International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET

ML – Based Diabetes Foretell Using SVM and Logistic Regression In Healthcare

Ayesha Siddiqua*[1], Ayesha Fatima*[2], Tahniyath Shaikh*[3], Dr. Ahmed Khan*[4]

*[1] BE Student, Department of Computer Science Engineering, ISL Engineering College , Hyderabad, Telangana, India

*[4] Associate Professor, Department of Computer Science Engineering, ISL Engineering College , Hyderabad, Telangana, India

ARTICLEINFO

Article History:

Accepted: 05 April 2023
Published: 20 April 2023

Publication Issue
Volume 10, Issue 2
March-April-2023

Page Number
553-559

ABSTRACT

Diabetes is one of the most grievous diseases in the world which has no remedy to cure it after a particular stage. Based on the survey of the last 20 years, the number of people having diabetes tripled. Over 422 million people in the world are diagnosed with diabetes. There are many factors that are responsible for the occurrence of diabetes. It is caused due to increased blood sugar level because of imbalance in insulin processing by the body, which leads to varieties of disorders like Coronary failure, blood pressure, etc and it can also effect other parts of the body. This project mainly focuses on the management of diabetes prediction, that will be approached using machine learning algorithms. Machine learning algorithms provide better results in diabetes detection by constructing models from patient datasets. The aim of this work is to make a prediction of diabetes more precisely with Logistic Regression (binary classification) and Support Vector Machine algorithm(SVM) in machine learning. It predicts the diabetes risk in early stages using symptoms and also predict using distinctive attributes of diabetes. Therefore, two different datasets of patients are used to train the models. This project work will function as an aid for the medical examiners in the diagnosis of diabetes of the patients. Thus, it can significantly help diabetes research and, ultimately, improve the quality of healthcare for diabetic patients.

Keywords : Support Vector Machine, Binary Classification

I. INTRODUCTION

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. There are two types of diabetes, namely, Type 1 and Type 2. Type 1 diabetes is characterized by deficient insulin production and requires daily administration of

insulin. The cause of Type 1 diabetes is not known, and it is not preventable with current knowledge. Type 2 diabetes results from the body's ineffective use of insulin. Type 2 diabetes comprises what the majority of people with diabetes around the world have and is mostly the result of excess body weight and physical inactivity. Changing lifestyles require deliberate effort. Therefore, diabetics must take the ultimate responsibility for their care and treatment using available technology-related systems. The advances in artificial intelligence (AI) and in particular machine learning and computer vision have made producing applications to automate tasks requiring intelligent behavior, learning, and adaptation possible, hence, providing solutions to real-life problems such as diabetes management. Many existing researches have handled for diabetes detection. Data mining approaches like clustering, classification using KNN were studied in existing system. Lot of work has been carried out to predict diabetic diseases using dataset. Different levels of accuracy have been attained using various machine learning techniques. The accuracy of the existing system is around 70-80%. It required more memory and processing time. Then it does not provide the history of patient diabetic report. Finally, we built up a diabetes prediction system based on the required inputs, with high accuracy and overcoming all above mentioned problems, by using Support Vector Machines (SVMs) and Logistic Regression ML algorithm. Also, the framework contain BMI, Insulin and Calorie calculation.

Project Statement: Producing a real time framework for diabetes management system to automate tasks required to predict diabetes by using a binary classification ML approach –Logistic Regression and SVM, in healthcare sector

II. LITERATURE SURVEY

- N. Sneha1 and Tarun Gangil has designed a model for Analysis of diabetes mellitus for early prediction using optimal features selection The dataset consists of 2500 entries and 15 attributes and 768 items used for testing and they have used 5 algorithms out of which support vector machine provides 77% accuracy.
- Raja Krishnamoorthi proposed a diabetes healthcare disease prediction framework using machine learning techniques. The dataset contains 768 rows and 9 columns and 90% of the 14 MAJOR PROJECT REPORT (2019-2023 Batch) Dept. of CSE, ISLEC data is used for training and 10% used for the testing purpose and they performed hyper parameter tuning to evaluate the Machine Learning models and used to increase the accuracy. Out of 5 algorithms best one is SVM and provide better accuracy as a result of 92%.
- Aishwarya and Vaidehi used several machine learning algorithms such as Support Vector Machines, Random Forest Classifier, Decision Tree Classifier, Extra Tree Classifier, Ada Boost algorithm, Perceptron, Linear Discriminant Analysis, Logistic Regression, K-NN, Gaussian Naïve Bayes, Bagging algorithm and Gradient Boost Classifier. They used two different datasets- the PIMA Indian and another Diabetes dataset for testing the various models. Logistic Regression gave them an accuracy value of 96%.

