PROJECT REPORT ON

Glucose Level Prediction and Diabetes Detection System

Submitted By

HARITHA KRISHNA R (CEC21CS050)

under the esteemed guidance of

Mrs.Preetha Theresa Joy

Professor

Department Of Computer Engineering

APRIL 2025

DEPARTMENT OF COMPUTER ENGINEERING
COLLEGE OF ENGINEERING, PALLIPPURAM P O, CHERTHALA,
ALAPPUZHA PIN: 688541,
PHONE: 0478 2553416, FAX: 0478 2552714
http://www.cectl.ac.in

PROJECT REPORT ON

Glucose Level Prediction and Diabetes Detection System

Submitted By

HARITHA KRISHNA R (CEC21CS050)

under the esteemed guidance of

Mrs.Preetha Theresa Joy

(Professor)

In partial fulfillment of the requirements for the award of the degree
of
Bachelor of Technology
in
Computer Science and Engineering
of
APJ Abdul Kalam Technological University

APRIL 2025
DEPARTMENT OF COMPUTER ENGINEERING
COLLEGE OF ENGINEERING, PALLIPPURAM P O, CHERTHALA,
ALAPPUZHA PIN: 688541,
PHONE: 0478 2553416, FAX: 0478 2552714
http://www.cectl.ac.in

DEPARTMENT OF COMPUTER ENGINEERING

COLLEGE OF ENGINEERING CHERTHALA

ALAPPUZHA-688541

CERTIFICATE

This is to certify that the project report titled **Glucose Level Prediction and Diabetes Detection System** is a bonafide record of the **CSD496 MINI PROJECT** presented by **HARITHA KRISHNA R** (CEC21CS050) Eight Semester B.Tech Computer Science & Engineering student, under our guidance and supervision, in partial fulfillment of the requirements for the award of the degree, **B. Tech. Computer Science & Engineering** of **A P J Abdul Kalam Technological University**.

HoD

Dr. Preetha Theresa Joy
Professor
Dept Of Computer Engg

ACKNOWLEDGEMENT

This work would not have been possible without the support of many people. First and the foremost, I would thanks to the Almighty God who gave the inner strength, resource and ability to complete project successfully.

I would like to thank **Dr. Jaya V L**, our Principal, who has provided with the best facilities and atmosphere for the project completion and presentation. I would also like to thank our HoD **Dr. Preetha Theresa Joy** (Professor, Department Of Computer Engineering) for the help extended and also for the encouragement and support given while doing the project.

Also I would like to thank our dear friends for extending their cooperation and encouragement throughout the project work, without which would never have completed the project this well. Thank you all for your love and also for being very understanding.

ABSTRACT

Diabetes is a widespread chronic condition caused by high blood glucose levels, leading to severe health complications such as heart disease, kidney failure, and vision loss if left untreated. Early detection is crucial for effective management of the disease and prevention of its associated risks. This project focuses on developing a Glucose Level Prediction and Diabetes Detection System using machine learning techniques. The system aims to predict the likelihood of diabetes and estimate glucose levels based on key health parameters such as glucose level, BMI, blood pressure, and insulin levels. By leveraging machine learning models, including classification and regression techniques, the system provides an accessible tool for early detection of diabetes. The project utilizes the Pima Indians Diabetes Dataset for training the models, with the objective of improving accuracy and providing actionable insights for individuals at risk. The models are integrated into a user-friendly web application developed using Flask, enabling users to input health data and receive real-time predictions. The findings demonstrate the efficacy of the machine learning models in predicting both diabetes risk and glucose levels with high accuracy, contributing to proactive healthcare management.

Keywords: Machine Learning, Glucose Level Estimation, Diabetes Prediction, Healthcare AI, Flask Web Application.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

Diabetes is a chronic disease that is becoming a global health crisis. It occurs when the body is unable to regulate blood glucose (sugar) levels effectively, either due to insufficient insulin production or the body's inability to use insulin properly. According to the World Health Organization (WHO), about 422 million people worldwide suffer from diabetes, with this number expected to rise sharply by 2030. India, with its population of over 1.4 billion, has a diabetes population exceeding 40 million. Diabetes can cause serious complications such as heart disease, kidney failure, vision impairment, nerve damage, and even amputations if not managed properly. There are three main types of diabetes:

- **Type 1 Diabetes:** This occurs when the immune system attacks the insulin-producing cells in the pancreas, leading to little or no insulin production. Type 1 diabetes is often diagnosed in childhood or adolescence, and individuals with this type must rely on insulin injections to manage their blood sugar levels.

- **Type 2 Diabetes:** The most common form of diabetes, Type 2 occurs when the body becomes resistant to insulin or when the pancreas does not produce enough insulin. It is closely linked with lifestyle factors such as obesity, physical inactivity, and poor diet. While it primarily affects adults, Type 2 diabetes is increasingly being diagnosed in younger individuals.

- **Gestational Diabetes:** This form of diabetes develops during pregnancy and typically resolves after childbirth. However, women who experience gestational diabetes are at a higher risk of developing Type 2 diabetes later in life.

The main cause of diabetes, particularly Type 2, is a combination of genetic predisposition and lifestyle choices such as poor diet, lack of physical activity, and obesity. Early detection and timely intervention are critical to preventing complications and managing the disease effectively. Key symptoms of diabetes include frequent urination, excessive thirst, unexplained weight loss, blurred vision, and slow-healing wounds.

This project aims to provide a solution for early prediction of diabetes as well as the prediction of glucose levels, which are vital indicators in the diagnosis of the disease. By predicting glucose levels, it is possible to assess the risk of diabetes before it fully develops, giving individuals an opportunity for early intervention.

