Diabetes and Retinopathy Classification: A Multi-Modal Approach

Course: I S 675 Deep Learning for Business

College of Business, California State University, Long Beach

Instructor: Prof. Basil Latif

Submission Date: 12/05/2024

Team:

Musharraf Shaikh (031878923)

Deepak Deokar (032225204)

Ajay Tupe (032188167)

Table of Contents

1. Introduction

Diabetes and diabetic retinopathy are critical global health challenges, affecting millions worldwide. Diabetes, a chronic disease characterized by elevated blood glucose levels, often leads to complications such as cardiovascular disease, kidney failure, and diabetic retinopathy—a condition causing damage to the retina, potentially resulting in blindness. Early detection and accurate classification of diabetes and its complications are essential for timely medical intervention, reducing healthcare costs, and improving patient outcomes.

In recent years, advancements in artificial intelligence and machine learning have paved the way for innovative solutions in healthcare. Specifically, deep learning, with its ability to analyze complex data patterns, has shown promise in diagnosing and predicting diseases from multi-modal datasets. Multi-modal approaches leverage both structured tabular data (e.g., health metrics like glucose levels, BMI, and age) and unstructured image data (e.g., retinal images) to improve diagnostic accuracy.

This project aims to utilize cutting-edge deep learning techniques to address two key objectives:

1. **Diabetes Status Prediction**: Classify individuals as Non-Diabetic, Pre-Diabetic, or Diabetic using tabular health data.
2. **Retinopathy Severity Prediction**: Determine the severity of diabetic retinopathy from retinal images.

The methodologies employed include state-of-the-art convolutional neural networks (CNNs), specifically ResNet-50, for image classification, and feedforward neural networks for tabular data analysis. By combining these approaches, the project seeks to provide actionable insights for healthcare professionals and contribute to improved patient care. Additionally, this study adheres to the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, ensuring a systematic approach to problem-solving, from data understanding and preparation to modeling and evaluation.

2. Problem Statement

Diabetes is a chronic disease with significant health and economic impacts, affecting over 400 million people globally. One of its severe complications, diabetic retinopathy, is the leading cause of blindness among working-age adults. Despite advances in medical technology, late diagnosis and delayed intervention remain major challenges. Early detection of diabetes and its associated complications can lead to better disease management and improved outcomes for patients. However, traditional diagnostic methods often rely on costly, time-consuming, and invasive procedures, creating a need for innovative, scalable solutions.

The primary aim of this project is to develop predictive models capable of addressing the following objectives:

1. **Diabetes Status Prediction**:
    - Predict the health status of individuals (Non-Diabetic, Pre-Diabetic, or Diabetic) using structured health data, including metrics such as glucose levels, BMI, and age.
    - Identify significant features influencing the predictions, providing healthcare professionals with insights into critical risk factors.
2. **Retinopathy Severity Prediction**:
    - Accurately classify retinal images based on the severity of diabetic retinopathy, assisting in prioritizing patients for timely medical intervention.
    - Explore advanced deep learning architectures like ResNet-50 for handling the complexity of image data.

This project addresses two significant gaps in healthcare:

- **Efficiency**: Automation of predictions reduces the reliance on manual, resource-intensive diagnostic procedures.
- **Accuracy**: Multi-modal deep learning approaches promise higher predictive accuracy by integrating structured tabular data with unstructured image data.

3. Datasets Selection, Overview and Summary

The datasets for this project were sourced from Kaggle, a popular platform for data science and machine learning resources. Two datasets were selected to address the multi-modal nature of the problem:

1. **Tabular Dataset**:
    - Contains 253,680 rows of health indicators, with each row representing a respondent's health metrics and lifestyle data.

    1. **HighBP**: Indicates whether the individual has high blood pressure (binary: 0 = No, 1 = Yes).
    2. **HighChol**: Indicates whether the individual has high cholesterol (binary: 0 = No, 1 = Yes).
    3. **CholCheck**: Whether the individual has had their cholesterol checked (binary: 0 = No, 1 = Yes).
    4. **BMI**: Body Mass Index, a measure of body fat based on height and weight.
    5. **Smoker**: Indicates smoking status (binary: 0 = No, 1 = Yes).
    6. **Stroke**: History of stroke (binary: 0 = No, 1 = Yes).
    7. **HeartDisease**: Presence of coronary heart disease or myocardial infarction (binary: 0 = No, 1 = Yes).
    8. **PhysActivity**: Indicates physical activity status in the past 30 days (binary: 0 = No, 1 = Yes).
    9. **Fruits**: Frequency of fruit consumption (binary: 0 = No, 1 = Yes).
    10. **Veggies**: Frequency of vegetable consumption (binary: 0 = No, 1 = Yes).
    11. **HvyAlcoho**: Heavy alcohol consumption (binary: 0 = No, 1 = Yes).
    12. **AnyHealth**: Whether the individual perceives their health as poor (binary: 0 = No, 1 = Yes).
    13. **NoDocbcCost**: Whether cost prevented a doctor visit in the past year (binary: 0 = No, 1 = Yes).
    14. **GenHlth**: General health rating (1 = Excellent to 5 = Poor).
    15. **MentHlth**: Number of days mental health was not good in the past 30 days.

16. **PhysHlth**: Number of days physical health was not good in the past 30 days.
17. **DiffWalk**: Difficulty walking or climbing stairs (binary: 0 = No, 1 = Yes).
18. **Sex**: Gender of the individual (binary: 0 = Female, 1 = Male).
19. **Age**: Age bracket of the individual (1 = 18–24, 2 = 25–29, ..., 13 = 80 or older).
20. **Education**: Level of education attained (1 = Never attended school to 6 = College graduate).
21. **Income**: Income bracket of the individual (1 = Less than $10,000 to 8 = $75,000 or more).

- **Target Variable:** Diabetes_012 (categorical: 0 = Non-Diabetic, 1 = Pre-Diabetic, 2 = Diabetic).

2. **Image Dataset**:
    - Comprises retinal images divided into training and testing sets.
    - Images are labeled based on the severity of diabetic retinopathy.
    - Target Variable: Retinopathy severity levels (e.g., 0 = No Retinopathy, 1–4 = Increasing Severity).

Dataset Overview:

Tabular Dataset

The tabular dataset provides a comprehensive view of health and lifestyle indicators relevant to diabetes prediction. Key details:

- **Features**: Includes 21 features, such as:
    - **Health Indicators**: Blood Pressure (HighBP), Cholesterol (HighChol), BMI, and General Health (GenHlth).
    - **Lifestyle Factors**: Physical Activity (PhysActivi), Smoking (Smoker), and Alcohol Use (HvyAlcoho).
    - **Demographics**: Age, Education, and Income.
- **Target Variable**: Diabetes_012 with three classes for diabetes status.
- **Data Size**: 253,680 rows, offering a robust sample for training machine learning models.

Image Dataset

The image dataset focuses on detecting diabetic retinopathy severity. Key details:

- **Training Data**: Contains a significant number of labeled retinal images for model training.
- **Testing Data**: Separate set of labeled images for model evaluation.
- **Image Labels**: Represent retinopathy severity (0 to 4).
- **Data Size**: Approximately 10GB.

Dataset Summary

Tabular Dataset Insights:

1. The dataset is well-balanced, with no missing values.
2. Features provide comprehensive coverage of potential diabetes risk factors.
3. Class distribution of the target variable (Diabetes_012) may need evaluation to address any imbalance.

Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	1	1	1	40	1	0	0	0	0	1	0	1	0	5	18	15	1	0	9	4	3
0	0	0	0	25	1	0	0	1	0	0	0	0	1	3	0	0	0	0	7	6	1
0	1	1	1	28	0	0	0	0	1	0	0	1	1	5	30	30	1	0	9	4	8
0	1	0	1	27	0	0	0	1	1	1	0	1	0	2	0	0	0	0	11	3	6
0	1	1	1	24	0	0	0	1	1	1	0	1	0	2	3	0	0	0	11	5	4
0	1	1	1	25	1	0	0	1	1	1	0	1	0	2	0	2	0	1	10	6	8
0	1	0	1	30	1	0	0	0	0	0	0	1	0	3	0	14	0	0	9	6	7
0	1	1	1	25	1	0	0	1	0	1	0	1	0	3	0	0	1	0	11	4	4
2	1	1	1	30	1	0	1	0	1	1	0	1	0	5	30	30	1	0	9	5	1
0	0	0	1	24	0	0	0	0	0	1	0	1	0	2	0	0	0	1	8	4	3
2	0	0	1	25	1	0	0	1	1	1	0	1	0	3	0	0	0	1	13	6	8
0	1	1	1	34	1	0	0	0	1	1	0	1	0	3	0	30	1	0	10	5	1
0	0	0	1	26	1	0	0	0	0	1	0	1	0	3	0	15	0	0	7	5	7
2	1	1	1	28	0	0	0	0	0	1	0	1	0	4	0	0	1	0	11	4	6
0	0	1	1	33	1	1	0	1	0	1	0	1	1	4	30	28	0	0	4	6	2
0	1	0	1	33	0	0	0	1	0	0	0	1	0	2	5	0	0	0	6	6	8
0	1	1	1	21	0	0	0	1	1	1	0	1	0	3	0	0	0	0	10	4	3
2	0	0	1	23	1	0	0	1	0	0	0	1	0	2	0	0	0	1	7	5	6
0	0	0	0	23	0	0	0	0	0	1	0	1	0	2	15	0	0	0	2	6	7

Fig. Sample of Tabular Dataset

Image Dataset Insights:

1. High-quality retinal images allow for robust feature extraction using convolutional neural networks.
2. Class-wise distribution of severity levels will be analyzed to handle imbalances, if present.
3. Augmentation techniques such as rotation, scaling, and flipping will be applied to improve generalization.

