International Research Journal of Computer Science (IRJCS)
Issue 08, Volume 09 (August 2022)
ISSN: 2393-9842
https://www.irjcs.com/archives

DIABETES PREDICTION USING MACHINE LEARNING

Ajay V
Department of Information Science and Engineering
Vemana Institute of Technology, Bangalore
Andrew S Dhas
Department of Information Science and Engineering
Vemana Institute of Technology, Bangalore
Anish Sanchith
Department of Information Science and Engineering
Vemana Institute of Technology, Bangalore
Chandan K R
Department of Information Science and Engineering
Vemana Institute of Technology, Bangalore
B V Vinay
Assistant Professor, Department of ISE
Vemana Institute of Technology, Bangalore

Publication History
Manuscript Reference No: IRJCS/RS/Vol.09/Issue08/SPAUCS10086
Research Article | Open Access
Peer-review: Double-blind Peer-reviewed
Article ID: IRJCS/RS/Vol.09/Issue08/SPAUCS10086
Volume 2022 | Article ID SPAUCS10086 http://www.irjcs.com/volumes/Vol09/iss-08/07.SPAUCS10086.pdf
Received: 25, July 2022 | Accepted: 04, August 2022 | Published: 12, August 2022
doi: https://doi.org/10.26562/irjcs.2022.v0908.007
Citation: Ajay,Andrew,Anish,chandan,Vinay(2022).Diabetes Prediction Using Machine Learning. International Research Journal of Computer Science (IRJCS), Volume IX, Issue VIII, 2022, Pgs190-194
BibTeX key:
Academic Editor-Chief: Dr.A.Arul Lawrence Selvakumar

Abstract: Diabetes is a fatal chronic illness that adversely affects every system in the body. This disease affects many people, and its adverse effects cause a sizable number of sufferers to pass away each year. High blood glucose levels are a concern for diabetic patients. It is quite challenging to make a robust and accurate prediction of diabetes due to the tiny amount of labeled data and outliers (or incomplete data) in the datasets for diabetes. We are putting up a reliable framework for diabetes prediction in this literature. Untreated diabetes may result in hearing loss, renal damage, heart and blood vessel disease, poor wound healing, and a few skin problems. For the classification, early diagnosis, and prediction of diabetes, a machine learning (ML)-based strategy has been put forth.
Keywords: Diabetes, Machine Learning, Supervised Mode, Random Forest Classifier, Prediction of Diabetes.

I. INTRODUCTION

Diabetes affects millions of individuals worldwide. Which is one of the deadliest degenerative illnesses that can adversely affect every system of the body. Diabetes affects 422 million people worldwide, or 8.5 percent of the world's population, according to WHO. Databases in the healthcare business are very vast. Structured, semi-structured, or unstructured data may be present in such databases. Big data analytics is a method that examines enormous data sets and finds hidden patterns and information to extract insight from the data. Given the current situation, diabetes mellitus (DM) has developed into a highly serious illness in developing nations like India. Many people suffer from diabetic mellitus, which is categorized as a no communicable disease (NCB). According to statistics from 2017, there are 425 million people worldwide who have diabetes. Every year, diabetes claims the lives of between 2 and 5 million sufferers. According to estimates, this number will reach 629 million by 2045. Types of diabetes mellitus include:

1. Insulin-Dependent Diabetes Mellitus, also referred to as Type-1 (IDDM). Because of this type of DM, a patient needs to receive insulin injections because their body is unable to produce enough insulin on its own.
2. Non-insulin-dependent diabetes mellitus, sometimes referred to as Type-2 (NIDDM). This form of diabetes manifests as improper insulin uses by bodily cells.
3. Type-3 Gestational Diabetes, which develops when a pregnant woman's blood sugar level rises, and the diabetes is not identified sooner. Long-term consequences are linked to DM. A diabetic person also has a significant risk of developing several health issues.

