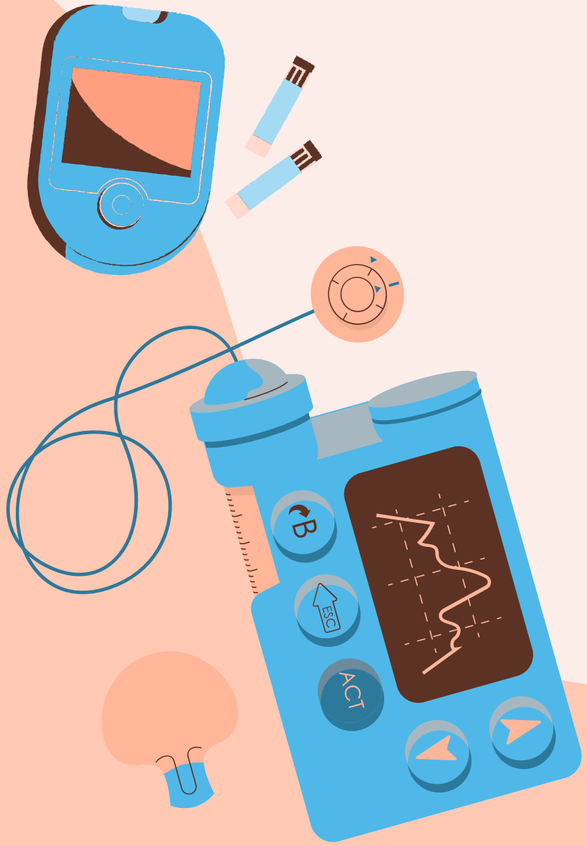
CAPSTONE 3: DIABETES PREDICTION PROJECT



Mengxiao Wang

Addressing Diabetes Prediction through Data Science

Purpose of project

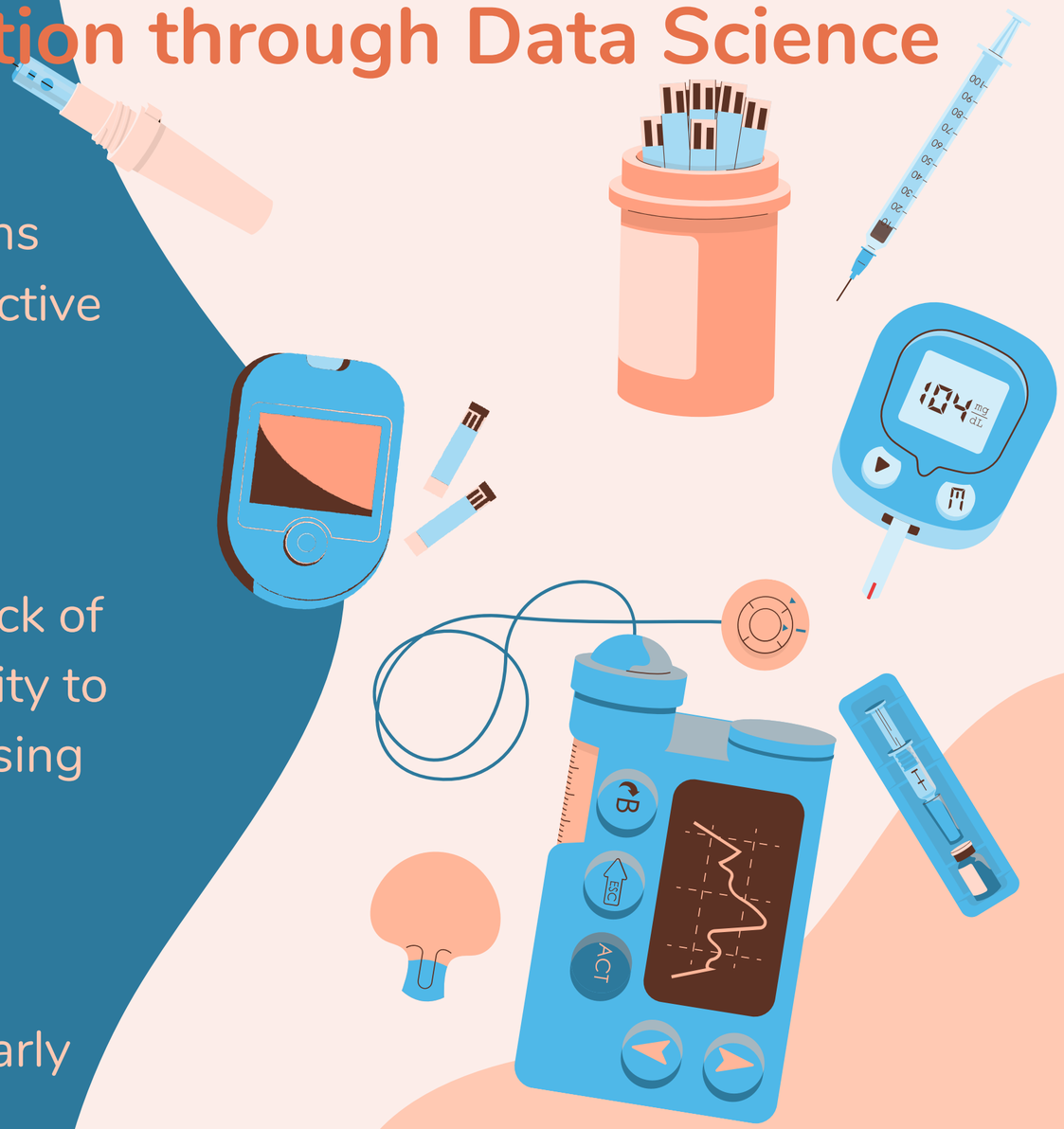
- Diabetes is a chronic disease that affects millions worldwide. Early detection is important for effective management and prevention.

