ISSN: 2582 - 6379
Orange Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2025, Volume 6 Issue 1
www.ijisea.org

Diabetes Prediction Using Machine Learning

N.Srinivasan[1], S.Afisha[2],S.Asma[3],K.Udayakshitha[4], R.Shanthi Priya[5],M.Lalitha Devi[6]
[1,2,3,4,5,6]Chaitanya Bharathi Institute of Technology, Proddatur, A.P, India.

shaikh.afisha.shaik@gmai.com[2], shaikasmashaikasma46@gmail.com[3], udayakondreddy@gmail.com[4], shanthipriyarushigalla@gmail.com[5], madisettylalitha2004@gmail.com[6]

ABSTRACT

Diabetes Mellitus is among critical diseases and lots of people are suffering from this disease. Age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause Diabetes Mellitus. People having diabetes have high risk of diseases like heart disease, kidney disease, stroke, eye problem and nerve damage, etc. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. Current practice in hospital is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis. In existing method, the classification and prediction accuracy are not so high. In this project, we have proposed a diabetes prediction models like Logistic Regression, KNN, SVM, Naivesbayes classifier, Decision tree, Random forest for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is boosted with new dataset compared to existing dataset. Further with imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification .The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for predicting the diabetes.

Keywords: Diabetes Mellitus, Glucose levels, Diabetic problems(complex) , Prediction Model.

I. INTRODUCTION

Healthcare sectors have large volume databases. Such databases may contain structured, semi-structured or unstructured data. Big data analytics is the process which analyses huge data sets and reveals hidden information, hidden patterns to discover knowledge from the given data. Considering the current scenario, in developing countries like India, Diabetic Mellitus (DM) has become a very severe disease. Diabetic Mellitus (DM) is classified as Non-Communicable Disease (NCB) and many people are suffering from it. Around 425 million people suffer from diabetes according to 2017 statistics. Approximately 2-5 million patients every year lose their lives due to diabetes. It is said that by 2045 this will rise to 629 million. .However, prevalence of diabetes is found among various Countries like Canada, China, and India etc. Population of India is now more than 100 million so the actual number of diabetics in India is 40 million. Diabetes is major cause of death in the world. Early prediction of disease like diabetes can be controlled and save the human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. F [1] Diabetes Mellitus (DM) is classified as- Type-1 known as Insulin-Dependent Diabetes Mellitus (IDDM). Inability of human's body to generate sufficient insulin is the reason behind this type of DM and hence it is required to inject insulin to a patient. Type-2 also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). This type of Diabetes is seen when body cells are not able to use insulin properly. Type-3 Gestational Diabetes, increase in blood sugar level in pregnant woman where diabetes is not detected earlier results in this type of diabetes. DM has long term complications associated with it. Also, there are high risks of various health problems for a diabetic person.

ISSN: 2582 - 6379
Orange Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2025, Volume 6 Issue 1
www.ijisea.org

Symptoms of Diabetes: Frequent Urination , Increased thirst , Tired/Sleepiness Weight loss , Blurred vision , Mood swings , Confusion and difficulty concentrating , frequent infections.

Causes of Diabetes : Genetic factors are the main cause of diabetes. It is caused by at least two mutant genes in the chromosome 6, the chromosome that affects the response of the body to various antigens. Viral infection may also influence the occurrence of type 1 and type 2 diabetes. Studies have shown that infection with virursessuch as rubella, Coxsackievirus, mumps, hepatitis B virus, and cytomegalovirus increase the risk of developing diabetes.

