CSE422 Lab Project Report: Diabetes Prediction Using Machine Learning Models

Md. Imam Hasan
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
md.imam.hasan@g.bracu.ac.bd

Wasif Azraf
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
wasif.azraf@g.bracu.ac.bd

***Abstract*—This paper investigates the application of machine learning techniques for diabetes prediction using a highly imbalanced dataset of 100,000 patient records with 9 features. Five classification models—Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Random Forest, and Neural Network—are implemented and evaluated. The dataset is pre-processed by categorizing BMI and age, imputing missing values, scaling numerical features, and encoding categorical variables. The class imbalance, with 78.37% non-diabetic and 7.27% diabetic instances, poses a significant challenge, addressed without oversampling techniques like SMOTE. Results show Random Forest and Decision Tree outperforming others, achieving accuracies of 0.9669 and 0.9663, respectively, with F1-scores of 0.76 for the diabetic class and AUCs of 0.9614 and 0.9631. The study highlights the challenges of imbalanced medical data and the effectiveness of tree-based models in such scenarios.**

***Index Terms*—Diabetes Prediction, Machine Learning, Imbalanced Dataset, Medical Diagnosis**

I. Introduction

Diabetes mellitus, a chronic condition marked by elevated blood glucose levels, affects over 400 million people globally, posing risks of severe complications such as cardiovascular disease and kidney failure if undiagnosed. Early detection is critical for effective management, and machine learning offers a promising approach to assist healthcare professionals in identifying at-risk patients. This project focuses on predicting diabetes using a dataset with significant class imbalance, aiming to achieve high recall for the diabetic class to minimize missed diagnoses, which are critical in medical contexts.

The dataset, sourced from Kaggle [1], includes 100,000 patient records with features like age, BMI, and blood glucose levels. We evaluate five machine learning models—Logistic Regression, Decision Tree, KNN, Random Forest, and Neural Network—on their ability to classify patients as diabetic or non-diabetic. This report details the dataset, preprocessing steps, model training, and performance evaluation, emphasizing the impact of class imbalance on model outcomes.

II. Dataset Description

A. Overview

The *Diabetes Dataset* contains 100,000 records, each with 9 features: 7 numerical (`age`, `hypertension`, `heart_disease`, `bmi`, `HbA1c_level`, `blood_glucose_level`, `diabetes`) and 2 categorical (`gender`, `smoking_history`). The target variable, `diabetes`, is binary, where 0 indicates non-diabetic and 1 indicates diabetic.

B. Dataset Details

The features are categorized as follows:

- **Numerical Features**:
  - `age` — 102 unique values
  - `hypertension` — binary (0 or 1)
  - `heart_disease` — binary (0 or 1)
  - `bmi` — 4,174 unique values
  - `HbA1c_level` — 18 unique values
  - `blood_glucose_level` — 18 unique values
  - `diabetes` — binary (0 or 1)
- **Categorical Features**:
  - `gender` — 3 categories: *male*, *female*, *other*
  - `smoking_history` — 6 categories: *never*, *former*, *current*, etc.

C. Correlation Analysis

A correlation matrix was computed to examine pairwise relationships among the numerical features. The resulting Pearson correlation coefficients are visualized in the heatmap shown in Fig. 1.

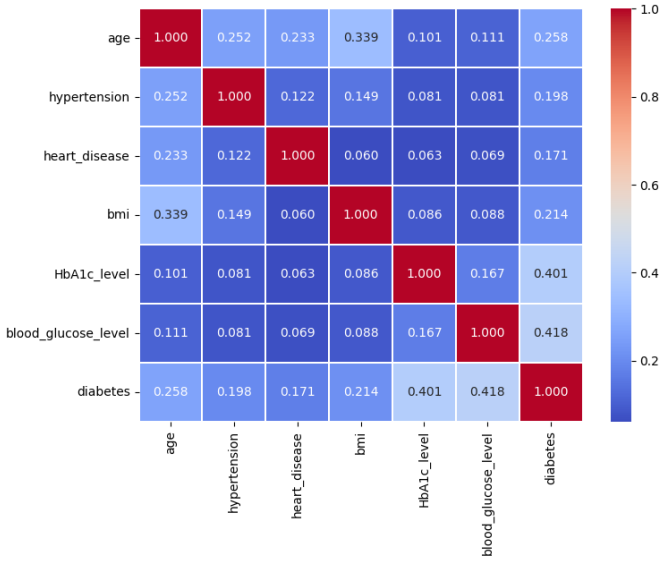

Fig. 1: Pearson correlation matrix for numerical features.