Problem Statement

- Many individuals remain undiagnosed due to lack of access to medical testing. There is an opportunity to leverage data science to predict diabetes risk using readily available health data.

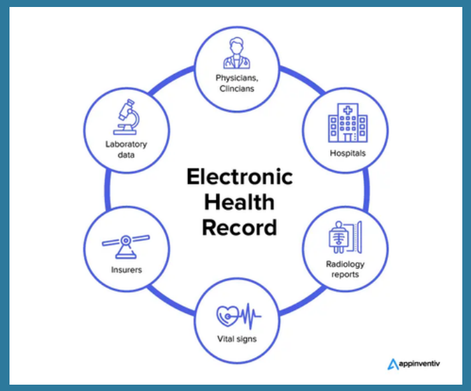
Opportunity:

- Implementing a predictive model can provide early warnings.

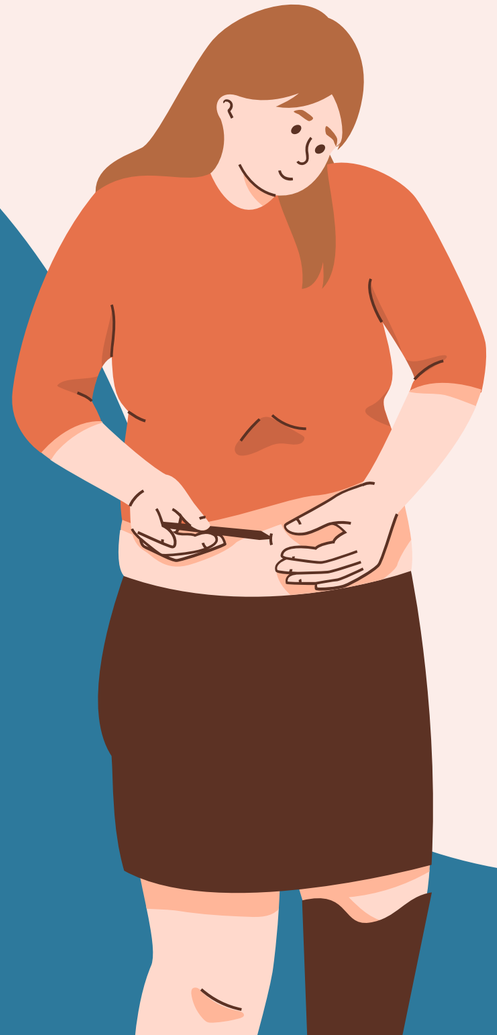


INTRODUCTION OF THE DATASET

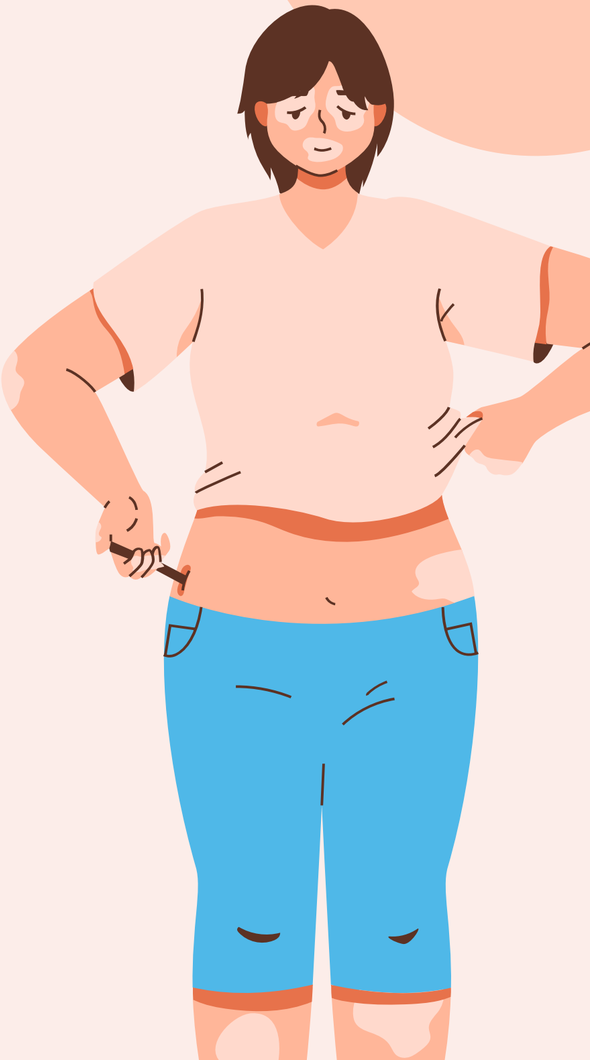
Resource: the Diabetes prediction dataset is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). Electronic Health Records (EHRs) are the primary source of data for the Diabetes Prediction dataset.



SHAPE OF DATASET

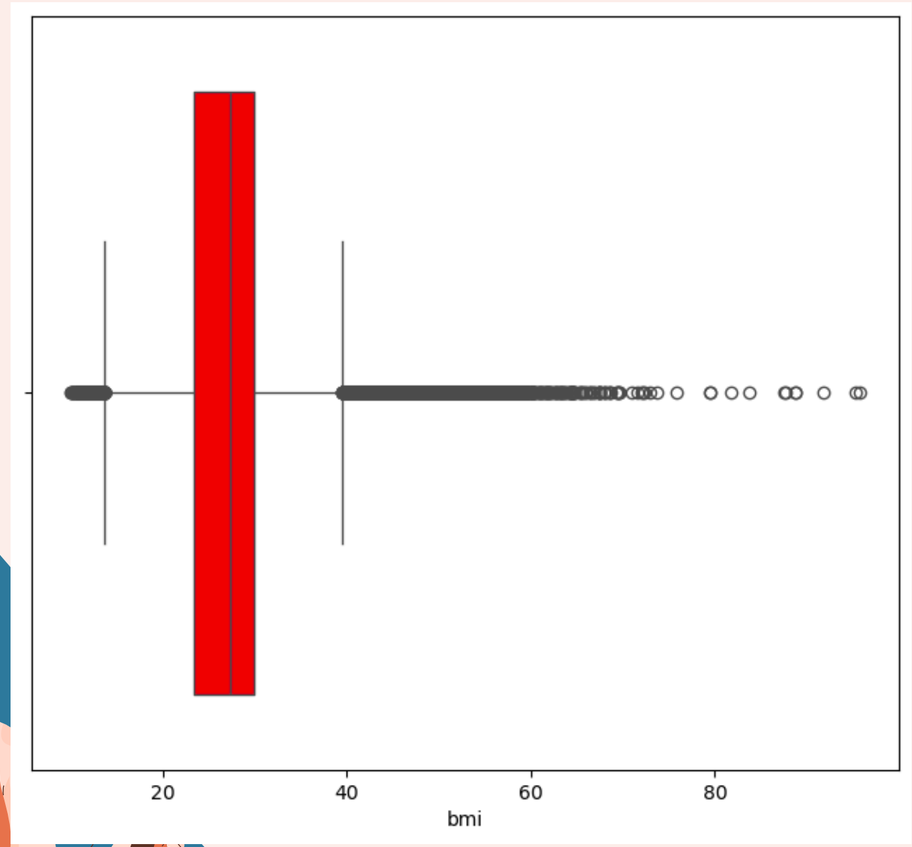


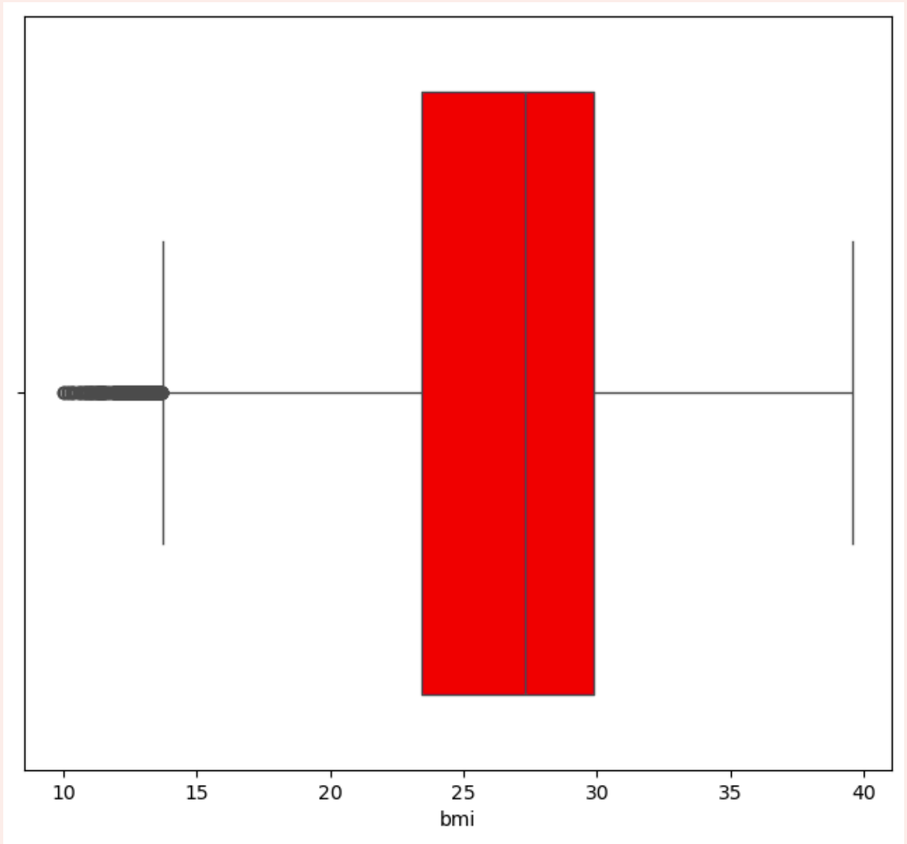
The diabetes dataset has 100,000 rows and 9 columns, with 3854 duplicate rows and no null values.

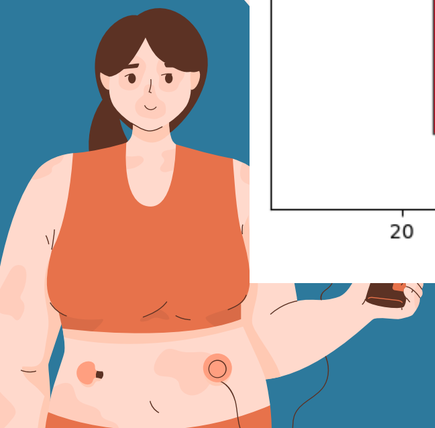


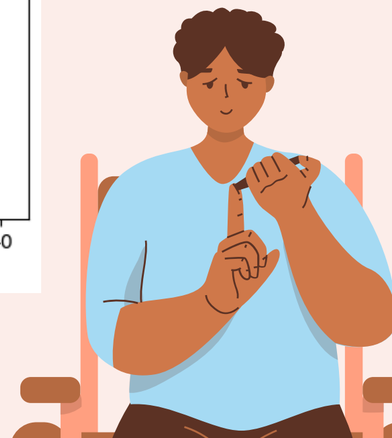
Preprocessing Procedures

1.Dropped 3854 duplicated rows
2.Mapping the smoking_history column for simplification reason
3.Handle outliners for BMI, HbA1c_level and blood_glucose_level

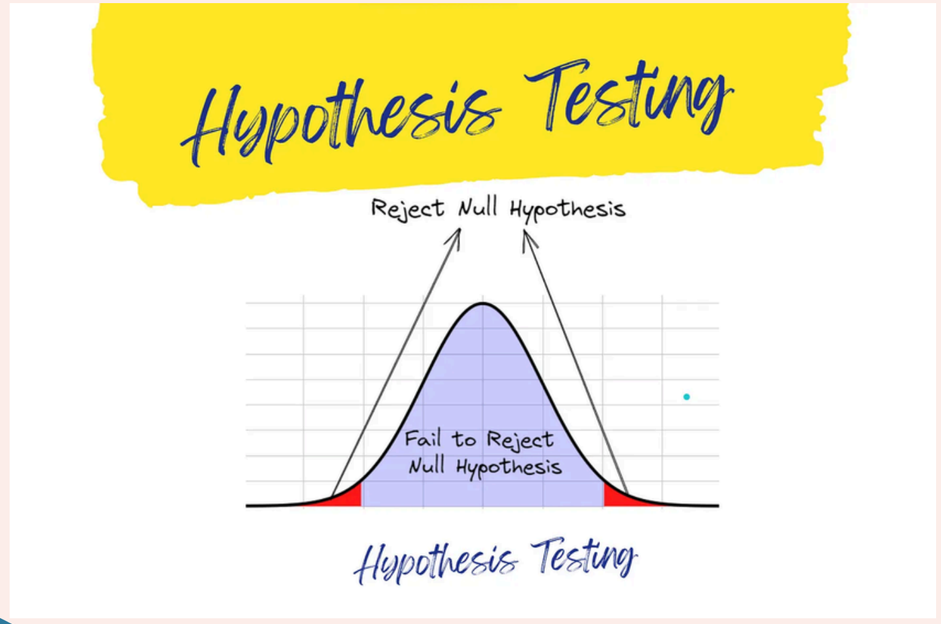








HYPOTHESIS TESTING

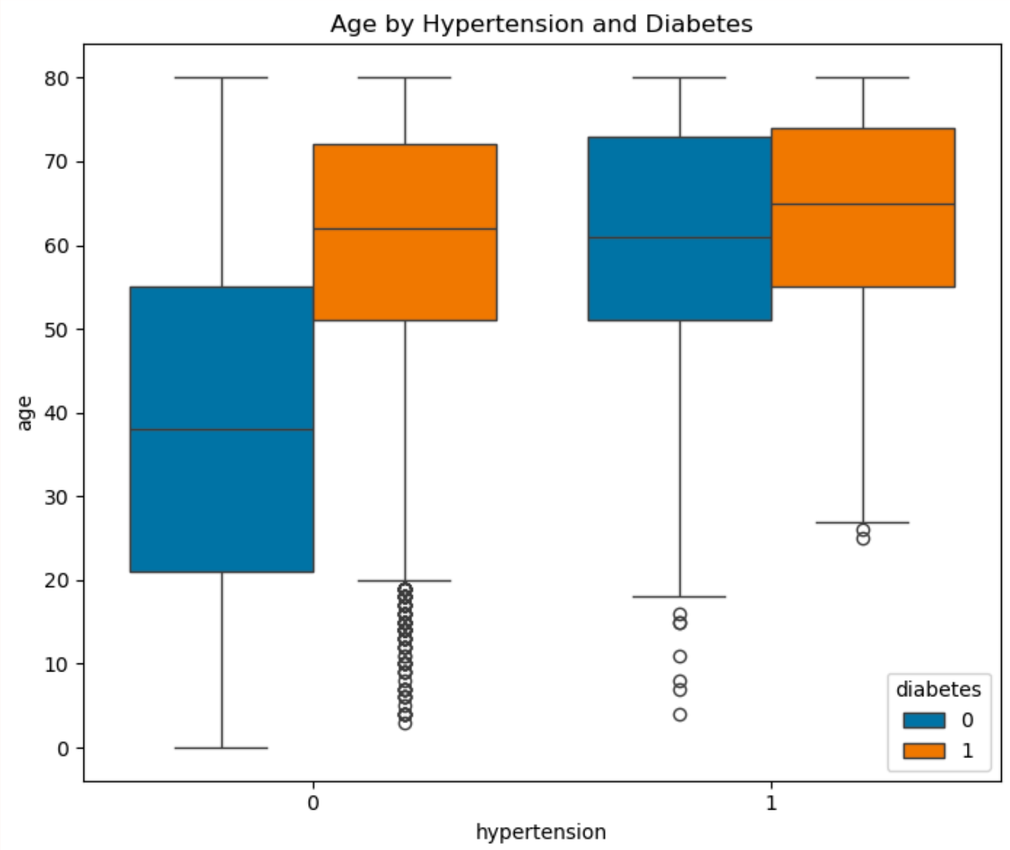


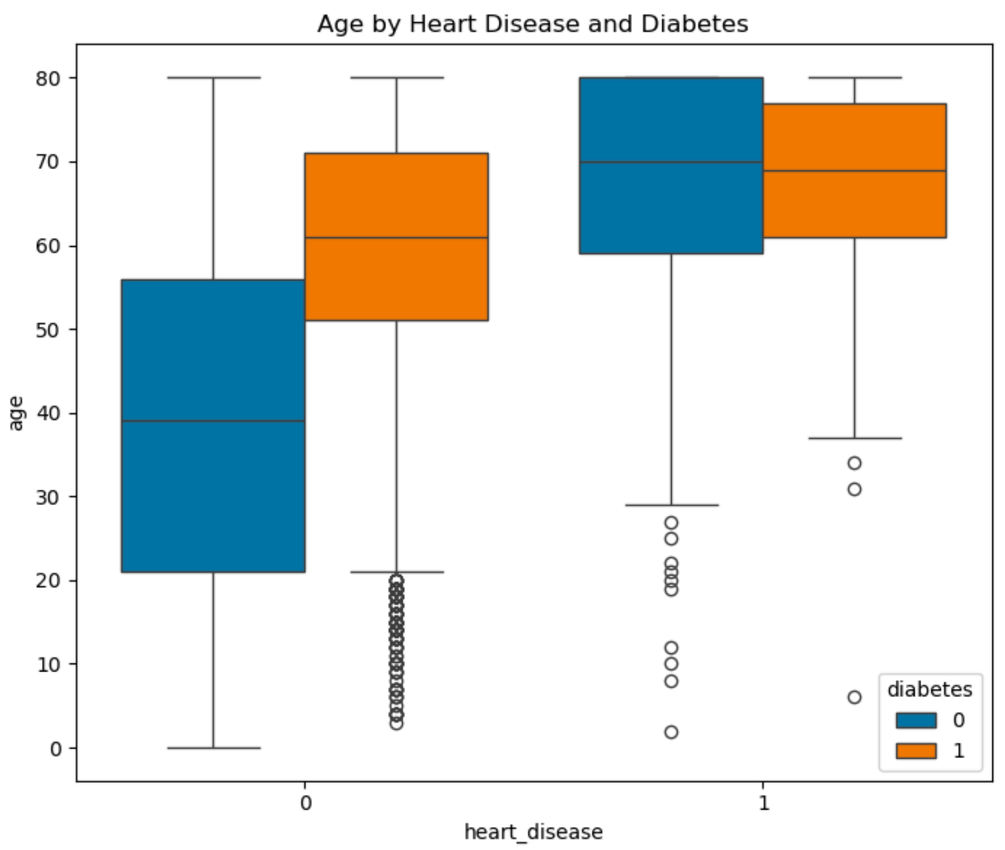
All tested features (gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, and blood glucose level) are highly significant for predicting diabetes.

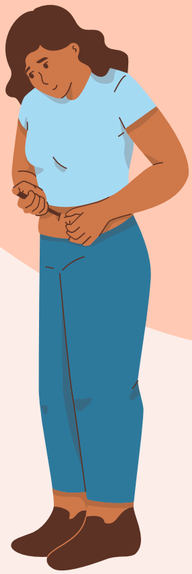
This suggests that these features should be considered important predictors in your diabetes prediction model.

Findings from EDA

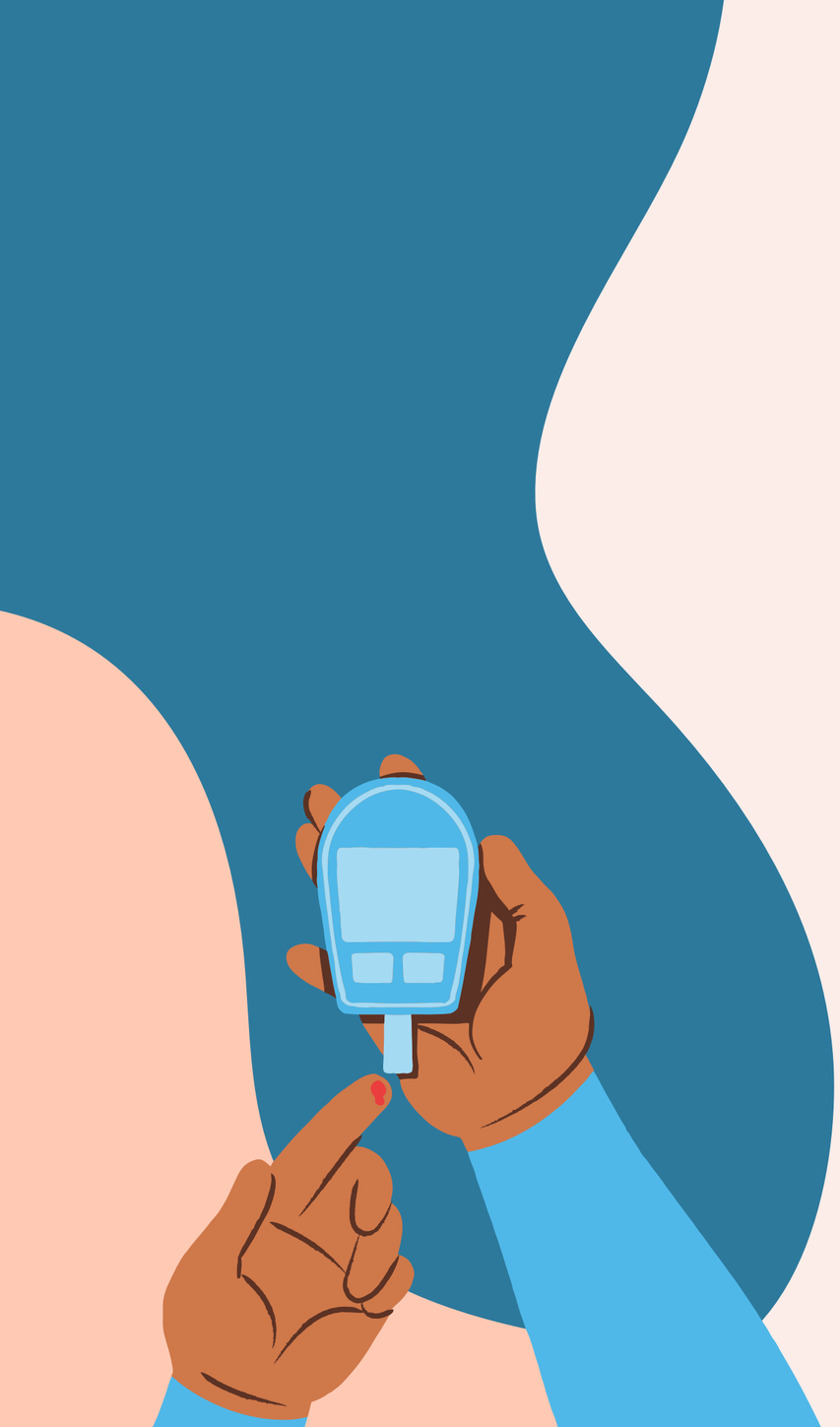
Hypertension and heart_disease may independently contribute to the risk of diabetes regardless of age.

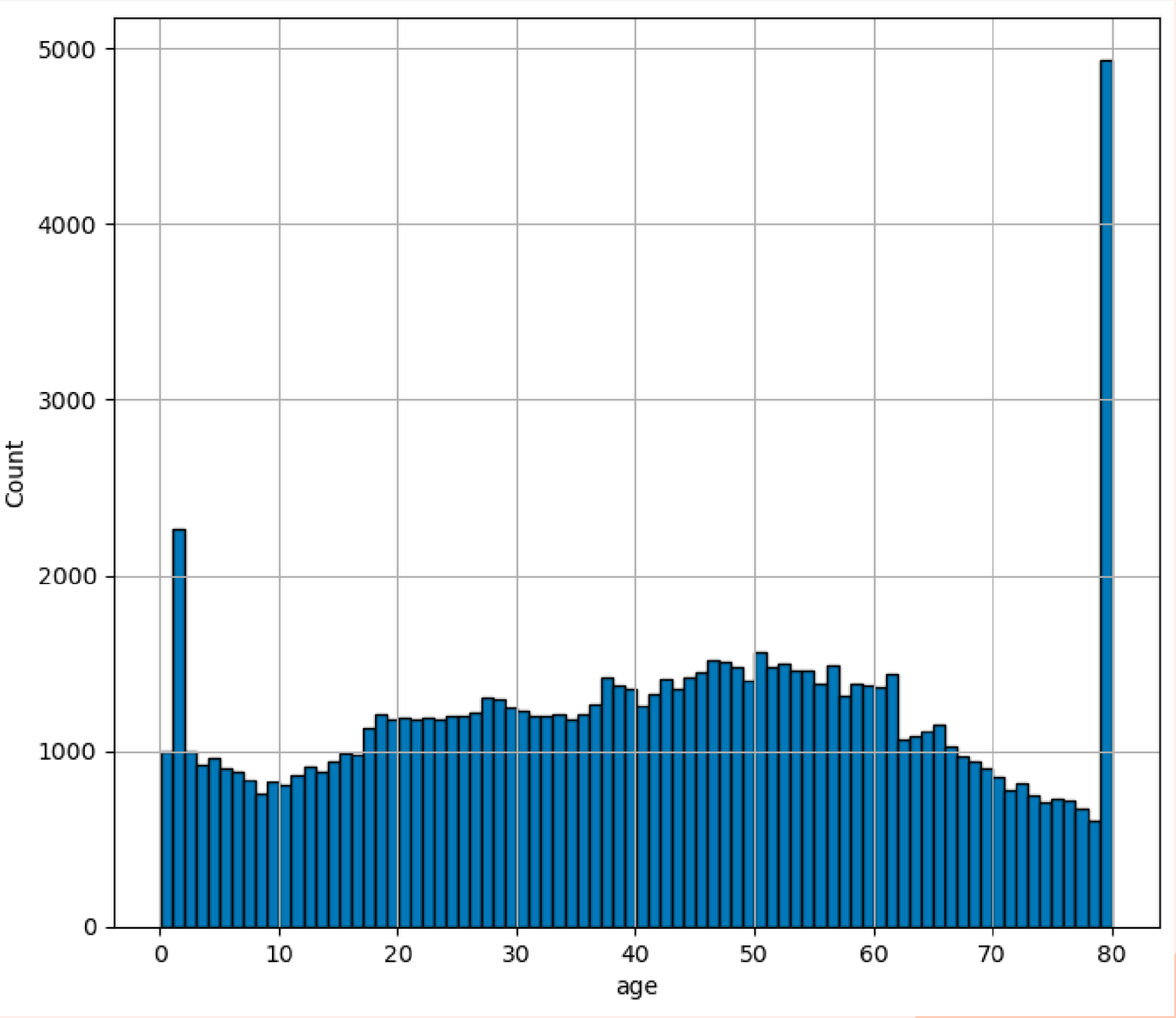


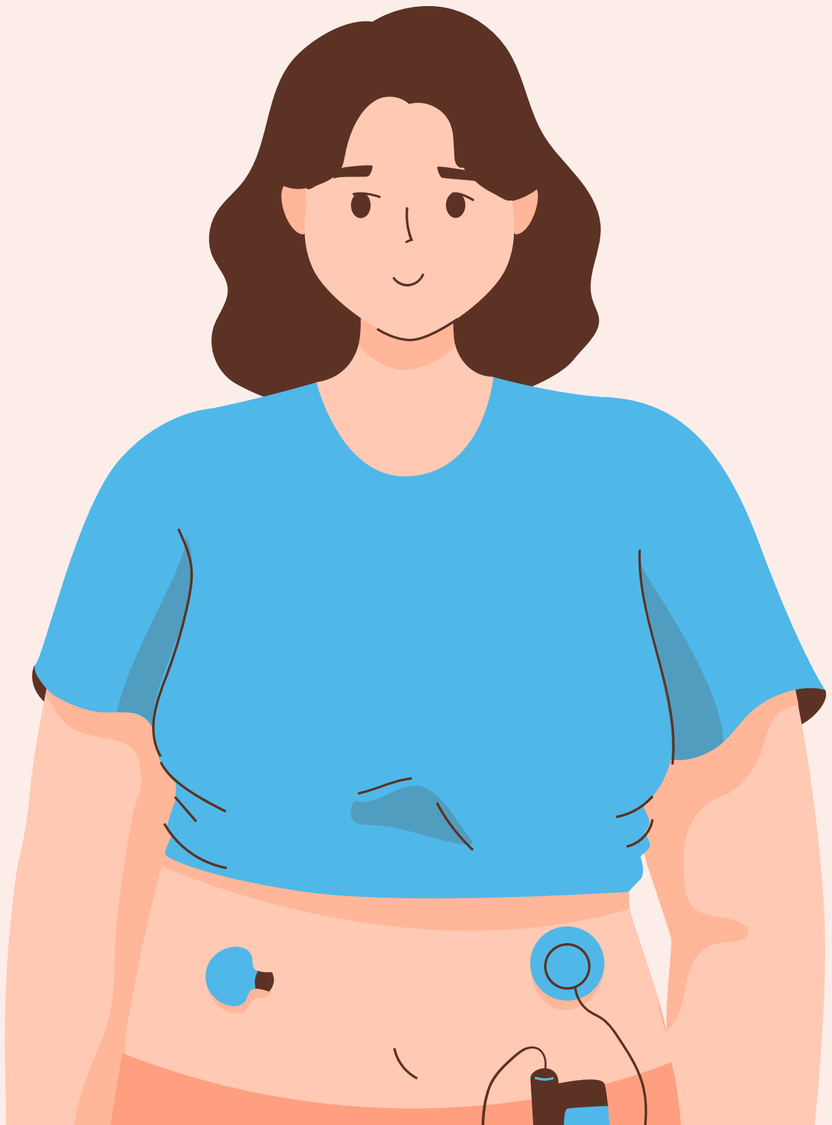




THERE ARE NOTICEABLE PEAKS AT AGE 0 AND AGE 80

