Female Diabetic Prediction in India Using Different Learning Algorithms

Payal Bose¹, Samir Kumar Bandyopadhyay^{1,*}, Amiya Bhaumik², Sandeep Poddar³

¹Faculty of Engineering, Lincoln University College, Malaysia

²Faculty of Business Administration, Lincoln University College, Malaysia

³Faculty of Research & Innovation, Lincoln University College, Malaysia

Received November 8, 2021; Revised December 10, 2021; Accepted December 26, 2021

Cite This Paper in the following Citation Styles

(a): [1] Payal Bose, Samir Kumar Bandyopadhyay, Amiya Bhaumik, Sandeep Poddar, "Female Diabetic Prediction in India Using Different Learning Algorithms," *Universal Journal of Public Health*, Vol. 9, No. 6, pp. 460 - 471, 2021. DOI: 10.13189/ujph.2021.090614.

(b): Payal Bose, Samir Kumar Bandyopadhyay, Amiya Bhaumik, Sandeep Poddar (2021). *Female Diabetic Prediction in India Using Different Learning Algorithms*. *Universal Journal of Public Health*, 9(6), 460 - 471. DOI: 10.13189/ujph.2021.090614.

Copyright©2021 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract Diabetes, also known as Diabetic Mellitus, is a metabolic disease that affects the body's natural blood glucose levels. It is a non-contagious illness with numerous serious health risks. The said illness is rapidly growing in India. It is a chronic disorder that happens when the human body unable to create enough insulin hormone to keep blood sugar levels under control. Several characteristics that cause diabetes were investigated in this study, and multiple machine learning techniques were used to predict whether or not an unknown substance had diabetes. PIMA diabetes detection for female patients was employed for this purpose. For prediction, six distinct classification models were applied. This research presented a comprehensive performance assessment of the multiple factors in the PIMA dataset. Based on all factors of the PIMA dataset, a full discussion was made to demonstrate how diabetes is affected. Finally, in order to forecast the best automated diabetic prediction model, a thorough analysis of many classification approaches was undertaken. It was discovered that the Support Vector Machine (SVM) model delivers the best prediction result, with a reliability of 83.5 percent. Interestingly, Random Forest (RF) Classifier produced the second-best prediction result, with a reliability of 82.76 percent. The study's findings demonstrate that machine learning models produce efficient solutions. The accuracy of the two best machine learning models is 82-83 percent, which can be used for subsequent improvement of the autonomous forecasting tool. The accuracy of these techniques can be

improved further by integrating diverse variables for prediction and classification.

Keywords Diabetics, Diabetic Mellitus, Diabetic Prediction, Female Diabetic Patients, Machine Learning(ML), PIMA Diabetic Dataset, Support Vector Machine, Random Forest

1. Introduction

Diabetes Mellitus, also known as simple diabetes, is a metabolic condition characterized by insufficient management of blood sugar levels. The hormone insulin aids in the transport of glucose from the bloodstream to the tissue and the absorption of sugar as fuels for subsequent use in the internal organs [1, 2]. Diabetics develop when the percentage of this glucose level climbs or drops from its usual level. Diabetes, or hyperglycemia, can affect the nerves, eyes, kidneys, skin, and many other organs [3-5]. This is a highly prevalent and incurable condition. However, modifying one's lifestyle and eating habits in conjunction with medication can assist to manage blood sugar levels and lead a healthy life.

Prediction of diabetics in their early stages is quite important due to the adverse effects. Machine Learning approaches are now being used in the healthcare sector to more correctly forecast diabetes. Machine Learning allows

a machine to become intelligent by learning from previous experiences or input attributes and predicting the next and its categorization. Researchers have presented numerous diabetes prediction algorithms in order to effectively predict PIMA Indians Female diabetics. The researchers conducted a comparative investigation of various machine learning techniques in order to assess their performance [6], calculate the different performance factors from the receiver operating characteristic curves [7]. Other researchers combine the PIMA diabetics dataset with data from the Bangladesh. They also compared several machine learning techniques for detecting diabetes individuals [8]. Deep learning approach also performs well in predicting diabetics [9]. The deep learning and machine learning algorithms achieved accuracy ranging from 90 to 98 percent. In one investigation, the deep learning strategy attained the highest accuracy, 98.04 percent [10, 11].