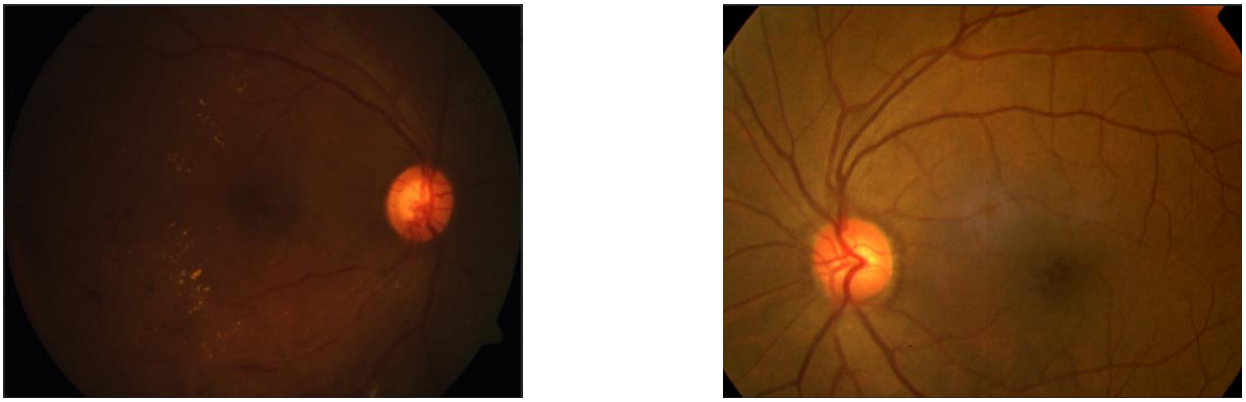


Fig. Sample of Train and Test Image Dataset

4. Methodology

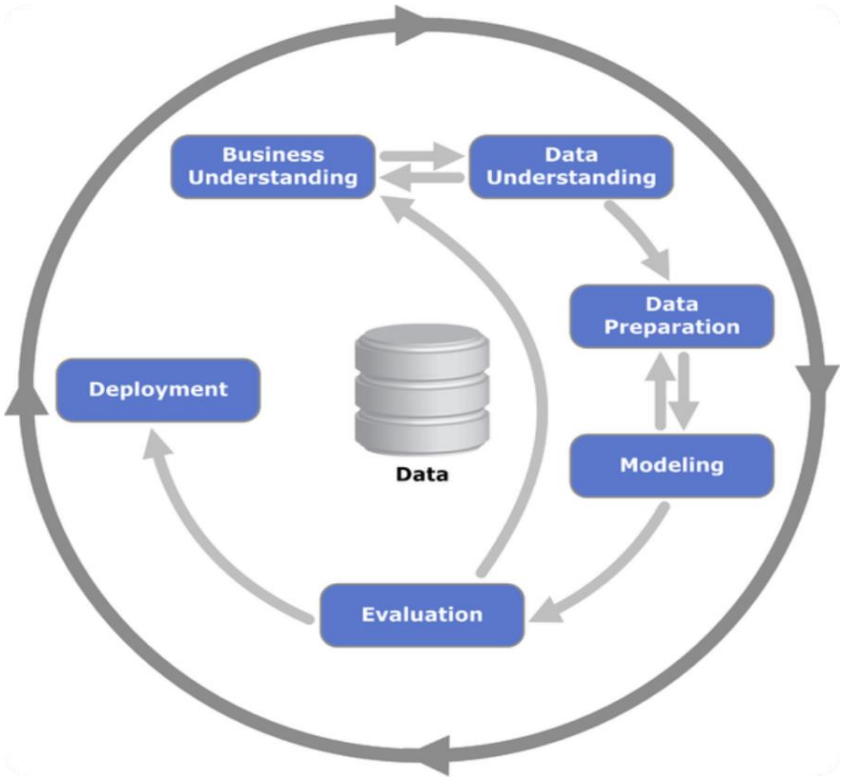


Fig. CRISP-DM Process Overview

CRISP-DM Framework:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

1. Business Understanding

This phase focuses on clearly defining the problem and identifying business goals to ensure alignment with project outcomes. In this project, the aim is to predict diabetes status and retinopathy severity to enable early detection and effective healthcare intervention. Key objectives

include building models that classify diabetes and retinopathy accurately, while identifying critical health indicators that assist in medical decision-making.

2. Data Understanding

This phase involves collecting and exploring datasets to understand their structure, quality, and relevance. For this project, the tabular dataset was analyzed to explore health indicators like BMI, cholesterol, and physical activity, while the image dataset was reviewed to assess the class distribution of retinal images. Exploratory data analysis revealed potential imbalances in the classes, correlations between features, and patterns critical for model training.

3. Data Preparation

This phase includes cleaning, transforming, and formatting data for optimal model performance. In the tabular dataset, missing values were addressed, numerical features were scaled, and categorical variables (e.g., education and income) were encoded. For the image dataset, preprocessing steps included resizing and normalizing the images, applying data augmentation techniques (e.g., rotation and flipping), and splitting the dataset into training, validation, and testing sets. These steps ensure the data is ready for effective modeling.

4. Modeling

This phase involves selecting and training machine learning models to address the problem. For tabular data, models like feedforward neural networks and traditional algorithms (e.g., Decision Trees, Random Forests) were used. For image data, a pretrained ResNet-50 model was fine-tuned to classify retinal images based on retinopathy severity. Hyperparameter optimization, such as adjusting the learning rate and batch size, was performed to enhance model performance. Comparisons between deep learning and traditional models highlighted the strengths of CNNs for image classification.

5. Evaluation

This phase assesses the model's performance to ensure it meets the business objectives. Metrics like accuracy, precision, recall, F1-score, and ROC-AUC were used to evaluate both

tabular and image models. Confusion matrices and class-wise accuracy helped identify areas where the models performed well and where improvements were needed. The evaluation revealed that the ResNet-50 model achieved higher accuracy for retinopathy severity classification, while the tabular model provided insights into key health indicators influencing diabetes predictions.

6. Deployment

This phase integrates the trained models into real-world applications for making predictions and supporting decision-making processes. For this project, the models are saved enabling healthcare practitioners to input patient data or retinal images and receive predictions in real time. Comprehensive documentation ensures the models can be reproduced, monitored, and updated as needed, making them suitable for long-term clinical use.

5. Data Understanding and Preprocessing

Data Understanding

This phase focuses on exploring and analyzing the datasets to gain insights and assess their suitability for modeling.

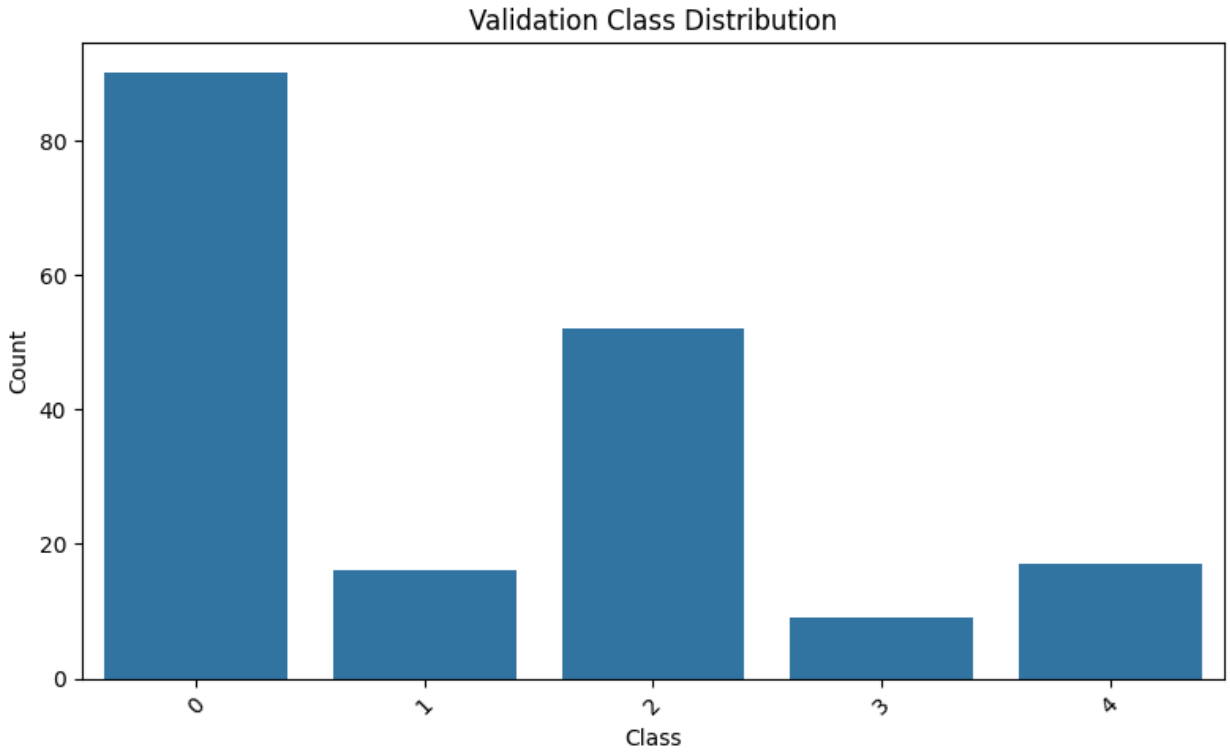
1. **Tabular Dataset**:
    - **Overview**: The tabular dataset consists of 253,680 rows and 22 features, including both health metrics (e.g., BMI, cholesterol levels) and lifestyle indicators (e.g., smoking, physical activity). The target variable, Diabetes_012, categorizes individuals into three classes: Non-Diabetic (0), Pre-Diabetic (1), and Diabetic (2).
    - **Initial Observations**:
        - The dataset is free of missing values and duplicates.
        - The features exhibit varying ranges (e.g., BMI vs. binary variables like HighBP), requiring normalization or scaling.
        - Class distribution analysis revealed a potential imbalance in the target variable, which could affect model training.
2. **Image Dataset**:
    - **Overview**: The image dataset contains retinal images split into training and testing sets. Each image is labeled with a severity score indicating the level of diabetic retinopathy (0 = No Retinopathy to 4 = Severe Retinopathy).
    - **Initial Observations**:
        - The images vary in resolution and quality, necessitating standard resizing and normalization.
        - Class imbalance in the severity levels is evident, with fewer examples of higher severity classes.