Using the Pima Indian Diabetes Dataset, which contains medical data such as glucose levels, body mass index (BMI), age, blood pressure, and insulin levels, the system will predict both the likelihood of diabetes and abnormal glucose levels. The prediction of glucose levels specifically helps in understanding how the body is managing sugar, which is crucial for diagnosing both prediabetes and diabetes. For this purpose, various machine learning classification and ensemble techniques, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machine (SVM), Gradient Boosting, will be applied. These models will be trained on the dataset to predict both diabetes and glucose levels, aiming to achieve higher accuracy in predicting individuals at risk of developing diabetes or experiencing abnormal glucose levels.

The goal of this project is to provide a reliable tool that can predict diabetes risk as well as glucose levels, enabling earlier diagnosis and better management of the disease. This will allow for improved health outcomes and potentially prevent the severe complications associated with diabetes.

Chapter 2

PROBLEM STATEMENT

2.1 PROBLEM STATEMENT

Diabetes is a major global health issue, affecting millions of people worldwide, and its prevalence is expected to continue rising. The disease can lead to severe complications, including heart disease, kidney failure, nerve damage, and blindness if not detected and managed early. Early detection plays a crucial role in preventing or delaying the onset of diabetes and its related complications. However, diagnosing diabetes typically involves complex tests that may not be readily available or accessible in all areas, particularly in low-income settings.

Moreover, abnormal glucose levels, often seen as an early indicator of diabetes, can be challenging to monitor consistently. Predicting glucose levels and identifying individuals at risk of developing diabetes before the disease fully manifests could significantly improve public health outcomes. This prediction can help in early interventions, lifestyle changes, and timely medical treatment.

The problem lies in efficiently predicting both diabetes and glucose levels using easily accessible data, without the need for invasive procedures or expensive tests. Traditional diagnostic methods rely heavily on a combination of clinical assessments and lab tests, which can be costly and time-consuming. In contrast, utilizing machine learning models on readily available health data, such as age, body mass index (BMI), blood pressure, and glucose levels, could provide a more affordable and accessible solution for early diagnosis.

This project aims to address this problem by developing a system that can predict the likelihood of diabetes as well as abnormal glucose levels. By using machine learning techniques, we intend to create a model that can accurately predict these outcomes based on input data, helping healthcare providers and individuals to make informed decisions earlier in the disease progression. By accurately predicting diabetes and glucose levels, the system will aid in preventing complications associated with undiagnosed or poorly managed diabetes, improving the quality of life for individuals worldwide.

2.2 OBJECTIVE

The objective of this project is to design and develop a predictive system using machine learning algorithms to assess an individual's risk of developing diabetes and to forecast their glucose levels based on health-related attributes. Diabetes is a chronic condition that can lead to severe complications if not managed properly, and early detection is crucial to preventing or delaying its onset. The primary aim of the project is to utilize various attributes such as age, BMI, blood pressure, insulin levels, family history, and previous glucose levels to predict the likelihood of an individual developing diabetes and to provide an estimate of their future glucose levels.

The project also aims to explore and evaluate the performance of different machine learning classification and regression models, including techniques such as logistic regression, decision trees, support vector machines, random forests, and ensemble methods. These models will be trained using the Pima Indian Diabetes dataset, which includes health data of individuals diagnosed with diabetes, and will be tested for accuracy and efficiency in both diabetes prediction and glucose level forecasting.

By focusing on the prediction of diabetes and glucose levels, the project seeks to offer a tool that can assist healthcare professionals and individuals in early intervention, ensuring better management of the disease. The end goal is to create an accurate and reliable system that can predict the risk of diabetes and monitor glucose levels, providing valuable insights for preventive healthcare measures.

Chapter 3

LITERATURE REVIEW

Application of data mining: Diabetes health care in young and old patients (2012): The study focuses on analyzing diabetes treatment effectiveness using predictive regression analysis based on data from the 2005 Non-Communicable Disease (NCD) risk factor report provided by the Saudi Arabian Ministry of Health. The dataset, stored in Oracle 10g, consists of six tables categorizing different treatment methods across various age groups. To achieve accurate predictions, the research follows a structured six-step data mining process, which includes data selection, data preparation, data analysis, pattern prediction, and model deployment. The Support Vector Machine (SVM) regression model was applied to evaluate treatment effectiveness, achieving an accuracy rate of 87%. The results indicate that drug therapy is more effective for older patients, whereas exercise and weight management significantly improve diabetes control in younger individuals. These findings highlight that predictive data mining techniques can enhance diabetes management by providing accurate treatment predictions. The study concludes that SVM-based regression analysis is a reliable method for optimizing treatment strategies, supporting data-driven decision-making in healthcare. [1]

An Effective Diabetes Prediction System Using Machine Learning Techniques (2020): In this paper, various approaches for diabetes prediction using the Pima Indians Diabetes Dataset (PIDD) were analyzed. The study emphasizes the role of data preprocessing techniques, such as mean imputation for handling missing values, Mutual Information-based Feature Selection (MI-FS) for selecting important features, Random Over-Sampling to balance the dataset, and Robust

Scaling for normalizing data, to improve model performance. Several machine learning classifiers were examined, including Decision Tree (DT), Random Forest (RF), Extra Trees (ET), and Adaptive Boosting (AdaBoost).The evaluation metrics used were F1-score, precision, recall, and ROC_AUC. Among the models, Random Forest achieved the highest accuracy of 84.1%, followed by Extra Trees with 83.2%, AdaBoost with 81.7%, and Decision Tree with 78.6%. The results demonstrated that ensemble learning methods, particularly Random Forest and Extra Trees, outperformed individual classifiers by reducing overfitting and improving generalization. Additionally, AdaBoost further enhanced classification performance by combining multiple weak learners into a stronger predictive model.Overall, this literature survey highlights that effective data preprocessing, feature selection, and ensemble learning techniques significantly improve the accuracy of diabetes classification models. The findings provide valuable insights into optimal preprocessing strategies and machine learning approaches, which can serve as a foundation for future research in diabetes prediction. [2]