This study reveals the behavior of many crucial parameters for diabetic individuals, as well as the interaction between the primary essential components, and is entirely focused on detecting diabetics in female patients. The most efficient classifier for better prediction is revealed by comparing six different machine learning methods.

2. Materials and Methods

2.1. Dataset

In this study, a dataset of Pima Indian women from the Kaggle repository [12] is utilized to predict a patient's diabetic proclivity. The dataset can be framed as a set of attributes containing many criteria for detecting diabetic proclivity. In this dataset, the attribute 'outcome' is used as the prediction's target class. This dataset is entirely dedicated to female diabetic patient prediction in India.

2.2. Dataset Details

This dataset [13] includes information about 768 female patients. There are 500 non-diabetic female patients and 268 diabetic female patients among them. The diagnosis

outcome, along with the other parameters, is saved in the dataset as binary numbers for each individual. The PIMA dataset's summary and diagnosis details are included in Table 1 and Figure 1. The dataset has 8 predictive variables and one responsive variable. Table 1 provides extensive information on how many patients participated in the investigation, how many disease criteria were evaluated for investigation, how many of the participants were diagnosed as diabetes and how many were not, and lastly what the ratio of diabetes and non-diabetes was. Figure 1 depicts a visual representation of the total amount of diabetics and non-diabetics patient. Table 2 lists the nine experimental characteristics or features that were utilized to predict diabetics, as well as their minimum and maximum values.

Table 1. Experimental data Summary

PIMA Diabetic Detection Dataset	
Observations	Total Number
Total No of Patients in the dataset	768
Total number of Features present the dataset	9
Sum of Diabetic Patients	268
Sum of Non-diabetic patients	500
The ration of Diabetic and Non-Diabetic	0.35 : 0.65

Figure 1. Distribution of Diabetic Patients

Table 2. Database variables Details

No.	Feature name	Feature Details	Min Value	Max value
1	Pregnancies	Frequency of pregnancies	0.0	17.0
2	Glucose	The tolerance test of oral glucose, plasma glucose concentration was measured after 2 hours.	0.0	199.0
3	Blood Pressure	The systolic blood pressure (mm Hg)	0.0	122.0
4	Skin Thickness	Thickness of the triceps skin fold (mm)	0.0	99.0
5	Insulin	Serum insulin after 2 hours (mu U/ml)	0.0	846.0
6	BMI (Body mass index)	Body weight (in kg/(height in m) ²)	0.0	67.1
7	Diabetes Pedigree Function	Diabetes pedigree function	0.078	2.42
8	Age	Calculated in years	21	81
9	Outcome	Class label (0 for No Diabetic or 1 for Diabetic)	-	-

Figure 2. The Distribution of Box Plot

2.2.1. Dataset Analysis

There are eight predicting factors in the dataset. Each predictive variable is crucial in the detection of diabetics. Figures 2 depicted the distribution of box plot and 3A to 3H depict the box plot of various predictive variables with responsive variables, respectively.

The boxplot plot is frequently used to depict data dispersion. It reflects if the data in the dataset are balanced, how neatly ordered the data is, and how biased the data are. It generates a five statistical summary: the minimal, first percentile (Q1), median, interquartile (Q3), and utmost, the minimal, first percentile (Q1), median, interquartile (Q3), and utmost. The 'Minimum' indicates the minimum values of the data, 'Median' indicates the median value of the data, 'Quartile Q1' indicates the middle values of minimum and median value of the data (25 Percentile data), similarly 'Quartile Q2' indicates the middle value of median and maximum values (75 Percentile data) of data, finally, 'Maximum' indicates the maximum values of the data.

