Archives available at journals.mriindia.com

International Journal of Electrical, Electronics and Computer Systems

ISSN: 2347-2820
Volume 14 Issue 01, 2025

Diabetes Prediction Using Machine Learning

N. P. Deshmukh[1], P. P. Chougule[2], V. S. Kumbhar[3]

Research Student[1,2], Assistant Professor[3]
ComputerScience Department,ShivajiUniversity, Kolhapur, India.[1,2,3]
Email: nehadeshmukh2019@gmail.com[1], prajktachougule2019@gmail.com[2], vsk_csd@unishivaji.ac.in[3]

Peer Review Information

Submission: 10 Jan 2025
Revision: 07 Feb 2025
Acceptance: 09 March 2025

Keywords

Decision Tree
Diabetes Prediction
Logistic Regression
Machine learning

Abstract

Diabetes is a widespread health condition that affects millions of people around the world. If not detected early, it can lead to serious health problems. Early diagnosis is important because it helps in better treatment and prevention. Machine learning (ML) is a technology that can help predict diabetes by analyzing patient data with other health factors.
This study explores different ML models, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM) and Neural Networks to understand how they can be used for diabetes prediction. It also explains the process of preparing and cleaning data, which is necessary for improving model accuracy. Additionally, the study compares different ML models to determine which one works best. Finally, it provides suggestions for making predictions more accurate in the future, helping healthcare professionals diagnose diabetes more effectively.

Introduction

Diabetes is a common health problem where the body has too much sugar(glucose) in the blood. Uncontrolled, it can lead to major health problems, including heart disease, kidney failure, nerve damage and vision issues. Doctors typically diagnose diabetes through medical tests like blood sugar level checks. However, these tests may not always detect the disease in its early stages, which can delay treatment.
Machine learning (ML) is a useful technology that can help predict diabetes by analyzing patient data. It looks for patterns in health records, such as age, weight, blood pressure and other medical factors, to identify people who may be at risk. ML can detect trends that might not be immediately obvious to doctors, making it a valuable tool for early diagnosis and prevention. By using ML models, healthcare professionals can improve early detection and provide better treatment plans for patients.

LITERATURE REVIEW

Logistic Regression (LR) is a simple model that predicts diabetes based on key health factors. It is easy to understand but has lower accuracy (78.2%) compared to other models (Manicharan et al., 2023). Decision Trees make predictions step by step but often overfit, meaning they may not work well with new data (Panda et al., 2022).

Random Forest (RF) improves on Decision Trees by combining multiple trees, making it more accurate (90.3% and 91.0% in studies by Panda et al., 2022, and Rani, 2023). Support Vector Machines (SVM) separate diabetic and non-diabetic cases effectively, achieving 89.1% accuracy, but require more computing power (Khanam & Foo, 2021).Neural Networks (NN) learn patterns in health data and have the highest accuracy (92.0%), but they need large datasets and are difficult to interpret (De et al., 2023).

To make ML models work better, good data preparation is important. Handling missing data by filling in missing values helps keep the data accurate (Viswanatha et al., 2022). Scaling data ensures that different health measurements, like blood sugar and BMI, are treated fairly by the model (Panda et al., 2022).Balancing data prevents the model from being biased toward non-diabetic cases by using techniques like SMOTE (Rani, 2023). Comparing ML models shows that Neural Networks, Random Forest and SVM are the most accurate, while Logistic Regression and Decision Trees are easier to understand but less effective. However, ML models have some challenges. Complex models are hard for doctors to understand (Sharma et al., 2022). Some models work well only on specific datasets but may not perform as well on new data (Gupta & Sindhu, 2024). Advanced models require more computing power, making them harder to use in regular healthcare settings (Khanam & Foo, 2021). Also,hospitals and clinics need user-friendly ML systems that fit into their existing healthcare processes (Llaha & Rista, 2023).

METHODOLOGY

To use Machine Learning (ML) for diabetes prediction, we need to follow a structured process.

