Abstract

This project replicates and extends a prior R-based statistical analysis by implementing a full end-to-end machine learning pipeline in Python to predict diabetes risk. The dataset is first explored through visualisation and preprocessing to understand feature behaviour and class imbalance. Hyperparameter tuning and structured model selection are then conducted to optimise predictive performance.

Four supervised learning models, K-Nearest Neighbour (KNN), Decision Tree (DT), Logistic Regression (LR), and Random Forest (RF), are evaluated using a diabetes dataset comprising 100,000 individuals with medical and demographic attributes. The DT, LR, and RF models are optimised using 5-fold cross-validation, while KNN is tuned through a heuristic-based search over neighbourhood sizes. Model performance is assessed using six metrics: Accuracy, Precision, Recall, False Positive Rate (FPR), F1 score, and Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

The results indicate that Logistic Regression is the most suitable model for diabetes screening, achieving the lowest False Negative Rate of 0.146 and the highest Recall of 0.854. These metrics are prioritised due to the high cost of missed diagnoses in a healthcare context. In contrast, Random Forest achieved the highest Accuracy of 0.970 and the largest ROC-AUC value of 0.967, reflecting superior overall predictive power. This highlights the trade-off between interpretability and raw performance in medical risk prediction.

Introduction

Diabetes mellitus is a prevalent metabolic disorder with significant global health implications. Early identification of individuals at risk enables timely lifestyle interventions and clinical follow-ups, potentially reducing the risk of chronic complications such as cardiovascular disease, renal failure, and neuropathy. Conventional screening approaches are costly, requiring specialized tests and equipment. Machine learning offers a cost-effective alternative for risk stratification and disease screening by leveraging large-scale medical and demographic data.

Recent literature highlights the growing role of machine learning in improving diagnostic accuracy and supporting clinical decision-making across diverse populations (Khokhar et al., 2025; Tan et al., 2021). In healthcare, sensitivity (true positive rate) and false negative rates are prioritised over overall accuracy, as failing to identify a patient with the disease carries significantly higher consequences than false positives (Luo et al., 2024).

The dataset used in this study, provided by Mohammed Mustafa on Kaggle, consists of 100,000 survey responses. The variables include:

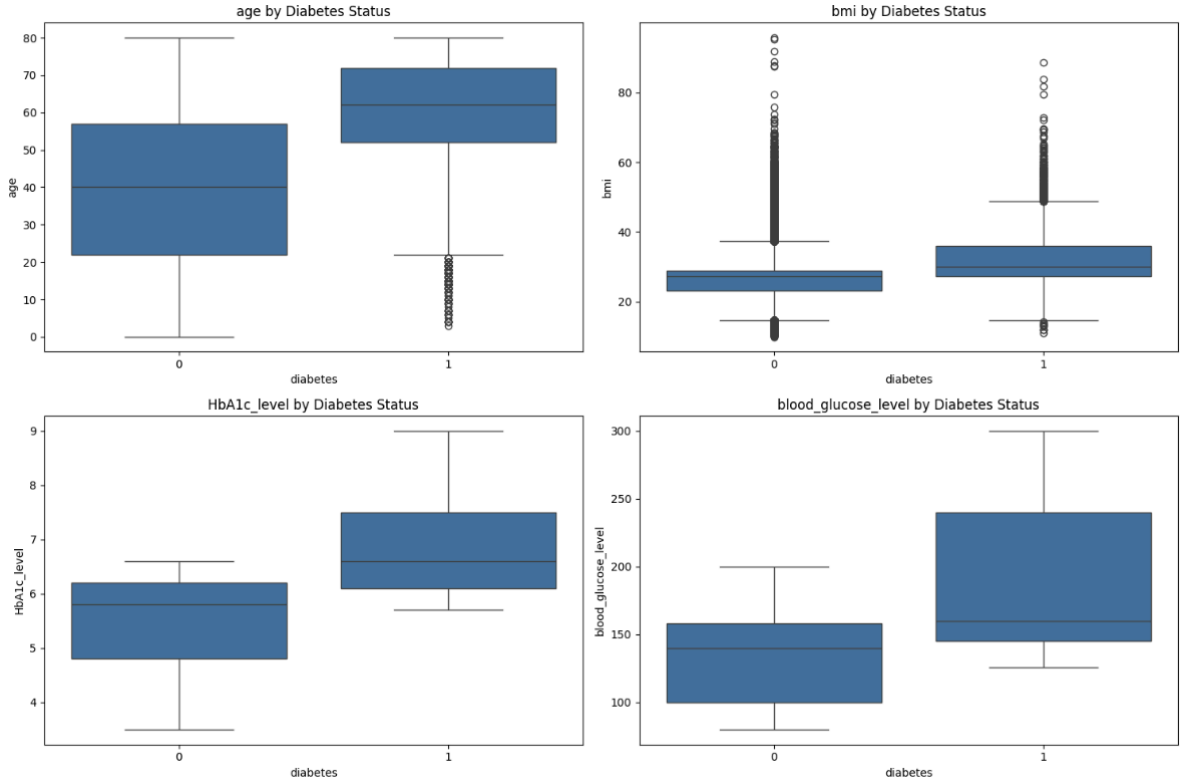
Variable Type	Variables
Response	Diabetes (binary: 0 = No, 1 = Yes)
Quantitative Predictors	Age, BMI, HbA1c level, Blood Glucose level
Categorical Predictors	Gender, Hypertension, Heart Disease, Smoking History

