IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Diabetic Mellitus Prediction Using Machine Learning

[1] Dr . Padmaja Pulicherla [2] Ch Poojitha, [3] K Bhanu laxman, [4] K AbhiRam, [5] U Vyshnavi

[1] Professor and HOD, Hyderabad institute of technology and management, Medchal, Telangana

[2]UG student, Hyderabad institute of technology and management, Medchal, Telangana

[3]UG student, Hyderabad institute of technology and management, Medchal, Telangana

[4]UG student, Hyderabad institute of technology and management, Medchal, Telangana

[5]UG student, Hyderabad institute of technology and management, Medchal, Telangana

Abstract: Diabetes is a chronic, irreversible disease in which the glucose level in the blood is elevated. It is associated with potential health complications, such as cardiovascular, renal, dermatological, and ocular complications. Monitoring blood glucose levels is costly and time-consuming for patients. The work in this paper develops a predictive model using a decision tree classifier to predict diabetes in women and offers an early intervention and lifestyle modification solution that has the potential to reduce the risk of complications due to diabetes. This model was developed and validated using a Kaggle dataset, showing high accuracy for differentiating between diabetic and non-diabetic cases. The model generalizes well when trained on a large dataset, and thereby, based on just one sample, reliable individual predictions can be made. Such an approach will be a practical, accessible screening tool for the assessment of risk for diabetes benefiting high-risk populations as it opens up opportunities for timely preventive measures.

Index Terms – Diabetes Mellitus, Decision tree classifier, Random Forest, SVM, Logistic regression, confusion matrixbench,

I. INTRODUCTION

Diabetes Mellitus (DM) is a chronic metabolic disorder that has emerged as one of the most significant public health challenges worldwide. Characterized by persistent hyperglycemia, diabetes occurs when the body either fails to produce sufficient insulin or cannot effectively utilize the insulin it produces. Insulin, a hormone secreted by the pancreas, is essential for regulating blood glucose levels and ensuring energy supply to the body's cells. When this regulation is disrupted, glucose accumulates in the bloodstream, leading to various physiological imbalances and long-term complications.

The global burden of diabetes is alarming, with an ever-increasing prevalence attributed to modern lifestyle changes, urbanization, and the global obesity epidemic. India, in particular, bears a disproportionate share of this burden, earning it the title of the "Diabetes Capital of the World." In 2000, India had 31.7 million individuals diagnosed with diabetes, a figure that escalated to 77 million by 2020, underscoring the rapid progression of the disease in the region. Contributing factors include dietary transitions, reduced physical activity, and a lack of awareness about the early signs and risks associated with diabetes.

Diabetes Mellitus is broadly classified into four categories: Type 1 Diabetes, Type 2 Diabetes, Gestational Diabetes, and Prediabetes. Each type has distinct etiologies, risk factors, and disease trajectories. Type 1 diabetes, often diagnosed in children and adolescents, is an autoimmune condition where the immune system mistakenly attacks insulin-producing cells. Type 2 diabetes, the most common form, is primarily associated with lifestyle factors and insulin resistance. Prediabetes serves as a precursor to Type 2 diabetes, while gestational diabetes is a temporary condition occurring during pregnancy that poses risks to both mother and child.

If left untreated, diabetes can result in severe complications affecting multiple organ systems. These include cardiovascular diseases, kidney damage, nerve dysfunction, and vision impairment, all of which contribute to increased morbidity and mortality rates. Despite advancements in medical science, the growing prevalence of diabetes indicates a need for enhanced public health measures, early detection, and sustainable lifestyle interventions to curb this epidemic.

This paper delves into the different types of diabetes, their pathophysiological mechanisms, the complications they entail, and the urgent need for preventive strategies. By understanding the nuances of this complex disease, efforts can be directed towards reducing its prevalence and improving the quality of life for those affected.

II. LITERATURE SURVEY

IPrabhu P, Selvabharathi S [1] - In this paper they proposed a deep belief neural network model to predict diabetes mellitus. Pima Indian Women Diabetes Dataset was used. A fully connected Deep belief network is constructed, the dataset is trained and validated. Pretraining is done and fine-tuning of parameters is also done to achieve the best results. Different accuracy measures like recall, precision is also used. The proposed solution hasbetter accuracy compared with other traditional models like Naïve Bayes, Decision Tree, Logistic Regression.

J. Vijayashree, J. Jayashree [2] – In this paper they proposed an expert system for diagnosis of diabetes using deep neural networks. Pima Indian diabetes dataset was used. Data is preprocessed. Different feature selection methods like recursive feature elimination and principal component analysis. Best features are then selected and classification is done by two algorithms and are deep neural networks and artificial neural networks. Performance of themodels is evaluated using sensitivity, specificity and ROC curve. Deep neural networksperformed well than artificial neural network.

