CSE422 Lab Project Report: Diabetes Prediction Using Machine Learning Models

1st Md. Imam Hasan
Dept. of Computer Science & Engineering
BRAC University
Dhaka, Bangladesh
md.imam.hasan@g.bracu.ac.bd

***Abstract*—This study investigates machine learning approaches for diabetes prediction using an imbalanced dataset of 100,000 patient records with nine clinical features. We evaluate five classification models—Logistic Regression, Decision Trees, K-Nearest Neighbors (KNN), Random Forests, and Neural Networks—following a rigorous preprocessing pipeline that includes BMI and age categorization, missing value imputation, feature scaling, and categorical encoding. Without employing synthetic oversampling techniques to address the significant class imbalance (78.37% non-diabetic versus 7.27% diabetic cases), our analysis using weighted evaluation metrics demonstrates that tree-based methods achieve superior performance. The Decision Tree classifier attains the highest accuracy (0.9670) and AUC score (0.9636), with Random Forest closely matching these results (accuracy: 0.9664, AUC: 0.9622). All models maintain robust weighted F1-scores between 0.95 and 0.96, though KNN shows relatively weaker discriminative ability (AUC: 0.8974). These findings highlight the effectiveness of tree-based algorithms in handling imbalanced medical data while preserving the natural class distribution, offering valuable insights for clinical decision support systems.**

***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. Exploratory Data Analysis (EDA)

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
  - `bmi` — 4,174 unique values
  - `HbA1c_level` — 18 unique values
  - `blood_glucose_level` — 18 unique values
- **Binary Categorical Features (represented numerically:)**:
  - `diabetes` — binary (0 or 1)
  - `hypertension` — binary (0 or 1)
  - `heart_disease` — binary (0 or 1)
- **Categorical Features**:
  - `gender` — 3 categories: *male*, *female*, *other*
  - `smoking_history` —`smoking_history` — 6 categories: *never*, *former*, *current*, *ever*, *not current*, *No Info*

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.

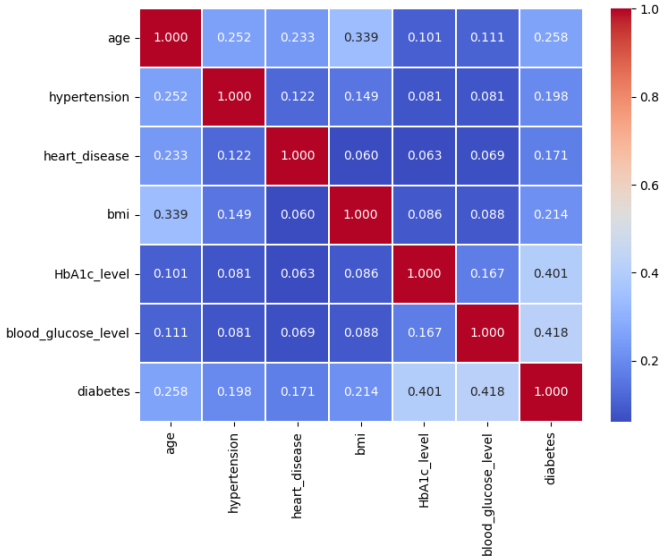

Fig. 1: Correlation Matrix.

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.

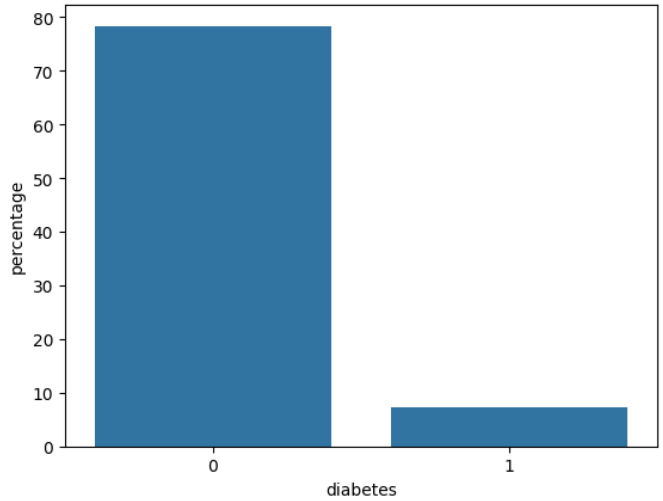

Fig. 2: Class distribution of `diabetes`.