II. LITERATURE SURVEY

In India, diabetes is a widespread problem as more than 70% of the adult population is suffering from this disease. Various researchers have worked to predict symptoms of diabetes by applying different approaches such as machine learning and data mining [11]. Few of them have also applied neural network and genetic algorithm. Since the problem of prediction of diabetes is supervised in nature, the supervised methods of machine learning, data mining and ANN have been applied by many. Some closely related works are discussed in this section. Many of the research studies have used Pima Indians Diabetes Dataset (PIDD) for diabetes prediction. Machine learning methods and Weka tool were applied by [13, 14, 16, 17, 20, 21, 23]. The different approaches applied by researchers can be broadly classified as machine learning methods, data mining techniques, hybrid methods and neural network or genetic algorithms. Swapna et. al. in [12] used deep learning methods on electrocardiogram (ECG) signals for detection of diabetes. Specifically, convolution neural network and long short-term memory has been used by them and then features were extracted by support vector machine. As a result, they found a very high accuracy of 95.7%. Meng et. al. in compared logistic regression, artificial neural network (ANN) and decision tree (DT) for identifying the risk of diabetes and prediabetes based on 12 risk factors which included education level, work stress, BMI, age, sleep duration, gender, marital status, family history of diabetes, coffee drinking, preference to salty foods, physical activity, and consumption of fish. DT was found to provide best results among the three methods. Choubey et. al. applied a hybrid algorithm using genetic algorithm (GA) and radial basis function neural network (RBFNN), wherein first GA is applied for feature selection then RBFNN is applied for classification. Their findings were that hybrid method performed better than RBFNN alone. Tigga et. al. in applied logistic regression on PIDD for diabetic prediction and found number of pregnancies, BMI and glucose level are most significant variables for diabetes prediction among all features in PIDD. Huang et. al. in did feature selection and classification of diabetes by applying naïve bayes, IB1 and C4.5 algorithms. The study concluded that patient age, diagnosis duration, need of insulin and diet control are most important factors for blood sugar control. Some other factors are also affecting results that are type of care, home monitoring and importance of smoking. Saravana et. al. in collected raw data from various places in form of Electronic Reports (EHR) that may be clinical reports, prescriptions given by doctors, diagnostic centre reports, pharmacy related information, and data asked by insurance personals. All this information collectively put in a map reduce to exact features which are directly related to diabetes. Nongyao et. al. in compared four classification techniques i.e. decision tree, ANN, logistic regression and naïve bayes. Further bagging and boosting were applied on all and random forest was also included. The maximum accuracy achieved by all was in between 84% and 86%. Zou et. al. in applied Random Forest, J48, ANN for classification after the feature reduction is done by unsupervised methods: Principal Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (mRMR) methods. Accuracy for mRMR is found to be better than PCA with all features. Perveen et. al. in were concerned in finding risk of metabolic syndrome and diabetes. For prediction Naïve Bayes and J48 (C4.5) decision tree model was applied and the balancing of training set was done by k-medoids sampling. In their study, NB outperformed the others. Rahman et. al. in summarises the effect of various data mining techniques for diabetes diagnosis. For the prediction purpose, Multilayer Perceptron (MLP), Bayes Classification, J48graft, JRip (RIP- PER), Fuzzy Lattice Reasoning (FLR) classification methods were applied. J48graft was found most accurate. Choi et. al. in applied machine learning algorithms on patients having history of non-diabetes having cardiovascular risk. Five years data have been collected in form of EMR from Korea University Guro Hospital. Then, machine learning methods were applied with 10-fold cross validation.

Highest accuracy was obtained in logistic regression model

ISSN: 2582 - 6379
Orange Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2025, Volume 6 Issue 1
www.ijisea.org

Aiswaryaet al. [2] aims to discover solutions to detect the diabetes by investigating and examining the patterns originate in the data via classification analysis by using Decision Tree and Naïve Bayes algorithms. The research hopes to propose a faster and more efficient method of identifying the disease that will help in well-timed cure of the patients. Using PIMA dataset and cross validation approach the study concluded that J48 algorithm gives an accuracy rate of 74.8% while the naïve Bayes gives an accuracy of 79.5% by using 70:30 split. after applying the resample filter over the data. The author emphasis on the class imbalance problem and the need to handle this problem before applying any algorithm to achieve better accuracy rates. The class imbalance is a mostly occur in a dataset having dichotomous values, which means that the class variable has two possible outcomes and can be handled easily if observed earlier in data preprocessing stage and will help in boosting the accuracy of the predictive model.

