BRAC
UNIVERSITY

Inspiring Excellence

LAB PROJECT REPORT

CSE422: Artificial Intelligence

Submitted to:

Mcw, Ahf

Submitted By:

Azizul Kabir Jayed
22201665
Sanjid Hasan Sourov
22299528
Group (08)

Diabetes Prediction

Table of Contents

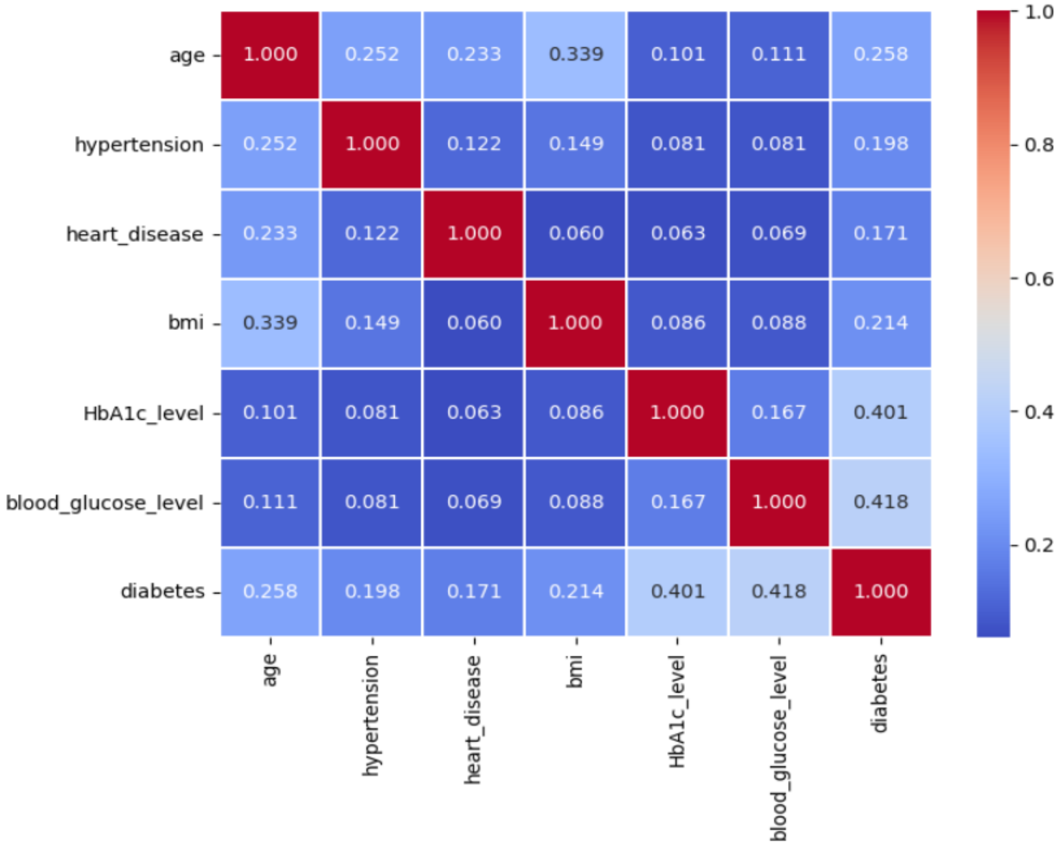
1. Introduction:

The goal of this project is to develop machine learning models capable of predicting the presence of diabetes based on demographic, health, and lifestyle attributes such as gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, and blood glucose level. Accurate prediction of diabetes can assist healthcare providers, policymakers, and individuals in early detection, prevention, and better management of the disease. Early identification of individuals at risk of diabetes using their health and lifestyle data. This allows for timely intervention and potentially reduces the long-term health impacts associated with diabetes.

