# Diabetes Detection

Aadi Krishna Vikram (211020402)
DSAI
International Institute of Information Technology
Naya Raipur, India
aadi21102@iiitnr.edu.in

Ashish Agrawal (211020414)
DSAI
International Institute of Information Technology
Naya Raipur, India
ashish21102@iiitnr.edu.in

**Abstract -** In order to determine if a patient has diabetes or not, we want to develop a logistic regression model using the dataset's diagnostic values. Diabetes is a common condition that can be prevented from progressing and causing problems by early detection. We investigate several methods, including as feature selection and data cleaning, to improve the model's performance and accuracy.

Using a confusion matrix, which gives a thorough breakdown of the model's predictions and enables us to compute metrics like precision and recall, we can assess the model's performance. The confusion matrix also aids in the visualisation of model performance and the detection of dataset imbalances that might impair model accuracy. Overall, our findings indicate that the logistic regression model can accurately detect 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 pre-processing techniques such as imputation, feature selection 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 pre-processing 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/mathc hi/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 |
| 5 | 5 | 116 | 74 | 0 | 0 | 25.6 | 0.201 | 30 | 0 |
| 6 | 3 | 78 | 50 | 32 | 88 | 31.0 | 0.248 | 26 | 1 |
| 7 | 10 | 115 | 0 | 0 | 0 | 35.3 | 0.134 | 29 | 0 |
| 8 | 2 | 197 | 70 | 45 | 543 | 30.5 | 0.158 | 53 | 1 |
| 9 | 8 | 125 | 96 | 0 | 0 | 0.0 | 0.232 | 54 | 1 |
| 10 | 4 | 110 | 92 | 0 | 0 | 37.6 | 0.191 | 30 | 0 |

Table 1. Dataset and its Features

- The diabetes data set consists of 768 data points, with 9 features each.
- “Outcome” is our Target Variable
- {0} indicates non-Diabetic and {1} indicates Diabetic

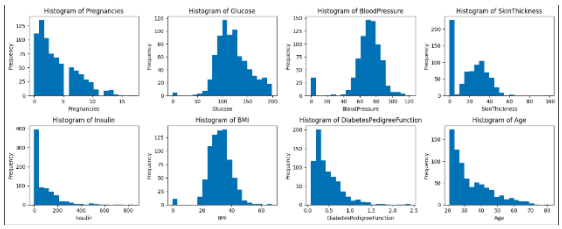
**Steps Performed:**

**Load Data**

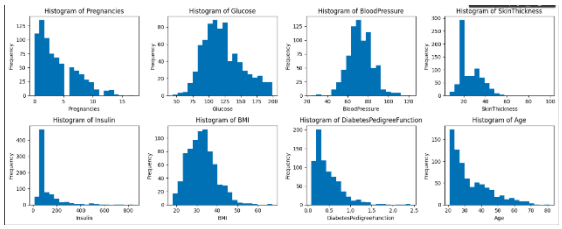
The data, which is in the CSV format, is loaded to a variable. There are 768 data points in Dataset

**Data Cleaning:** 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 of inaccurate predictions, this is a crucial step in improving prediction accuracy.

**BEFORE:**



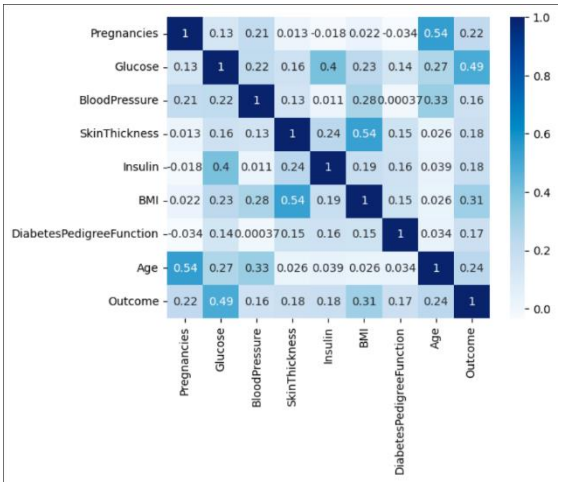

**AFTER:**




**Feature Selection**: We have used correlation analysis, feature importance analysis, and domain knowledge to select the most appropriate features. and inconclusion we have used all the features.

**Visualization:** Visualization is an important tool in exploratory data analysis and helps to understand patterns and relationships in the data. Different types of visualizations, such as scatter plots, histograms, and heat maps are used to visualize the data and relationships between variables and also help in removing the outliers. Visualization can provide insights into the data and guide the selection of features and models for machine learning tasks.