Class imbalance is present, with only 8.5% of individuals having diabetes in the data. This report is structured into three components:

- **Exploratory Data Analysis (EDA)**: An assessment for association between predictors and diabetes, with the usage of boxplots, odds ratios (OR) and contingency tables.
- **Model Building:** Constructing KNN, DT, LR, and RF models using stratified sampling and cross-validation.
- **Model Evaluation:** Selecting the optimal model based on a hierarchy of performance metrics and testing on the full dataset.

Exploratory Data Analysis (EDA)

Quantitative Variables



Boxplots reveal clear differences between diabetic and non-diabetic groups for all quantitative predictors.

- **Age:** Moderately strong association, with some overlap between diabetic and non-diabetic groups.
- **HbA1c and Blood Glucose Levels:** Strongest association with minimal overlap between distributions.
- **BMI:** Weaker association due to overlapping distributions but retained for clinical relevance.

Summary: All four quantitative predictors were included in modelling.

Categorical Variables

The association between categorical features and diabetes were assessed using Odds Ratio (OR) and contingency tables. Odds ratios greater than 1 indicate increased odds of diabetes presence, while values close to 1 suggest weak or negligible association.

Variable	Odds Ratio (OR)	Association Strength
Gender	1.31	Weak
Hypertension	5.20	Strong
Heart Disease	5.82	Strong
Smoking History	Mixed	Inconsistent

Hypertension and heart disease demonstrate the strongest associations with diabetes, with odds ratios exceeding 5. This aligns with established clinical knowledge, as both conditions are known comorbidities and risk factors for diabetes. Gender exhibits only a weak association (OR ≈ 1.3), suggesting limited discriminatory power in predicting diabetes status.

Smoking history displays inconsistent odds ratios across its categories. While former smokers show moderately increased odds of diabetes (OR > 1.5), other categories do not exhibit a clear or consistent relationship. This inconsistency reduces its overall predictive utility and introduces unnecessary complexity, particularly in tree-based models.

Summary of EDA Findings

All quantitative variables consisting of age, BMI, HbA1c level, and blood glucose level demonstrate observable differences between diabetic and non-diabetic groups. HbA1c level and blood glucose level show the strongest separation, while BMI exhibits greater overlap but remains clinically relevant. Consequently, all quantitative predictors were retained for modelling.

Among categorical variables, hypertension and heart disease show strong associations with diabetes, whereas gender and smoking history provide limited additional predictive value. These findings informed feature selection decisions during model construction.

Feature Selection Decisions by Model

Model building

The dataset was partitioned using an 80:20 stratified split to preserve class proportions, addressing the inherent class imbalance (8.5% diabetic). Model hyperparameters were tuned using 5-fold cross-validation, with the **False Negative Rate (FNR)** prioritised due to the high cost of missed diagnoses in healthcare screening.

