Diabetes Multi-Class Classification and Clustering

Name: Anish Maharjan
Assignment: Classification_Takeaway_Assignment

1. Introduction

Diabetes is a chronic metabolic disorder that can be categorized into non-diabetic, prediabetic, and diabetic stages. Early and accurate identification of these stages is critical for preventing disease progression and complications.

This project aims to:

- Analyze a clinical diabetes dataset using Exploratory Data Analysis (EDA)
- Build supervised classification models to predict diabetic status
- Apply unsupervised clustering to explore natural groupings
- Compare supervised vs unsupervised approaches
- Interpret results from both medical and machine learning perspectives

2. Dataset Description

The dataset contains 1000 patient records with 14 attributes, including demographic information, biochemical test results, and a target class indicating diabetic status.

Columns:

- Demographics: Gender, Age
- Blood sugar marker: HbA1c
- Lipid profile: Cholesterol, Triglycerides, HDL, LDL, VLDL
- Kidney function: Urea, Creatinine
- Body composition: BMI
- Target: CLASS (N, P, Y)

The target variable CLASS represents:

- N: Non-diabetic
- P: Prediabetic
- Y: Diabetic

	ID	No_Pation	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS
0	502	17975	F	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	N
1	735	34221	M	26	4.5	62	4.9	3.7	1.4	1.1	2.1	0.6	23.0	N
2	420	47975	F	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	N
3	680	87656	F	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	N
4	504	34223	M	33	7.1	46	4.9	4.9	1.0	0.8	2.0	0.4	21.0	N

Figure 1: *Sample rows of the diabetes dataset*

```
(1000, 14)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 14 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   ID         1000 non-null   int64
 1   No_Pation  1000 non-null   int64
 2   Gender     1000 non-null   object
 3   AGE        1000 non-null   int64
 4   Urea       1000 non-null   float64
 5   Cr         1000 non-null   int64
 6   HbA1c      1000 non-null   float64
 7   Chol       1000 non-null   float64
 8   TG         1000 non-null   float64
 9   HDL        1000 non-null   float64
 10  LDL        1000 non-null   float64
 11  VLDL       1000 non-null   float64
 12  BMI        1000 non-null   float64
 13  CLASS      1000 non-null   object
dtypes: float64(8), int64(4), object(2)
memory usage: 109.5+ KB
```

Figure 2: *Dataset structure and data types*

3. Data Cleaning and Preprocessing

3.1 Label Cleaning

During initial analysis, inconsistencies were found in the target labels due to trailing whitespace (e.g., "Y" vs "Y "). These were cleaned using string normalization to ensure correct class grouping.

This step was crucial to avoid incorrect class counts and misleading model performance.