Predictive analysis is a method for gathering information and forecasting future occurrences that combines several machine learning techniques, data mining techniques, and statistical tools. Important judgments and using predictive analysis on healthcare data, predictions may be generated. Predictive analytics can employ regression and machine learning techniques. Predictive analytics strives to improve healthcare outcomes by increasing patient care, improving resource use, and making the best possible disease diagnosis. One of the most crucial aspects of artificial intelligence is machine learning, which enables the creation of computer systems with the capacity to learn from the past without the need for programming in every situation. To assist automation with the fewest defects and remove human jobs, machine learning is an urgent necessity in today's world. Lab procedures including oral glucose tolerance testing and fasting blood glucose testing are currently used to diagnose diabetes. This approach takes a lot of time, though. This research focuses on developing a predictive model for diabetes prediction utilizing machine learning algorithms and data mining approaches. Patients who are at risk for diseases or other health issues can be found using disease prediction systems. The quality of care can then be improved, and potential hospital admissions can be avoided by clinicians taking the necessary steps to avoid or limit the risk. Recent developments in data analytics tools and methods have made it possible to use a wealth of semantic data for disease risk prediction, including demographics, clinical diagnosis and measurement, health habits, test results, prescriptions, and service utilization. In this sense, creating disease prediction models could be done using electronic health data. Large-scale electronic health databases, various techniques, and healthcare factors have all been used in a substantial number of these disease prediction model that have been put forth in the literature over time. This thorough literature review's objective was to discuss several risk predictions models that have been put out using electronic health data. The search criteria were created to locate pertinent research publications that identified illness risks using electronic health data. The results were analyzed and compared in terms of the technique utilized, disease kind, and prediction accuracy after being retrieved from online scholarly databases. In-depth analysis of the use of electronic health data for risk prediction models is provided by this project. This study also compared the outcomes of several modelling approaches for three commonly modelled diseases utilizing electronic health data. Additionally, the benefits and drawbacks of several risk prediction models were discussed, along with how well they performed. The prediction of diseases has frequently made use of electronic health data. Using such data, a few modelling systems demonstrate extremely high accuracy in forecasting certain diseases. To improve results, the clinical decision-making process has been informed by these modelling approaches.

II. LITERATURE REVIEW

Umair Muneer Butt et.al[1] A significant amount of vital and sensitive healthcare data has been produced because of the tremendous breakthroughs in infrastructure for public healthcare and biotechnology. Many intriguing patterns are discovered using intelligent data analysis tools for the early and onset diagnosis and prevention of various fatal diseases. Because it causes damage to the heart, kidneys, and nerves, diabetes mellitus is a condition that poses a serious risk to life. In this study, a machine learning-based method for classifying, identifying, and predicting diabetes in its early stages has been developed. Additionally, it offers a fictitious Internet of Things-based diabetes monitor system for a healthy and affected individual to track his blood glucose level. Three distinct classifiers-random forest (RF), multilayer perceptron, and logistic-regression have been used to categorize diabetes (LR). Long short-term memory (LSTM), moving averages (MA), and linear regression have all been used for predictive analysis (LR).A benchmark PIMA Indian Diabetes dataset is used for experimental evaluation. The investigation shows that MLP performs better than other classifiers with 86.08 percent accuracy, and LSTM enhances the significant prediction of diabetes with 87.26 percent accuracy. The applicability of the suggested strategy in numerous public healthcare applications is further illustrated by a comparison analysis of the proposed approach with current state-of-the-art methodologies.

Vaibhav Kulkarniet.al [2] The system known as "disease prediction using machine learning" uses data provided by patients or other users to make disease predictions. The user's symptoms are processed by the algorithm, which then outputs the likelihood that the user will develop the condition. The supervised machine learning method known as the Naive Bayes classifier is utilized in the disease prediction process. The Nave Bayes method determines the likelihood of the disease. Accurate analysis of medical data helps with early illness identification and patient care as biological and healthcare data volumes rise. We are forecasting diseases like Jaundice, Malaria, Dengue, Diabetes and Tuberculosis using decision trees and linear regression.

KM Jyoti Rani et.al[3] Diabetes mellitus is a disease that could lead to a global health care disaster. 382 million people worldwide have diabetes, according to the International Diabetes Federation. This will double to 592 million by 2035. Diabetes is a condition brought on by elevated blood glucose levels. The symptoms of this elevated blood sugar level include excessive urination, extreme thirst, and decreased appetite. One of the main causes of stroke, heart problems, kidney problems, disfigurement, myopia, and organ failure. When we eat, our bodies turn the food we ingest into sugars like glucose. Our pancreas is then expected to release insulin. Insulin acts as a key to unlock our cells, allowing glucose to enter and be used by us as fuel. However, this mechanism does not function in diabetes.

The most prevalent forms of the disease are type 1 and type 2, but there are other varieties as well, including gestational diabetes, which develops during pregnancy. 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. Support vector machine, K nearest neighbors, Random Forest, Decision tree, and Logistic Regressions are just a few of the methods that are employed. Each algorithm's accuracy is calculated along with the model's accuracy. The model for predicting diabetes is then chosen from those with good accuracy.

III. PROPOSED SYSTEM

Our method can anticipate many diseases. It allows the machine learning model to take structured data as input. Patients and other users who are end users use this system. The user will enter all the symptoms they are experiencing into this system. The machine learning model will then be given these symptoms in order to predict the disease. The best accuracy is then determined by applying the appropriate algorithms. The system will then use symptoms to forecast disease. The technology used by this system is machine learning (ML).The Naive Bayes algorithm is used to predict the disease using symptoms, the KNN algorithm is used for classification, the features with the greatest impact value are extracted using logistic regression, and the decision tree is utilized to break the large dataset into manageable chunks. The model's projected disease will be the system's ultimate output.

SYSTEM ARCHITECTURE

An orderly description of the proposed system's structural views, operational functions, and system views. The system architecture outlines the key elements, their interconnections, organizational structures, and interactions with major key components. Software architecture and design involve a wide range of contributing factors, including business strategy, quality attributes, and many more.
Software Architecture and Design can be mainly classified into two phases

- Software Architecture
- Software Design.

Software architecture defines the basic components of a programme system, each of which is made up of the software's components as well as the relationships and qualities between them.

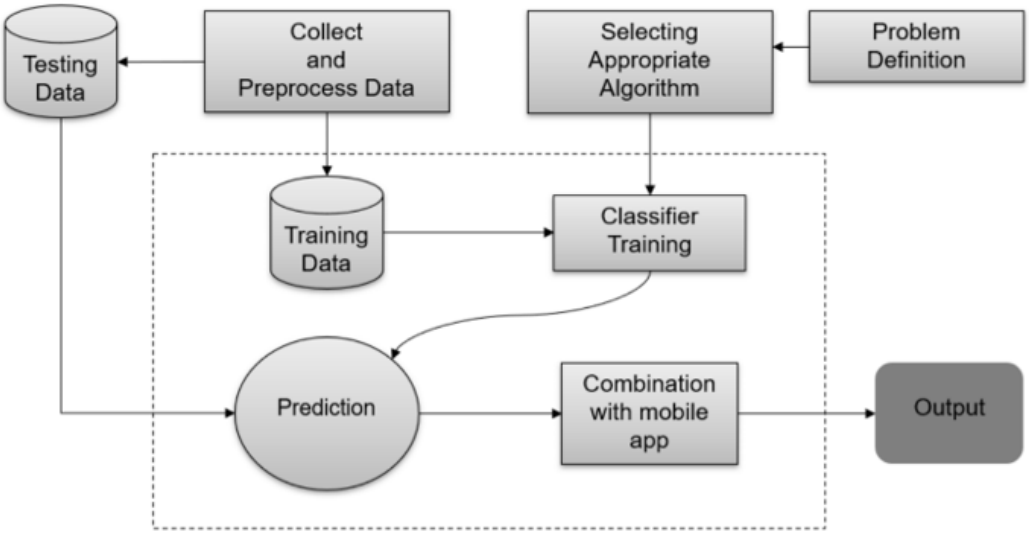

Figure 1 System Dataflow

Implementation of system means the process of making an updated or new system design operational based on proposed model. The implementation requires the following:

Design Approach

The use of the Diabetes Prediction System, which is being developed for early diabetes detection and advice on how to live a healthy life. As a result, the system's prediction feature should play a key role. There are three main phases in this system's fundamental process.

1. The user must first enter the required data into the mobile app.
2. The system's backend offers a precise forecast based on the previously trained model. Data gathering and pre-processing, statistical analysis, machine learning model creation, and front-end development are the four main steps that make up the total development process.

3. The system offers advice to the user on how to maintain their health. It is important to create models that perform accurately on test datasets.

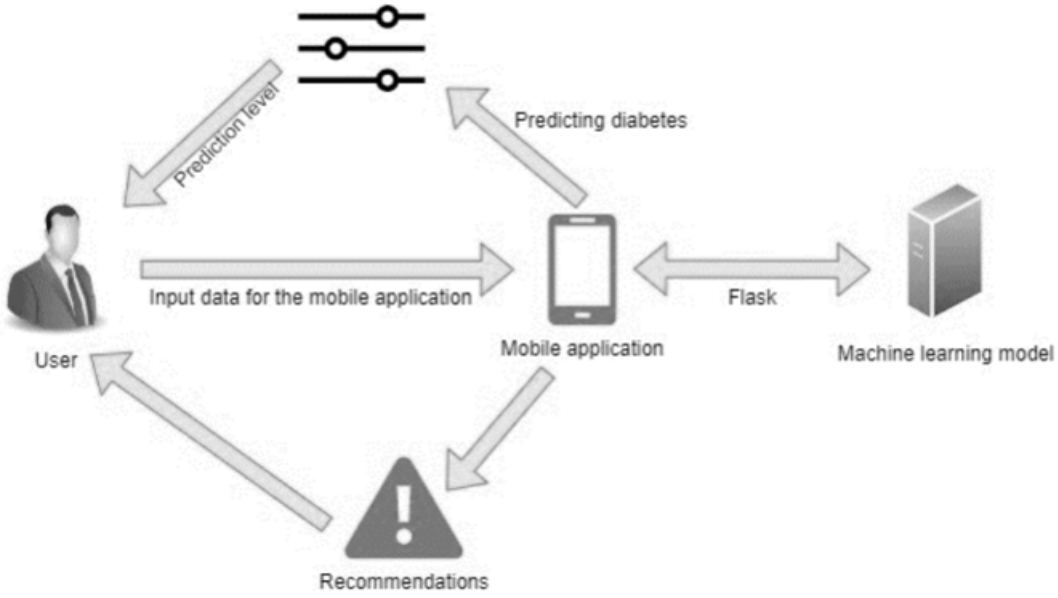

Figure 2 Design Approach

Data Collection and Pre-Processing

To have accurate results, the machine learning algorithm has to be fed with a substantial amount of data. The dataset is obtained from authenticate sources like health organizations. To balance the dataset, both male and female information from multiple age groups were acquired.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	62	35	0	33.6	0.127	47	1
1	0	84	82	31	125	38.2	0.233	23	0
2	0	145	0	0	0	44.2	0.630	31	1
3	0	135	68	42	250	42.3	0.365	24	1
4	1	139	62	41	480	40.7	0.536	21	0

Table1 Collaborative Data obtained from PIMA and FrankFurt Hospital

Statistical Analysis

In order to determine whether the chosen risk factors are appropriate for predicting diabetes, a correlation analysis has been performed. A popular method for finding intriguing connections in data is correlation analysis. These connections aid in our understanding of how relevant certain characteristics is to the target class under prediction.

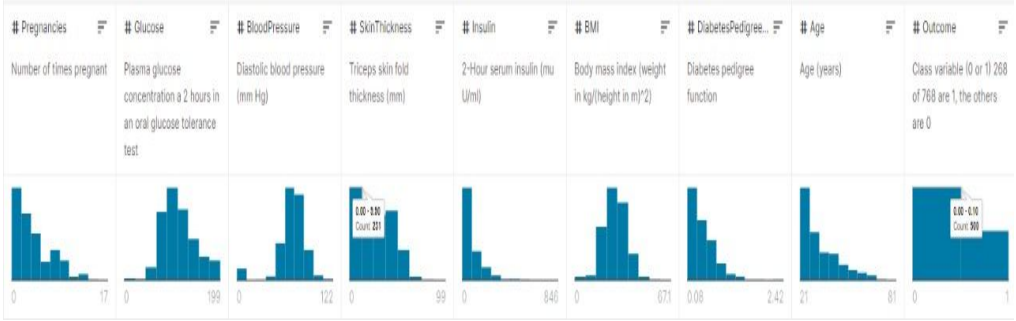

Figure 3 Correlation Analysis between Attributes

Development of ML Model and Application

At this stage, the machine learning technique based on artificial neural networks has been utilized to predict diabetes. Algorithms like Decision Tree Classifier and Logistic Regression to analysis with suitable real world data sets. Since this is application-based research, the front-end is being developed using flutter which provides a fast and expressive way for developers to build native apps on both IOS and Android. The application would serve a purpose to the public, hence there would be an effort to deploy the application into real world.

