Prediction of Diabetes Mellitus Using Machine Learning

G. Arunalatha

Assistant Professor, Department of Computer Science and Engineering, Perunthalaivar Kamarajar Institute of Engineering and Technology (PKIET), Karaikal. Puducherry

ABSTRACT

Diabetes is considered one of the most deadly chronic diseases caused by high blood sugar. There are several reasons for the increase in the number of diabetics, such as obesity, autoimmune reaction, poor diet, lifestyle changes, dietary habits,, etc. In the long run, diabetics can develop many complications. such as heart disease, kidney disease, nerve damage, diabetic retinopathy, etc. But this risk can be reduced if it is anticipated early. Machine learning is a technique used to deliberately train computers. This study attempts to predict diabetes using three different machine learning algorithms: SVM, MLP and decision tree.

Keywords: Diabetes, SVM, MLP, Decision tree

INTRODUCTION

Diabetes is a chronic disease characterized by high blood sugar. It occurs when the body does not produce enough insulin or is unable to use the insulin it does produce effectively. Insulin is a hormone produced by the pancreas that helps glucose enter cells for energy. Without adequate insulin, glucose absorption is inhibited, leading to elevated blood sugar levels that can damage multiple organs over time. Although diabetes cannot be cured, it can be managed with a careful diet, physical activity, medication and regular screening for complications. If left untreated, diabetes can lead to serious complications such as cardiovascular disease, diabetic ketoacidosis, chronic kidney disease and leg ulcers..

There are three main types of diabetes:1. Type 1 Diabetes: An autoimmune condition where the immune system attacks and destroys the insulin-producing beta cells in the pancreas. It typically develops in childhood or adolescence but can occur at any age. People with type 1 diabetes need to take insulin daily. 2. Type 2 Diabetes: The most common form of diabetes, often associated with obesity and a sedentary lifestyle. It occurs when the body becomes resistant to insulin or when the pancreas fails to produce enough insulin. It usually develops in adults, but increasing numbers of children and adolescents are being diagnosed with it. Management includes lifestyle changes, oral medications, and sometimes insulin. 3. Gestational Diabetes: This type develops during pregnancy and usually goes away after childbirth. However, it increases the risk of developing type 2 diabetes later in life. It requires careful monitoring and management to protect the health of both mother and baby.

Diabetes can lead to serious health complications if not managed properly, including heart disease, stroke, kidney disease, nerve damage, and vision problems. Management involves monitoring blood sugar levels, maintaining a healthy diet, regular physical activity, and medication or insulin therapy as needed. Regular check-ups and education are essential for effective diabetes management and prevention of complications.

Generally, for a normal human being, glucose levels range from 70 to 99 milligrams per deciliter. A person is considered diabetic only if the fasting glucose level is found to be more than 126 mg/dL. In the medical practice, a person having a glucose concentration of 100 to 125 mg/dL is considered as pre-diabetic [4]. Such a person is prone to the development of type 2 diabetes. Over the years, it has been found that people with the following health characteristics face a greater risk against diabetes:

- A Body Mass Index value greater than 25
- Members of the family suffering from diabetes
- People with HDL cholesterol concentration in the body less than 40 mg/dL
- Prolonged hypertension having gestational diabetes
- People who have suffered from polycystic ovary disorder in the past

- People belonging to ethnic groups like African American, or Native American, or Latin American, or Asianpacific aged over 45 years
- Having an inactive lifestyle

LITERATURE SURVEY

A case study of Indian pregnant women with diabetes[1]. The information is taken from the PIMA database of the UCI archive and used. Each theme is characterized by eight attributes. It was confirmed in 768 patients. A general MLP classifier is used for the attributes and the experiment is learned on the R studio platform.

Kopitar et al. [2] used the concept of artificial intelligence to improve the accuracy of disease prognosis. It consists of three steps: preprocessing, feature selection and feature classification. Methods such as HSA (Harmonic Search Algorithm), GA (Genetic Algorithm) and PSO (Particle Swarm Optimization) and k-means clustering are used for feature selection. KNN is used for classification procedures. Evaluation measures used to predict accuracy include sensitivity, specificity, recall, and precision. Accuracy is 91.65%.