- value_counts() before cleaning

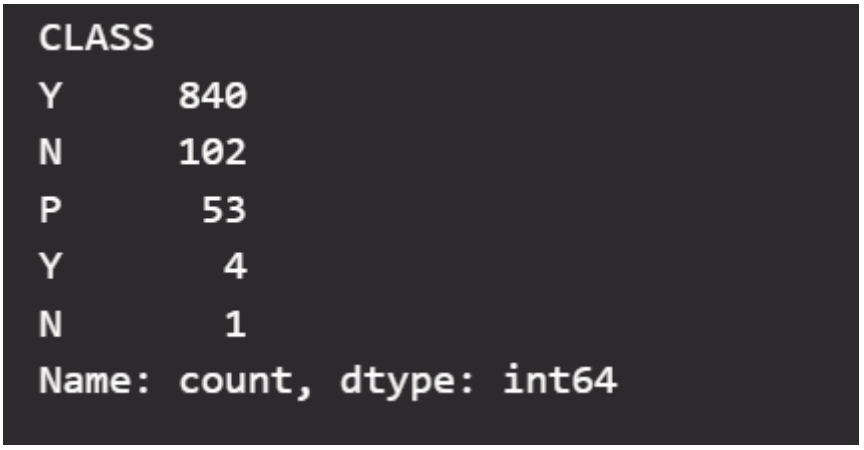

Figure 3: *Target class distribution before label cleaning*

- value_counts() after cleaning

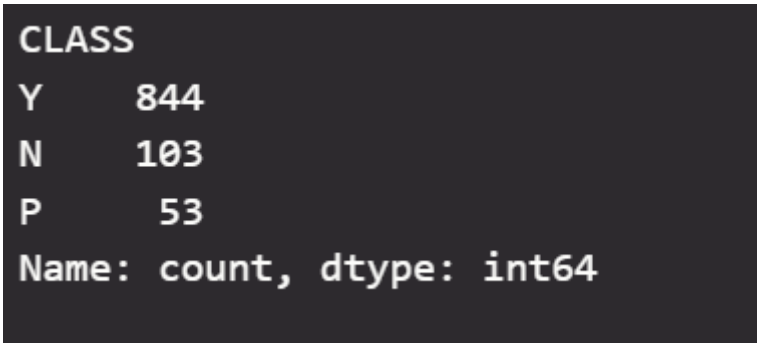

Figure 4: *Target class distribution after label cleaning*

3.2 Removal of Identifier Columns

Columns such as ID and No_Pation were removed before modeling. These columns serve only as identifiers and do not carry medical meaning. Including them could lead to data leakage, where the model memorizes records instead of learning patterns.

3.3 Preprocessing Pipeline

A preprocessing pipeline was used to ensure:

- Missing value handling (median imputation)
- Feature scaling (standardization)

- Categorical encoding (Gender)

Using a pipeline guarantees consistent preprocessing for both training and testing data and prevents data leakage.

4. Exploratory Data Analysis (EDA)

EDA was performed to understand data quality, distributions, relationships, and class imbalance.

4.1 Target Class Distribution

The dataset is highly imbalanced:

- Diabetic (Y): majority class
- Prediabetic (P): minority class
- Non-diabetic (N): minority class

This imbalance motivated the use of Balanced Accuracy and Macro F1-score for evaluation.

Figure 5: *Target class distribution after preprocessing*

4.2 Feature Distributions

Histograms of numeric features showed:

- Most patients are middle-aged or older
- BMI values indicate a largely overweight/obese population
- HbA1c values span normal to severe diabetic ranges
- Several biochemical features exhibit right-skewed distributions

This suggests a high-risk clinical population.