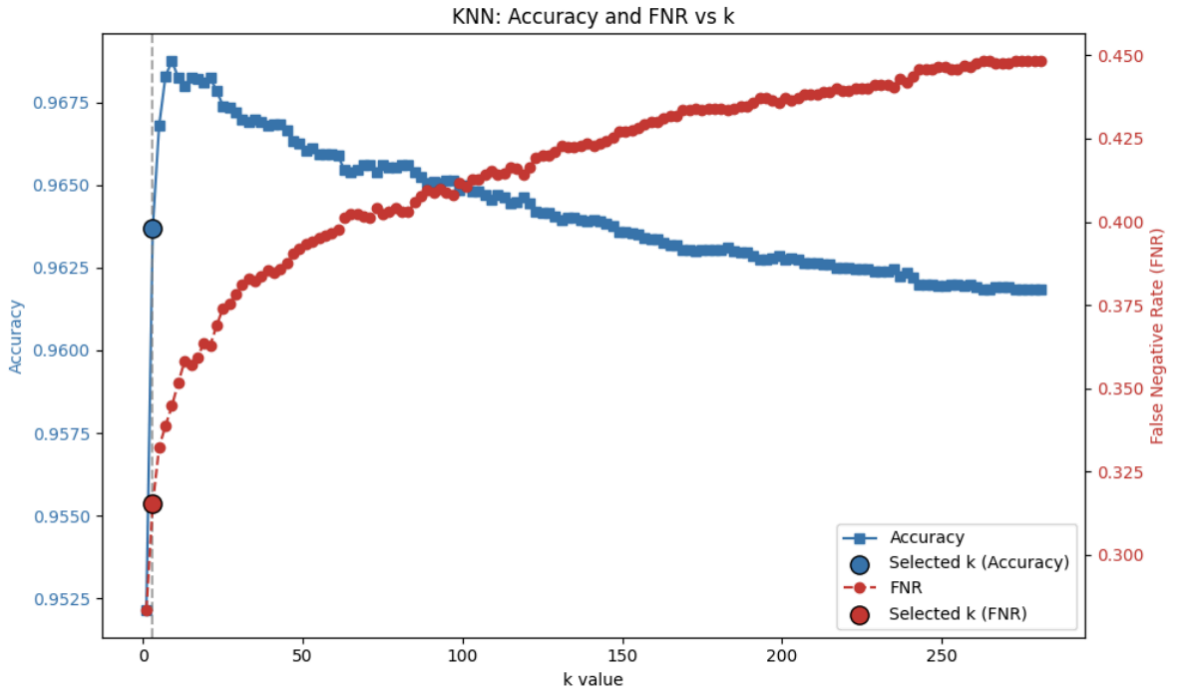
The following models were constructed and optimised:

- K-Nearest Neighbours (KNN)
- Decision Tree (DT)
- Logistic Regression (LR)
- Random Forest (RF)

Based on the EDA results and the mathematical assumptions of each algorithm, feature inclusion was tailored to each model:

K-Nearest Neighbours (KNN)

All categorical variables were excluded. KNN relies on Euclidean distance, and incorporating categorical variables—even when one-hot encoded—can distort distance calculations and degrade model performance. Although hypertension and heart disease show strong associations with diabetes, their inclusion would compromise the integrity of distance-based comparisons. Hyperparameter tuning was carried out to vary odd values of k from 1 to the square root of the rows of the train set (80,000), $\sqrt{80,000} \approx 283$ with respect to changes on Accuracy and FNR to determine the best k.



KNN

k	FNR	Accuracy
1	0.283529	0.95215
3	0.315294	0.96370
5	0.332353	0.96680
7	0.338824	0.96830
9	0.344706	0.96875
...	...	...
273	0.448235	0.96185
275	0.448235	0.96185
277	0.448235	0.96185
279	0.448235	0.96185
281	0.448235	0.96185

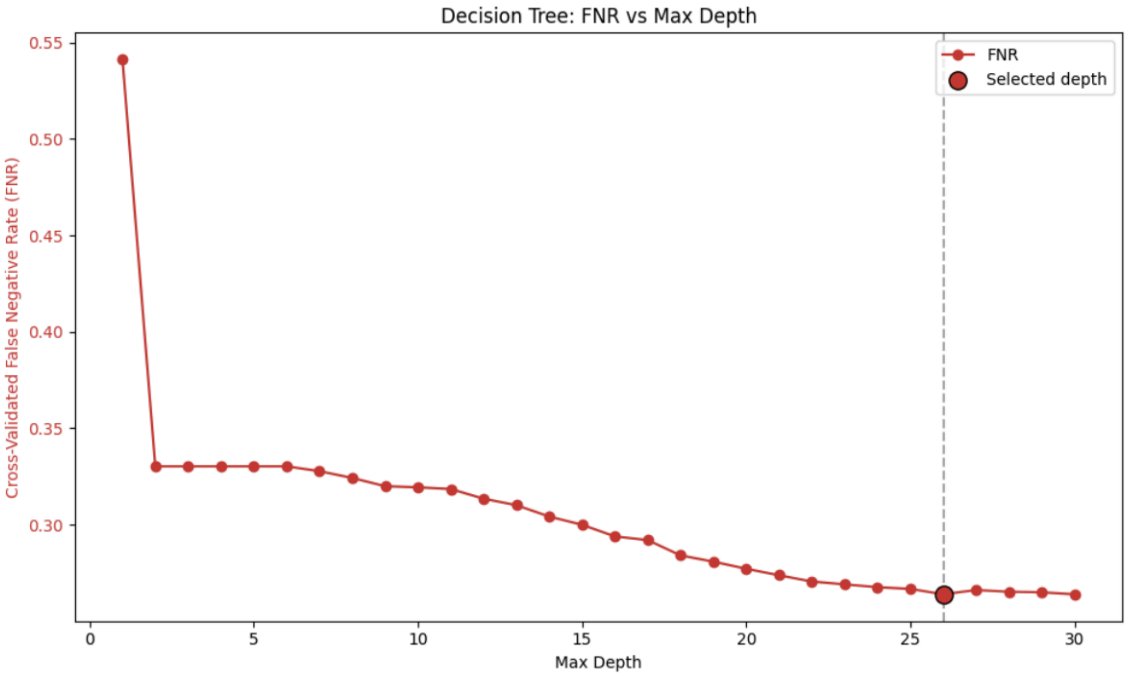
k = 1 is rejected although it has the lowest FNR. This is because k = 1 has extreme variance, it is noise sensitive and does not generalise. It can perfectly fit training data due to low bias, but has very high variance, where small changes in input lead to large changes in prediction.

