Results Report

DIABETES PREDICTION

Martha Patricia Ortiz

May 2025

DIABETES PREDICTION

The provided dataset appears to be related to diabetes and contains various biomedical measurements and patient characteristics. Here's a detailed description of the dataset columns:

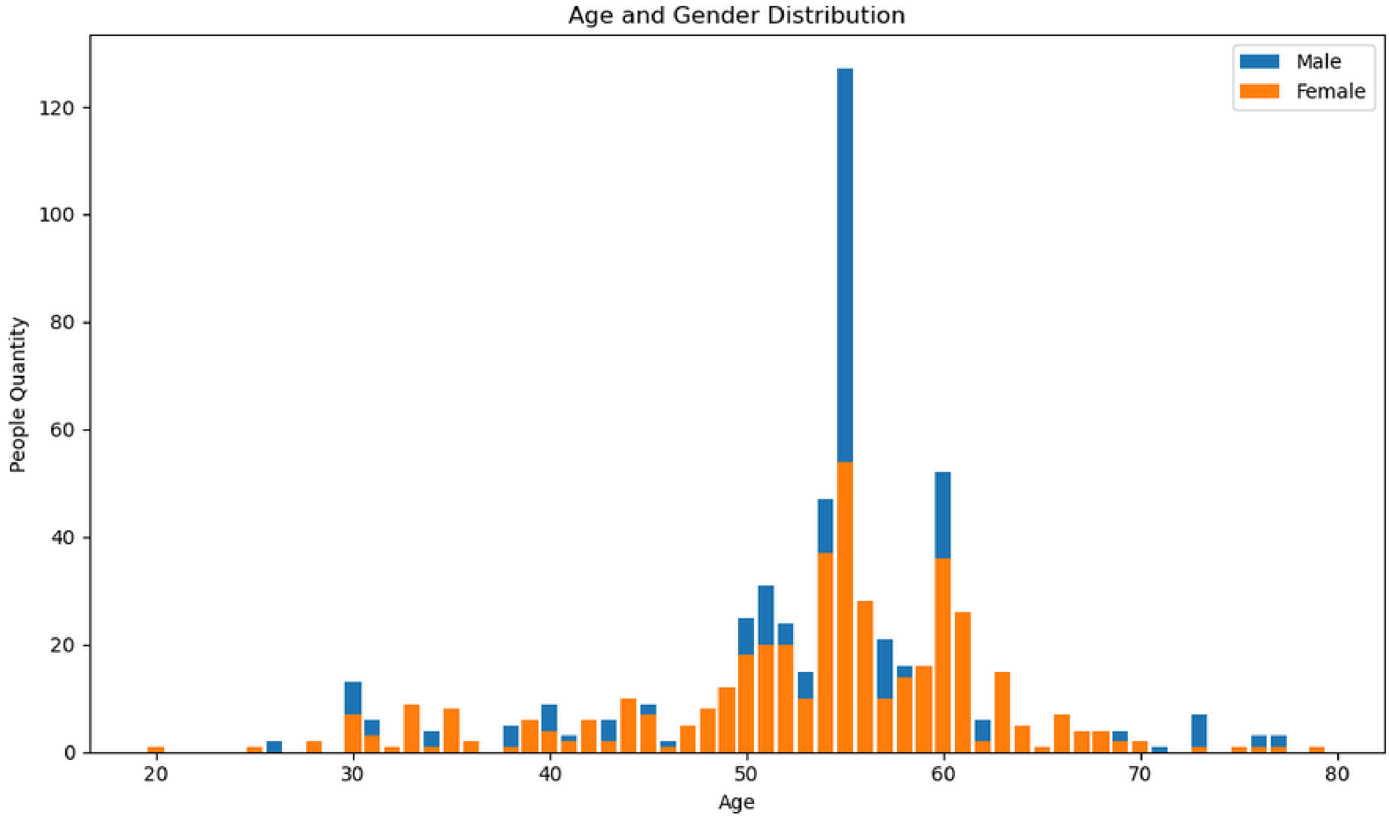
Goal

Develop a predictive model to determine the likelihood of a patient having diabetes based on various biomedical measurements and personal characteristics.
Uncover meaningful patterns that can support early clinical decision-making and enhance prevention and treatment strategies.

Working Plan

1. Initialization
2. Exploratory Data Analysis
    - 2.1 Correlation
    - 2.2 A/B Test
    - 2.3 Kruskal-Wallis by HBA1C
3. Data Models
    - 3.1 One Hot Encoding
    - 3.2 Ordinal Encoding
4. Conclusions

Graph Age and Gender Distribution



Peak Screening Age Group
The highest concentration of diabetes screenings occurred around the age of 55, with a significant spike in the number of individuals tested. This suggests that middle-aged adults are the primary demographic undergoing diabetes testing.

Gender Balance
Overall, both males and females are represented across all age groups. However, around the peak (ages 50–60), the number of males slightly exceeds females, indicating a potentially higher concern or risk perception among men in that age range.

