Predictive Study and Classification of Diabetes Using Machine Learning Techniques

Krishan Kumar and Sanjay Patidar

Abstract Diabetes mellitus is a common but deadly disease in humans. It is caused by having excessive sugar levels existing for a long time. It causes around 30–40 lakh deaths worldwide each year. Technology plays a consequential role in the medical industry to assess diabetes prediction. In this research, we trained four machine learning techniques so as to make predictions on whether a person is diabetic or not. The Pima Indian Diabetes dataset is used, which consists of 768 samples, and each sample contains 8 attributes and one target class attribute. Data preprocessing techniques are used to update the raw dataset into a dataset which is suitable for the machine learning models. KNN, Logistic Regression, Support Vector Machine, and Artificial Neural Networks are the techniques used for prediction of diabetes in this research. As a result, K-Nearest Neighbor performed the best, with an accuracy of 76.17%.

Keywords Diabetes prediction · Machine learning · Neural network · Classification · Data preprocessing

1 Introduction

In today's world, technology surrounds us in every possible way, usage of modern day technologies to get some ease in our work, some of the examples we can see are driverless cars, maps GPS navigation, voice controlled devices and many others. All these modern technologies use the data of the real world to train their models and finally test them with the help of data again and again to generate better accuracy. The degree of ease and accuracy of the technology is directly proportional to its usage. Machine learning is also a similar technique in which various algorithms are used to train the model using some part of data and then test the model with some other part of data to generate the outcome. This modern day technique can be used in the prediction of some disease with the help of associated symptoms. Algorithms can

K. Kumar · S. Patidar (✉)
Department of Software Engineering, Delhi Technological University, New Delhi, India
e-mail: sanjaypatidar@gmail.com

S. Smys et al. (eds.), *Inventive Computation and Information Technologies*, Lecture Notes in Networks and Systems 563, https://doi.org/10.1007/978-981-19-7402-1_32

train the model in which symptoms values will be used as inputs and generation of outcome will be tested to check the best accuracy and some other different measures, which will help the patient and the medical field to deal with the patient's condition in the earlier stage [1].

Diabetes is a major cause of death, metabolic disorders in humans as well as leads to commercial and productivity loss throughout the world due to lower levels of efficiency of man power. It is a metabolic disorder, characterized by high blood sugar levels which is caused by low insulin production in the pancreas. It increases the risk of long-term complications. It increases the chances of heart disease and about 75% people having this disease die due to coronary artery disease.

In this research, different machine learning algorithms are compared in order to predict risk of someone having diabetes. Classification algorithms are used to classify the target outcomes (1 for diabetic or 0 for non-diabetic) independently. Our study is structured in the following order—Sect. 2 contains literature review. The next, Sect. 3 explains the procedural approach, the machine learning techniques used and the model evaluation. Section 4 discusses about the final result obtained from the research. The last, Sect. 5 contains the conclusion and future work.

2 Literature Review

Diabetes diagnosis and treatment has been a crucial topic in medical research from a very long period of time. With the help of Machine learning, a really good progress has been made in the process of predicting diabetes in people. This prediction is made by the help of machine learning models, which are trained on the dataset consisting of medical information of patients, along with the information whether patients are diabetic or not. After the training phase the model is evaluated by passing the testing data to the model, to check how efficiently the model is working. Kahramanli and Allahverdi [2] used amalgamation of Artificial Neural Networks and fuzzy logics to make a model with good accuracy to predict diabetes. Kumar Dwivedi [3] compared five machine learning algorithms to predict diabetes. The algorithms used were artificial neural networks, classification tree, KNN, SVM, and logistic regression. The author in [4] used two classification algorithms, deep neural networks and artificial neural networks. And also used principal component analysis. Using deep neural networks they achieved better accuracy of 82.67%. Khan and Mohamudally [5] used k-means clustering, neural networks and C4.5 decision tree algorithm to predict diabetes in patients. Bayesian network, Artificial neural network, SVM, Decision tree, and KNN were used to predict diabetes, by Heydari et al. [6]. Temurtas et al. [7] made a model which was trained by Levenberg–Marquardt (LM) algorithm, and the model was combined with multilayer neural network structure.

Rajesh and Sangeetha used classification technique. They used C4.5 decision tree algorithm to find hidden patterns from the dataset for classifying efficiently [8]. Butwall and Kumar proposed a model using Random Forest Classifier to forecast diabetes behavior [9]. Ashiquzzaman et al. [10] proposed a prediction framework for

the diabetes mellitus using deep learning approach where the overfitting is diminished by using the dropout method. Patil proposed Hybrid Prediction Model which includes Simple K-means clustering algorithm, followed by application of classification algorithm to the result obtained from clustering algorithm. In order to build classifiers C4.5 decision tree algorithm is used [11].

Patients can have several symptoms and some of the symptoms and factors are included in the data set like Age, Insulin level, Glucose level, Diabetes Pedigree Function, Blood Pressure level, Skin Thickness, and BMI. Prediction of the outcome from data has been done using various traditional machine learning techniques and artificial neural networks. In order to apply these algorithms, we need to preprocess the data which includes cleaning of the data [12]. Then proposed algorithms are applied and their performances are validated. These prerequisite actions are necessary so that optimal levels of accuracy, precision, and recall can be obtained.

In this paper, classification algorithms are used on the diabetic patient's data set to predict the outcome of diabetes presence in patients and we achieved a success rate on the test set of 76%. Moreover, we were able to obtain this much accuracy with traditional machine learning approaches, by adding some data preprocessing techniques.

3 Procedural Approach and Methodology

The procedural approach is as follows, as shown in Fig. 1.

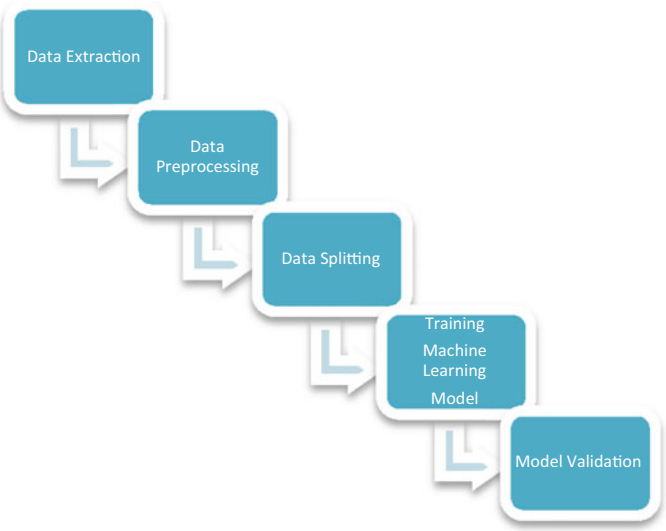

Fig. 1 Proposed architecture

Table 1 List of attributes present in the dataset and their data type

Attribute name	Data type
Pregnancies	Integer
Glucose	Integer
BloodPressure	Integer
SkinThickness	Integer
Insulin	Integer
DiabetesPedigreeFunction	Float
BMI	Float
Age	Integer
Outcome	Integer

3.1 Data Extraction

The data used is the PIMA Indian diabetes dataset. Aim of which is to predict whether or not a patient is diabetic, on the basis of several attributes included in the dataset. Different criteria were used on the selection of these values from the database.

The dataset contains the medical details of 768 different patients, and these medical details were used for classifications. These medical details were stored in 768 rows and 9 columns. Nine columns consisting of 8 attributes and one class column 'Outcome' (diabetic or non-diabetic), as shown in Table 1.

3.2 Data Preprocessing

After the data is collected, it cannot be directly used for the study, therefore it needs to be processed and cleaned to gather suitable information from the raw data useful for the study. The raw data is expected to have many inconsistencies, anomalies, out of bound values, missing values or a format not suitable for our model. Hence, the data needs to be processed in order to use it for our study. Moreover, vast data in present day business, science, industry, and academia scenarios needs complex mechanisms to analyze it. It includes data cleaning, transformation of data; and irregular data reduction tasks, used to reduce the convolution of data, determine and eliminate irrelevant and boisterous elements from the data through feature selection or discretization processes.

Elimination of Null Values

The data was checked for any null values across all features and secondly in individual feature columns, Elimination of Not a number (NaN) values: We replaced the null values of 'glucose' and 'blood pressure' by the mean of the respective attributes, and replaced the null values of the 'skin thickness', 'insulin,' and 'BMI' with the median of the respected attributes.

Table 2 Null values count

Attribute name	Null values	
	Before elimination	After elimination
Pregnancies	0	0
Glucose	5	0
BloodPressure	35	0
SkinThickness	227	0
Insulin	374	0
DiabetesPedigreeFunction	0	0
BMI	11	0
Age	0	0
Outcome	0	0

Table 2 shows the count of number of null values present in the data set before elimination of null values, and also after the elimination of null values.

Evaluation of Class Distribution

The data was checked to be distributed evenly between the target variable outcomes.

3.3 Data Splitting

Entire data was divided into training and testing data. Two-thirds of main dataset was the training data and the rest one-third was used for testing. The training data is the dataset which is given to the model in the beginning for model's training purpose, which is, to learn from the dataset about the input attributes and the output attribute. The testing data is the dataset which given to the model to model after the training of the model is complete, to check of efficiently the model is working.

So, here the testing data is 1/3rd of the whole data set, and the training dataset is 2/3rd of the whole dataset, as shown in Table 3.

Table 3 Verification of data splitting

Dataset	Percentage of data w.r.t original dataset
Dataset before splitting	768 (100%)
Training dataset	512 (66.6%)
Testing dataset	33 (33.3%)

3.4 Machine Learning Methodologies

K-Nearest Neighbor

Aim of the algorithm is to find the class for the given input. It is a supervised machine learning algorithm. K is the number of neighbors with which we will compare the given input. The input will be assigned to the class whose maximum number of data will be near to the input itself. And the calculation is done with the help of Euclidean distance KNN formula, where x and y are the values of the independent attributes of the neighbor and the new point, and m is the number of independent attributes:

$$\text{dist}(A, B) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \tag{1}$$

Support Vector Machine

It is a labeled training data algorithm that creates a hyperplane that separates the points according to their classes. This hyperplane can be seen in 2D space as a plane splitting line into two pieces, one for each segment. Linear SVM is a technique for generating a classifier that can distinguish between labeled datasets. Given two sorts of points, it tries to maximize the margin geometrically. The letter 'Z' is utilized to solve the problem of maximum margin and the reparability limitation.

Logistic Regression

It is an algorithm for calculating binary outcomes like zero and one (in our case diabetic or non-diabetic). A linear regression is ineffective for categorizing a binary variable because it predicts continuous values that are beyond the range.

Artificial Neural Networks

The output layer, hidden layer, and input layer are the three layers of an ANN, which are made up of interconnected neurons. The hidden layer has multi-layered structure. The nodes in successive layers are all linked together. Every neuron has an activation function, which is a transformation function that is applied to the node before it is sent to the next layer as input. The result of a node is computed as in Fig. 2.

3.5 Model Validation

The methodologies were performed on the jupyter notebook. The data was analyzed using data visualization techniques and conformed via performance evaluation metrics such as accuracy, precision, recall, $F1$-score. Cross-validation method was used for evaluation. In k-fold cross-validation, we broke the data into k distinct sets

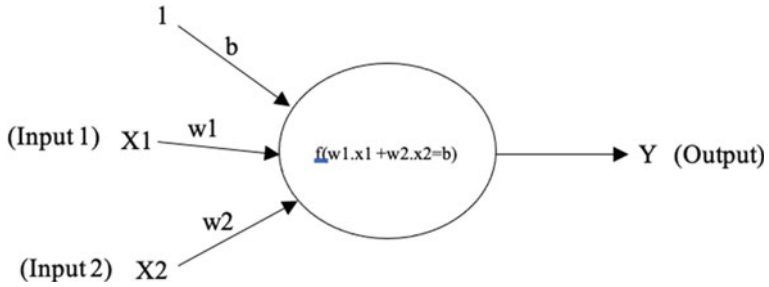

Fig. 2 Neuron in ANN

which are exclusive in nature and have equal size, with one set used for training purpose and other for testing.

Evaluation Metrics

The study is evaluated/validated via confusion matrix using metrics such as accuracy, precision, recall, and $F1$-score.

True Positives (TP): Total predicted diabetic cases, validated as diabetic.
True Negative (TN): Total predicted non-diabetic cases, validated as non-diabetic.
False Positives (FP): Total predicted diabetic cases, validated as non-diabetic.
False Positives (FN): Total predicted non-diabetic cases, validated as diabetic.

$$\text{Presision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad F1\text{-score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

In this study for model validation confusion matrix have been used. In Fig. 3, confusion matrix of KNN model is shown in Table 4. Figure 4 shows the learning curve of KNN model which represents the training score and cross-validation of the KNN model. Table 5 shows the accuracy, precision, recall, and $F1$-score of the KNN model. Similarly, all the other algorithms were validated.

4 Results

In this research, we have performed diabetes prediction on PIMA Indian dataset, to predict diabetes a person is diabetic or not. First data is preprocessed by eliminating all the Not a number (NaN) values, by replacing them by the mean or the median of the respective attribute. Then the prediction was made by using four different machine learning algorithms KNN, SVM, logistic regression, and artificial neural networks. And among all the four algorithms, KNN showed the best accuracy of

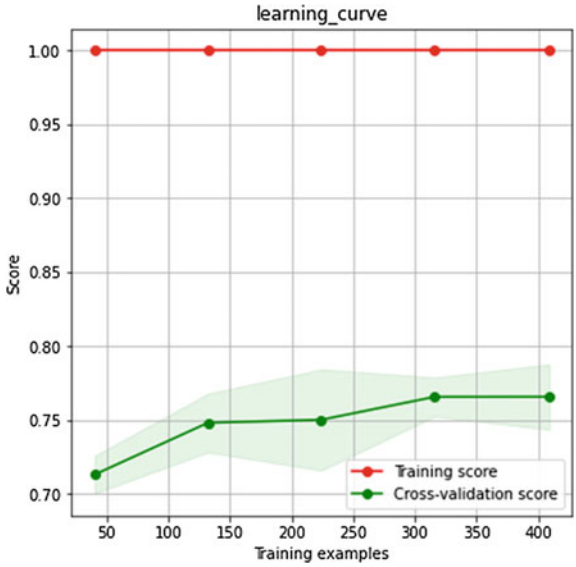

Fig. 3 Learning curve of KNN

Table 4 Confusion matrix

Output		Predicted values	
		Diabetic	Non-diabetic
Actual values	Diabetic	TP	FN
	Non-diabetic	FP	TN