Diabetes Prediction using ML with Feature Selection and Dimensionality Reduction (2022): Sivaranjani S., Ananya S., Aravinth J., and Karthika R. propose a machine learning approach to diabetes prediction, focusing on feature selection and dimensionality reduction techniques. Feature selection helps identify the most relevant features from a dataset, eliminating unnecessary or redundant information that could slow down or reduce the accuracy of the model. Dimensionality reduction further simplifies the data by projecting it into a lower-dimensional space, making it easier to process without losing critical information. The combination of these techniques improves the performance of machine learning models, resulting in faster and more accurate predictions. This study highlights the significance of preprocessing techniques like feature selection and dimensionality reduction in improving machine learning models for healthcare applications. By optimizing the input data, these methods enhance the overall efficiency and accuracy of diabetes prediction systems. [3]

Aiswaryaet al. 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. [4]

Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming (2014): The reviewed studies highlight the effectiveness of machine learning and deep learning techniques in improving diabetes detection. Traditional ML models, such as K-Nearest Neighbor (K-NN), Decision Trees (DT), and Logistic Regression (LR), have shown promising results, with K-NN achieving an accuracy of 79.6%. Additionally, ensemble models and boosted decision trees have demonstrated enhanced robustness in classification tasks. Deep learning models, including Convolutional Neural Networks (CNNs) and Variational Autoencoders (VAEs), have further improved accuracy, with CNN-based approaches achieving up to 92.31% accuracy. Techniques such as data augmentation, feature selection, and batch normalization have proven to be crucial in optimizing model performance. However, data imbalance, interpretability, and real-world applicability remain challenges that require further exploration. Future research should focus on integrating large-scale datasets, refining deep learning architectures, and enhancing model interpretability to improve clinical adoption and real-time diabetes screening accuracy. [5]

Diabetes Prediction using Optimization Techniques with Machine Learning Algorithms(2023): Diabetes is a chronic disease that requires early detection for effective management. In the study *"Diabetes Prediction Using Machine Learning Techniques,"* various machine learning models were analyzed for their ability to predict diabetes based on clinical data. The research utilized the Pima Indians Diabetes Dataset (PIDD), which includes key health indicators such as glucose levels, BMI, and insulin levels.The methodologies employed included data preprocessing (handling missing values, normalization), feature selection, and model training using supervised learning techniques. The study evaluated multiple models, including Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Artificial Neural Networks (ANN). The performance of these models was assessed using accuracy, precision, recall, and F1-score.Random Forest (accuracy $\sim 87\%$) and ANN (accuracy $\sim 85\%$) outperformed other models, demonstrating strong predictive capabilities. In contrast, simpler models like Logistic Regression

($\sim 78\%$) and Decision Tree ($\sim 75\%$) showed moderate accuracy. The study highlights that ensemble methods and deep learning approaches enhance diabetes prediction accuracy, making them effective tools for early diagnosis and healthcare decision-making. [6]

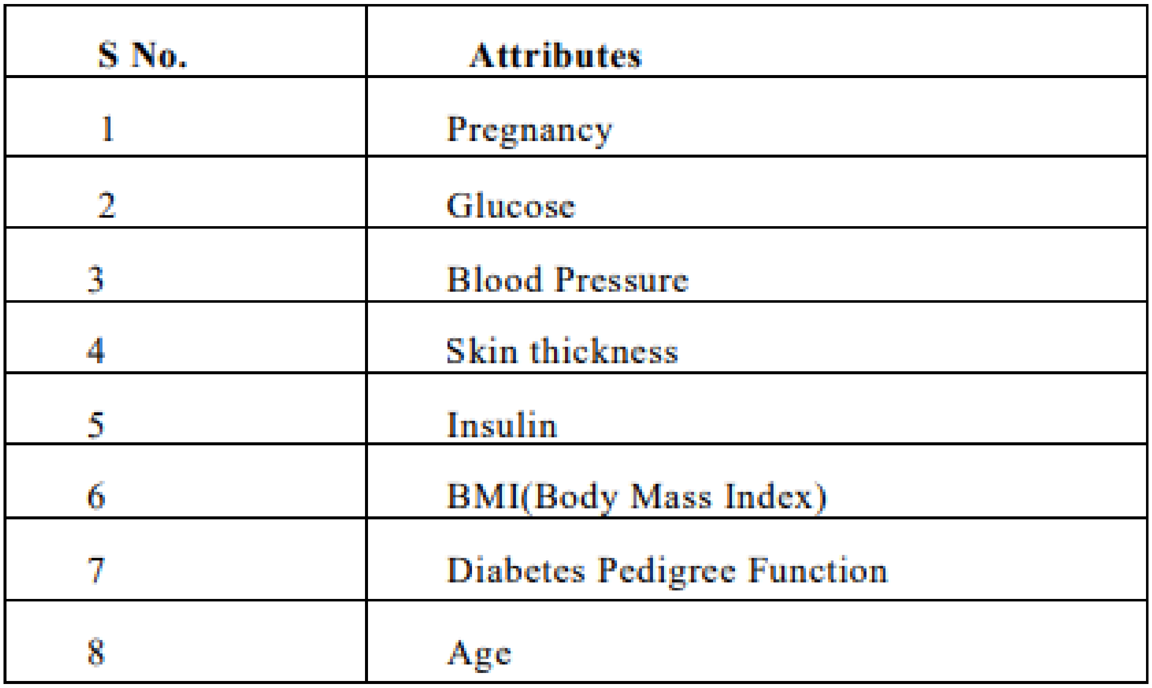
Chapter 4

METHODOLOGY

The methodology for this project involves several stages, from data collection to model implementation and evaluation. The goal is to predict both the likelihood of an individual developing diabetes and to forecast their glucose levels. The project uses the Pima Indian Diabetes dataset, a well-known dataset in diabetes research, which includes various health attributes such as age, BMI, blood pressure, insulin levels, and more.

