Sri Lanka Institute of Information Technology

IT2011

Artificial Intelligence and Machine Learning

2025

Year 2 Semester 1

Group ID: 2025-Y2-S1-MLB-B10G1-09

Project Topic: Diabetes Prediction Using Machine Learning

Group Members

Student ID	Name
IT24102649	Liyanage D C J
IT24102651	Disanayaka R M K S
IT24103838	Dissanayaka D. M. H
IT24102605	Punchihewa S.D
IT24102643	Anuraheesara U.A.S
IT24102575	Madhusanka S.P

1. Introduction and Problem Statement

Diabetes is increasingly prevalent globally, with both healthcare costs and personal expenses rising as a result. Early detection of diabetes is crucial as it helps to:

- **Reduce medical expenses**
- **Prevent chronic complications**
- **Facilitate lifestyle changes for better health outcomes**

Machine learning (ML) can aid in the early detection of diabetes by analyzing various health predictors such as age, BMI, physical activity, and smoking history. This project seeks to develop a machine learning model that predicts the risk of diabetes in individuals based on these attributes.

2. Dataset Description

- **Dataset Name and Source**:
  - **Healthcare Diabetes Dataset** from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).
  - Available at: Kaggle Healthcare Diabetes Dataset
- **Dataset Size**:
  - **Number of Records**: 2,768 records
  - **Number of Features**: 9 attributes
- **Target Variable**:
  - **Outcome (Binary Classification)**
    - **0** = Non-diabetic
    - **1** = Diabetic
- **Key Features**:
  - **Demographic**:
    - Age
    - Gender
  - **Clinical/Health-related**:
    - BMI (Body Mass Index)
    - Blood Pressure
    - Insulin Levels
    - Glucose
    - Physical Activity
    - Smoking History
    - Cholesterol Levels

3. Preprocessing & EDA

3.1 Data Cleaning and Missing Values

The dataset used for this project was clean, with no missing values in any of the 9 features. This ensured that no imputation or data preprocessing steps related to missing values were required.

3.2 Exploratory Data Analysis (EDA)

A thorough EDA was conducted to explore the characteristics of the data and identify any trends or anomalies:

- **Age Distribution**:
  The age distribution was fairly balanced with no significant outliers. The dataset represented a wide range of ages, making it suitable for diabetes prediction across different age groups.
- **Outliers**:
  Several features exhibited potential outliers, including BMI, Blood Pressure, and Cholesterol levels. These were carefully treated using outlier handling techniques such as **Winsorization**, where extreme values were capped at the 5th and 95th percentiles to prevent them from skewing the analysis.
- **Feature Correlation**:
  A correlation matrix was used to explore relationships between features. The analysis revealed:
  - **BMI and Diabetes**: A significant positive correlation between BMI and the likelihood of being diabetic. Higher BMI values were associated with higher chances of diabetes.
  - **Glucose Levels and Outcome**: Higher glucose levels were strongly associated with a higher probability of diabetes.
  - No strong **multicollinearity** was observed between the features, ensuring that the models could work effectively without issues of feature redundancy.

3.3 Cross Analysis (Target vs. Features)

A deeper analysis was performed to understand the relationship between the features and the target variable (Outcome):

- **BMI**: Higher BMI values showed a strong correlation with diabetes, with non-diabetic individuals typically having lower BMI.
- **Blood Pressure**: Elevated blood pressure was more common in diabetic individuals, suggesting a possible link between hypertension and diabetes.
- **Glucose Levels**: Elevated glucose levels were a strong predictor of diabetes, with individuals who had higher blood glucose values being more likely to have diabetes.
- **Age**: Older individuals had a higher likelihood of being diabetic. The age distribution of diabetic individuals showed a gradual increase in prevalence with age.

3.4 Feature Transformation

Several transformations were applied to improve the performance of the machine learning models:

- **Outlier Treatment**:
  Winsorization was used to limit the influence of extreme values in features like **BMI**, **Blood Pressure**, and **Cholesterol**.

- **Feature Encoding**:
  Categorical features like **Gender**, **Physical Activity**, and **Smoking History** were encoded using **One-Hot Encoding** to convert them into numerical representations.

- **Feature Scaling**:
  All numerical features were standardized using **StandardScaler** to ensure that all features had the same scale, which is crucial for machine learning models like SVM and Logistic Regression.

- **Dimensionality Reduction**:
  Although **Principal Component Analysis (PCA)** was explored as a dimensionality reduction technique, it was decided to retain all features, as they all contributed meaningfully to the model.

4. Model Design and Implementation

The prediction task is a **binary classification problem**. The goal is to predict whether an individual is diabetic (Outcome = 1) or not (Outcome = 0) based on various health and demographic features. The models were evaluated on a standard **80/20 train-test split** (test_size = 0.2).

Class imbalance mitigation

To correct the class imbalance identified in the EDA, the **Synthetic Minority Over-sampling Technique (SMOTE)** was applied to the training data. This technique generated synthetic instances of the minority class (diabetic individuals) to achieve a balanced training set, thus ensuring that the models are not biased towards the majority class (non-diabetic individuals).