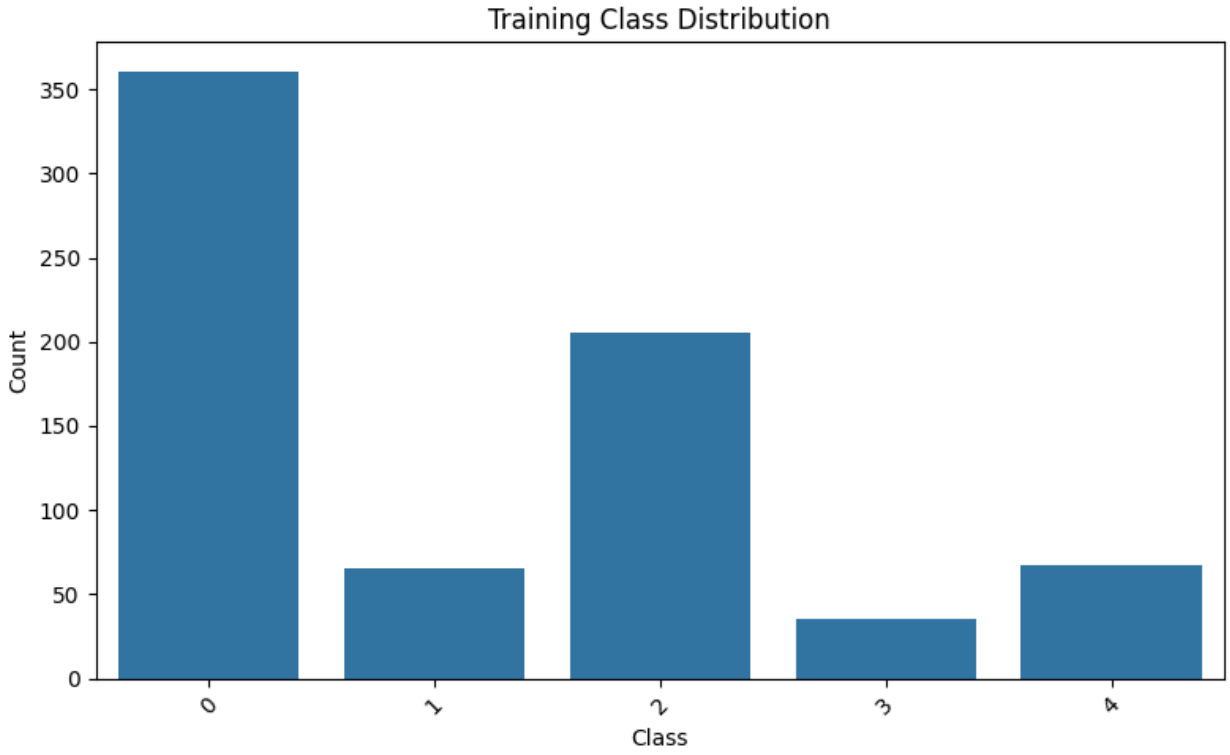
- Visual inspection of sample images confirmed the presence of distinguishable patterns for classification.

Data Preprocessing

This phase involves preparing the data for model training by addressing quality issues and transforming it into a usable format.

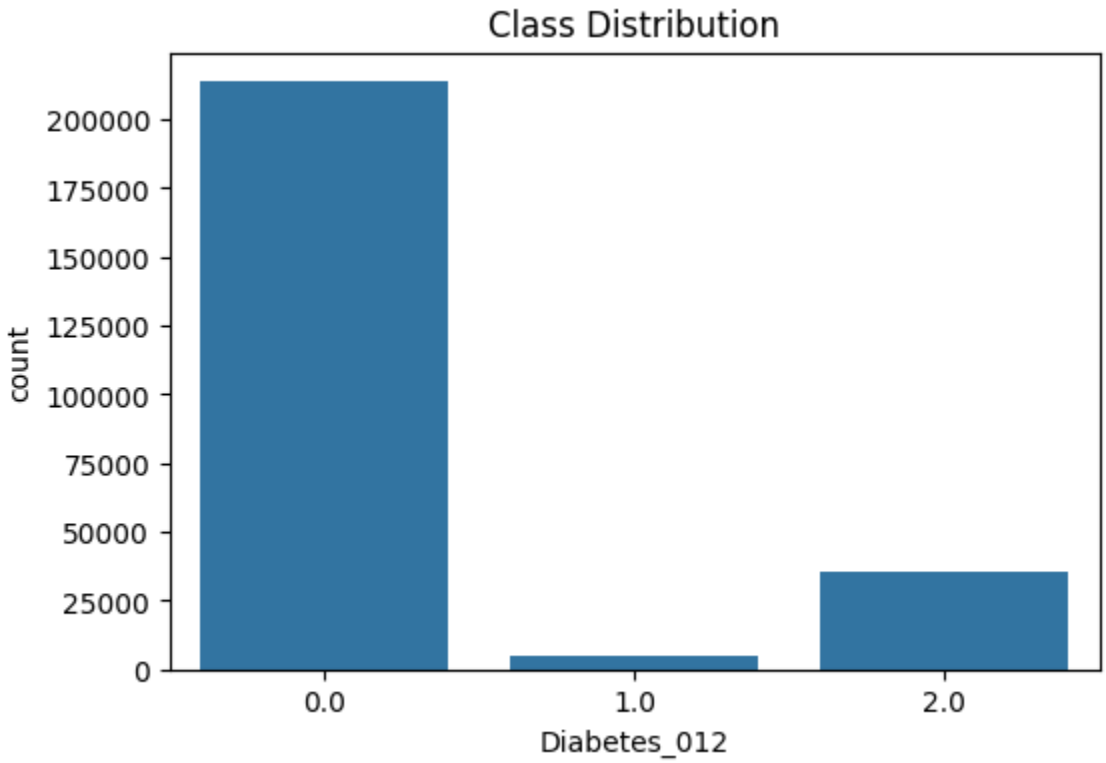
1. **Tabular Dataset Preprocessing**:
    - **Feature Scaling**: Standard Scaling was applied to numerical features to ensure all features were on a similar scale.
    - **Feature Exclusion**: Categorical features, like Education and Income, were encoded to represent them numerically.
    - **Class Balancing**: SMOTE (Synthetic Minority Over-sampling Technique) was used to handle the class imbalance in the target variable, which consists of three classes: Non-diabetic (0), Pre-diabetic (1), and Diabetic (2).
    - **Data Splitting**: The dataset was divided into training (80%) and validation (20%) subsets, ensuring stratification to maintain the target variable's class proportions.
2. **Image Dataset Preprocessing**:
    - **Image Resizing and Normalization**: All images were resized to a fixed dimension (e.g., 224x224) suitable for the ResNet-50 model and normalized to a pixel range of [0, 1].
    - **Data Augmentation**: Techniques like rotation, flipping, and zooming were applied to artificially increase the training dataset size and improve model generalization.
    - **Train-Test Split**: Images were organized into separate training and testing directories, maintaining a consistent distribution of classes in both subsets.


Validation Class Distribution
Count
Class
0
20
40
60
80
0
1
2
3
4


Training Class Distribution
Count
Class
0
50
100
150
200
250
300
350
0
1
2
3
4

6. Data Visualization

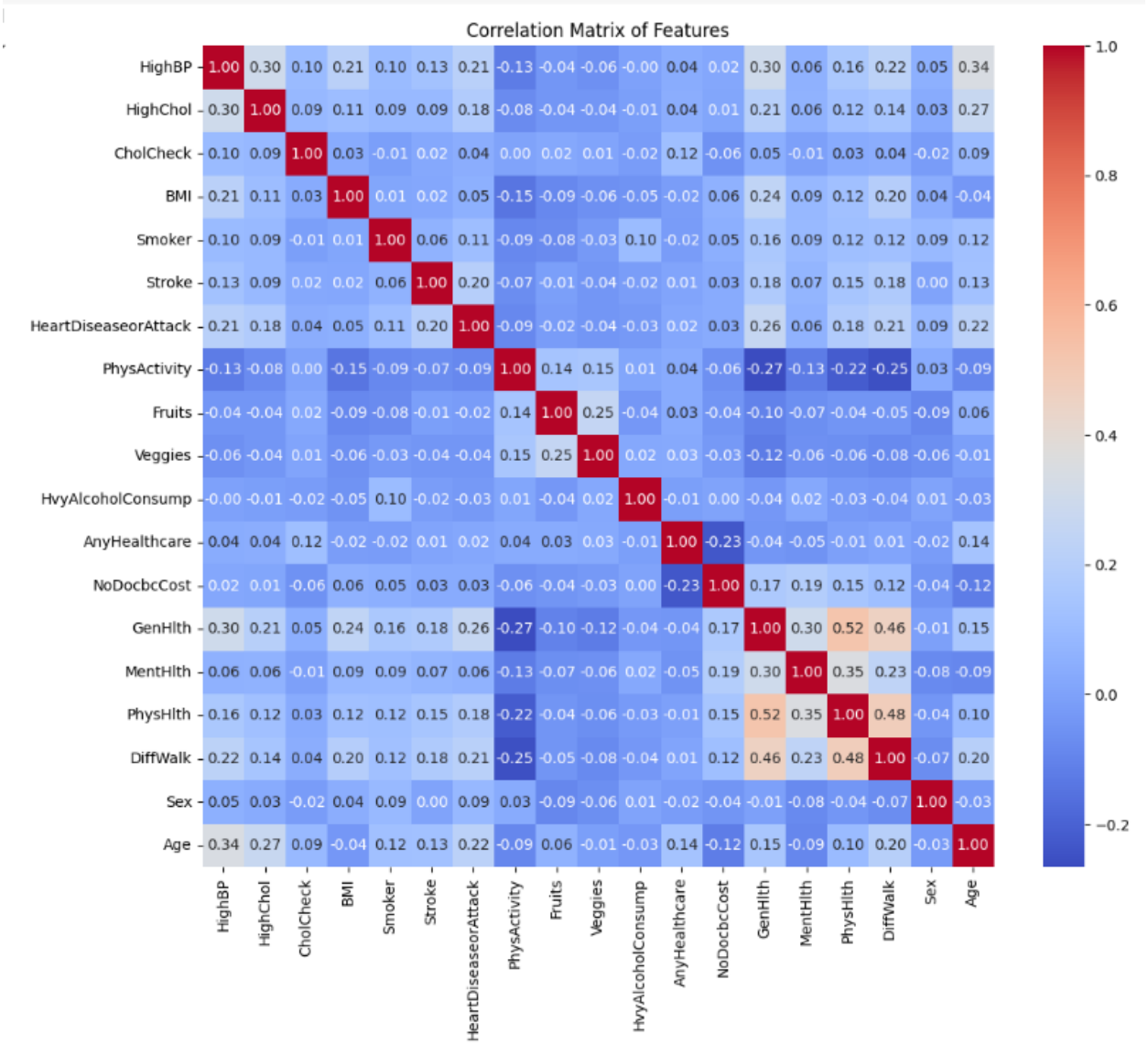
Tabular Dataset:



Class Distribution

- **Description**: This bar plot visualizes the distribution of the target variable Diabetes_012, which represents the diabetes status of individuals. The classes are:
    - 0: Non-Diabetic
    - 1: Pre-Diabetic
    - 2: Diabetic
- **Insights**:
    - The dataset is highly imbalanced, with a majority of the samples belonging to the Non-Diabetic class.

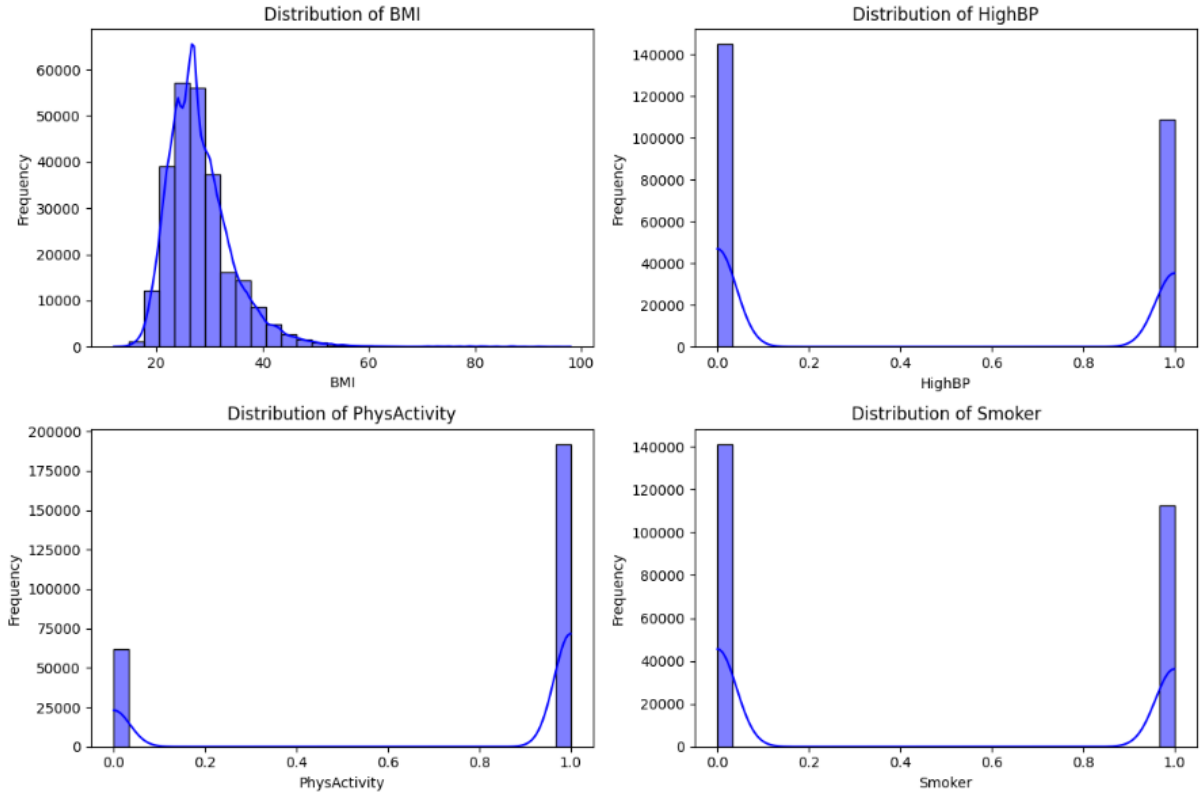
- This imbalance may lead to bias in model predictions, requiring techniques like oversampling, under sampling, or class weighting during training.



Correlation Matrix

- **Description**: This heatmap displays the correlation coefficients between numerical features in the dataset.
- **Insights**:
    - Features like HighBP (High Blood Pressure) and BMI show moderate positive correlations with diabetes status.

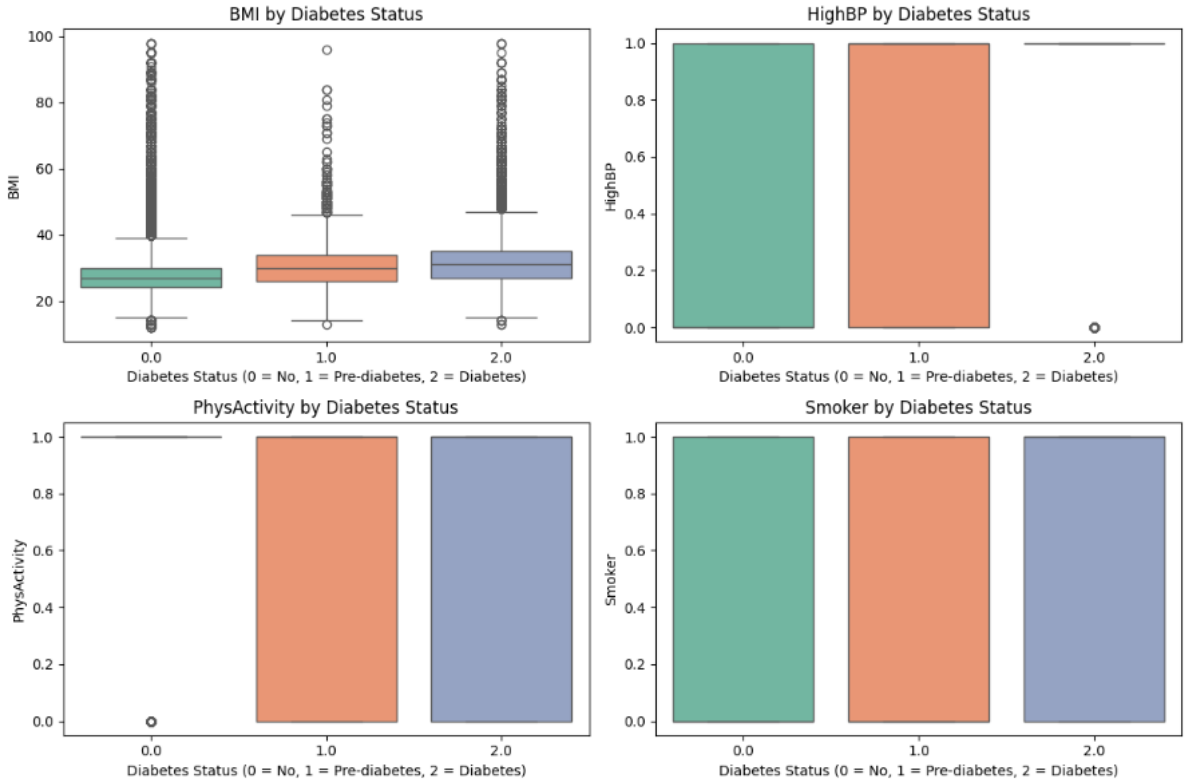
- Strong correlations between some features (e.g., PhysHlth and MentHlth) suggest potential redundancy, which could impact feature selection.
- The matrix helps identify independent features that contribute most to the prediction.



Distribution of Key Features

- **Graphs**: The histograms display the distributions of critical features, including BMI, HighBP, PhysActivity, and Smoker.
- **Insights**:
  - **BMI**: The distribution is positively skewed, with a concentration of individuals having a BMI between 20 and 40. Outliers at higher BMI levels are observed.

- **HighBP**: Binary feature showing a near-even split between individuals with and without high blood pressure.
- **PhysActivity**: Most individuals reported engaging in physical activity, suggesting its significance in health metrics.
- **Smoker**: Many individuals are non-smokers, but the presence of smokers provides a key lifestyle factor for analysis.



Feature Distributions by Diabetes Status

- **Graphs**: Boxplots and bar plots compare feature distributions across diabetes classes.
- **Insights**:
  - **BMI**: Diabetic individuals tend to have higher BMI values on average, emphasizing its role in diabetes risk.

- **HighBP**: A larger proportion of diabetic individuals report having high blood pressure compared to non-diabetic individuals.
- **PhysActivity**: Lower physical activity is associated with diabetic individuals, highlighting its preventive role.
- **Smoker**: Smoking status does not show significant variation across diabetes classes, suggesting it may have a lesser direct impact.

Image Dataset:

Diagnosis 0 (No Retinopathy):

- Healthy retinal image with no visible signs of diabetic retinopathy.
- Clear blood vessels and absence of abnormalities.

Diagnosis 1 (Mild Retinopathy):

- Early-stage retinopathy with minimal signs, such as microaneurysms or small hemorrhages.
- Blood vessels are slightly altered but mostly intact.

Diagnosis 2 (Moderate Retinopathy):

- More noticeable changes, including increased microaneurysms and potential exudates.
- Visible signs of retinal damage, indicating disease progression.

Diagnosis 3 (Severe Retinopathy):

- Significant retinal damage with extensive hemorrhages and exudates.
- Blood vessels show severe abnormalities, requiring immediate attention.

