DSA1101 Assignment 2: Diabetes Prediction
AY25/26 Sem 1

Ng Yao Teck

December 2, 2025

1 Introduction

The global average proportion of people with diabetes is about 11%, and needs insulin to regulate their blood glucose level. It is paramount for a person to be screened for diabetes in order to be prescribed insulin, and to prevent further complications if detected early. Using the dataset diabetes-dataset.csv of 100000 people, the article will analyse the data given, performance and suitability of 3 different machine learning models in predicting diabetes.

2 Exploratory Data Analysis

The dataset of 100000 data has 9 variables:

Table 1: Dataset features and their characteristics

Feature	**Type**	**Classification**	**Description**
gender	Categorical	Nominal	Female, Male, Other
age	Numerical	Continuous	Range: 0.08–80 years
hypertension	Categorical	Binary	0 = No, 1 = Yes
heart_disease	Categorical	Binary	0 = No, 1 = Yes
smoking_history	Categorical	Nominal	never, former, current, ever, not current, No Info
bmi	Numerical	Continuous	Body Mass Index (kg/m^2)
HbA1c_level	Numerical	Continuous	Hemoglobin A1c level (%)
blood_glucose_level	Numerical	Continuous	Blood glucose level (mg/dL)
diabetes	Categorical	Binary	0 = No, 1 = Yes (Response Variable)

Noted that HbA1c level might be related related to blood glucose level as it is the average blood glucose level in the past 2 to 3 months. Furthermore, there is a data with age = 0.08 year.

From the data, it shows median age of 43 years (range: 0.08–80), median BMI of 27.32 kg/m^2 (10.01–95.69), average HbA1c of 5.53% (3.5–9.0%), and median blood glucose of 140 mg/dL (80–300). Hypertension affects 7.49% and heart disease 3.94%. Gender distribution is 58.55% female, 41.43% male, and 0.02% other. Figures 1 and 2 show diabetic individuals have higher HbA1c (≈6.5% vs 5.5%) and blood glucose (≈200 vs 140 mg/dL), with BMI outliers, and highest diabetes proportion among those with heart disease (32.1%) and hypertension (27.9%).