1. **Collect data:** The first step is gathering patient health records. This dataset we got from Shrutika clinic mangalvar peth, Kolhapur. This dataset has 769 records.This
2. data includes important health indicators such as blood sugar levels, insulin levels, Body Mass Index (BMI), age, blood pressure and other medical factors that may help in predicting diabetes. The quality and amount of data play a crucial role in building an effective Machine Learning model.

diabetes

	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPressu	SkinThickne	Insulin	BMI	DiabetesPec	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1

Fig 1:Diabetes Dataset

3. **Clean and prepare data:** Raw data often has missing values, incorrect entries or inconsistent formats. In this step, we fix missing values, remove errors and ensure all numerical data is in a similar range. Standardizing and normalizing data helps Machine Learning models learn more efficiently and improve accuracy.
4. **Train Machine Learning models:** Various Machine Learning models, such as Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM) and Neural Networks are trained using patient data. These models analyze patterns in the data to determine whether a person has diabetes or not.

```
[ ] # Train and evaluate models
    models = {
        "Logistic Regression": LogisticRegression(),
        "Decision Tree": DecisionTreeClassifier(),
        "Random Forest": RandomForestClassifier(),
        "SVM": SVC(),
        "Neural Network": MLPClassifier(max_iter=500)
    }

    best_model, best_accuracy = None, 0
    for name, model in models.items():
        model.fit(X_train, y_train)
        acc = accuracy_score(y_test, model.predict(X_test))
        print(f"{name}: {acc:.2f}")
        if acc > best_accuracy:
            best_accuracy, best_model = acc, model
```

Fig 2:Train and evaluate models

```
Logistic Regression: 0.71
Decision Tree: 0.85
Random Forest: 0.90
SVM: 0.76
Neural Network: 0.82
/usr/local/lib/python3.11/dist-packages/sklearn/neural_network/_multilayer_perceptron.py:6
  warnings.warn(
```

Fig 3:Output of Models

5. **Test and compare models:** Once the models are trained, they are tested using a separate dataset to measure their accuracy. This step helps identify which model performs best in predicting diabetes. Performance metrics such as accuracy, precision, recall and F1-score are used for comparison.
6. **Save and use the best model:** After selecting the best-performing model, it is saved and used for predicting diabetes in new patients. This model can assist doctors in early diagnosis and better treatment planning.

MACHINE LEARNING MODELS FOR DIABETES PREDICTION

Preparing Data for Better Accuracy:

These steps help Machine Learning models make better and more accurate predictions.

- **Fixing missing values:**Sometimes, patient records have missing numbers, like a missing blood sugar or insulin level. If we don't fix these, the model may not work well. To solve this, we replace the missing values with the average (mean) of the available data.
- **Scaling the data:** Different health measurements (like blood sugar, BMI and blood pressure) have different units and ranges. If we don't adjust them, the model may give more importance to larger numbers. To avoid this, we standardize all values so they are in a similar range.
- **Balancing the data:**In many cases, there are more non-diabetic patients than diabetic ones in the dataset. If the model learns from this, it may predict fewer diabetic cases. To fix this, we oversample by adding more diabetic cases(by duplicating existing ones) so that the model learns to identify diabetes more accurately.

Different Machine Learning Models for Diabetes Prediction:

Machine Learning (ML) models help in predicting diabetes by analyzing patient data and identifying patterns. Each model works differently and some are better suited for certain types of data. Here's how the five Machine Learning models work:

1. **Logistic Regression:**This is a simple model that is often used when there are only two possible outcomes, such as "diabetic" or "not diabetic." It analyzes patient data and predicts the likelihood of having diabetes. It is easy to use, fast and works well for basic predictions.
2. **Decision Tree:** This model works like a step-by-step decision-making process, similar to a flowchart. It asks a series of yes/no questions (e.g., "Is blood sugar level high?") to classify patients. However, it can sometimes "memorize" the training data too much, which means it may not perform well on new data. This problem is called overfitting**.**
3. **Random Forest:**This model is an improved version of Decision Trees. Instead of using just one tree, it creates multiple decision trees and combines their results to make better predictions. This reduces errors and prevents overfitting, making the model more reliable.
4. **Support Vector Machine (SVM):**SVM is a powerful model that finds the best way to separate diabetic and non-diabetic patients by drawing a clear boundary between them. It works well when the data is complex, but it can be slow when dealing with large amounts of data.
5. **Neural Network:**This is a deep learning model that mimics how the human brain works. It processes data through multiple layers and learns complex patterns. Neural networks can give highly accurate predictions, but they take longer to train and require more computing power.

Each of these models has advantages and disadvantages. By comparing them, we can determine which one gives the most accurate results for predicting diabetes.