Diagnosis 4 (Proliferative Retinopathy):

- Advanced stage with the presence of neovascularization and severe retinal damage.
- Potential for vision loss if not treated urgently.

7. Modeling

We implemented and compared the following models to predict diabetes status:

a. Logistic Regression:

Logistic Regression was used as a simple baseline model for multi-class classification. Despite being a linear model, it performed reasonably well.

Accuracy: We observed that the accuracy of logistic regression was decent but not the highest among the models.

```
Logistic Regression Accuracy: 0.5242938029355337
Classification Report:
              precision    recall  f1-score   support

         0.0       0.60      0.66      0.63     42688
         1.0       0.43      0.32      0.37     42676
         2.0       0.51      0.59      0.55     42858

    accuracy                           0.52    128222
   macro avg       0.51      0.52      0.52    128222
weighted avg       0.51      0.52      0.52    128222
```

b. Decision Tree:

Decision Tree is another simple classifier that splits the data based on feature thresholds. It performed better than logistic regression but still struggled to capture complex relationships.

Accuracy: While better than Logistic Regression, it had high variance, meaning it could overfit the data easily.

```
Decision Tree Accuracy: 0.8459390744178066
Classification Report:
              precision    recall  f1-score   support

         0.0       0.86      0.86      0.86     42688
         1.0       0.88      0.90      0.89     42676
         2.0       0.80      0.78      0.79     42858

    accuracy                           0.85    128222
   macro avg       0.85      0.85      0.85    128222
weighted avg       0.85      0.85      0.85    128222
```

c. Random Forest:

Random Forest, an ensemble method, significantly outperformed other models. It used multiple decision trees to reduce overfitting and gave the most robust performance across all metrics.

Accuracy: Random Forest had the highest accuracy compared to all models. It handled the class imbalance better than individual decision trees.

Key Insight: Random Forest was able to better capture the complex relationships between features, especially those like BMI and HighBP, that are indicative of diabetes.

```
Random Forest Accuracy: 0.9118092059085025
Classification Report:
              precision    recall  f1-score   support

         0.0       0.89      0.90      0.89     42688
         1.0       0.96      0.96      0.96     42676
         2.0       0.88      0.87      0.88     42858

    accuracy                           0.91    128222
   macro avg       0.91      0.91      0.91    128222
weighted avg       0.91      0.91      0.91    128222
```

d. Gradient Boosting:

Gradient Boosting also performed well and showed that boosting methods can handle complex data with multiple weak learners.

Accuracy: Like Random Forest, Gradient Boosting performed quite well, though it was slightly less accurate than Random Forest.

Key Insight: Boosting methods like Gradient Boosting are effective when working with imbalanced datasets, though Random Forest was the top performer.

```
Gradient Boosting Accuracy: 0.7185974325778728
Classification Report:
              precision    recall  f1-score   support

         0.0       0.80      0.85      0.83     42688
         1.0       0.73      0.69      0.71     42676
         2.0       0.62      0.62      0.62     42858

    accuracy                           0.72    128222
   macro avg       0.72      0.72      0.72    128222
weighted avg       0.72      0.72      0.72    128222
```

e. Neural Network:

We implemented a Complex Feedforward Neural Network (FNN) with multiple layers and dropout for regularization. This model struggled with accuracy due to the tabular data's structure, which is less suited to CNN-style models.

Accuracy: The neural network had a relatively low accuracy (57%) compared to traditional machine learning models, confirming that neural networks may not always outperform simpler methods for tabular data.

```
Neural Network Accuracy: 0.6843287423375084
Classification Report:
              precision    recall  f1-score   support

         0.0       0.73      0.62      0.67     42688
         1.0       0.64      0.81      0.72     42676
         2.0       0.70      0.62      0.66     42858

    accuracy                           0.68    128222
   macro avg       0.69      0.68      0.68    128222
weighted avg       0.69      0.68      0.68    128222
```

7. Hyperparameter Tuning

Hyperparameter tuning involves optimizing parameters set before training to improve model performance. Here's a brief overview:

Key Hyperparameters:

1. **Learning Rate**: Controls step size during optimization. Start with 0.01, 0.001, or 0.0001.
2. **Batch Size**: Number of samples per training step. Common sizes: 32, 64, 128.
3. **Number of Epochs**: Determines how many full datasets passes the model makes. Use early stopping to avoid overfitting.
4. **Optimizer**: Algorithms like Adam (adaptive learning rates) or SGD (momentum-based).
5. **Dropout Rate**: Regularization to prevent overfitting (e.g., 0.1, 0.3, 0.5).
6. **Hidden Layers/Neurons**: Adjust architecture complexity for better performance.

Tuning Techniques:

1. **Grid Search**: Exhaustive search over fixed parameter sets.
2. **Random Search**: Randomly sample parameters for efficiency.
3. **Manual Tuning**: Adjust parameters based on insights.
4. **Learning Rate Schedulers**: Dynamically reduce learning rate during training.

8. Results and Performance Metrics

We evaluated each model using the following metrics:

Accuracy: To measure the overall classification performance.

Classification Report: To assess precision, recall, and F1-score for each class (Non-diabetic, Pre-diabetic, Diabetic). This was critical for understanding how well each model differentiated between classes.