A. Mary Posonia, S. Vigneshwari, D. Jamuna Rani [3]- gives a study on gestational diabetes using the decision tree J48 algorithm. This algorithm has good performance hence it is Chosen. Pre-processing was done on the data, and a classifier is used on it. The data was trained and tested. Performance is measured, and this algorithm gave better accuracy compared to others in predicting gestational diabetes.

Santi Wulan Purnami, Abdullah Embong, Jasni Mohd Zain, and 1 S.P. Rahayu [4] discussed smooth support vector machine (SSVM) and Multiple Knots Spline-SSVM and their applications of it in diabetes diagnosis. MKS-SSVM has a new smooth function that is a multiple knot smooth function. Two approaches were used and classified. MKS-SSVM used optimal parameters using uniform design and 10fold cross-validation to select the best feature. This is solved with the Newton Armijo algorithm and then a separating plane is achieved. The model evaluated the outcomes for new input values with good accuracy. MKSSSVM has a significant rise in its training accuracy and hence has better accuracy than SSVM.

Yashi Srivastava, Pooja Khanna, and Sachin Kumar's [6] – In their paper, they presented a model with two algorithms – two-class Logistic regression and a two-class boosted decision tree which use glucose levels, pregnancies, and blood pressure as the main factors to detect diabetes. These classification models are created using Microsoft Azure AI services. In view of ROC curve, Logistic regression classification was preserved and it is applied to cross-validate the model to predict the diabetes. The model they presented was crossed validated, and two algorithms were perfectly capable of analyzing and predicting the output. Results itself proved this model can work better than the other models presented.

Suyash Srivastava, Lokesh Sharma, Vijeta Sharma, Ajai Kumar, and Hemant Darbari [8] – presented the discussion related to the prediction of the diabetes using ANN. Data is pre-processed and ANN is applied. ANN It consists of the input layer, hidden layers, and output layer where data is processed in hidden layers. There are random weights, feedforward and back propagation, and application of sigmoid function to the obtained data. The error is calculated and accuracy is measured. In this case, better accuracy was obtained.

III. EXISTING SYSTEM

Previously many researchers implemented this project using many different approaches like machine learning, data mining, and deep learning. The models were trained and tested on the Pima dataset that is available on Kaggle.

3.1.1 Limitations

- The existing approaches did not have the promising performance of the model.

IV. PROPOSED SYSTEM

In this project, we used four classification supervised models to predict the disease with a good dataset that consists of 2000 records of positive and negative samples.

3.2.1 Advantages

- Used optimization techniques like feature selection for model performance.
- High Accuracy

METHODOLOGY

1. Data Collection and Loading

The first step in our methodology involves collecting and loading the diabetes dataset. The dataset used in this study was sourced from [Kaggle or specify the source if applicable]. Once collected, the data is loaded into the chosen software environment for analysis and preprocessing.

2. Data Analysis

After loading the data, the next step is data analysis. This stage involves importing all necessary libraries and reading the dataset to understand its structure. Information such as the number of attributes (features) and the distribution of samples (positive and negative cases) is explored. This preliminary analysis helps us understand how the data is organized and the proportions of diabetic (positive) versus non-diabetic (negative) samples. Data visualization techniques are also employed to gain insights into relationships between features and to detect any initial patterns that might inform later steps in model training.

3. Data Preprocessing

Data preprocessing is critical for handling inconsistencies in the dataset, such as missing values. Missing data can adversely affect machine learning model performance, so it must be handled appropriately. In our dataset, missing values were represented as zeros, which do not align with the actual physiological ranges for certain features. To address this, we examined each attribute (column) for null values, replacing them with either the mean or median of that attribute. This approach helps create a complete and consistent dataset, improving the model's reliability by ensuring that each data point is adequately represented.

4. Feature Selection

Feature selection is essential to enhance model accuracy and performance by removing irrelevant or redundant data. This technique allows us to isolate the most informative features that have the greatest impact on predicting diabetes. For this study, we applied the chi-square test method to identify and select the best subset of features. The chi-square test evaluates the association between categorical features and the target variable, helping us determine which features are most relevant for classification. Using this method, we selected four

key features that best support the model's predictive accuracy: **Glucose levels**, **Insulin**, **Age**, and **Body Mass Index (BMI)**.

The chi-square test formula is as follows:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Formula for chi square test
where O - observed values
E - expected values

5. Train and Test Model

Once the data is preprocessed and essential features are selected, the dataset is divided into training and testing subsets. The dataset is split into 70% training data and 30% testing data, ensuring that the model is trained on a substantial portion of the data while reserving enough data for unbiased evaluation of the model's performance.

