Diabetes Prediction

Manish Kumar Sahu(211000030)

Computer Science & Engineering

International Institute Of Information Technology

Naya Raipur, India

manish21100@iiitnr.edu.in

Ritesh Kumar(211000046)

Computer Science & Engineering

International Institute Of Information Technology

Naya Raipur, India

ritesh21100@iiitnr.edu.in

Abstract - In this project, we aim to build a logistic regression model to predict whether a patient has diabetes or not based on certain diagnostic measurements included in the dataset. Diabetes is a prevalent disease worldwide, and early detection can help prevent its progression and avoid complications. We explore various techniques to enhance the performance and accuracy of the model, including data preprocessing and feature selection.

To evaluate the performance of the model, we use a confusion matrix, which provides a detailed breakdown of the model's predictions and allows us to calculate metrics such as precision, recall. The confusion matrix also helps us visualize the model's performance and identify any imbalances in the dataset that may affect the model's accuracy. Overall, our results show that the logistic regression model can accurately predict whether a patient has diabetes or not.

INTRODUCTION - Diabetes is a chronic metabolic disorder that affects a significant portion of the population worldwide. It is a complex disease that requires careful management to prevent or delay its long-term complications. Early detection and intervention are essential to improve patient outcomes and quality of life. Machine learning algorithms have emerged as powerful tools to aid in the early detection and diagnosis of diabetes based on clinical and demographic data.

We use logistic regression, a popular binary classification algorithm, to predict the likelihood of diabetes based on the demographic and clinical data. We also use a confusion matrix to evaluate the performance of the model and identify areas of improvement. Additionally, we investigate the impact of data preprocessing techniques such as imputation, feature selection, and feature scaling on the performance of the logistic regression model.Significant clinical implications by providing a useful tool for early detection and management of diabetes.

Moreover, our study contributes to the growing body of research on the use of machine learning algorithms in clinical analysis and highlights the importance of data preprocessing techniques in model performance.

Methodology

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

Dataset Description

The diabetes data set was originated from https://www.kaggle.com/datasets/mathchi/diabetes-data-set

	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

Table 1. Dataset and its Features.

➔ The diabetes data set consists of 768 data points, with 9 features each.

➔ “Outcome” is the feature we are going to predict, 0 means No diabetes, 1 means diabetes.

Proposed Model Diagram

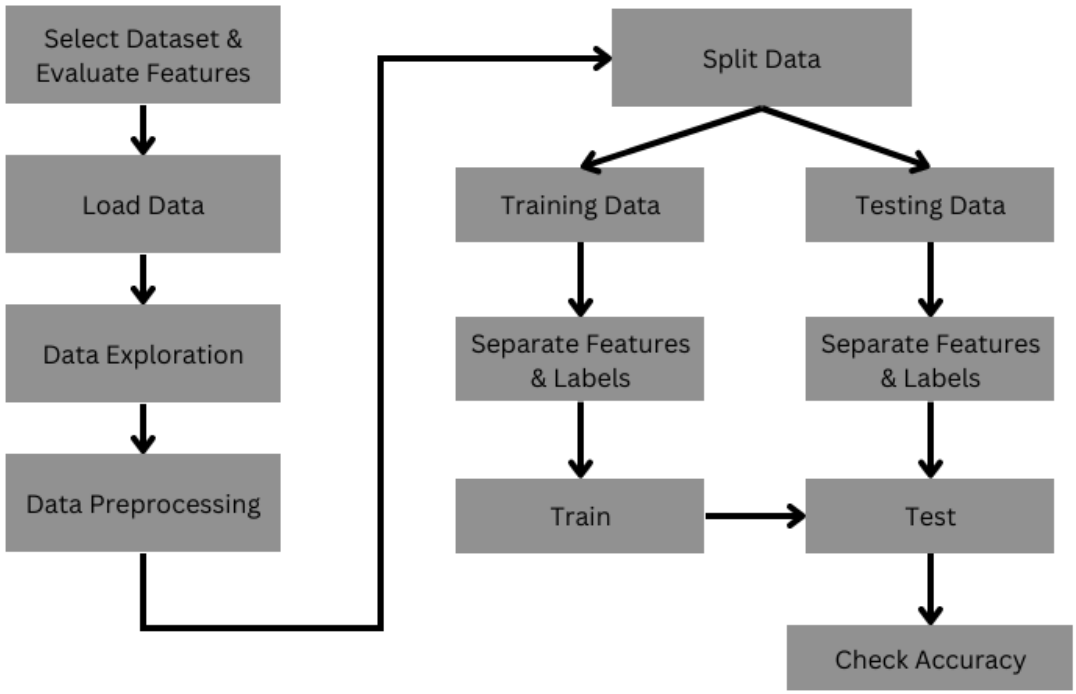

Fig. 1. Flowchart of the Diabetes Prediction Model.

