DSA1101 Statistical Report

JING JIAYUN

November 06, 2025

Part 0: Introduction

This report aims to analyze a survey dataset of 100,000 observations to predict diabetes status. We begin with EDA to understand variable–outcome relationships. Next, we use a 5-fold cross-validation on a stratified 10% sample, and then evaluate the best versions on the full dataset using ROC, TPR, Precision, and AUC. Finally, we recommend the most suitable classifier for a medical screening context.

Part I: EDA - Exploring Variables and Associations

1.1 Variable Description

Response Variable

- `diabetes`: A binary variable (`1` = diabetic, `0` = non-diabetic).
  Only **8.5%** of individuals have diabetes, indicating a **strong class imbalance** that must be considered during modeling.

Input Variables (Predictors)

Demographic Information

- `gender`: Categorical variable for gender.
- `age`: Numerical variable for age in years.

Medical History & Lifestyle

- `hypertension`: Binary indicator for hypertension.
- `heart_disease`: Binary indicator for heart disease.

- **`smoking_history`**: Categorical variable for smoking status.

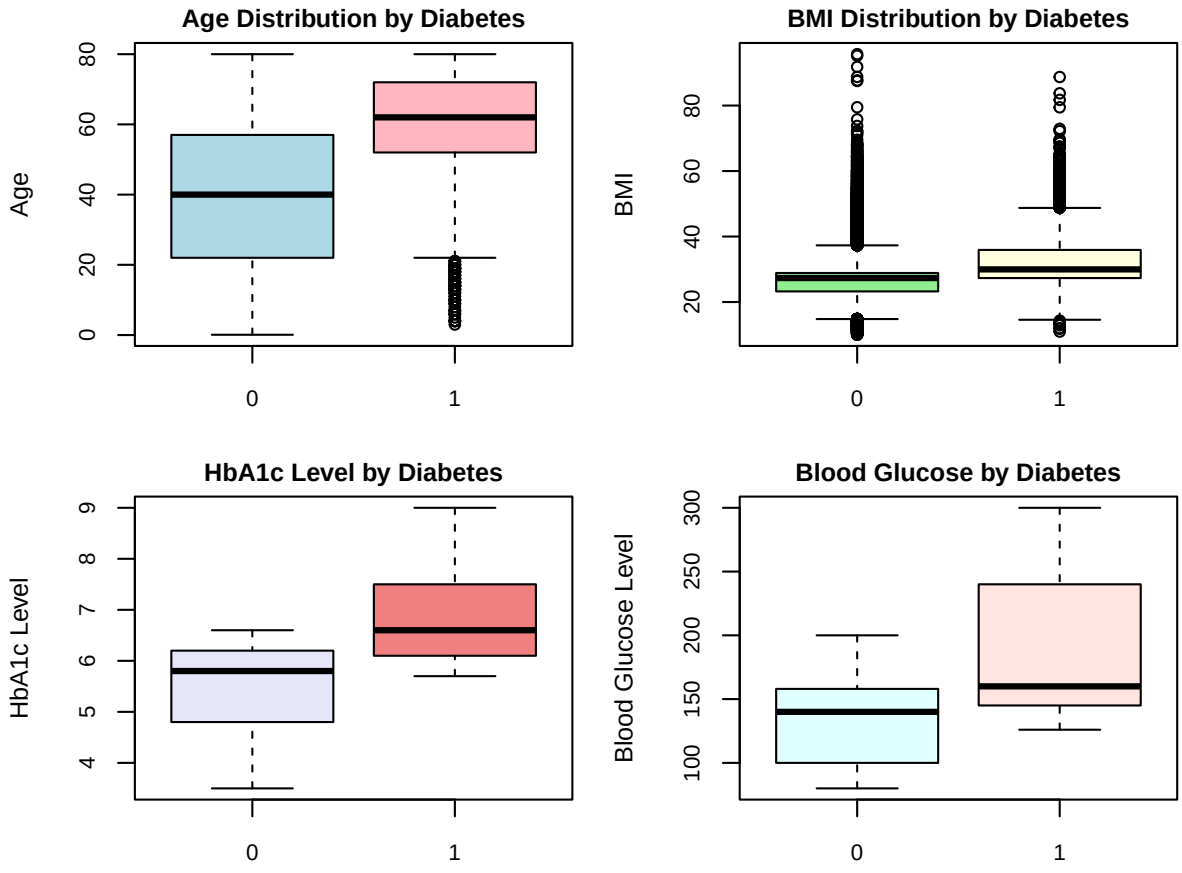
Clinical Measurements

- **`bmi`**: Body Mass Index.