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.

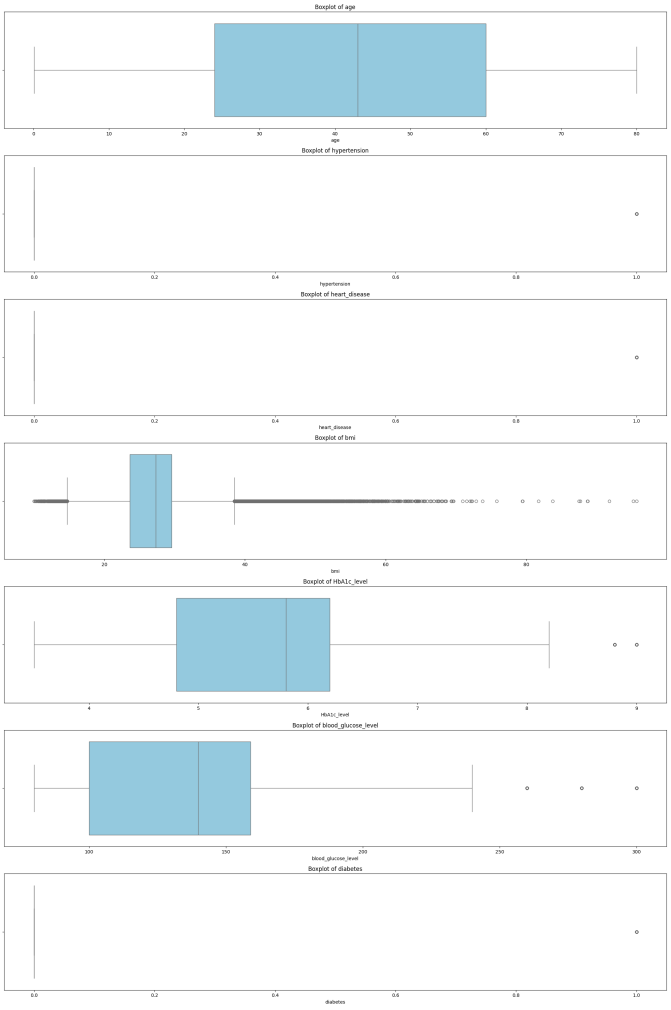

Fig. 3: Box-and-whisker plots showing skewness of numerical features.

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.

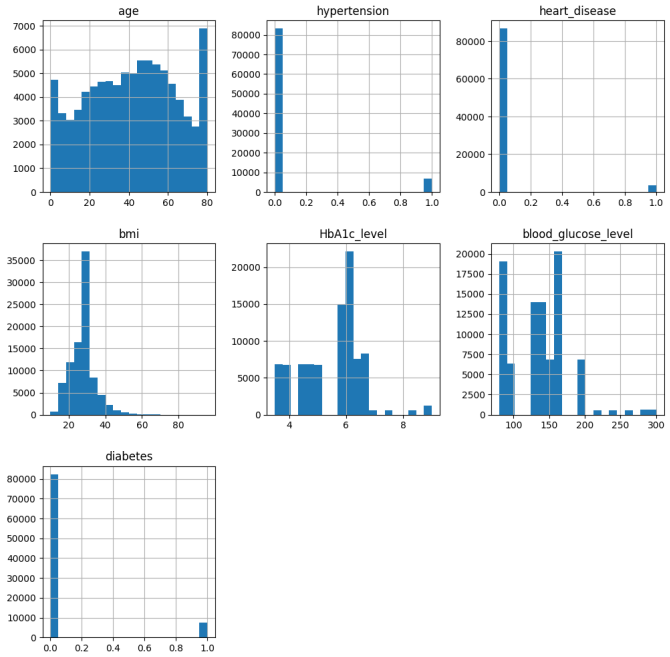

Fig. 4: Histogram of numerical features.