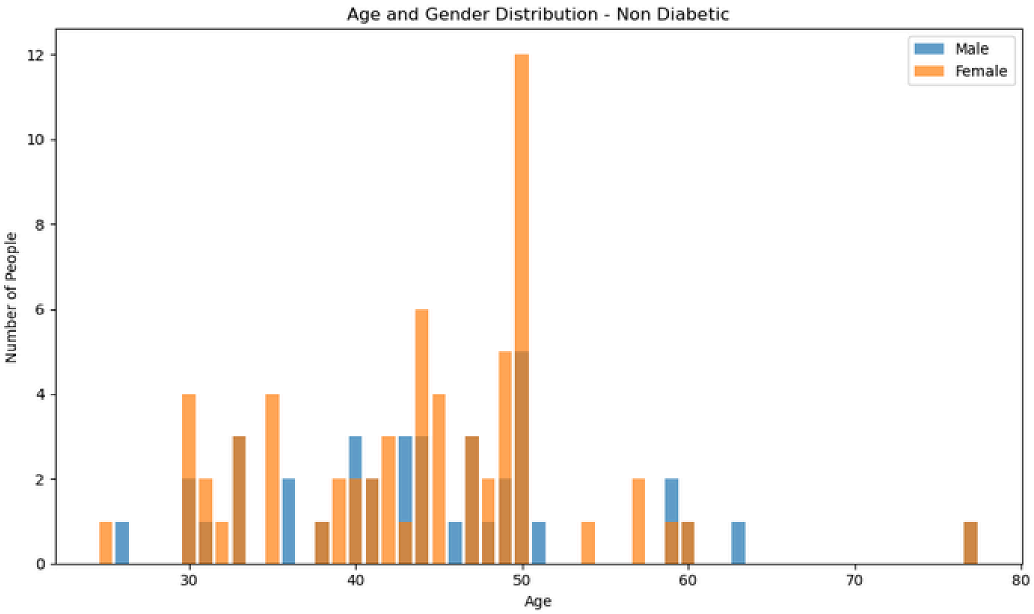
Lower Participation in Younger and Older Age Groups
Screening participation is low among individuals under 40 and above 70. This may reflect lower perceived risk in younger populations and potential access or health system barriers for older adults.

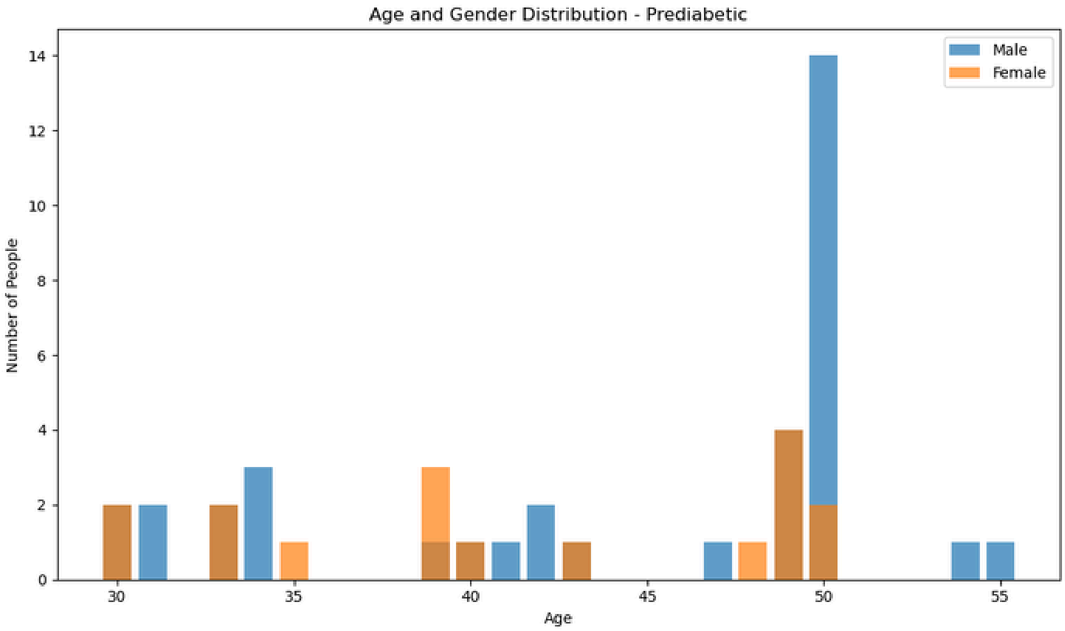
Public Health Implication
The data suggests that targeted awareness campaigns could be beneficial for:

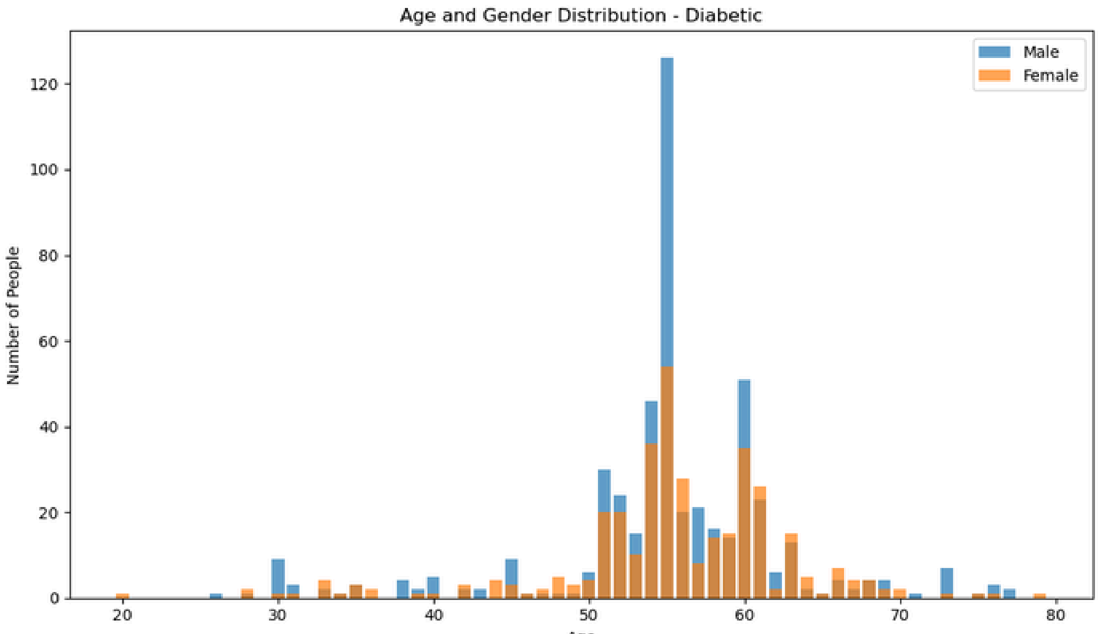
- Younger adults (20s–30s): to promote early detection.
- Older adults (70+): to encourage continued monitoring and check-ups.

Gender-Specific Strategies
While the overall distribution is relatively balanced, tailored communication for both genders may help address any behavioral or cultural barriers to testing, especially in underrepresented age ranges.