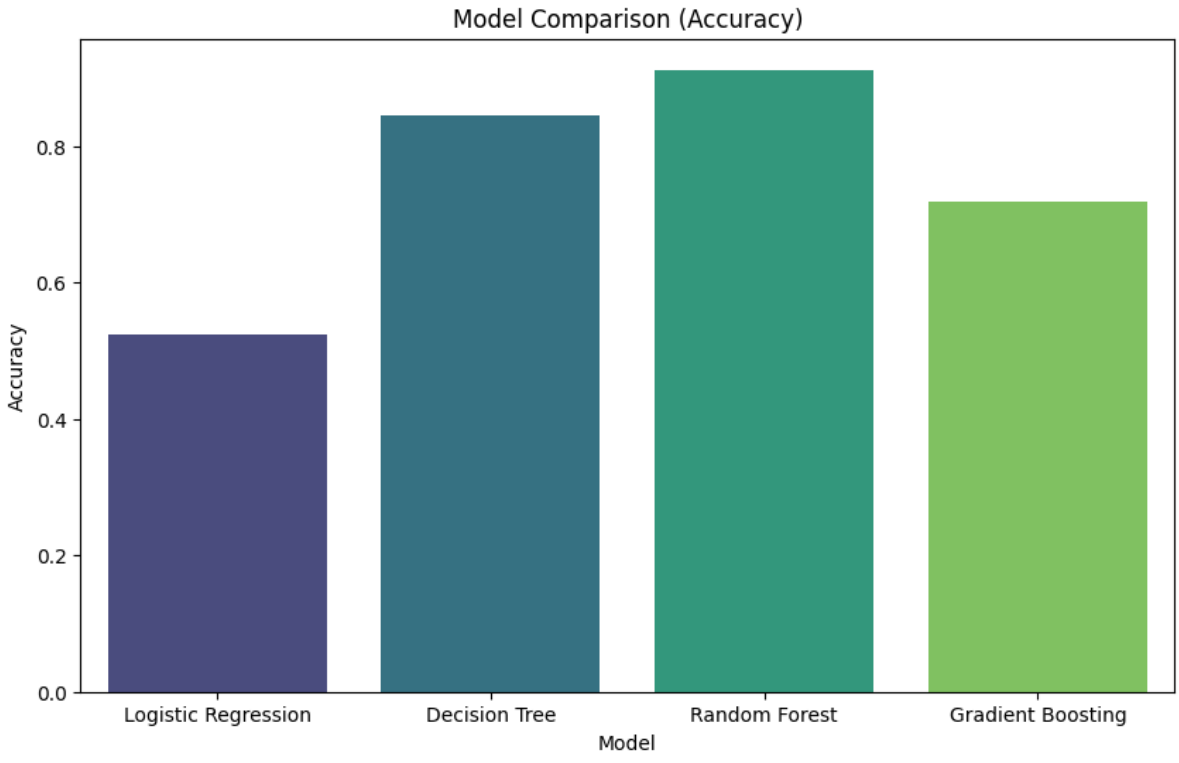
Logistic Regression: 0.52 or 52%

Decision Tree: 0.85 or 85%

Random Forest: 0.91 or 91%

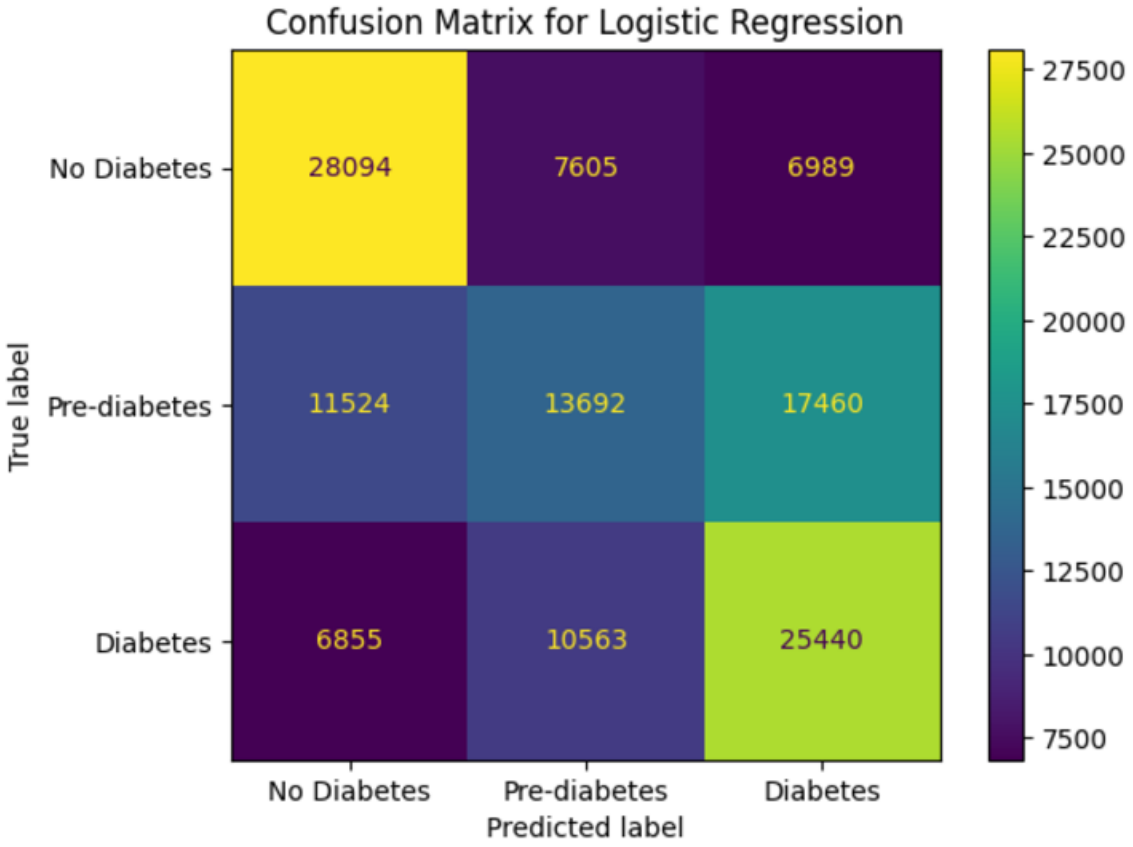
Gradient Boosting: 0.72 or 72%

Neural Network: 0.67 or 67%



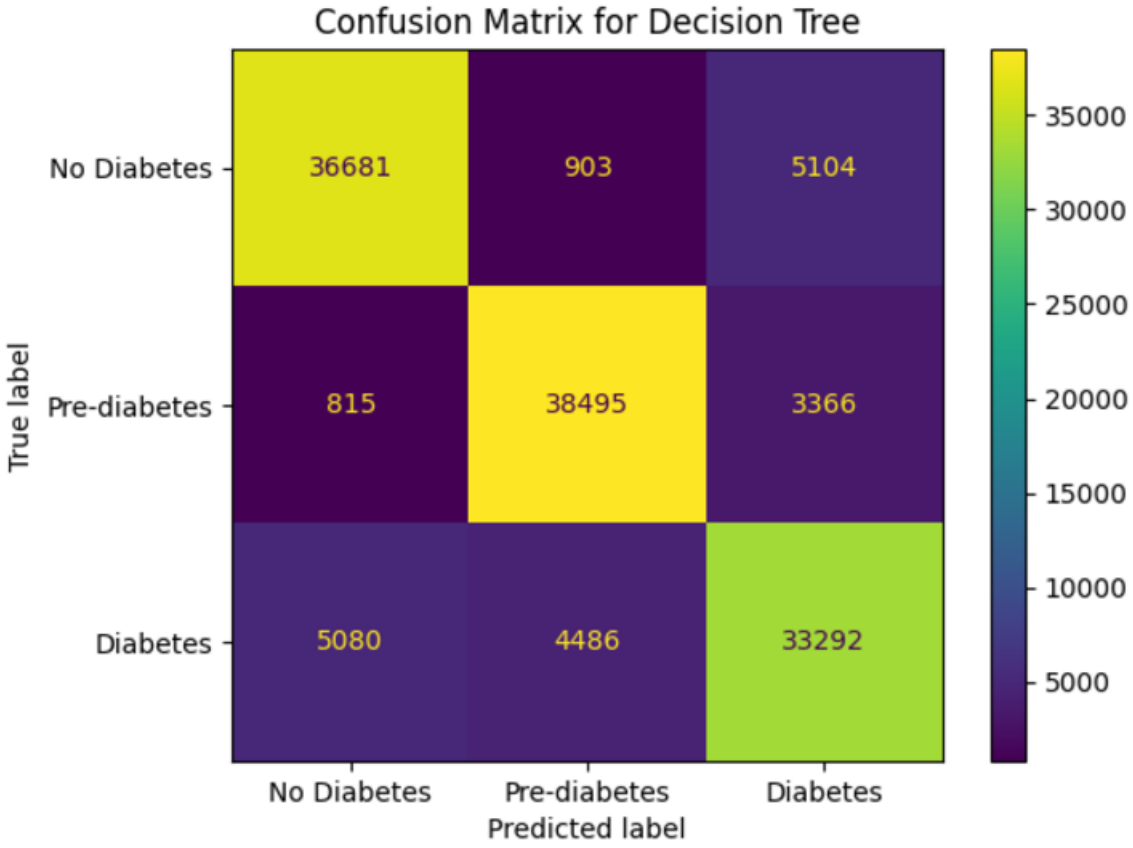
Confusion Matrix: We visualized the confusion matrix for each model to show how many instances of each class were correctly or incorrectly predicted.

1. Confusion Matrix for Logistic Regression



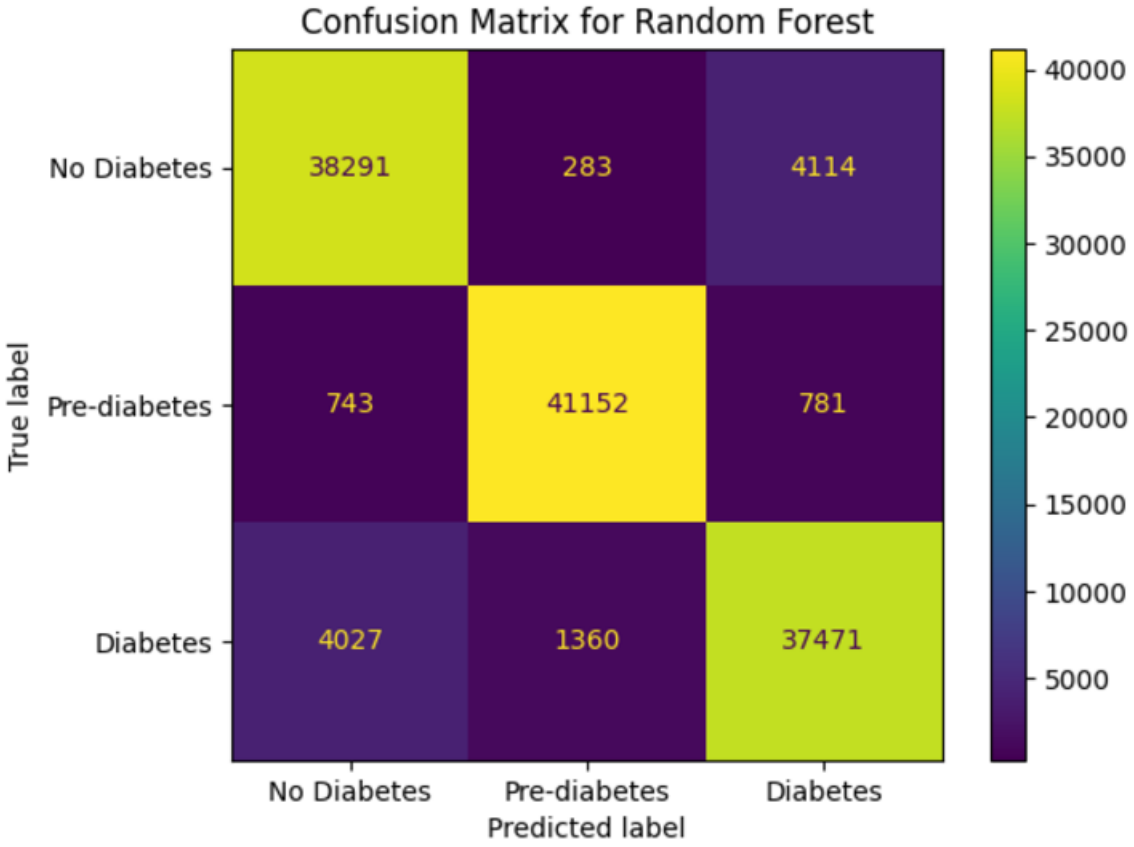
The Logistic Regression model, serving as the baseline, achieved an accuracy of 52%, with significant misclassifications across all classes. A considerable number of Pre-Diabetic (Class 1) cases were misclassified as Non-Diabetic (Class 0) or Diabetic (Class 2), indicating limited model capability in distinguishing between intermediate cases. Similarly, Non-Diabetic samples showed moderate confusion with both Pre-Diabetic and Diabetic classes. Overall, the model struggled to effectively separate the three classes, likely due to its simplistic linear decision boundary.