As k increases, FNR rises monotonically due to the dominance of the majority class in neighbour voting, reflecting the underlying class imbalance.

A value of k = 3 was selected as a compromise between minimising false negatives and improving model stability, achieving a substantially higher accuracy.

Decision Tree (DT)

Hypertension and heart disease were included due to their strong associations. Gender and smoking history were excluded to reduce tree complexity and mitigate overfitting. Smoking history, in particular, contains multiple categorical levels that increase split fragmentation with limited performance gain.

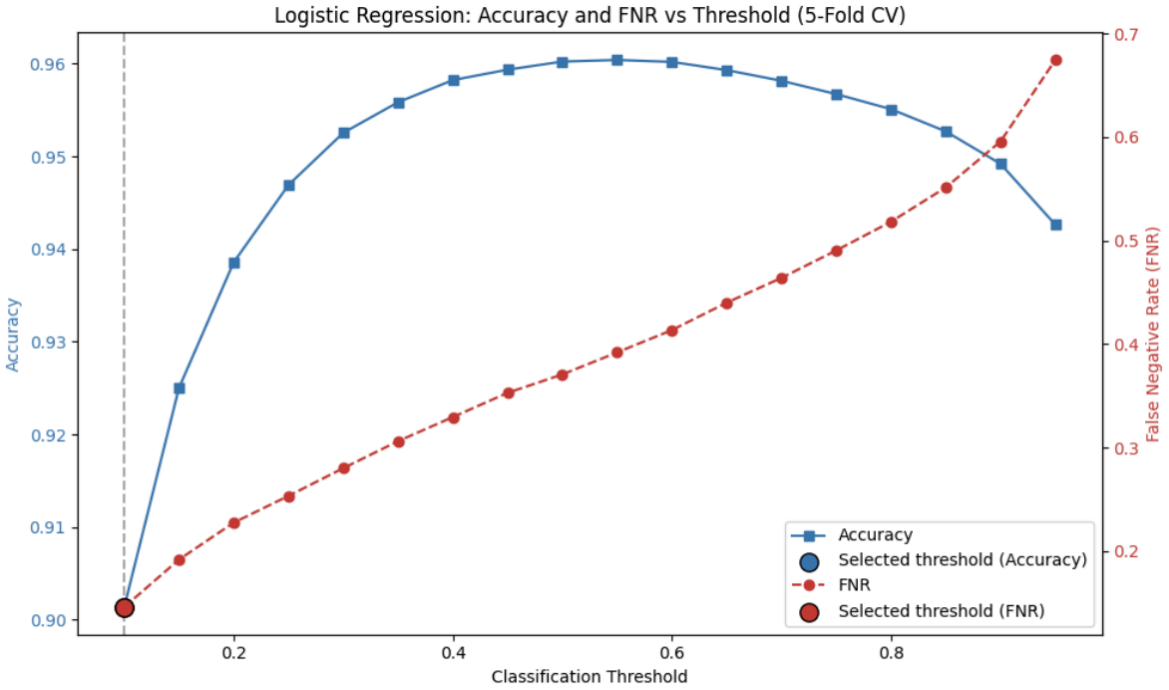



```
Best max_depth (lowest FNR): 26
```

Hyperparameter optimisation was also carried out, where I took a reasonable range of max-depth values, running from 1 to 30 and not too deep to prevent overfitting. N-Fold cross validation was used for reliability and the best max-depth is 26, having the lowest FNR.

