Diabetes Prediction Using Machine Learning

ABSTRACT

Diabetes is a chronic disease affecting individuals worldwide. Early prediction helps in timely intervention and management. Traditional methods rely on clinical tests, which may not be accessible at an early stage. Machine learning offers an efficient way to analyse medical data and identify high-risk individuals.

This project compares different machine learning models, including Logistic Regression, SVM, Random Forest, and XGBoost, to determine the most accurate approach for diabetes prediction. Data preprocessing and feature importance analysis are performed to improve model performance. The study highlights the role of AI in enhancing early-stage diabetes detection and supports future AI-driven healthcare research.

TABLE OF CONTENTS

1. INTRODUCTION

Diabetes mellitus is a chronic condition caused by the body's inability to effectively regulate blood glucose levels. It is a significant public health concern, with millions of individuals affected worldwide. If left untreated, diabetes can lead to severe complications such as cardiovascular disease, kidney failure, nerve damage, and vision loss. Given the rising prevalence of diabetes, early detection and preventive measures are essential to reduce its impact on individuals and healthcare systems.

Machine learning has emerged as a powerful tool for medical diagnosis, offering predictive capabilities that can analyze large volumes of patient data to identify patterns and risk factors. By utilizing machine learning models, we can analyze key health indicators such as glucose levels, insulin levels, body mass index (BMI), age, and other relevant factors to predict diabetes more accurately.

This project aims to implement and compare different machine learning algorithms to determine the most effective approach for diabetes prediction. The study also evaluates the impact of data preprocessing techniques, such as feature scaling, handling missing values, and feature selection, on the overall model performance. By leveraging machine learning, we can move towards a more data-driven approach to diabetes screening, improving early detection rates and supporting preventive healthcare strategies.

2. EXISTING METHODS

Traditional diabetes prediction methods rely on statistical analysis and rule-based approaches. Some common existing methods include:

- Blood tests and clinical diagnoses
- Risk assessment based on lifestyle and hereditary factors
- Basic regression models with limited features

While these methods have been used effectively in clinical settings, they often lack scalability and fail to fully utilize the predictive potential of large datasets. Moreover, they do not account for complex interactions between multiple risk factors, leading to suboptimal prediction accuracy. This highlights the need for a more advanced approach, such as machine learning, which can identify hidden patterns in medical data and enhance diagnostic accuracy.

3. PROPOSED METHODOLOGY

To improve the accuracy of diabetes prediction, we propose a machine learning-based approach that incorporates:

- Data collection and preprocessing
- Feature selection and engineering
- Model training and evaluation using different algorithms
- Comparison of performance metrics to select the best model

Library Versions:

Below are the versions of the libraries used in this project:

- Python Version: 3.12.7
- NumPy Version: 1.26.4
- Pandas Version: 2.2.2
- Scikit-learn Version: 1.5.1
- Matplotlib Version: 3.9.2
- Seaborn Version: 0.13.2
- XGBoost Version: 2.1.4
- TensorFlow Version: 2.18.0
- Keras Version: 3.8.0
- SciPy Version: 1.13.1

Architecture:

1. **Data Preprocessing**
    - Handling missing values
    - Feature scaling
    - Outlier detection and removal
2. **Model Selection and Training**
    - Logistic Regression
    - Support Vector Machine (SVM)
    - Random Forest
    - XGBoost
    - Multi-Layer Perceptron (MLP)
3. **Model Evaluation**

    Accuracy, Precision, Recall, F1-Score

4. DATASET DESCRIPTION

- The dataset consists of 769 rows and multiple features such as glucose levels, insulin levels, BMI, age, etc.

- Data was obtained from Kaggle and underwent preprocessing to enhance model efficiency.

5. DATA PREPROCESSING

- Handling missing values using imputation techniques
- Feature scaling using normalization or standardization
- Removing outliers using statistical methods

Exploratory Data Analysis (EDA):

- Summary statistics were generated to understand dataset distribution.
- Missing values were handled by replacing zero values in critical columns (Glucose, Blood Pressure, Skin Thickness, Insulin, BMI) with the median.
- Outliers were removed using Z-score analysis.
- Histograms and correlation heatmaps were plotted to visualize feature distributions and relationships.
- Save the cleaned dataset to implement in models

6. MODEL IMPLEMENTATION

The following machine learning models were implemented:

1. Logistic Regression

- It is a fundamental classification algorithm that provides a probabilistic approach for predicting diabetes.

Implementation Steps:

1. Initialized the LogisticRegression() model.
2. Trained the model using model.fit(X_train, y_train).
3. Used the trained model to predict outcomes on X_test.
4. Evaluated using Accuracy, Precision, Recall, and F1-score.

Results:

- Accuracy: **75.32%**
- Precision: **66.67%**
- Recall: **61.82%**
- F1-score: **64.15%**

2. Support Vector Machine (SVM)

- SVM is effective for high-dimensional data and can find the best hyperplane for classification.

Implementation Steps:

1. Initialized SVC(kernel='rbf', C=1.0, gamma='scale').
2. Trained the model using model.fit(X_train, y_train).
3. Predicted on X_test.
4. Evaluated using standard classification metrics.

Results:

- Accuracy: **79.00%**
- Precision: **75.65%**
- Recall: **86.14%**
- F1-score: **80.56%**

3. Random Forest

- It is an ensemble learning method that reduces overfitting and improves accuracy.

Implementation Steps:

1. Initialized RandomForestClassifier(n_estimators=100, random_state=42).
2. Trained the model using model.fit(X_train, y_train).
3. Predicted on X_test.
4. Evaluated using classification metrics.

