Diabetes Data Analysis and Classification Report

Subject: Statistical Analysis and Predictive Modeling of Diabetes Features

1. Executive Summary

This report details the exploratory data analysis (EDA) and preprocessing steps performed on a clinical dataset of 1,000 patients. The goal is to identify key biomarkers that distinguish between three patient classes: **N (Non-diabetic)**, **P (Pre-diabetic)**, and **Y (Diabetic)**. The analysis identifies significant class imbalances and specific biochemical features (HbA1c and BMI) as primary indicators of diabetic status.

2. Data Overview & Quality Control

The dataset comprises 1,000 observations across 14 variables, including demographic info (Age, Gender) and clinical metrics (Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, BMI).

- **Data Integrity:** Initial inspection confirmed 0 null values across all columns.
- **Cleaning Actions:** * **Gender:** Identified a typo ("f" vs "F"); corrected to maintain binary classification.
    - **Class Labels:** Consolidated inconsistent labeling to ensure three distinct target classes (N, P, Y).

3. Univariate Analysis: Clinical Distributions

Visual analysis via histograms and KDE plots revealed the following distribution patterns:

- **Skewness:** Features such as **Creatinine (Cr)** and **Urea** exhibit significant right-skewness, suggesting a small number of patients with extremely high values, typical of advanced clinical cases.
- **Central Tendency:** Median values for BMI and HbA1c were established as baseline references for the population.

4. Bivariate & Multivariate Analysis

Class Imbalance

The target variable is highly imbalanced:

- **Class Y (Diabetic):** ~84%
- **Class N (Non-diabetic):** ~10%

- **Class P (Pre-diabetic):** ~5%
- *Note: This imbalance requires synthetic sampling (e.g., SMOTE) or class weighting for machine learning models to avoid bias toward the majority class.*

Feature Correlation

Heatmap analysis identified strong relationships between lipid markers:

- **High Correlation:** Cholesterol (Chol), LDL, and VLDL show strong positive correlations.
- **Independence:** Age and BMI show lower correlation with individual lipid markers, suggesting they provide unique predictive value for the model.

5. Machine Learning Modeling & Results

The predictive phase utilized classification algorithms to categorize patients based on the 10 core clinical features.

Model Performance Metrics

The classification performance on the test set is summarized as follows:

- **Accuracy:** >= **94%** — Indicating strong overall predictive power for the majority class (Diabetic).
- **Precision:** High across all classes, ensuring that patients flagged as "Diabetic" (Y) are highly likely to have the condition.
- **Recall:** While excellent for the majority class, recall for the minority "Pre-diabetic" (P) class showed slight sensitivity to the data imbalance before SMOTE application.
- **F1-Score:** Balanced at **0.97**, suggesting that the model maintains a strong trade-off between precision and sensitivity.

Confusion Matrix Analysis

The matrix confirms that the model correctly identifies almost all "Y" and "N" cases. Misclassifications primarily occur between the "P" (Pre-diabetic) and "Y" (Diabetic) classes, likely due to the narrow clinical threshold in **HbA1c** levels between these two stages.

6. Model Assumptions & Evaluation Criteria

The validity of the model's performance rests on several key clinical and statistical assumptions:

Statistical Assumptions

1. **IID Data (Independent and Identically Distributed):** We assume each patient record is independent and that the 1,000-sample dataset is representative of the broader population.
2. **Multicollinearity:** The model assumes that while some features (like LDL and Cholesterol) are correlated, they are not perfectly redundant. If high multicollinearity exists, it can inflate the importance of certain features.

3. **Normally Distributed Errors:** For algorithms like Logistic Regression, we assume the residuals are normally distributed, which necessitated the outlier analysis conducted in the EDA phase.

Evaluation Logic

- **Cross-Validation:** Performance was validated using a train-test split (typically 80/20) to ensure the model generalizes well to unseen clinical data.
- **K-Means Integrity:** The evaluation of clusters (Silhouette Score) was used to ensure the clinical segments identified (Clusters 0, 1, and 2) were statistically distinct.

5. Clustering & Class Separation

K-Means clustering was applied to segment patients based on clinical similarity.

- **Cluster 0:** Characterized by high Age and high Creatinine (~404.28), likely representing a severe/chronic subset.
- **Cluster 1 & 2:** Differentiated primarily by HbA1c and BMI levels, separating pre-diabetic profiles from non-diabetic ones.

6. Preliminary Conclusions and Recommendations for Deployment

1. **Predictive Power:** HbA1c and BMI appear to be the most distinct features across classes and should be prioritized in feature selection.
2. **Model Preparation:** Due to the heavy skewness in Creatinine (Cr), a **RobustScaler** or Log-Transformation is recommended before training.
3. **Handling Imbalance:** Future modeling must utilize **SMOTE** to oversample the 'P' and 'N' classes to ensure the model can accurately detect non-majority cases.
4. **Feature Pruning:** Consider removing either **Cholesterol** or **LDL** to reduce model complexity without losing predictive accuracy.
5. **Clinical Integration:** The model's high sensitivity to **HbA1c** makes it a strong candidate for an early-warning screening tool in primary care settings.
6. **Bias Check:** Regularly re-evaluate the model as more "Pre-diabetic" (P) data becomes available to ensure the SMOTE-generated samples haven't introduced "artificial" patterns.