**Correlation heatmap:**




**Logistic regression**: Logistic regression is a popular statistical method used to model the relationship between a binary response variable (i.e., a variable that can take on only two values, such as yes or no) and one or more predictor variables.

**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

**Performance Evaluation:**

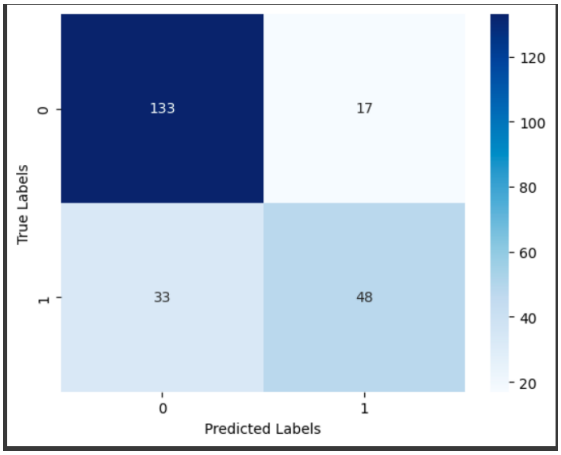
Performance evaluation is an essential step in machine learning to assess the quality of the trained models. Multiple performance metrics are used to evaluate the model's accuracy, such as Precision, Recall, Accuracy, Etc.

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

| | Metric | Value |
|---|---|---|
| 0 | PRECISION | 0.738462 |
| 1 | RECALL | 0.592593 |
| 2 | ACCURACY | 0.783550 |

**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




**Why Linear Regression is not used?**

Linear regression is a statistical method used to model the relationship between a continuous dependent variable and one or more independent variables.

However, diabetes detection typically involves predicting a binary outcome (i.e., whether a patient has diabetes or not), which is better suited for logistic regression. Linear regression can still be used in diabetes research to explore the relationship between continuous predictor variables and measures of diabetes severity or risk factors,

After seeing the Correlation between the attributes, we found that the Glucose attribute had the highest value. so, we used Glucose for linear regression.

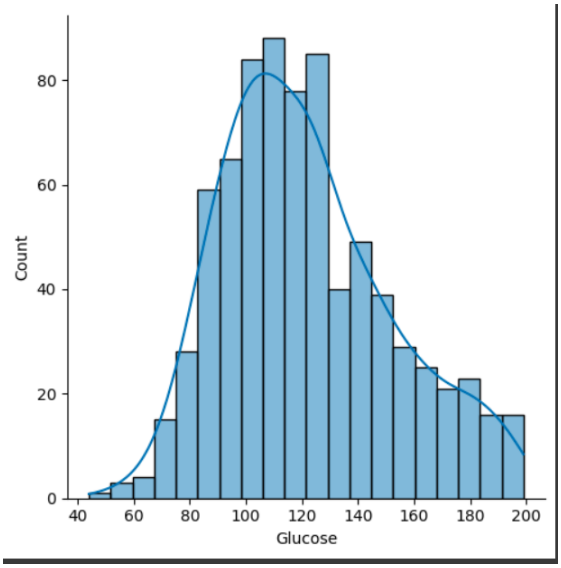



Fig: Glucose Attribute

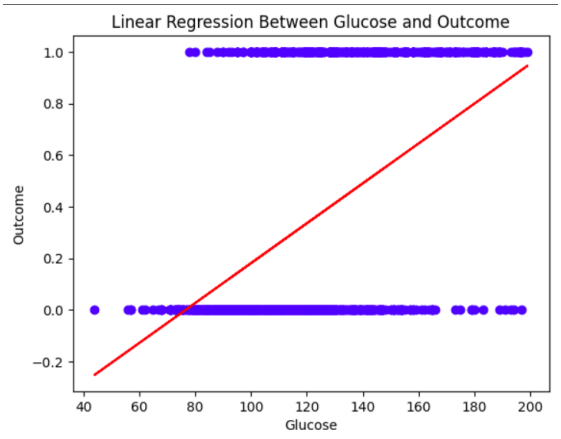



Fig: Regression Line

**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-roo m/fact-sheets/detail/diabetes.

2.https://www.researchgate.net/profile/Pramila-Chawan/publication/326416823_LOGISTIC_REGRESSION_AND_SVM_BASED_DIABETES_PREDICTION_SYSTEM/links/5b4c80ddaca272c609 47858a/LOGISTIC-REGRESSION-AND-SVM-BASED-DIABETES - PREDICTION-SYSTEM.pdf

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

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