III. EXISTING SYSTEM

In existing system, many researches have handled for diabetes detection. Data mining approaches like clustering, classification using KNN were studied in existing system. Lot of work has been carried out to predict diabetic diseases using dataset. Different levels of accuracy have been attained using various machine learning techniques. The accuracy of the existing system is around 70-80%. It required more memory and processing time. Then it does not provide the history of patient diabetic report. Finally, we built up a diabetes prediction system based on the required inputs, with high accuracy and overcoming all above

mentioned problems, by using Support Vector Machines (SVMs) and Logistic Regression ML algorithm. Also, the framework contains BMI, Insulin and Calorie calculation.

IV. PROPOSED SYSTEM

The proposed method uses the SVM algorithm and Logistic Regression. The main idea behind the proposed system after reviewing the existing paper is to create a diabetic prediction system based on the required inputs, with high accuracy.

Purpose of Proposed System:

1.Developing a user-friendly web-based system for users and remote hospitals , to diagnose the diabetic and take correct treatment plan.

2. Recognizing diabetes diseases accurately from required inputs.

3. Maintains and store the information of patients.

Architecture diagrams:

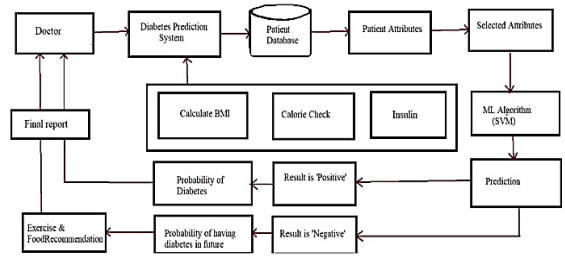


Fig 1. Architecture diagram for Module 1

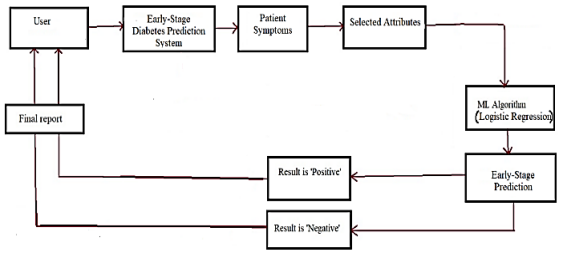


Fig 2. Architecture diagram for Module 2

V. IMPLEMENTATION

Importing necessary libraries: The required libraries like numpy, pandas, sklearn, numpy, matplotlib, etc of ML, Python are imported.

A. Data Collection

Dataset1: Diabetes Prediction Dataset

In the first step, we collect the data from a reliable source like the PIMA Indian diabetes dataset which is available in the csv format. The 8 parameters used are the number of times pregnant, Body mass index, plasma glucose, diastolic blood pressure, triceps skin fold thickness, diabetic pedigree function.

Dataset Description- the data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset have many attributes of 768 patients.

Table 1: Dataset Description

S No.	Attributes
1	Pregnancy
2	Glucose
3	Blood Pressure
4	Skin thickness
5	Insulin
6	BMI(Body Mass Index)
7	Diabetes Pedigree Function
8	Age

The 9th attribute is class variable of each data points. This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics.

Dataset 2: Early stage diabetes risk prediction dataset.

This dataset contains the sign and symptom data of newly diabetic or would be diabetic patient.

This has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor.

Data Set Characteristics:	Multivariate	**Number of Instances:**	520
Attribute Characteristics:	N/A	**Number of Attributes:**	17
Associated Tasks:	Classification	**Missing Values?**	Yes

B. Data Preprocessing

After loading the data, preprocessing is performed. Data preprocessing is the processing of a dataset in which data is transformed and encoded in a form such that the machine learning algorithm can parse it and only useful information is being extracted from the

dataset. The values are then read sequentially for further training.

C. Feature extraction

It is the process of converting raw data into numerical features that may be processed while maintaining the information in the original data set. It yields better results than simply applying machine learning to raw data. This is an important categorizing feature.

D. Model Creation

The SVM algorithm, which stands for Support Vector Machine, is used in this project for module 1 and Logistic Regression is used for module 2.

SVM is a classification and regression supervised machine learning algorithm. The Sci-kit Learn library has four SVM kernels. SVM creates a hyperplane that separate two classes. It can create a hyperplane or set of hyperplane in high dimensional space. This hyper plane can be used for classification or regression also. SVM differentiates instances in specific classes and can also classify the entities which are not supported by data. Separation is done by through hyperplane performs the separation to the closest training point of any class.

Algorithm-

• Select the hyper plane which divides the class better.

• To find the better hyper plane you have to calculate the distance between the planes and the data which is called Margin.

• If the distance between the classes is low then the chance of miss conception is high and vice versa.

• So we need to select the class which has the high margin. Margin = distance to positive point + Distance to negative point.

Logistic regression is a statistical tool that can be used in classification modelling about the presence or absence of diabetes. The logistic regression model transforms the linear regression function continuous value output into categorical value output using a sigmoid function, which maps any real-valued set of independent variables input into a value between 0 and 1. This function is known as the logistic function.

E. Training & Testing

Training: We split the data into training and testing datasets. During the training process we trained the machine from data source. We fit the SVM model for each kernel to our training set. We make predictions on our training set to see which kernel will give us the highest accuracy score. We call this Hyper-Parameter Optimization. The test data is transformed and predicts the accurate result.

Testing: Training data set will be validated using the test dataset model. The test data is transformed and predicted accurate result will be achieved, 90-92%.

F. Prediction

This module predicts the user is suffer from early-stage diabetes or not using SVM algorithm and predict the diabetes of patients using medical parameters using Logistic Regression and produce a final report.

In the future, this hierarchical framework combined with machine learning algorithms would be used to predict or analyze various disorders. Other ML computations can be used to enhance and improve the job for diabetes examination.

System Flow Diagrams:

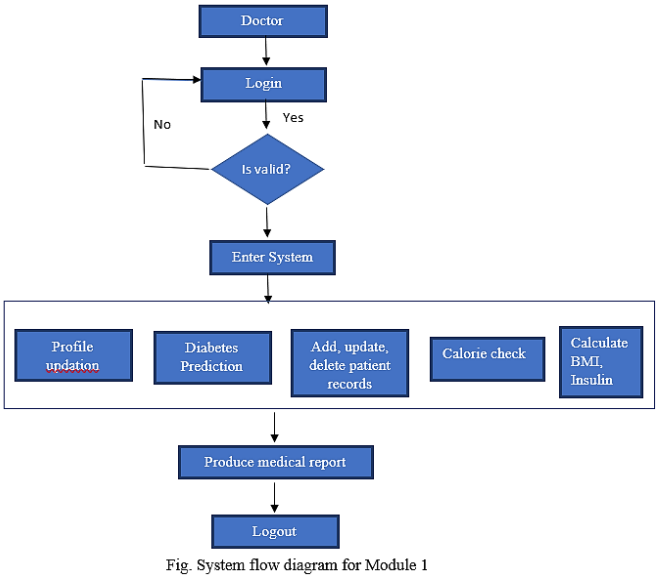


Fig. System flow diagram for Module 1

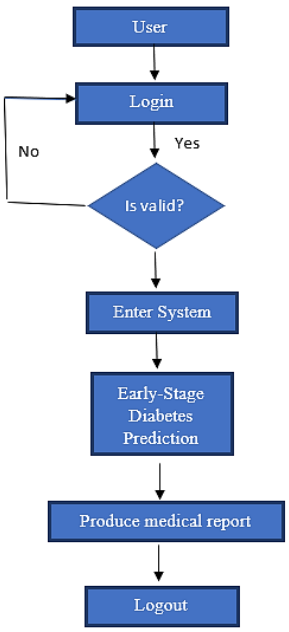
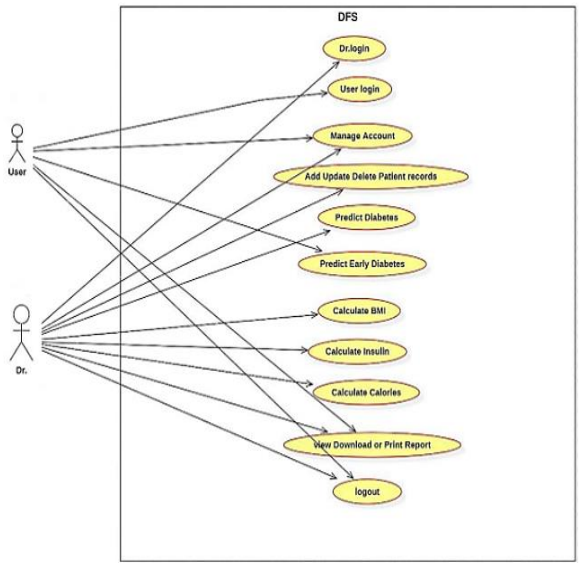


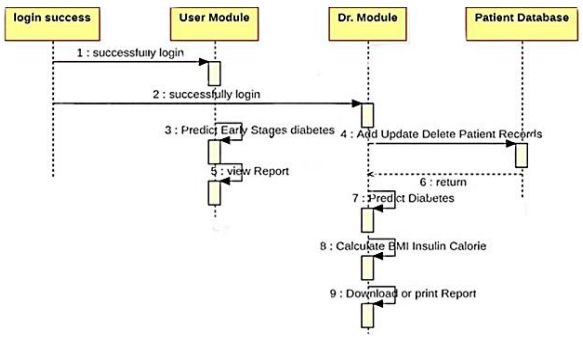
Fig . System flow diagram for Module 2

UML Diagrams:

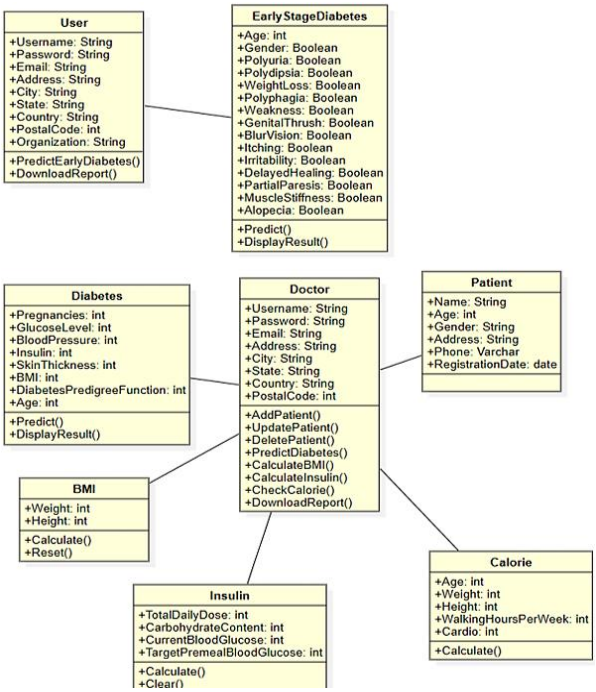
a. Use case Diagram



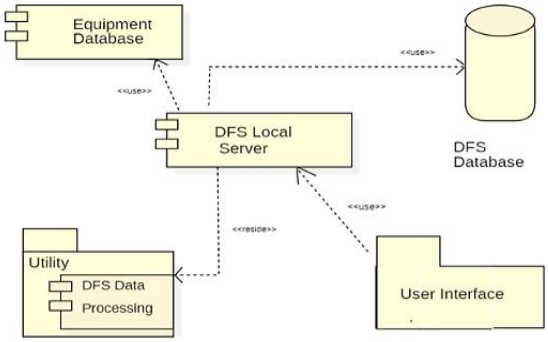
b. Sequence Diagram



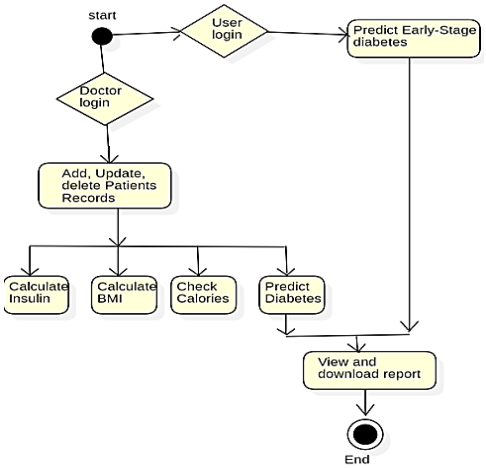
c. Class Diagram



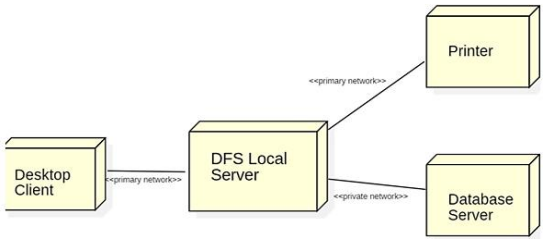
d. Component Diagram



e. Activity Diagram



f. Deployment Diagram



VI. REQUIREMENT SPECIFICATIONS

SOFTWARE TOOLS

OPERATING SYSTEM : Windows: 7 or newer
LANGUAGES USED : Python, HTML, CSS, JS
DBMS : MY SQL
IDE : PyCharm
Framework : Flask, Flask-MySQL
Web Server : XAMPP

FUNCTIONAL REQUIREMENTS

- The proposed system shall provide a platform to analyze the dataset for new patients.
- The system shall provide a framework to predict the early-stage diabetes using symptoms for the patients and predict the diabetes using medical parameters .
- The system maintains a database of patients information and medical records.