2. Confusion Matrix for Decision Tree



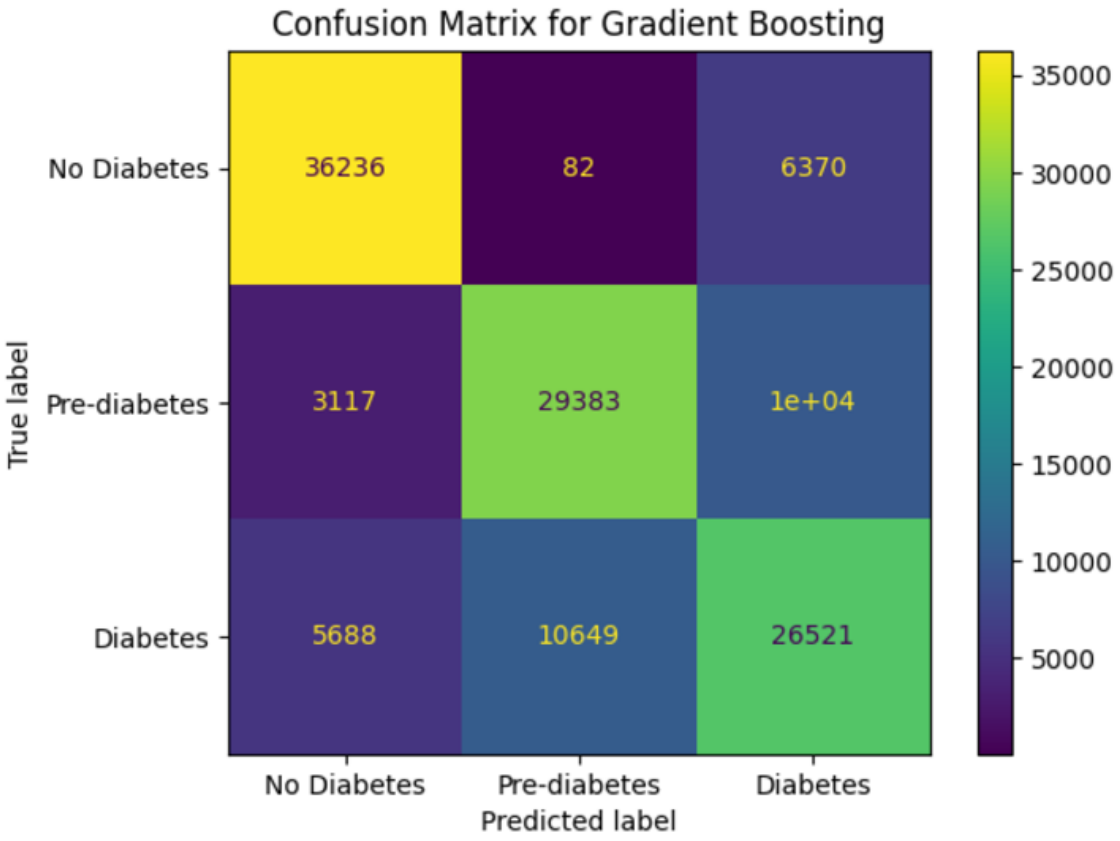
The Decision Tree model performed better than Logistic Regression, achieving an accuracy of 85%. It demonstrated improved classification across all three classes, with fewer misclassifications between Pre-Diabetic and Diabetic samples compared to Logistic Regression. However, there were still some notable errors, such as Diabetic samples being classified as Non-Diabetic. While the Decision Tree model showed substantial improvement, its performance was slightly below that of Random Forest, as it lacked the ensemble's ability to reduce overfitting and handle complex relationships in the data.

3. Confusion Matrix for Random Forest



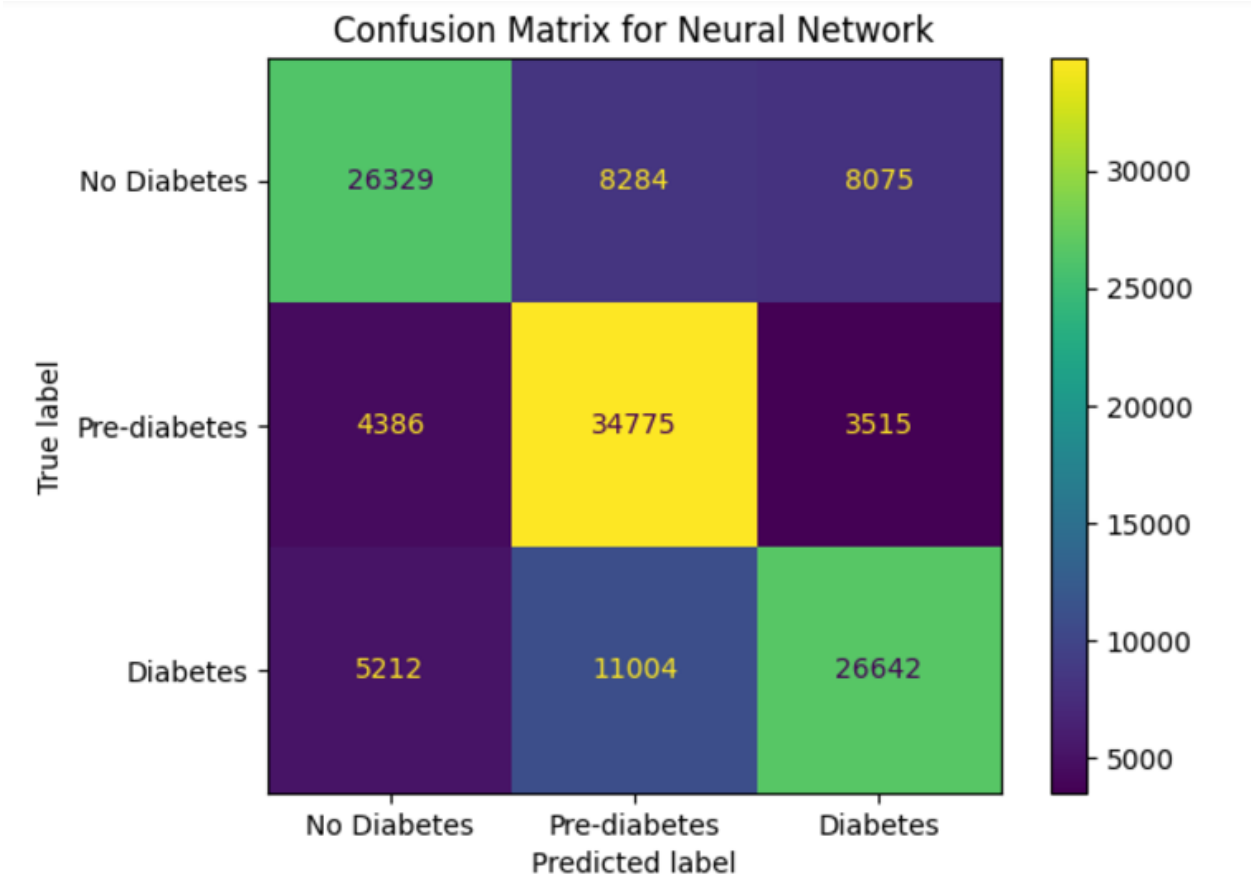
The confusion matrix for the Random Forest model, which achieved the highest accuracy of 91%, demonstrates its effectiveness in classifying all three classes—Non-Diabetic (Class 0), Pre-Diabetic (Class 1), and Diabetic (Class 2). Most predictions align with true labels, particularly for Pre-Diabetic and Diabetic classes, showing strong precision and recall. Misclassifications are minimal, with a few Non-Diabetic samples misclassified as Diabetic and vice versa. This highlights the model's robustness and its ability to handle the complexities of the dataset better than other models.

4. Confusion Matrix for Gradient Boosting



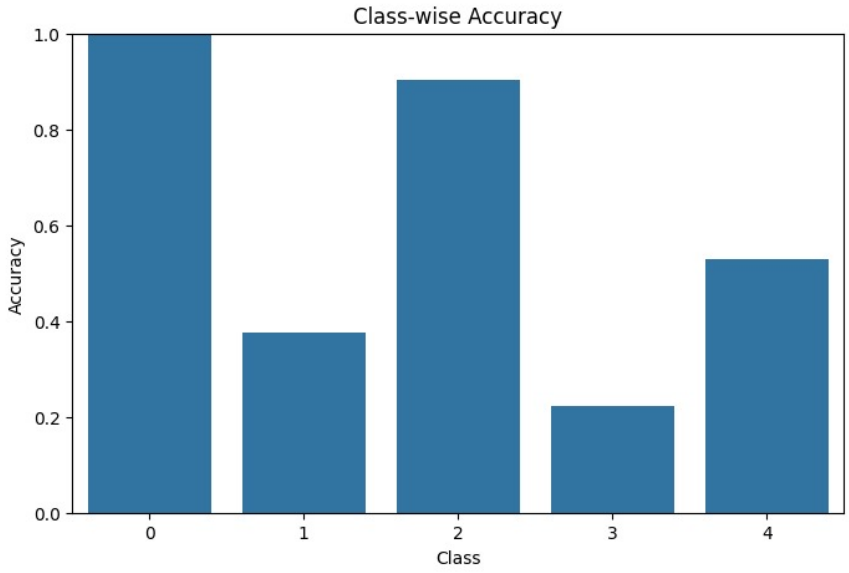
The Gradient Boosting model achieved an accuracy of 72%, showing an improvement over Logistic Regression but falling short of the performance of Decision Tree and Random Forest. This model displayed better distinction between Non-Diabetic and Pre-Diabetic classes, but it struggled with Diabetic samples, with notable misclassifications into both Non-Diabetic and Pre-Diabetic classes. While the model provided more consistent predictions compared to Logistic Regression, it lacked the robustness seen in ensemble-based methods like Random Forest.

5. Confusion Matrix for Neural Network



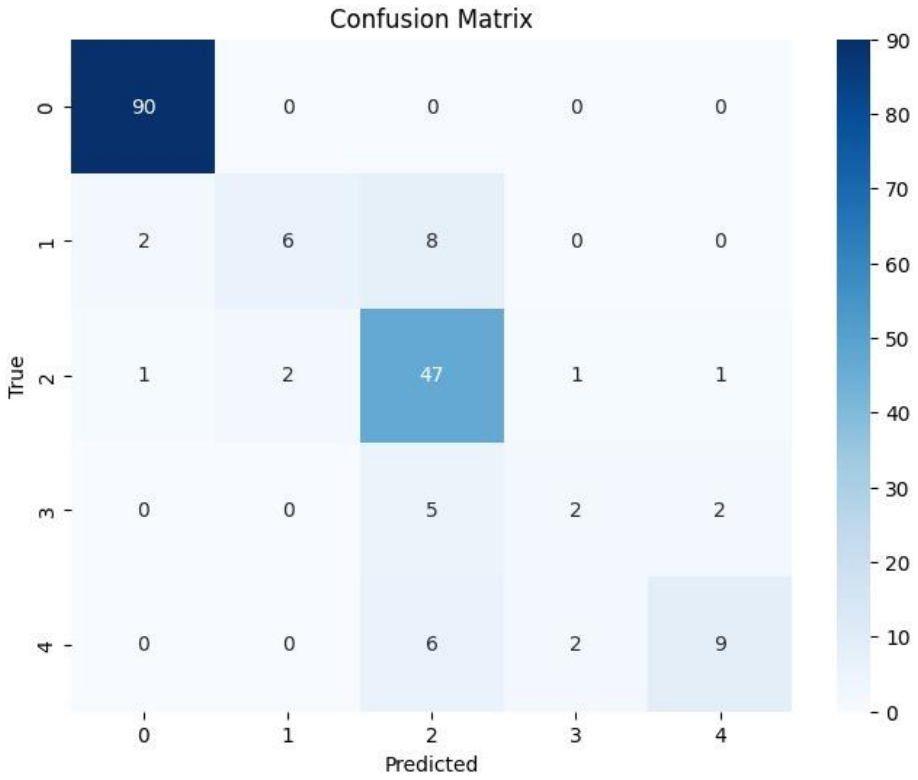
The confusion matrix for the neural network shows moderate performance. While it accurately classifies most Pre-Diabetic samples (34,775), it struggles with boundary cases, misclassifying a significant number of Diabetic cases as Pre-Diabetic (11,004) and Non-Diabetic cases into other classes. Although it captures some patterns, the model underperforms compared to Random Forest and could benefit from further fine-tuning or architecture adjustments.