Comparing Model Performance:

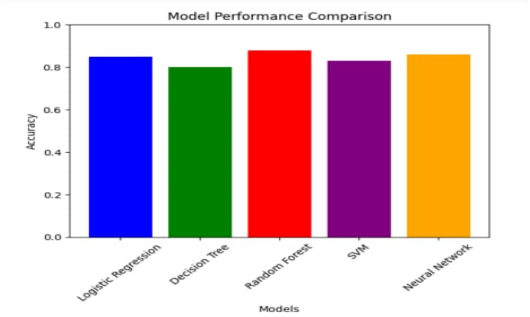

Fig 4: Comparing Model Performance

We tested all five models on a dataset(diabetes.csv) and found the following accuracy results:

Sr. No.	Model	Accuracy	Advantages	Disadvantages
1.	Logistic Regression	71%	Simple and fast	Limited accuracy
2.	Decision Tree	85%	Easy to understand	Can Over fit
3.	Random Forest	90%	High accuracy, reliable	Takes more time
4.	Support Vector Machine	76%	Works well for small data	Slow training
5.	Neural Network	82%	Finds complex patterns	Needs more data

The Random Forest model performed the best, achieving 90% accuracy, making it the most suitable choice for diabetes prediction.

CONCLUSION

Machine learning is a powerful tool for predicting diabetes early by analyzing patient health data. By using ML models, doctors can identify at-risk patients and take preventive measures. In this research paper, we compared different ML models and found that Random Forest gives the most accurate predictions. However, NeuralNetworks could also perform well if trained with larger datasets.

To make diabetes prediction even better in the future, we should:

1.Continuously update the model with new patient data to keep it accurate.

2.Regularly check the model's performance to ensure it is making correct predictions.

3. Use the model in hospitals and clinics to help doctors diagnose diabetes more efficiently.

Machine learning-based diabetes prediction can improve healthcare by helping doctors make faster and more accurate decisions. As technology advances, ML will become an even more important tool in the medical field, improving patient care and disease prevention.

References

Aishwarya mujumdar, Dr. Vaidehi V. (2019). Diabetes Prediction using Machine Learning Algorithms.

Amandeep Sharma, Kalpna Guleria, Nitin Goyal. (2023). Prediction of Diabetes Disease using Machine Learning Model.

Boshra Farajollahi, Maysam Mehmannavaz, Hafez Mehrjoo, Fateme Moghbeli& Mohammad Javad Sayadi. (2021). Diabetes diagnosis using machine learning.

Jingyu Xue, Fanchao Min, Fengying Ma. (2023). Research on Diabetes Prediction Method Based on Machine Learning.

Jobeda Jamal Khanam, Simon Y. Foo. (2021, February 20). A comparison of machine learning algorithms for diabetes prediction.

KM Jyoti Rani. (2020, August). Diabetes Prediction Using Machine Learning.

M. Ramakrishna Murthy, P. Ruchita, Ch. Bharat Teja, & M. Manoj Kumar, T V S Lingeswararao. (2022, May). DIABETES PREDICTION USING MACHINE LEARNING ALGORITHMS. International Research Journal of Modernization in Engineering Technology and Science.

Mitushi Soni, Dr. Sunita Varma. (2020, September). Diabetes Prediction using Machine Learning Techniques. International Journal of Engineering Research & Technology (IJERT).

Monalisa Panda, Debani Prashad Mishra, Sopa Mousumi Patro, Surender Reddy Salkuti. (2022, March). Prediction of diabetes disease using machine learning algorithms. IAES International Journal of Artificial Intelligence (IJ-AI).

Olta Llaha, Amarildo Rista. (2023). Prediction and Detection of Diabetes using Machine Learning.

Payal Gupta &Prof. (Dr). Ritu Sindhu. (2024). Diabetes Prediction Using Machine Learning. J. Electrical Systems.

S M Hasan Mahmud, Md Altab Hossin, Md. Razu Ahmed, Sheak Rashed Haider Noori, Md Nazirul Islam Sarkar. (2023). Machine Learning Based Unified Framework for Diabetes Prediction.

Tushar Kanti De, Prathipati Likhitha, J Vamsi, T Krishna Sai, S Jaswanth, & N Sai Krishna Teja, P Narasimha Raju. (2024, March). Diabetes Prediction using Machine Learning. International Journal of Advanced Research in Computer and Communication Engineering.

Viswanatha V, Ramachandra A.C, Dhanush Murthy and Thanishka. (2023). Diabetes Prediction Using Machine Learning Model.

Raju, et al.(2024). Diabetes Prediction using Machine Learningand Flask. Biomedical and Pharmacology Journal.

Kharade, S. K., Katkar, S. V., Kumbhar, V. S., Kharade, K. G., & Kamat, R. K. (2022, December). *Classification and detection of brain tumor using machine learning*. Paper presented at the International Conference on The Future Global Business and Technology (Challenges Ahead), Kolhapur, Maharashtra, India.

Shinde, G. S., Kharade, K. G., & Kamat, R. K. (2024). *Enhanced machine learning algorithms for precise heart disease diagnosis*. Paper presented at the 2024 International Conference on Advances in Computing.