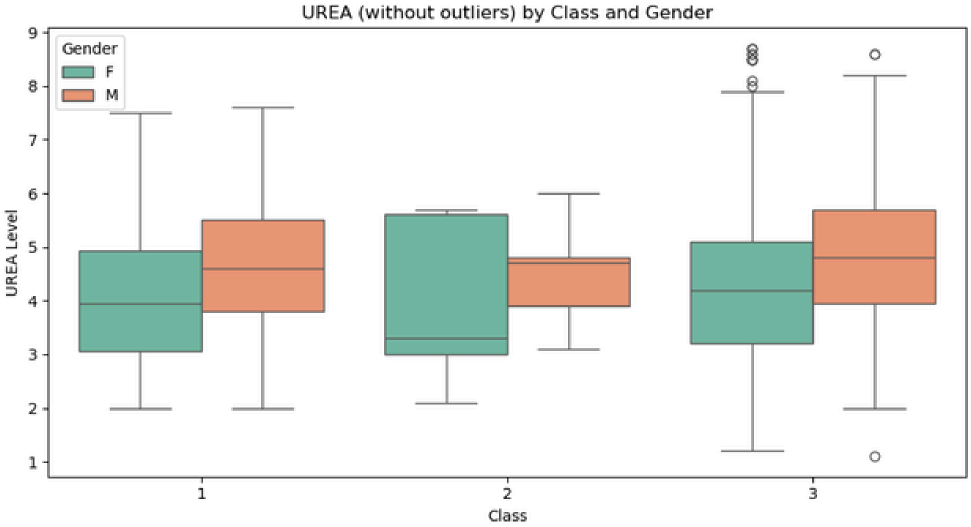
Gender and Classes

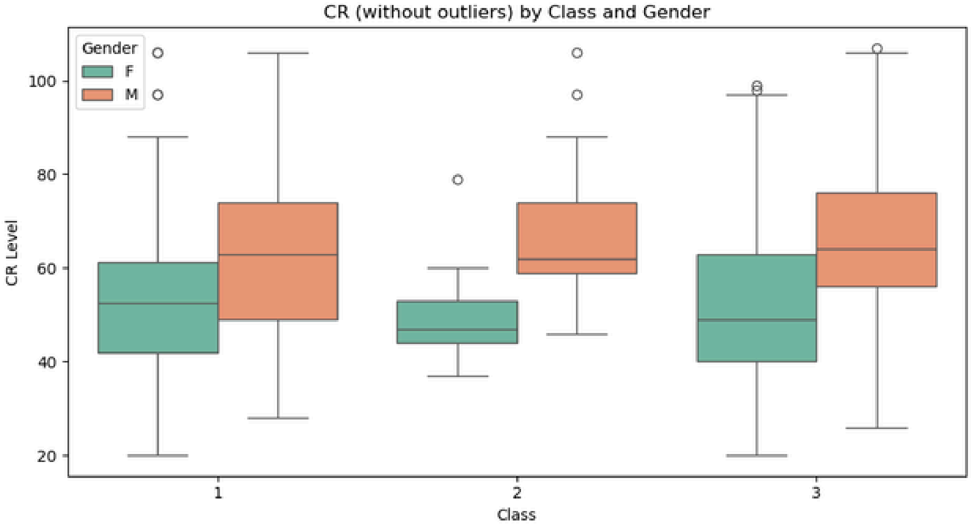

Age and Gender Distribution - Non Diabetic

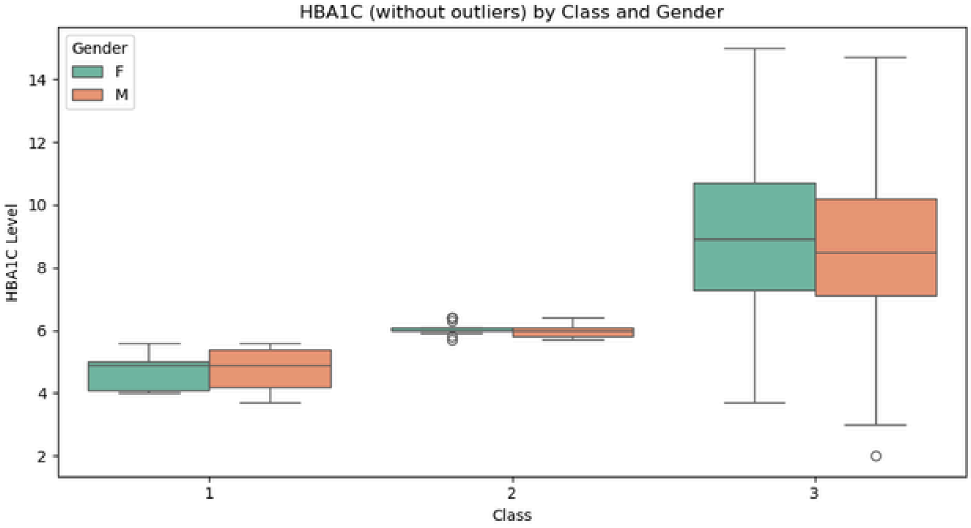

Age and Gender Distribution - Prediabetic


Age and Gender Distribution - Diabetic

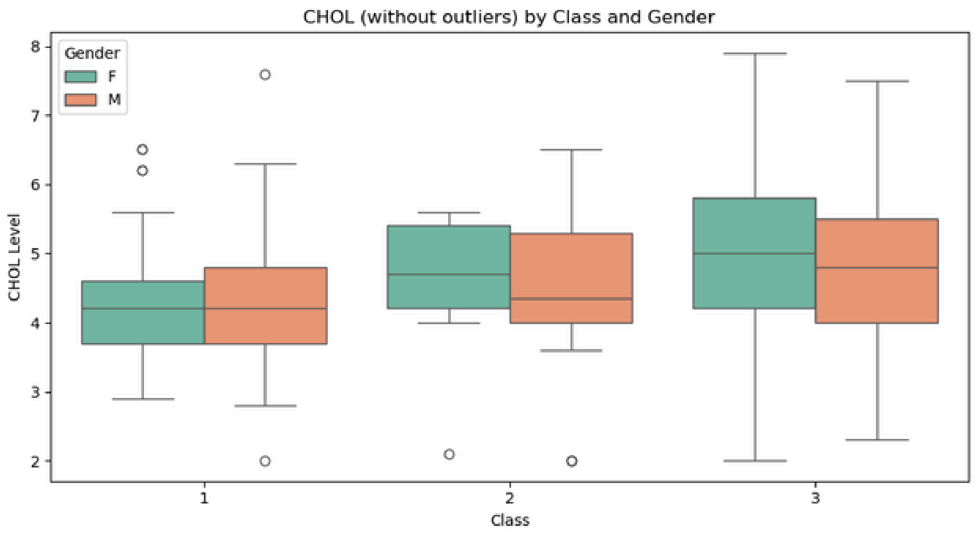