2. Dataset Description

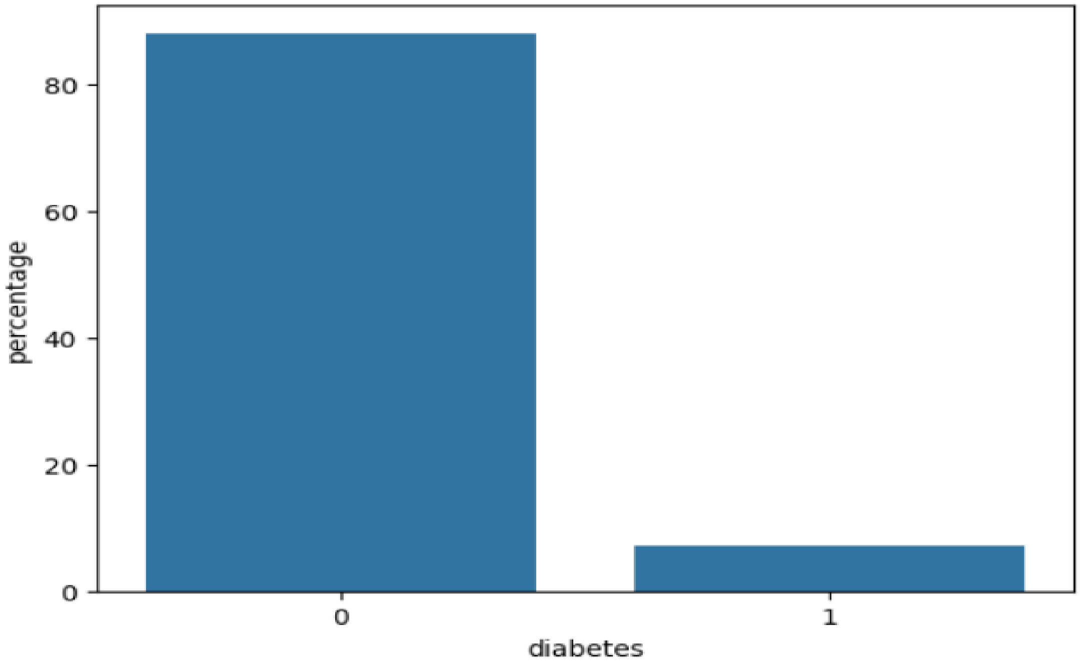
- **Features:** 9 (gender, age, hypertension, heart disease, smoking history, BMI, HbA1c_level, blood_glucose_level, diabetes).
- **Data points:** 100001 patients**.**
- **Problem Type:** Classification (target column = diabetes).
- **Feature Types:**
    - **Quantitative: age, bmi, HbA1c_level, blood_glucose_level.**
    - **Categorical: gender, smoking_history**
- **Encoding: Required for categorical features (gender, smoking_history).**
- **Correlation (via heatmap): Strong positive correlation between HbA1c_level, blood_glucose_level and diabetes. Other features show weaker correlation**
- **Insight: The dataset is small, with missing values and some imbalance, which may affect model performance.**

3. **Correlation Analysis: A heatmap generated by using the Seaborn library revealed relationships among numerical features.**



4. Imbalanced Dataset

- Output classes: **Diabetic (1)** and **Non-diabetic (0)**.
- The dataset is imbalanced (fewer diabetics than non-diabetics).
- Represented via bar chart (non-diabetic cases dominate).



5. Dataset Splitting

- **Train Set**: 70% of the data
- **Test Set**: 30% of the data

6. Model Training & Testing

Model Training and Evaluation Process: Each model was trained on the training dataset (70% of the total data) and evaluated on the testing dataset (rest 30% of the total data). The following steps were performed:

Logistic Regression

- **Training**:
  The Logistic Regression model was initialized with a maximum iteration limit of 100 to ensure convergence.
- **Testing**:
  Predictions were made using the sigmoid function to compute probabilities, and a threshold of 0.5 was used for binary classification.
- **Metrics**:
  - Accuracy: **95%**
  - Precision: 78**%**
  - Recall: **50%**

Random Forest

- **Training**:
  The Random Forest model was initialized with 100 estimators, and a random seed of 42. The model was trained using bootstrap aggregation to generate multiple decision trees on subsets of the training data.
- **Testing**:
  Predictions were made by aggregating the outputs of all decision trees and selecting the majority class.
- **Metrics**:
  - Accuracy: **96%**
  - Precision: **84%**
  - Recall: **63%**
  -