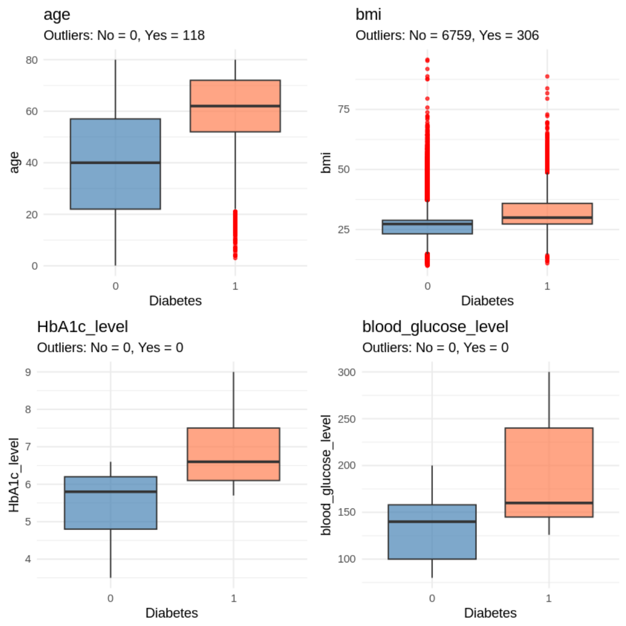


Figure 1: Distribution by diabetes status

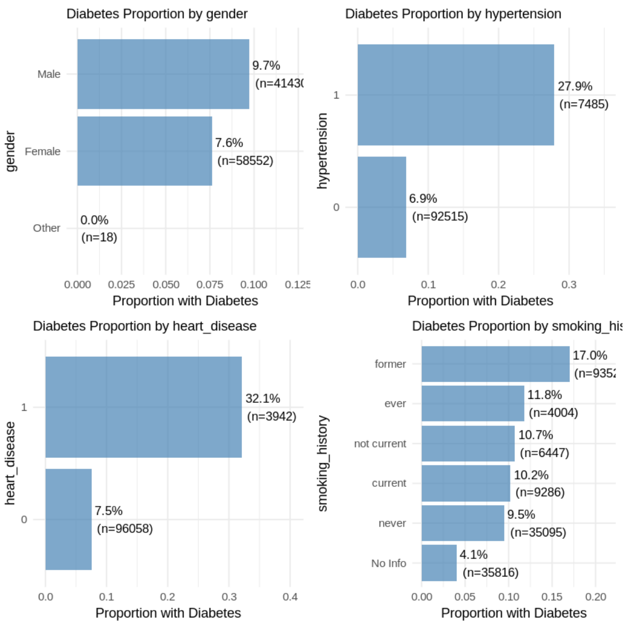
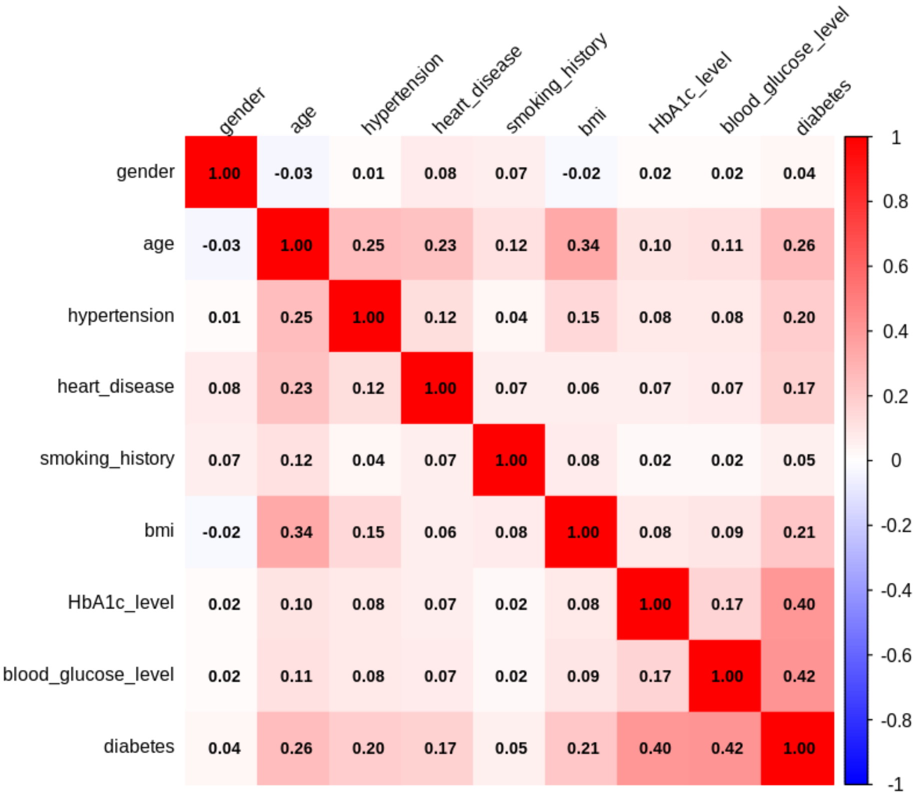


Figure 2: Diabetes proportion by category

Table 1 shows blood glucose level has the strongest association (Pearson's $r = 0.42$), followed by HbA1c level ($r = 0.40$), both moderate correlations. Age ($r = 0.26$) and BMI ($r = 0.21$) show weaker associations. The correlation heatmap (Figure 3) reveals all variables are significantly associated with diabetes, with metabolic markers showing strongest effects.

Table 2: Association between predictor variables and diabetes status

Variable	**Test**	**Statistic**	**Effect**	**Strength**
blood_glucose_level	Pearson r	0.420	0.420	Moderate
HbA1c_level	Pearson r	0.401	0.401	Moderate
age	Pearson r	0.258	0.258	Weak
bmi	Pearson r	0.214	0.214	Weak
hypertension	Chi-square	3910.71	0.198	Weak
heart_disease	Chi-square	2945.85	0.172	Weak



	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
gender	1.00	-0.03	0.01	0.08	0.07	-0.02	0.02	0.02	0.04
age	-0.03	1.00	0.25	0.23	0.12	0.34	0.10	0.11	0.26
hypertension	0.01	0.25	1.00	0.12	0.04	0.15	0.08	0.08	0.20
heart_disease	0.08	0.23	0.12	1.00	0.07	0.06	0.07	0.07	0.17
smoking_history	0.07	0.12	0.04	0.07	1.00	0.08	0.02	0.02	0.05
bmi	-0.02	0.34	0.15	0.06	0.08	1.00	0.08	0.09	0.21
HbA1c_level	0.02	0.10	0.08	0.07	0.02	0.08	1.00	0.17	0.40
blood_glucose_level	0.02	0.11	0.08	0.07	0.02	0.09	0.17	1.00	0.42
diabetes	0.04	0.26	0.20	0.17	0.05	0.21	0.40	0.42	1.00

Figure 3: Correlation heatmap of all variables

3 Classification Methods

3.1 Hyperparameter Tuning

Since the class ratio matches similarly to the global diabetes percentage, no under or over sampling methods were used. Type 1 error is to be minimised as false diabetes diagnosis will require the purchase of insulin and to live in a more restrictive lifestyle.