- **`HbA1c_level`**: Hemoglobin A1c level.
- **`blood_glucose_level`**: Current blood glucose concentration.

1.2 Analysis of Numerical Variables (Boxplots):

Table 1: Association of Numerical Variables with Diabetes

Variable	Association Strength	Key Observations
Age	Strong	Median: 60 (diabetic) vs 40 (non-diabetic). Distributions are slightly higher for diabetics, but there is a large overlap.
BMI	Moderate-Weak	Median slightly higher for diabetics, but there is a large overlap.
HbA1c_level	Very Strong	Distributions are clearly separated; IQR(non-diabetic) lies well below IQR(diabetic).
Blood_Glucose_Level	Very Strong	Similar to HbA1c. IQRs show little to no overlap.



1.3 Analysis of Categorical Variables (Contingency Tables)

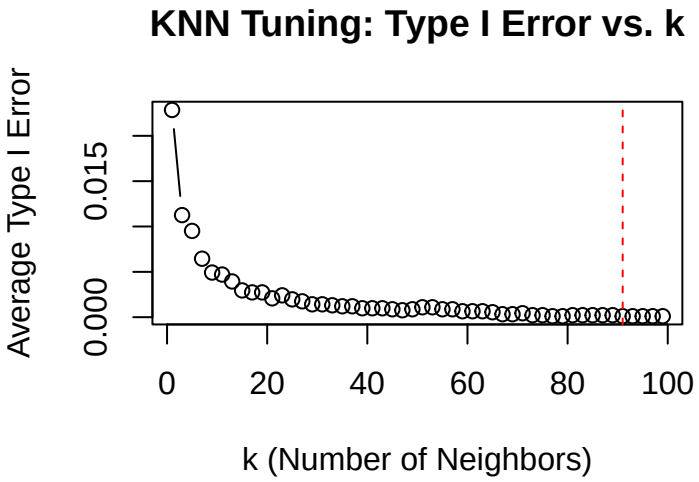
Table 2: Association of Categorical Variables with Diabetes

Variable	Association_Strength	Key_Observations
Gender	Very Weak	9.7 percent (males) vs 7.6 percent (females).
Hypertension	Very Strong	27.9 percent (Yes) vs 6.9 percent (No).
Heart Disease	Very Strong	32.1 percent (Yes) vs 7.5 percent (No).
Smoking History	Moderate	"Former" smokers (17.0 percent) show the highest rate, while "No Info" (4.1 percent) is lowest.