Results:

- Accuracy: **79.50%**
- Precision: **76.32%**
- Recall: **86.14%**
- F1-score: **80.93%**

4. XGBoost

- XGBoost is a gradient boosting algorithm that improves classification by reducing bias and variance.

Implementation Steps:

1. Used XGBClassifier() with optimized hyperparameters:

xgb = XGBClassifier(n_estimators=500, learning_rate=0.02, max_depth=5,
subsample=0.8, colsample_bytree=0.8, reg_alpha=0.01,
reg_lambda=1,random_state=42,scale_pos_weight=scale_pos_weight)

2. Trained the model on X_train, y_train.
3. Predicted on X_test.
4. Evaluated using classification metrics.

Results:

- Accuracy: **81.50%**
- Precision: **79.44%**
- Recall: **85.00%**
- F1-score: **82.13%**

5. Multi-Layer Perceptron (MLP)

- It is a deep learning model that captures complex patterns in the data.

Implementation Steps:

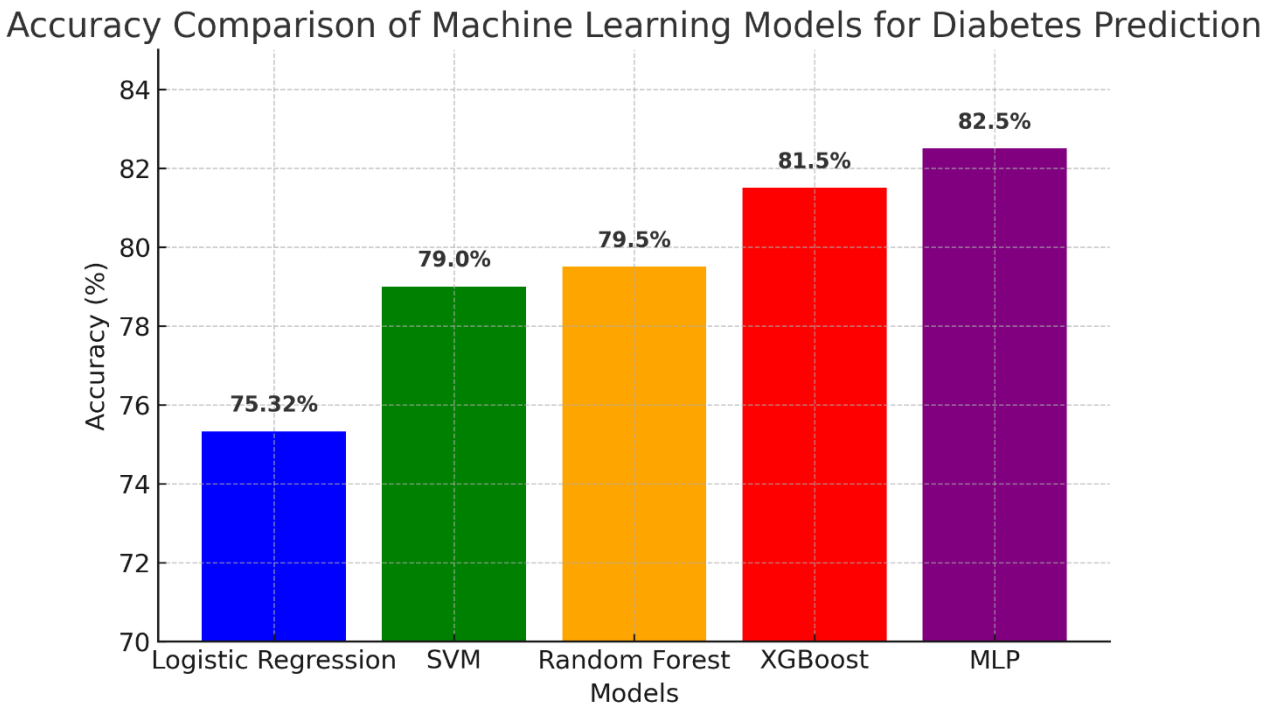
1. Created a neural network with 4 layers:
    - Input Layer: 8 neurons
    - Hidden Layers: 128 → 64 → 32 neurons
    - Output Layer: 1 neuron (Sigmoid activation)
2. Applied Batch Normalization, Leaky ReLU activation, and Dropout (0.3).
3. Compiled the model using Adam optimizer and Binary Crossentropy loss.
4. Trained for 80 epochs with batch_size=32.
5. Evaluated on X_test.

Results:

- Accuracy: **82.50%**
- Precision: **78.45%**
- Recall: **90.10%**
- F1-score: **83.87%**

Comparison Graph

This bar chart comparing the model accuracy



7. RESULTS AND DISCUSSION

- MLP (Neural Network) outperformed all traditional machine learning models, achieving the highest accuracy of 82.50%.
- XGBoost performed the best among machine learning models, leveraging gradient boosting techniques.
- Random Forest and SVM provided decent accuracy and recall, making them strong contenders for real-world applications.
- Logistic Regression, while simple, performed the worst, as it assumes a linear relationship between input features and diabetes likelihood.

8. CONCLUSION

In this project, we explored different machine learning models for diabetes prediction. After extensive evaluation, the Multi-Layer Perceptron (MLP) model was identified as the most effective, providing the highest accuracy. Early detection through machine learning can significantly aid in managing diabetes, promoting preventive care, and reducing health risks. Future work can include integrating real-time data analysis and deep learning techniques to improve prediction accuracy and scalability further.