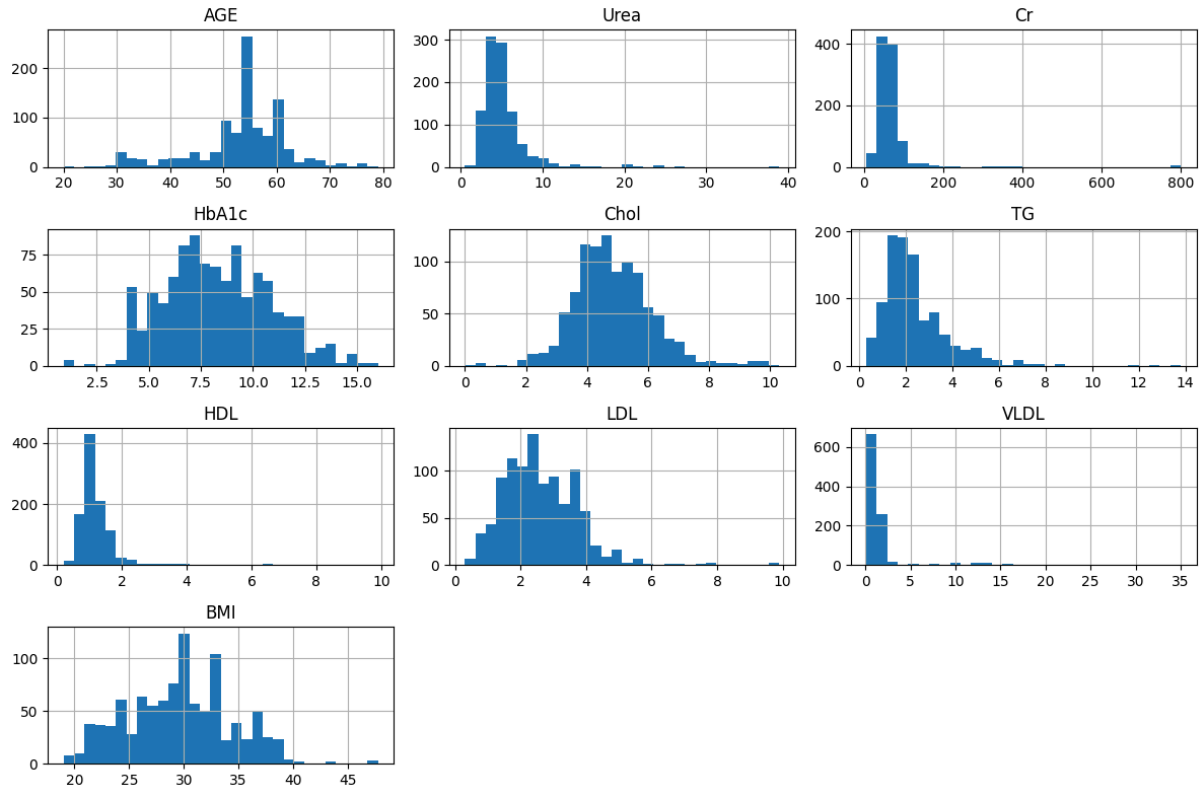

Figure 6: *Histogram of numeric feature distributions*

4.3 Correlation Analysis

Correlation analysis revealed:

- Expected relationships (e.g., Urea–Creatinine, TG–VLDL)
- No extreme multicollinearity
- HbA1c remains largely independent, indicating strong diagnostic value