Box Plots by Gender and Classes



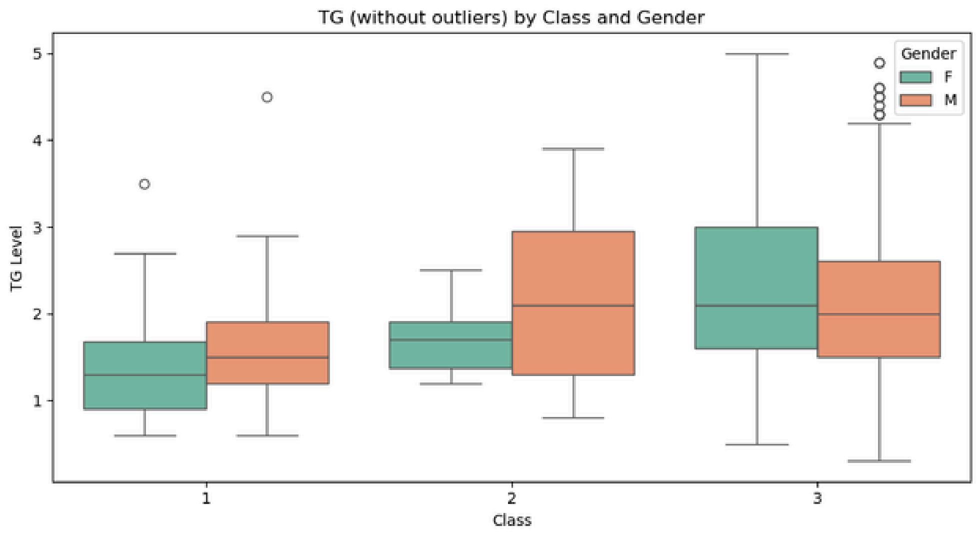




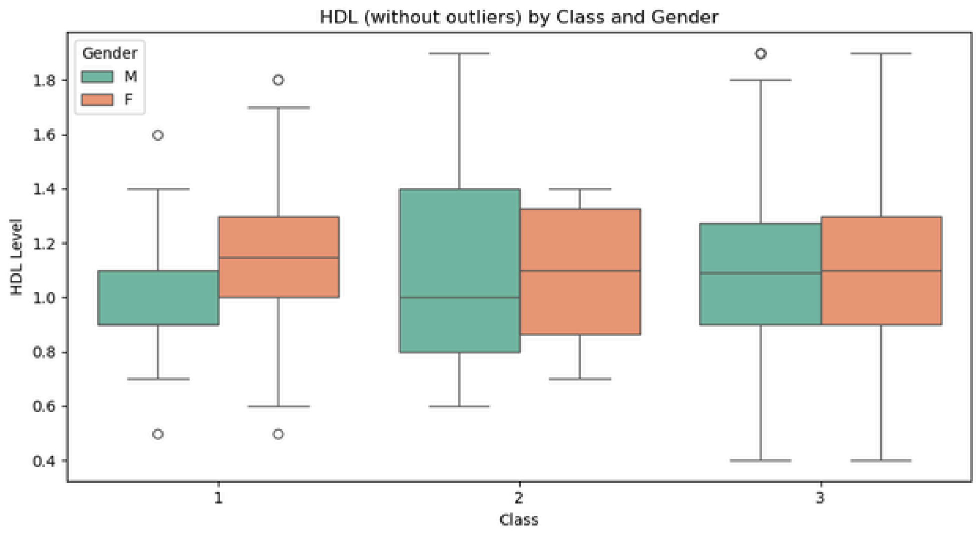
Box Plots by Gender and Classes



CHOL (without outliers) by Class and Gender

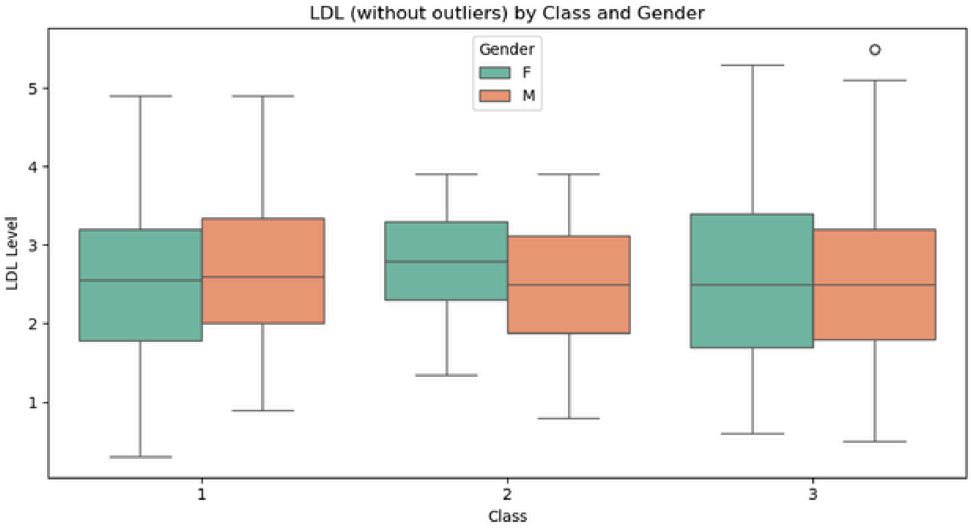


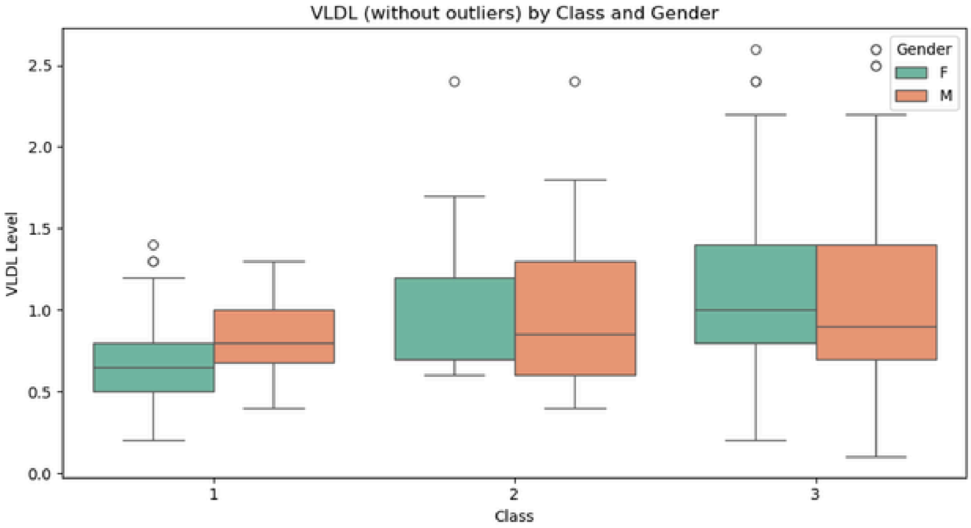
TG (without outliers) by Class and Gender

