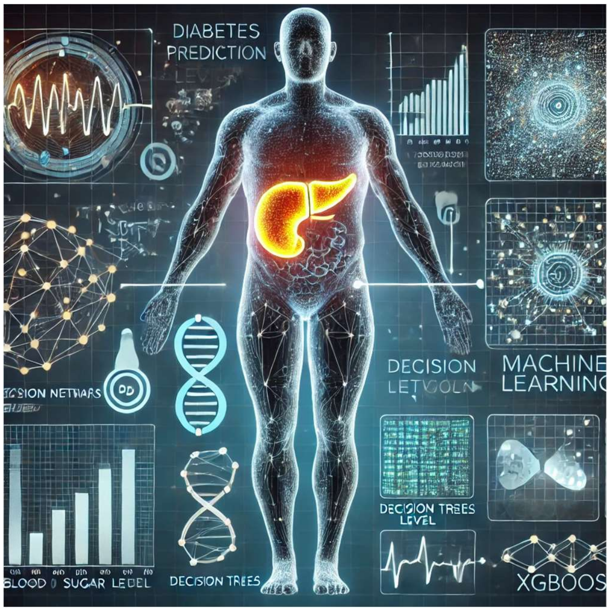
Diabetes Prediction Project Report

1. Introduction



Diabetes is a chronic medical condition that affects how the body processes blood sugar (glucose). If left untreated, diabetes can cause many complications. Acute complications can include diabetic ketoacidosis, hyperosmolar hyperglycemic state, or death. Serious long-term complications include cardiovascular disease, stroke, chronic kidney disease, foot ulcers, and damage to the eyes.Early detection and effective management are crucial for improving patient outcomes. This project aims to develop a predictive model for diabetes diagnosis using various machine learning algorithms on a dataset comprising medical predictor variables.

2.Dataset Overview

- This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

- **Features**: The dataset includes the following variables:
  - **Pregnancies**: Number of pregnancies.
  - **Glucose**: Plasma glucose concentration.
  - **BloodPressure**: Blood pressure (diastolic).
  - **SkinThickness**: Skin fold thickness.
  - **Insulin**: Insulin levels.
  - **BMI**: Body mass index.
  - **DiabetesPedigreeFunction**: A function that scores likelihood of diabetes based on family history.
  - **Age**: Age of the patient.
  - **Outcome**: Class variable (0 or 1) indicating diabetes presence.

3.Data Preprocessing

The data preprocessing phase involved several critical steps to ensure the dataset's quality and usability for modeling.

First, outliers were identified and removed using the Local Outlier Factor method, enhancing the integrity of the data. Next, new variables were created based on existing features, categorizing BMI, insulin levels, and glucose levels into meaningful groups.

Categorical variables were then transformed into numerical values through One Hot Encoding, which helped prevent the dummy variable trap. To ensure the models could perform optimally, the dataset was standardized using the RobustScaler method, making the feature distributions more uniform. This comprehensive preprocessing approach laid the foundation for effective model training and evaluation.

4. Model Selection and Evaluation

4.1 Models Implemented

A variety of machine learning models were implemented to assess their performance in predicting diabetes outcomes based on the processed dataset. The following models were utilized:

- **Logistic Regression (LR)**: A statistical model that predicts the probability of a binary outcome based on one or more predictor variables, using a logistic function to model the relationship.
- **K-Nearest Neighbors (KNN)**: A non-parametric method used for classification that predicts the class of a data point based on the classes of its nearest neighbors in the feature space.
- **Decision Tree Classifier (CART)**: A model that uses a tree-like structure to make decisions based on feature splits, effectively handling both classification and regression tasks.
- **Random Forest Classifier (RF)**: An ensemble learning method that combines multiple decision trees to improve classification accuracy and control overfitting by averaging the predictions of individual trees.
- **Support Vector Machine (SVM)**: A supervised learning model that finds the optimal hyperplane to separate classes in high-dimensional spaces, effective for both linear and non-linear classification tasks.
- **Gradient Boosting Classifier (XGB)**: An ensemble technique that builds models sequentially, where each new model attempts to correct the errors made by the previous ones, particularly useful for handling complex datasets.
- **LightGBM Classifier**: A gradient boosting framework that uses tree-based learning algorithms, designed for high efficiency and performance, particularly on large datasets.

4.2 Cross-Validation Scores

To evaluate the models, 10-fold cross-validation was performed, ensuring robust performance metrics by partitioning the dataset into ten subsets. Each model was trained and validated multiple times to obtain a reliable estimate of its accuracy. The summarized results of the cross-validation process are as follows

Model	**Mean Accuracy**	**Standard Deviation**
Logistic Regression	0.8487	0.0369
KNN	0.8408	0.0239
CART	0.8579	0.0248
Random Forest	0.8816	0.0263
SVM	0.8539	0.0365
XGB	0.8908	0.0204
LightGBM	0.8855	0.0243

Overall, the model evaluation phase demonstrated that the ensemble methods, particularly Random Forest and Gradient Boosting, outperformed simpler models, confirming their effectiveness for the diabetes prediction task.

4.3 Model Tuning

Model tuning is a crucial step in the machine learning process that involves optimizing the hyperparameters of models to enhance their performance. For this project, hyperparameter tuning was conducted specifically for the Random Forest, LightGBM, and Gradient Boosting (XGBoost) classifiers, which showed the best performance in the initial evaluations.

The tuned models were then validated again using cross-validation to ensure their performance improvements were consistent and reliable.

Overall, the tuning process significantly enhanced the models' predictive capabilities, ensuring they were well-suited for the diabetes prediction task.

5. Reporting

The aim of this study was to develop classification models for a diabetes dataset to predict whether an individual has diabetes, with a focus on maximizing validation scores. The project encompassed several key stages:

- **Data Acquisition**: The diabetes dataset was read and prepared for analysis.
- **Exploratory Data Analysis (EDA)**: Structural properties of the dataset were examined, including variable types and size. Notably, zero values in the dataset were identified as missing values and replaced with NaN. Descriptive statistics were generated to better understand the data distribution.

- **Data Preprocessing**: Missing values were addressed by filling NaN entries with the median values based on whether the individual was diabetic or not. Outliers were detected using the Local Outlier Factor (LOF) method and subsequently removed. The feature variables (X) were standardized using the RobustScaler method to ensure optimal performance of the models.
- **Model Building**: Various machine learning models were implemented, including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), CART, Random Forest, XGBoost, and LightGBM. Cross-validation scores were calculated to assess model performance. Following initial evaluations, hyperparameter tuning was conducted for the Random Forest, XGBoost, and LightGBM models to enhance their accuracy.
- **Results**: The final model, resulting from the hyperparameter optimization of the XGBoost classifier, achieved the highest cross-validation score of **0.90**, indicating its effectiveness in predicting diabetes. This outcome demonstrates the model's potential for reliable classification, paving the way for future applications in medical diagnostics.

In conclusion, the project successfully developed robust classification models for diabetes prediction, highlighting the importance of data preprocessing and model tuning in achieving high performance. The **XGBoost model** stands out as the most effective approach, showcasing the power of advanced machine learning techniques in healthcare applications.