Bavkar et al. [3] developed a pipeline-based model that uses deep learning (DL) techniques to predict diabetes. This includes data augmentation using a variable autoencoder (VAE), feature augmentation using a sparse encoder (SAE), and a convolutional neural network for classification. The dataset used is the Pima dataset taken from the UCI repository. The obtained accuracy is 92.31% using a CNN classifier and training with SAE to boost features compared to a well-balanced dataset.

Machine learning techniques[4] are used with the pima Indian diabetes dataset to develop trends and identify patterns associated with risk factors using the R data processing tool. To classify patients as diabetic and non-diabetic, we developed and analyzed five different prediction models using the R data manipulation. To do this, we used supervised machine learning algorithms, ie. linear kernel support vector machine (SVM-linear), radial basis function (RBF) kernel support vector machine, k-nearest neighbor (k-NN), artificial neural network (ANN), and multifactorial. dimensionality reduction (MDR).

PROPOSED WORK

Detecting diabetes using machine learning involves developing models that can analyze patient data and predict whether or not an individual has diabetes. Here is a high-level overview of the process:

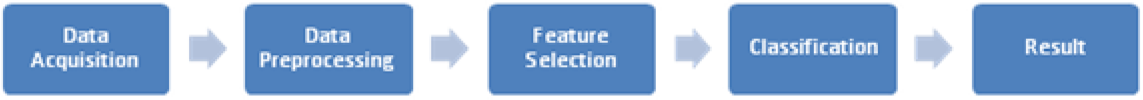

Fig:1 Proposed Work

Data Collection:
In this study used the PIMA India Diabetes dataset, a publicly available dataset collected and compiled by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) is utilized. The database contains information on 768 patients, including 268 with diabetes and 500 without diabetes. Datasets consist of several medical predictor variables and one target variable, Eight physiological parameters were recorded for each patient, namely pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, diabetes family history, smoking, alcohol, hours of sleep and age. These parameters were used to predict the occurrence of diabetes in each individual.

Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
Blood Pressure: Diastolic blood pressure (mm Hg)
Skin Thickness: Triceps skin fold thickness (mm)
Insulin: 2-Hour serum insulin (mu U/ml)
BMI: Body mass index (weight in kg/(height in m)^2)
Diabetes Pedigree Function: Diabetes pedigree function
Age: Age (years)
Outcome: Class variable (0 or 1)

PIMA Indian dataset is analyzed for normality using a whisker plot. This is a statistical tool often used in explanatory data analysis. Tt can be observed that the values of some features are skewed which is an indication of a violation of normality assumptions. Considering that the PIMA Indian dataset does not satisfy normality assumptions, we approached data preprocessing differently. As an initial step to the preprocessing, the dataset is relabeled to include the prediabetes class. This is because the existing diabetes research is limited to only the prediction of normal and diabetes classes. However, if the research is targeted towards diagnosing diabetes, then there is a need for the prediabetes class. As a result, the dataset is relabeled based on the levels of glucose to conform to medical charts provided online1 on clinical practices in the diagnosis of diabetes. Then, SC and PR methods are employed for feature importance selection, and missing value imputation, respectively

Data Preprocessing:
Handle missing values: Techniques include imputation, removing rows with missing data, or using algorithms that handle missing values. Feature scaling: Normalize or standardize features to ensure they contribute equally to the model. - Encode categorical variables: Convert categorical variables into a numerical format.

This phase of model handles inconsistent data in order to get more accurate and precise results. This dataset contains missing values. So we imputed missing values for few selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age because these attributes cannot have values zero. Then we scale the dataset to normalize all values.

Feature Selection:
Feature selection is the process of selecting a subset of significant features or variables from a larger feature set to improve model performance. This is an important step in machine learning and data analysis because it reduces the number of inputs and focuses on the most important ones that can provide meaningful insights and improve model accuracy. The purpose of feature selection is to remove unimportant, redundant, or noisy features that may cause overfitting.

The basic idea behind Principal Component Analysis (PCA) is to reduce the dimensionality of a data set consisting of many correlated variables while preserving as much variation as possible within the data set. The same is done by transforming the variables into a new set of variables called the principal component. PCA helps to identify the relationship and dependencies between the features of a data set. The covariance matrix expresses the correlation between different variables in a data set. It is important to identify highly dependent variables because they contain biased and redundant information that degrades the overall performance of the model. of.Classification Algorithms

Classification
Support Vector Machine Classifier (SVM)
Support Vector Machine (SVM) is a supervised machine learning technique that is useful for both classification and regression problems. This is a technique that best separates two classes by a hyperplane or line. It works under the assumption that only support vectors are important, while other training samples can be ignored. This classifier is very efficient in large spaces [28]. In addition, a radial basis function (RBF) kernel was used in the experiments.

MLP (Multilayer Perceptron)
The concept of a multilayer perceptron is inspired by the human nervous system [29]. The advantages of MLP are that it is: (i) highly fault tolerant, ie. if neurons and the connections between them are broken, they continue to work (ii) it is non-linear in nature, so it is suitable for all types. of real problems. In our experiments, we used 100 hidden layers, the activation function is ReLU (Recified Linear Unit), and the learning rate is 0.01.

Decision tree (DT)
Decision trees are used for decision analysis ". Decision trees whose objective values can take continuous values are known as regression trees. When you think of a tree, [input values] are represented as a [path] from the root to the leaves, [where] each leaf represents a target variable" [31]. The steps in DT consist of: (i) "Construct the tree and its nodes as functions [(ii)] Select [one] function to predict the input output, where the root node is the one containing the highest information gain [(iii))] Repeat the above steps to create subtrees based on properties not used in the nodes above"

PERFORMANCE PARAMETERS

Four evaluation parameters are taken into consideration which are as follows:

Accuracy
It is the basis of measuring the quality of any predictive model. The accuracy measures the ratio of correct predictions to the total number of data points evaluated. This paper consists of the best accuracies that were obtained by various machine learning models after applying the feature selection and K-Fold techniques.
Accuracy= (TP+TN)/(TP +TN+ FP+FN)

Precision

The Precision of a model is the fraction of relevant occurrences among the retrieved occurrences. It is also referred to as a positive predictive value. It is calculated by taking the ratio of true positives by the total positives in a model. In simple words, a high precision means that the algorithm returns more relevant results than the irrelevant ones.
Precision = TP/(TP + FP)

Recall

The Recall is also known as the sensitivity of the model. It is the fraction of relevant occurrences that have been retrieved over the total amount of relevant occurrences. A high recall means that most of the occurrences returned were relevant. It is measured as the ratio of true positives to the summation of true positives and false negatives,
Recall = TP/(TP+FN)

F-Score

The F-Score is a measure that combines the precision and recall by taking its harmonic mean. It is approximately the average of the two when they are close, else their harmonic means. The harmonic mean is the ratio of the square of the geometric mean divided by the arithmetic mean.
F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Where,
True positive (TP): correct positive prediction
False positive (FP): incorrect positive prediction
True negative (TN): correct negative prediction
False negative (FN): incorrect negative prediction

Algorithms	**Accuracy**
SVM	65%
MLP	76%
DT	87%

Fig:2 Accuracy Table

CONCLUSION

Diabetes prediction is a classification technique with two mutually exclusive possible outcomes, either the person is diabetic or not diabetic. Today, diabetes has become a common disease of mankind from young to old. Therefore, early prediction of diabetes is very important to save human life from diabetes. Machine learning is an emerging field of computer science that deals with the ways in which machines learn from experience. This work aims to predict diabetes using three different supervised machine learning methods such as SVM, MLP and decision tree.

REFERENCES

[1]. Mohapatra, Saumendra Kumar, Jagjit Kumar Swain, and Mihir Narayan Mohanty. "Detection of diabetes using multilayer perceptron." International Conference on Intelligent Computing and Applications: Proceedings of ICICA 2018. Springer Singapore, 2019.

[2]. Kopitar L., Kocbek P., Cilar L., Sheikh A., and Stiglic G., Early detection of type 2 diabetes mellitus using machine learning-based prediction models, Scientific Reports. (2020) 10, no. 1, 11981, https://doi.org/10.1038/s41598-020-68771-z.

[3]. Bavkar V. C. and Shinde A. A., Machine learning algorithms for Diabetes prediction and neural network method for blood glucose measurement, Indian Journal of Science and Technology. (2021) 14.

[4]. Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. Applied computing and Informatics. 2018. Available from: https://doi.org/10.1016/j.aci.2018.12.004