4.1 Dataset Description

The dataset used in this project is the Pima Indian Diabetes Dataset, which consists of 768 instances with 9 attributes. These attributes represent various health metrics that are relevant for diabetes prediction. The features include Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI (Body Mass Index), DiabetesPedigreeFunction, Age, and the target variable Outcome. The Outcome attribute indicates whether the patient is diabetic (1) or not (0), and it serves as the key variable for prediction. The dataset is slightly imbalanced, with more instances of non-diabetic individuals than diabetic ones, which requires careful handling of class imbalance during model training. Although there are no missing values for most of the attributes, some features like Glucose, BloodPressure, SkinThickness, and Insulin contain zero values, which are likely indicative of missing or invalid data. These zero values need to be addressed during the data preprocessing phase to ensure accurate model predictions.



S No.	Attributes
1	Pregnancy
2	Glucose
3	Blood Pressure
4	Skin thickness
5	Insulin
6	BMI(Body Mass Index)
7	Diabetes Pedigree Function
8	Age

Fig. 4.1: Dataset Description

Distribution of Diabetic patient- 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.

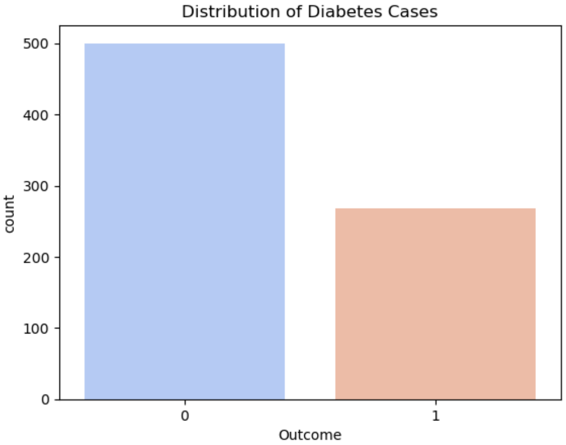


Fig. 4.2: Distribution of Diabetic patient

4.2 Data Preprocessing

Data preprocessing is a crucial step in any machine learning project as it ensures that the data is clean, consistent, and formatted in a way that enhances the performance of the models. For this diabetes prediction project, the preprocessing steps include handling missing values, removing outliers, and normalizing the data to improve the model's predictive accuracy.

- **Handling Missing or Invalid Values:** The dataset contains columns like Glucose, BloodPressure, SkinThickness, Insulin, and BMI, where zero values are not valid. These zero values are likely placeholders for missing or invalid data.To handle this, the code replaces the zero values with the median of the respective columns. The median is chosen because it is less sensitive to outliers than the mean, ensuring the data remains robust.
- **Outlier Detection and Removal:** Boxplots are generated for all features to visually identify outliers. Boxplots provide a simple and clear way to detect extreme values in the dataset.The Interquartile Range (IQR) is calculated for each feature. The IQR is the difference between the 75th percentile (Q3) and the 25th percentile (Q1). It helps in identifying the "normal" range of values for each feature.Using the IQR, lower and upper bounds are defined (Q1 - 1.5 * IQR, Q3 + 1.5 * IQR). Data points outside these bounds are considered outliers.Rows containing outliers are removed from the dataset. This ensures that the model doesn't get biased by extreme values that may not represent the typical data distribution.
- **Final Dataset:**After preprocessing, the dataset is cleaned, and any invalid zero values and outliers are handled. The final dataset size is reduced from 768 entries to 375 entries after outlier removal.The cleaned dataset is now ready for feature scaling and training the machine learning models.

4.3 Machine Learning Model Implementation

This project involves two key tasks:

- Diabetes Prediction (Classification Task)
- Glucose Level Prediction (Regression Task)

4.3.1 Diabetes Prediction (Classification Task)

4.3.1.1 Support Vector Machine (SVM)

Support Vector Machine is a supervised machine learning algorithm primarily used for classification tasks. It is one of the most powerful classification techniques that works well for both linearly separable and non-linearly separable data. SVM creates a hyperplane that separates data points into two distinct classes. This hyperplane can be in a high-dimensional space, making SVM effective even for complex datasets. The key idea is to maximize the margin, which is the distance between the hyperplane and the closest data points from either class. These closest points are called support vectors, and they play a crucial role in defining the decision boundary. **Algorithm**

- Select the hyper plane which divides the class better.
- To find the better hyper plane you have to calculate the distance between the planes and the data which is called Margin.
- If the distance between the classes is low then the chance of miss conception is high and vice versa.
- Select the class which has the high margin.Margin = distance to positive point + Distance to negative point.

4.3.1.2 K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a supervised machine learning algorithm that can be used for both classification and regression tasks. It is a *lazy learning* technique, meaning it does not create a model during training but instead stores all the training data and makes predictions by comparing new instances with existing ones.

KNN assumes that similar data points exist in close proximity in the feature space. The algorithm classifies a new data point based on the majority class of its nearest neighbors. The number of neighbors, denoted as K, is a hyperparameter that affects the model's accuracy. A smaller value of K makes the model more sensitive to noise, while a larger K results in smoother decision boundaries.

Distance Calculation To determine the nearest neighbors, Euclidean distance is commonly used. The Euclidean distance between two points $P(p_1, p_2, \ldots, p_n)$ and $Q(q_1, q_2, \ldots, q_n)$ is given by:

$$d(P, Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \tag{4.1}$$

Algorithm

- Take a sample dataset with multiple attributes and instances, such as the Pima Indian Diabetes dataset.
- Take a test dataset containing new attribute values that need to be classified.
- Calculate the Euclidean distance between the test data point and all points in the training dataset using the formula given above.
- Select a value for K, which represents the number of nearest neighbors.
- Identify the K nearest neighbors by selecting the smallest distance values.
- Determine the majority class among the selected K neighbors.
- Assign the predicted class to the new data point:
  - If most neighbors belong to the diabetic class (1), classify the instance as diabetic.
  - Otherwise, classify it as non-diabetic (0).

