diabetes

"Type 2 diabetes (adult diabetes or Non-insulin-dependent diabetes) is one of the common sorts of diabetes. It constitutes around 90 percent of the patients. Unlike type 1 diabetes, the body produces insulin in type 2 diabetes, but the insulin produced by the pancreas isn't enough, or the body cannot use insulin properly. When there's not enough insulin or the body does not use insulin, glucose (sugar) within the body cannot move to the body's cells and causes an accumulation of glucose within the body; therefore, the body would be in trouble and deficiencies. Unfortunately, there's no cure for this disease, but a healthy diet, exercise, and keeping fit can enhance it. If diet and exercise aren't enough, you would like medication or insulin treatment. Figure 1 is an analysis of diabetes"[10].

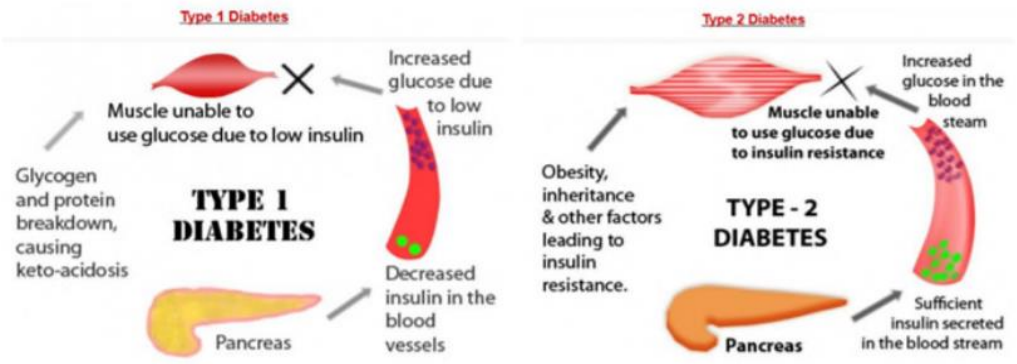

Figure 1: Analysis of Diabetes

DOI: https://doi.org/10.54216/FPA.010201
Received: January 02, 2020 Accepted: March 19, 2020

2. Related Work

Shetty et al. used KNN and "the Naïve Bayes technique for predicting diabetes. Their technique was implemented as a software program, where users provide input regarding patient records and find whether the patient is diabetic"[21] or not.

"Singh et al. applied different algorithms to datasets of different types. They used KNN, random forest, and Naïve Bayesian ML algorithms. The K-fold cross-validation technique was then used for evaluation. Ahmed utilized the patient's information and treatment plan to classify diabetes. Three algorithms applied were Naïve Bayes, logistic, and J48 algorithms"[18].

"Antony et al. utilized medical data for the Prediction of diabetes. Naïve Bayes, function-based multilayer perceptron (MLP), and decision tree-based random forest (RF) algorithms were applied after preprocessing of the data. A correlation-based feature selection method was employed to remove the extra features. A learning model then predicted whether the patient is diabetic or not. Using preprocessing techniques improved results when applying Naïve Bayes compared to other"[11] machine learning algorithms.

"Amina et al. compared different data mining techniques by using the PID dataset for the early Prediction of diabetes. Sellappan Palaniappan et al. proposed a heart disease prediction system by using various algorithms like Naïve Bayes, ANN"[12], and decision trees.

"Shadab Adam Pattekari and Asma Parveen [14] developed a web-based application for the Prediction of myocardial infarction using Naive Bayes. Anuja Kumari and R. Chitra used"[15] an SVM model to diagnose diabetes using a high-dimensional medical dataset.

Md. Kamrul Hasan et al. used a proposed ensemble model on the PIDD dataset to predict diabetes. It was concluded that the highest accuracy was achieved using the combination of boosting type classifiers (AdaBoost and XGBoost) when the proposed preprocessing (i.e., outlier rejection + filling missing values) is applied.

Aishwarya Mujumdar et al. implemented a diabetes prediction system using various ML algorithms like Gradient Boost, AdaBoost, Logistic Regression, KNN, GaussianNB, Perceptron, LDA, and SVC. Out of all these algorithms, the pipeline application gave AdaBoost classifier the best model with maximum accuracy.

Quan Zou et al. predicted diabetes mellitus using Random forest, Decision tree, Neural Network, and Decision tree algorithms. They used two datasets, namely the Luzhou dataset and Pima Indians Dataset. It was concluded that the Random forest gave the maximum accuracy, and the Pima Indians dataset gave the best performance. ML algorithms can predict diabetes efficiently provided that suitable attributes, classifiers & data mining methods are found properly.

Safial Islam Ayon et al. implemented a diabetes prediction system using deep neural networks based on several medical predictor variables. The highest accuracy was achieved for five-fold cross-validation.