Part II: Methods: Building KNN, DT and LR Classifiers

2.1 K-Nearest Neighbors (KNN)

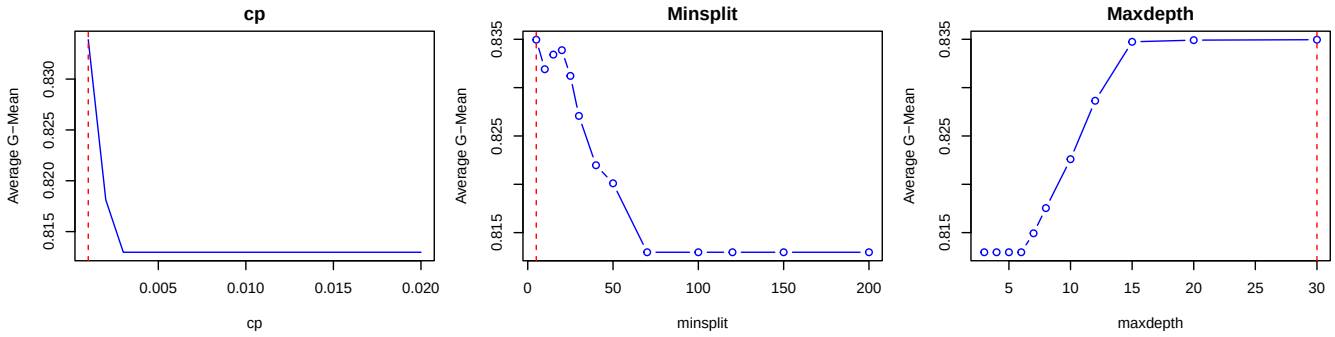
Before running the KNN model, the non-numerical variables *Gender* (very weak association) and *Smoking History* (moderate association) were removed. To keep the model simple and avoid the high dimensionality that *one-hot encoding* would introduce, we used only the original numerical predictors.Using 5-fold cross-validation, we tuned KNN over odd k values from 1 to 100 using **Type I Error** as the criterion.
The lowest average Type I Error occurred at $\mathbf{k = 91}$, indicating the best model.



2.2 Decision Tree (DT)

We first tuned the Decision Tree by minimizing the **cross-validated Type I Error**, but this caused a trivial classifier that always predicted the majority class due to severe imbalance (only 8.5 % positive cases). To address this, we switched to optimizing **the Geometric Mean (G-Mean)**,

which balances sensitivity and specificity. This approach yielded the optimal hyperparameters: **cp = 0.001, minsplit = 5, and maxdepth = 30**.



2.3 Logistic Regression (LR)

The logistic regression model was refined using **backward stepwise selection**, removing **variables with high p-values** (Gender and Smoking History). The final model maintains similar Type I Error performance while improving overall statistical significance and simplicity.

Because the dataset is imbalanced, we optimized the classification threshold from **0.5 to 0.07** to improve the detection of diabetic cases. This adjustment increased sensitivity from **0.63 to 0.88**, with lower specificity and precision as a trade-off. In medical contexts, where **missing a true case is riskier than generating a false alarm**, the optimized model at threshold 0.07 provides a more suitable balance for early diabetes prediction.

Table 3: Comparison of LR Models (Type I Error)

Model	Type1_Error
Full	0.0091
Model2	0.0095
Best	0.0093

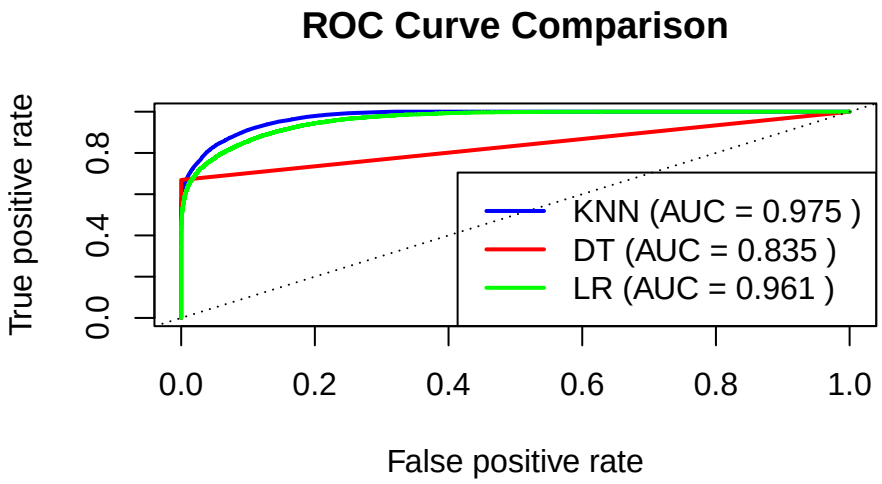
Table 4: Best LR Model: Default vs Optimal Thresholds

	Accuracy	Sensitivity	Specificity	Type1_Error	Precision
LR with Default Threshold (0.5)	0.9604	0.6341	0.9907	0.0093	0.8638
LR with Optimal Threshold (0.07)	0.8731	0.8847	0.8720	0.1280	0.3911