NON-FUNCTIONAL REQUIREMENTS

Scalability: System should be able to handle a large number of users. The system is capable enough to work properly.

- Speed: The application should be fast. It should not slow down with the increase of number of users. Search functionality should be fast to enable better end-user experience. The system should be quick enough to be able to respond to user actions with a short period of time.
- Usability: User interface should be simple and clear to break to understand to any user. At every step of this project user seems to be familiar with the interfaces as they are easy to use.
- Availability: The system should be available at every moment to the user. It should be ensured that there should be minimum or no downtime to ensure better user experience for students.
- Reliability: The system should be reliable and yield correct results if a user performs any actions. Also, if the farmer uploads a image, the system should ensure that the correct message is delivered to the correct destination without any loss of content.
- Testability: The application is tested for validation, uploading images, message structures and works fine.

VII. ADVANTAGES

- Early detection of diabetes: One of the primary advantages of this project is the ability to detect diabetes early. Early detection allows for more effective treatment, reducing the potential for health loss.
- This diabetic analysis will be helpful for patients, remote hospitals and remote doctors to diagnose the diabetic and take correct treatment plan.
- Better accuracy: The use of technology can improve the accuracy of diabetes detection. This can reduce false positives and false negatives, allowing farmers to take appropriate action based on accurate information.
- It is more accurate than existing system.

VIII. VIII. DISADVANTAGES

- No backup system.
- High initial costs.

IX. APPLICATIONS

- Monitor person's health and detect diabetes in early stages.

- It can be used in remote hospitals to monitor the medical parameters and detect the diabetes of patients, allowing for more effective treatment.
- Research

X. CONCLUSION

In this project, a diabetes foretell is made in which we predicted probabilities of diabetes and generated the results based on the required input and at the last provided a final diabetes report to be downloaded.

By this, the medical doctors can make decisions for further treatments. Besides that, accuracy is increased for the system.

It improves the quality of healthcare for diabetic patients.

Therefore, this system can be used as a support to medical decision making in healthcare environments.

XI. FUTURE WORK

- Different types of diabetes levels along with risk levels using an image dataset can be predicted.
- Features like exercise recommendation system can be added to the prediction environment.
- A structured dataset has been selected in the model but in the future, unstructured data will also be considered, and these methods will be applied to other medical domains for prediction, such as for different types of cancer, psoriasis, and Parkinson's disease.
- Other attributes including physical inactivity, family history of diabetes, and smoking habit, will also be planned to be considered in the future for the diagnosis of diabetes.

XII.REFERENCES

[1]. Mitushi Soni, Dr. Sunita Varma, “Diabetes Prediction using Machine Learning Techniques”, International Journal of Engineering Research & Technology, Volume 9, pp. 921-925, 2020.

[2]. Raja Krishnamoorthi, Shubham Joshi, and Hatim Z. Almarzouki, “A Novel Diabetes Healthcare Disease Prediction Framework using Machine Learning Techniques,” Journal of Healthcare Engineering, pp. 1- 10 2022.

[3]. Desmond Bala Bisandu, Godwin Thomas "Diabetes Prediction using Data mining Techniques,” International journal of research and Innovation in Applied Sciences, volume 4, pp. 103-111, 2019.

[4]. Salliah Shafi, Prof. Gufran Ahmad Ansari, “Early Prediction of Diabetes Disease & Classification of Algorithms Using Machine Learning Approach”, International Conference on Smart Data Intelligence, 2021.

[5]. Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques" .Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13.

[6]. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance’s for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[7]. K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019

Cite this article as :

Ayesha Siddiqua, Ayesha Fatima, Tahniyath Shaikh, Dr. Pathan Ahmed Khan, "ML - Based Diabetes Foretell Using SVM and Logistic Regression In Healthcare", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 10 Issue 2, pp. 553-559, March-April 2023.

Journal URL : https://ijsrset.com/IJSRSET2310279