881177 SIL Project - Diabetes Prediction

Victoria Grosu May 2025

Introduction to the Data

Dataset (2023 survey):

- 440 columns
- 68% with missing values
- Avg. missingness per column: 52%
- Only 32% of columns fully complete

Missing Data Pattern:

- Not missing at random (NMAR)
- Follows survey logic (e.g., follow-ups only shown after "Yes" responses)

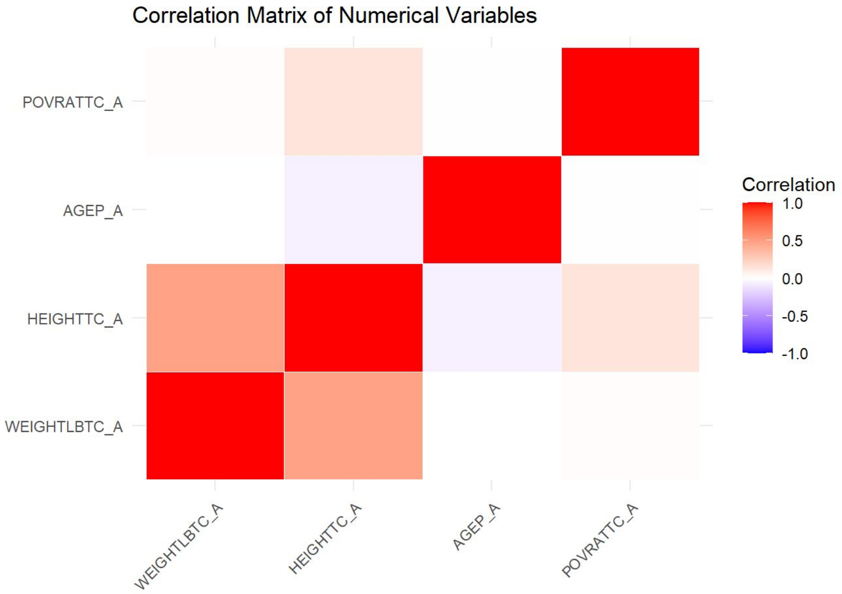
Data Cleaning:

- Removed columns with any missing data
- Dropped uninformative or ID-like columns
- Excluded survey-explanatory fields
- Handling Perfect Separation in Categorical Predictors

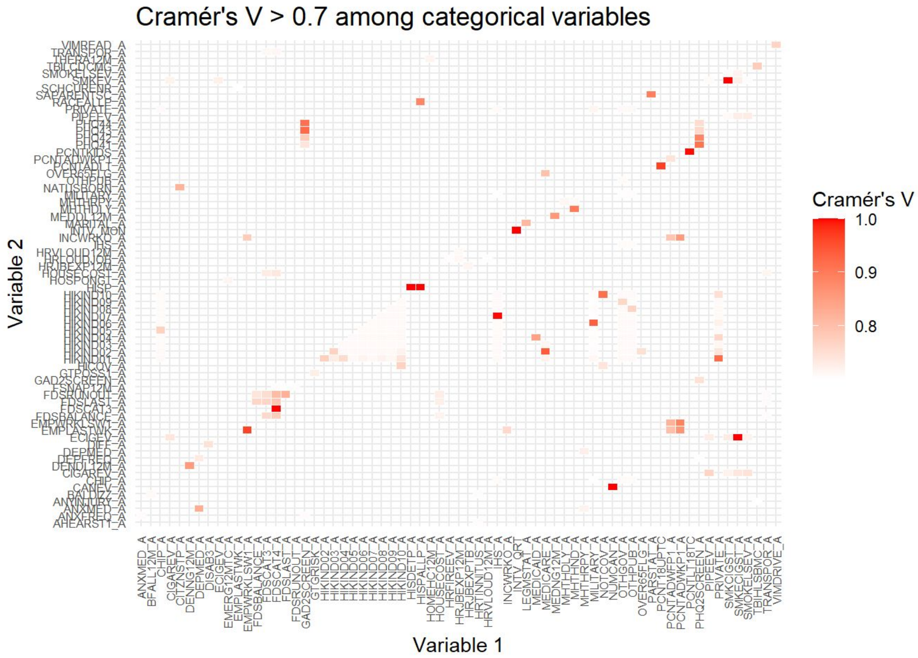
Result:

- Variables reduced from 440 → ~200

Correlation and Collinearity Analysis

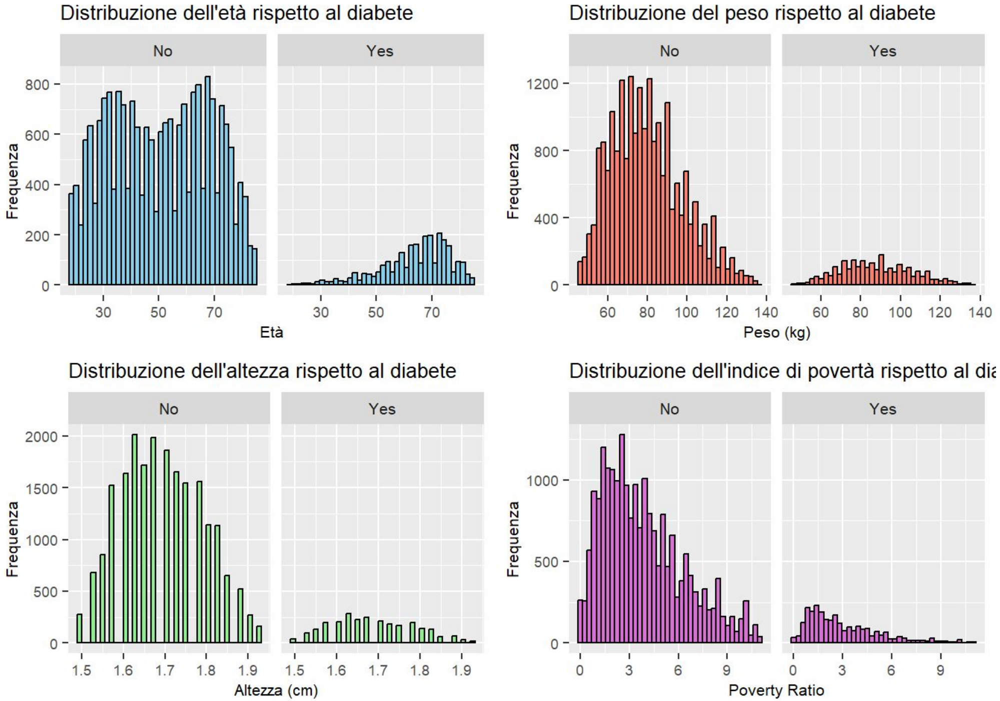


Correlation Matrix of Numerical Variables

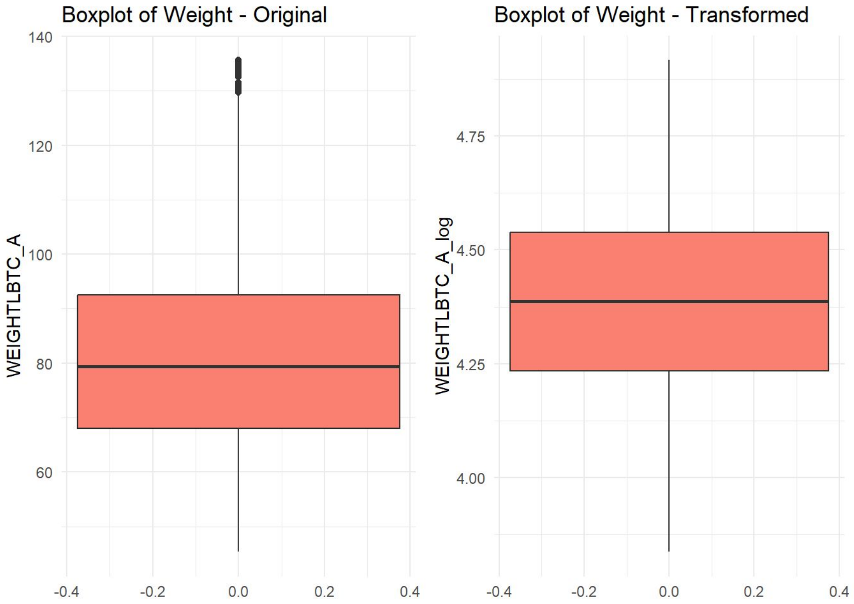


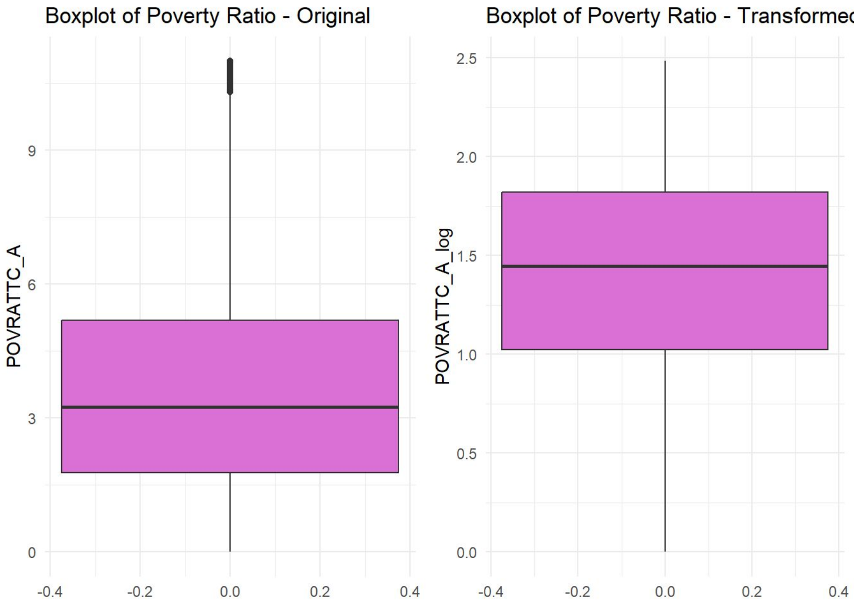
Cramér's V > 0.7 among categorical variables

Exploratory Analysis of Numeric Predictors

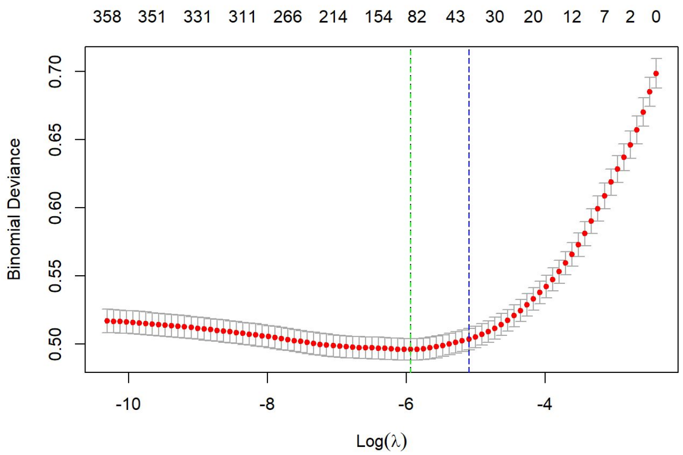


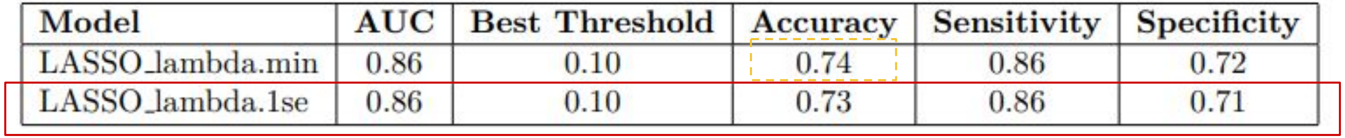
Log Transformation of Numeric Variables





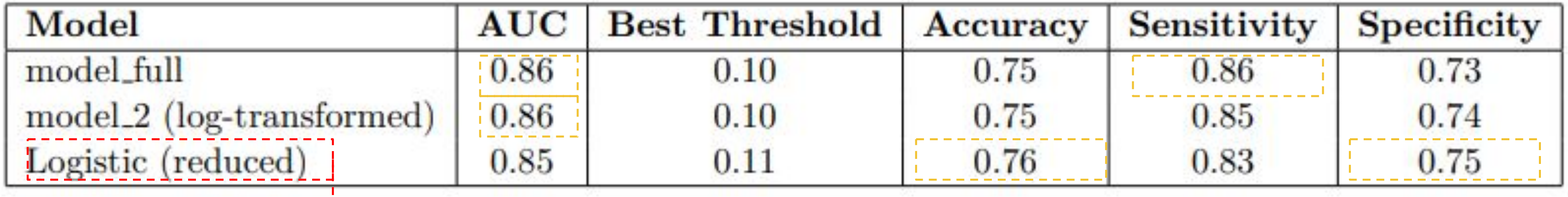
LASSO Model





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

Logistic Regression Modeling



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

- **Ethnicity (HISPALLP_A)**
  Ref: Hispanic.
  Non-Hispanic White and AIAN → **Lower risk**
  Non-Hispanic Asian → **Higher risk**

- **Age (AGEP_A)**
  Older age → **Higher risk**

- **Eye Exam Timing (AVISEXAM_A)**
  Ref: Never.
  Recent/Unknown exam → **Higher risk**

- **No CHD (CHDEV_A2)**
  Ref: Diagnosed CHD.
  No CHD → **Lower risk**
- **No High Cholesterol (CHLEV_A2)**
  Ref: Diagnosed.
  No cholesterol → **Lower risk**

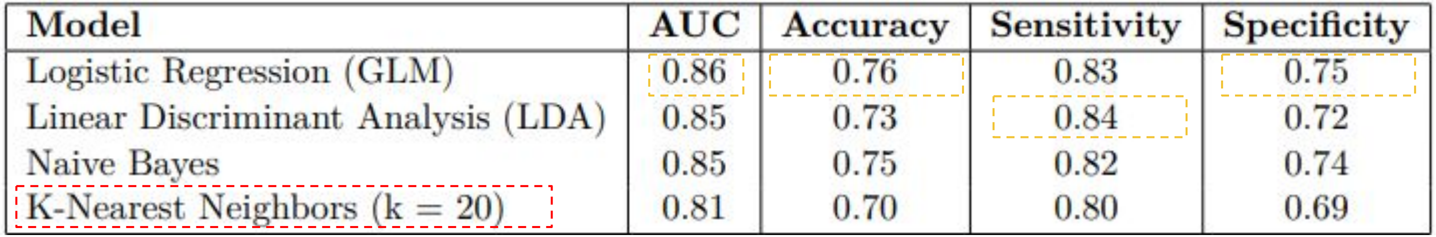
- **No Hypertension (HYPEV_A2)**
  Ref: Diagnosed.
  No hypertension → **Lower risk**

- **Self-Rated Health (PHSTAT_A)**
  Ref: Excellent.
  Poorer health → **Higher risk** (dose-response)

- **Log-Weight (WEIGHTLBTC_A_log)**
  Higher weight → **Higher risk**

- **Poverty Ratio (POVRATTC_A_log)**
  Higher income → **Lower risk**

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

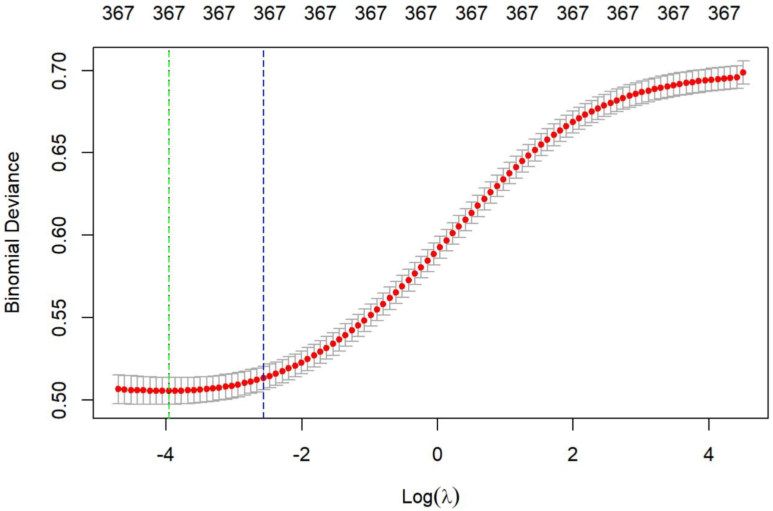


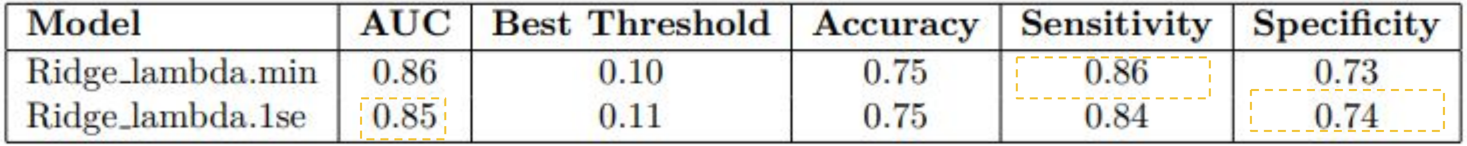
Model	AUC	Accuracy	Sensitivity	Specificity
Logistic Regression (GLM)	0.86	0.76	0.83	0.75
Linear Discriminant Analysis (LDA)	0.85	0.73	0.84	0.72
Naive Bayes	0.85	0.75	0.82	0.74
K-Nearest Neighbors (k = 20)	0.81	0.70	0.80	0.69

--- Confusion Matrix: KNN ---

	No	Yes
No	6351	795
Yes	0	0

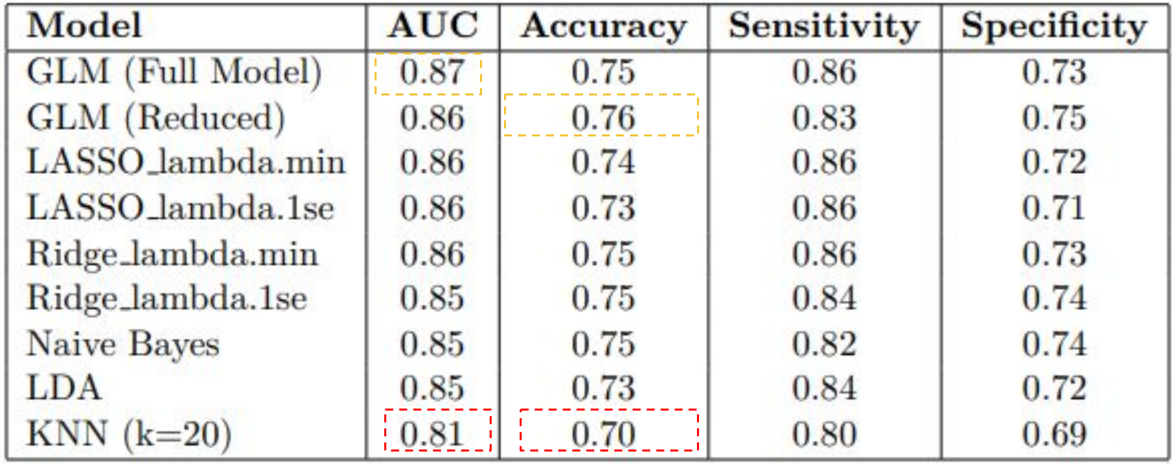
Ridge Model





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

Classification Model Comparison



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

Best Model: GLM (Reduced)

Worst Model: KNN (k = 20)