During training, the selected classifiers (decision tree, random forest, support vector machine, and logistic regression) are used to build the predictive model. Each classifier provides probabilistic outcomes in terms of binary classification: 1 for diabetic patients and 0 for non-diabetic cases. The model is then tested on the reserved test dataset, allowing us to evaluate its accuracy, precision, recall, and other performance metrics.

ALGORITHMS USED

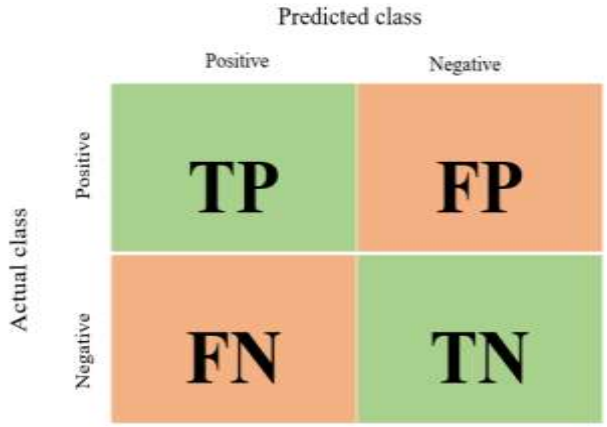
1. **Decision tree classifier:**
   Decision tree classifier is a popular supervised learning approach used for classification. Decision tree is a hierarchical structure which has nodes and branches. Each internal node consists the attribute or variable and the branches mean the decision rule that predicts yes or no and finally the leaf node will have the final output. According to the features the tree further splits and gives final outcome at the leaf node. To develop a tree in this classifier first root node has to be chosen. Picking the root node is the most challenging part. The process of picking the root node is the attribute selection. It can be done by two measures and are information gain and Gini index. a. Information gain identifies the root node and further splits the tree. It does this by measuring the attribute how useful it is. It uses entropy to get root node. b. Gini index is other method to determine the root node. Gini index is that which measures the impurity of an attribute. That is it measures about how frequently a randomly picked attribute would be mistakenly recognized. Hence, root node attribute is picked up which has least Gini index value. Gini Formula Using Gini, the tree is then formed and is capable of predicting whether a person has diabetes or not.
2. **Logistic regression:**
   Logistic regression is also a supervised approach in machine learning. It is a statistical analysis mostly used for binary classification problems that means it has only two kinds of classes like yes or no. 0 and 1. It is like predicting the probability in between classes 0 and 1. Furthermore, it can be used on any kind of data like discrete data, continuous data etc. Logistic regression works based on some assumptions. This algorithm is based on logistic function which is also called sigmoid function. This algorithm used a decision boundary to classify the output into classes. The input is passed to the logistic or sigmoid activation function and is processed. The output is based on this sigmoid function's output as if value given by sigmoid function is greater than 0.5 it takes positive that is 1 as output else 0 as output.
3. **Random forest:**
   Random forest is additionally a supervised approach for classification and regression problems. It is based on ensemble learning which suggests many classifiers are grouped together so as to enhance the accuracy of the model. It's the group of trees made up of subsets of the dataset. Based on the prediction of each tree, checks for majority of votes for predictions and based on this final output is predicted. Hence the model performance is increased. First it selects the subsets from dataset randomly then trees are formed. Each tree's output is predicted and stored. Finally based on this the final output is obtained. It repeats this step for different subsets of dataset with different subtrees.

4. **Support vector machine:**

A supervised learning algorithm that can be used for classification and regression. Within the SVM calculation, we plot every datum as a degree in ndimensional space (where n is various attributes you have) with the value of each attribute being the value of a specific coordinate. Then, classification is performed by observing the hyper-plane that separates the two classes quite well. If there are two attributes hyperplane is a line and if there are three then hyperplane is two dimensional. This becomes complex as the number of attributes increase. Hyperplane is a sort of decision boundary where it can segregate between two classes so drawing the hyperplane could be a challenging part in this algorithm. The data points that are nearer to hyperplane are called support vectors and help in the inclination of the hyperplane. The output of new data points is classified with the help of hyperplane.

5. **Finding accuracy:**

The test data is evaluated using accuracy measures to check how effectively the model is performed. Different algorithms accuracy is measured and then comparative analysis is done between the algorithms used. Accuracy is measured using the confusion matrix for each algorithm.



Confusion matrix is a table that is used to determine how well a model is performed. True positive, true negative, false positive and false negative are the terms used in it. After the analysis based on confusion matrix Decision tree classifier has given the better accuracy amongst other algorithms.