Figure 4A to 4H depicts the barplots of each predictive variable with responsive variables, respectively. The barplot diagram describes the correlation between a numerical and a categorical variable. Each categorical

variable is represented by a bar. The quantitative value is represented by the size of the bar.

In this investigation, the barplots for each explanatory factor explain how diabetes and non-diabetes fluctuate with a particular range of outcomes. The bars in each plot are organized by a specific range of the associated characteristic. The length of each bar is measured as a percentage of its related component.

To explain the statistical relations between variables, a correlation matrix is utilised. Correlation is a measure of how closely one parameter is dependent upon the other. A correlation matrix is a table that displays the correlation coefficients between parameters. Each entry in the table or matrix represents the relationship between two variables. It is used to consolidate the incoming data in a more complex manner.

Figure 5 depicts the connection of the PIMA diabetes dataset's eight predictive variables and one responsive (Outcome) variable of female patients in India. The correlation matrix plot demonstrates that the dataset contains four significant predictor variables. They are as follows: 1) the number of pregnancies related with age, 2) the glucose level related with insulin, 3) the BMI (Body

Mass Index) related with the thickness of the skin, and 4) the BMI (Body Mass Index) related with the level of blood pressure. Figure 5A to 5D depicts the association between all the highest correlated variables.

Every figure from 6A through 6D demonstrates how the four dependent components in the experimental dataset are interrelated.

(A)

(B)

(C)

(D)

Figure 3. Boxplot of Predictive Variables for Female Diabetic Patients, A) Number of Pregnancies vs Diabetes status dataset visualization, B) Glucose Level vs Diabetes status dataset visualization, C) Blood Pressure vs Diabetes status dataset visualization, D) Skin Thickness vs Diabetes status dataset visualization, E) Insulin Level vs Diabetes status dataset visualization, F) BMI vs Diabetes status dataset visualization, G) Diabetic Pedigree Function vs Diabetes status dataset visualization, H) Age Group vs Diabetes status dataset visualization

Figure 4. Barplot of distinct predicting factors with response variable (Outcome), A) Number of Pregnancies vs Diabetes status visualization, B) Glucose Level vs Diabetes status visualization, C) Blood Pressure vs Diabetes status visualization, D) Skin Thickness vs Diabetes status visualization, E) Insulin Level vs Diabetes status visualization, F) BMI vs Diabetes status visualization, G) Diabetic Pedigree Function. vs Diabetes status visualization, H) Age Group vs Diabetes status visualization

Figure 5. Correlation Plot of PIMA Diabetics Dataset

Figure 6. The graphical illustration of the relationship between the Significant correlated indicators for Diabetic Prediction factors, A) Relation Between Age and No. of Pregnancies, B) Relation between Insulin Level with Glucose Level, C) Relation between BMI with Skin Thickness, D) Relation Between BMI with Blood Pressure Level

2.3. Classification Models

A classifier that uses training data to translate input variables to target classes. The purpose of using a classifier throughout this investigation is to predict whether or not a patient has a diabetes predisposition. The classification model predicts the test dataset using the trained model.

2.3.1. Decision Tree (DT)

Decision Tree (DT) [14, 15] is known as a machine learning classifier. For prediction, this model was utilised to generate a tree-like structure. It acquires knowledge from classification expertise. Each outcome variable is represented as a DT leaf node, while DT non-leaf nodes are used as test class decision nodes. The results of such tests are determined by one of the decision node's branches. This model generates categorization results by growing the tree branches from the parent node to the leaf node. This model is used to forecast the target depending on certain criterion.

2.3.2. Random Forest (RF)

This classification [15] system is made up of numerous decision trees. Bagging and feature unpredictability are used to construct each one-of-a-kind tree. This method generates a statistically independent forest of trees, and the whole forecast is far more precise than any individual tree's forecast. It predicts using ensemble learning, which is a technique that combines numerous weak classifiers to handle complicated problems.

