**Prediction of diabetes using machine learning**

FATIMA INAMDAR VIVEK JADHAV

ASSISTANT PROFESSOR Vishwakarma institute of information technology

Vishwakarma institute of information technology

[fatima.inamdar@viit.ac.in](mailto:fatima.inamdar@viit.ac.in) SUNIKET KHAIRNAR

Vishwakarma institute of information technology

OM LIMBHARE

Vishwakarma institute of information technology

SATYAM MAURYA

Vishwakarma institute of information technology

***ABSTRACT***

***Diabetes is an illness that is rapidly becoming more prevalent, even in youngsters. Chronic diseases like diabetes have the potential to wreck the world's health care system. According to the International Diabetes Federation, there are 382 million diabetics worldwide. By 2035, this will increase to 592 million. Elevated blood glucose levels cause diabetes, also referred to as diabetes mellitus or simply diabetes. On the basis of physical and chemical exams, a number of standard approaches can be used to diagnose diabetes. Nevertheless, since diabetes affects human organs including the kidney, eye, heart, nerves, foot, and others, early diabetes prediction is a difficult assignment for medical professionals.***

***Data science has an emerging topic called machine learning that studies how machines learn from experience. The goal of this study is to create a system that, by fusing the findings of several machine learning approaches, can more accurately conduct early diabetes prediction for a patient. Through the use of SVM, logistic regression, and ANN, this study tries to predict diabetes using three different supervised machine learning techniques. Another goal of this effort is to suggest a reliable method for detecting diabetes early on.***

***KEYWORD***

***Training and testing, Diabetes prediction, Machine Learning, Supervised learning, Logistic Regression, SVM: Support vector machine.***

**I INTRODUCTION**

Even among children, diabetes is a condition that is spreading quickly. We must comprehend what occurs in the body without diabetes if we are to comprehend diabetes and how it arises. The glucose is the source of sugar (glucose). The beta cells in the pancreas create the hormone insulin.

Diabetes symptoms include:

* increased thirst,
* frequent urination,
* fatigue,
* sleepiness,
* loss of weight, and
* blurred vision.

Diabetes causes include: -

The primary cause of diabetes is genetics. It is brought on by at least two defective genes on chromosome 6, a chromosome that influences how the body reacts to different antigens. The development of type 1 and type 2 diabetes possibly affected by viral illness as well. According to studies, having viruses such hepatitis B, CMV, mumps, rubella, and coxsackievirus increases the risk of acquiring diabetes.

In order to determine if a person has diabetes or not, we applied the SVM algorithm. SVM is a supervised machine learning technique used for regression and classification. SVM will be used for classification in this case because our main objective is to determine from the provided dataset whether a person has diabetes or not.

**II LITERATURE SURVEY**

1) Mithusi Soni compared a number of techniques, including gradient boosting, decision trees, logistic regression, random forests, KNN, and SVM. The model will only be applicable to female patients because the author only considered data for females[5].

2) K. Vijaykumar suggested the random forest algorithm for diabetes prediction. A machine learning algorithm called random forest may accurately forecast diseases[2].

Data mining is a notion that Dheeraj Sheety utilised to analyse diabetes using patient databases[3].

4) Tejas Joshi used three different machine learning techniques to predict diabetes in his paper: SVM, logistic regression, and ANN.

[8]

After reading numerous papers, it is evident that predicting diabetes is a growing topic of research interest for scientists who want to teach a computer programme to recognise if a patient has diabetes or not. According to prior research, accuracy has not yet increased. Therefore, a system is needed to accurately forecast diabetes.

**III PROPOSED METHODOLOGY:**

**SYSTEM ARCHITECTURE**

The purpose of the paper is to look into models that can more accurately forecast diabetes. To test this hypothesis, we used a support vector machine (SVM) to predict diabetes. The period is briefly covered in the sections that follow.

Pima Indian Diabetes Dataset, a UCI repository, is where the data for this dataset was collected (from KAGGLE). Numerous features of the dataset are addressed below. [9]

|  |  |
| --- | --- |
| S. NO. | ATTRIBUTE |
| 1 | Glucose |
| 2 | Blood level |
| 3 | Skin thickness |
| 4 | Insulin |
| 5 | Bmi |
| 6 | Diabetes pedigree function |
| 7 | Age |
| 8 | Label |

Machine learning techniques were used to predict if a person has diabetes or not after the data had been processed.

Support Vector Machine, also referred to as Support Vector Machine, is an algorithm for supervised machine learning. The most common method of categorization is support vector machine. A hyperplane that divides two classes is produced using a support vector machine. In high-dimensional space, it can produce a hyperplane or collection of hyperplanes. This hyperplane can also be utilised for regression or classification. Support vector machines can distinguish between examples in particular classes and can also categorise entities for whom there is no supporting data. The nearest training point for any class is used for separation, which is carried out using a hyperplane. [1]

Algorithm-

• Decide which hyperplane best divides the class.

• You must determine the Margin, or the distance between the planes and the data, in order to determine the best hyperplane.

• Low distances between courses increase the likelihood of missed conception, and vice versa.

• As a result, we must choose the class with the highest margin. Margin is equal to the sum of the distances to the positive and negative points.