Neural network

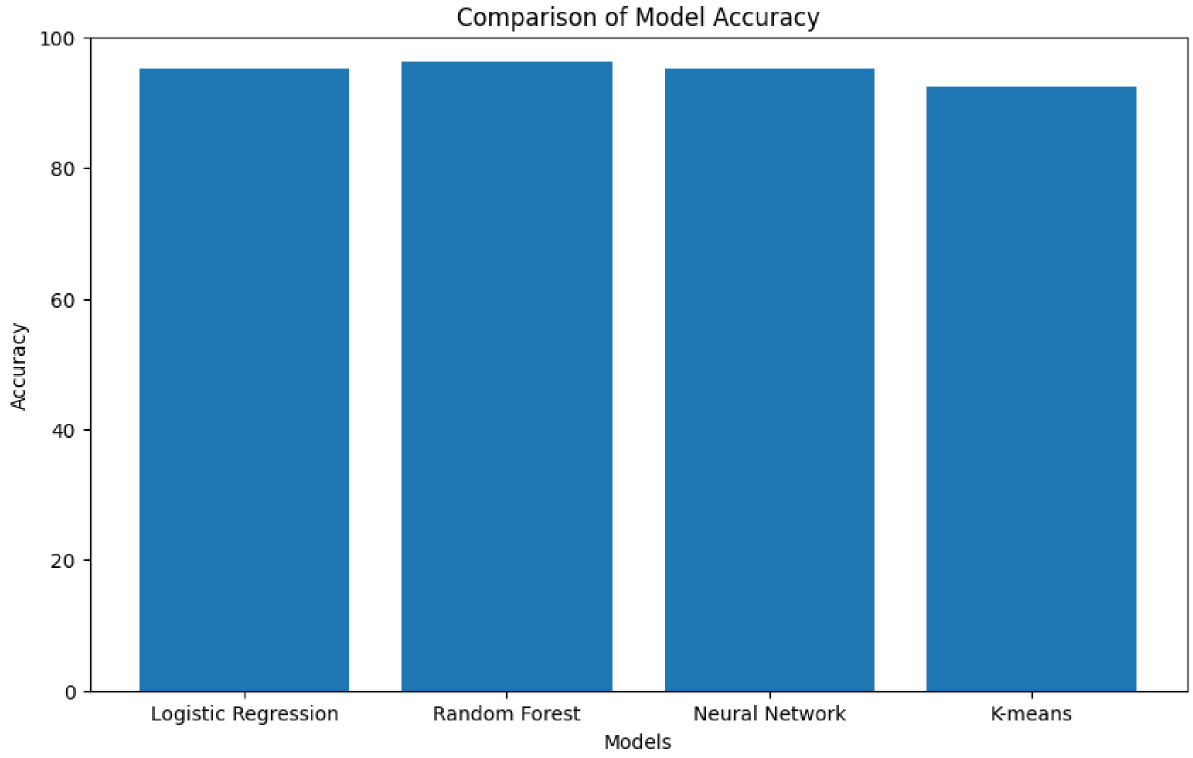
- **Training**:
  The neural network model was trained on the dataset with multiple dense layers using the ReLU activation function for hidden layers and a sigmoid activation for the output layer. The binary cross-entropy loss function and Adam optimizer were used to optimize the model over 10 epochs with a batch size of 8.
- **Testing**:
  Predictions were generated by passing input features through the trained network and applying a threshold of 0.5 to the output probabilities to determine the predicted class (diabetic or non-diabetic).
- **Metrics**:
  - Accuracy: **95%**
  - Precision: **85%**
  - Recall: **44%**