Since there are 100000 data, a 10% stratified subset 5 cross fold validation is used as there is not much loss of information while ensuring representation of diabetes in each fold, and to save computational power.

3.2 KNN

To reduce the dimensionality of the data and computational complexity, nominal responses and smoking history were removed. Variable weights were introduced to account for their contribution to diabetes statistics. For simplicity sake, weights are based on effect value in Table 2 after normalisation.

To save computational power and avoid tie breakers, tested odd $k = 1$–149 with weighted/unweighted schemes.

Optimal: weighted KNN with $k = 31$ (Type I Error $= 0.0001$), while offering efficiency with fewer neighbors (Figure 4).

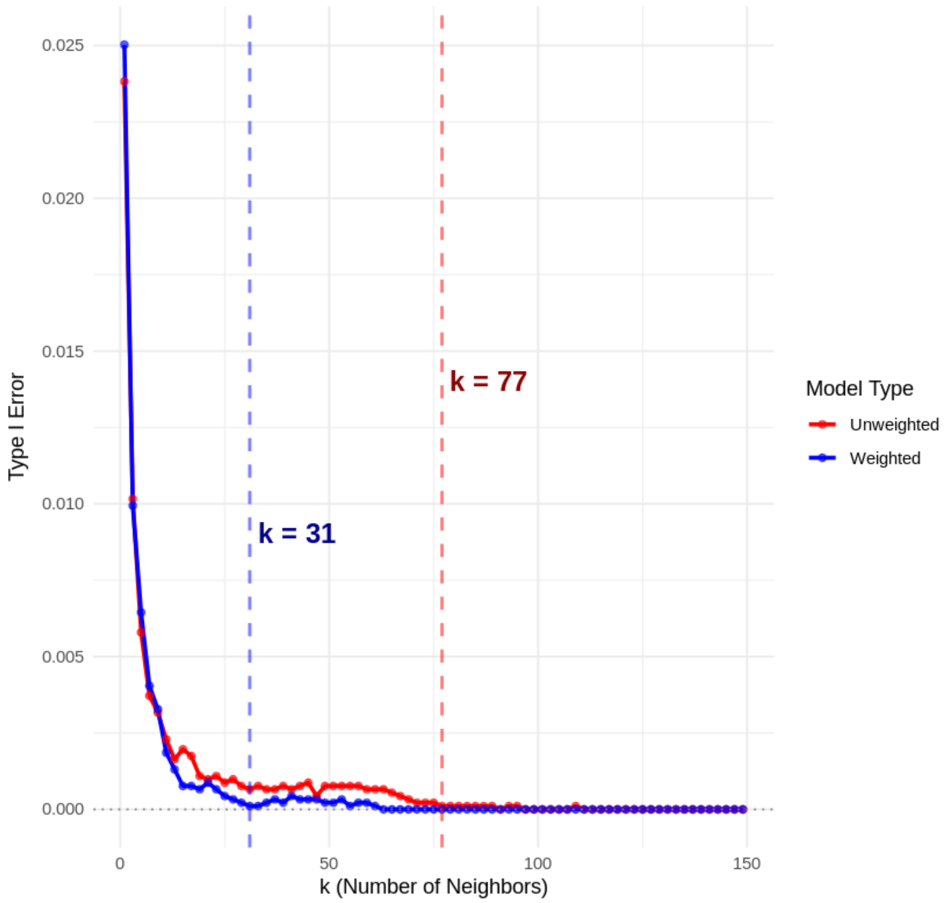


Figure 4: KNN Type I Error comparison

3.3 Decision Tree

Tested max depth 2–20. Optimal depth = 2 with zero Type I error (Figure 5). The tree (Figure 6) uses HbA1c $\geq$ 6.7% as primary split and blood glucose $\geq$ 210 mg/dL as secondary, aligning with clinical criteria for the diagnosis of diabetes.