Image Dataset:



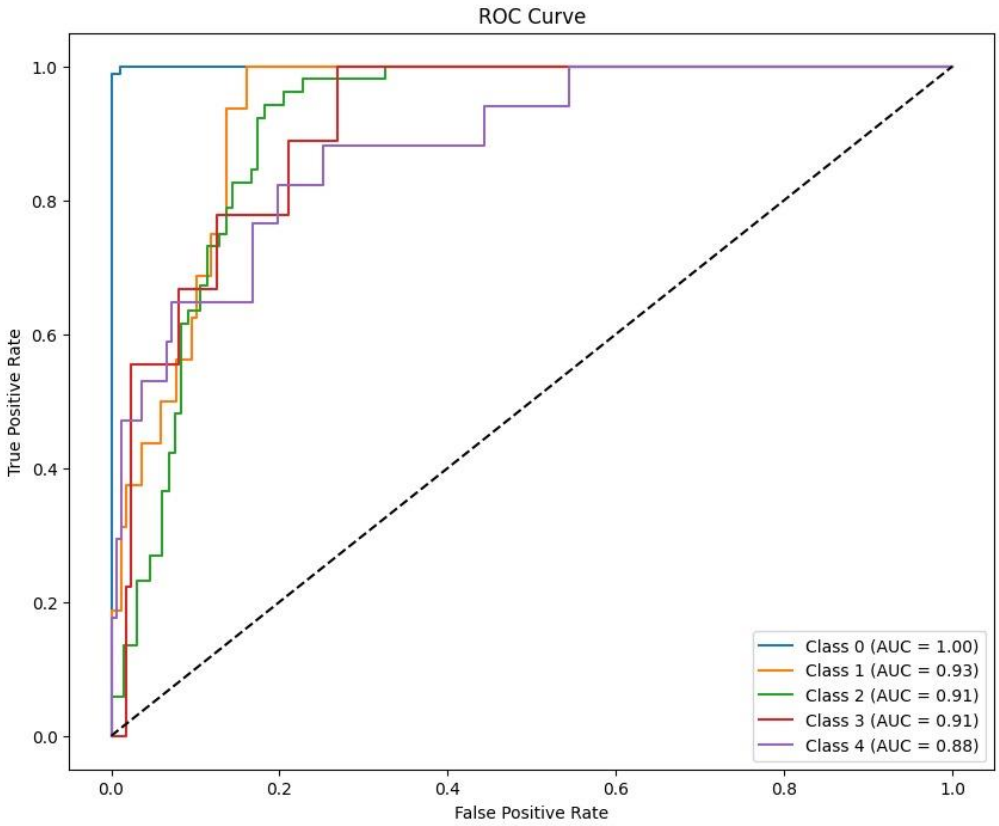
Class-wise Accuracy:

The bar chart shows the accuracy of the model for each class. The model performs exceptionally well on class 0 with near-perfect accuracy and moderately well on class 2. However, classes 1, 3, and 4 have lower accuracy, indicating that the model struggles to distinguish these classes, possibly due to data imbalance or overlap in features between these classes.



Confusion Matrix:

This heatmap illustrates how the predictions are distributed across actual and predicted labels. The diagonal values represent correct predictions, while off-diagonal values indicate misclassifications. The model excels in correctly predicting class 0, but misclassifies some samples from classes 1, 3, and 4 into neighboring classes like 2.



ROC Curve: The ROC curves for each class display the trade-off between the true positive rate and false positive rate. The AUC (Area Under the Curve) values are high, especially for class 0 (AUC = 1.00) and class 1 (AUC = 0.93), indicating strong performance. However, classes like 4 (AUC = 0.88) show slightly weaker discrimination.

9. Insights and Challenges

Key Insights from Model Results:

- Random Forest outperformed other models in terms of accuracy, precision, recall, and F1-score.
- Logistic Regression was the simplest and the fastest but did not capture the complexities in the data as well as Random Forest and Gradient Boosting.
- Decision Tree performed better than Logistic Regression but overfit the data, leading to poor generalization.
- Gradient Boosting was highly effective but slightly slower and less accurate than Random Forest.
- Neural Network underperformed compared to the other models due to the dataset's tabular nature.

Challenges:

- Class imbalance required SMOTE and weighted loss.
- Neural networks struggled with tabular data due to the data's structured nature.

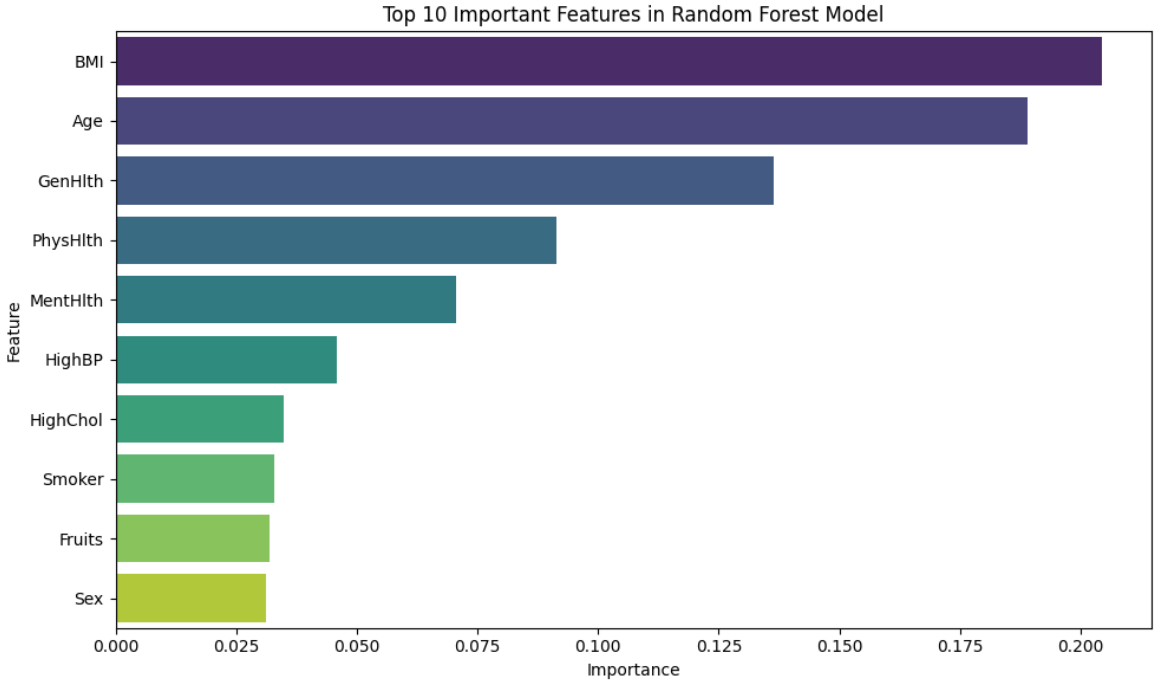
10. Findings and Conclusion

Random Forest was identified as the best model for this dataset based on accuracy and classification performance. It was particularly good at handling the class imbalance.

The diabetes-related features (such as BMI, HighBP, and Physical Activity) played an important role in distinguishing between Non-diabetic, Pre-diabetic, and Diabetic individuals.

The imbalanced nature of the dataset required careful attention during preprocessing, and SMOTE helped balance the classes for better model training.

Feature relationships, as revealed through EDA, confirmed the critical role of BMI and HighBP in diabetes classification.



11. Achievements and Future Work

Achievements

The project achieved notable success in predictive modeling:

- 91% accuracy with Random Forest for tabular data, highlighting its ability to handle structured features effectively.
- 82% accuracy with ResNet50 for image data, leveraging transfer learning to classify retinopathy severity with high precision.

Future Work

a. **Hybrid Model:** Combine predictions from tabular and image datasets into a unified framework to provide more comprehensive diagnostic insights.
b. **Advanced Architectures:** Test models like XGBoost and TabNet to further improve accuracy and interpretability for tabular data.
c. **Real-World Applications:** Deploy models in clinical settings with APIs or user-friendly tools, ensuring scalability and performance on real-world data.

11. References

- Kaggle, GitHub for datasets.
- PyTorch and Scikit-learn documentation for implementations.
- Deep Learning Techniques for Diabetic Retinopathy Classification: A Survey
- Diabetes Prediction Using Machine Learning Algorithms with Feature Selection and Hyperparameter Optimization
- Diabetic Retinopathy Detection and Severity Classification Using Optimized Deep Learning with Explainable AI
- Machine Learning in Precision Diabetes Care and Cardiovascular Risk Management
- A Systematic Review on Diabetic Retinopathy Detection Using Deep Learning Techniques
- Computationally Efficient Deep Learning Models for Diabetic Retinopathy Detection: A Systematic Literature Review
- Diabetes Prediction Using Machine Learning Approach
- Deep Learning in Automatic Diabetic Retinopathy Detection and Grading: A Comprehensive Review
- Machine Learning and Deep Learning Predictive Models for Type 2 Diabetes: A Systematic Review
- Detection and Classification of Diabetic Retinopathy Using Deep Learning Algorithms

Dataset Link: Deep Learning Dataset Link