4.3.1.3 Logistic Regression

Logistic Regression is a supervised learning classification algorithm used to predict the probability of a binary outcome based on one or more independent variables. It is widely used for binary classification problems, such as predicting whether a patient has diabetes (1) or does not have diabetes (0). Unlike linear regression, which predicts continuous values, logistic regression models the probability that a given instance belongs to a particular class.

The main goal of logistic regression is to find the best-fit function that describes the relationship between the predictor variables (features) and the target variable (outcome). The model is based on the principles of linear regression, but instead of fitting a straight line, it uses a sigmoid (logistic) function to restrict the output to values between 0 and 1. The sigmoid function is used in logistic regression to transform linear outputs into probability values:
Sigmoid Function: The sigmoid function is used in logistic regression to transform linear outputs into probability values:

$$P = \frac{1}{1 + e^{-(a+bx)}} \tag{4.2}$$

where P is the probability that the output belongs to class 1 (e.g., diabetic), a is the intercept, b is the coefficient (weight) of the predictor variable, x is the input feature, and e is the mathematical constant (Euler's number).

The sigmoid function ensures that the output is always between 0 and 1, making it suitable for probability estimation.

Algorithm

- Load the Pima Indian Diabetes dataset, which contains multiple features such as Glucose level, BMI, Blood Pressure, Age, and more.

- Preprocess the data by handling missing values, removing outliers, and normalizing the features.

- Split the dataset into training and testing sets to evaluate the model's performance.
- Apply logistic regression to the training data:
  - Compute the linear combination of input features and model coefficients.
  - Apply the sigmoid function to obtain probability values.
  - Classify instances based on a probability threshold (default = 0.5):
    - If $P \geq 0.5$, classify as diabetic (1).
    - If $P < 0.5$, classify as non-diabetic (0).
- Evaluate the model using accuracy, precision, recall, and F1-score to measure its effectiveness in predicting diabetes.

4.3.1.4 Naive Bayes

Naive Bayes is a supervised machine learning algorithm used for classification tasks. It is based on Bayes' Theorem and assumes that the features are independent given the class label, which simplifies the calculations. Despite this assumption often being unrealistic, Naive Bayes is widely used in problems like spam classification and sentiment analysis because of its simplicity and effectiveness. The core of the Naive Bayes algorithm is Bayes' Theorem, which calculates the probability of a class given the observed features:

$$P(C \mid X) = \frac{P(X \mid C) \cdot P(C)}{P(X)} \tag{4.3}$$

Where $P(C \mid X)$ is the probability of class C given the features X, $P(X \mid C)$ is the likelihood of observing X given class C, $P(C)$ is the prior probability of class C, and $P(X)$ is the total probability of the features.

The algorithm classifies a data point by computing this probability for all possible classes and selecting the class with the highest probability.

Algorithm

- **Load the Dataset:** Start with a labeled dataset containing features and corresponding class labels.
- **Preprocess the Data:** Handle missing values, normalize features, and clean the data if necessary.
- **Calculate Prior Probabilities:** For each class C, calculate $P(C)$, the fraction of instances belonging to class C.
- **Calculate Likelihoods:** For each feature, compute the likelihood $P(X_i \mid C)$, the probability of observing feature X_i given class C.
- **Apply Bayes' Theorem:** Use Bayes' Theorem to calculate the posterior probability for each class.
- **Make Prediction:** Assign the class with the highest posterior probability to the data point.
- **Evaluate the Model:** Assess performance using metrics like accuracy and precision.

4.3.1.5 Random Forest

Random Forest is an ensemble learning method used for classification and regression. It works by creating multiple decision trees during the training phase and combining their results. This approach improves prediction accuracy by reducing overfitting, which is common in individual decision trees.

Algorithm

1. Select a random subset of features R from the total features M.
2. For each tree, find the best split using the selected features.
3. Split the nodes recursively until a predefined number of nodes is reached.
4. Repeat steps 1-3 to create multiple trees in the forest.

5. For classification, the final prediction is made based on the majority vote from all trees. For regression, the output is the average prediction from all trees.

Prediction:

- For classification, the most voted class by all trees is chosen as the final prediction.
- For regression, the average of all tree predictions is used as the final output.

4.3.1.6 XGBoost (Extreme Gradient Boosting)

XGBoost is an optimized machine learning algorithm based on the gradient boosting framework. It is widely used for classification and regression tasks due to its efficiency, speed, and high accuracy. XGBoost improves predictive performance by combining multiple weak learners (decision trees) and reducing errors through gradient boosting. **Algorithm**

1. Initialize the dataset and define the objective function.
2. Create an initial weak model (decision tree).
3. Compute the loss function to measure errors.
4. Calculate gradients (direction of error reduction).
5. Train new trees to correct errors from previous models.
6. Update weights and improve predictions iteratively.
7. Repeat steps 3-6 until a stopping condition is met (e.g., no further improvement).
8. Make the final prediction by summing the weighted outputs of all trees.

4.3.2 Glucose Level Prediction (Regression Task)

4.3.2.1 Linear Regression

Linear Regression is one of the simplest and most widely used regression techniques in machine learning. It models the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a linear equation to the observed data. The goal of Linear Regression is to minimize the difference between the predicted values and the actual values of the target variable. Linear regression finds the best-fitting line (also called the regression line) through the data points. The line is determined by the equation:

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \cdots + b_n \cdot x_n \tag{4.4}$$

where y is the predicted value (glucose level in this case), b_0 is the intercept (the point where the line crosses the y-axis), $b_1, b_2, \ldots, b_n$ are the coefficients (weights) of the features $x_1, x_2, \ldots, x_n$, $x_1, x_2, \ldots, x_n$ are the input features.