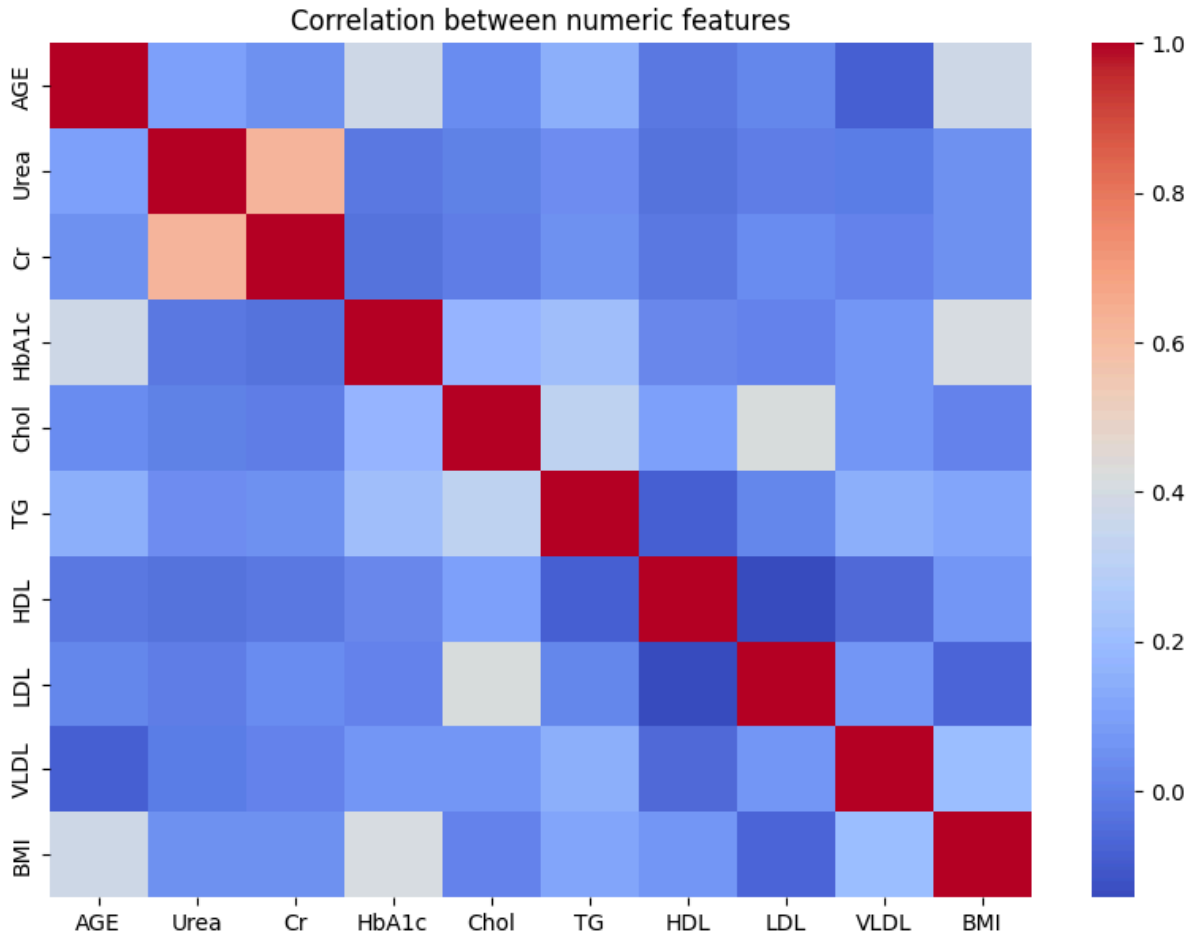

Figure 7: *Correlation heatmap of numeric features*

4.4 Feature vs Class Analysis

HbA1c by Class

HbA1c shows clear separation between classes, confirming it as the strongest indicator of diabetes.

BMI by Class

BMI shows an increasing trend from non-diabetic to diabetic patients, supporting its role as a risk factor.

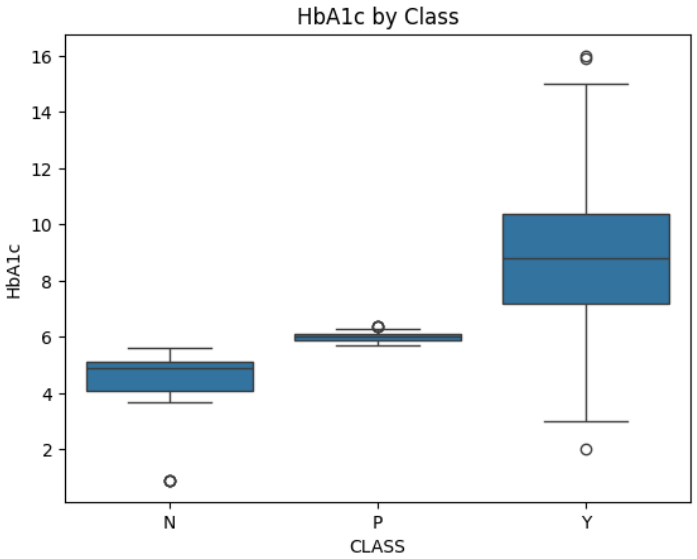

Figure 8: *HbA1c distribution by diabetic class*

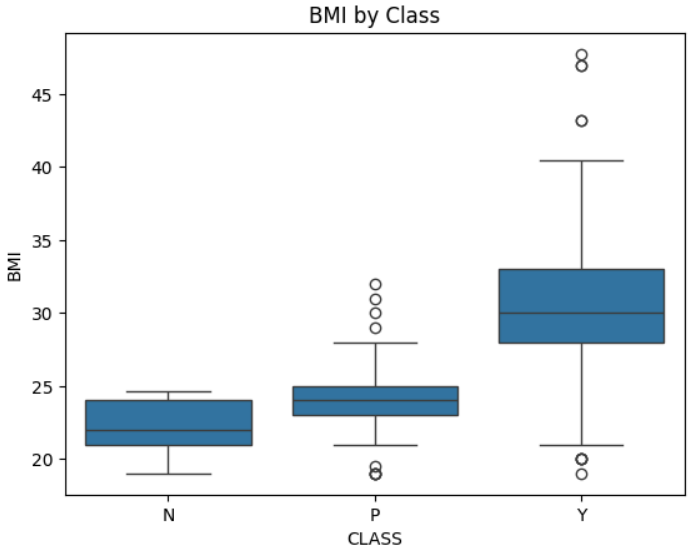

Figure 9: *BMI distribution by diabetic class*

5. Evaluation Metrics

Due to the imbalanced nature of the dataset, multiple evaluation metrics were used to assess model performance fairly and comprehensively. Relying solely on accuracy can lead to misleading conclusions, particularly when one class dominates the dataset. Therefore, the following metrics were employed.

5.1 Accuracy

Accuracy measures the proportion of correctly classified instances out of all predictions.

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions}$$

While accuracy provides an overall performance measure, it does not account for class imbalance. In this dataset, where diabetic cases form the majority, a model predicting only the majority class could still achieve high accuracy without being clinically useful.