Bala Manoj Kumar et al. implemented diabetes prediction using a Deep Neural Networks classifier. For feature attribute selection, the proposed model made use of feature importance. The best accuracy was achieved using DNN-FI compared with random forest & decision tree algorithms.

3. Procedure and Entities

3.1. Dataset

PIDD (Pima Indians Diabetes Dataset)

DOI: https://doi.org/10.54216/FPA.010201
Received: January 02, 2020 Accepted: March 19, 2020

The database is known as PIDD and is extracted from the National Institute of Diabetes and Digestive and Order Conditions, a. It aims to predict if a case suffers from diabetes-mellitus or not, valid for specific individual criteria of which the database consists. For this dataset, all cases of women are at least 21 years aged.
Pima Indian Diabetes (PID) data set has the following characteristics 9 = 8 1 (Class Identifier), 768 records describing womanish cases (in the case of 500 adverse events), and 268 positive cases. A detailed description of the features is given in **Table 1**.

Table 1: Dataset description and characteristics.

Sr.	Attribute Name	Attribute Description	Mean ± S.D
1	Pregnancies	Number of times a woman got pregnant	3.8 ± 3.3
2	Glucose(mg/dl)	Glucose concentration in oral glucose tolerance test for 120 min	120.8 ± 31.9
3	Blood Pressure(mmHg)	Diastolic Blood Pressure	69.1 ± 19.3
4	Skin Thickness (mm)	Fold Thickness of Skin	20.5 ± 15.9
5	Insulin (mu U/mL)	Serum Insulin for 2 h	79.7 ± 115.2
6	BMI (kg/m2)	Body Mass Index (weight/(height)^2)	31.9 ± 7.8
7	Diabetes Pedigree Function	Diabetes pedigree Function	0.4 ± 0.3
8	Age	Age (years)	33.2 ± 11.7
9	Outcome	Class variable (class value 1 for positive 0 for Negative for diabetes)	

Many obstacles have been placed in selecting these conditions on a large database.

3.2 Data Preparation

Its procedure is carried out in the following way:

3.2.1 Data Exploration

The process is begun by locating connections and correlations between the two factors (and the difference in the result) and visualizing the connections using a heatmap (see **Table 2**).

DOI: https://doi.org/10.54216/FPA.010201
Received: January 02, 2020 Accepted: March 19, 2020

Table 2: Product of feature (and outcome) relationship/correlations

	Pregnancies	**Glucose**	**blood pressure**	**skin thickness**	**Insulin**	**BMI**	**DPF**	**Age**	**Outcome**
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
Mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
Std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
Min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.0000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	23.000000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	32.000000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	99.000000	67.100000	2.420000	81.000000	1.000000

In the below heatmap, the bright colors show some correlations. An important correlation can be seen in the table along with the heatmap of glucose levels, age, BMI, and gestation rate with the outgrowth variability. Also, a relationship between dyads of factors, like age and gestation, or insulin and skin firmness, as shown in Fig 2.

DOI: https://doi.org/10.54216/FPA.010201
Received: January 02, 2020 Accepted: March 19, 2020