	True Positive	True Negative	False Positive	False Negative
Decision tree classifier	388	189	9	14
Logistic regression	348	111	49	92
Random forest	350	153	47	50
Support vector machine	358	100	39	103

Table: Comparison of Confusion matrix values for different classifiers

Accuracy can be found by formula

Accuracy= (No. of true positives+ No. of true negatives) / Total samples

Accuracies for different classifiers is as

: Decision tree classifier: 96.1%

Logistic regression: 76.5%

Random forest: 83.8%

Support vector machine: 76.3%

Precision, recall and f-score are the other accuracy measures for classification problems

. Precision= (No. of true positives) / (No. of true positives + No. of false positives)

Recall= (No. of true positives) / (No. of true positives+ No. of false negatives)

F-measure= (2*precision*recall) / (precision+ recall)

	Precision	Recall	F-measure
Decision tree classifier	0.97	0.96	0.96
Logistic regression	0.87	0.79	0.82
Random forest	0.88	0.87	0.87
Support vector machine	0.90	0.77	0.82

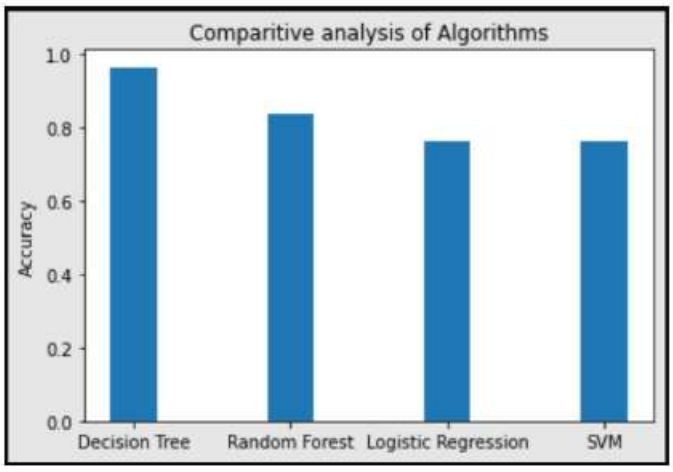
Table: Comparison of precision, recall and f-measure for different classifiers

Results and discussions:

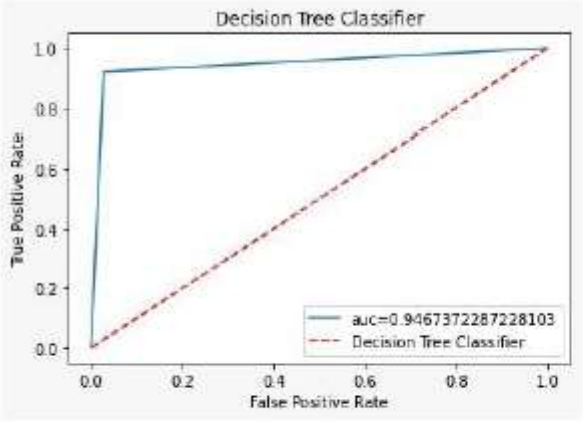
Different classifiers were tested for performance, accuracy scores representing those are as shown on Figure [X]. The highest accuracy is represented by the decision tree classifier shown on Figure [X], which was recorded as 96.1%. The Random Forest classifier came second at 83.8%. Most probably, these high performances are a result of steps in the optimization form: data preprocessing and feature selection steps in ensuring that only a clean and relevant dataset was used to train.

We also compared ROC curves across the classifiers confirming superiority of Decision Tree in the distinction ability between diabetic and not diabetic cases.

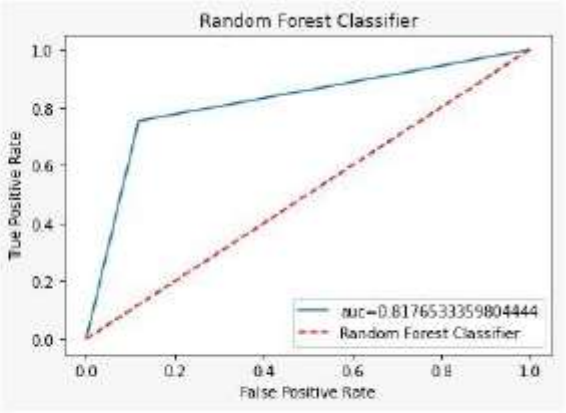
Above results can be considered an important indicator of proper data preparation which significantly enhances model accuracy, especially with Decision Tree and Random Forest classifiers