Purpose:
Used as a general performance indicator, but not relied upon alone.

5.2 Balanced Accuracy

Balanced Accuracy computes the average recall obtained for each class.

$$Balanced\ Accuracy = \frac{Recall_N + Recall_P + Recall_Y}{3}$$

This metric ensures that each class contributes equally to the final score, regardless of class size.

Why it was used:
To fairly evaluate performance across non-diabetic, prediabetic, and diabetic classes in the presence of class imbalance.

5.3 Precision

Precision measures how many of the predicted instances for a class were actually correct.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

High precision indicates a low false-positive rate.

Clinical relevance:
Important for understanding how reliable a positive prediction is, especially when mislabeling a healthy patient as diabetic.

5.4 Recall (Sensitivity)

Recall measures how many actual instances of a class were correctly identified.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

High recall ensures that fewer true cases are missed.

Clinical relevance:
Missing diabetic or prediabetic patients can delay treatment, making recall a critical metric in medical contexts.

5.5 F1-Score

The **F1-score** is the harmonic mean of precision and recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

It provides a balanced measure when both false positives and false negatives are important.

5.6 Macro F1-Score

Macro F1-score calculates the F1-score independently for each class and then averages them equally.

$$Macro\ F1 = \frac{F1_N + F1_P + F1_Y}{3}$$

Why it was the primary metric:

- Treats all classes equally
- Prevents majority class dominance
- Provides a fair comparison across models

5.7 Confusion Matrix

The confusion matrix provides a detailed breakdown of correct and incorrect predictions for each class.

It enables:

- Identification of specific misclassification patterns
- Analysis of which classes are commonly confused
- Better understanding of model errors

This visualization was especially useful for interpreting errors involving prediabetic cases.

5.8 Clustering Evaluation Metrics

For clustering analysis, where true labels are not used during training, different metrics were applied:

- **Silhouette Score**: Measures how well-separated the clusters are
- **Adjusted Rand Index (ARI)**: Measures similarity between clusters and true labels, adjusted for chance
- **Normalized Mutual Information (NMI)**: Measures shared information between cluster assignments and true labels

These metrics help assess how well unsupervised clusters align with known diabetic categories.

6. Classification Models

Supervised classification was used because the target variable consists of known categorical labels.

Models evaluated:

1. Logistic Regression (baseline)
2. Random Forest
3. Gradient Boosting

The dataset was split into:

- 80% training data
- 20% testing data
  using stratified sampling.

6.1 Logistic Regression

Logistic Regression served as a baseline linear model.

- Good overall accuracy
- Reduced performance on the prediabetic class
- Limited ability to model complex interactions

```
Logistic Regression
Accuracy: 0.935
Balanced Acc: 0.8525
Macro F1: 0.8045

Report:
              precision    recall  f1-score   support

           N       0.86      0.90      0.88        21
           P       0.47      0.70      0.56        10
           Y       0.99      0.95      0.97       169

    accuracy                           0.94       200
   macro avg       0.77      0.85      0.80       200
weighted avg       0.95      0.94      0.94       200
```

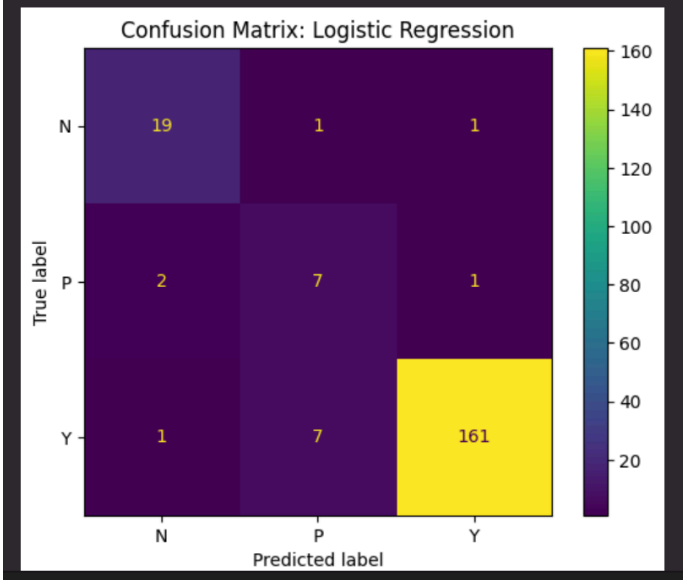

Figure 10: *Confusion matrix for Logistic Regression*

6.2 Random Forest (Best Model)

Random Forest achieved the strongest performance:

- Excellent accuracy and Macro F1-score
- Near-perfect classification across all classes
- Robust to noise and non-linear relationships

It was selected as the best model due to:

- High performance
- Stability
- Minimal hyperparameter tuning

- Better interpretability compared to boosting

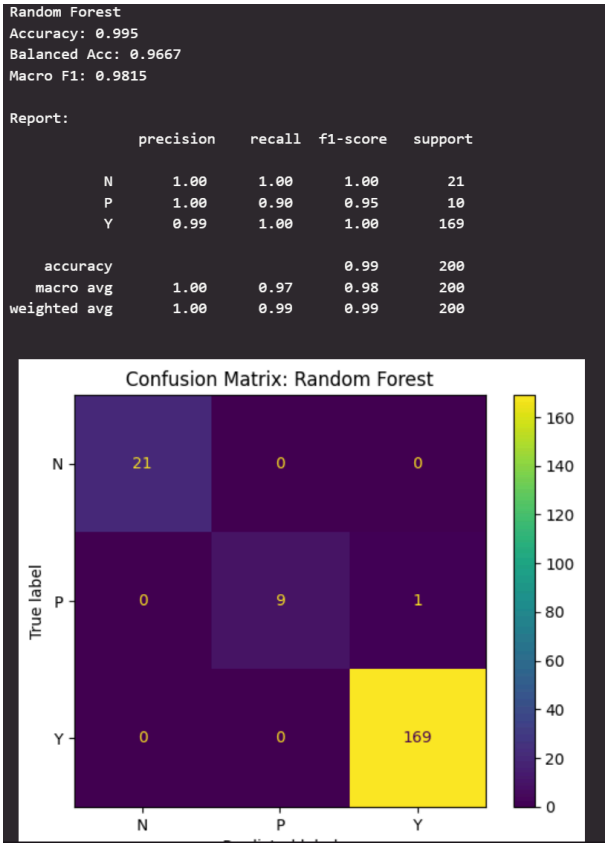

Figure 11: *Confusion matrix for Random Forest*

6.3 Gradient Boosting

Gradient Boosting achieved performance similar to Random Forest. However, due to its sequential nature and higher tuning complexity, Random Forest was preferred as the final model.

```
Gradient Boosting
Accuracy: 0.995
Balanced Acc: 0.9667
Macro F1: 0.9815

Report:
              precision    recall  f1-score   support

           N       1.00      1.00      1.00        21
           P       1.00      0.90      0.95        10
           Y       0.99      1.00      1.00       169

    accuracy                           0.99       200
   macro avg       1.00      0.97      0.98       200
weighted avg       1.00      0.99      0.99       200
```

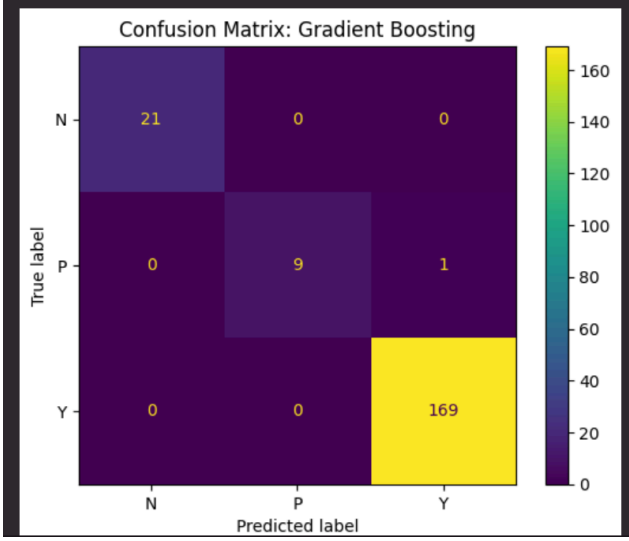

Figure 12: *Confusion matrix for Gradient Boosting*

7. Clustering Analysis (Unsupervised Learning)

KMeans clustering was applied with k = 3 to explore whether diabetic states naturally form groups without labels.