III. EXISTING SYSTEM

Existing systems for diabetes prediction primarily leverage machine learning algorithms trained on datasets of patient data, aiming to identify individuals at risk of developing the disease or predict its onset. These systems utilize various algorithms like Random Forest, logistic regression,KNN and Support Vector Machines, often achieving high accuracy in prediction.

Machine learning Algorithms in diabetes prediction are to analyze patient data like blood sugar levels, BMI, age, family history, and other relevant medical parameters, allowing them to predict the likelihood of someone developing diabetes based on their individual risk factors; these models can be particularly valuable for early detection and preventative measures.

To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying various Machine Learning Techniques. Machine learning techniques Provide better result for prediction by constructing models from datasets collected from patients. In order to conduct the experiment, 768 instances have been collected through an online Pima Indian database. In this work we will use Machine Learning algorithms. The Project work gives the accurate or higher accuracy model shows that the model is capable of predicting diabetes effectively. Our Result shows that we have achieved good accuracy.

Advantages of existing system:

Established infrastructure: Existing systems have established infrastructure, making it easier to implement and integrate diabetes prediction models.
Large datasets: Existing systems often have access to large datasets, which can be used to train and validate diabetes prediction models.
Clinical expertise: Existing systems have clinicians and healthcare professionals who can provide valuable insights and expertise in developing and implementing diabetes prediction models.
Patient engagement: Existing systems often have established patient engagement platforms, making it easier to integrate diabetes prediction models and provide personalized recommendations.

Disadvantages of existing system:

Limited scalability: Existing systems may have limited scalability, making it challenging to integrate diabetes prediction models that require large computational resources.
Data silos: Existing systems often have data silos, making it challenging to integrate data from different sources and provide a comprehensive view of patient data.
Legacy systems: Existing systems may have legacy systems that are not compatible with modern diabetes prediction models, making integration challenging.

ISSN: 2582 - 6379
Orange Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2025, Volume 6 Issue 1
www.ijisea.org

Resistance to change: Existing systems may have resistance to change, making it challenging to implement new diabetes prediction models and workflows.

Proposed System:

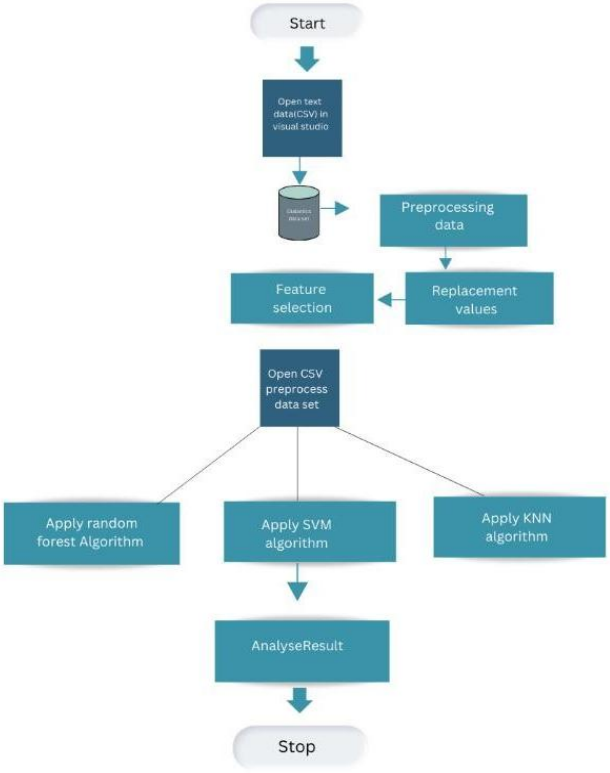


Fig 1. shows architecture diagram for diabetes prediction model

This model has five different modules. These modules include:
i.Dataset Collection ii. Data Pre-processing iii. Classification iv.
Build Model
Evaluation for Result
In this section we shall learn about the various classifiers used in machine learning to predict diabetes. We shall also explain our proposed methodology to improve the accuracy. Five different methods were used in this paper. The different methods used are defined below. The output is the accuracy metrics of the machine learning models. Then, the model can be used in prediction. Dataset Description. Diabetes dataset containing 2000 cases. The objective is to predict based on the measures to predict if the patient is diabetic or not.
Goal of the project is to investigate for model to predict diabetes with better accuracy. We experimented with classification and algorithms as kNN to predict diabetes. the data is gathered from UCI repository which is named as Pima
Indian Diabetes Dataset. The dataset has many attributes of 768 patients

ISSN: 2582 - 6379
Orange Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2025, Volume 6 Issue 1
www.ijisea.org

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 767 entries, 0 to 766
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               767 non-null    int64
 1   Glucose                   767 non-null    int64
 2   BloodPressure             767 non-null    int64
 3   SkinThickness             767 non-null    int64
 4   Insulin                   767 non-null    int64
 5   BMI                       767 non-null    float64
 6   DiabetesPedigreeFunction  767 non-null    float64
 7   Age                       767 non-null    int64
 8   Outcome                   767 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
None
```

This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics. We made a model to predict diabetes however the dataset was slightly imbalanced having around 500 classes labeled as 0 means negative means no diabetes and 268 labeled as 1 means positive means diabetic.