Load Data

The data, which is in the CSV format, is loaded to a variable. There are 768 data points in Dataset

Data Exploration

In this we have examined and analyzed datasets to extract meaningful insights and patterns. It involves using statistical and visualization techniques to understand the data, identify relationships between variables, and gain an overall understanding of the data set.

Data pre-processing

It is necessary to locate in the dataset any predictor variables with missing values, null values, or values of zero. Except for a few features, like the "Pregnancies" feature in Dataset, none of the predictor variables or features can have a zero value. The column's mean values should be used to replace these values. As inaccurate values increase the likelihood of inaccurate predictions, this is a crucial step in improving prediction accuracy.

Training and testing

The data is split into training and testing sets. The common split ratios are 80:20 and 70:30. In this project, the data is split in the ratio of 70:30, i.e. 70% for training and 30% for testing. A Logistic Regression algorithm is used to make the predictions and check for the accuracy. The execution time is also calculated.

Confusion matrix

confusion matrix visualizes and summarizes the performance of a classification algorithm. There are four important terms: - True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). TP and TN represent the cases when the actual outcome and the result are the same, whereas FP and FN are the cases when the opposite results are obtained. A classification report is generated which includes Precision, Recall, and Support. The Precision metric shows what percent of predictions are correct. Recall describes what percent of positives are correctly identified. The F1 score is the percent of positive predictions that are correct. Support is the count of actual occurrences of the class in the specified dataset

Precision = TP/(TP + FP)
Accuracy=(TP+TN)/(TP+TN+FN+FP)
Recall = TP/(TP+FN)
F1 Score = 2*(Recall * Precision) / (Recall + Precision)

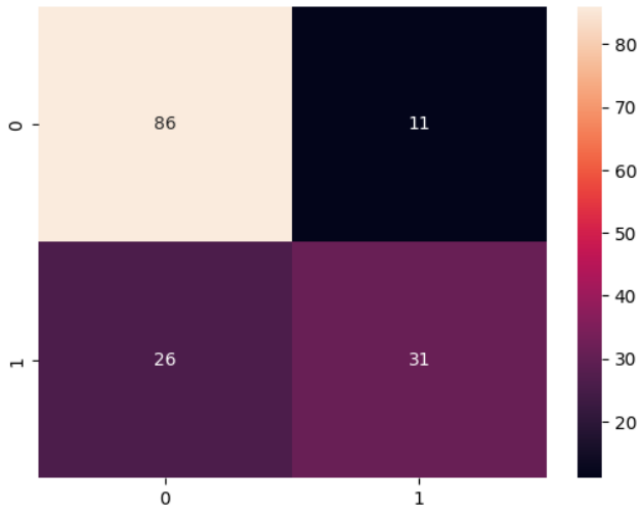

Fig 2.Confusion Matrix

	precision	recall	f1-score	support
0	0.77	0.89	0.82	97
1	0.74	0.54	0.63	57
accuracy			0.76	154
macro avg	0.75	0.72	0.72	154
weighted avg	0.76	0.76	0.75	154

Table 2. Classification report after Feature Selection.

CONCLUSION

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of diabetes. During this work, a logistic regression machine model and Confusion matrix were used and experimented Experimental results determined the adequacy of the designed system with an achieved accuracy of 75.97% Logistic regression model. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.

FUTURE SCOPE

There are many potential future directions for this project, and further research could help improve the accuracy and applicability of the developed model in real-world settings.

Other machine learning algorithms: In this project, we used logistic regression for diabetes prediction. However, there are many other machine learning algorithms that can be used for this task, such as decision trees, random forests, support vector machines, and neural networks. Future work could explore these algorithms and compare their performance with logistic regression.

Larger datasets: In this project, we used a relatively small dataset. Gathering more data could help improve the model's accuracy and generalizability.

Real-time prediction: In this project, we used a dataset of historical data to predict diabetes. However, future work could involve developing a real-time prediction system that can predict diabetes in patients as new data becomes available.

Clinical decision support: The developed model can be integrated with clinical decision support systems to help healthcare professionals in diagnosing diabetes in patients.

REFERENCES

1. **World Health Organization, 2021 https://www.who.int/news-room/fact-sheets/detail/diabetes.**

2. **https://www.researchgate.net/profile/Pramila-Chawan/publication/326416823_LOGISTIC_REGRESSION_AND_SVM_BASED_DIABETES_PREDICTION_SYSTEM/links/5b4c80ddaca272c60947858a/LOGISTIC-REGRESSION-AND-SVM-BASED-DIABETES-PREDICTION-SYSTEM.pdf**

3. **Dataset from Vanderbilt, 2021 https://data.world/informatics-edu/diabetes-prediction. Google Scholar**

4. **Diabetes prediction using machine learning algorithms International Conference on Recent Trends in Advanced Computing", 2019, ICRTAC (2019)**