Key observations:

- Patients with `HbA1c_level` > 6.5% and `blood_glucose_level` > 140 mg/dL are more likely to have diabetes.
- Individuals over 50 years show higher diabetes prevalence.
- `BMI` > 30 kg/m^2 indicates increased risk, linking obesity to insulin resistance.
- Former smokers and males exhibit slightly higher diabetes prevalence.

D. Imbalanced Dataset

The dataset exhibits significant class imbalance, with 82,284 non-diabetic instances (78.37%) and 7,634 diabetic instances (7.27%), a ratio of approximately 11:1. This is visualized in Fig. 2, highlighting the disparity.

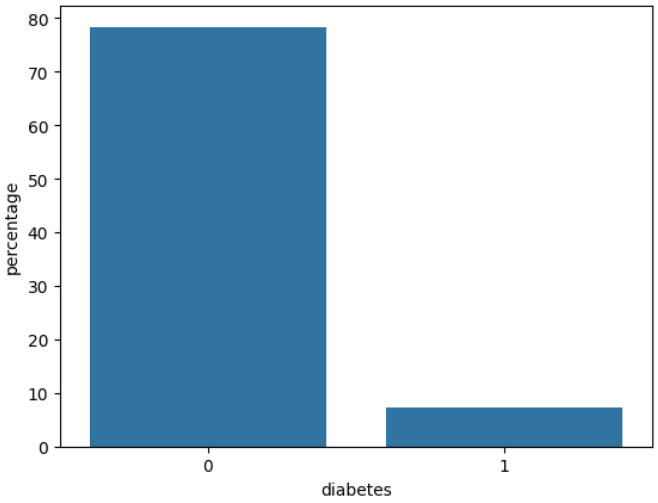

Fig. 2: Class distribution of `diabetes`.

E. Data Skewness

Box-and-whisker plots were used to examine the skewness of numerical features (see Fig. 3). The plots reveal that:

- `diabetes`, `hypertension`, and `heart_disease` are highly skewed (skewness values: 2.9785, 3.2178, and 4.7528, respectively), reflecting their binary/imbalanced nature.
- `bmi` (skewness 1.0496) and `blood_glucose_level` (skewness 0.8161) exhibit moderate positive skew.

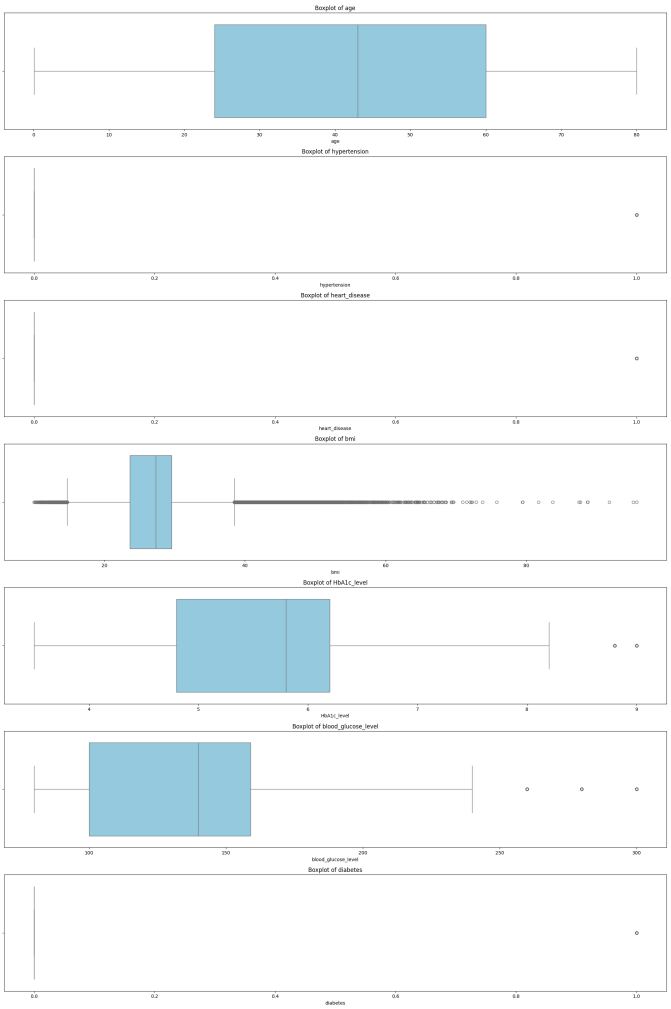

Fig. 3: Box-and-whisker plots showing skewness of numerical features.