2.4 Model Performance Comparison on Full Data Set

After tuning the hyperparameters, the best KNN, DT, and LR models were tested on the full dataset.

a) ROC Curve and AUC



All three models performed substantially better than random guessing, as their ROC curves lie close to the top-left corner. KNN and LR perform almost identically and slightly better than DT in terms of AUC.

b) True Positive Rate and Precision

Table 5: Final Performance of KNN, DT, and LR

Model	AUC	TPR	Precision
K-Nearest Neighbors	0.9751347	0.5840000	0.9821923
Decision Tree	0.8345294	0.6690588	1.0000000
Logistic Regression	0.9607036	0.8894118	0.3918520

KNN: Detects **58.40%** of diabetic cases but achieves **very high precision (98.22%)**, indicating it's reliable in positive predictions, though it misses many true diabetics.

DT: Reaches **66.91% sensitivity** with **perfect precision (100.00%)**. It misses about one-third of diabetics but makes **no false positive predictions**.

LR: The optimized model, evaluated on the full dataset, identifies **88.94%** of diabetic cases (the highest TPR) but with **lower precision (39.19%)**, meaning it tends to flag more false positives.

2.5 Comments on Pros and Cons of Each Classifier

Classifier	Pros	Cons
K-Nearest Neighbors (KNN)	**High Precision (98.22%):** Positive predictions are extremely reliable, useful for confirmatory diagnosis. **Model Simplicity:** Conceptually simple and non-parametric.	**Poor Sensitivity (58.40%):** Fails to identify over half of diabetic cases, making it unsuitable for initial screening. **Computational Cost:** Expensive for large datasets (100,000 observations). **Handles Only Numerical Data:** Required exclusion of categorical variables like `smoking_history`.
Decision Tree (DT)	**Interpretability:** Rules are easy for medical professionals to understand and validate. **Handles Mixed Data Types:** Natively uses both numerical and categorical predictors. **Balanced Performance:** Achieved a good balance between sensitivity (66.91%) and precision (100.00%).	**Lower Sensitivity than LR:** Misses more diabetic cases compared to the optimized LR model. **Instability:** Small changes in data can lead to a completely different tree structure.
Logistic Regression (LR)	**Highest Sensitivity (88.94%):** Best at identifying true diabetic cases, which is the primary goal of screening. **Probabilistic Output & Flexibility:** Allows threshold adjustment for clinical needs. **Interpretability of Coefficients:** Coefficients reveal the strength and direction of each risk factor.	**Low Precision (39.19%):** High sensitivity comes at the cost of many false positives. **Linearity Assumption:** Assumes a linear relationship between predictors and the log-odds of the outcome.

Part III Conclusion: The Best Model/Classifier

After comparing the **K-Nearest Neighbors (KNN)**, **Decision Tree (DT)**, and **Logistic Regression (LR)** models, the **optimized Logistic Regression (LR)** model is the **best classifier** for this screening task.

The main goal of a screening model is **high sensitivity**, detecting as many true diabetic cases as possible. Missing a diabetic patient (a *false negative*) has more serious consequences than flagging a healthy person (*false positive*).

The **Logistic Regression model**, with a **threshold set to 0.07**, achieved the **highest sensitivity (88.94%)**, correctly identifying *nearly nine out of ten* diabetic individuals. Although this came with a **lower precision of 39.19%**, the trade-off is acceptable for an *initial screening stage*, where flagged individuals can later undergo follow-up diagnostic tests.

In comparison:

KNN showed **extremely high precision (98.22%)** but **low sensitivity (58.40%)**, missing over half the diabetic cases.

DT achieves **66.91% sensitivity and 100% precision**, producing no false positives. This makes it suitable as **a secondary confirmatory model**, not as the primary screener.

Overall, the **Logistic Regression model** offers the best balance for medical screening: it **detects the majority of true cases**, allows **flexible threshold adjustment** for clinical needs, and remains **highly interpretable**.

Hence, it is the **most effective and responsible choice** for diabetes prediction in this dataset.