Models Implemented :

1. Random Forest
2. Neural Network(MLP)
3. Decision Tree
4. Support Vector Machine (SVM)
5. Logistic Regression
6. Xg boost

5. Evaluation and Comparison

1.Random Forest

- Accuracy: ~98%
- Precision: ~99%
- Recall: ~97%
- F1 Score: ~98%
- AUC Score: 99%

2.Neural Network(MLP)

- Accuracy: ~73%
- Precision: ~64%
- Recall: ~84%
- F1 Score: ~80%

3.Decision Tree

- Accuracy: ~ 99%
- Precision: ~ 100%
- Recall: ~98%
- F1 Score: ~99%

4. Support Vector Machine (SVM) Model

- Accuracy: ~ 97%
- Precision: ~ 99%
- Recall: ~93%
- F1 Score: ~96%

5.Logistic Regression

- Accuracy: ~ 78%
- Precision: ~68%
- Recall: ~73%
- AUC: ~87%
- F1 Score: ~71%

6. XG Boost

- Accuracy: ~ 93%
- Precision: ~93%
- Recall: ~93%
- AUC: ~98%
- F1 Score: ~93%

6.Evaluation Summary

The models were evaluated based on key metrics such as accuracy, precision, recall, F1-score, and AUC. **Random Forest** and **Decision Tree** models showed the best performance, with high accuracy, precision, and recall, making them highly effective for predicting diabetes. The **Decision Tree (Tuned)** model, in particular, achieved perfect recall, ensuring that all diabetic cases were correctly identified, which is critical in healthcare applications.

Neural Network (MLP) had lower precision, indicating it misclassified more non-diabetic individuals. **SVM** and **XGBoost** also performed well but did not outperform the decision tree in terms of recall and interpretability.

The **Decision Tree (Tuned)** was chosen as the final model due to its perfect recall, simplicity, and transparency, which are vital for healthcare settings where model decisions need to be clearly understood.

7. Ethical Considerations and Bias Mitigation

Ethical considerations

Since the project involves sensitive health-related data, **data privacy** is of paramount importance. All personal information in the dataset, such as **age**, **gender**, and **health conditions** (BMI, blood pressure, glucose levels), were anonymized to prevent any identification of individual participants. In addition, all data handling was done in compliance with international data privacy regulations such as **GDPR (General Data Protection Regulation)** and **HIPAA (Health Insurance Portability and Accountability Act)**.

All patient data were anonymized to ensure compliance with data privacy regulations, such as GDPR.

Bias Mitigation

- **Class Imbalance**: SMOTE was used to mitigate class imbalance, ensuring that the minority class (diabetic individuals) was adequately represented during training.
- **Algorithmic Bias**: Performance metrics like **Recall** and **F1-Score** were prioritized over **Accuracy** to ensure the model did not favor the majority class (non-diabetic individuals).
- **Demographic Bias**: The dataset includes features like **Age**, **Gender**, and **BMI**, ensuring that predictions account for different demographic groups. However, further work should include more diverse datasets for better generalization.

8.Reflections and Lessons Learned

Challenges Faced

1. **Class Imbalance**:
   The dataset had more non-diabetic individuals than diabetic ones, which could lead to biased predictions. This imbalance was addressed using **SMOTE** to generate synthetic data for the minority class.
2. **Overfitting and Underfitting**:
   Some models overfitted the training data, performing well on training but poorly on test data. Finding the right balance between model complexity and generalization was a challenge.
3. **Model Selection**:
   Choosing the best model involved testing multiple algorithms. While **Logistic Regression** was the baseline, more complex models like **Random Forest** and **SVM** required fine-tuning to avoid overfitting while maintaining high

Key Learnings

1. **Class Imbalance Handling**:
   SMOTE helped improve model performance by balancing the dataset, with **Recall** becoming the most important metric to ensure diabetic individuals were identified.
2. **Feature Importance**:
   BMI, **Glucose levels**, and **Blood Pressure** were the most important features for predicting diabetes, based on feature analysis.
3. **Model Evaluation**:
   Recall and **F1-Score** were prioritized over **Accuracy** to reduce the risk of missing diabetic patients, even if it slightly lowered the model's overall accuracy

Improvements for Future Work

1. **External Validation**:
   The model should be validated on a larger and more diverse dataset to improve generalization.
2. **Model Explainability**:
   XGBoost and **Random Forest** are "black-box" models. Using techniques like **SHAP** or **LIME** would help make the models more interpretable for healthcare professionals.
3. **Integration of Other Predictors**:
   Future models could include additional features like **family history** and **lifestyle factors** to make the predictions more comprehensive.

9.References

1. Public Health Dataset ,“Heart Disease Dataset’’ Kaggle.
   Available : https://www.kaggle.com/datasets/nanditapore/healthcare-diabetes

2. World Health Organization (WHO), "The Top 10 Causes of Death," 2024.