F. Numerical Histogram

Histograms of numerical features (`age`, `bmi`, `HbA1c_level`, `blood_glucose_level`) provide insights into their distributions, as shown in Fig. 4. These distributions guide preprocessing decisions, such as scaling.

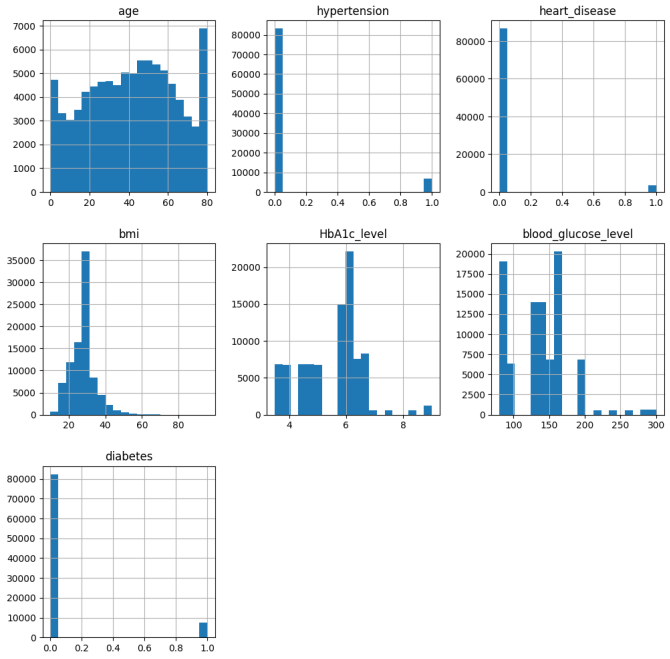

Fig. 4: Histogram of numerical features.

G. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to uncover patterns, relationships, and anomalies within the dataset. Key observations include:

- **Medical Indicators and Diabetes:**
  - Patients with `HbA1c_level` > 6.5% and `blood_glucose_level` > 140 mg/dL show a significantly higher likelihood of having diabetes, aligning with medical diagnostic thresholds.
  - `BMI` values above 30 kg/m^2 are associated with increased diabetes risk, suggesting a strong link between obesity and insulin resistance.
- **Demographic Insights:**
  - Individuals aged over 50 demonstrate higher diabetes prevalence, supporting age as a risk factor.
  - Males show a slightly higher proportion of diabetes cases compared to females and others.
- **Lifestyle Factors:**
  - Former smokers tend to have higher diabetes incidence than current or never smokers, possibly reflecting delayed health impacts from smoking.

These insights informed feature selection and emphasized the importance of both clinical and lifestyle variables in predicting diabetes.

III. Dataset Pre-processing

A. Feature Engineering

Two new categorical features were created based on domain knowledge and data distribution:

- **BMI Category**: This feature categorizes individuals based on their Body Mass Index (BMI) into the following groups:
  - Underweight: BMI < 18.5
  - Normal: 18.5 –24.9
  - Overweight: 25 –29.9
  - Obese: BMI >= 30
- **Age Group**: This feature categorizes individuals into age groups to observe potential age-related trends:
  - Young: Age < 30
  - Middle-Aged: 30 –49
  - Senior: Age >= 50

B. Handling Missing Values and Encoding

To handle missing values and encode features appropriately, the following strategies were applied:

- **Numerical Features**: Features such as `age`, `bmi`, `HbA1c_level`, and `blood_glucose_level` were imputed using the median to handle missing values. Standard scaling was applied to normalize their range.
- **Categorical Features**: Features like `gender`, `smoking_history`, `bmi_category`, and `age_group` were imputed with the most frequent value (mode). One-hot encoding was applied to convert them into a suitable format for modeling.

A `ColumnTransformer` pipeline was implemented for efficient preprocessing:

- **Numerical Pipeline**: Median imputation followed by standard scaling using `StandardScaler`.
- **Categorical Pipeline**: Most-frequent imputation followed by one-hot encoding using `OneHotEncoder`.