G. Summary of Key Findings

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. Dropping Missing Target Values

To ensure reliable training of the machine learning models, all rows with missing values in the target column `diabetes` were removed from the dataset.

B. Feature Engineering

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

- **BMI Category**: Categorizes individuals based on Body Mass Index (BMI) into:
  - Underweight: BMI < 18.5
  - Normal: 18.5 –24.9
  - Overweight: 25 –29.9
  - Obese: BMI >= 30
- **Age Group**: Groups individuals by age to capture potential age-related patterns:
  - Young: Age < 30
  - Middle-Aged: 30 –49
  - Senior: Age >= 50

C. Handling Missing Values and Encoding

The dataset contained missing values in both numerical and categorical features. The following preprocessing steps were applied:

- **Scaled Numerical Features** (`HbA1c_level`, `blood_glucose_level`): Missing values were imputed using the median. Features were standardized using `StandardScaler`.
- **Unscaled Numerical Features** (`hypertension`, `heart_disease`): Missing values were imputed using the most frequent value, and no scaling was applied due to their binary nature.
- **Categorical Features** (`gender`, `smoking_history`, `bmi_category`, `age_group`): Missing values were imputed using the most frequent category. One-hot encoding was applied using `OneHotEncoder` with `handle_unknown='ignore'` to manage unseen categories during inference.

D. Preprocessing Pipeline

A modular preprocessing pipeline was implemented using `Pipeline` and `ColumnTransformer` to streamline transformation:

- **Scaled Numerical Pipeline**: Median imputation followed by standard scaling.
- **Unscaled Numerical Pipeline**: Most frequent imputation only.
- **Categorical Pipeline**: Mode imputation followed by one-hot encoding.

E. Feature Scaling

Standardization was applied to continuous numerical features, specifically `HbA1c_level` and `blood_glucose_level`, to ensure each has a mean of 0 and a standard deviation of 1. This scaling is crucial for distance-based algorithms such as K-Nearest Neighbors (KNN), where feature magnitude directly affects distance calculations, as well as for gradient-based models like Logistic Regression and Neural Networks, which converge faster and perform more reliably when input features are on a comparable scale.

F. Class Imbalance

The dataset exhibits a significant class imbalance, with non-diabetic cases outnumbering diabetic cases by approximately 10.8:1. To mitigate bias during model evaluation, stratified sampling was applied during the train-test split to preserve class distribution. Oversampling techniques (e.g., SMOTE) were not used in this study.

IV. Dataset Splitting

After removing rows with missing values in the target variable, the dataset contained 89,918 samples across 11 features. The data was then split into training (70%, approximately 62,942 samples) and testing (30%, approximately 26,976 samples) sets using stratified sampling to preserve the original class imbalance ratio of approximately 11:1 (non-diabetic to diabetic).

V. Model Training and Evaluation

To classify the preprocessed dataset, five supervised learning models were developed and evaluated. The models, along with their respective configurations, are described below:

- **Logistic Regression**: Employed as a linear baseline model to establish a performance reference.
- **Decision Tree Classifier**: Configured with a maximum depth of 10, a minimum of 2 samples required to split an internal node (`min_samples_split=2`), and a minimum of 2 samples per leaf node (`min_samples_leaf=2`). The Gini impurity criterion was used for split quality measurement.
- **K-Nearest Neighbors (KNN)**: Implemented with 20 neighbors (`n_neighbors=20`), using the Manhattan distance metric and distance-based weighting (`weights='distance'`).
- **Random Forest Classifier**: Configured with a maximum depth of 15, `min_samples_split=8`, `min_samples_leaf=1`, and the entropy criterion for evaluating split quality.
- **Neural Network (MLPClassifier)**: Constructed with two hidden layers of 50 neurons each (`hidden_layer_sizes=(50, 50)`). The model used the stochastic gradient descent (`solver='sgd'`) optimization algorithm with an adaptive learning rate, `activation='tanh'`, and regularization parameter `alpha=0.0001`.

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 (Weighted Averages for precision, recall, f-1)

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.9554	0.95	0.96	0.95	0.9438
Decision Tree	**0.9670**	0.97	**0.97**	0.96	**0.9636**
KNN	0.9553	0.95	0.96	0.95	0.8974
Random Forest	0.9664	**0.97**	**0.97**	0.96	0.9622
Neural Network	0.9563	0.95	0.96	0.95	0.9435

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.