The goal of Linear Regression is to find the optimal values of the coefficients $b_0, b_1, b_2, \ldots, b_n$ that minimize the prediction error.

4.3.2.2 Random Forest Regressor

Random Forest Regressor is an ensemble learning algorithm that combines multiple decision trees to make predictions. It works by constructing a multitude of decision trees during training, and the final prediction is made by averaging the predictions from all the individual trees. This approach improves the model's accuracy and robustness, as the randomization of both data samples and feature selection helps in reducing overfitting.

The key idea behind Random Forest is bagging (Bootstrap Aggregating), where several decision trees are trained independently on different subsets of the data, and each tree contributes to the final output. The model's prediction for a given input is obtained by averaging the predictions of all the individual trees. The process can be broken down as follows:

Training the Model: A random subset of the training data is selected for each tree, with replacement (i.e., some data points may be repeated). For each decision tree, a random subset of features is selected at each split to ensure diversity and reduce correlation between trees.

Prediction: For regression tasks, the prediction of the Random Forest model is the average of the predictions made by all individual trees:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{4.5}$$

Where $\hat{y}$ is the final predicted value (e.g., glucose level), y_i is the prediction from the i-th tree, and N is the total number of trees in the forest.

4.3.2.3 XGBoost (Extreme Gradient Boosting)

XGBoost is an efficient and powerful machine learning algorithm used for regression tasks. It is based on the gradient boosting framework, which builds multiple decision trees in a sequential manner to improve prediction accuracy. Each tree in the sequence attempts to correct the errors made by the previous trees.

The primary goal of XGBoost is to minimize a loss function, which measures the difference between the predicted and actual values. The algorithm combines the predictions of all trees to make the final prediction, resulting in a strong and accurate model. The key equation for XGBoost in regression is:

$$\hat{y} = \sum_{i=1}^{T} f_i(x) \tag{4.6}$$

Where $\hat{y}$ is the predicted value (in this case, the glucose level), $f_i(x)$ represents the prediction made by the i-th tree, and T is the total number of trees in the model. The goal is to minimize the residual errors at each step by fitting new trees to the errors of the combined predictions of previous trees.

XGBoost also includes regularization terms to prevent overfitting, ensuring that the model generalizes well to unseen data.

4.4 Model Building

- Import necessary libraries such as pandas, numpy, sklearn, and xgboost, and load the diabetes dataset.
- Clean the dataset by removing missing or irrelevant values and handle missing data appropriately.
- Split the data into training (80%) and testing (20%) sets using a standard train-test split.
  - Glucose Level Prediction: Choose Linear Regression, Random Forest Regressor, and XGBoost Regressor for predicting continuous glucose levels.
  - Diabetes Prediction: Choose KNearest Neighbor, Support Vector Machine, Naives Bayes,Logistic regression, Random Forest and XGBoost algorithm.
- Train each selected model on the training dataset using respective algorithms.
  - Evaluate each model's performance on the test dataset for glucose level prediction using metrics such as Mean Squared Error (MSE) and R^2 score.
  - For diabetes prediction, evaluate the models using classification metrics like Accuracy, Precision, Recall, F1-score, and AUC-ROC.
- Compare the models' results and select the best-performing model based on accuracy and performance metrics.

Chapter 5

Experimental Results

5.1 Diabetes Prediction Results

To classify whether a person has diabetes or not, multiple classification models were tested. The table below summarizes their performance in terms of Accuracy and ROC AUC.

Model	Accuracy	ROC AUC
Logistic Regression	0.7733	0.8822
Support Vector Machine (SVM)	0.7733	0.8777
Naïve Bayes	0.7600	0.8169
Random Forest	0.7600	0.7895
XGBoost	0.7333	0.7690
K-Nearest Neighbors (KNN)	0.6933	0.7264

Table 5.1: Comparison of Model Performance

From the results, Logistic Regression achieved the highest Accuracy (77.33%) and ROC AUC (88.22%), making it the best-performing model for diabetes prediction. Hyperparameter tuning was performed on Logistic Regression to further improve its performance.