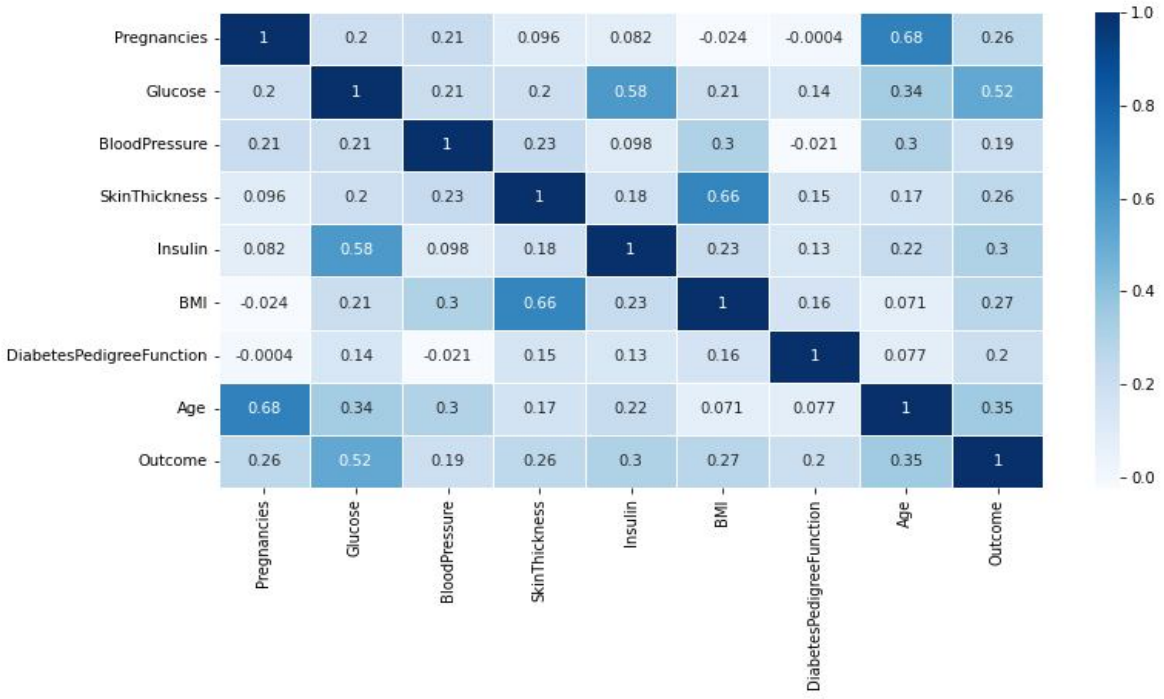

Figure 4 Correlation Matrix between Attributes

CONCLUSION AND FUTUREWORK

Different classification algorithms like naïve bayes, KNN, decision tree, ANN and random forest classifiers were applied on our dataset, and results for all techniques were slightly different as the working criteria of each algorithm is different. The results were evaluated based on accuracy by comparing with different algorithms and found that random forest Algorithm gives 99% accuracy based on the given classification and prediction of dataset. An application is developed for predicting diabetes in an early stage and giving recommendations to maintain a healthy life. This project has helped us in understanding, experiencing the challenges faced in real life projects. Working for this project has helped us learn about team building; helped us improve our communication skills and made us understand the challenges like understanding the literature survey, designing the architecture, finding out appropriate algorithms and figuring out the methodology to work with. Overcame those problems by going through the base papers, which gave us a clear idea for further proceedings. In future this application would serve a purpose to the public, hence there would be an effort to deploy the application into real world, can be further extended to consult doctors by taking appointments through this application. Charts can be integrated to visualize the historical reports of the patients. Google and mobile number integration can be done for user authentication. Use cases such as integrating IOT devices to read values from glucometer and other smart devices instead of inputting manual values.

REFERENCES

1. Mitushi Soni,, “Diabetes Prediction using Machine Learning Techniques”, Dept of Computer Science and Engineering Shri G.S. Institute of Technology and Science Indore, India.
2. Katulanda, P., Sheriff, M., Matthews,Diabetes://www.who.int/westernpacific/healthtopics/diabetes (accessed 10.19.19), 2009.
3. https://flutter.dev/ (Widgets, Dependencies such as http, firebase_auth, provider)
4. https://firebase.google.com/docs (Crud operations, implementations, authentication, firestore)
5. Aminul, Md., Jahan, N., 2017. “Prediction of Onset Diabetes using Machine Learning Techniques”. Int. J. Comput. Appl. 180, 7–11.
6. Predictive modelling and analytics for diabetes using a machine learning approach. Appl. Comput. Inform. S221083271830365X
7. Sisodia, D., Sisodia, D.S., 2018. “Prediction of Diabetes using Classification Algorithms”. Procedia Comput. Sci. 132, 1578–1585.
8. Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. “Application of data mining: Diabetes health care in young and old patients”., 127–136. doi:10.1016/j. jksuci.2012.10.003.