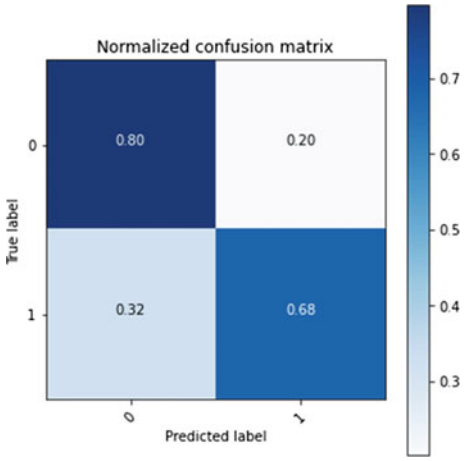

Fig. 4 Confusion matrix of KNN

Table 5 Accuracy, precision, recall, and $F1$-score of the KNN model

Accuracy	0.76
Precision	0.67
Recall	0.59
$F1$-score	0.63

Table 6 Performance measures of all the four algorithms

	Accuracy	Precision	Recall	$F1$-score
KNN	0.76	0.67	0.59	0.63
SVM	0.75	0.69	0.53	0.60
Logistic regression	0.73	0.63	0.52	0.57
Neural network	0.73	0.62	0.61	0.62

76%. Table 5 shows all the values of accuracy, precision, recall, and $F1$-score of all the four machine learning algorithms are shown in Table 6.

5 Conclusion and Future Work

In this study we ought to resolve the complications occurred during diagnosis of diabetes disease. The study put forwards an light on different machine learning algorithm such as the SVM, KNN, logistic regression, and ANN for predicting whether a patient is diabetic or not. It was concluded that out of all KNN performed best with an accuracy of 76%, hence it is a better option for classifying complex data.

In future we will try to come up with much better mechanisms and a much larger data set in order to increase the accuracy to help medical practitioners to treat patients and overcome this deadly disease.

References

1. Kalyankar GD, Poojara SR, Dharwadkar NV (2017) Predictive analysis of diabetic patient data using machine learning and Hadoop. In: International conference on I-SMAC. 978-1-5090-3243-3
2. Kahramanli H, Allahverdi N (2008) Design of a hybrid system for the diabetes and heart disease. Expert Syst Appl Int J 35(1–2)
3. Kumar Dwivedi A (2017) Analysis of computational intelligence techniques for diabetes mellitus prediction. Neural Comput Appl 13(3):1–9
4. Vijayashree J, Jayashree J (2017) An expert system for the diagnosis of diabetic patients using deep neural networks and recursive feature elimination. Int J Civ Eng Technol 8:633–641
5. Khan DM, Mohamudally N (2011) An integration of K-means and decision tree (ID3) towards a more efficient data mining algorithm. J Comput 3(12)

6. Heydari M, Teimouri M, Heshmati Z, Alavinia SM (2015) Comparison of various classification algorithms in the diagnosis of type diabetes in Iran. Int J Diabetes Dev Ctries 1–7
7. Temurtas H, Yumusak N, Temurtas F (2009) A comparative study on diabetes disease diagnosis using neural networks. Expert Syst Appl 36(4):8610–8615. https://doi.org/10.1016/j.eswa.2008.10.032
8. Rajesh K, Sangeetha V (2012) Application of data mining methods and techniques for diabetes diagnosis. Int J Eng Innov Technol (IJEIT) 2(3)
9. Butwall M, Kumar S (2015) A data mining approach for the diagnosis of diabetes mellitus using random forest classifier. Int J Comput Appl 120(8)
10. Ashiquzzaman A, Tushar AK, Islam M, Kim J-M et al (2017) Reduction of overfitting in diabetes prediction using deep learning neural network. arXiv preprint arXiv:1707.08386
11. Patil BM, Joshi RC, Toshniwal D (2010) Association rule for classification of type-2 diabetic patients. In: ICMLC'10 proceedings of the 2010 second international conference on machine learning and computing, 09–11 Feb 2010
12. Fatima M, Pasha M (2017) Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl 09(01):1–16. https://doi.org/10.4236/jilsa.2017.91001