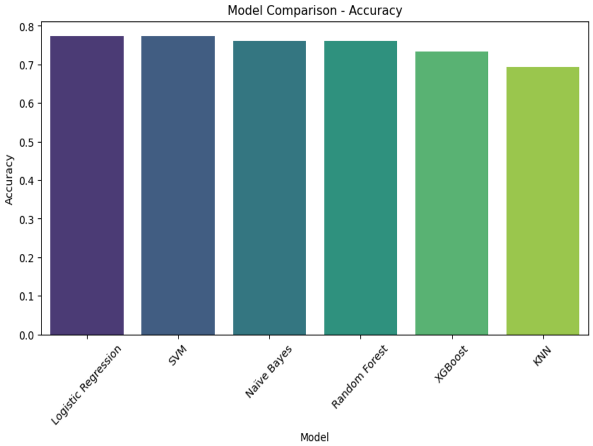


Fig. 5.1: Model Accuracy Comparison

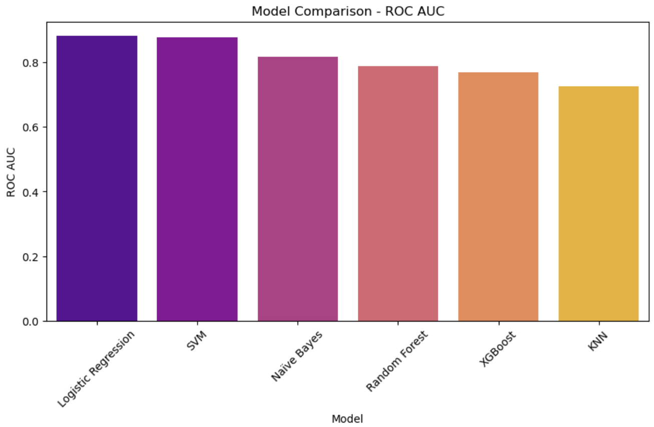


Fig. 5.2: ROC AUC Comparison

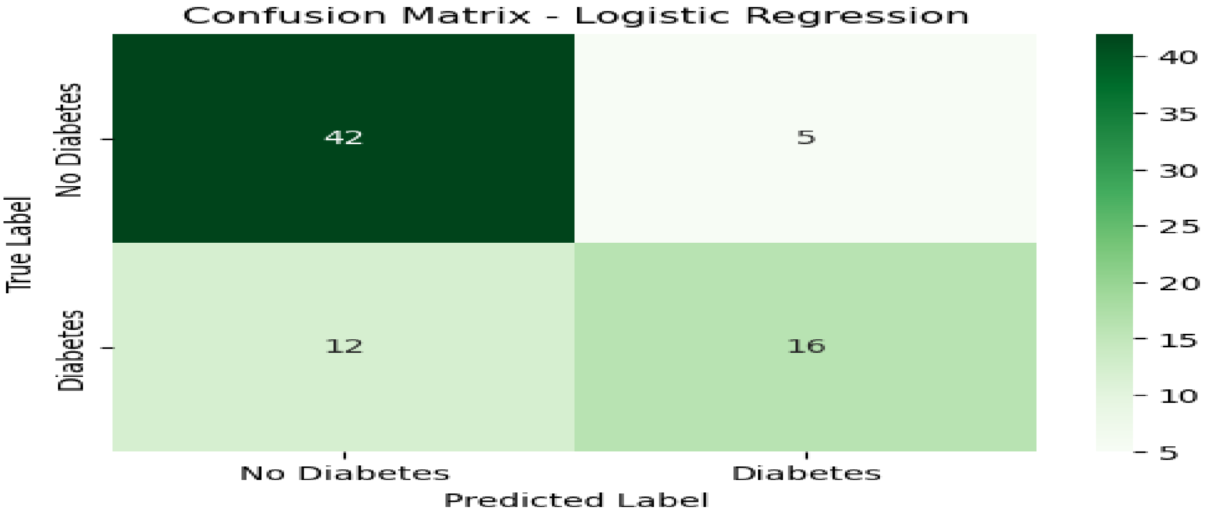


Fig. 5.3: Confusion Matrix

5.2 Glucose Level Prediction Results

For predicting glucose levels as a continuous variable, Linear Regression, Random Forest Regressor, and XGBoost Regressor were evaluated using Mean Squared Error (MSE) and R² Score.

Model	MSE	R² Score
Linear Regression	650.85	0.0692
Random Forest Regressor	771.36	-0.1031
XGBoost Regressor	873.65	-0.2494

Table 5.2: Comparison of Regression Model Performance

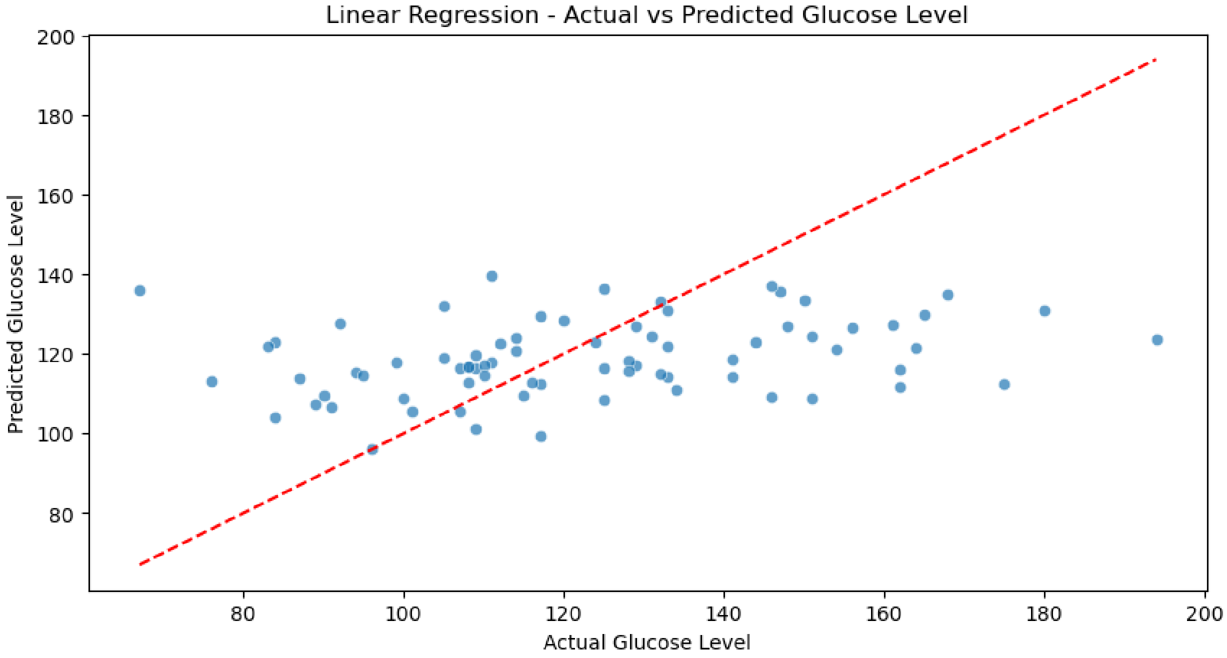
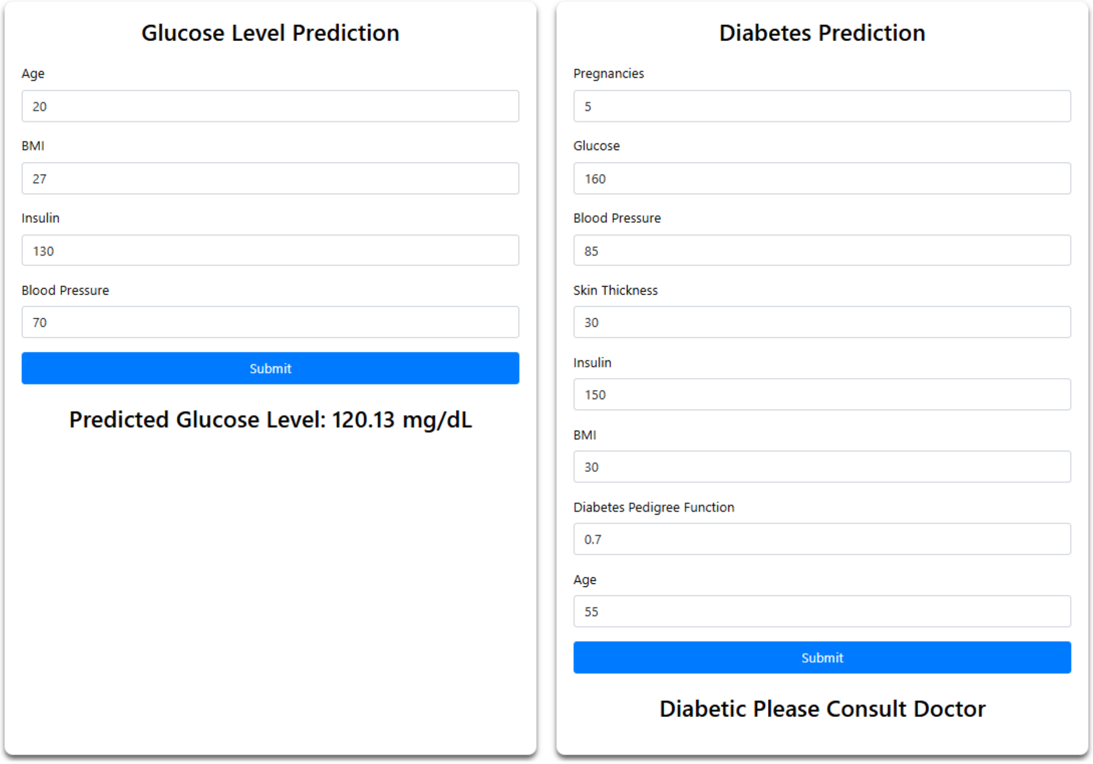