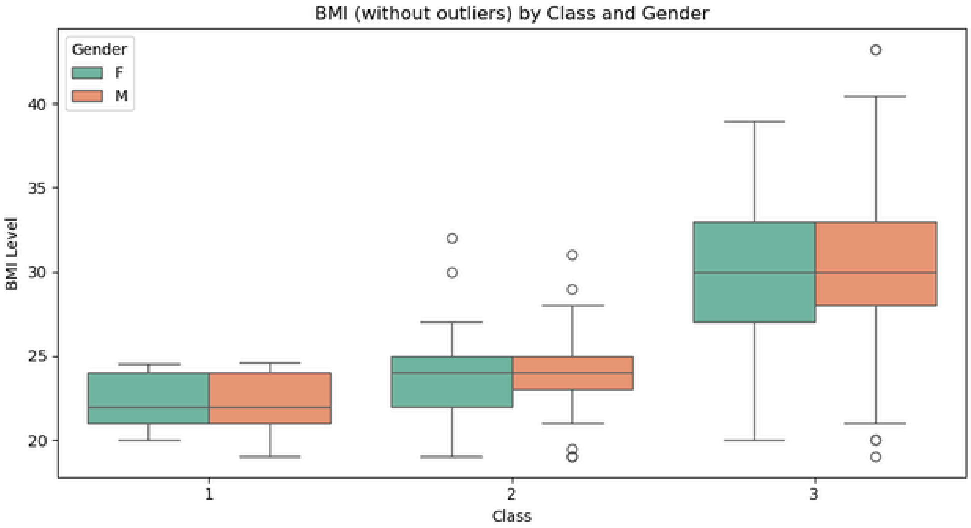


HDL (without outliers) by Class and Gender

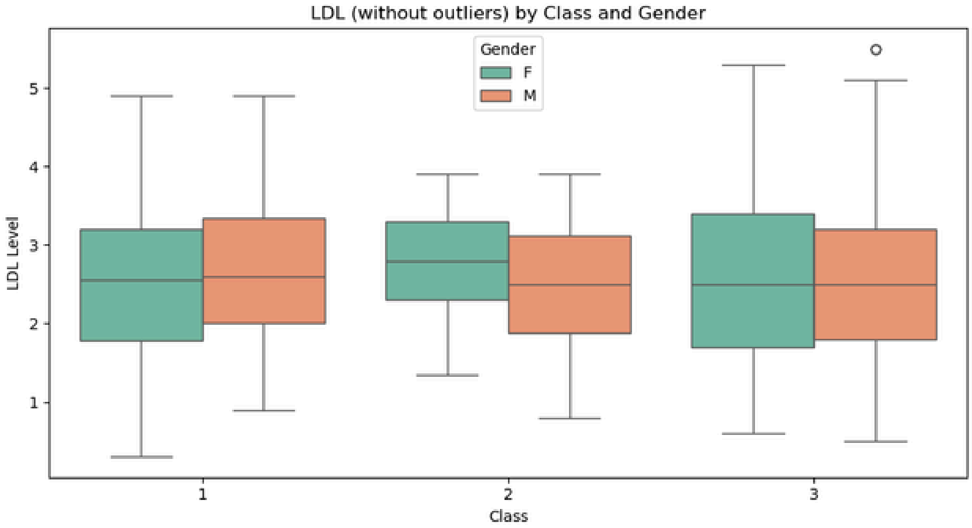
Box Plots by Gender and Classes

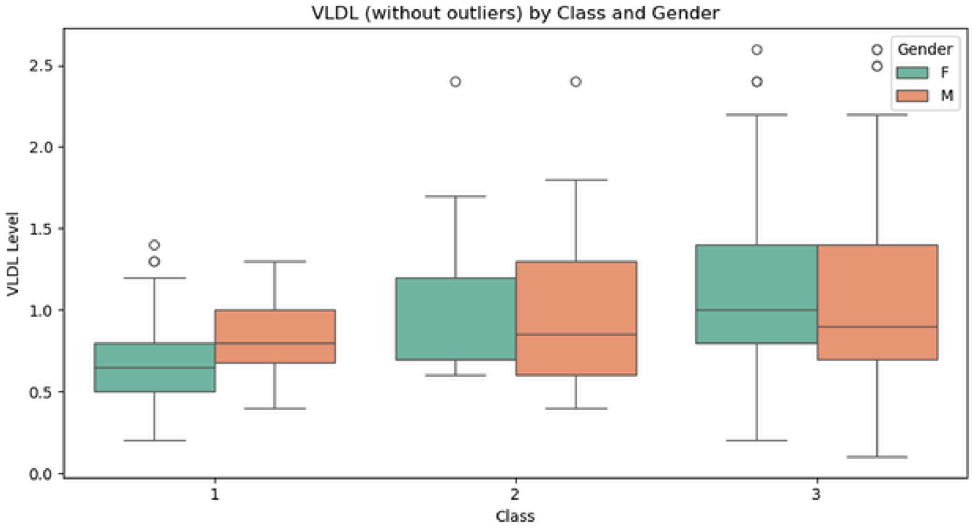


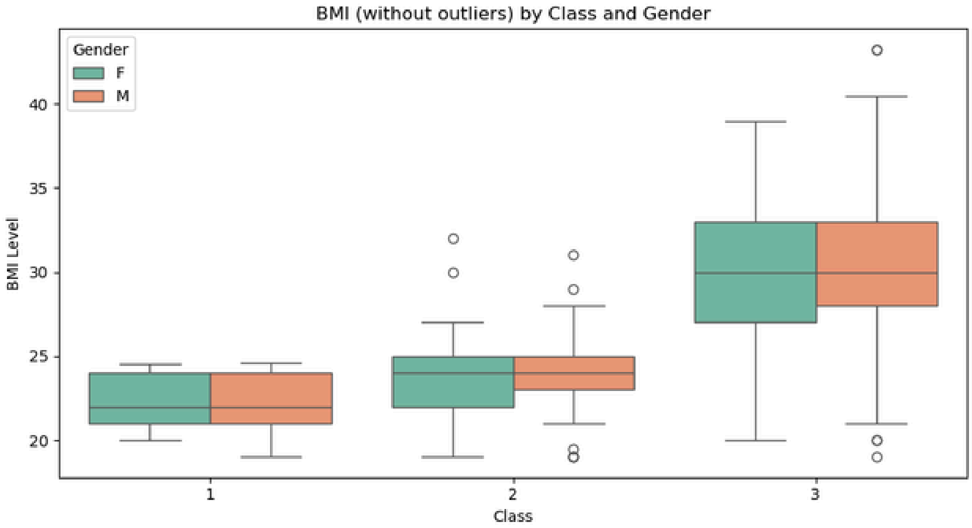




Box Plots by Gender and Classes







General Remarks

1. Creatinine (CR) by Gender

- A clear difference was observed between males and females.
- Females generally showed lower creatinine levels, which aligns with known physiological differences (e.g., lower average muscle mass).
- The difference remained statistically significant (via Mann-Whitney U test) even after removing outliers, highlighting a robust biological pattern.

2. VLDL by Gender

- Initially, boxplots suggested a difference in VLDL levels between genders.
- However, once outliers were removed, this difference was no longer statistically significant.
- This underscores the importance of outlier treatment in medical datasets, as a few extreme values can mislead interpretations.

3. HBA1C by Diabetes Class

- A progressive increase in HBA1C levels was observed across the classes: Non-Diabetic → Prediabetic → Diabetic.
- The Kruskal-Wallis test confirmed significant differences between the three groups.