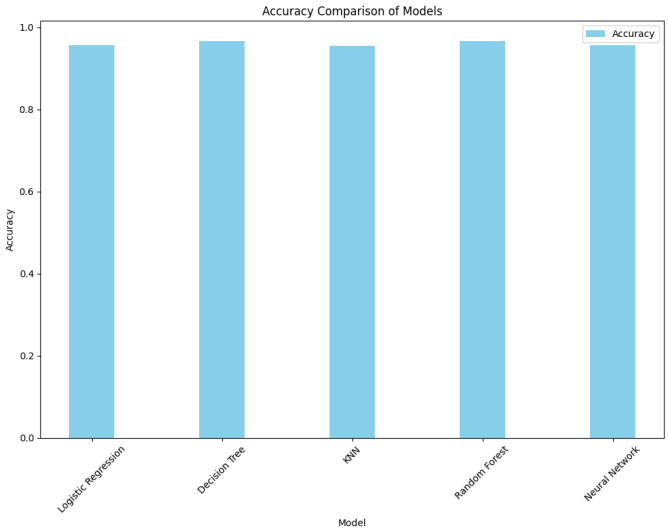

Fig. 5: Accuracy comparison of all models.

C. Precision, Recall, F-1 Comparison

The weighted average precision, recall, and F1-scores across all classes are shown in Fig. 6. All models demonstrate strong performance, with Decision Tree achieving the highest scores across all metrics

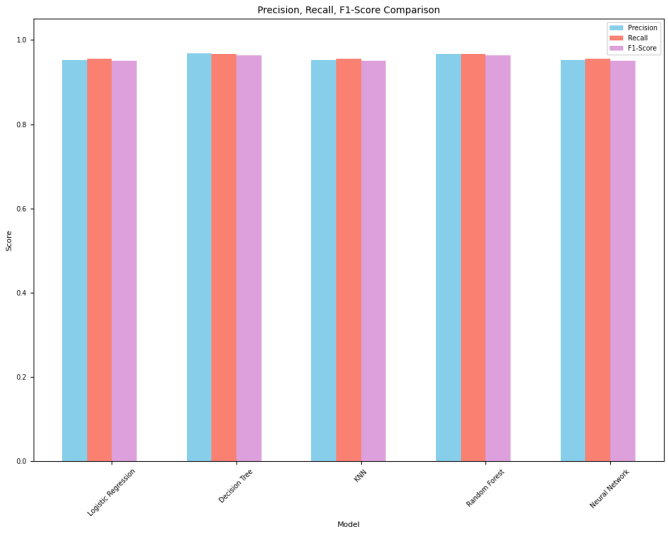

Fig. 6: Weighted average precision, recall, and F1-scores comparison across all classifier

D. AUC Score Comparison

The model performance is further evaluated using Area Under the Curve (AUC) scores, as shown in Fig. 7. The Decision Tree classifier achieves the highest AUC score of 0.9636, demonstrating excellent class separation capability.

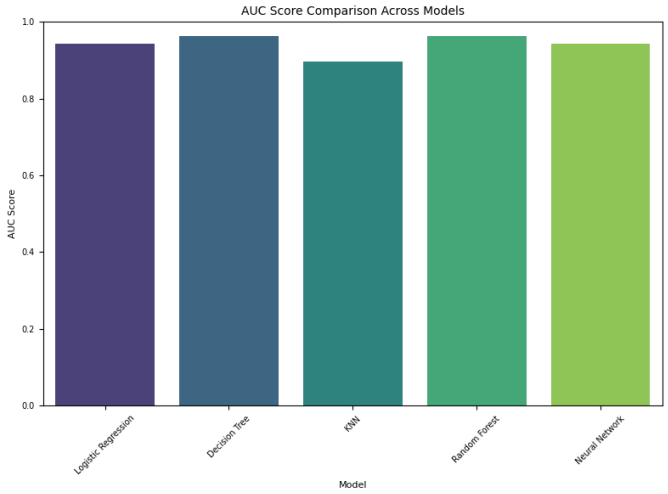

Fig. 7: AUC scores comparison across for class 1.