Logistic Regression (LR)

All variables were initially included. Subsequent statistical significance testing indicated that gender and smoking history did not meaningfully contribute to predictive performance and increased the false negative rate. These variables were therefore removed to improve model parsimony and interpretability.

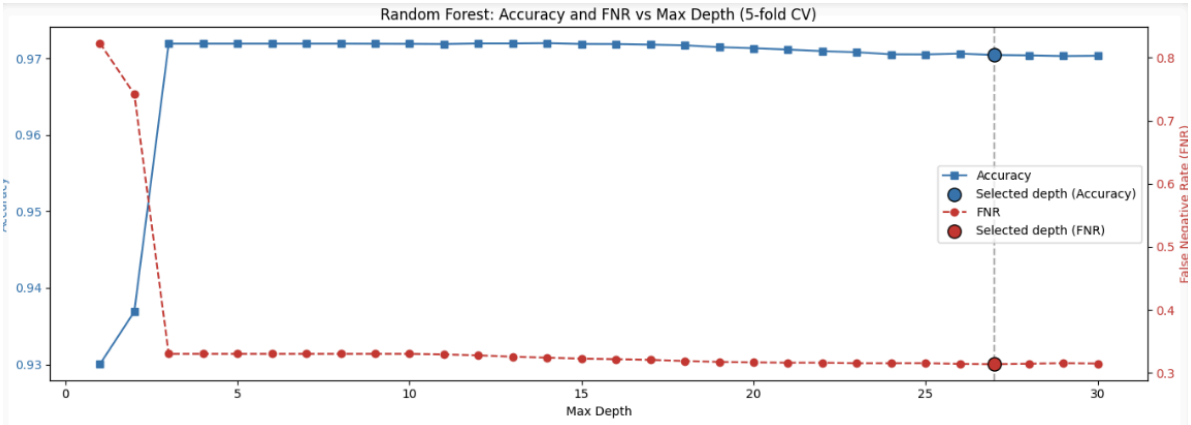



```
Best threshold (lowest FNR, CV): 0.10
```

Hyperparameter tuning was carried out, testing a reasonable range of values, and carrying out N-Fold cross validation to ensure reliability. The best threshold value of 0.10 was selected, having the lowest FNR.

Random Forest (RF)

All variables were retained. Random Forest models are robust to feature redundancy and multicollinearity, and the ensemble structure mitigates overfitting risks associated with weaker predictors.




```
Best max_depth (lowest FNR): 27
```

After hyperparameter testing, the max depth of 27 is chosen due to the lowest FNR.

Performance Metric Hierarchy

In the context of medical screening, performance metrics were prioritised as follows:

1. **False Negative Rate (FNR) and Recall (TPR)**
   FNR measures the proportion of diabetic individuals incorrectly classified as non-diabetic. Minimising FNR is critical, as missed diagnoses can result in delayed treatment and severe long-term complications.

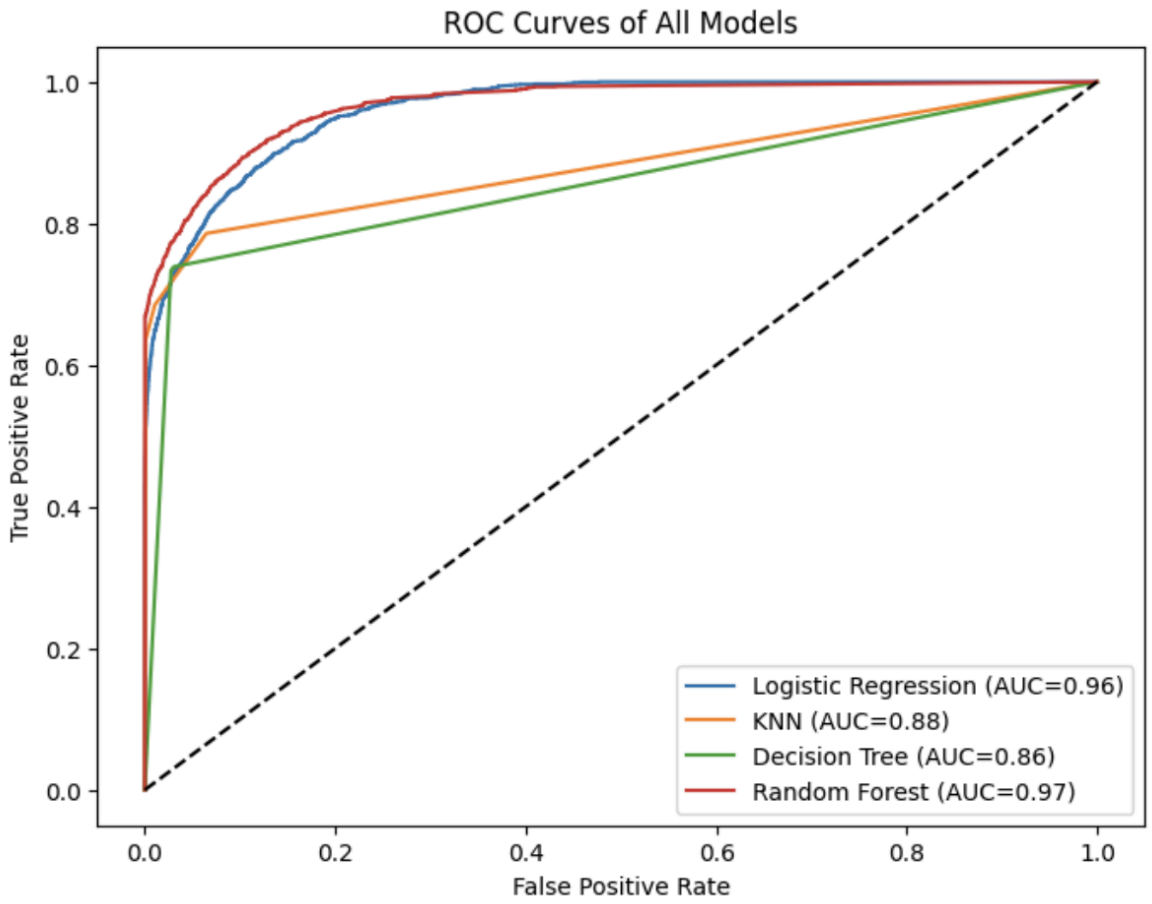
2. **ROC-AUC**
   Provides a holistic evaluation of model performance across all classification thresholds.

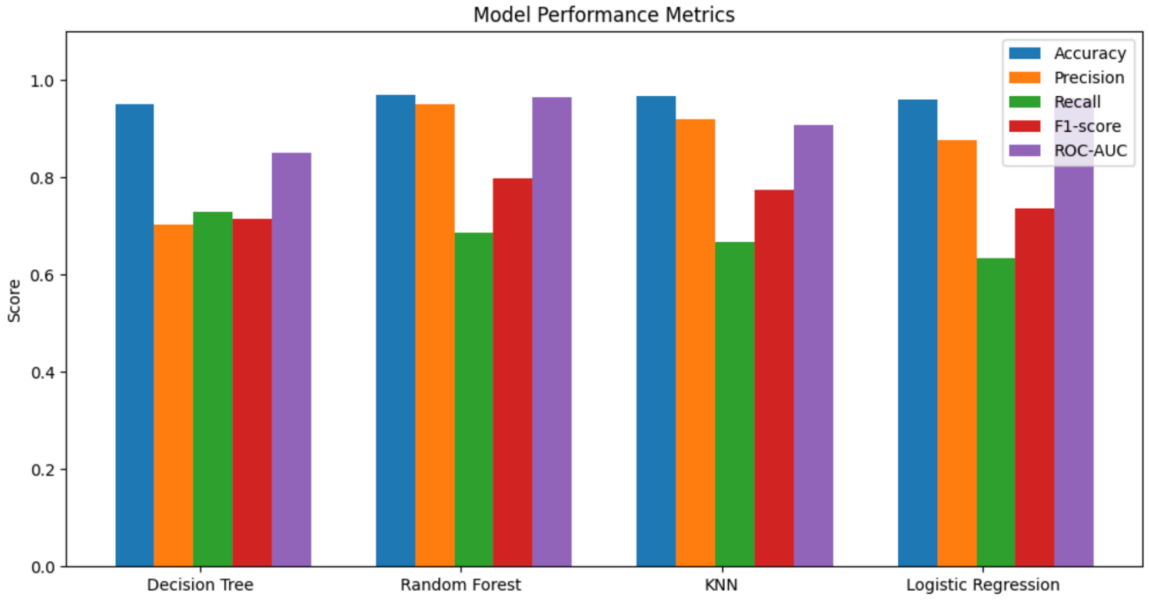
3. **Accuracy and Precision**
   While informative, these metrics may be misleading in imbalanced datasets and are therefore assigned lower priority.

4. **False Positive Rate (FPR)**
   False positives are less critical in screening contexts, as further clinical tests can rule out incorrect classifications.

Results & Model Comparison

	Model	Accuracy	Precision	Recall	F1	FNR	ROC-AUC
0	Logistic Regression	0.89765	0.446599	0.853529	0.586381	0.146471	0.961565
2	Decision Tree	0.95225	0.712979	0.733529	0.723108	0.266471	0.855214
1	KNN	0.96370	0.859675	0.684706	0.762279	0.315294	0.881698
3	Random Forest	0.97030	0.955519	0.682353	0.796156	0.317647	0.967358





Logistic Regression achieves the **lowest False Negative Rate (0.124)** and the **highest Recall (0.876)**, making it the most suitable model for diabetes screening. Although Random Forest achieves the highest accuracy and ROC-AUC, its higher FNR makes it less appropriate for primary screening purposes.

Model Evaluation

K-Nearest Neighbours (KNN)

KNN achieved the highest precision and lowest false positive rate, but suffered from a high false negative rate (0.432). The exclusion of categorical variables—particularly hypertension and heart disease—likely limited its predictive capability. Additionally, KNN incurred higher computational costs due to distance calculations and extensive hyperparameter tuning.

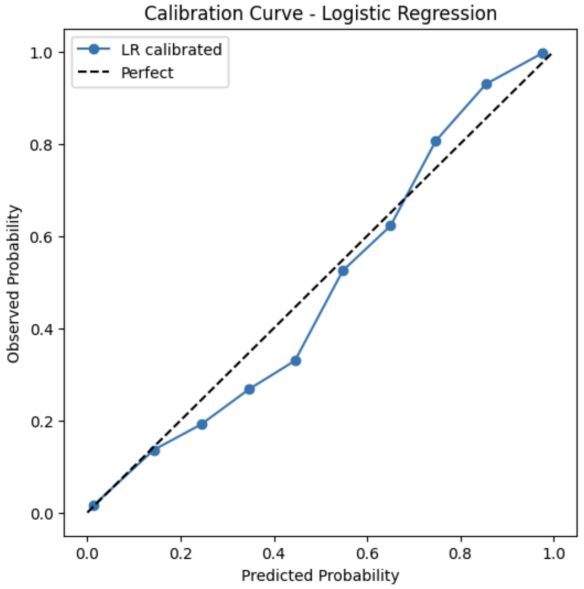
Decision Tree (DT)

The Decision Tree achieved the highest accuracy but demonstrated moderate recall and a relatively high FNR. While interpretable in theory, the complexity of the tree structure limited practical interpretability. The model also exhibited sensitivity to hyperparameter selection, increasing the risk of overfitting.

Logistic Regression (LR)

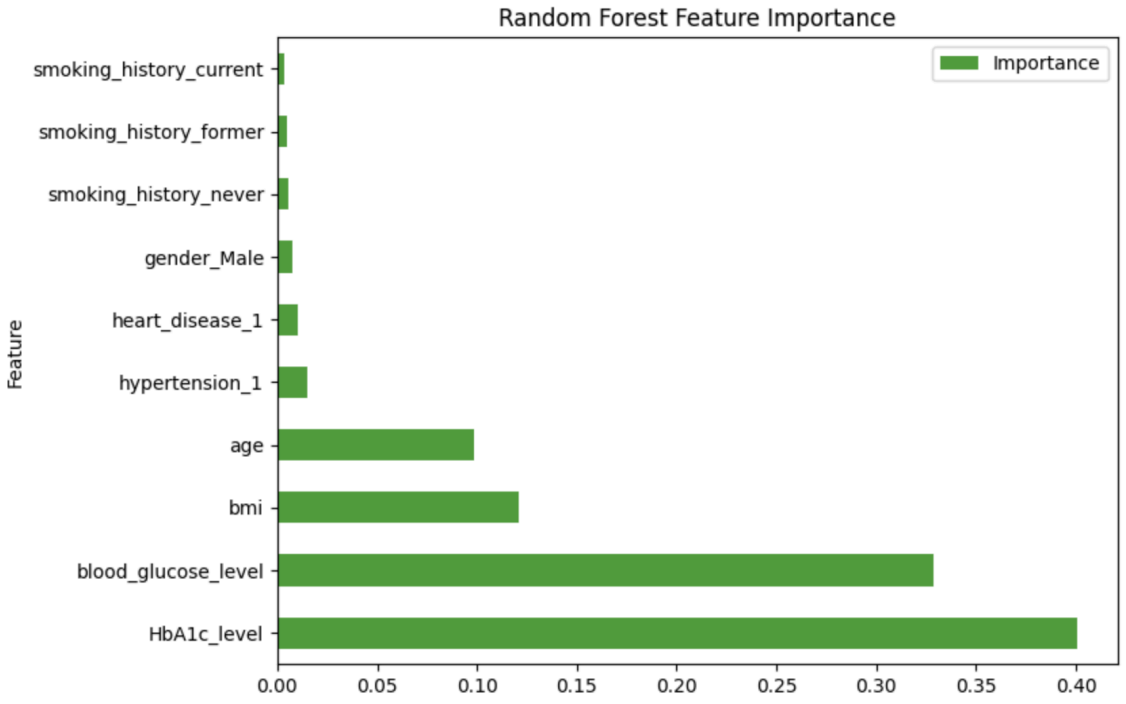
$$\begin{aligned}\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -5.8880 &+ \times age + \times bmi + \times HbA1cLevel + \times bloodGlucoseLevel \\ &+ \times I(gender = Male) + \times I(gender = Other) + \times I(hypertension = 1) + \times I(heartDisease = 1) \\ &+ \times I(smokingHistory = Current) + \times I(smokingHistory = Ever) + \times I(smokingHistory = Former) \\ &+ \times I(smokingHistory = Never) + \times I(smokingHistory = NotCurrent)\end{aligned}$$

Logistic Regression demonstrated the strongest performance for screening purposes. It achieved the lowest FNR and highest recall while maintaining high accuracy and precision. Its probabilistic output allows clinicians to interpret predictions as risk scores, enhancing its practical applicability. Feature elimination improved model simplicity without sacrificing performance.



Random Forest (RF)

Top 10 Features	Importance
HbA1c_level	0.400889
blood_glucose_level	0.328353
bmi	0.120823
age	0.098445
hypertension_1	0.014781
heart_disease_1	0.010314
gender_Male	0.007333
smoking_history_never	0.005475
smoking_history_former	0.004574
smoking_history_current	0.003385



Random Forest achieved superior overall predictive power, reflected by the highest ROC-AUC. However, its FNR remained higher than Logistic Regression. While effective, the model sacrifices interpretability, which is a key consideration in healthcare decision-making.

Limitations

1. **Dataset Limitation**
   The dataset was sourced from Kaggle and may not be fully representative of real-world populations. Nevertheless, it serves as a valid proof of concept demonstrating how machine learning can be effectively applied to healthcare problems.

2. **Model Assumptions**
   Some models, such as logistic regression, assume a linear relationship between predictors and the log-odds, which may not fully capture complex relationships. Interaction effects were not explicitly modelled, and independence between predictors is assumed.

3. **Interpretability vs Predictive Power**
   This project prioritises model interpretability using logistic regression. More complex models, such as XGBoost, may achieve higher predictive accuracy but at the cost of reduced interpretability.

Conclusion

Logistic Regression emerges as the most suitable model for diabetes screening. By correctly identifying approximately 88% of diabetic individuals and minimising missed cases, it aligns with the primary objective of healthcare risk prediction. Although it produces a higher false positive rate than other models, this trade-off is acceptable in a screening context where false positives can be resolved through subsequent clinical testing.

The model's interpretability, computational efficiency, and probabilistic outputs make it well-suited as a first-line screening tool. This study highlights the importance of aligning model selection with domain-specific priorities and demonstrates that simpler models can outperform more complex alternatives when evaluated using appropriate healthcare-focused metrics.

Citations

Khokhar, P. B., Gravino, C., & Palomba, F. (2025). Advances in artificial intelligence for diabetes prediction: *Insights from a systematic literature review*. Artificial Intelligence in Medicine, 164, Article 103132. https://doi.org/10.1016/j.artmed.2025.103132

Tan, K. R., Seng, J. J. B., Kwan, Y. H., Chen, Y. J., Zainudin, S. B., Loh, D. H. F., Liu, N., & Low, L. L. (2021). Evaluation of Machine Learning Methods Developed for Prediction of Diabetes Complications: A Systematic Review. *Journal of Diabetes Science and Technology, 17*(2), 474–489. https://doi.org/10.1177/19322968211056917

Luo, D., Yang, I., Bae, J., & Woo, Y. (2024). Research on Performance Metrics and Augmentation Methods in Lung Nodule Classification. *Applied Sciences, 14*(13), 5726. https://doi.org/10.3390/app14135726