2.3.3. Gradient Boosting (GB)

Gradient boosting [16] is a sort of boosting in machine learning algorithms. This algorithm combines a number of poor machine learning techniques to produce a powerful predictive model. It is predicated on the idea that the best next model, when combined with past models, reduces

total prediction error.

2.3.4. Support Vector Machine (SVM)

It is a supervised classification model [17]. When the training set contains a collection of labelled data, this approach can generalise between two separate classes. The primary job of SVM is to look for a hyperplane that can differentiate between the two classes.

2.3.5. K-Nearest Neighbor (KNN)

K-Nearest Neighbour Classifiers [18] are referred to as slow learners. It recognises objects in the features space based on the closest distance between training and testing samples. The classifier considers k objects before deciding on a class based on the closest object. The biggest difficulty with this classification strategy is determining the appropriate value of k.

2.3.6. Logistic Regression (LR)

This regression model [19, 20] is a classification approach derived from statistics by machine learning. This statistical method is used to examine a dataset that has one or many independent factors that influence the outcome. The purpose of logistic regression is to find the optimal model for describing the relationship between the dependent and independent variables.

2.4. Method

Figure 7 depicts the complete process of predicting diabetic or non-diabetic patients from an unknown sample using the PIMA Indian female diabetes dataset. In this investigation, six different categorization methods were used. The algorithm for diabetic prediction is provided in the algorithm section.

Figure 7. Workflow of the Principal Method

2.4.1. Input Dataset

The experiment dataset was obtained from the Kaggle archive [13].

2.4.2. Dataset Pre-Processing

The input dataset contains some empty fields. To build a clean dataset, null values must be removed. To fill in the omitted or null data, the mean value of the associated feature is computed.

2.4.3. Split the Dataset

The second step is to prepare the dataset for categorization. To categorize the data, the dataset must be divided into two parts: training and testing. In this experiment, 80 percent of the data in the input dataset is chosen for training, while the remaining 20% has been used for testing as unknown samples.

2.4.4. Train the Data for Classification

Following that, six distinct machine learning algorithms are applied one by one on the training dataset to train the data models.

2.4.5. Testing the Data and Make the Prediction

Afterwards, use the trained data models to evaluate the test data. For each classification model, the performance of the test data was estimated. Next, find the highest and second highest prediction models precision.

2.4.6. Algorithm

Step 1: Collect the input dataset from Kaggle repository.

Step 2: Examine the entire dataset to discover if there is any missing data. If there is any, the null data should be replaced with the field's average scores. The dataset is now capable for the prediction model after the missing data has been filled in.

Step 3: Divide the dataset in 2:1 ratio. The vast bulk of the data was utilised for training, with the remaining being used for testing.

Step 4: Train the dataset using a classifier.

Step 5: Make a prediction on test data using the trained model.

Step 6: Calculate the accuracy for both training and testing data.

3. Results and Discussions

3.1. Results

Table 3 displays the outcomes of all classification models. The testing accuracy, sensitivity, specificity, and kappa values of all classifiers were used to evaluate their performance.

Table 3. Performance Evaluation of All Classification techniques

Methods	Accuracy	Specificity	Sensitivity	Cohen – Kappa Value
Decision Tree	0.7470588	0.6207	0.7699	0.3842
Random Forest	0.8276471	0.6824	0.8253	0.4118
Gradient Boosting	0.7917447	0.7708	0.7279	0.5410
Support Vector Machine	0.8352941	0.7500	0.8689	0.6036
K-Nearest Neighbor	0.8000000	0.7292	0.8279	0.5302
Logistic Regression	0.7294118	0.8333	0.8197	0.6128

3.2. Discussions

Diabetes is a prevalent [21, 22] and incurable disease. It is a protracted health condition that affects the human body's ability to produce energy from food. There are various factors that can be intended to identify the diabetic status of a human body. The input dataset for this investigation contains the eight essential key parameters for diagnosing diabetics, which are listed in table 2.

Figure 4 show that each predictive variable suggests different values for predicting diabetics in test samples.