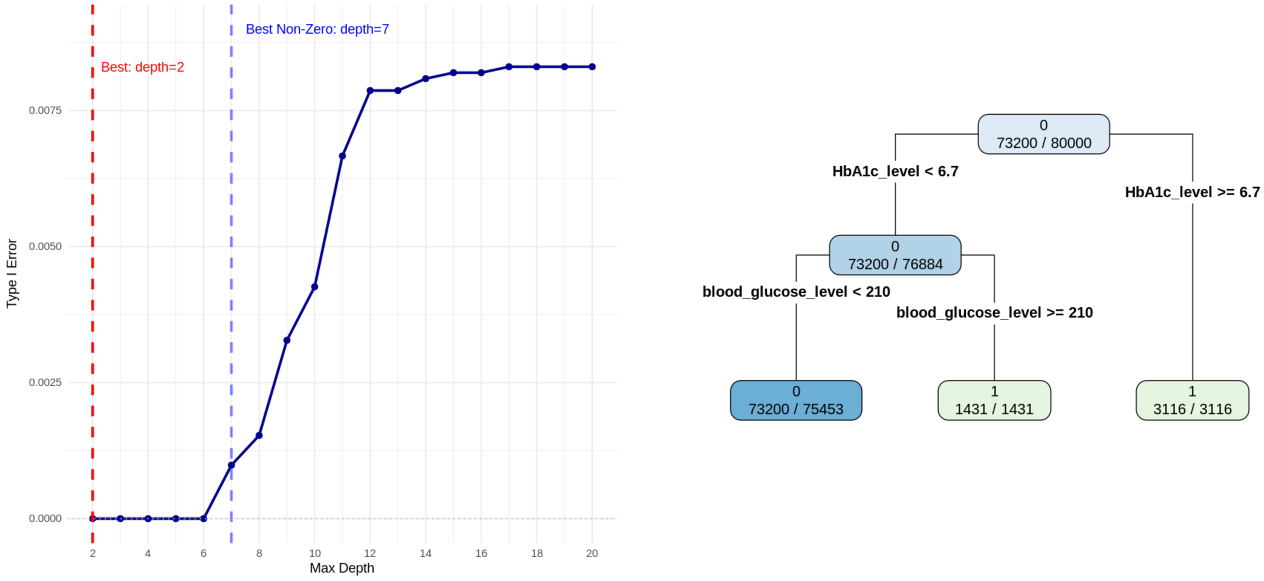


Figure 5: Decision Tree optimization (left) and structure (right)

3.4 Logistic Regression

To account for HbA1c and blood glucose level relationship, and age bmi correlation in Table 2, 4 different models were tested at different threshold.

- $HbA1c\_level + blood\_glucose\_level + age + bmi + hypertension$
- $HbA1c\_level \times blood\_glucose\_level + age + bmi + hypertension$
- $HbA1c\_level + blood\_glucose\_level + age \times bmi + hypertension$
- $HbA1c\_level \times blood\_glucose\_level + age \times bmi + hypertension$

The optimal relationship is $HbA1c\_level \times blood\_glucose\_level + age \times bmi + hypertension$ at threshold = 0.9.

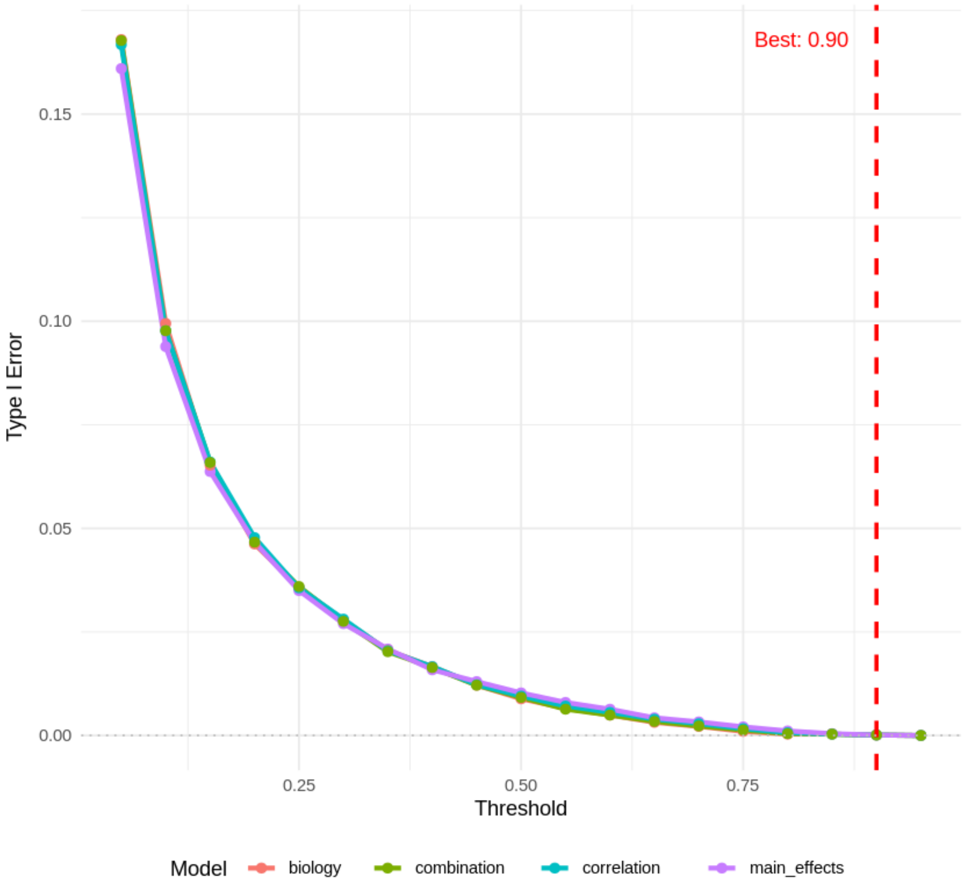


Figure 6: Logistic Regression Type I Error vs threshold

4 Model Evaluation

To evaluate all of the model's performance, a 80/20 train test split is conducted, using the hyperparameter found above, to receive their performance metrics in Table 2.

KNN achieves highest TPR (96.5%) and AUC (0.965) with excellent precision (98.8%) and minimal FPR (0.1%). Decision Tree shows perfect precision (100%) but lower TPR (83.5%), being most conservative. Logistic Regression balances performance with TPR 95.9% and AUC 0.959. Figure 8 shows KNN and LR have superior ROC curves.

Table 3: Performance comparison on full dataset

Method	**TPR**	**FPR**	**Precision**	**AUC**
KNN ($k = 31$, weighted)	0.965	0.001	0.988	0.965
Decision Tree (depth=2)	0.835	0.000	1.000	0.835
Logistic Regression	0.959	0.002	0.980	0.959

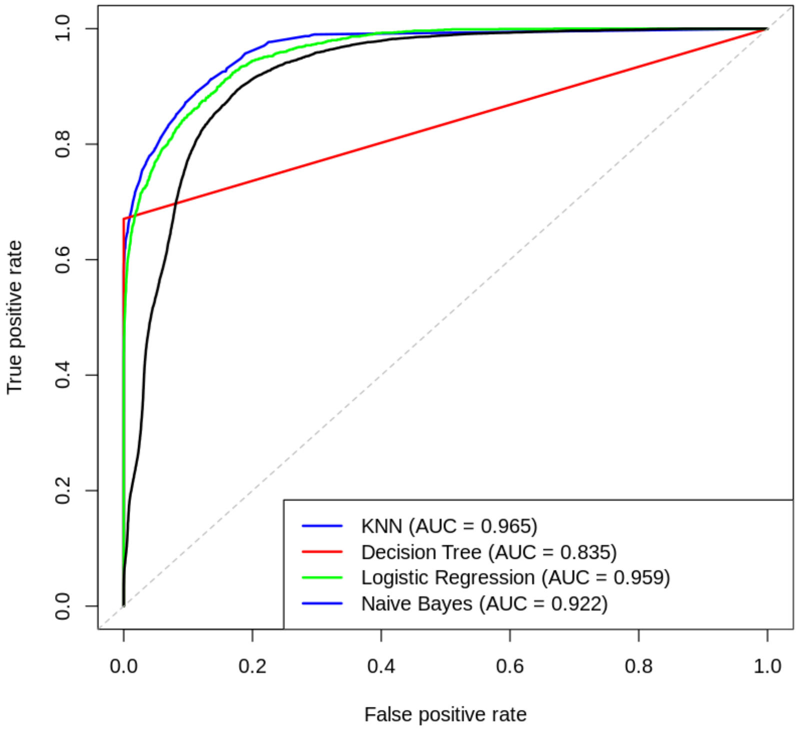


Figure 7: ROC curves comparison for all methods

5 Conclusion

Pros and Cons: KNN does well in discrimination but is computationally intensive and sensitive to scaling. Decision Tree offers interpretability and clinical validity but has lower sensitivity. Logistic Regression provides balanced performance with probabilistic outputs and coefficient interpretability but requires a suitable linearised model.

Best Classifier: Based on the metrics, KNN is indeed the best model for prediction diabetes. However, there might be a Logistic Regression equation that offers better metrics than KNN.

6 AI Declaration

AI use: I have used AI to search up the implications of diabetes, lifestyle changes after diagnosis and cost+benefit of false diagnosis of diabetes.

AI use in R code: I have AI to debug continuous data analysis, some ggplot and association analysis + to fit multiple plots in 1 window + improve aesthetics of plots.

AI use in LaTeX code: I have used AI to debug my code, format side by side figures (1+2 and 5), and setting vertical spaces.