E. Confusion Matrix

The confusion matrices for all models are shown in Fig. 8. The Random Forest model has no false positives, while the KNN model shows a higher number of false negatives, reflecting its lower recall.

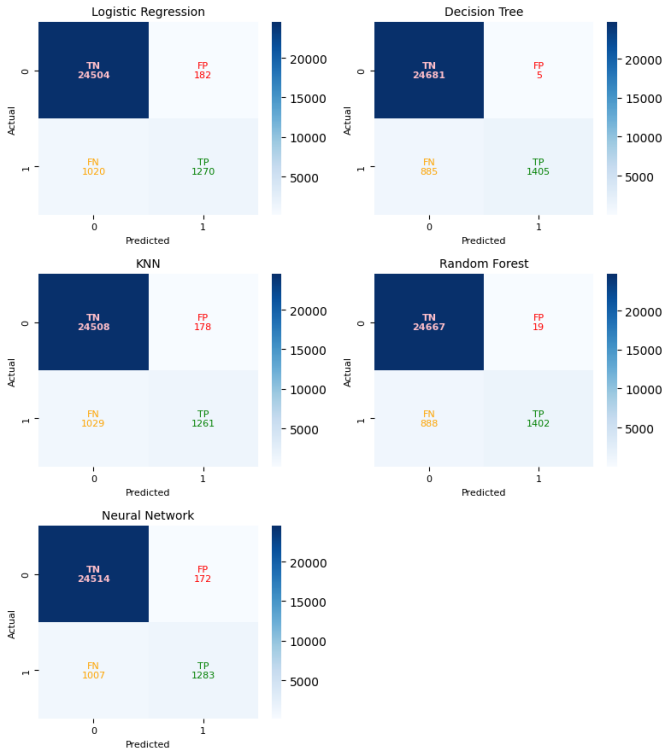

Fig. 8: Confusion matrices for all models.

F. AUC Score and ROC Curve

ROC curves are shown in Fig. 9. Decision Tree and Random Forest have the highest AUCs (0.9631 and 0.9614), while KNN has the lowest (0.9249).

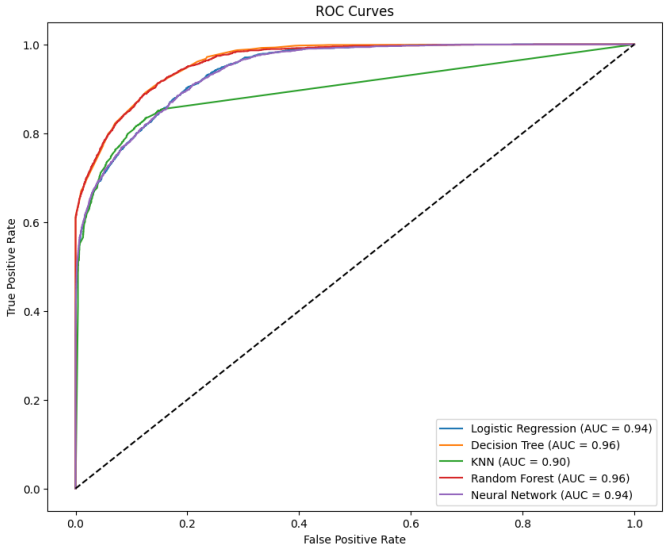

Fig. 9: ROC curves for all models.

VII. Conclusions

The weighted metrics demonstrate consistently strong performance across all models, with Decision Tree achieving the highest scores (Accuracy: 0.9670, F1: 0.96, AUC: 0.9636). Random Forest closely follows (Accuracy: 0.9664, F1: 0.96, AUC: 0.9622), while Logistic Regression and Neural Network show identical weighted F1-scores (0.95). KNN maintains competitive accuracy (0.9553) but lags in discriminative power (AUC: 0.8974). All models exhibit balanced precision-recall tradeoffs (0.95-0.97 weighted averages), though the 11:1 class imbalance suggests potential for further optimization of minority-class recall.

References

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