- If a female patient has more than six pregnancies, her chances of developing diabetes increase.
- If the glucose concentration level is greater than 140, diabetics are more likely to be detected.
- As blood pressure rises in conjunction with the other variables, the likelihood of becoming diabetes rises as well.
- There is a link between skin thickness and insulin level. When insulin levels fall, the likelihood of becoming diabetes rises. Similarly, as insulin levels fall, skin thickness increases.
- As BMI and Diabetic Pedigree levels rise, so does the likelihood of diabetic detection.
- At a certain age, the majority of females are diagnosed as diabetes. According to the graph, the majority of those above the age of 60 fall into this category.

The comprehensive performance study of six alternative categorization models is shown in Table 3. Depending on the testing results, each model measures the accuracy and other quality metrics.

According to the performance of each model, the support vector machine classifier has the highest accuracy of 83.5%, followed by the random forest classifier with the second highest accuracy of 82.8%.

This strategy is developed to create a comparison model. This model aids in visualising the relationship between the numerous predictive and responsive factors. It also aids in the identification of the optimal classification model for effectively and efficiently predicting diabetic patients.

4. Conclusions

Diabetes mellitus is a chronic, noncommunicable disease with no complete cure. It can grow slowly in the human body and increase the risk of various linked diseases over time. Several risk factors, such as a poor lifestyle, obesity, poor nutrition, and so on, can be the most serious concern for a diabetic. Preventive actions and increased knowledge can assist to lower the risk of this disease. In a developing country like India, most individuals lack basic information of a healthy lifestyle and are unaware that they are already suffering from diabetes or are on the verge of getting the ailment. Anticipating diabetics at an early stage allows a patient to take precautions and prevent the illness from intensifying. Whenever it comes to own households, women in India have traditionally played an essential role. They emphasize the health of their family over their own. As a result, due to the various factors like poor nutrition, obesity, poor health, etc., the number of female diabetes cases has increased compared to male diabetes cases.

In this study, an automated diabetic forecasting system was developed that can diagnose this disease efficiently and precisely. This technology can detect the condition at an early stage, allowing people with prediabetic signs to better manage their health.

This study included a thorough examination of eight predictive variables and one responsive variable. To show the important variables and their dependencies in the dataset, a correlation matrix was calculated. The prediction is accomplished by a number of machine learning methods. Finally, the algorithms were subjected to a thorough comparison. According to the comparative evaluation, the support vector machine classifier achieved the highest accuracy of up to 83.5 percent. Similarly, the random forest classifier achieves the second highest accuracy of up to 82.8 percent. In this case, only one dataset is used, and the sample sizes in that dataset are quite minimal. As a result, future research will incorporate the use of a large dataset as well as the use of deep learning techniques to increase performance.

REFERENCES

- [1] Diabetes A, diabetes W. What is diabetes, Online available from: <https://www.idf.org/aboutdiabetes/what-is-diabetes.html> (accessed 16 September 2021)
- [2] What is Diabetes? Centers for Disease Control and Prevention. Online available from: <https://www.cdc.gov/diabetes/basics/diabetes.html> (accessed 16 September 2021)
- [3] Diabetes: Symptoms, Causes, Treatment, Prevention, and More. Online available from: <https://www.healthline.com/health/diabetes> (accessed 16 September 2021)
- [4] diabetes A, diabetes T. Type 2 diabetes, Online available from: <https://www.idf.org/aboutdiabetes/type-2-diabetes.html> (accessed 16 September 2021)
- [5] Diabetes, Who.int. Online available from: <https://www.who.int/news-room/fact-sheets/detail/diabetes> (accessed 16 September 2021)
- [6] M, Sharma P. "Survey on Prediction and Analysis of Diabetic Data using Machine Learning Techniques". *International Journal of Computer Sciences and Engineering*. Vol. 6, no. 6, pp.1032-1038, 2018.
- [7] Kumar S, Kumar S, Kumar K, Abhisekh PA. "Prediction of Diabetes in Females of Pima Indian Heritage: A Complete Supervised Learning Approach". Vol. 12, no. 10, pp. 3074–84, 2021.
- [8] Pranto, B., Mehnaz, S. M., Mahid, E. B., Sadman, I. M., Rahman, A., & Momen, S. "Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh". *Information (Switzerland)*, vol. 11, no. 8, 2020. <https://doi.org/10.3390/INFO11080374>
- [9] Swapna G., Vinayakumar R., and Soman K. P., "Diabetes detection using deep learning algorithms," *ICT Express*, vol. 4, no. 4, pp. 243–246, 2018, doi: 10.1016/j.icte.2018.10.005.
- [10] Ahmad H. F., Mukhtar H., Alaqail H., Seliaman M., and Alhumam A. "Investigating health-related features and their impact on the prediction of diabetes using machine learning," *Appl. Sci.*, vol. 11, no. 3, pp. 1–18, 2021, doi: 10.3390/app11031173.
- [11] Naz H. and Ahuja S. "Deep learning approach for diabetes prediction using PIMA Indian dataset," *J. Diabetes Metab. Disord.*, vol. 19, no. 1, pp. 391–403, 2020, doi: 10.1007/s40200-020-00520-5.
- [12] Pima Indians Diabetes Database. Kaggle.com. Online available from: <https://www.kaggle.com/uciml/pima-indians-diabetes-database> (accessed 6 September 2021)
- [13] Smith J. W., Everhart J. E., Dickson W. C., Knowler W. C., and Johannes R. S., "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," *Proc. - Annu. Symp. Comput. Appl. Med. Care*, pp. 261–265, 1988.
- [14] Patel H. H. and Prajapati P., "Study and Analysis of Decision Tree Based Classification Algorithms," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 10, pp. 74–78, Oct. 2018, doi: 10.26438/ijcse/v6i10.7478.
- [15] Ali J., Khan R., Ahmad N., and Maqsood I. "Random Forests and Decision Trees," *Int. J. Comput. Sci. Issues*, vol. 9, no. 5, pp. 272–278, 2012.
- [16] Natekin A. and Knoll A. "Gradient boosting machines, a tutorial," *Front. Neurobot.*, vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.
- [17] Dai H. "Research on SVM improved algorithm for large data classification," 2018 IEEE 3rd Int. Conf. Big Data

- Anal. ICBDA 2018, no. 1, pp. 181–185, 2018, doi: 10.1109/ICBDA.2018.8367673.
- [18] Wang L. “Research and Implementation of Machine Learning Classifier Based on KNN,” IOP Conf. Ser. Mater. Sci. Eng., vol. 677, no. 5, 2019, doi: 10.1088/1757-899X/677/5/052038.
- [19] Zewude B. T. and Ashine K. M. “Binary Logistic Regression Analysis in Assessment and Identifying Factors That Influence Students” Academic Achievement: The Case of College of Natural and Computational,” J. Educ. Pract., vol. 7, no. 25, pp. 1–6, 2016, [Online]. Available: <https://eric.ed.gov/?id=EJ1115855>.
- [20] Fernandes A. A. T., Figueiredo Filho D. B., da Rocha E. C., and Nascimento W. da S., “Read this paper if you want to learn logistic regression,” Rev. Sociol. e Polít. vol. 28, no. 74, pp. 1/1-19/19, 2020, doi: 10.1590/1678-987320287406en.
- [21] Rusiman, M. S., Md Nor, S. N., Suparman, Mohd Razali, S. N. A.,” Robust Method in Multiple Linear Regression Model on Diabetes Patients”, Mathematics and Statistics; vol. 8, no. 2A, pp. 36-39, 2020, doi: 10.13189/ms.2020.081306
- [22] Mallampati, T.,” Risk Factors, Knowledge, Attitudes and Practices of Type 2 Diabetics Attending Industry, Campbellville Health Centers and West Demerara Hospital, Guyana”, Universal Journal of Public Health, vol. 9, no. 3, pp. 131-139, 2021, doi: 10.13189/ujph.2021.090304.