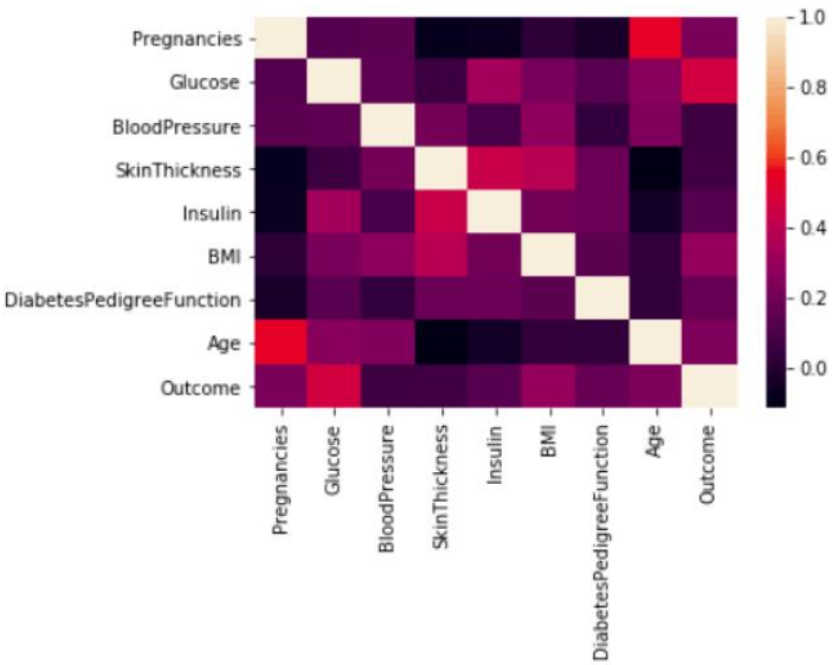

Figure 2: Heatmap of feature (and effect) correlation

3.2.2 Data Preprocessing

Sometimes, data in real life can be noisy or inconsistent, and it might also contain missing values. When such degraded quality data is used, the quality of results also degrades. Thus, data reprocessing becomes necessary to gain results of good quality. Drawing, incorporating, modifying, reducing, and separating data is used in pre-data processing. Thus, it's significant to make the data correlated to mining in terms of efficiency of time taken, production cost, and data standard.

3.2.3 Data Cleaning

Sanctification involves fulfilling missing quantities and reducing unwanted data. Data should contain excerpts to resolve inconsistencies. For this database, glucose, Blood Pressure, Skin Consistency, Insulin, and BMI have zero or null values (0). Therefore, every null attribute's value is replaced by the average value of that trait to remove inconsistencies. **Fig 3** and **Fig 4** show the outlier in the PIDD dataset and outlier junking, respectively.

DOI: https://doi.org/10.54216/FPA.010201
Received: January 02, 2020 Accepted: March 19, 2020

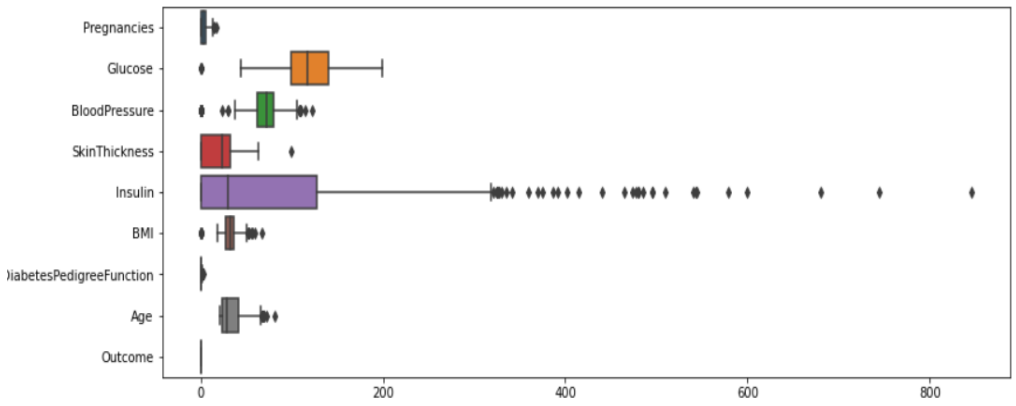

Figure 3: Outliers in the dataset

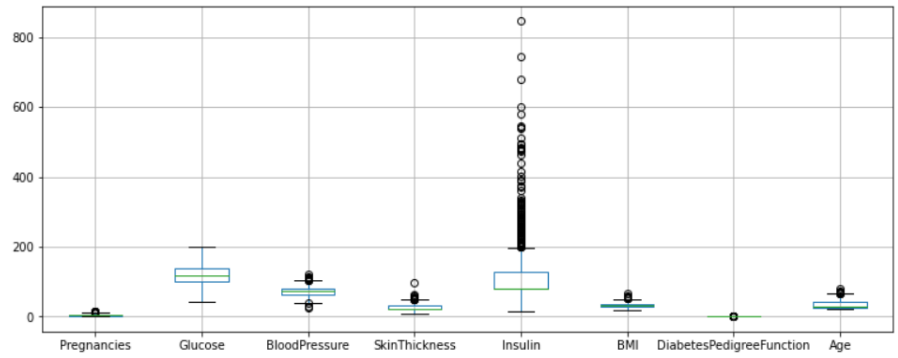

Figure 4: Outlier Junking

3.2.4. Data Reduction

Data reduction reduces the representation of datasets with much less volume but still gives the same (or nearly the same) results. Dimensionality and size are reduced to decrease the number of attributes in a database. The crucial element figuring system used, the prize's the crucial values from the entire database. Glucose, BMI, diastolic blood pressure, and age were the most important factors in the dataset after visualization.

3.2.5. Data Metamorphosis

Data revision includes smoothness, familiarity, and integration of data [18]. To smooth the data, a combination system was used. The age factor has helped divide the five orders, as shown in Table 3.

Table 3: Binning of age.

Age (years)	Age Bins

DOI: https://doi.org/10.54216/FPA.010201
Received: January 02, 2020 Accepted: March 19, 2020

≤ 30	Youngest
31-40	Younger
41-50	Middle-aged
51-60	Older
≥ 60	Oldest

Blood glucose uptake in non-diabetic cases differs from diabetic cases. Glucose values are divided into 5 orders (19), as shown in Table 4.