Evaluation Metrics:

- Silhouette Score
- Adjusted Rand Index (ARI)
- Normalized Mutual Information (NMI)

Results:

- Low silhouette score indicates weak natural cluster separation
- Moderate ARI and NMI show partial alignment with true labels

- Prediabetic patients were spread across clusters

This suggests diabetes progression is continuous, not discretely clustered.

```
Silhouette: 0.17597531048231388
ARI: 0.5657459682189259
NMI: 0.432558408911456
```

Figure 13: *Clustering evaluation metrics (Silhouette, ARI, NMI)*

TrueClass	N	P	Y
Cluster			
0	1	1	12
1	2	1	743
2	100	51	89

Figure 14: *Cluster vs true class distribution for KMeans clustering*

8. Feature Importance and Interpretation

Permutation feature importance was used to interpret the Random Forest model.

Key contributing features included:

- Cholesterol
- Triglycerides
- Urea
- BMI
- Gender

Although HbA1c showed low permutation importance, it remains clinically dominant. This behavior is a known limitation of permutation importance in tree-based models, where strong features are used early and become insensitive to permutation.

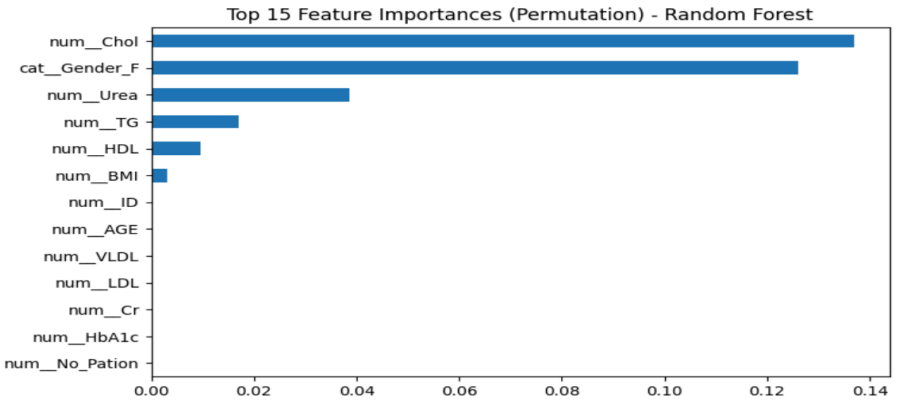

Figure 15: *Top feature importances using permutation importance (Random Forest)*

9. Error Analysis

Only a very small number of misclassifications were observed. Errors primarily involved borderline prediabetic cases being classified as diabetic.

These cases had:

- HbA1c values near clinical thresholds
- Elevated BMI or lipid values

This reflects real-world diagnostic ambiguity, not model failure.

Number of misclassified samples: 1

	ID	No_Pation	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	true	pred
152	135	45397	M	54	4.0	88	5.7	4.4	2.9	0.6	2.5	1.3	28.0	P	Y

Mean feature values for common confusion pairs (numeric only):

		ID	No_Pation	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI
true	pred												
P	Y	135.0	45397.0	54.0	4.0	88.0	5.7	4.4	2.9	0.6	2.5	1.3	28.0

Figure 16: *Misclassified samples and feature values from Random Forest model*

10. Comparison: Classification vs Clustering

Aspect	Classification	Clustering
Uses labels	Yes	No
Performance	Very high	Moderate
Prediabetes handling	Strong	Weak
Clinical usability	High	Exploratory

Table 1: *Comparison between classification and clustering approaches*

Supervised learning is more appropriate for diagnostic tasks, while clustering provides exploratory insights.

11. Conclusion

This study demonstrates that diabetes classification using routine clinical measurements is highly effective when supervised machine learning techniques are applied. HbA1c is the strongest predictor, with BMI, lipid profile, and kidney markers providing additional context. Clustering analysis highlights the continuous nature of diabetes progression and explains the difficulty of isolating prediabetes without labels.

12. Future Enhancements

- Apply SMOTE or advanced imbalance handling techniques
- Perform hyperparameter tuning using GridSearchCV
- Use SHAP for deeper model explainability
- Test additional clustering algorithms (GMM, hierarchical)
- Deploy the model via a web application or API