C. Feature Scaling

For numerical features, standardization was applied to ensure they have a mean of 0 and a standard deviation of 1. This is critical for algorithms such as K-Nearest Neighbors (KNN) and Logistic Regression, which are sensitive to the scale of the input features.

D. Class Imbalance

The dataset exhibits a significant class imbalance with an 11:1 ratio between non-diabetic and diabetic instances. However, this imbalance was not addressed via oversampling techniques such as SMOTE. Instead, stratified sampling was employed during train-test splits to ensure the class distribution was preserved in both training and testing datasets.

IV. Dataset Splitting

The dataset was split into training (70%, 70,000 samples) and testing (30%, 30,000 samples) sets using stratified sampling to maintain the 11:1 class ratio.

V. Model Training and Testing

Five classification models were trained using the preprocessed data:

- **Logistic Regression**: A linear baseline model.
- **Decision Tree**: `min_samples_split=2`, `min_samples_leaf=2`, `max_depth=10`, `criterion='gini'`.
- **K-Nearest Neighbors (KNN)**: `weights='distance'`, `n_neighbors=20`, `metric='manhattan'`.
- **Random Forest**: `min_samples_split=8`, `min_samples_leaf=1`, `max_depth=15`, `criterion='entropy'`.
- **Neural Network**: `solver='sgd'`, `learning_rate='adaptive'`, `hidden_layer_sizes=(50,50)`, `alpha=0.0001`, `activation='tanh'`.

VI. Model Selection and Comparison Analysis

A. Evaluation Scores

The models were evaluated on accuracy, precision, recall, F1-score, and AUC. Results are summarized in Table I.

TABLE I: Evaluation Scores for All Models

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.96	0.88	0.56	0.68	0.95
Decision Tree	**0.97**	**0.98**	0.61	0.76	**0.96**
KNN	0.95	0.94	0.49	0.65	0.92
Random Forest	**0.97**	**1.00**	0.61	0.76	0.96
Neural Network	0.96	0.89	0.56	0.69	0.94

B. Accuracy Comparison

Accuracy is visualized in Fig. 5. Random Forest and Decision Tree lead with accuracies above 0.96, while KNN shows the lowest performance.

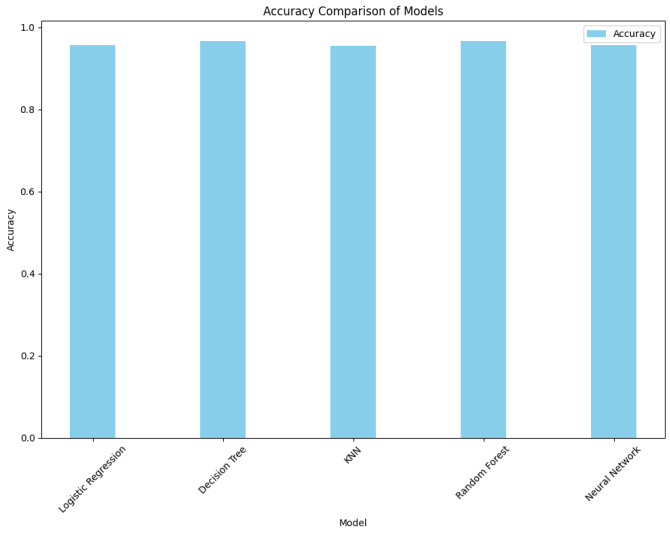

Fig. 5: Accuracy comparison of all models.

C. Precision and Recall Comparison

Precision and recall for the diabetic class are shown in Fig. 6. Random Forest's perfect precision and Decision Tree's near-perfect precision are notable, but their recall (0.61) indicates room for improvement.

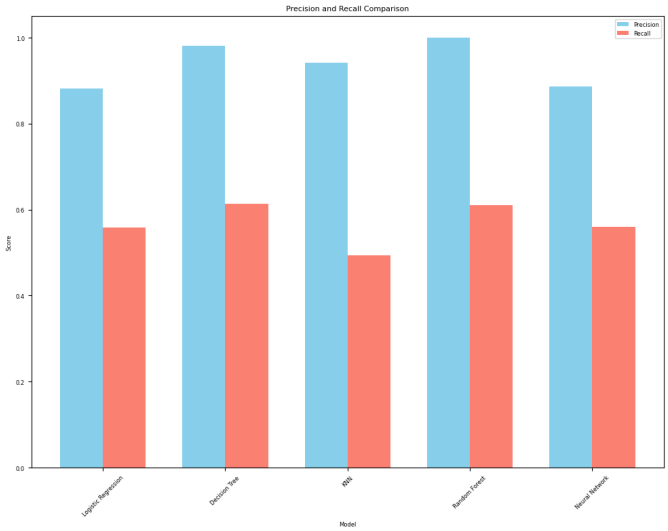

Fig. 6: Precision and recall for class 1.