Table 4: Binning of glucose.

Glucose	**Glucose Bins**
≤ 60	Very Low
61-80	Low
81-140	Normal
141-180	Early Diabetes
≥ 181	Diabetes

A strong correlation was planted between non-diabetic and diabetic cases regarding their pressure situations. Blood pressure is categorized into five distinct orders, as shown in Table 5.

DOI: https://doi.org/10.54216/FPA.010201
Received: January 02, 2020 Accepted: March 19, 2020

Table 5: Types of diastolic blood pressure.

Blood Pressure	**Diastolic Blood Pressure Bins**
≤ 61	Very Low
61-75	Low
75-90	Normal
91-100	High
≥ 100	Hypertension

A relationship between the body mass index and diabetes is found. The original study concludes that BMI is the most dangerous factor in determining type 2 diabetes. BMI values are divided into 5 classes, as shown in Table 6.

Table 6: Binning of BMI.

BMI	**BMI Bins**
≤ 19	Starvation
19-24	Normal
25-30	Overweight
31-40	Obese
≥ 40	Very Obese

DOI: https://doi.org/10.54216/FPA.010201
Received: January 02, 2020 Accepted: March 19, 2020

3.2.6. Dataset Splitting and Normalization

This process is begun by unyoking the data into one training set and one test set. The database contains records of 767 cases in aggregate. Only 614 (80%) records will be used to train the model. The remaining records will be used to test and estimate the model.

4. Experimental Result

Sample 1

Table 7: Patient Data

	pregnancies	glucose	bp	skin thickness	insulin	BMI	dpf	age
0	3	120	70	20	79	20	0.4700	33

Visualized Patient Report

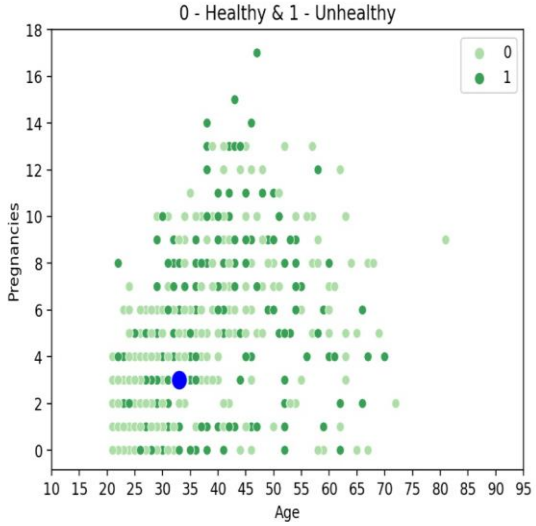

Figure 5: Pregnancy Count Graph (Others vs. Yours)

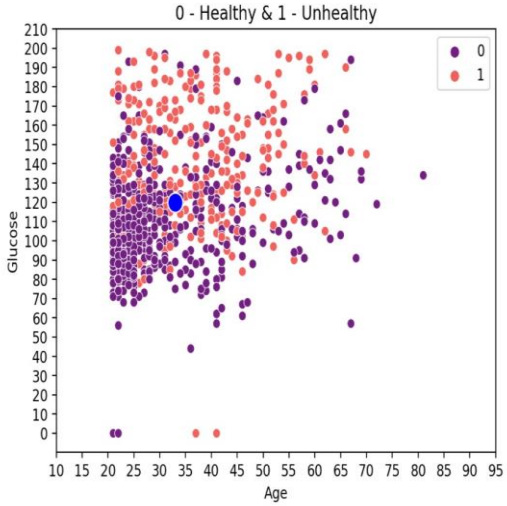

Figure 6: Glucose Value Graph (Others vs. Yours)

DOI: https://doi.org/10.54216/FPA.010201
Received: January 02, 2020 Accepted: March 19, 2020

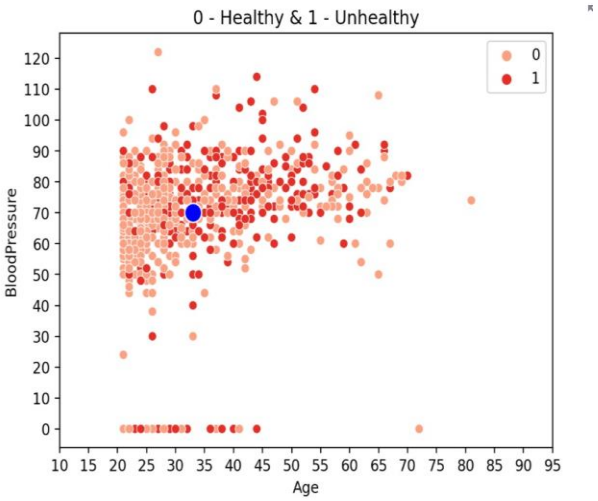

Figure 7: Blood Pressure Graph (Others vs. Yours)

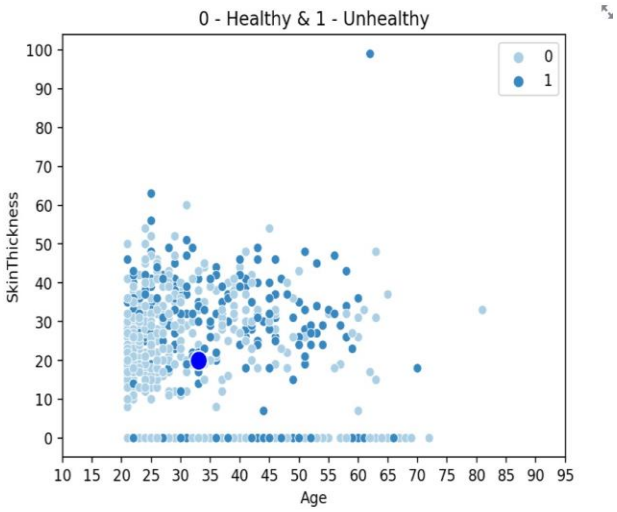

Figure 8: Skin Thickness Value Graph (Others vs. Yours)

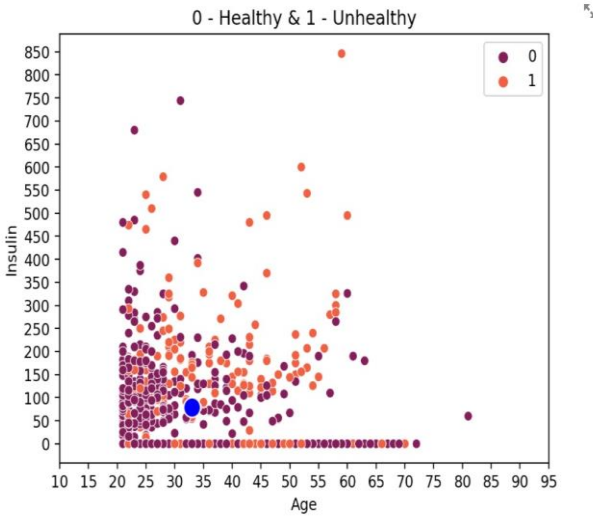

Figure 9: Insulin Value Graph (Others vs. Yours)

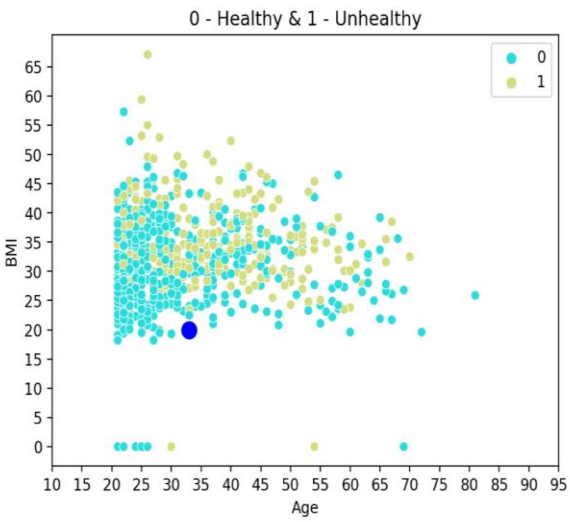

Figure 10: BMI Value Graph (Others vs. Yours)

DOI: https://doi.org/10.54216/FPA.010201
Received: January 02, 2020 Accepted: March 19, 2020

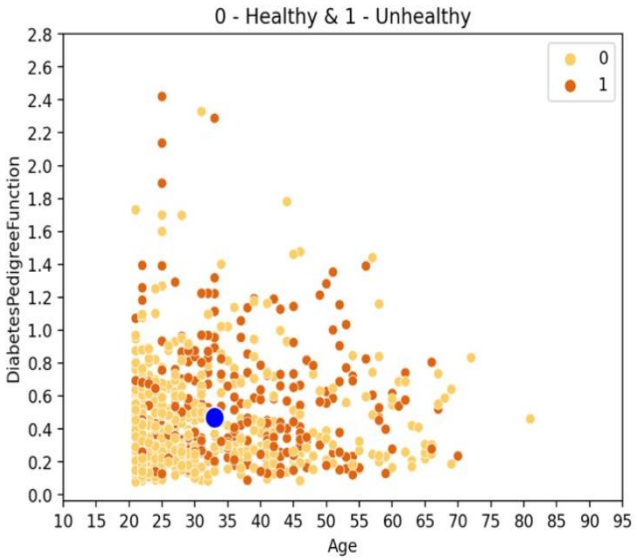

Figure 11: DPF Value Graph (Others vs. Yours)

Table 8: Report of sample 1

RESULT	**ACCURACY**
Non-Diabetic	80.5614

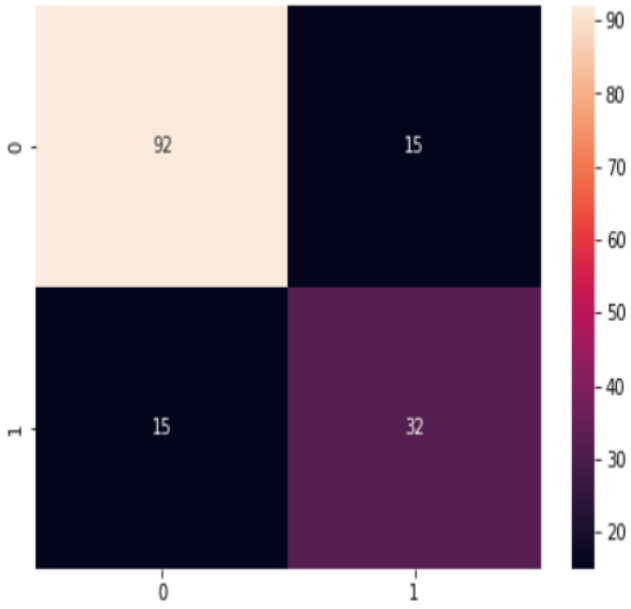

Figure 12: Confusion matrix for Random Forest Classifier

True Positive (TP) = 92 , False Positive (FP) = 15, False Negative (FN) = 15, True Negative (TN) = 32

DOI: https://doi.org/10.54216/FPA.010201
Received: January 02, 2020 Accepted: March 19, 2020

. In order to find the exact accuracy the following measures have been calculated as depicted in **Table 9**:

Accuracy Rate = (TP + TN) / (TP + TN + FN + FP)

Error Rate = (FN + FP) / (TP + TN + FN + FP)

Sensitivity = TP / (TP + FN)

Precision = TP / (TP + FP)

F-Measure = 2 * (Precision * Sensitivity) / (Precision + Sensitivity)

From the obtained confusion matrices following measure given in the equation can be calculated. These matrices gave True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP). The TN is higher than TP in both datasets because both datasets have non-diabetic cases more than diabetic ones. Thus, all the methods are giving good results.

Table 9: Values for different measures for Random Forest Classifier

	RANDOM FOREST
Accuracy Rate	80.52 %
Error Rate	19.40 %
Sensitivity	0.86
Precision	0.86
F-measure	0.86

DOI: https://doi.org/10.54216/FPA.010201
Received: January 02, 2020 Accepted: March 19, 2020

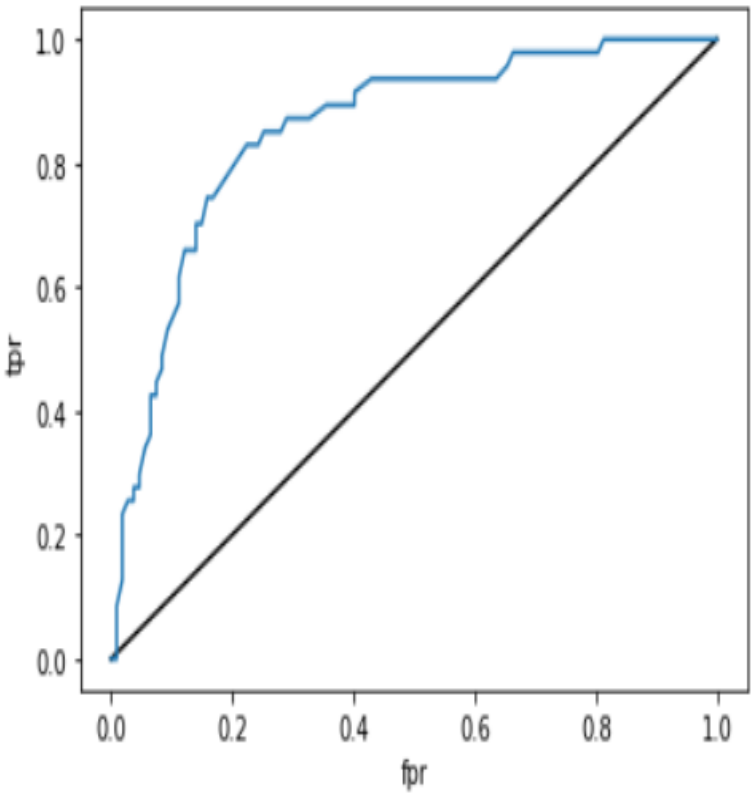

Figure 13: ROC curve for Random Forest

The Area Under the Receiver Operating Characteristic Curve (**ROC AUC**) score is 85.35 %.

The graph is plotted to depict the true vs. predicted value of the obtained result, as shown in Fig 14, and the same has been analyzed along with the error percentage in **Table 10**.

DOI: https://doi.org/10.54216/FPA.010201
Received: January 02, 2020 Accepted: March 19, 2020

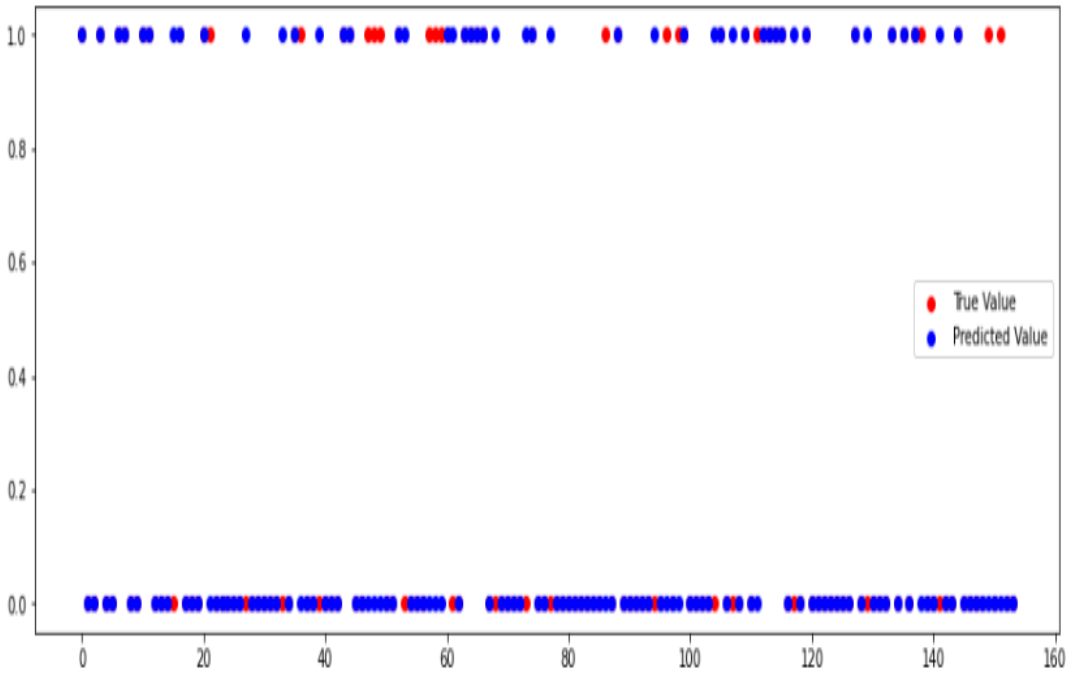

Figure 14: True Value vs. Predicted Value

Table 10: Result Analysis

	Actual Value	**Experimental Value**	**Error Percentage**
Diabetic	268	216	19.40 %
Non-Diabetic	500	403	19.40 %

5. Conclusion

Machine learning and deep learning techniques are profound learning strategies that are significant in health diagnosis. The capacity to anticipate diabetes at an early stage is essential for the at-risk individuals' proper treatment system. The study implemented logistic regression, decision trees, and ANN techniques to process Pima Indian diabetes datasets and predict whether people at risk have diabetes. The dataset contains 9 = 8 + 1 (class attribute) credits, 768 datasets representing female patients (of which 500 negative cases (65.1%) and 268 positive cases (34.9%)).

The study implemented logistic regression, decision trees, and ANN techniques to process Pima Indian diabetes datasets and predict whether people at risk have diabetes. It was analyzed that random forest had the best accuracy of **80.52 %.** Out of 500 negative records & 268 positive records, our model correctly analyzed 403 records & 216 records, respectively.

DOI: https://doi.org/10.54216/FPA.010201
Received: January 02, 2020 Accepted: March 19, 2020

The impediment of this review is that an organized dataset has been chosen yet, later, unstructured information will likewise be thought of. These strategies will be applied to other clinical areas for expectation. Various factors, including actual idleness, family background of diabetes, and smoking propensity, are likewise intended to be considered in the foreseeable future for the analysis of diabetes.