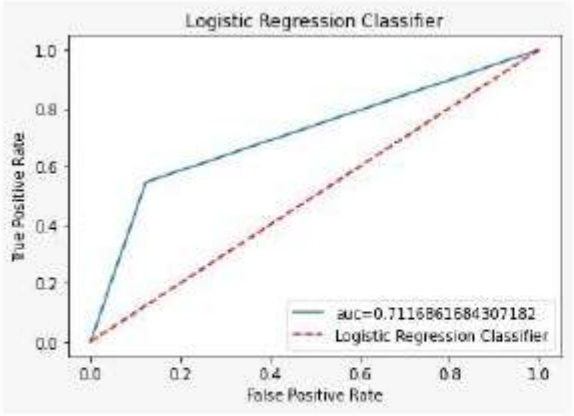
Accuracy comparison



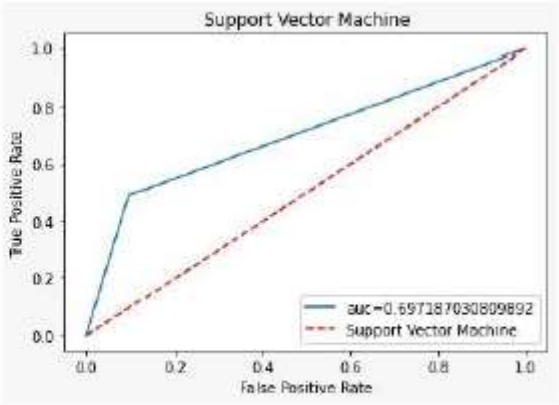
ROC for decision tree



ROC for random forest



ROC for Logistic regression



ROC for Support vector machine

CONCLUSION AND FUTURE WORK

As diabetes is a long-term chronic disease it must be detected in its early stages and should be prevented. In this paper, we used different machine learning models to predict diabetes in women. From our experimental analysis decision tree classifier has given the best accuracy. The accuracy was measured with the help of a confusion matrix and compared. The prediction of diabetes using this model helps in time savings such that one can use this model to predict diabetes instead of moving to the hospital and taking tests and waiting for reports. We would like to extend this work in the future in such a way that optimization methods like tuning the parameters, normalizing data and other techniques are used to improve the accuracy of the other models that gave less accuracy and also work on other datasets to improve performance of all the models

IX. ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Mrs. krishna Jyothi.k , Professor and HOD at HITAM, for her constant guidance and valuable assistance throughout the development of the DIABETIC PREDICTION USING MACHINE LEARNING project. I would also like to thank Dr.Padmaja Pulicherla, Assistant Professor at HITAM for his guidance. Finally, I would like to express my gratitude to the Hyderabad Institute of Technology and Management for providing me with the opportunity to complete this project

REFERENCES

1. Prabhu P, Selvabharathi S "Deep Belief Neural Network Model for Prediction of Diabetes Mellitus" 3rd International Conference on Imaging, Signal Processing and Communication, 19 December 2019
2. J. Vijayashree, J. Jayashree "An expert system for the diagnosis of diabetic patients using deep neural networks and recursive feature elimination" International Journal of Civil Engineering and Technology (IJCIET) Volume 8, 12 December 2017
3. A. Mary Posonia, S. Vigneshwari, D. Jamuna Rani "Machine Learning based Diabetes Prediction using Decision Tree J48" Third International Conference on Intelligent Sustainable Systems [ICISS 2020]
4. Santi Wulan Purnami, Abdullah Embong, Jasni Mohd Zain and 1 S.P. Rahayu "A New Smooth Support Vector Machine and Its Applications in Diabetes Disease Diagnosis" Journal of Computer Science (Science publications) ,2009
5. PHILIP H. SWAIN and HANS HAUSKA "The Decision Tree Classifier: Design and Potential" IEEE TRANSACTIONS ON GEOSCIENCE ELECTRONICS, VOL. GE-IS, NO. 3, JULY 1977
6. Yashi Srivastava, Pooja Khanna, Sachin Kumar "Estimation of Gestational Diabetes Mellitus using Azure AI Services" IEEE Amity International Conference on Artificial Intelligence 29 April 2019
7. Deepak Gupta1 & Ambika Choudhury 1 & Umesh Gupta 1 & Priyanka Singh2 & Mukesh Prasad "Computational approach to clinical diagnosis of diabetes disease: a comparative study" ACM Digital library- Multimedia Tools and Applications Volume 80, 04 January 2021
8. Suyash Srivastava, Lokesh Sharma, Vijeta Sharma, Ajai Kumar, Hemant Darbari "Prediction of Diabetes using Artificial Neural Network approach" Springer Engineering vibration, communication and information processing 2019.