Fig 3.1 SVM hyperplane

Overview of proposed methodology

1. Data collection
2. Data pre-processing
3. Splitting the data into training and testing test
4. Applying machine learning techniques to train the data
5. result

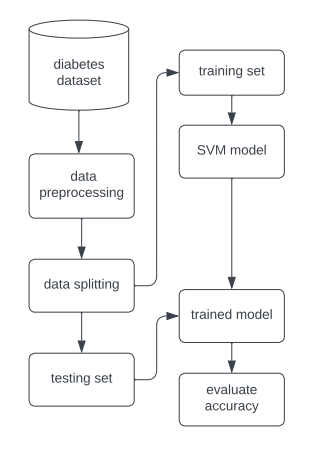


Fig 3.2 system architecture

**IV IMPLEMENTATION**

This is the most crucial stage, during which a model for diabetes prediction is built. The numerous machine learning techniques for diabetes prediction that were outlined above have been used in this. [3]

Algorithm

Step 1: Import diabetic dataset and necessary libraries.

Pre-process data to omit missing data in step two.

Step 3: Divide the dataset by 80% into a training set and 20% into a test set.

Apply the machine learning in step four.

Step 5: Create the classifier model based on the training set for the aforementioned machine learning technique.

Step 6: Use a test set to evaluate the Classifier model for the aforementioned machine learning algorithm. Comparative evaluation of the experimental performance outcomes for each classifier is step seven.

Step 8: Select the highest performing algorithm after analysis based on various metrics.

Data pre-processing

The most crucial phase is data pre-processing. Most data pertaining to healthcare has missing values and other contaminants that can affect the data's usefulness. Data pre-processing is done to increase the quality and efficacy of the results from the mining process. This procedure is crucial for accurate results and good prediction when applying machine learning techniques to the dataset. the results information from the initial diabetes. To train the data only on the specified parameters, csv was separated. [6]

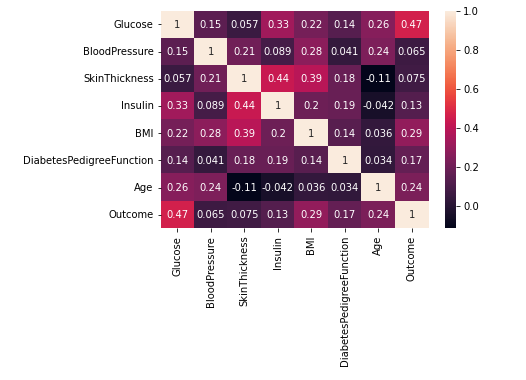


Fig 4.1 Heatmap after pre-processing data

Training and Testing

Training

Before training the data we need to split the data into train dataset and test dataset which was done using the following line of code

*X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X,Y, test\_size =  0.2, stratify=Y, random\_state=2)*

After splitting the data, now we have to train the model for train dataset so that the model with the help of training can predict diabetes result for various other data’s. The model was trained using support vector machine algorithm. The final training of the model was done using fitting the data to our model.

*classifier = svm.SVC(kernel='linear')*

*classifier.fit(X\_train, Y\_train)*

Testing

so as of now the model has been trained and now its time to test the model on trained dataset as well as test dataset and to compare the accuracies of both the result t avoid any kind of overfitiing problem.

*X\_train\_prediction = classifier.predict(X\_train)*

*training\_data\_accuracy = accuracy\_score(X\_train\_prediction, Y\_train)*

*X\_test\_prediction = classifier.predict(X\_test)*

*test\_data\_accuracy = accuracy\_score(X\_test\_prediction, Y\_test)*

Building the model

Now as our model is trained so we must write a code snippet which will evaluate whether the person has diabetes or not.

**V RESULTS AND DISCUSSION**

The accuracy on training data was 78% and the accuracy on testing data was 77% as we can see the difference between accuracies of training and testing data is very less so we can say that this is not the case of overfitting and our model is good to go to find the result. The accuracy which was achieved on sample data set was nearly 78%.

We have used SVM algorithm to predict whether a person is diabetic or not. SVM is a supervised machine learning algorithm used for both classification and regression. Our main goal here is to classify from given dataset whether a person is not.

**VI CONCLUSION**

Designing and implementing a Diabetes Prediction Using Machine Learning Method was the primary goal of this project (SVM). Data validation accuracy was 78%. For the primary detection of diabetes, the accuracy is adequate. Now that the model has been extracted, it may be used in a user interface. A person may use it primarily depending on their needs.

**VII REFERENCES**

[1] fig 3.1 javatpoint

[2] K.vijaykumar, B.lavanya, I.Nirmala, S.sofia caroline, “random forest algorithm for prediction of diabetes”

[3] Jingyu xue, Fancho min, Fengying ma, “research on diabetes prediction based on machine learning”

[4] Aishwarya Mujumdar, Dr. vaidehi V, “diabetes prediction using machine learning algorithm”

[5] Mitushi Soni, Dr. sunita verma “diabetes prediction using various machine learning techniques”

[6] Nahla B, Andrew “intelligible support vector machines for diagnosis of diabetes mellitus”

[7] Dheeraj sheety “diabetes detection”

[8] Tejas joshi, Pramila chawan, “diabetes prediction using ML”

[9] Kaggle (diabetes dataset)