```
[66]
data.head()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

```
[67]
data.tail()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
762	9	89	62	0	0	22.5	0.142	33	0
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1

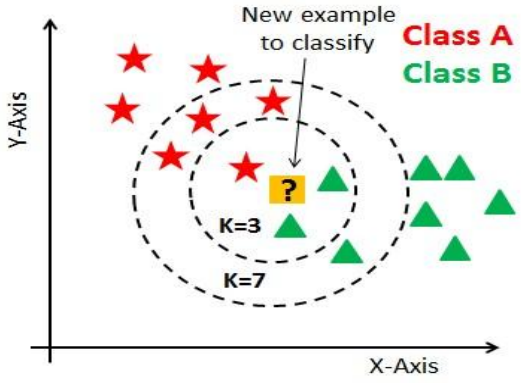
```
[68]
print(data.shape)  ### Return the shape of data
(767, 9)
```

When data has been ready, we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes. The methods applied on Pima Indians diabetes dataset. Main objective to apply Machine Learning Techniques to analyse the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in prediction.

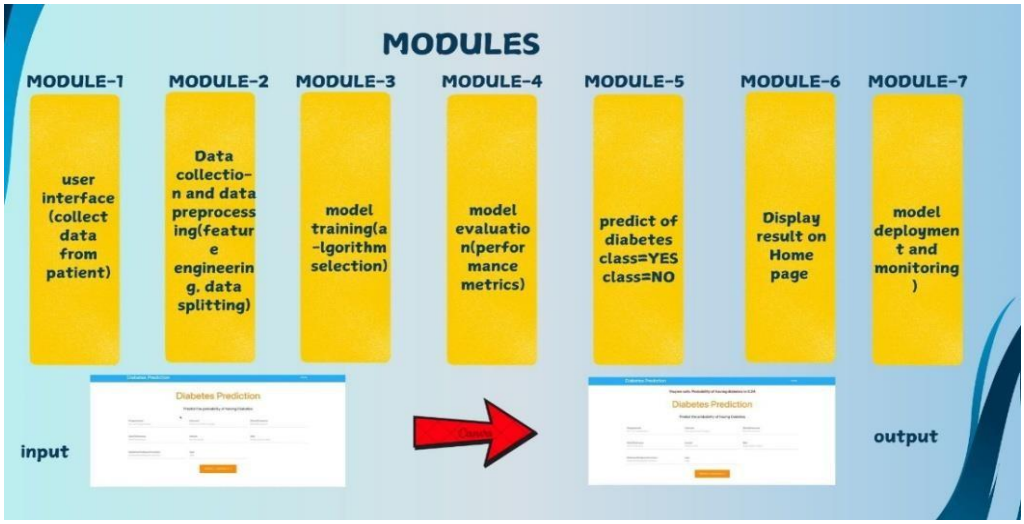
K-Nearest Neighbor- KNN is also a supervised machine learning algorithm. KNN helps to solve both the classification and regression problems. KNN is lazy prediction technique. KNN assumes that similar things are near to each other. Many times, data points which are similar are very near to each other. KNN helps to group new work based on similarity measure. KNN algorithm record all the records and classify them according to their similarity measure. For finding the distance between the points uses tree like structure. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set — it's nearest neighbors. Here K= Number of nearby neighbors, it's always a positive integer. Neighbor's value is chosen from set of class. Closeness is mainly defined in terms of Euclidean

ISSN: 2582 - 6379
Orange Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2025, Volume 6 Issue 1
www.ijisea.org

distance. The Euclidean distance between two points P and Q i.e. P (p1,p2, …. Pn) and Q (q1, q2,..qn) is defined by the following equation:



MODULES:



1. Data Collection and Preprocessing:

Data Sources: Gathered relevant data from various sources, such as medical records, patient surveys, and electronic health records.

Data Cleaning: Handle missing values, outliers, and inconsistencies in the dataset.

Feature Engineering: Creating new features or transform existing ones to improve model performance.

Data Splitting: Divide the dataset into training, validation, and testing sets.

2. Feature Selection:

Identify Relevant Features:

Determine which features (e.g., age, BMI, blood glucose levels, family history) are most predictive of diabetes. Use Techniques:

Employ feature selection techniques like information gain, chi-squared test, or recursive feature elimination.

3. Model Training:

Choose Algorithms: Select appropriate machine learning algorithms, such as:

Logistic Regression: A linear model for predicting binary outcomes.

Random Forest: An ensemble method that combines multiple decision trees.

Support Vector Machines (SVMs): A powerful algorithm for classification and regression.

ISSN: 2582 - 6379
Orange Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2025, Volume 6 Issue 1
www.ijisea.org

Train Models: Train the chosen models on the training data.
Hyperparameter Tuning: Optimize the model's parameters for better performance.

4. Model Evaluation:

Evaluate Performance:
Assess the model's accuracy, precision, recall, F1-score, and AUC-ROC on the validation and testing sets.
Compare Models:
Compare the performance of different models to identify the best-performing one.

5. Model Deployment and Monitoring:

Deploy Model: Integrate the trained model into a system for predicting diabetes risk.
Monitor Performance: Continuously monitor the model's performance and retrain it as needed.

This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms which are discussed above for diabetes prediction.
Step1: Import required libraries, Import diabetes dataset.
Step2: Pre-process data to remove missing data.
Step3: Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.
Step4: Select the machine learning algorithm Le. K- Nearest Neighbor
Step5: Build the classifier model for the mentioned machine learning algorithm based on training set.
Step6: Test the Classifier model for the mentioned machine learning algorithm based on test set
Step7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.
Step8: After analyzing based on various measures conclude the best performing algorithm.

IV. CONCLUSION

The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Method and Performance Analysis of that method and it has been achieved successfully. The proposed approach uses classification and ensemble learning method. High classification accuracy has been achieved. The Experimental results can be asst. health care to take early prediction and make early decision to cure diabetes and save humans life. Machine learning algorithms (Logistic regression, K-NN, Random Forest, Support vector machine) and clustering (k-means) are used to predict the diabetics disease in early stages.

REFERENCES:

[1] Strategies to make online teaching more effective Project by "Pooja Rathi" from Researchgate.in publication.
[2] Sci-kit learn Python Module
[3] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance's for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.
[4] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes "Proceeding of International Conference on Systems Compu-tation Automation and Networking, 2019.
[5] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Perfor-mance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 Feb-ruary, 2019.
[6] [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.

[8] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". Intemational Conference onElectrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

[9] 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

[10] NonsoNnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.