Funding: "This research received no external funding."
Conflicts of Interest: "The authors declare no conflict of interest."

6. Future Scope

The proposed methodology combines Neural networks and a Logistic regression model.
Our proposed method will consist of leading artificial neural network input and a logistic regression statistical model. Given previous research, we have found that the error of artificial neural networks combined with logistic regression is far more reduced. Thus a better accuracy was analyzed in a combinational model rather than a simple method of artificial neural network or a simple method of logistic regression.

The model first uses the regression coefficients to determine the value of each variable. Next, consider the output potential of each rule (the input of the neural network), the effect on the output that triggers the input of the proposed model, and each result to accurately predict the potential.

References

[1] Temurtas, H., Yumusak, N., Temurtas, F., "A relative study on diabetes complain opinion using neural networks", Expert Syst, Vol. 36, pp. 8610 – 15, 2009.

[2] Chavey, A., Kroon, M., Bailbé, D., "programming of beta-cell diseases and the intergenerational threat of type 2 diabetes Diabetes", Motherly Diabetes, Vol. 40, No. 5, pp. 323-30, 2014.

[3] Manzella,D., Grella,R., Abbatecola,AM., Paolisso,G.,"Repaglinide Administration Improves Brachial Reactivity in Type 2 Diabetic Cases", Diabetes Care, Vol. 28, pp. 366 –71, 2005.

[4] Mohamed, E.I., Linde, rR., Perriello,G., Di Daniele, N., Pöppl,S.J., De Lorenzo, A.," Predicting type 2 diabetes using an electronic nose - grounded artificial neural network analysis", Diabetes nutrition & metabolismVol. 15, No. 4, pp.222-215, 2002.

[5] Volley, J.C., WilliamsG., (Eds.)., Textbook of diabetes, Blackwell Science, Oxford, 2003.

[6] Ahmadi KGuideline & book review. The internal (endocrine and lung). Ahmadi Cultural Institute, 2009.

[7] Morteza, Afsaneh, et al., "Inconsistency in albuminuria predictors in type 2 diabetes a comparison between neural network and tentative logistic retrogression", Translational

[8] Marateb, HamidR., et al."A cold-blooded intelligent system for diagnosing microalbuminuria in type 2", pp. 34-42, 2014.

[9] Torkestani, Javad, Akbari., and Elham, GhanaatPisheh., "A literacy automata- grounded blood glucose regulation medium in type 2 diabetes", Control Engineering Practice, Vol. 26, pp.151-159, 2014.

[10] Metz, CE., Wang, P-L., Kronman, HB., A new approach for testing the significance of differences between ROC angles measured from identified data. In DeconinckF. (editor) Information processing in medical imaging. The Hague Nijhoff, pp. 432-445, 1984.

DOI: https://doi.org/10.54216/FPA.010201
Received: January 02, 2020 Accepted: March 19, 2020

[11] Nielsen, D., Krych, L., Burchard, K., "Beyond Genetics Influence of salutary factors and gut microbiota on type 1 diabetes", FEBS Lett, Vol. 588, pp. 4234 – 43, 2014.

[12] Pei, E., Li, J., Lu, C., Xu, J., Tang, T., Ye, M., et al," Goods of lipids and lipoproteins on diabetic bottom in people with type 2 diabetes mellitus a meta-analysis", J Diabetes Complications,Vol. 28,pp. 559 – 64, 2014.

[13] Livingstone, D., Totowa, NJ, Artificial Neural Networks Styles, and Operation. 1st ed Totowa, NJ Humana Press; 2008.

[14] Dunne, RA., Wiley, J., Inc, S., "A Statistical Approach to Neural Networks for Pattern Recognition", New Jersey John Wiley & Sons Inc; 2007.

[15] Zini,G., d'Onofrio,G.,"Neural network in hematopoietic malice", Clin Chim Acta, Vol. 333,No. 2,pp.195-201, 2003.

[16] Ruczinski,I., Kooperberg,C., etal., Logic Regresion. Journal of Computational and Graphical statistic, Vol. 12, No. 3, pp.475-511, 2003.

[17] Danesh-Pour, MS., Mehrabi, Y., Hedayati, M., Azizi, F., "Multivariable check of factors identified with the metabolic pattern using factor analysis (Persian)", Iranian Journal of Endocrinology and Metabolism, Vol. 30, pp.139-46, 2006.

[18] Talha Mahboob Alam, Muhammad Atif Iqbal, Yasir Ali, Abdul Wahab et al. "A model for early prediction of diabetes", Informatics in Medicine Unlocked, 2019.

DOI: https://doi.org/10.54216/FPA.010201
Received: January 02, 2020 Accepted: March 19, 2020