Fig. 5.4: Linear Regression

Chapter 6

IMPLEMENTATION



Glucose Level Prediction

Age: 20

BMI: 27

Insulin: 130

Blood Pressure: 70

Predicted Glucose Level: 120.13 mg/dL

Diabetes Prediction

Pregnancies: 5

Glucose: 160

Blood Pressure: 85

Skin Thickness: 30

Insulin: 150

BMI: 30

Diabetes Pedigree Function: 0.7

Age: 55

Diabetic Please Consult Doctor

Fig. 6.1: Predictions

Chapter 7

CONCLUSION AND FUTURE SCOPE

The proposed methodology effectively utilized machine learning models for diabetes prediction and glucose level estimation, demonstrating the potential of AI in healthcare. Among the classification models tested, Logistic Regression emerged as the best-performing algorithm for diabetes prediction, achieving an accuracy of 77.33% and an ROC AUC score of 88.22%, making it a reliable choice for early detection. For glucose level prediction, Linear Regression provided the most stable results, with an MSE of 650.85 and an R^2 score of 0.0692, indicating moderate predictive performance. While the results are promising, improvements can be made to enhance model accuracy and robustness. The study highlights the importance of data-driven approaches in assisting medical professionals with early diagnosis and proactive management of diabetes.

Future work can focus on enhancing prediction accuracy by incorporating additional features such as HbA1c levels, dietary habits, and real-time physiological data. The use of deep learning techniques, including Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN), can further improve predictive capabilities. Hyperparameter tuning and ensemble methods can be explored to optimize model performance. Additionally, integrating IoT-based real-time glucose monitoring can enable continuous tracking and early intervention. Developing a web or mobile application to provide users with personalized risk assessment and glucose level predictions can make the system more accessible and impactful. By implementing these advancements, the predictive capability of the models can be significantly improved, leading to more effective diabetes diagnosis, monitoring, and management.

References

[1] Aljumah, Abdullah & Ahamad, Mohammed & Siddiqui, Mohammad Khubeb. (2012). Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University – Computer and Information Sciences, 25. https://doi.org/10.1016/j.jksuci.2012.10.003.

[2] Mahedy Hasan, S. M., Rabbi, M. F., Champa, A. I., & Zaman, M. A. (2020). An Effective Diabetes Prediction System Using Machine Learning Techniques.2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT). https://doi.org/10.1109/icaict51780.2020.9333497.

[3] S. Sivaranjani, S. Ananya, J. Aravinth, and R. Karthika, "Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS),Coimbatore, India, 2021, pp. 141-146. https://doi.org/10.1109/ICACCS51430.2021.9441935.

[4] Arora, R., Suman. (2012). Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications, 54, 21–25. https://doi.org/10.5120/8626-2492.

[5] Sharief, Ahlam & Sheta, Alaa. (2014). Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming. International Journal of Advanced Computer Science and Applications (IJACSA),3, 54-59. https://doi.org/10.14569/IJARAI.2014.031007.

[6] Jaiswal, Mansi & Tiwari, Harsh & Kumar, Sanjeev. (2023). Diabetes Prediction using Optimization Techniques with Machine Learning Algorithms. International Journal of Electronic Healthcare, 13,1. https://doi.org/10.1504/IJEH.2023.10054229