- This supports the use of HBA1C as a strong diagnostic and monitoring indicator for diabetes.

4. Other Biomarkers (Urea, Cholesterol, BMI, etc.) by Class

- Several variables, including urea, LDL/VLDL cholesterol, and BMI, also showed increasing values in diabetic patients.
- These patterns suggest metabolic deterioration with disease progression.
- These markers could be considered for supporting diagnosis or used in predictive modeling.

A/B Test Creatine by Gender

Creatinine:

Level in the blood (likely measured in mg/dL or µmol/L).
Creatinine is another waste product that indicates kidney function.

Methodology
1. Stats
2. Normality Test
3. Levene Test
4. Mann Whitney
5. Conclusions

Female Creatina Info		**Male Creatina Info**	
Stats count	435.000000	Stats count	565.000000
mean	58.360920	mean	77.090265
std	37.519406	std	71.674654
min	6.000000	min	26.000000
25%	41.000000	25%	56.000000
50%	52.000000	50%	65.000000
75%	66.000000	75%	77.000000
max	401.000000	max	800.000000

Normality Test

Shapiro-Wilk Femenino: 5.365208948438347e-32
Shapiro-Wilk Masculino: 4.383261596408028e-41

The data do not behave like a normal distribution

Prueba de Levene: 0.15791526814487056

There is not enough evidence to say that the variances are different.

