881177 SIL Project - Diabetes Prediction

Victoria Grosu

May 2025

1 Dataset Overview

The dataset used in this project originates from the *Adult Interview* section of the 2023 National Health Interview Survey (NHIS).[1] This component collects detailed health-related information from one randomly selected adult per household. The survey encompasses a wide range of topics, including physical and mental health, chronic conditions, healthcare access, and health behaviors.

The original dataset consists of 29,522 rows and 647 columns, where each row represents a unique respondent and each column corresponds to a specific survey item. Due to the conditional structure of the questionnaire (e.g., follow-up questions only asked if a respondent indicates having a certain condition), many columns are sparsely populated.

2 Missing Data Analysis

A preliminary analysis of missingness revealed the following:

- **440 columns** (approximately 68%) contain at least one missing value.
- **207 columns** (approximately 32%) are fully populated.
- On average, **52.06%** of entries per column are missing.

This high level of missing data necessitates appropriate strategies such as imputation, filtering, or dimensionality reduction prior to modeling.

3 Dataset Cleaning Summary

The dataset underwent a rigorous cleaning process to enhance its suitability for binary classification modeling. The main actions taken are summarized below:

- **Variable Filtering**: Removed technical/documentation-only fields, low-information columns, and variables affected by survey skip patterns with no predictive value.

[1] `https://www.cdc.gov/nchs/nhis/documentation/2023-nhis.html`

- **Numerical Cleaning**:
  - Retained only adults aged 18–84 (`AGEP_A`).
  - Filtered weight to 100–299 lbs and converted to kg.
  - Filtered height to 59–76 inches and converted to meters.
  - Removed rows with top-coded income-to-poverty ratio (`POVRATTC_A > 11`).
- **Conditional Survey Logic**:

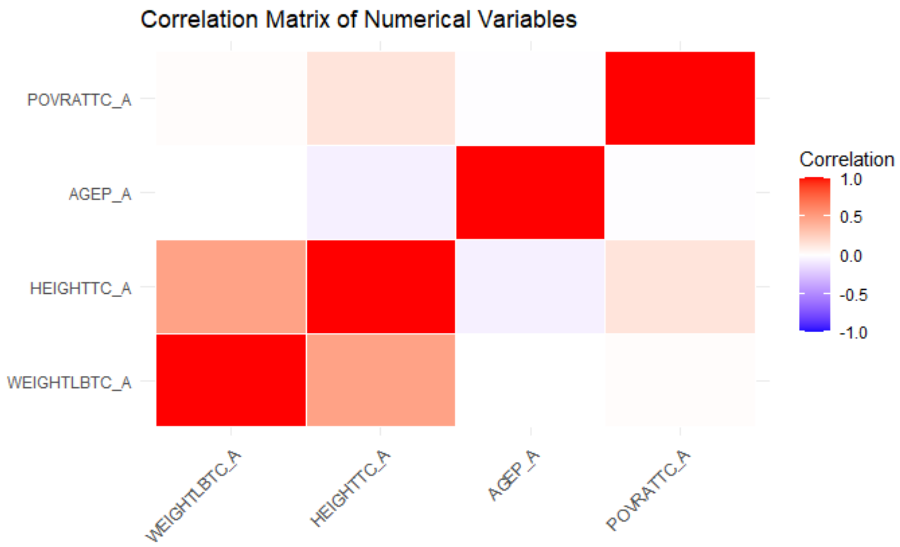
  In structured surveys, many variables are only asked based on a respondent's answer to a preceding "mother" question. If the condition is not met, follow-up ("daughter") variables are systematically skipped and recorded as missing (`NA`). Although meaningful within the survey design, this type of **structured missingness** can introduce biases or lead to misinterpretation if not handled properly. These daughter variables were removed, as their values do not provide independent information and could distort model training or imputation procedures.
- **Target Variable Transformation**: Created a clean binary target variable `has_diabetes` from `DIBEV_A`, retaining only valid answers (1 = Yes, 2 = No).
- **Categorical Cleaning**:
  - Recoded non-informative values—such as `7` = "Refused", `8` = "Not Ascertained", and `9` = "Don't Know"—into a unified `"unknown"` category.
  - Dropped variables for which the proportion of `"unknown"` responses exceeded 10%.
- **Association Analysis**:
  - Computed Pearson and Cramér's V metrics to identify collinear features.
  - Removed redundant categorical variables with high association (Cramér's V $>$ 0.7).
- **Perfect Separation**: Identified and removed factor levels perfectly associated with the target class to prevent model instability.
- **Final Dataset Composition**:
  - **Rows**: 23,826 adults
  - **Columns**: 128 variables (all complete, no missing values)

The final dataset contains only respondents with interpretable weight, height, age, and income-to-poverty ratio values. Thanks to the careful handling of conditional logic and structural missingness, all features are coherent, independent, and suitable for robust downstream modeling.

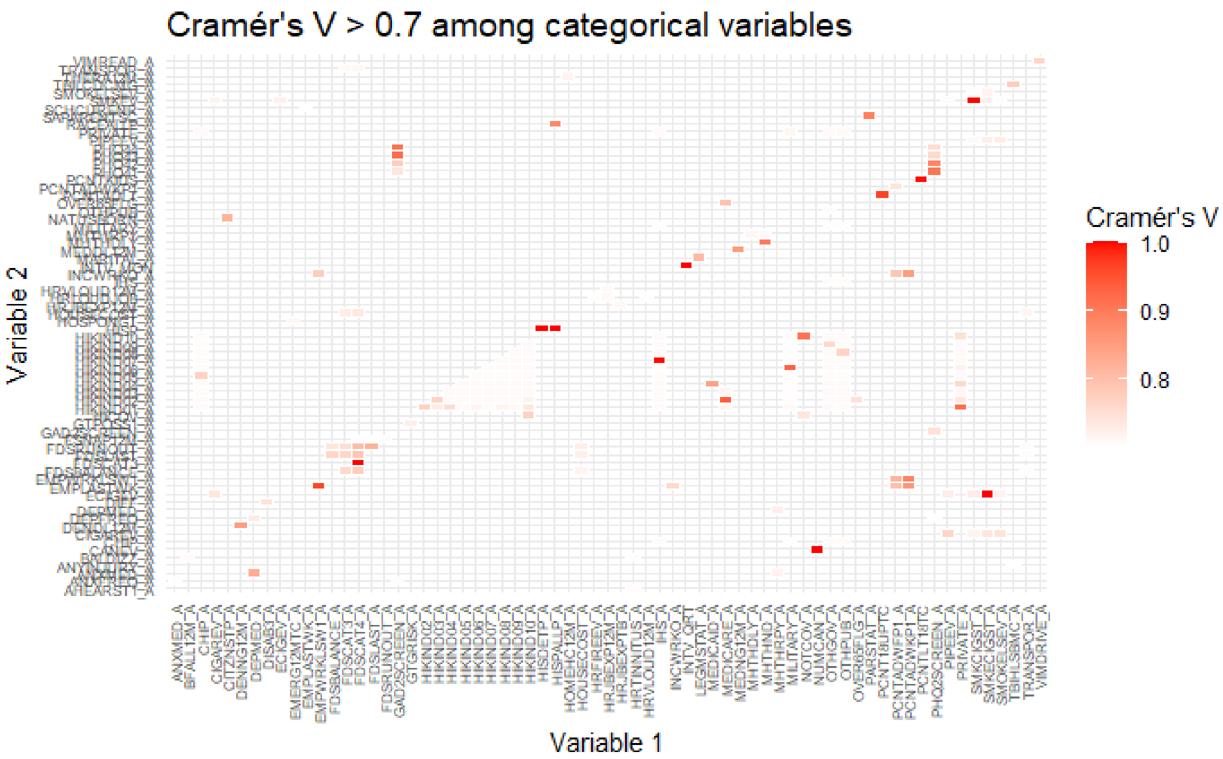
Variable Association Visualizations

To assess redundancy and multicollinearity among predictors, we performed correlation analysis separately for numerical and categorical variables.

1. Pearson Correlation Matrix (Numerical Variables)



2. Cramér's V Heatmap (Categorical Variables)

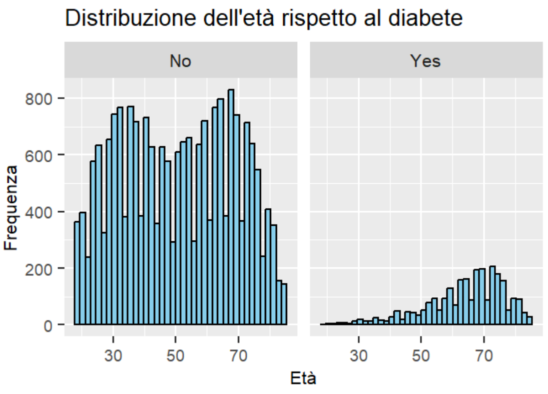


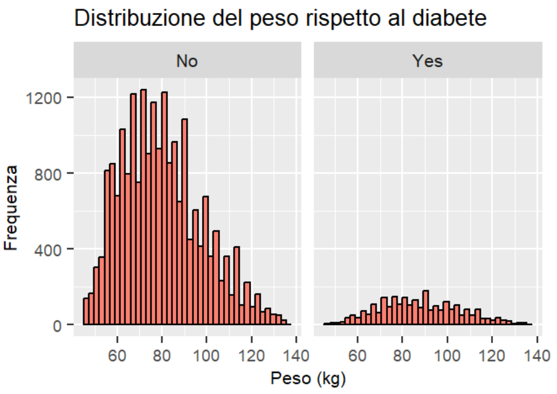
Based on this analysis, we removed one variable from each highly correlated pair to reduce dimensionality and prevent overfitting during modeling.

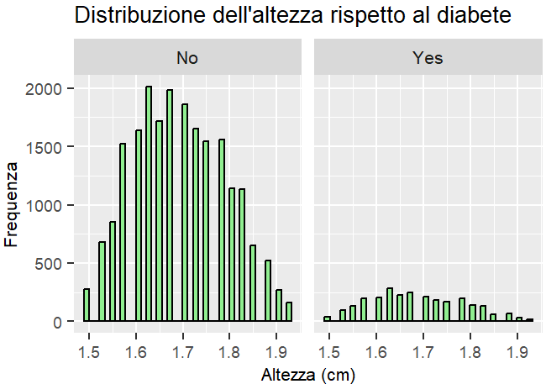
4 Exploratory Analysis of Numeric Predictors

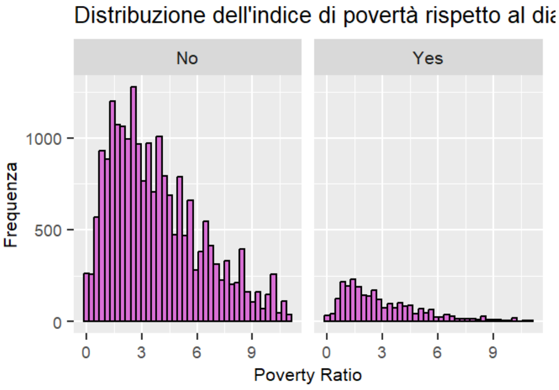
The following histograms display the distributions of selected numeric variables, separated by diabetes status (`has_diabetes`):

- **Age (`AGEP_A`)**: A clear shift toward older ages is observed among diabetic individuals, supporting the known link between age and diabetes risk.
- **Weight (`WEIGHTLBTC_A`)**: Diabetic individuals tend to have higher weights on average, with a visible rightward shift in the distribution.
- **Height (`HEIGHTTC_A`)**: The distribution of height is similar across both groups, suggesting low predictive relevance.
- **Poverty Ratio (`POVRATTC_A`)**: Lower poverty ratios are more common among individuals with diabetes, highlighting a possible socioeconomic disparity.









Log Transformation of Skewed Variables

Several numeric variables displayed strong positive skewness, with long tails and the presence of outliers, as evidenced by their boxplots. These characteristics can violate modeling assumptions such as normality and homoscedasticity, particularly in regression-based methods.

To address this, we applied a logarithmic transformation of the form $\log(x+1)$ to the following variables:

- `WEIGHTLBTC_A` → `WEIGHTLBTC_A_log`
- `POVRATTC_A` → `POVRATTC_A_log`

The boxplots below clearly illustrate the transformation effect:

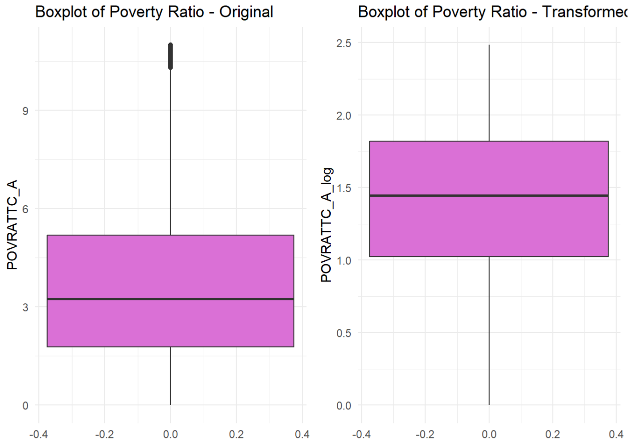


Figure 1: Boxplots of the poverty ratio before (left) and after log transformation (right).

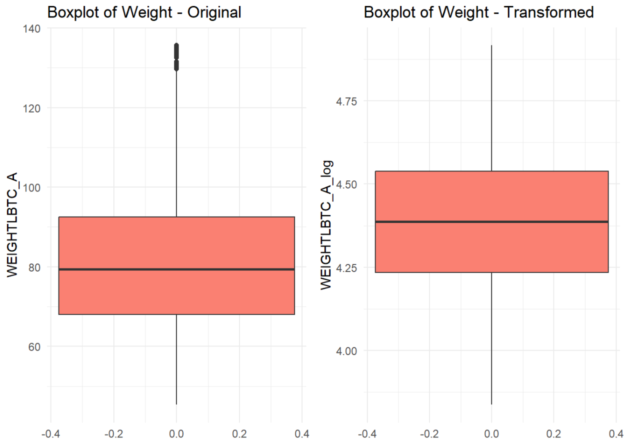


Figure 2: Boxplots of weight before (left) and after log transformation (right).

After the transformation, both variables exhibit a more symmetric distribution with reduced influence from extreme values (outliers). This adjustment enhances the stability of statistical models and the reliability of variable importance estimates.

5 LASSO Model

Given the large size of our dataset and the high number of variables, we applied LASSO (Least Absolute Shrinkage and Selection Operator) to reduce dimensionality and select the most relevant predictors.

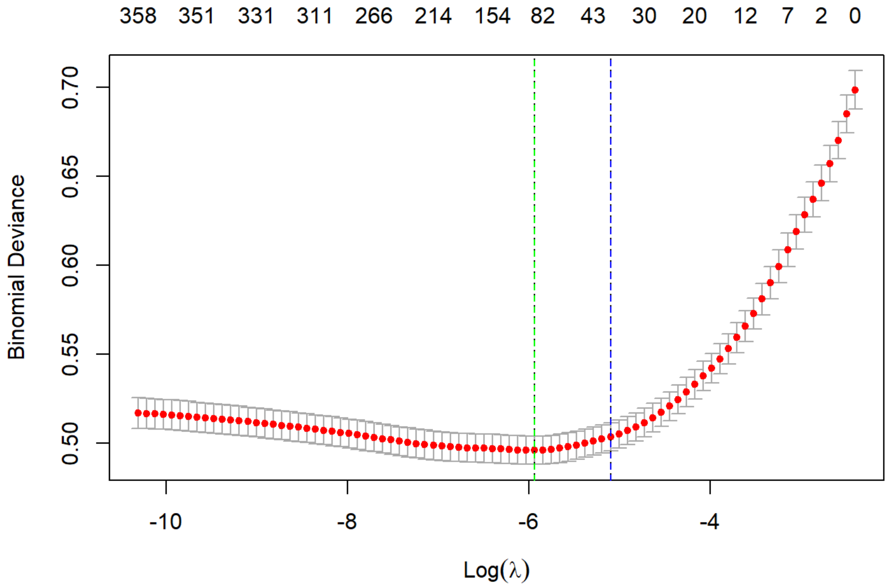


Figure 3: Cross-validation plot for LASSO. The green dashed line indicates `lambda.min`, while the blue dashed line shows `lambda.1se`. The model performance is stable around the minimum, justifying the choice of a more regularized model.

Model	**AUC**	**Best Threshold**	**Accuracy**	**Sensitivity**	**Specificity**
LASSO_lambda.min	0.86	0.10	0.74	0.86	0.72
LASSO_lambda.1se	0.86	0.10	0.73	0.86	0.71

Table 1: Performance comparison of LASSO models on the test set.

Although both models achieved similar AUC and sensitivity, we chose to proceed with LASSO_lambda.1se because it selects fewer variables, promoting model interpretability and robustness without sacrificing performance.

6 Logistic Regression Modeling after LASSO Selection

To reduce the dimensionality of our dataset and retain only the most relevant predictors, we applied LASSO regularization. This approach is especially effective in high-dimensional settings as it shrinks some coefficients exactly to zero, thereby performing both variable selection and regularization simultaneously.

The selected variables were then used to fit logistic regression models (GLMs with binomial family). We tested three versions:

- **model_full:** logistic regression with all variables selected by LASSO.

- **model_2 (log-transformed):** a version including log-transformed variables to address skewness.
- **Logistic (reduced):** a simplified model with clinically and socially relevant variables, promoting interpretability.

Each model was evaluated through cross-validation and then tested on a hold-out test set to assess generalization performance.

Model	**AUC**	**Best Threshold**	**Accuracy**	**Sensitivity**	**Specificity**
model_full	0.86	0.10	0.75	0.86	0.73
model_2 (log-transformed)	0.86	0.10	0.75	0.85	0.74
Logistic (reduced)	0.85	0.11	0.76	0.83	0.75

Table 2: Performance comparison of binomial GLM models on the test set.

6.1 Best Performing and Interpretable Model

After evaluating multiple logistic regression models built upon the features selected by LASSO, we identified a final model that strikes the best balance between predictive performance, interpretability, and clinical relevance. This model includes only 9 variables and achieves high performance metrics with reduced complexity:

Model	**AUC**	**Best Threshold**	**Accuracy**	**Sensitivity**	**Specificity**
Logistic (reduced)	0.85	0.11	0.76	0.83	0.75

Table 3: Performance comparison of the final simplified GLM model on the test set.

The variables were selected based on both statistical significance and clinical relevance in predicting diabetes risk. The coefficients from the logistic regression model offer meaningful, interpretable insights:

- **Age (`AGEP_A`)**: Each additional year increases the log-odds of being diagnosed with diabetes, reflecting the natural decline in insulin sensitivity as people age.
- **Weight (`WEIGHTLBTC_A_log`)**: A higher log-transformed weight significantly increases diabetes risk. For instance, going from 70kg to 90kg (a 28% increase) corresponds to a large rise in predicted probability.
- **Health status (`PHSTAT_A`)**: Individuals who self-report "fair" or "poor" health show progressively higher risk. There is a clear dose-response pattern: the worse the self-perceived health, the higher the diabetes probability.
- **Recent eye exam (`AVISEXAM_A`)**: Having had an eye exam within the past year is associated with higher diabetes diagnosis, likely due to detection of related complications such as retinopathy.
- **Chronic conditions**: Absence of coronary heart disease (`CHDEV_A2`), high cholesterol (`CHLEV_A2`), and hypertension (`HYPEV_A2`) all act as strong protective factors, consistent with the metabolic links between these conditions and diabetes.

- **Poverty ratio (`POVRATTC_A_log`)**: A higher poverty ratio, indicating better socioeconomic conditions, significantly reduces diabetes risk, underlining the impact of income inequality on health.
- **Ethnicity (`HISPALLP_A`)**: The risk of diabetes varies across ethnic groups, with some—such as non-Hispanic White—showing statistically lower odds compared to the Hispanic reference group.

The final model enables concrete reasoning about risk mitigation. Consider the following examples:

- If a person improves their self-rated health from "poor" to "fair," the model predicts a notable reduction in diabetes risk.
- A weight loss from 90kg to 70kg corresponds to a sharp decline in predicted risk due to the log transformation emphasizing proportional change.
- An increase in the poverty ratio—reflecting improved financial status—also leads to lower estimated diabetes probability.

7 Classification Model Comparison: GLM, LDA, Naive Bayes, and KNN

To ensure a fair and consistent evaluation, all models were trained and tested using the same set of 9 predictors. The goal was to assess whether alternative classification techniques could improve predictive performance beyond that of the generalized linear model (GLM) with binomial family.

Model	**AUC**	**Accuracy**	**Sensitivity**	**Specificity**
Logistic Regression (GLM)	0.86	0.76	0.83	0.75
Linear Discriminant Analysis (LDA)	0.85	0.73	0.84	0.72
Naive Bayes	0.85	0.75	0.82	0.74
K-Nearest Neighbors (k = 20)	0.81	0.70	0.80	0.69

Table 4: Performance comparison of classification models using the same 9-variable input. Metrics are rounded to two decimal places.

Main Observations:

- **Logistic Regression** (GLM) provides the most balanced performance, achieving the highest accuracy and a strong AUC.
- **LDA** attains slightly higher sensitivity but at the cost of lower specificity, which may lead to more false positives.
- **Naive Bayes** is competitive, especially in recognizing true positives, with robust AUC and balanced metrics.

- **KNN** shows the weakest performance overall. Its sensitivity to the curse of dimensionality and the unbalanced class distribution (with only ∼10% positive cases) likely contributes to this underperformance.

Among all classifiers tested, the GLM-based logistic regression model emerges as the most effective and reliable choice. It not only maintains a high level of interpretability but also delivers superior performance metrics, confirming its suitability for predictive diagnostics in the context of diabetes classification.

8 Ridge Model

Ridge regression was also applied as an alternative regularization technique.

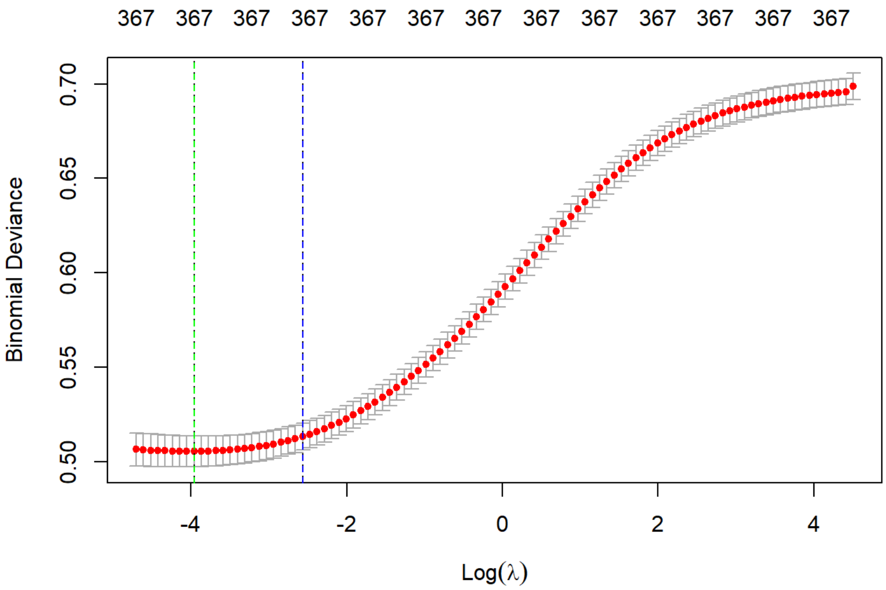


Figure 4: Cross-validation plot for Ridge. The green dashed line indicates `lambda.min`, while the blue dashed line shows `lambda.1se`. The deviance curve is smooth, and the difference between the two lambdas is minimal, making either a viable choice.

Model	AUC	Best Threshold	Accuracy	Sensitivity	Specificity
Ridge_lambda.min	0.86	0.10	0.75	0.86	0.73
Ridge_lambda.1se	0.85	0.11	0.75	0.84	0.74

Table 5: Performance comparison of Ridge models on the test set.

Both Ridge models achieved strong predictive performance. Ridge_lambda.min slightly favors sensitivity, making it well suited for minimizing false negatives. However, Ridge_lambda.1se offers a more balanced trade-off between sensitivity and specificity.

9 Classification Model Comparison

To evaluate the performance of different classification techniques, we compared multiple models trained on different subsets of predictors.

- The **GLM (Full Model)** and **LASSO models** (`lambda.min`, `lambda.1se`) were trained using all 31 variables selected by LASSO with non-zero coefficients.
- The **Reduced GLM**, as well as **Naive Bayes**, **LDA**, and **KNN**, were trained on a simplified set of 9 predictors. These were chosen based on clinical relevance, socioeconomic interpretability, and statistical robustness.
- The **Ridge models** used their own optimized variable encoding based on the regularization path and may include one-hot encoded forms, potentially differing from LASSO.

The table below summarizes the test set performance of all models:

Model	**AUC**	**Accuracy**	**Sensitivity**	**Specificity**
GLM (Full Model)	0.87	0.75	0.86	0.73
GLM (Reduced)	0.86	0.76	0.83	0.75
LASSO_lambda.min	0.86	0.74	0.86	0.72
LASSO_lambda.1se	0.86	0.73	0.86	0.71
Ridge_lambda.min	0.86	0.75	0.86	0.73
Ridge_lambda.1se	0.85	0.75	0.84	0.74
Naive Bayes	0.85	0.75	0.82	0.74
LDA	0.85	0.73	0.84	0.72
KNN (k=20)	0.81	0.70	0.80	0.69

Table 6: Performance comparison of all classification models. Metrics are rounded to two decimal places.

Conclusion: Logistic regression models—both the full and the reduced version—demonstrate superior and more consistent performance compared to all other classifiers tested. In particular, the **reduced GLM model**, which uses only 9 carefully selected variables, achieves high accuracy (76%), excellent sensitivity (83%), and strong AUC (0.86). This makes it not only effective but also highly interpretable, a critical aspect for clinical applications.

Regularized approaches such as **LASSO** and **Ridge** regression also performed competitively, validating the utility of penalized models in high-dimensional settings. These methods effectively handle one-hot encoded variables by automatically excluding non-informative levels, contributing to model parsimony and stability.

In contrast, models such as **KNN**, **Naive Bayes**, and **LDA** showed weaker or unstable performance. This is likely due to the unbalanced nature of the dataset (with only about 10% positive cases) and the complexity introduced by high-dimensional and non-linear relationships among features.