D. Confusion Matrix

The confusion matrices for all models are shown in Fig. 7. The Random Forest model has no false positives, while the KNN model shows a higher number of false negatives, reflecting its lower recall.

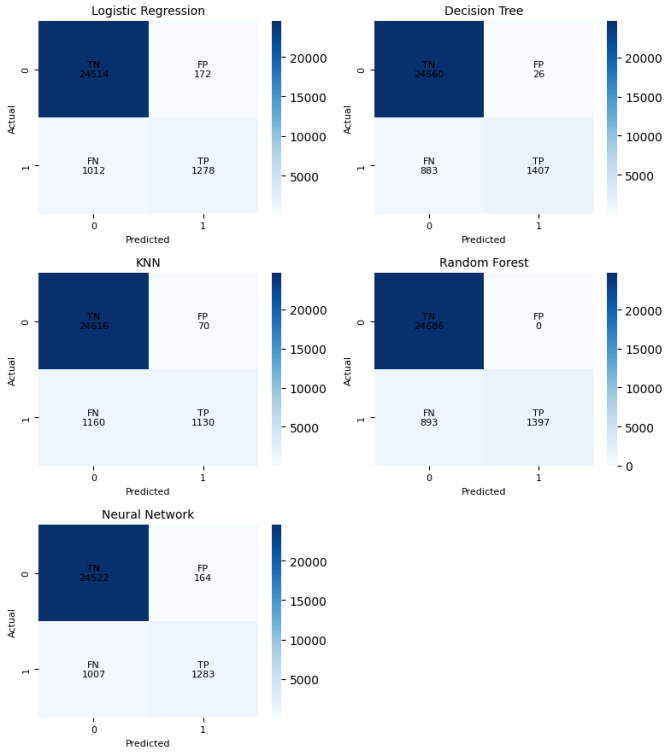

Fig. 7: Confusion matrices for all models.

E. AUC Score and ROC Curve

ROC curves are shown in Fig. 8. Decision Tree and Random Forest have the highest AUCs (0.9631 and 0.9614), while KNN has the lowest (0.9249).

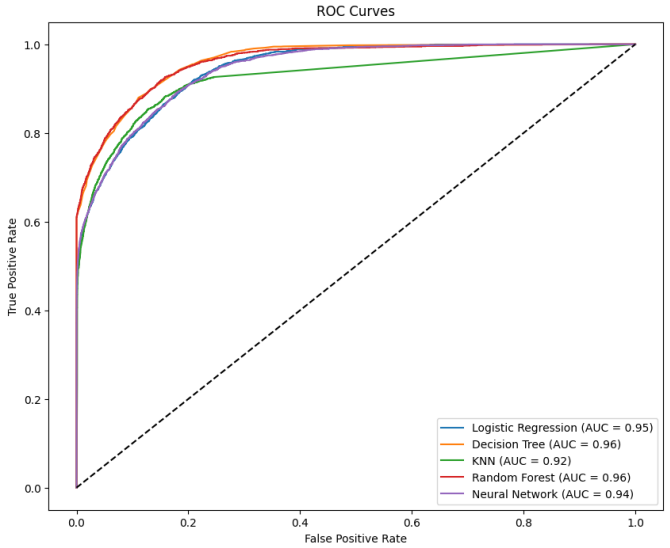

Fig. 8: ROC curves for all models.

VII. Conclusions

Random Forest and Decision Tree outperform others, with accuracies of 0.9669 and 0.9663, F1-scores of 0.76, and AUCs of 0.9614 and 0.9631. Random Forest's perfect precision ensures no false positives, though its recall (0.61) misses some diabetic cases. Logistic Regression and Neural Network show moderate performance (accuracies 0.956, F1-scores 0.68–0.69), while KNN struggles with recall (0.49). The 11:1 class imbalance, without SMOTE, highlights the challenge of prioritizing the minority class. Future work could explore class weighting or additional diabetic data to improve recall while maintaining precision.

References

[1] Mustafa T., "Diabetes Prediction Dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset