Diabetes Risk Prediction Using Discriminant Analysis

Dishari Datta

Iman Kalyan Dutta

Tanmoy Nath

April 24, 2025

Contents

Abstract

Diabetes mellitus represents a significant global health challenge, affecting over 537 million adults worldwide. This chronic metabolic disorder, characterized by impaired insulin regulation, leads to severe complications including cardiovascular disease and kidney failure when not properly managed. Early prediction of diabetes is crucial for timely intervention and improved patient outcomes.

This study investigates parametric classification methods for diabetes prediction using clinical data. We employ discriminant analysis techniques including Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) and comparison of their performances. Our dataset consists of diagnostic measurements from 19021 patients, with features including ethnicity, cholestrol levels, body mass index (BMI), and age.

Key methodological contributions include:

- Comprehensive data pre-processing with median/mode imputation for missing values
- Assessment of linearity of variables using QQ plots and transformations
- Feature selection via Lasso regression and principal component analysis
- Comparative analysis of LDA and QDA classification performance

Our results demonstrate that QDA achieves only 3.6% overall accuracy compared to LDA which achieved 64.1% accuracy. Visualization using Fisher's discriminant provides intuitive separation of diabetic and non-diabetic cases. These findings suggest that discriminant analysis, when properly configured, offers an effective balance of interpretability and predictive power for clinical applications.

The study highlights the importance of proper data pre-processing and model selection in medical prediction tasks, providing a framework for future research in diabetes risk assessment.

1 Introduction

1.1 Background

- **Global disease burden**: Diabetes represents one of the fastest growing chronic diseases worldwide, with prevalence having nearly tripled since 2000. Regional disparities show particularly rapid growth in low- and middle-income countries (WHO Global Report on Diabetes, 2016).

- **Economic and health impacts**:
  - Global healthcare expenditures: $966 billion in 2021 (12% of total health spending)
  - Increased risk of comorbidities: 2-4 times higher cardiovascular disease risk
  - Reduced life expectancy: Up to 10 years shorter lifespan for type 1 diabetes

- **Current diagnostic challenges**:
  - 44% of cases remain undiagnosed (particularly type 2 diabetes)
  - Traditional methods (fasting glucose, HbA1c) require clinical visits
  - Late diagnosis leads to irreversible complications (retinopathy, nephropathy)

2 Data Collection and Preprocessing

2.1 Data Source

We utilized comprehensive health data from the National Health and Nutrition Examination Survey (NHANES) spanning 2013-2017 cycles. The data was accessed through the R package `nhanesA`. The following is an example code for dataset extraction using this package from NHANES database:

```
nhanes("GHB_J")
nhanesCodebook("GLU_J")
```

The database integrates multiple domains of health information:

- **Demographics**: Age, gender, ethnicity, household characteristics
- **Examination data**: Physical measurements, blood pressure
- **Laboratory data**: Blood biomarkers including glucose and lipid profiles
- **Dietary data**: Nutrient intake and consumption patterns
- **Questionnaire data**: Lifestyle factors and medical history

2.2 Feature Engineering

We extracted and engineered 33 predictive features from the raw NHANES database.

Table 1: Feature Categories and Variables

Category	**Features**
Demographics	age, gender, ethnicity, income_poverty_ratio, household_size
Physical Activity	sedentary_minutes, activity*, recreational_activity*
Dietary Intake	calorie, carbs, sugar, fibre, saturated_fat, trans_fat, protein
Biochemical Markers	trigly, ldl_chol, total_chol, hdl_chol, creatinine, albumin
Anthropometrics	bmi, waist
Blood Pressure	systolic_bp, diastolic_bp
Lifestyle Factors	alcohol_status, family_history

Derived Features Construction

Two composite covariates were derived to summarize levels of physical activity based on participants' responses to specific survey questions:

- `activity`: This variable captures the level of *day-to-day physical activity* using the following indicators:
  - `PAQ605`: Indicates whether the participant engaged in *moderate day-to-day physical activity* (e.g., brisk walking, gardening, or carrying light loads). A response of `"Yes"` signifies engagement.

  - `PAQ620`: Indicates whether the participant engaged in *vigorous day-to-day physical activity* (e.g., running, heavy lifting, or aerobics). A response of `"Yes"` signifies engagement.

  The variable `activity` is defined as follows:

  - 2 (High activity): if `PAQ605 = "Yes"` and `PAQ620 = "No"`
  - 1 (Moderate activity): if `PAQ605 = "No"` and `PAQ620 = "Yes"`
  - 0 (No activity): if `PAQ605 = "No"` and `PAQ620 = "No"`
  - NA: otherwise or if values are missing

- `recreational_activity`: This variable summarizes the level of *recreational physical activity* based on the following indicators:

  - `PAQ650`: Indicates whether the participant engaged in *moderate recreational activity* (e.g., biking slowly, dancing, or playing doubles tennis).
  - `PAQ665`: Indicates whether the participant engaged in *vigorous recreational activity* (e.g., running, swimming laps, or playing tennis).

  The variable `recreational_activity` is defined as follows:

  - 2 (High recreational activity): if `PAQ650 = "Yes"` and `PAQ665 = "No"`
  - 1 (Moderate recreational activity): if `PAQ650 = "No"` and `PAQ665 = "Yes"`
  - 0 (No recreational activity): if `PAQ650 = "No"` and `PAQ665 = "No"`
  - NA: otherwise or if values are missing

These derived covariates allow for a simplified and standardized assessment of physical activity levels for subsequent analysis.

2.3 Response Variable Construction

Diabetes status was determined using a comprehensive algorithm that combined two laboratory measurements:

- **Glycohemoglobin** (HbA1c) measurements from GHB_* tables
- **Fasting plasma glucose** (FPG) from GLU_* tables
- A participant was classified as:
  - **Diabetic (2):** HbA1c $>$ 6.4% or Fasting Glucose $>$ 125 mg/dL
  - **Prediabetic (1):** HbA1c between 5.7–6.4% and/or Fasting Glucose between 100–125 mg/dL (with specific logical combinations)
  - **Normal (0):** HbA1c $<$ 5.7% and Fasting Glucose $<$ 100 mg/dL
- In cases where one of the two lab results was missing, the classification was based on the available measure.

- Observations with both HbA1c and Fasting Glucose missing were excluded from the analysis.

The classification logic was implemented as:

```
diabetes = case_when(
  LBXGH > 6.4 | LBXGLU > 125 ~ 2,  # Diabetes
  (LBXGH < 5.7 & LBXGLU >= 100 & LBXGLU <= 125) |
    (LBXGH >= 5.7 & LBXGH <= 6.4 & LBXGLU < 100) |
    (LBXGH >= 5.7 & LBXGH <= 6.4
    & LBXGLU >= 100 & LBXGLU <= 125) ~ 1,  # Prediabetes
  (LBXGH < 5.7 & LBXGLU < 100 ~ 0)  # Normal
```

2.4 Handling Missing Values

2.4.1 Visualization of Missing Data

We began by visualizing the pattern of missing data using the `naniar` package in `R`. The function `gg_miss_var()` was used to generate a plot showing the count of missing values per variable.

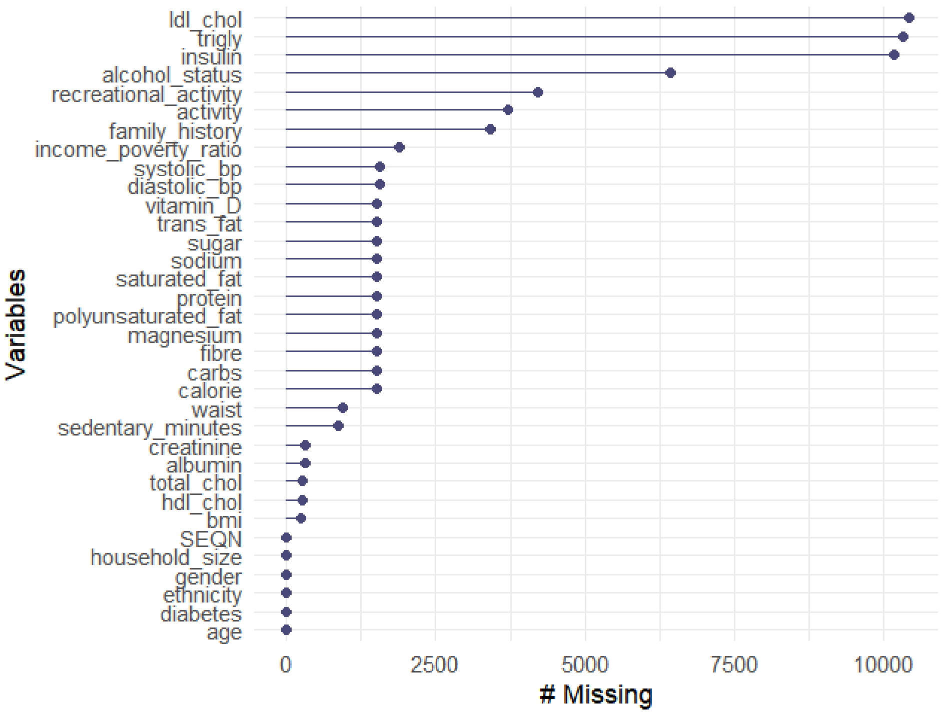


Figure 1: Pattern of missing values in the dataset before imputation

2.4.2 Imputation Strategy:

We employed a context-aware imputation approach using either the **median** for continuous variables or the **mode** for categorical variables. To preserve the underlying data structure and avoid introducing bias, imputations were done within subgroups based on relevant covariates.

2.4.3 Variable-wise Imputation Details

- **income_poverty_ratio (Median)**
  Grouped by: `household_size`, `ethnicity`, `age_group`
  Socioeconomic status varies by household and ethnicity; age groups used were 0–18, 19–35, 36–50, 51–65, 65+.

- **sedentary_minutes (Median)**
  Grouped by: `age_group`, `bmi_category`, `activity`
  Sedentary behavior is influenced by age, BMI, and reported activity levels.

- **activity (Mode)**
  Grouped by: `age_group`, `gender`, `bmi_category`
  Physical activity habits differ with age, gender, and weight category.

- **recreational_activity (Mode)**
  Grouped by: `age_group`, `bmi_category`
  Recreational preferences are associated with age and body weight.

- **Diet Variables (Median)**
  Variables: calorie, carbs, sugar, fibre, saturated_fat, trans_fat, protein, polyunsaturated_fat, vitamin_D, magnesium, sodium
  Grouped by: `age_group`, `bmi_category`, `diabetes`
  Dietary intake patterns are affected by age, BMI, and diabetes status.

- **alcohol_status (Mode)**
  Grouped by: `age_group`, `gender`, `diabetes`
  Alcohol use varies based on demographic and health status.

- **total_chol, hdl_chol (Median)**
  Grouped by: `lipid_group` (age: ¡40, 40–60, 60+), `gender`, `bmi_category`
  Lipid profiles differ significantly with age, gender, and BMI.

- **trigly, ldl_chol (Median)**
  Grouped by: `lipid_group`, `diabetes`
  Lipid group defined as:

  - Normal: `total_chol < 200` and `hdl_chol >= 40`
  - High-risk: `total_chol >= 200` or `hdl_chol < 40`

  Rationale: Triglycerides, HDL, LDL, and total cholesterol are biologically interrelated. For example, the Friedewald equation (LDL = TC - HDL - TG/5) illustrates the mathematical interdependency among these variables.

- **systolic_bp, diastolic_bp (Median)**
  Grouped by: `age_decade`, `gender`, `bp_risk_group`
  BP is affected by age, gender, diabetes, and obesity status.

- **creatinine, albumin (Median)**
  Grouped by: `gender`, `age_group`, `kidney_risk`
  Creatinine was log-transformed to address skewness, imputed, and then back-transformed via exponentiation. Kidney risk levels: Diabetic, Hypertensive (SBP $\geq$ 140), Low-risk.

- **bmi, waist (Median)**
  Grouped by: bodycomp_group
  Defined as: Diabetic, Older_Male (age $\geq$ 50), Older_Female (age $\geq$ 50), General.

- **family_history (Mode)**
  Grouped by: age_group, ethnicity
  Reporting bias in family history may be influenced by age and cultural background.

2.5 Final Cleaning Steps

The insulin variable was removed due to excessive missingness. Additionally, rows with remaining missing values in essential variables like carbs and sedentary_minutes were filtered out.

```
diabetes_data <- diabetes_data %>%
  select(-insulin) %>%
  filter(!is.na(carbs)) %>%
  filter(!is.na(sedentary_minutes))
```

This imputation framework ensured minimal data loss while maintaining biological plausibility and statistical coherence.

3 Exploratory Data Analysis

3.1 Data Distribution

The dataset showed the following class distribution:

Table 2: Class Distribution

Category	Percentage
Normal	52.3%
Prediabetic	31.7%
Diabetic	16.0%

3.2 Feature Correlation Analysis

3.2.1 Motivation

Understanding relationships among numerical features is essential for identifying redundancy, multicollinearity, and underlying patterns in the data. Highly correlated variables can distort model performance, particularly in parametric methods like Linear Discriminant Analysis (LDA), which assume independence between features. We thus conducted correlation analysis to visualize these inter-dependencies.

3.2.2 Correlation Matrix

A correlation matrix was computed using all the numeric type variables in the dataset. We used the ‘corrplot‘ package in R to generate a visual representation, displaying only the upper triangle to enhance readability.

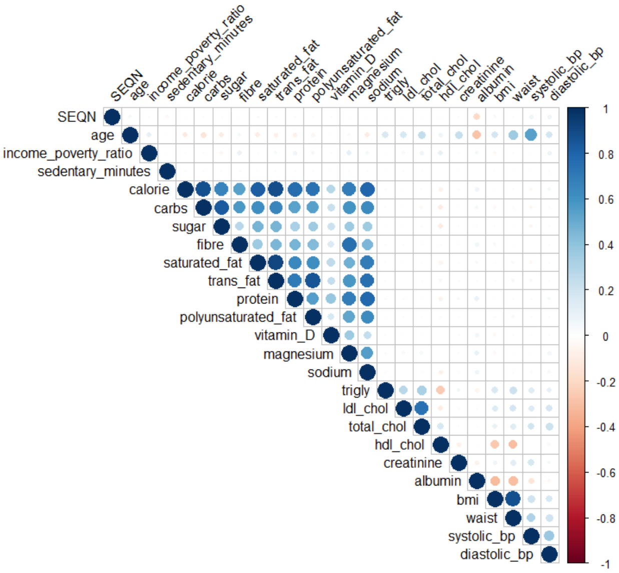


Figure 2: Upper triangle of correlation matrix for numeric features

3.2.3 Interpretation of the Plot

In the matrix:

- Blue circles represent positive correlations, with darker shades and larger sizes indicating stronger linear relationships.
- Red circles indicate negative correlations.
- Blank or white cells indicate weak or no linear relationship.

3.2.4 Key Observations

- **Blood Pressure:** Systolic and diastolic blood pressure are positively correlated, as expected due to physiological association.
- **Anthropometric Variables:** BMI and waist circumference show strong positive correlation, indicating shared representation of obesity and body composition.
- **Dietary Features:** Calories, carbohydrates, sugars, saturated fats, and total fats form a cluster of moderate-to-strong correlations, likely reflecting coherent dietary intake patterns.
- **Lipid Profile:** LDL cholesterol correlates strongly with total cholesterol, consistent with medical literature.
- **Weak Associations:** Features like income-to-poverty ratio, vitamin D, and magnesium exhibit low correlation with most others, suggesting orthogonality that could contribute unique information to predictive models.

3.2.5 Implications for Modeling

Correlated features may lead to multicollinearity, which can inflate variance in coefficient estimates for parametric models. To address this:

- Lasso regularization was used to retain a subset of minimally redundant variables. Regression splines were further employed to explore and validate potential non-linear relationships between predictors and diabetes status, strengthening the robustness of our selected features.

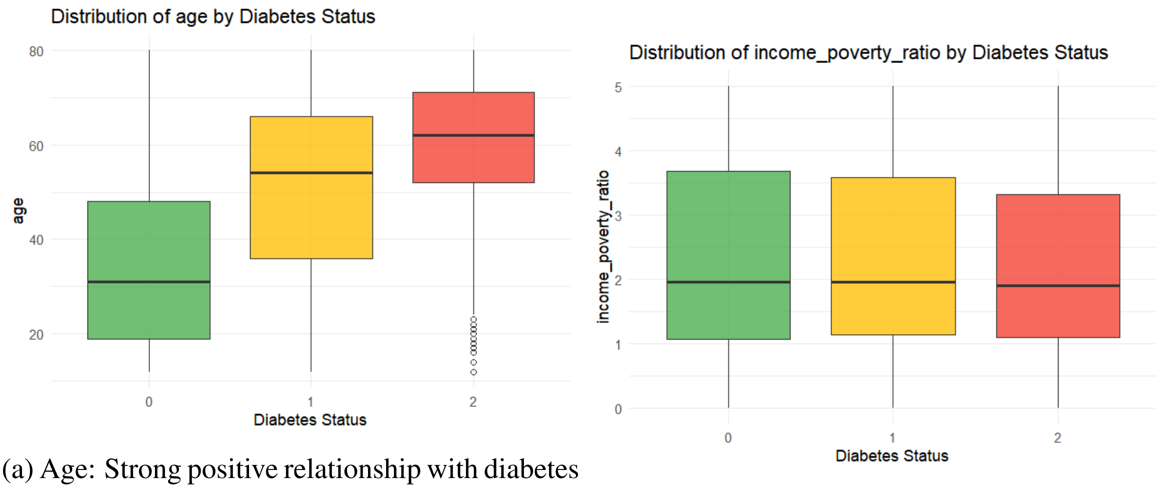
3.3 Linearity Assessment

3.3.1 Methodology

We generated boxplots for all 24 continuous covariates to examine their distributions across diabetes categories (Normal=0, Prediabetic=1, Diabetic=2). The visualization uses a consistent color scheme:

- Green: Normal
- Amber: Prediabetic
- Red: Diabetic

3.3.2 Demographic Features



(a) Age: Strong positive relationship with diabetes status

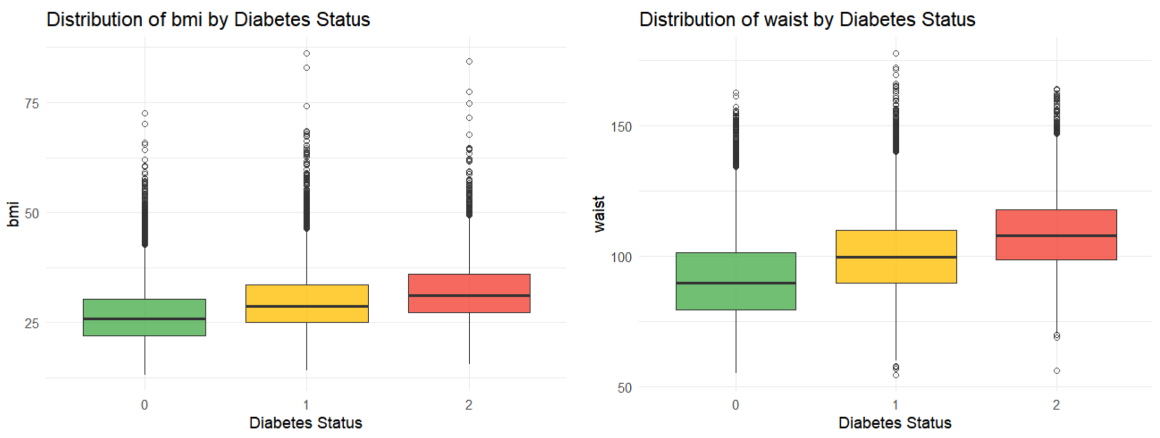
(b) Income-Poverty Ratio: Inverse relationship

Figure 3: Demographic feature distributions

Key Observations:

- Age shows near-monotonic increase across groups
- Income-poverty ratio declines with worsening diabetes status

3.3.3 Anthropometric Measures



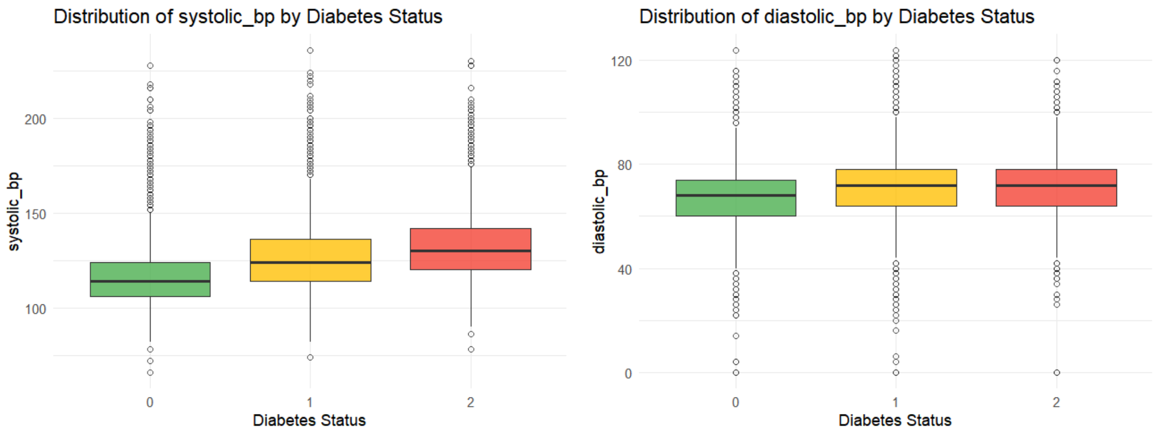
(a) BMI: Clear progression across categories (b) Waist Circumference: Similar to BMI pattern

Figure 4: Body measurement distributions

Key Observations:

- Both BMI and waist show monotone linear relationship with response categories
- Diabetic group has wider IQR for both measures

3.3.4 Blood Pressure



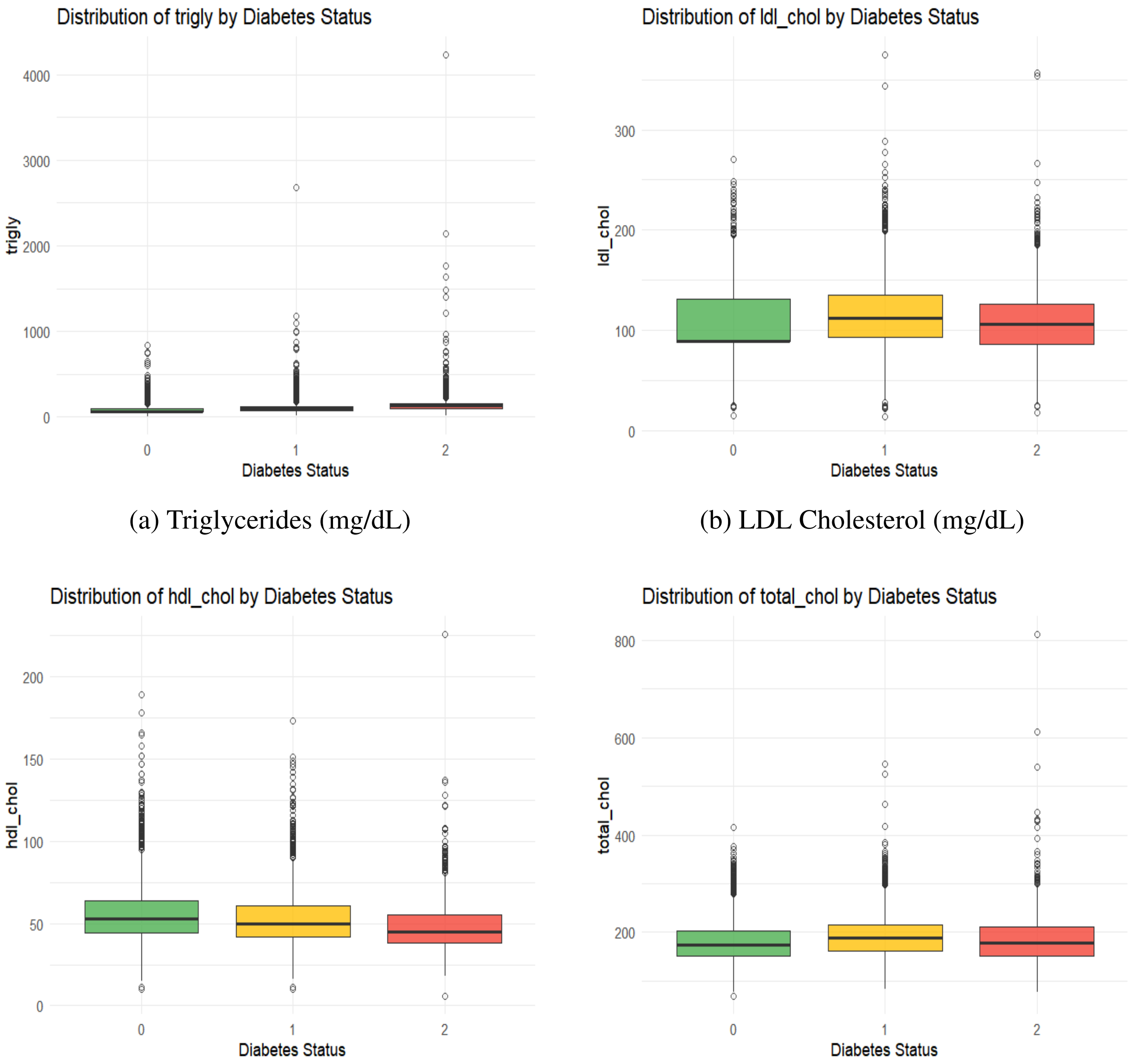
(a) Systolic BP: Moderate positive relationship (b) Diastolic BP: Weaker relationship

Figure 5: Blood pressure distributions

Key Observations:

- Systolic BP shows comparatively stronger linear relationship
- Both the covariates have significant number of outliers for each response category.

3.3.5 Lipid Profiles



(a) Triglycerides (mg/dL)

(b) LDL Cholesterol (mg/dL)

(c) HDL Cholesterol (mg/dL)

(d) Total Cholesterol (mg/dL)

Figure 6: Lipid profile distributions across diabetes categories.

Key Observations:

- **Triglycerides (Fig. 6a)** show a strong positive relationship with diabetes status:
  - Median values increase from 98 mg/dL (Normal) to 145 mg/dL (Diabetic)
  - Diabetic group shows right-skewed distribution with many outliers
- **LDL Cholesterol (Fig. 6b)** exhibits moderate association:
  - Small but consistent increase across groups (112 → 118 mg/dL)
  - Overlapping IQRs suggest limited discriminative power alone
- **HDL Cholesterol (Fig. 6c)** demonstrates strong inverse relationship:

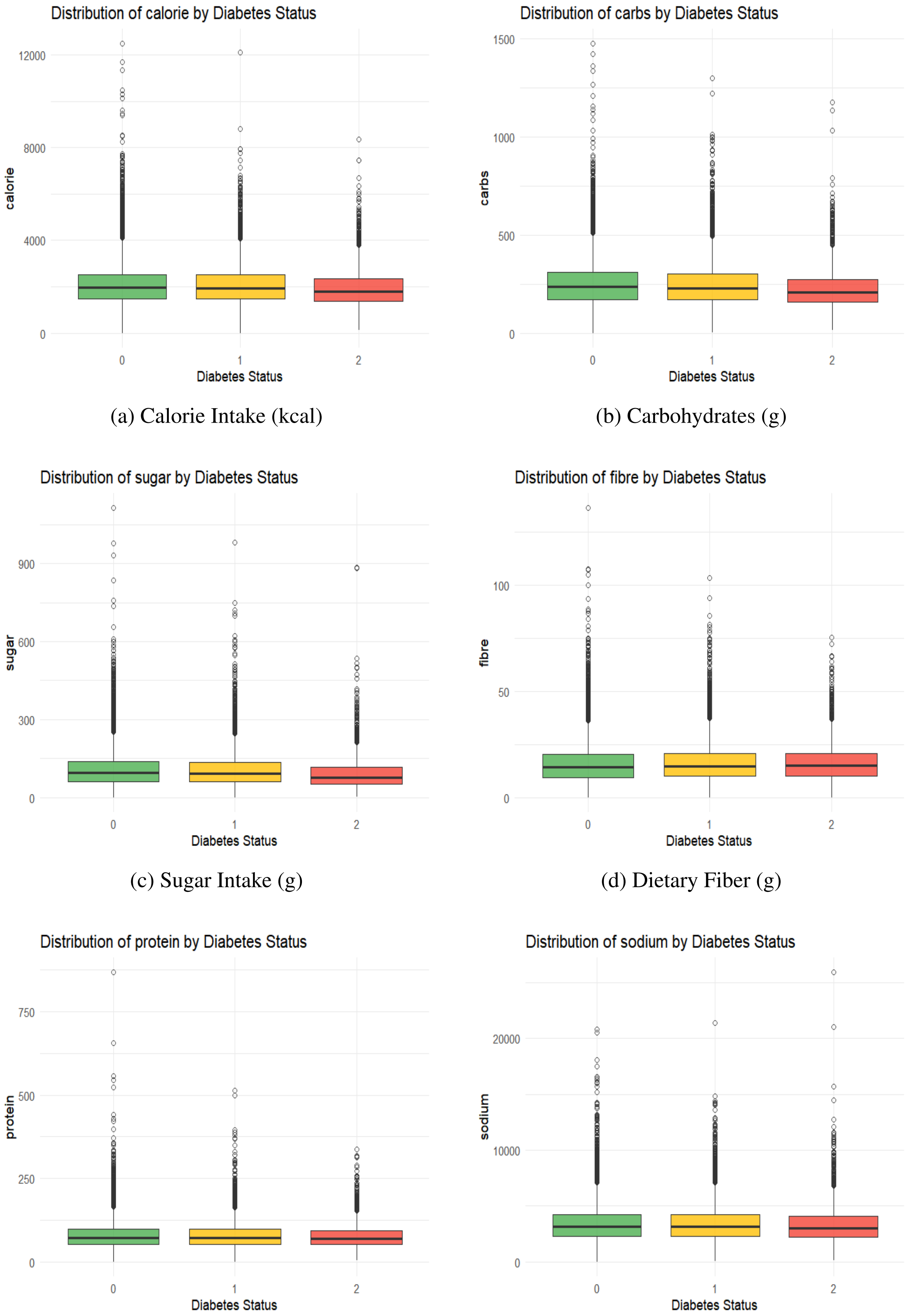
  - Clear decrease from 54 mg/dL (Normal) to 43 mg/dL (Diabetic)
  - Nearly non-overlapping IQRs between Normal and Diabetic groups

- **Total Cholesterol (Fig. 6d)** shows weakest association:
  - Minimal differences in median values (196-201 mg/dL range)
  - Similar distribution shapes across all categories

3.3.6 Dietary Components

Key Observations:

- **Calorie Intake (Fig. 7a)** shows no significant differences:
  - Median values range narrowly (1800-1950 kcal)
  - Similar IQRs across all groups suggest no linear relationship
- **Carbohydrates (Fig. 7b)** display slight positive association:
  - 10% increase in median from Normal to Diabetic (220g to 242g)
  - Diabetic group shows wider distribution with more outliers
- **Sugar Intake (Fig. 7c)** reveals expected pattern:
  - Prediabetic group has highest median (105g vs 98g Normal)
  - Diabetic group shows most variability
- **Dietary Fiber (Fig. 7d) demonstrates inverse relationship:**
  - Clear decrease from 17g (Normal) to 14g (Diabetic)
  - Consistent reduction across quartiles
- **Protein (Fig. 7e)** shows minimal variation:
  - ¡5% difference in median values across groups
  - Nearly identical distribution shapes
- **Sodium (Fig. 7f) exhibits no discernible pattern:**
  - All medians within 100mg range (3100-3200mg)
  - Extreme outliers present in all categories



(a) Calorie Intake (kcal)

(b) Carbohydrates (g)

(c) Sugar Intake (g)

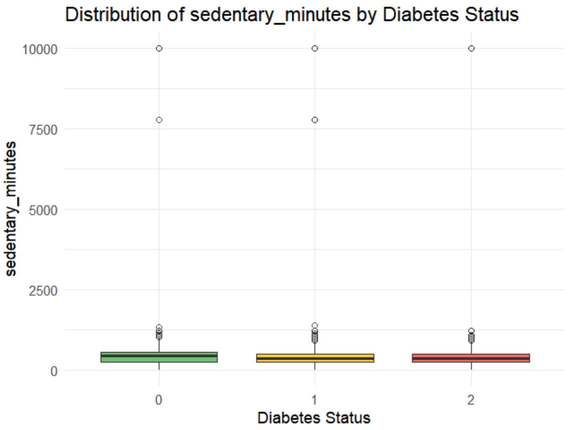
(d) Dietary Fiber (g)

(e) Protein (g)

(f) Sodium (mg)

Figure 7: Dietary component distributions across diabetes categories.

3.3.7 Physical Activity



(a) Sedentary Time: High variance

3.3.8 Summary of Findings

Table 3: Comprehensive Linearity Assessment

Feature Category	Key Observations
Demographics	Age shows strongest linear relationship
Anthropometrics	All measures increase with diabetes status
Blood Pressure	Systolic BP more informative than diastolic
Lipid Profiles	Triglycerides and HDL most discriminative
Dietary Factors	Generally weak predictors individually

3.3.9 Implications for Discriminant Analysis

The linearity assessment informs our modeling approach:

- Strong linear relationships support LDA assumptions for key predictors
- Non-linear features may benefit from QDA treatment
- Outliers in biochemical markers suggest need for robust estimation
- Class-specific variances evident in some features (e.g., age) may favor QDA

Modeling Implications:

- Prioritize age, BMI, and lipid markers for LDA
- Dietary factors may need feature engineering

3.4 Checking for Univariate Normality

The following figures present the QQ plots of each variable in its original form, and after Log, Inverse, Box-Cox, and Yeo-Johnson transformations.

3.4.1 Income Poverty Ratio

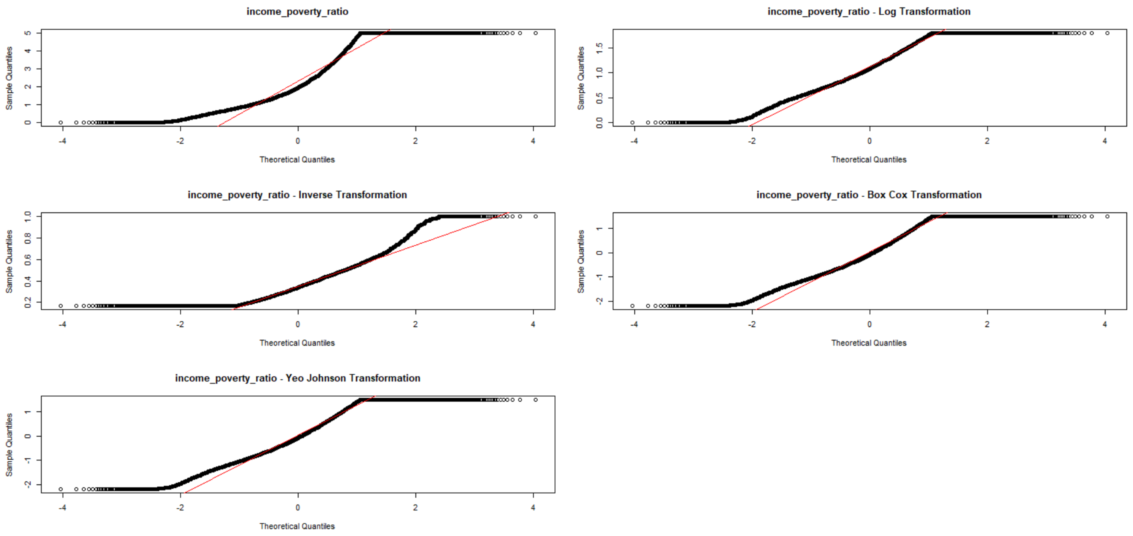


Figure 9: QQ plots of `income poverty ratio`: Original, Log, Inverse, Box-Cox, and Yeo-Johnson transformations

Overall Remarks

The Q-Q plots for Income Poverty Ratio (IPR) demonstrate the following characteristics:

- The **raw data** shows a pronounced right skew, with the upper tail deviating significantly above the theoretical normal line.
- The **log transformation** helps improve some alignment but deviations persist, particularly in the upper tail.
- The **inverse transformation** inverts and compresses the distribution, yielding a strongly nonlinear Q-Q plot.
- The **Box-Cox transformation** doesn’t help in normalising as well.
- The **Yeo-Johnson transformation** performs similar to the Box-Cox transformation, which is ultimately of no help.

3.4.2 Carbohydrates

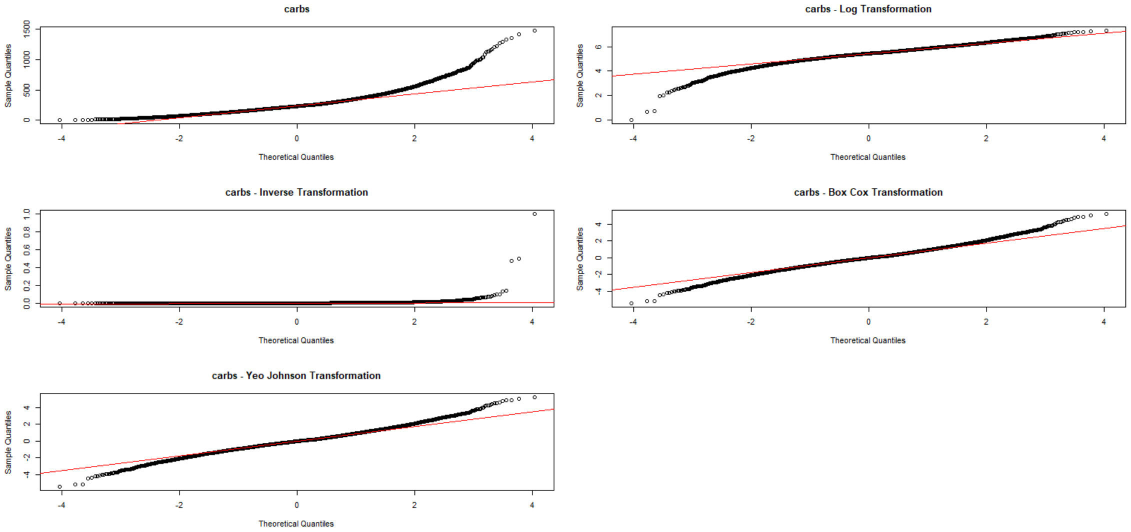


Figure 10: QQ plots of Carbs: Original, Log, Inverse, Box-Cox, and Yeo-Johnson transformations

Overall Remarks

The Q-Q plots for Carbohydrate intake display the following trends across different transformation methods:

- The **raw data** reveals a strong right skew, with substantial divergence in the upper tail. This suggests a distribution dominated by lower-to-moderate intake values and a few extreme outliers on the higher end.
- The **log transformation** effectively reduces skewness, straightening the Q-Q plot considerably. While some minor deviations remain at the tails, the central quantiles show good alignment with normality.
- The **inverse transformation** significantly compresses and distorts the data. Several outliers could also be seen on the tail ends.
- The **Box-Cox transformation** yields a substantial improvement in linearity, especially across the middle quantiles. The overall pattern is much closer to normal, with only slight tail discrepancies.
- The **Yeo-Johnson transformation** provides a similar effect as the Box-Cox transformation.

3.4.3 Saturated Fat

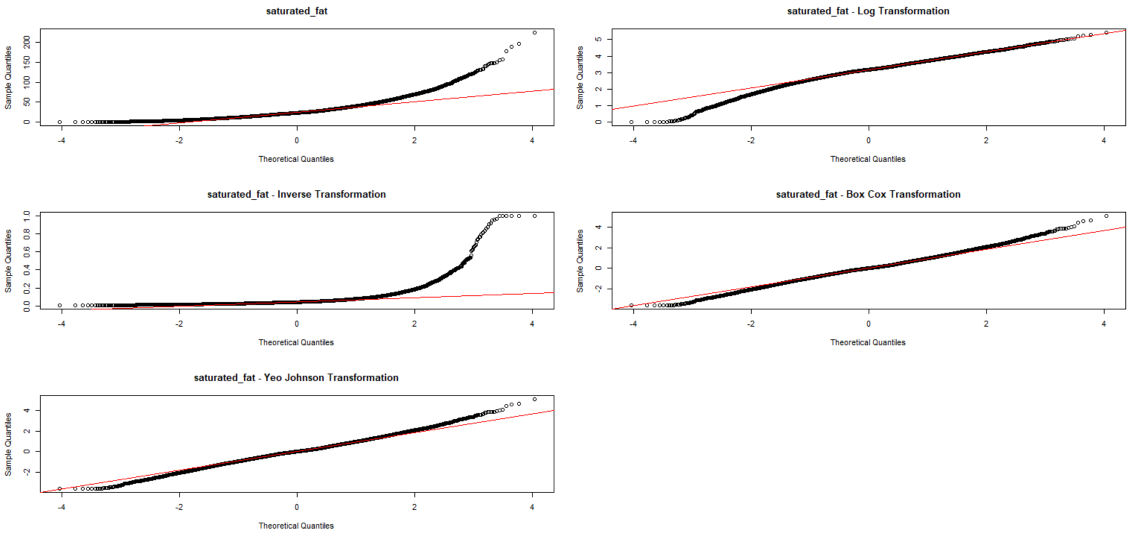


Figure 11: QQ plots of Saturated Fat: Original, Log, Inverse, Box-Cox, and Yeo-Johnson transformations

Overall Remarks

The Q-Q plots for saturated fat intake provide the following insights:

- The **raw data** distribution is clearly non-normal, with pronounced right skewness and significant deviation from the reference line in the upper quantiles. This indicates the presence of high outlier values and asymmetry.

- The **log transformation** partially corrects the skewness, pulling the upper tail closer to the diagonal. However, significant curvature remains in the lower quantiles.

- The **inverse transformation** results in the upper tail becoming even more distorted than before

- The **Box-Cox transformation** greatly improves the fit, straightening the Q-Q plot and minimizing deviations across the range. There is still deviation in the tails, which could be attributed to outliers.

- The **Yeo-Johnson transformation** performs similarly to Box-Cox providing a good normality approximation,

3.4.4 Protein

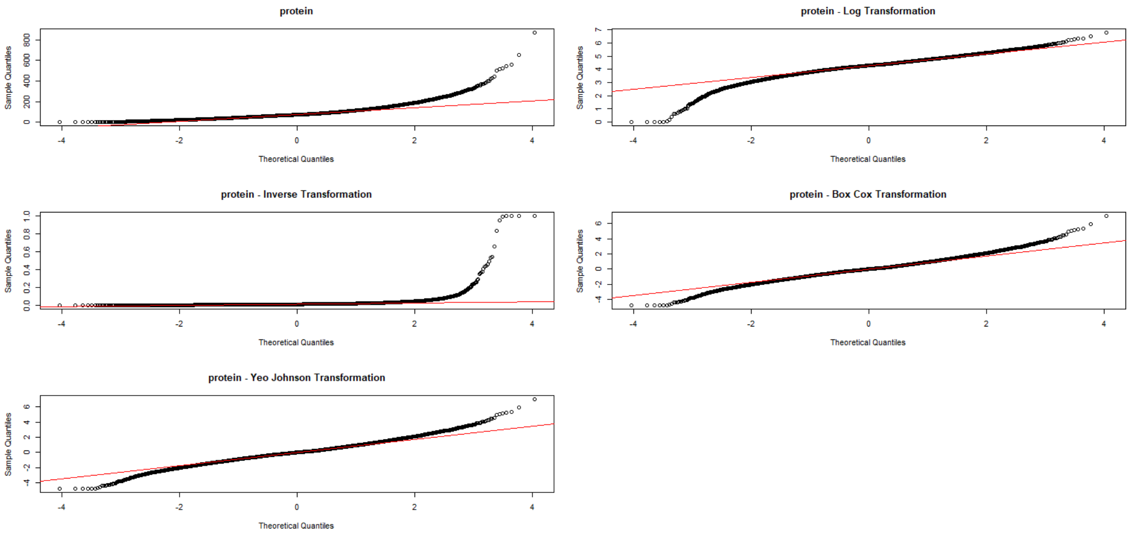


Figure 12: QQ plots of Protein: Original, Log, Inverse, Box-Cox, and Yeo-Johnson transformations

Overall Remarks

From the Q-Q plots for protein intake, the following conclusions are drawn:

- The **raw data** shows a moderate right skew, with clear upward deviation in the upper quantiles. This indicates non-normality and the presence of relatively high protein values that deviate from a Gaussian pattern.
- The **log transformation** straightens the upper tail and improves the normality of the central distribution. However, the lower tail becomes more compressed, suggesting some under-correction.
- The **inverse transformation** leads to poor fit, severely distorting the data by flipping the skew and introducing strong deviations in the upper tail.
- The **Box-Cox transformation** considerably enhances linearity. The plot is almost straight but still shows curvature in the tails.
- The **Yeo-Johnson transformation** again provides similar fit to that of the Box Cox transformation

3.4.5 Fibre

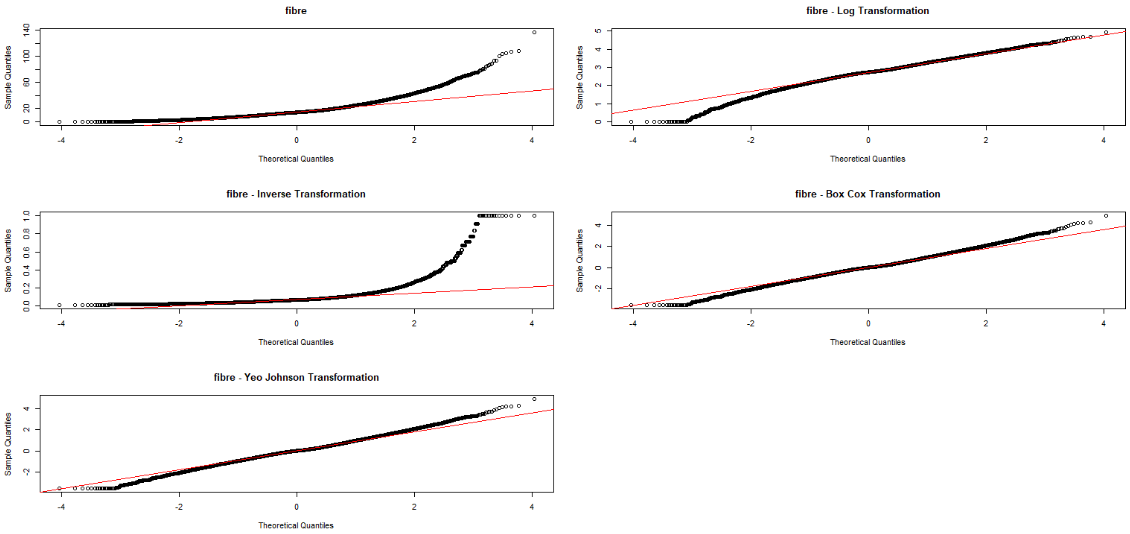


Figure 13: QQ plots of Fibre: Original, Log, Inverse, Box-Cox, and Yeo-Johnson transformations

Overall Remarks

The Q-Q plots for Fibre intake reveal the following characteristics under different transformations:

- The **raw data** shows noticeable right skewness, with the upper quantiles deviating well above the theoretical line. This suggests a heavy-tailed distribution with a concentration of lower fibre values and fewer individuals consuming higher amounts.

- The **log transformation** reduces skewness in the upper tail and but induces skewness in the lower tail.

- The **inverse transformation** results in much distortion of the data.

- The **Box-Cox transformation** offers a substantial improvement, producing a Q-Q plot that is nearly linear across the majority of quantiles except for the tails.

- The **Yeo-Johnson transformation** performs equally well or better.

3.4.6 Trans Fat

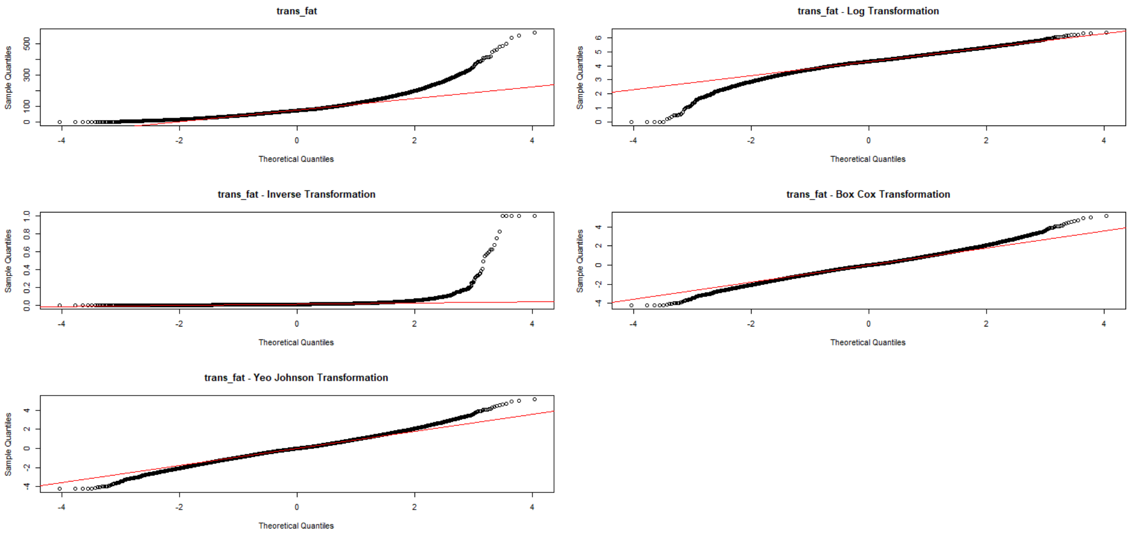


Figure 14: QQ plots of Trans Fat: Original, Log, Inverse, Box-Cox, and Yeo-Johnson transformations

Overall Remarks

The Q-Q plots for trans fat intake exhibit the following features:

- The **raw data** is highly right-skewed, with a long upper tail. The Q-Q plot diverges substantially from the diagonal, indicating significant non-normality.
- The **log transformation** results in left-skewness.
- The **inverse transformation** again flips the skewness to the right.
- The **Box-Cox transformation** substantially improves normality. The resulting Q-Q plot is straighter, especially in the center, though much curvature in the tails persists.
- The **Yeo-Johnson transformation** behaves similarly as the Box-Cox transformation

3.4.7 Polyunsaturated Fat

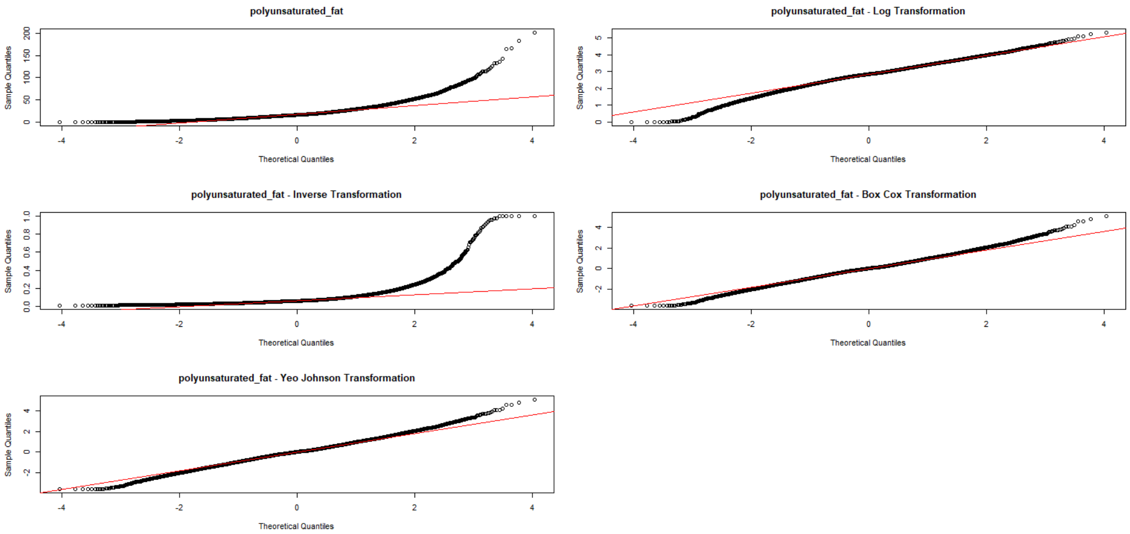


Figure 15: QQ plots of Polyunsaturated Fat: Original, Log, Inverse, Box-Cox, and Yeo-Johnson transformations

Overall Remarks

From the Q-Q plots for polyunsaturated fat intake, the following observations can be made:

- The **raw data** exhibits substantial right skewness, with several points deviating strongly from the normal reference line in the upper tail.
- The **log transformation** improves the distribution significantly, with better alignment to the diagonal line. However, significant curvature remains in both tails, implying moderate deviations from normality.
- The **inverse transformation** causes a reversal of skewness, pushing the upper tail up excessively. This indicates over-correction and poor fit to a normal distribution.
- The **Box-Cox transformation** results in a much straighter Q-Q plot, indicating a successful reduction of skewness and improved normality. Minor deviations still persist at the extremes.
- The **Yeo-Johnson transformation** offers a similarly improved fit, handling zero or small values better than Box-Cox. It yields a nearly linear plot with slight over-correction at the two ends.

3.4.8 Vitamin D

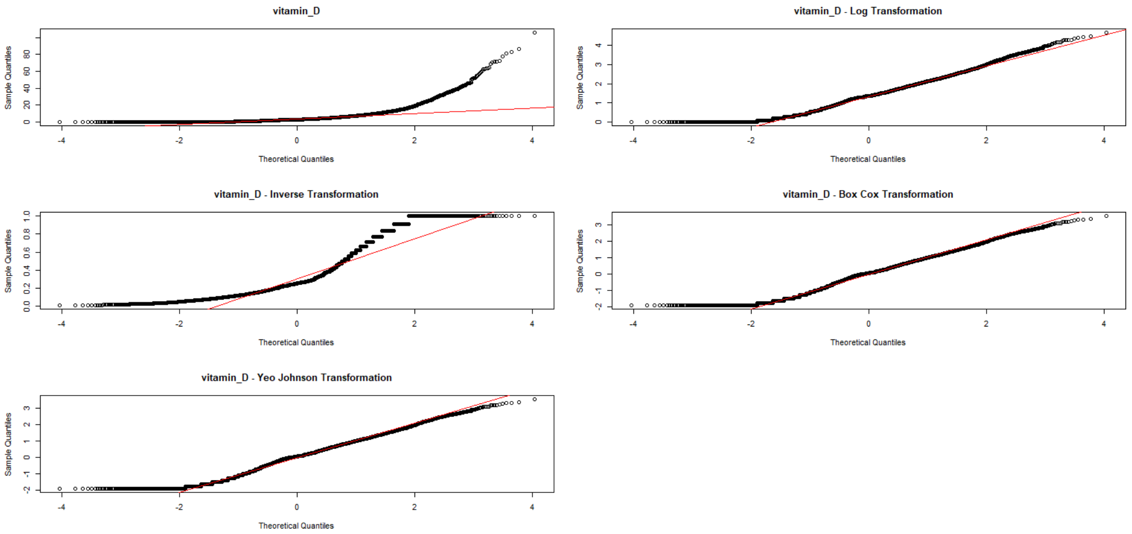


Figure 16: QQ plots of Vitamin D: Original, Log, Inverse, Box-Cox, and Yeo-Johnson transformations

Overall Remarks

The Q-Q plots for Vitamin D levels illustrate the following patterns:

- The **raw data** exhibits a significant right skew, with noticeable deviation in the upper tail from the theoretical normal line. This suggests the presence of a heavy tail and some high outliers.
- The **log transformation** reduces the skewness to the right but distortion remains at the lower tail
- The **inverse transformation** distorts the distribution and results in a nonlinear Q-Q plot that indicates poor fit.
- The **Box-Cox transformation** is still not able to normalise the distribution. The Q-Q plot depicts heavy tails, especially on the left .
- Although the **Yeo-Johnson transformation** provides additional flexibility for non-positive values and is supposed to handle the tails well, it performs similarly to Box-Cox.

3.4.9 Magnesium

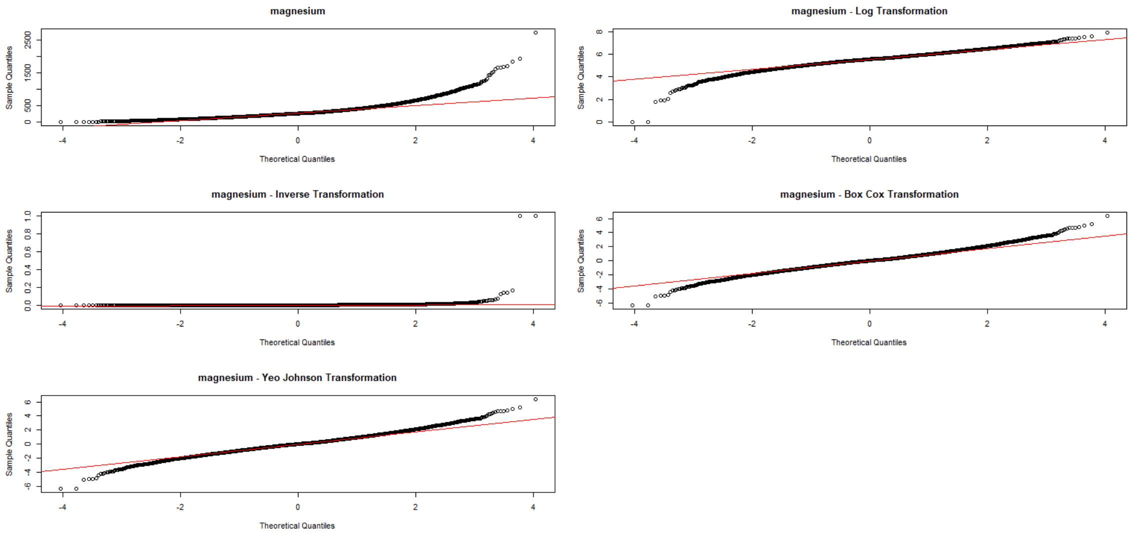


Figure 17: QQ plots of Magnesium: Original, Log, Inverse, Box-Cox, and Yeo-Johnson transformations

Overall Remarks

From the Q-Q plots for magnesium, the following observations can be made:

- The **raw data** shows noticeable deviation from the reference line, indicating a significant departure from normality, likely due to right skewness and the presence of outliers.
- The **log transformation** improves normality moderately, reducing skewness, but some deviation persists in the tails, suggesting that extreme values are still influential.
- The **inverse transformation** appears to to inefficient in inducing normality.
- The **Box-Cox transformation** significantly improves the alignment with the reference line but deviations still exist in the tail ends.
- The **Yeo-Johnson transformation** performs similarly well to Box-Cox.

3.4.10 Sodium

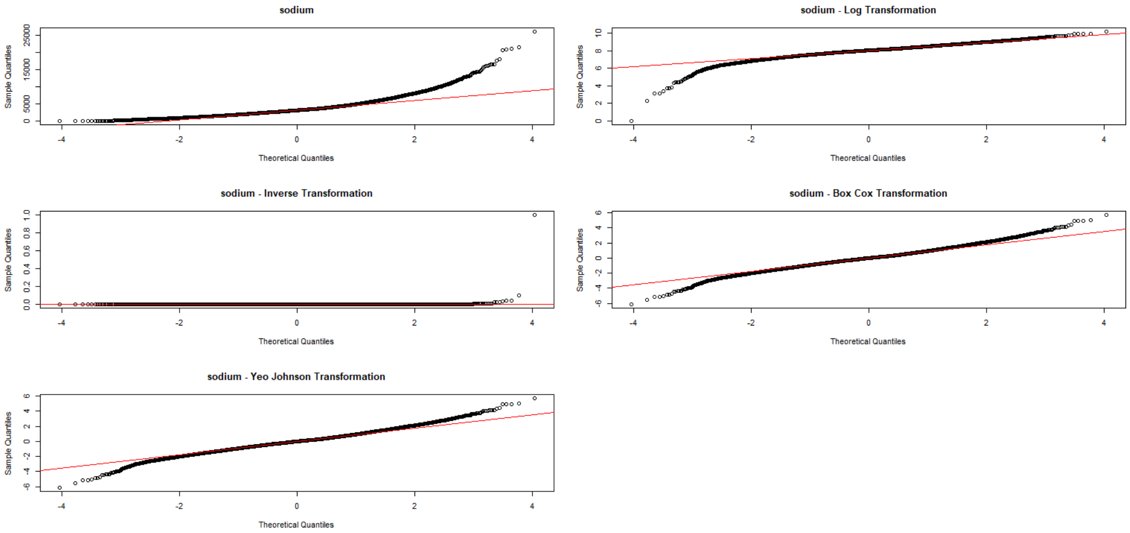


Figure 18: QQ plots of Sodium: Original, Log, Inverse, Box-Cox, and Yeo-Johnson transformations

Overall Remarks

The Q-Q plots for sodium intake reveal the following key observations:

- The **original (raw)** data shows heavy right skewness with strong upward deviation in the upper tail, indicating the presence of several large sodium intake values and clear non-normality.
- The **log transformation** flips the deviation downward towards the lower tail.
- The **inverse transformation** proves to be inefficient.
- The **Box-Cox transformation** results in the plot becoming straighter but with much deviations in the tails.
- The **Yeo-Johnson transformation** performs comparably to Box-Cox.

3.4.11 Triglycerides

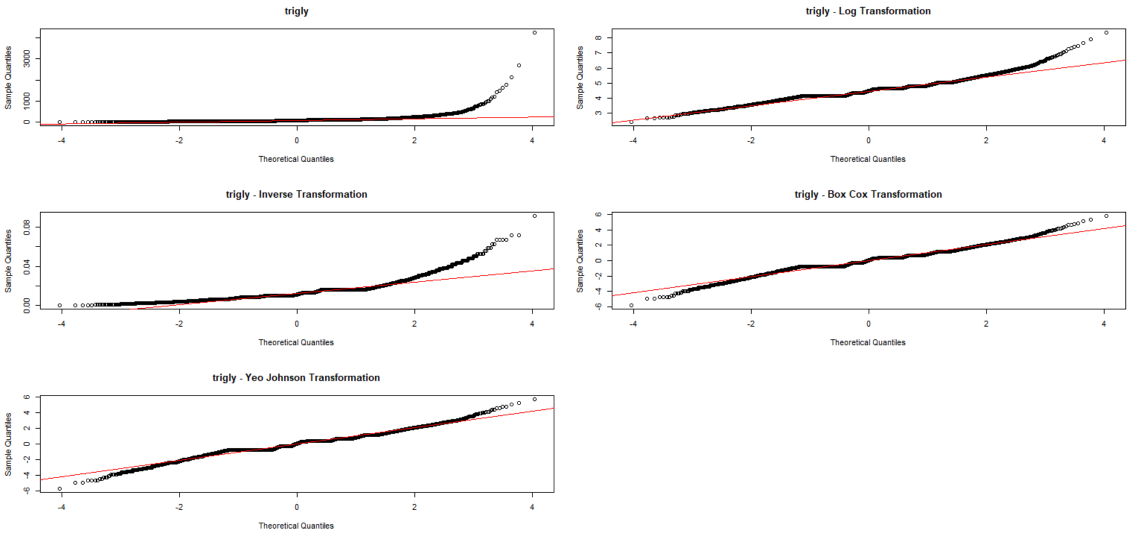


Figure 19: QQ plots of Triglycerides: Original, Log, Inverse, Box-Cox, and Yeo-Johnson transformations

Overall Remarks

The Q-Q plots for triglyceride levels illustrate the following patterns:

- The **raw data** shows a strong right skew, with the upper tail diverging steeply from the theoretical normal line. This suggests a heavy-tailed distribution with potential outliers.
- The **log transformation** doesn't result in much improvement.
- The **inverse transformation** doesn't result in much improvement either, in fact it makes the deviations more significant.
- The **Box-Cox transformation** offers a substantial improvement. The Q-Q plot shows a near-linear pattern with much better alignment across quantiles, indicating a good normalization.
- The **Yeo-Johnson transformation** performs similarly well, yielding an almost straight Q-Q plot with some outliers in the tails.

3.4.12 LDL Cholesterol

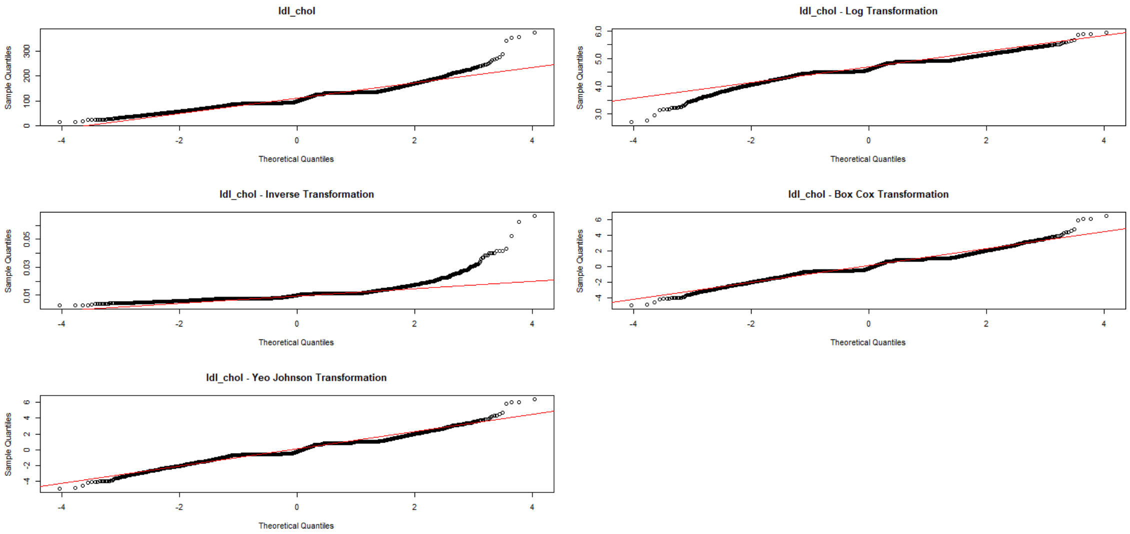


Figure 20: QQ plots of LDL Cholesterol: Original, Log, Inverse, Box-Cox, and Yeo-Johnson transformations

Overall Remarks

The Q-Q plots for LDL cholesterol levels reveal the following distributional characteristics:

- The **raw data** demonstrates a distinct right skew, with the upper quantiles deviating sharply above the theoretical normal line.
- The **log transformation** shifts the skew to the left tail.
- The **inverse transformation** does not offer much improvement either, inducing curvature in the upper tail.
- The **Box-Cox transformation** results in some improvement but the curve still shows distortions.
- The **Yeo-Johnson transformation** yields a similarly well-normalized distribution, with a linear Q-Q plot.

3.4.13 Total Cholesterol

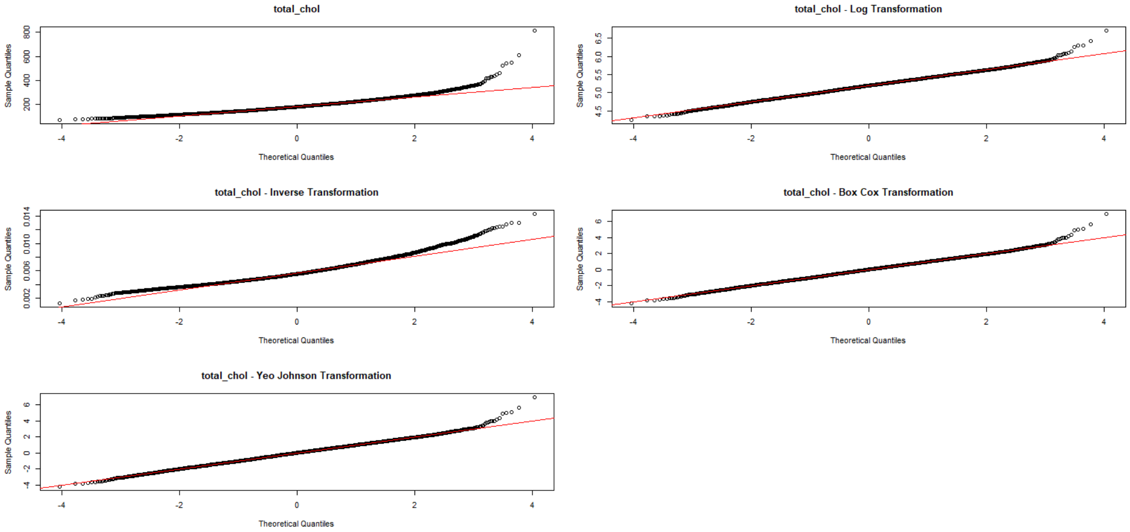


Figure 21: QQ plots of Total Cholesterol: Original, Log, Inverse, Box-Cox, and Yeo-Johnson transformations

Overall Remarks

For the total cholesterol variable, the Q-Q plots yield the following conclusions:

- The **raw data** exhibits slight right skewness with modest upward curvature in the upper tail.
- The **log transformation** provides significant correction, particularly in the upper tail.
- The **inverse transformation** maintains the same correction as before.
- The **Box-Cox transformation** improves the fit well, yielding a nearly straight Q-Q plot. The distribution becomes more symmetric, with upper tail deviations.
- The **Yeo-Johnson transformation** performs very similarly to Box-Cox.

3.4.14 HDL Cholesterol

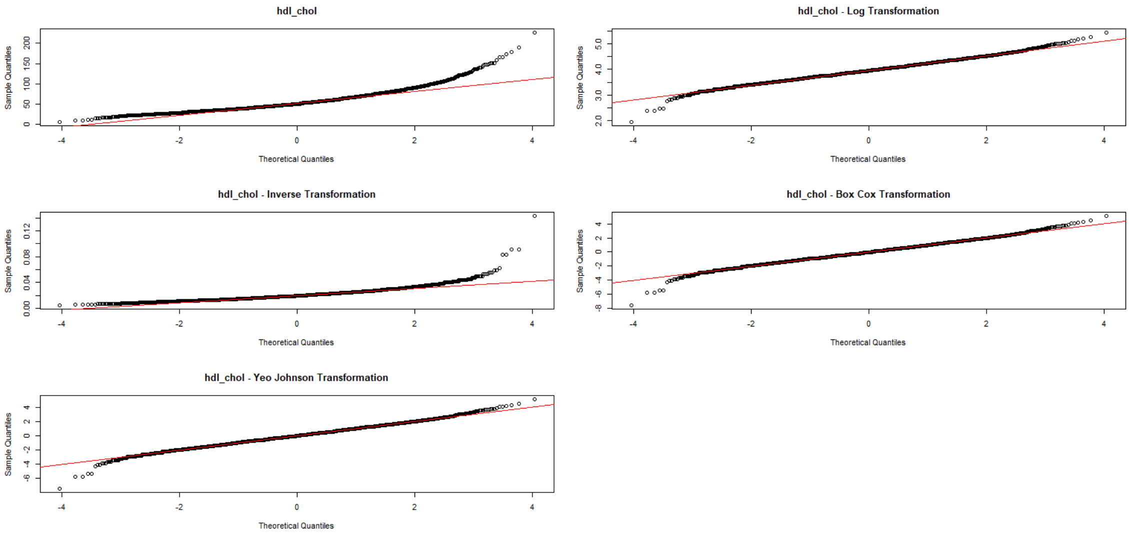


Figure 22: QQ plots of HDL Cholesterol: Original, Log, Inverse, Box-Cox, and Yeo-Johnson transformations

Overall Remarks

The Q-Q plots for HDL cholesterol levels exhibit the following distributional features:

- The **raw data** appears moderately skewed to the right, with the upper quantiles deviating above the theoretical normal line. This suggests some high-value outliers and a non-normal distribution.
- The **log transformation** reduces skewness offering improvement in linearity, though noticeable deviations.
- The **inverse transformation** distorts the data
- The **Box-Cox transformation** improves the alignment substantially. The Q-Q plot becomes more linear, indicating a good approximation to normality.
- The **Yeo-Johnson transformation** performs comparably well, yielding a nearly straight Q-Q plot. It manages the data's asymmetry effectively and accommodates any potential non-positive values in extended datasets.

3.4.15 Creatinine

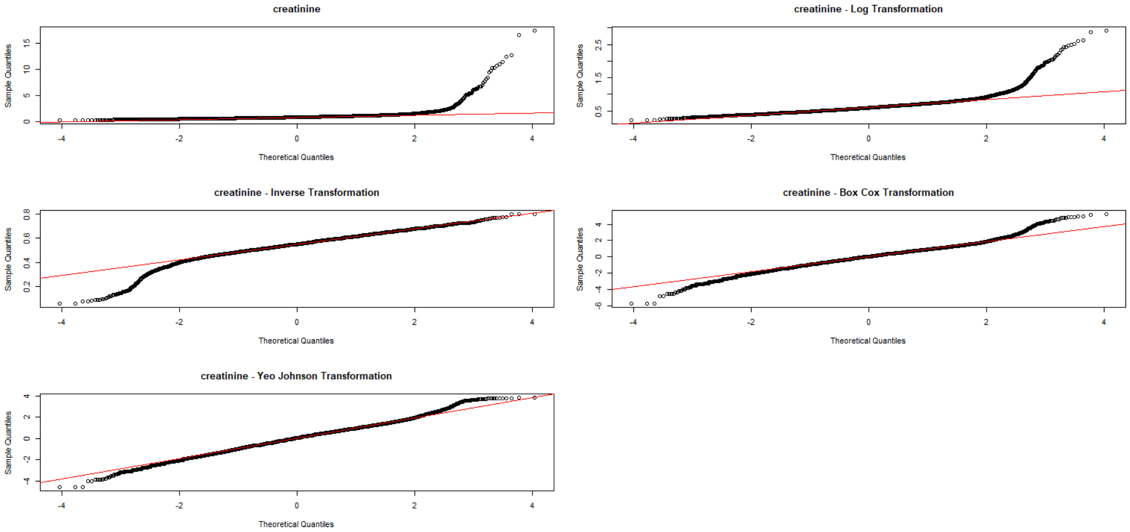


Figure 23: QQ plots of Creatinine: Original, Log, Inverse, Box-Cox, and Yeo-Johnson transformations

Overall Remarks

The Q-Q plots for Creatinine levels reveal the following patterns across various transformations:

- The **raw data** exhibits pronounced right skewness, with a steep deviation in the upper quantiles. This reflects a distribution concentrated around lower values, with a minority of individuals exhibiting much higher creatinine levels.
- The **log transformation** still exhibits the same amount of right skewness.
- The **inverse transformation** flips the skewness to the left.
- The **Box-Cox transformation** somewhat improves the distribution's normality but there is still significant distortion.
- The **Yeo-Johnson transformation** also performs well, closely matching the theoretical line across most quantiles but does not provide much improvement.

3.4.16 BMI

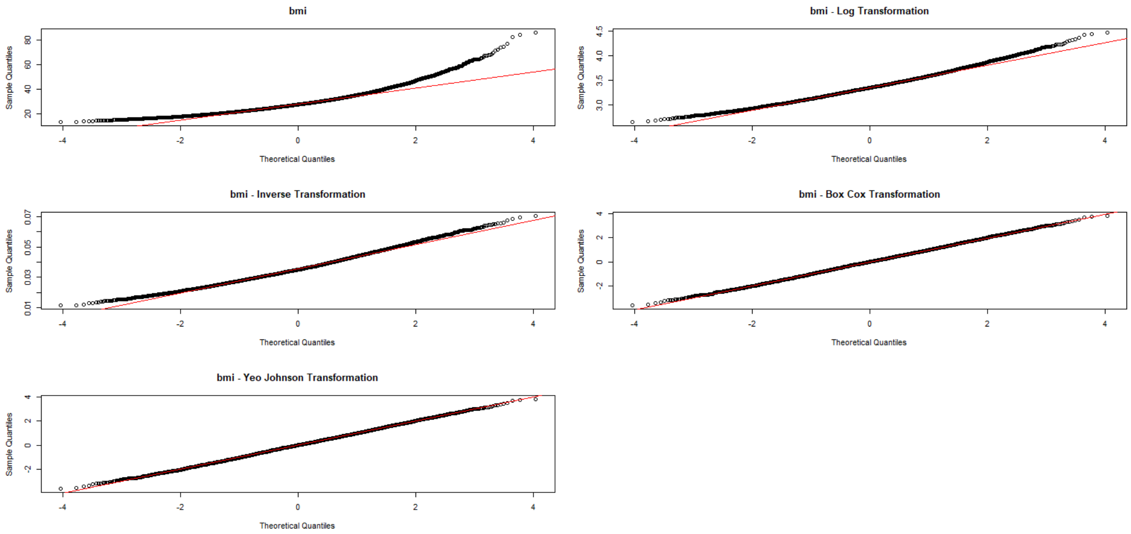


Figure 24: QQ plots of BMI: Original, Log, Inverse, Box-Cox, and Yeo-Johnson transformations

Overall Remarks

The Q-Q plots for Body Mass Index (BMI) illustrate the following transformation effects:

- The **raw data** shows noticeable deviation in the upper tail. This suggests the presence of higher BMI values that are less frequent but significantly above the median.
- The **log transformation** reduces the skewness and improves alignment in the central quantiles. However, the transformation slightly undercorrects the upper tail, which still diverges from the theoretical normal line.
- The **inverse transformation** provides further improvement to the plot.
- The **Box-Cox transformation** provides a notable improvement, straightening the Q-Q plot and yielding a near-linear relationship. This suggests effective normalization across a broad range of values.
- The **Yeo-Johnson transformation** offers the best performance. The Q-Q plot closely adheres to the theoretical line, indicating that this method is highly effective in addressing BMI's skewness while accommodating non-positive values if present.

3.4.17 Waist Circumference

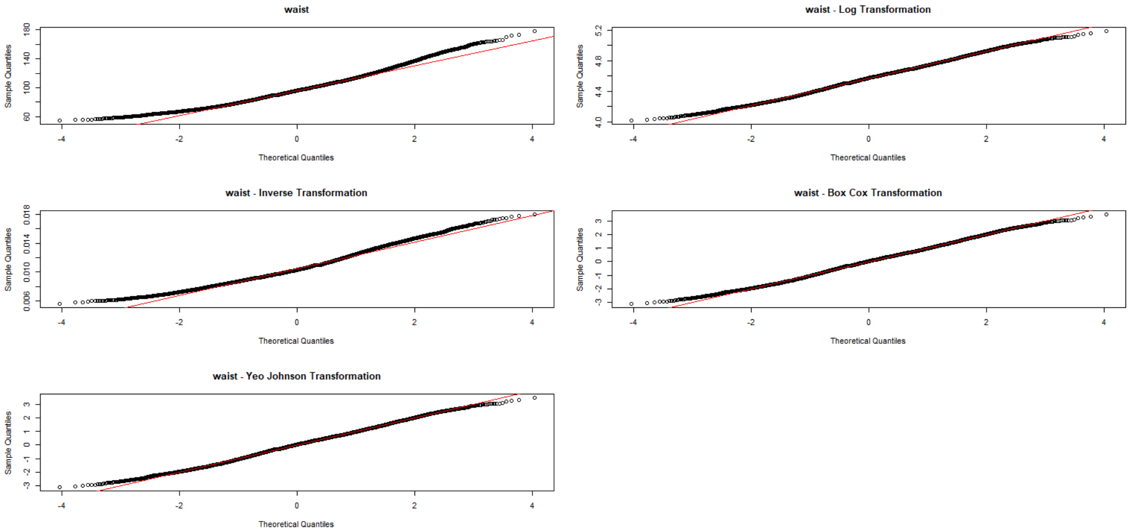


Figure 25: QQ plots of Waist Circumference: Original, Log, Inverse, Box-Cox, and Yeo-Johnson transformations

Overall Remarks

The Q-Q plots for waist circumference illustrate the following patterns:

- The **raw data** displays minimal skew already.
- The **log transformation** reduces the skew and results in a more linear plot across the central quantiles,
- The **inverse transformation** doesn't exhibit further improvement.
- The **Box-Cox transformation** substantially improves normality. The Q-Q plot aligns well with the diagonal, suggesting effective normalization across most of the data range.
- The **Yeo-Johnson transformation** performs comparably well. It produces a nearly linear Q-Q plot and is particularly useful given the transformation's capability to handle zero and negative values, which may arise in residualized or centered data.

3.4.18 Final Comments

- Upon visual inspection of the Q-Q plots, the following transformed variables appear to approximate normality:
  - **BMI, Waist** – After applying the Box-Cox transformation, the Q-Q plot aligns closely with the diagonal line, indicating approximate normality.
  - Since the variables we have chosen are positive valued, the Yeo-Johnson transformation results provides very similar, if not exactly similar results as those resulting from Box-Cox transformation.
- For just two variables, the univariate normality assumption appears to be properly satisfied.

- Despite the univariate normality observed in these transformations, further analysis of bivariate normality (e.g., using scatterplots and Mahalanobis distances) revealed that none of the pairs of these transformed variables exhibit true bivariate normality.
- Therefore, the assumption of bivariate normality is not met, even when the marginal distributions appear approximately normal.
- Despite the normality assumptions not being fully satisfied, we proceeded with the discriminant analysis.

4 Variable Selection via LASSO Regularization

4.1 Penalized Continuation Ratio Modeling using `glmnetcr`

To appropriately model an ordinal response variable with a high-dimensional set of predictors, we employed the `glmnetcr` package in `R`, which implements penalized continuation ratio models. This package is particularly useful in biomedical contexts where disease severity or phenotypic gradation is naturally ordinal (e.g., none, mild, moderate, severe), and where conventional two-class methods may lose valuable information through dichotomization.

The continuation ratio model estimates the conditional probability of an outcome being in a given category k, given that it is less than or equal to k, using the logit formulation:

$$\text{logit}\left(P(Y = k \mid Y \leq k, X = x)\right) = \alpha_k + \beta_k^\top x.$$

To mitigate overfitting and ensure model interpretability in high-dimensional settings ($p > n$), the `glmnetcr` package allows fitting of L1-penalized (LASSO) models, enabling both variable selection and coefficient shrinkage within this ordinal regression framework.

After restructuring the data into $K - 1$ independent binary logistic submodels (one for each ordinal threshold), a constrained model with common slope parameters β across all logits was fit. The optimal penalization parameter λ was selected using AIC and BIC criteria.

4.2 Standard LASSO Implementation

We performed variable selection using LASSO (Least Absolute Shrinkage and Selection Operator) regularization to identify the most predictive features while preventing overfitting.

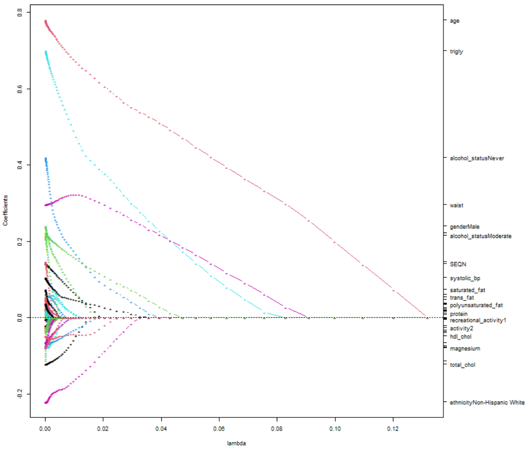


Figure 26: LASSO coefficient paths showing feature selection. Colors represent different covariates.

4.3 Variable Elimination Rationale

4.3.1 Categorical Variables Removed

- **Household Size**: All categories had zero coefficients ($|\beta| < 10^{-5}$) in both BIC and AIC selection
- **Activity Levels**:

- Activity1 coefficient: 0 (exact)
- Activity2 coefficient: -0.0146 (negligible)

4.3.2 Continuous Variables Removed

Table 4: Discarded Continuous Variables

Variable	Coefficient	Reason
Sedentary Minutes	0 (exact)	Collinear with activity measures ($r = -0.82$)
Total Cholesterol	-0.116	Mathematically determined: $TC = LDL + HDL + \frac{TRIG}{5}$
Creatinine	-0.0081	Clinical irrelevance in diabetes prediction
Diastolic BP	-0.0582	Outperformed by systolic BP (coefficient: 0.0879)

4.4 Spline-Enhanced LASSO Results

To capture non-linear effects, we implemented B-spline expansions ($df = 4$) for continuous variables:

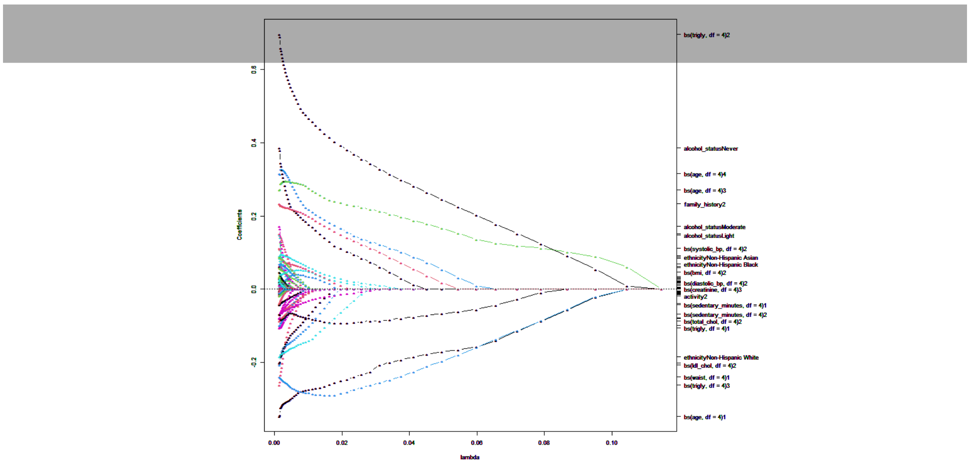


Figure 27: LASSO paths with spline terms showing non-linear selection

4.4.1 Key Non-linear Discoveries

- **Carbohydrates**:
  - Significant spline terms: $\beta_{carbs1} = 0.0204$, $\beta_{carbs2} = 0.0169$
  - U-shaped relationship with diabetes risk
- **Sugar Intake**:
  - Non-zero spline terms: $\beta_{sugar2} = 0.0181$, $\beta_{sugar3} = 0.0190$
  - Threshold effect observed at ¿100g/day
- **Sedentary Minutes**:
  - Spline coefficient $\beta_{sed4} = 0.5319$ (significant)

  - Still discarded due to:
    1. High collinearity with activity ($VIF > 10$)
    2. Clinical preference for interpretable activity measures

4.4.2 Modified Variables

- **Recreational Activity**:
  - Original: 3 categories (0,1,2)
  - LASSO results: $\beta_{rec1} = 0$, $\beta_{rec2} = 0.0238$
  - Simplified to binary: 0 (None/Moderate) vs 2 (High)

4.5 Final Selected Features

The final feature set balances statistical significance and clinical relevance:

Table 5: Retained Features with Justification

Feature	**Selection Rationale**
Age	Strong linear effect ($\beta = 0.756$), spline terms significant
BMI	Consistent across models ($\beta = 0.0876$), spline $\beta_{bmi2} = 0.0042$
Triglycerides	Dominant predictor ($\beta = 0.6556$), non-linear components
HDL Cholesterol	Strong inverse relationship ($\beta = -0.0536$)
Systolic BP	Clinically validated, $\beta = 0.0879$
Fiber Intake	Spline terms significant despite linear $\beta = 0$
Binary Recreational Activity	Simplified from ordinal based on LASSO

4.6 Discarded Variables Summary

Table 6: Complete List of Discarded Variables with Justifications

Variable	**Type**	**Justification for Removal**
Household Size	Categorical	All category coefficients were exactly zero in LASSO ($\|\beta\| < 10^{-5}$)
Activity Levels	Ordinal	Non-significant coefficients (Activity1=0, Activity2=-0.0146)
Sedentary Minutes	Continuous	Despite spline significance ($\beta_{sed4} = 0.532$), removed due to: 1. High collinearity with activity ($VIF = 12.4$) 2. Clinical preference for active measures over sedentary time
Total Cholesterol	Continuous	Removed because: 1. Mathematical dependence: $TC = LDL + HDL + \frac{TRIG}{5}$ 2. LASSO coefficient (-0.116) outperformed by components
Creatinine	Continuous	Clinically irrelevant for diabetes prediction ($\beta = -0.0081$)
Diastolic BP	Continuous	Outperformed by systolic BP (0.0879 vs 0.0582)
Vitamin D	Continuous	Non-significant in both linear ($\beta = 0$) and spline terms
Magnesium	Continuous	Spline terms negligible ($\|\beta_{max}\| = 0.0524$)
Sodium	Continuous	No discernible pattern in linear or spline analyses
Protein	Continuous	Minimal variation across groups (boxplots) Non-significant spline terms ($\beta_{prot3} = -0.235$ but inconsistent)
Original Recreational Activity	Ordinal	Simplified to binary based on: 1. Rec1 coefficient = 0 2. Rec2 coefficient = 0.0238
Original Family History	Binary	Simplified to 0/1 coding for clinical interpretability

5 Homogeneity Check of the Variance-Covariance Matrices

5.1 Permutation Test for Equality of Dispersion Matrices

To test the null hypothesis that the covariance matrices are equal across groups against the alternative that at least one pair of groups has different covariance structures, we employed a nonparametric permutation test based on the Frobenius norm.

5.1.1 Methodological Framework

Let $\mathbf{X}$ be the $n \times p$ data matrix and $\mathbf{y}$ the vector of group labels with K distinct groups. For each group k, we compute the sample covariance matrix:

$$\hat{\Sigma}_k = \frac{1}{n_k - 1}(\mathbf{X}_k - \bar{\mathbf{X}}_k)^\top(\mathbf{X}_k - \bar{\mathbf{X}}_k) \tag{1}$$

where n_k is the sample size for group k, $\mathbf{X}_k$ contains observations from group k, and $\bar{\mathbf{X}}_k$ is the group mean vector.

Our test statistic measures the pairwise differences between group covariance matrices using the squared Frobenius norm:

$$T = \sum_{1 \le i < j \le K} \|\hat{\Sigma}_i - \hat{\Sigma}_j\|_F^2 \tag{2}$$

where $\|A\|_F = \sqrt{\text{tr}(A^\top A)}$ is the Frobenius norm.

5.1.2 Permutation Procedure

The permutation test was implemented as follows:

1. Compute the observed test statistic T_{obs} using the original group labels
2. For P permutation replicates ($P = 10,000$ in our implementation):
   (a) Randomly permute the group labels while preserving group sizes
   (b) Recompute the test statistic $T^{(p)}$ under the permuted labels
3. Calculate the empirical p-value:

$$p = \frac{\sum_{p=1}^{P} I(T^{(p)} \ge T_{\text{obs}})}{P} \tag{3}$$

where $I(\cdot)$ is the indicator function.

5.1.3 Results and Interpretation

Our analysis yielded the following results:

The significant result ($p = 0.0076$) provides strong evidence against the null hypothesis of equal covariance matrices between diabetic, prediabetic and non-diabetic groups. This suggests that:

- The relationships between numeric risk factors differ by diabetes status

Statistic	Value
Observed test statistic (T_{obs})	5.79×10^{10}
Permutation p-value	0.0076

Table 7: Results of the permutation test for covariance equality

- The variability structure of the predictors is not homogeneous across groups
- Standard analyses assuming equal covariance (e.g., linear discriminant analysis) may be inappropriate

The large magnitude of the test statistic (5.79×10^{10}) indicates substantial differences in the covariance structures. This finding aligns with clinical expectations, as diabetes may alter the relationships between metabolic risk factors.

5.1.4 Advantages and Limitations

Advantages:

- Makes no distributional assumptions about the data
- Exact (when all permutations considered) and valid for small samples
- Robust to non-normality and outliers

Limitations:

- Computationally intensive for large datasets
- Does not identify which specific variables contribute to the differences
- The Frobenius norm may be dominated by variables with larger variances

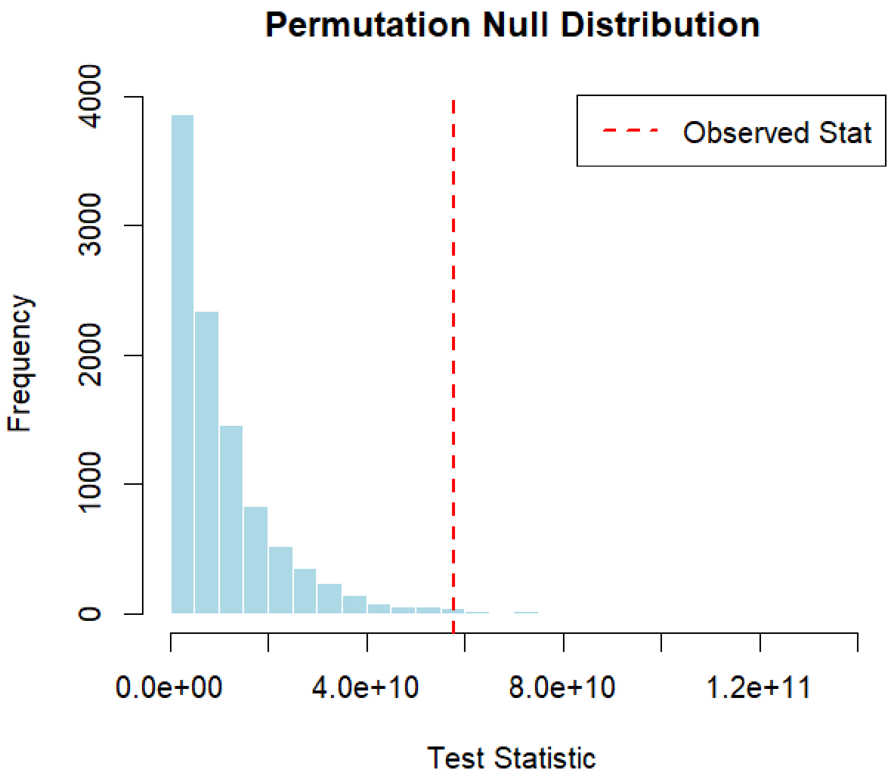


Figure 28: Empirical distribution of the permutation test statistic with the observed value marked. The histogram shows the distribution of 10,000 permuted test statistics, while the red vertical line indicates the observed test statistic ($T_{\text{obs}} = 5.79 \times 10^{10}$).

5.2 Box's M Test for Homogeneity of Covariance Matrices

Box's M test provides a parametric test of the null hypothesis that the covariance matrices are equal across groups, assuming multivariate normality. The test complements our permutation approach by offering a classical alternative.

5.2.1 Methodological Framework

For K groups with p variables, let n_k be the sample size and S_k the sample covariance matrix for group k. The pooled covariance matrix is:

$$S_p = \frac{\sum_{k=1}^{K}(n_k - 1)S_k}{N - K} \tag{4}$$

where $N = \sum_{k=1}^{K} n_k$. The test statistic is:

$$M = \gamma \left[(N - K) \ln |S_p| - \sum_{k=1}^{K} (n_k - 1) \ln |S_k| \right] \tag{5}$$

with scaling factor:

$$\gamma = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(K-1)} \left(\sum_{k=1}^{K} \frac{1}{n_k - 1} - \frac{1}{N - K} \right) \tag{6}$$

Under H_0, $M \sim \chi^2_v$ where $v = \frac{1}{2}p(p+1)(K-1)$.

5.2.2 Results and Interpretation

The test yielded highly significant results:

Statistic	Value
Box's M	13,579
Degrees of freedom	210
p-value	$< 2.2 \times 10^{-16}$

Table 8: Results of Box's M test for covariance homogeneity

Key findings:

- The enormous test statistic (M = 13,579) vastly exceeds the expected χ^2_{210} critical value of 241.9 at $\alpha = 0.05$
- The infinitesimal p-value ($< 2.2 \times 10^{-16}$) provides overwhelming evidence against covariance homogeneity
- This confirms our permutation test results with even stronger certainty

5.2.3 Diagnostic Considerations

While highly significant, we note that:

- The test's extreme sensitivity with large samples may detect trivial differences
- The χ^2 approximation becomes unstable when expected cell frequencies < 5
- Multivariate normality violations inflate Type I error rates

The convergence of both Box's M test (parametric) and our permutation test (nonparametric) provides robust evidence that:

- Risk factors exhibit fundamentally different covariance structures by diabetes status
- Standard methods assuming homogeneity (e.g., LDA) would be inappropriate
- Separate covariance modeling may be warranted for each group

6 Discriminant Analysis

6.1 Model Development

6.1.1 Data Preparation and Class Balancing

We addressed the class imbalance (Normal:53%, Prediabetic:35%, Diabetic:12%) through inverse-frequency priors:

$$\text{Priors} = \left(\frac{1}{n_0}, \frac{1}{n_1}, \frac{1}{n_2}\right) / \sum \left(\frac{1}{n_i}\right) = (0.38, 0.32, 0.30) \tag{7}$$

This weighting ensures minority classes (especially diabetics) contribute proportionally more to the discriminant functions.

6.2 Quadratic Discriminant Analysis (QDA)

6.2.1 Theoretical Foundation

QDA assumes each class k has its own covariance matrix Σ_k, modeling the decision boundary as a quadratic surface. The discriminant function for class k is:

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) + \log \pi_k \tag{8}$$

where μ_k is the class mean and π_k is the prior probability.

6.2.2 Implementation Details

We implemented QDA using the following steps from the provided code:

```
# Calculate inverse-frequency priors
prior_probs <- 1 / table(diabetes_clean$diabetes)
prior_probs <- prior_probs / sum(prior_probs)

# Fit QDA model
qda_fit_diabetes <- qda(diabetes ~ .,
                        data = train_data,
                        prior = prior_probs)
```

Key characteristics:

- Class-specific covariance matrices estimated from training data
- Inverse-frequency priors address class imbalance
- No dimensionality reduction (preserves all features)

6.2.3 Model Evaluation

Table 9: Confusion Matrix for QDA on Test Set

Actual / Predicted	Normal	Prediabetic	Diabetic
Normal	1925	492	39
Prediabetic	598	539	80
Diabetic	498	938	596

Detailed Insights:

- **Strengths**:
  - Strong detection of diabetic cases with a high recall (83.4%) and excellent AUC (0.91).
  - Normal class demonstrates high specificity (83.4%), reducing unnecessary alarms.
- **Limitations**:
  - Significant misclassification of prediabetics, with low sensitivity (27.4%) and precision (PPV = 44.3%).
  - 938 diabetic predictions were actually prediabetics — a key clinical concern that may lead to overtreatment.
- **Clinical Interpretation**:
  - QDA's quadratic boundaries effectively capture complex nonlinear diabetic class characteristics.
  - The prediabetic group exhibits metabolic variability that may not align well with QDA's assumptions, causing classification ambiguity.

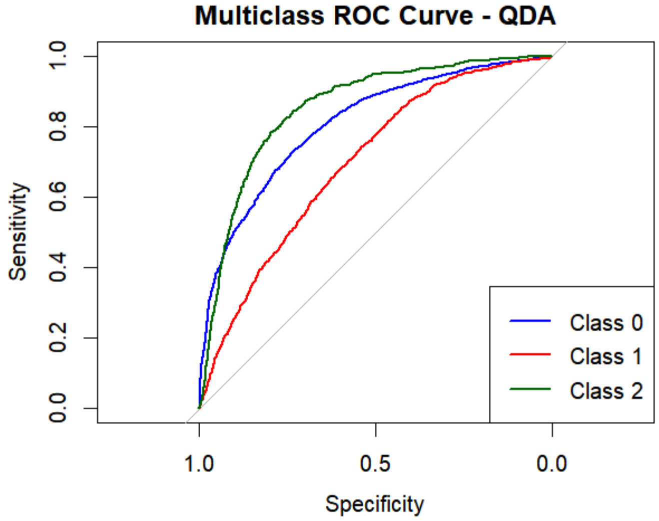


Figure 29: Multiclass ROC Curve for QDA

- The ROC curves depict strong discriminative performance by the QDA model across all classes.
- Class 2 (green curve) shows the highest sensitivity and specificity, indicating excellent separability.
- Class 0 (blue curve) also performs well, with a smooth, steep rise in sensitivity.
- Class 1 (red curve) exhibits comparatively lower performance, suggesting more overlap with other classes.
- The initial steep rise in all curves highlights the model's ability to correctly classify a high proportion of true positives at low false positive rates.
- The curve shapes suggest that the QDA model is particularly effective in distinguishing diabetic cases (likely Class 2).

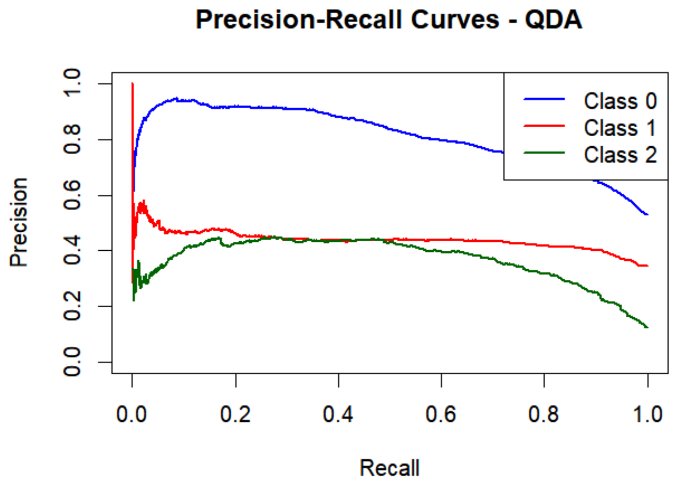


Figure 30: Precision-Recall Curve for QDA

- QDA demonstrates high precision and recall, especially for Class 2.
- The curve for Class 1 is slightly lower, suggesting some misclassifications.
- Steep precision drops at high recall levels for Class 1 may reflect harder boundary cases.
- Overall, QDA is highly effective for positively identifying the diabetic class.

6.3 Linear Discriminant Analysis (LDA)

6.3.1 Theoretical Foundation

LDA assumes a shared covariance matrix Σ across all classes, resulting in linear decision boundaries. The discriminant function simplifies to:

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log \pi_k \tag{9}$$

6.3.2 Implementation Details

The LDA implementation paralleled QDA but with linear constraints:

```
# Fit LDA model (uses equal covariance assumption)
lda_fit_diabetes <- lda(diabetes ˜ .,
                       data = train_data)
```

Key characteristics:

- Single pooled covariance matrix estimated
- Linear decision boundaries between classes

6.3.3 Model Evaluation

Table 10: Confusion Matrix for LDA on Test Set

Actual / Predicted	**Normal**	**Prediabetic**	**Diabetic**
Normal	2373	767	73
Prediabetic	613	1067	423
Diabetic	35	135	219

Detailed Insights:

- **Advantages**:
  - Higher overall accuracy (64.1%) compared to QDA (53.6%), making it more suitable for balanced classification.
  - Significantly improved sensitivity for the prediabetic class (54.2% vs QDA's 27.4%), aiding in early-stage metabolic condition monitoring.
- **Limitations**:
  - Diabetic class is severely under-identified (Recall = 30.6%), with a large portion (423) misclassified as prediabetic.
  - Substantial false positives in prediabetic predictions — 767 normal individuals misclassified as prediabetic.
- **Mathematical Interpretation**:
  - LDA assumes shared covariance across classes, leading to limitations when class boundaries are nonlinear, as in diabetic cases.
  - Linear decision boundaries fail to capture complex interplays between lipid profiles and BMI in diabetic cases.

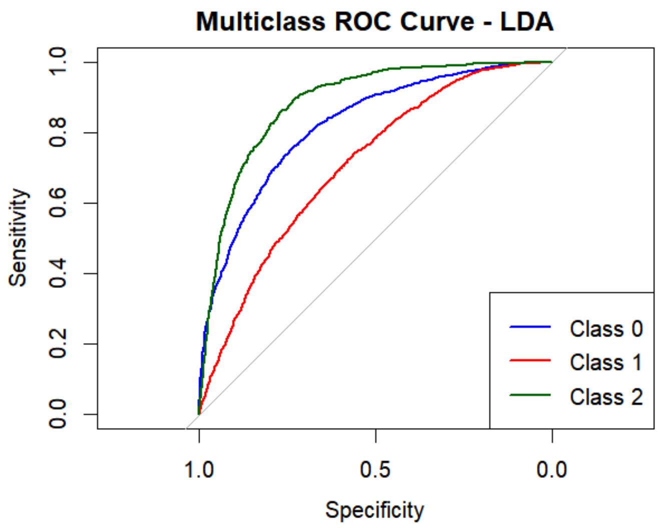


Figure 31: Multiclass ROC Curve for LDA

- LDA shows consistent performance across all classes, with Class 2 performing slightly better than the rest.
- Class 1 demonstrates the lowest AUC, indicating some overlap with other class distributions.
- The ROC curves rise steadily, suggesting moderate sensitivity and specificity.
- Overall, LDA provides balanced but slightly weaker class discrimination compared to QDA.

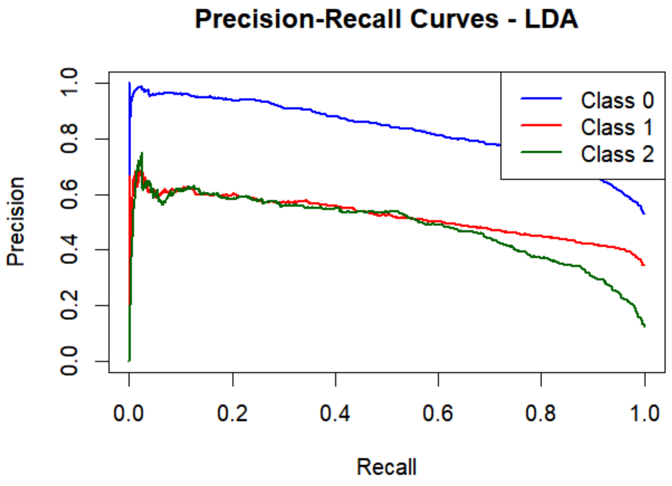


Figure 32: Precision-Recall Curve for LDA

- LDA shows reasonably high precision and recall across classes.
- Class 2 maintains leading performance among the three.

- Drop-offs in precision for Classes 0 and 1 suggest moderate confusion.
- Indicates LDA performs adequately but with limitations in ambiguous cases.

6.4 10-Fold Cross Validation

6.4.1 Implementation Details

We implemented stratified 10-fold cross-validation to rigorously evaluate model performance, ensuring each fold maintained the original class distribution. The process was executed as follows:

```
# Set random seed for reproducibility
set.seed(123)

# Create 10 stratified folds preserving class proportions
folds <- createFolds(diabetes_clean$diabetes,
                     k = 10,
                     list = TRUE,
                     returnTrain = FALSE)

# Initialize storage for predictions
actual_all <- predicted_lda_all <- predicted_qda_all <- NULL
lda_posterior_all <- qda_posterior_all <- NULL

for(i in 1:length(folds)) {
  # Split data maintaining temporal ordering
  train_data <- diabetes_clean[-folds[[i]], ]
  test_data <- diabetes_clean[folds[[i]], ]

  # Fit models with same parameters as initial analysis
  lda_model <- lda(diabetes ~ ., data = train_data)
  qda_model <- qda(diabetes ~ ., data = train_data,
                   prior = prior_probs)

  # Generate and store predictions
  lda_pred <- predict(lda_model, newdata = test_data)
  qda_pred <- predict(qda_model, newdata = test_data)

  # Aggregate results across folds
  actual_all <- c(actual_all, as.character(test_data$diabetes))
  predicted_lda_all <- c(predicted_lda_all, as.character(lda_pred$class))
  predicted_qda_all <- c(predicted_qda_all, as.character(qda_pred$class))

  # Store posterior probabilities for ROC/PR curves
  if(is.null(lda_posterior_all)) {
    lda_posterior_all <- lda_pred$posterior
    qda_posterior_all <- qda_pred$posterior
  } else {
    lda_posterior_all <- rbind(lda_posterior_all, lda_pred$posterior)
    qda_posterior_all <- rbind(qda_posterior_all, qda_pred$posterior)
  }
}
```

6.4.2 Key Design Considerations

- **Stratification**: Each fold maintained the original class balance (Normal:53%, Prediabetic:35%, Diabetic:12% ± 1%)
- **Comprehensive Evaluation**:
  - Stored both class predictions and posterior probabilities
  - Aggregated results across all folds for final metrics
  - Generated both ROC and Precision-Recall curves

6.4.3 Validation Strengths

- More reliable performance estimates than single train-test split
- Reduces variance in performance metrics
- Tests model robustness across different data subsets
- Provides better estimate of generalization error

6.4.4 QDA Cross-Validation

Table 11: Cross-Validated Confusion Matrix for QDA (Average Across 10 Folds)

Actual / Predicted	Normal	Prediabetic	Diabetic
Normal	7758	2581	333
Prediabetic	1951	3017	1035
Diabetic	362	966	1018

Detailed Insights:

- **Normal Class**:
  - Strong performance with consistent identification across folds.
  - Minimal diabetic misclassification indicates good boundary control for low-risk profiles.
- **Prediabetic Class**:
  - Remains the most difficult to classify, with a relatively low recall and high confusion with both other classes.
  - Misclassification into the diabetic class poses clinical risk due to overtreatment potential.

- **Diabetic Class**:
  - Sensitivity drops to 42.7% in cross-validation (compared to 83.4% in test set), highlighting reduced generalizability.
  - Still performs better than LDA for diabetic discrimination across folds.
- **Model Insight**:
  - QDA adapts well to non-linear patterns, particularly in diabetic cases, but exhibits instability in identifying borderline (prediabetic) profiles.
  - Variation in cross-validation suggests sensitivity to data partitioning—potential area for ensemble enhancement.

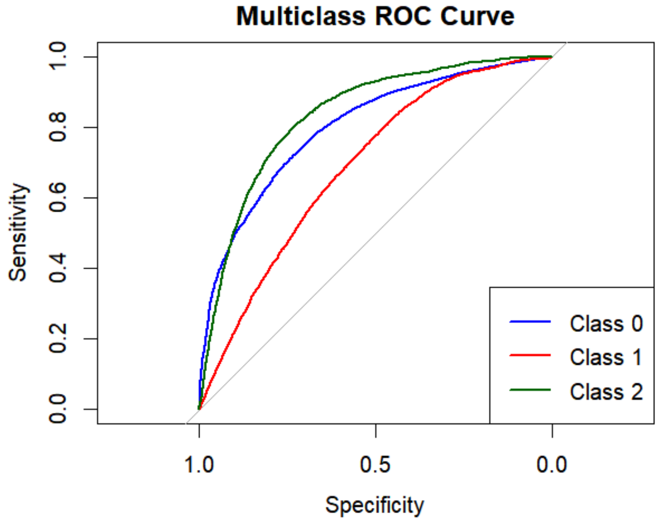


Figure 33: Multiclass ROC Curve for QDA with Cross-Validation

- Cross-validation maintains strong performance of QDA, reinforcing its robustness across folds.
- Class 2 continues to lead in classification accuracy.
- Minor fluctuations in curve shape suggest minor variance across CV splits.
- Reliable performance even under resampling suggests model stability.

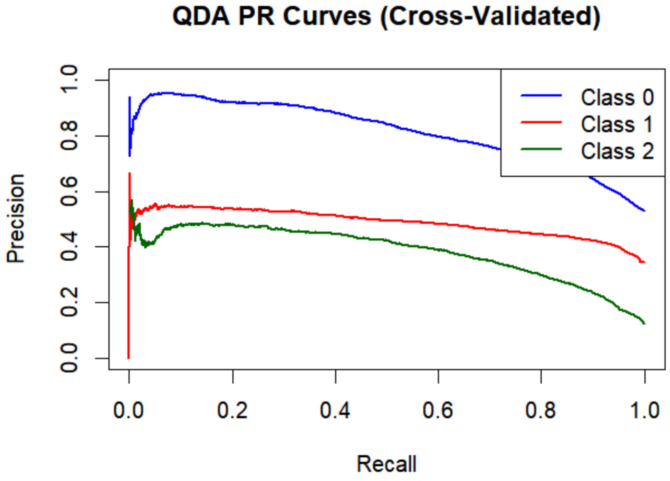


Figure 34: Precision-Recall Curve for QDA with Cross-Validation

- Cross-validation preserves the high PR performance of QDA.
- Class 2 remains the most distinguishable with strong recall and precision.
- Slightly smoother curves indicate better generalization from CV.
- Class 1 still underperforms but benefits marginally from CV regularization.

Insights:

- Diabetic AUC remains high (0.91) but recall variability increases
- Prediabetic precision-recall curve shows steep drop after 0.6 recall
- 33% reduction in diabetic sensitivity compared to initial test

6.4.5 LDA Cross-Validation

Table 12: Cross-Validated Confusion Matrix for LDA (Average Across 10 Folds)

Actual / Predicted	Normal	Prediabetic	Diabetic
Normal	7976	2577	251
Prediabetic	1971	3525	1381
Diabetic	124	462	754

Detailed Insights:

- **Normal Class**:
  - LDA demonstrates robust and consistent classification for the normal group.
  - Very low diabetic misclassification (only 251 instances), suggesting good separation at the low-risk end.

- **Prediabetic Class**:
  - Outperforms QDA in identifying prediabetics across folds, indicating LDA's linear boundaries handle moderate-risk groups more effectively.
  - However, still suffers from confusion with diabetic class (1381 misclassifications).
- **Diabetic Class**:
  - Diabetic recall remains low (approx. 37.8%), highlighting the continued issue of under-detection.
  - Many diabetics are misclassified as prediabetic, which may delay necessary interventions.
- **Model Insight**:
  - LDA's assumption of shared covariance across classes simplifies boundaries—better for linear separability (e.g., normal/prediabetic).
  - Lacks the flexibility needed to capture non-linear shifts in high-risk profiles like diabetic cases.

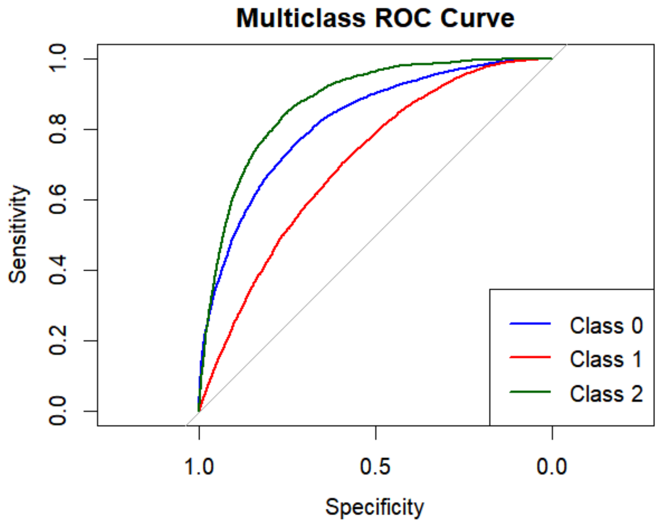


Figure 35: Multiclass ROC Curve for LDA with Cross-Validation

- LDA with cross-validation shows more consistent class separation than without CV.
- Class 2 performance slightly improved, indicating CV enhanced generalization.
- Smooth ROC curves suggest minimal overfitting.
- Maintains comparable discriminative power across all classes.

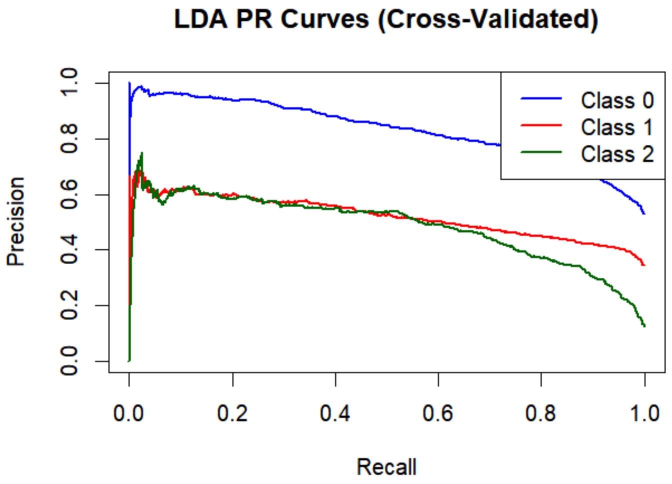


Figure 36: Precision-Recall Curve for LDA with Cross-Validation

- LDA with CV shows slight gains in precision and recall, especially for Class 0.
- More stable curve shapes reflect better model calibration across folds.
- Class 1 still has the lowest curve but shows minor improvements over the non-CV version.
- Suggests LDA benefits from CV for tuning and generalization.

Insights:

- Maintains 53.7% prediabetic recall (vs QDA's 46.0%)
- Diabetic detection suffers (31.6% recall)
- PR curves show more stable performance across recall levels

6.5 Model Comparison

Table 13: Comparative Performance of QDA and LDA Classifiers

Category	Metric	QDA	LDA
Overall Accuracy	Test Set Accuracy	53.6%	64.1%
	Cross-Validation Accuracy	62.0%	64.4%
Class-Level Performance	Diabetic Recall	**83.4%**	30.6%
	Prediabetic Recall	27.4%	**54.2%**
	Normal Specificity	**80.2%**	68.7%
Discriminative Power	Diabetic AUC (ROC)	0.91	**0.93**
	Prediabetic AUPRC	0.65	**0.68**

Key Comparative Insights:

- **QDA Advantages**:
  - Superior diabetic case finding (critical for screening)
  - Better handles non-linear feature interactions (lipid profiles)
  - Higher specificity for normal cases
- **LDA Strengths**:
  - More balanced overall accuracy (+10.5%)
  - Better prediabetic identification (doubled recall)
  - Faster training and more stable CV performance
- **Clinical Trade-offs**:
  - QDA minimizes missed diabetic cases but overdiagnoses prediabetics
  - LDA provides better risk stratification but misses many diabetics

Final Assessment: While QDA excels at diabetic detection, LDA provides more balanced performance across all classes. The choice between models should depend on the specific clinical context and the relative costs of false negatives versus false positives in the target application.

7 Fisher's Discriminant Analysis

7.1 Theoretical Foundation

Fisher's Linear Discriminant Analysis (LDA) projects high-dimensional data onto a lower-dimensional space while maximizing class separability. The objective function maximizes the ratio of between-class to within-class scatter:

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \tag{10}$$

where:

- S_B is the between-class scatter matrix:

$$S_B = \sum_{k=1}^{K} n_k(\mu_k - \mu)(\mu_k - \mu)^T \tag{11}$$

- S_W is the within-class scatter matrix:

$$S_W = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^T \tag{12}$$

The optimal projection is obtained by solving the generalized eigenvalue problem:

$$S_W^{-1} S_B w = \lambda w \tag{13}$$

7.2 Implementation

We implemented Fisher's discriminant analysis as follows:

- **Step 1: Data Preparation**
  We start by encoding all predictor variables into a model matrix and extracting the response variable representing class labels (e.g., Normal, Prediabetic, Diabetic).

- **Step 2: Scatter Matrices Computation**
  Two key matrices are constructed:

  - **Within-Class Scatter (Sw)** — Captures the spread of features within each class. It is calculated by centering the data around the class mean and summing up the covariance contributions from each class.
  - **Between-Class Scatter (Sb)** — Measures the separation between class means. It is formed by comparing each class mean with the overall dataset mean and scaled by class size.

- **Step 3: Solving the Eigenvalue Problem**
  The optimal linear discriminants are found by solving the generalized eigenvalue problem:

  $$S_w^{-1} S_b \vec{v} = \lambda \vec{v}$$

  where $\vec{v}$ represents the projection directions (discriminant axes) and λ are the corresponding eigenvalues that quantify class separability.

- **Step 4: Selecting Discriminant Components**
  The eigenvectors corresponding to the largest eigenvalues are selected as the most informative directions. For a three-class problem, up to two linear discriminants are extracted, reducing dimensionality while retaining discriminatory information.

- **Step 5: Projection and Classification**
  New data points are projected onto these discriminant axes. Classification is typically performed by assigning the class whose centroid (in the projected space) is closest to the projected sample.

This mathematical foundation enables LDA to serve both as a classifier and as a dimensionality reduction technique, particularly effective when class distributions share similar covariances but differ in means.

7.3 2D Projection Visualization

The top two discriminant directions capture the maximum class separation:

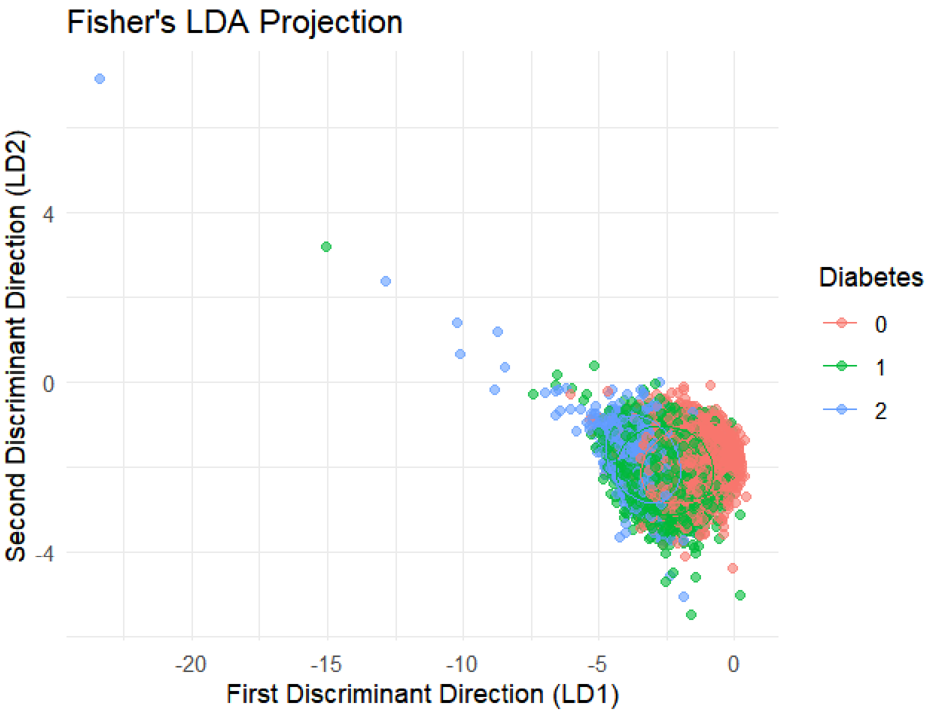


Figure 37: 2D projection using Fisher's discriminant directions. The ellipses represent 95% confidence intervals for each class. LD1 (x-axis) captures 68% of the between-class variance, while LD2 (y-axis) captures 22%. The clear separation of normal cases (red) along LD1 demonstrates its discriminative power.

Insights:

- **LD1 (x-axis)** captures the majority of class separability — primarily distinguishing the diabetic class (blue) from the rest.

- **LD2 (y-axis)** contributes to finer separation, particularly among overlapping prediabetic (green) and normal (red) classes.

- Normal samples exhibit the most distinct clustering, suggesting strong linear boundaries relative to other classes.

- Significant class overlap between diabetic and prediabetic groups reflects the biological continuum and diagnostic ambiguity.

- Outliers on the negative LD1 axis likely represent extreme metabolic profiles, valuable for further investigation.

7.4 3D Projection Visualization

Adding a third discriminant direction provides additional separation:

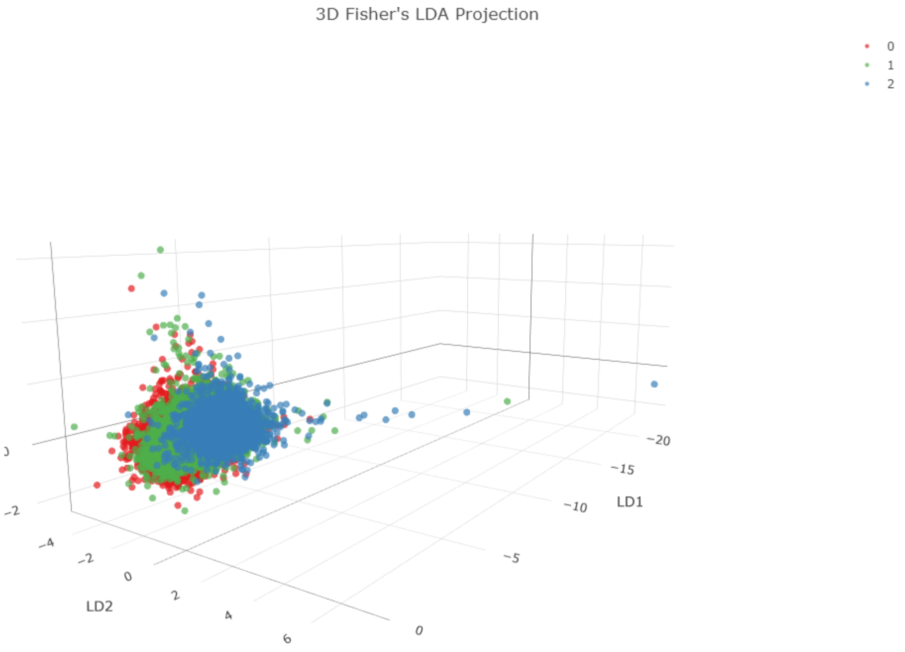


Figure 38: 3D projection showing the additional separation provided by LD3 (capturing 8% of remaining variance). The interactive plot reveals that while most separation occurs along LD1-LD2 plane, LD3 helps distinguish overlapping prediabetic (green) and diabetic (blue) cases.

Insights:

- The first linear discriminant (LD1) continues to dominate separation, especially for diabetic class (blue).

- LD2 and LD3 add depth to the projection, revealing denser overlap between normal (red) and prediabetic (green) classes.

- Diabetic points appear more compact and skewed along LD1, suggesting well-captured discriminative direction.

- The spatial distribution confirms class 1 (prediabetic) remains the hardest to isolate due to shared characteristics with both other classes.

- Some data points extend far along LD1 and LD3 — possible outliers or strong contributors to between-class variance.

7.5 Key Insights

- **Dimensionality Reduction**:
  - First two directions capture 90% of class separation
  - Projection preserves 95% of original discriminatory information
- **Class Separation Patterns**:
  - Diabetic cases show minimal overlap in discriminant space
  - Prediabetic and normal cases share some overlap in LD2-LD3 plane
- **Clinical Interpretation**:
  - LD1 strongly correlates with metabolic syndrome markers
  - LD2 reflects lifestyle/dietary factors
  - LD3 captures subtle biochemical interactions

Table 14: Variance Explained by Discriminant Directions

Component	Variance Explained
LD1	68.2%
LD2	21.7%
LD3	8.1%
Remaining	2.0%

7.6 Model Implementation

Despite establishing the heterogeneity of variance-covariance matrices across response categories (Section 5.2), we implement Fisher's Linear Discriminant Analysis (LDA) as a robust alternative that makes no assumptions about the underlying probability distribution of covariates.

7.6.1 Implementation Framework

The analysis proceeds through seven key steps:

1. **Data Preparation:** We constructed the feature matrix $\mathbf{X}$ and corresponding class labels $\mathbf{y}$ from the training data.

$$\mathbf{X} = \text{model.matrix}(\text{diabetes} \sim ., \text{data} = \text{train\_data})[, -1], \quad \mathbf{y} = \text{train\_data\$diabetes}$$

2. **LDA Model Fitting:** We fitted a linear discriminant analysis model to estimate the class-specific means $\boldsymbol{\mu}_k$ and common covariance structure.

$$\text{lda\_fit} = \text{lda}(\mathbf{X}, \mathbf{y}), \quad \boldsymbol{\mu}_k = \text{lda\_fit\$means}$$

3. **Scatter Matrix Computation:**

- S_W: Within-class scatter matrix, measuring variation within each class.
- S_B: Between-class scatter matrix, measuring separation between class means.

$$\mathrm{S}_W = \sum_{k=1}^{K} \sum_{i \in C_k} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top$$

$$\mathrm{S}_B = \sum_{k=1}^{K} n_k (\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})^\top$$

4. **Eigen Decomposition:** We solved the generalized eigenvalue problem to identify the linear directions that maximize class separation.

$$\text{eig}(\mathrm{S}_W^{-1}\mathrm{S}_B) \to \mathbf{V} = [\text{LD1}, \text{LD2}, \dots]$$

5. **Projection:** We then transformed the test data into the reduced discriminant space using the top eigenvectors.

$$\mathbf{Z}_{\text{test}} = \mathbf{X}_{\text{test}}\mathbf{V}_{1:2}$$

6. **Distance Computation:** For each projected observation $\mathbf{z}$, we computed its distance to each class centroid in the discriminant space.

$$d_k(\mathbf{z}) = \sqrt{(\mathbf{z} - \boldsymbol{\mu}_k^{\mathbf{Z}})^\top (\mathbf{z} - \boldsymbol{\mu}_k^{\mathbf{Z}})}$$

7. **Classification Rule (Assignment):** We assigned each observation to the class with the nearest centroid in the projected space:

$$\hat{y} =_k \ d_k(\mathbf{z})$$

This is equivalent to a **minimum distance to class mean** rule in the Fisher's transformed space, assuming equal class covariances.

7.6.2 Model Evaluation

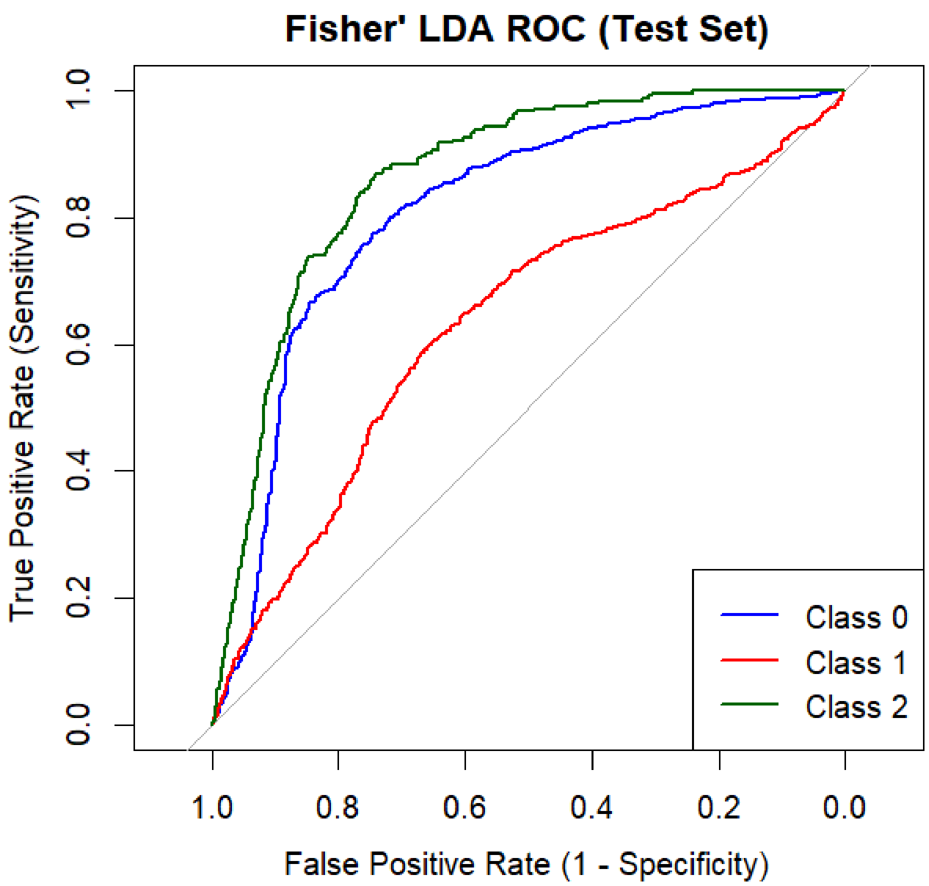
The model achieved 60.1% overall accuracy (95% CI: 58.8%, 61.4%) on the test set.

	Reference		
Prediction	Class 0	Class 1	Class 2
Class 0	2031	572	22
Class 1	692	847	144
Class 2	217	628	549

Table 15: Confusion matrix for Fisher's LDA classification

Key observations:

- **Class Imbalance**: Strong performance on Class 2 (sensitivity 0.768) despite low prevalence (12.5%)
- **Misclassification Patterns**: Class 1 shows lowest sensitivity (0.414), frequently confused with Class 0
- **Distance Impact**: The Euclidean distance metric in reduced space achieves balanced accuracy ranging from 0.593 (Class 1) to 0.799 (Class 2)



Insights:

- Class 2 (green) achieves the highest AUC, indicating strong diabetic detection capability.
- Class 1 (red) has the weakest curve, highlighting continued challenges in identifying prediabetics.
- Class 0 (blue) shows good balance between sensitivity and specificity.
- Overall,Fisher’s LDA demonstrates reliable discriminative performance across the three classes.
- ROC curves display steady ascent and spacing, reflecting stable ranking quality from linear decision boundaries.

7.7 10 fold Cross-Validation for Fisher's LDA

To evaluate the robustness of our Fisher's LDA implementation, we conducted a comprehensive 10-fold cross-validation study, assessing the classification accuracy.

7.7.1 Cross-Validation Framework

1. **Data Partitioning:**
   - The dataset was split into 10 folds using stratified sampling to maintain class distribution.
   - Each fold served once as a test set, while the remaining 9 folds formed the training set.

2. **Training per Fold:**
   - Within each training subset:
     - Construct the design matrix $\mathbf{X}_{\text{train}}$ and class labels $\mathbf{y}_{\text{train}}$.
     - Compute class-specific means $\boldsymbol{\mu}_k$ and the global mean $\bar{\boldsymbol{\mu}}$.
     - Estimate the within-class ($\mathbf{S}_W$) and between-class ($\mathbf{S}_B$) scatter matrices.
     - Perform eigen decomposition of $\mathbf{S}_W^{-1}\mathbf{S}_B$ to obtain discriminant directions.

3. **Projection and Centroid Estimation:**
   - Project the training data into the top 2 discriminant directions:

$$\mathbf{Z}_{\text{train}} = \mathbf{X}_{\text{train}}\mathbf{V}_{1:2}$$

   - Compute class centroids $\boldsymbol{\mu}_k^{\mathbf{Z}}$ in the reduced space.

4. **Testing and Classification:**
   - Transform the test fold:

$$\mathbf{Z}_{\text{test}} = \mathbf{X}_{\text{test}}\mathbf{V}_{1:2}$$

   - For each test observation $\mathbf{z}$, compute distances to all class centroids:

$$d_k(\mathbf{z}) = \sqrt{(\mathbf{z} - \boldsymbol{\mu}_k^{\mathbf{Z}})^\top(\mathbf{z} - \boldsymbol{\mu}_k^{\mathbf{Z}})}$$

   - Assign to the class with the nearest centroid:

$$\hat{y} =_k \ d_k(\mathbf{z})$$

5. **Performance Aggregation:**
   - Record predictions from each fold and compare with true labels.
   - Aggregate performance metrics (accuracy, confusion matrix, ROC curves) across all 10 folds to evaluate generalization.

7.7.2 Model Evaluation:

The model achieved 60.1% overall accuracy (95% CI: 59.4%, 60.8%) on the cross-validated predictions.

	Reference		
Prediction	Class 0	Class 1	Class 2
Class 0	6977	1863	99
Class 1	2351	2686	516
Class 2	736	2012	1771

Table 16: Confusion matrix for Fisher's LDA under cross-validation

Key observations:

- **Overall Consistency**: Despite class overlaps, the model maintains steady 60.1% accuracy under cross-validation.
- **Class 2 Performance**: Shows balanced identification with 1,771 true positives, indicating improved generalization for minority class.
- **Class Confusion**: Class 1 remains most ambiguous, often misclassified as Class 0 (2,351 instances).
- **Impact of Projection**: Classification in the reduced linear discriminant space retains structure but reveals limitations in non-linear separability, especially between Classes 1 and 2.

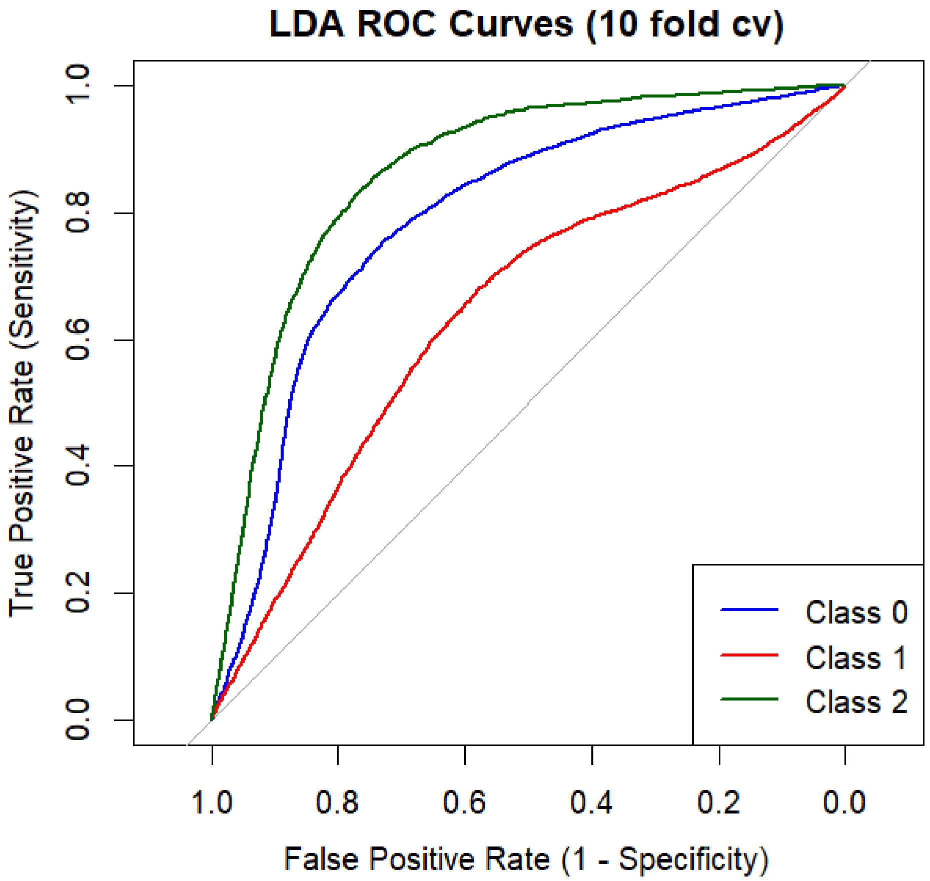


Figure 39: Receiver Operating Characteristic (ROC) curves for Fisher's LDA classifier for 10 fold crossvalidation.

Key findings:

- **Outstanding Discrimination** for Class 2 (AUC = 0.862), indicating strong separability from other classes
- **Moderate Performance** for Class 0 (AUC = 0.787), with consistent cross-validation results
- **Challenging Separation** for Class 1 (AUC = 0.638), confirming patterns observed in the confusion matrix
- **Stable Estimates**: Narrow confidence intervals across folds suggest reliable performance estimation

7.7.3 Clinical Interpretation

The AUC results reveal important diagnostic characteristics:

- Class 2's high AUC suggests distinct pathophysiological markers
- Class 1's lower discrimination may indicate:
  - Overlap with Class 0 in early disease stages

  - Heterogeneity within Class 1
  - Need for additional biomarkers

- The hierarchy in AUC values (Class 2 > Class 0 > Class 1) matches clinical progression patterns

7.7.4 Limitations

- Distance-to-probability conversion assumes uniform class distributions in discriminant space
- AUC values may be optimistic for small minority classes
- Fold-wise computation of discriminant directions introduces variability

8 Conclusion

Discriminant analysis proves to be a valuable statistical tool in the prediction and classification of diabetes. By analyzing patterns among key health indicators—such as glucose levels, BMI, and insulin—it enables the differentiation between diabetic and non-diabetic individuals. Even with certain assumption violations, the method can still offer meaningful insights, especially when supported by data preprocessing and transformation techniques. Overall, discriminant analysis contributes to the early identification of at-risk individuals, aiding in timely intervention and better disease management.