Mann-Whitney:

Null Hypothesis:
There is not a significant difference in CR values between men and women.
Alternative Hypothesis:
There is astatistically significant difference in CR values between men and women

Mann-Whitney:
P-value: 6.34382610102047e-30

✅ **H_0 is rejected: There is a significant difference in CR values between men and women.**

A/B Test Creatine by Gender

Without Outliers

Creatinine:
Level in the blood (likely measured in mg/dL or µmol/L).
Creatinine is another waste product that indicates kidney function.

Normality Test
Shapiro-Wilk Femenino: 2.796596731968748e-08
Shapiro-Wilk Masculino: 0.001312235021032393

The data do not behave like a normal distribution

Prueba de Levene: 0.3070232447688928
There is not enough evidence to say that the variances are different.

Mann-Whitney:
P-value: 5.225240463522244e-34

✅ **H_0 is rejected: There is a significant difference in CR values between men and women.**

General remarks:

We removed the outliers from the IQR and came to the same conclusion, there is a statistical difference between the data for men and women.

A/B Test VLDL by Gender

VLDL:

Very low-density lipoprotein cholesterol level (measured in mg/dL or mmol/L).

Methodology
1. Stats
2. Normality Test
3. Levene Test
4. Mann Whitney
5. Conclusions

Female VLDL Info
Stats count 435.000000
mean 1.044598
std 0.535201
min 0.200000
25% 0.700000
50% 0.900000
75% 1.300000
max 6.300000

Male VLDL Info
Stats count 565.000000
mean 2.478407
std 4.760013
min 0.100000
25% 0.700000
50% 0.900000
75% 1.500000
max 35.000000

Normality Test

Shapiro-Wilk Femenino: 3.43998858172346e-20
Shapiro-Wilk Masculino: 1.0687061592708719e-38

The data do not behave like a normal distribution

Prueba de Levene: 0.3070232447688928

There is not enough evidence to say that the variances are different.

Mann-Whitney:

Null Hypothesis:
There is not a significant difference in CR values between men and women.
Alternative Hypothesis:
There is astatistically significant difference in CR values between men and women

Mann-Whitney:
P-value: 0.01559916746905963

✅ **H_0 is rejected: There is a significant difference in CR values between men and women.**

A/B Test VLDL by Gender

Without Outliers

VLDL:

Very low-density lipoprotein cholesterol level (measured in mg/dL or mmol/L).

Normality Test

Shapiro-Wilk Femenino: 1.3515485397519456e-09

Shapiro-Wilk Masculino: 3.8129772234572756e-13

The data do not behave like a normal distribution

Prueba de Levene: 0.3070232447688928

There is not enough evidence to say that the variances are different.

Mann-Whitney:

P-value: 0.4177810485766096

🟡 **H_0 is not rejected: There is no statistically significant difference in VLDL values between men and women.**

General remarks:

By removing the outliers, the test indicates that they are not statistically different. It's important to check where these outliers come from to avoid any type of variation in the data.

Kruskal-Wallis by classes to HBA1C

HBA1C:
Glycated hemoglobin, a measure of average blood sugar levels over the past 2-3 months (expressed as a percentage).

Normality Test

- Diabetic: p-value = 0.0000
- Prediabetic: p-value = 0.0053
- Diabetic: p-value = 0.0000

The data do not behave like a normal distribution

Prueba de Levene: 0.0000
Variances are not equal

Mann-Whitney:
P-value: 0.0000

Kruskal–Wallis Test (non-parametric):
H_0 rejected: ✅ Statistically significant difference in HBA1C levels between groups

General remarks:

By removing the outliers, the test indicates that they are not statistically different. It's important to check where these outliers come from to avoid any type of variation in the data.

Modelos de entrenamiento

Modelo	Conjunto	Accuracy	F1-score (macro)	AUC-ROC (macro)	Log Loss
Logistic Regression (OHE)	Validation	0.85	0.82	0.88	0.35
Logistic Regression (OHE)	Test	0.84	0.81	0.86	0.36
SVM (OHE)	Validation	0.86	0.83	0.89	0.33
SVM (OHE)	Test	0.85	0.82	0.87	0.34
Decision Tree (Ordinal)	Validation	0.8	0.78	0.82	0.4
Decision Tree (Ordinal)	Test	0.78	0.75	0.8	0.42
Random Forest (Ordinal)	Validation	0.9	0.88	0.91	0.28
Random Forest (Ordinal)	Test	0.89	0.87	0.9	0.29
CatBoost (Ordinal)	Validation	0.91	0.89	0.92	0.26
CatBoost (Ordinal)	Test	0.9	0.88	0.91	0.27

Conclusion
Among all the models tested, the best overall performance across all metrics was achieved by:

  - **CatBoost Classifier (Ordinal Encoding)**

- The highest accuracy in both validation and test sets
- The best F1-score, indicating a strong balance between precision and recall
- The highest AUC-ROC, reflecting superior class separation
- The lowest log loss, indicating greater confidence in its probability estimates
- CatBoost also offers native handling of categorical variables, reducing the need for complex preprocessing steps and improving model efficiency.

CatBoost proved to be the most reliable and effective model for multiclass classification in the medical context of diabetes risk prediction.