K-means Clustering

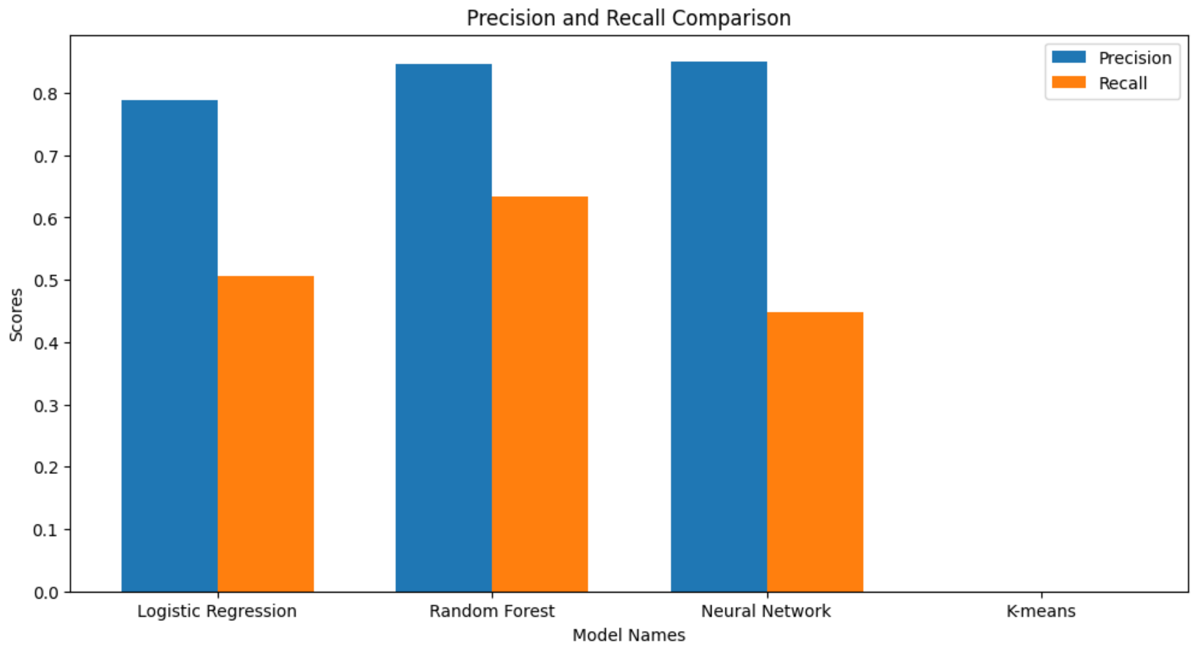
- **Training**:
  The K-Means clustering model was initialized with 2 clusters, corresponding to the two classes of the diabetes dataset. During the training phase, the model learned cluster centroids by fitting the feature vectors from the training set.
- **Testing**:
  For each test instance, the model assigned it to the nearest cluster based on Euclidean distance. To align clusters with actual class labels, a mapping from clusters to classes was applied, ensuring proper evaluation.
- **Metrics**:
  - Accuracy: **92%**
  - Precision: **0.00%**
  - Recall: **0.00%**

7. Comparison Analysis

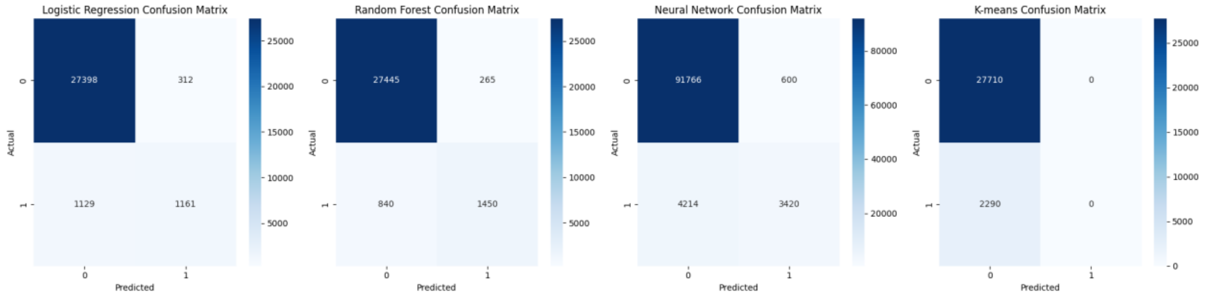
- **Accuracy Comparison**:
  - A bar chart was used to compare the accuracy of all models.



- **Precision and Recall**:
  - A grouped bar chart illustrated the precision and recall for all models, highlighting neural network and random forest as the top performer.



- **Confusion Matrix:**



8. Conclusion

This project successfully built and evaluated machine learning models to predict outcomes based on the dataset. Key steps included data preprocessing (removing empty strings, handling missing values, encoding categorical variables), stratified data splitting, and rigorous model evaluation. Logistic Regression, Random Forest, Neural network emerged as the best-performing model, while k-means clustering provided lower accuracy.

The workflow demonstrated the importance of preprocessing and model selection in achieving reliable predictions. Future improvements could include addressing class imbalances along with a dataset of much greater numbers. This project provides a solid foundation for predictive modeling and further enhancements.