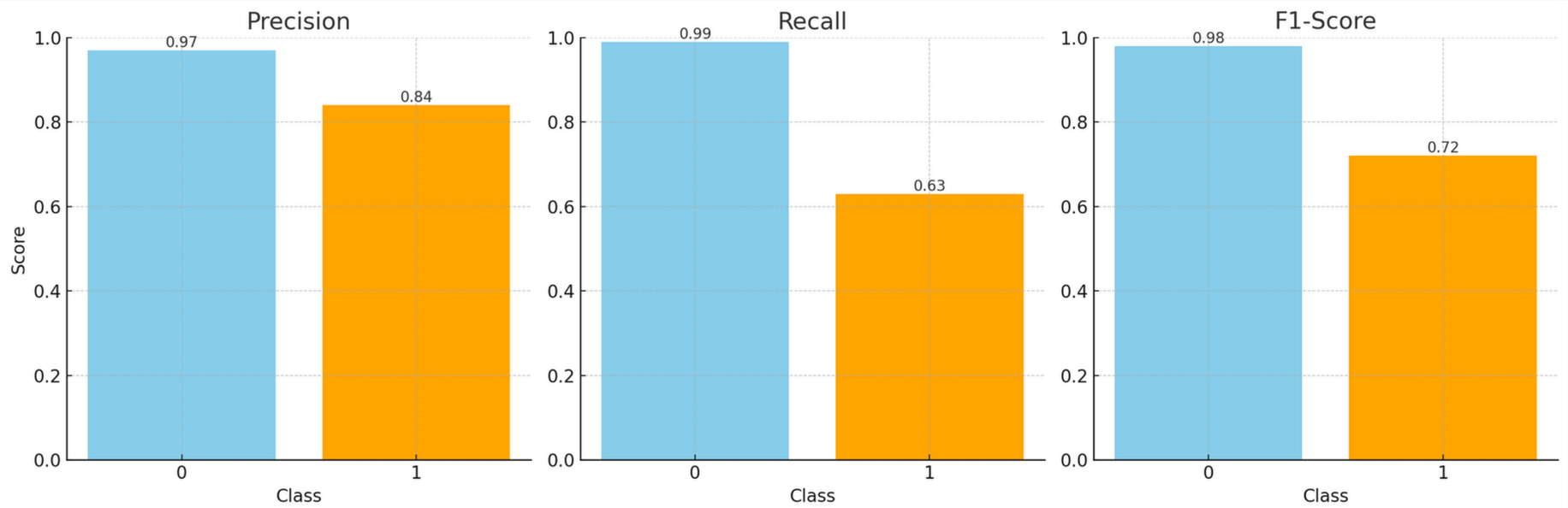


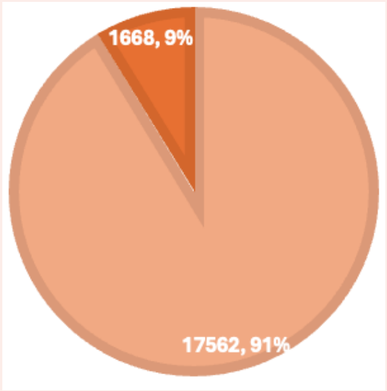




MODEL COMPARISON AND INTERPRETATION

LOGISTIC REGRESSION

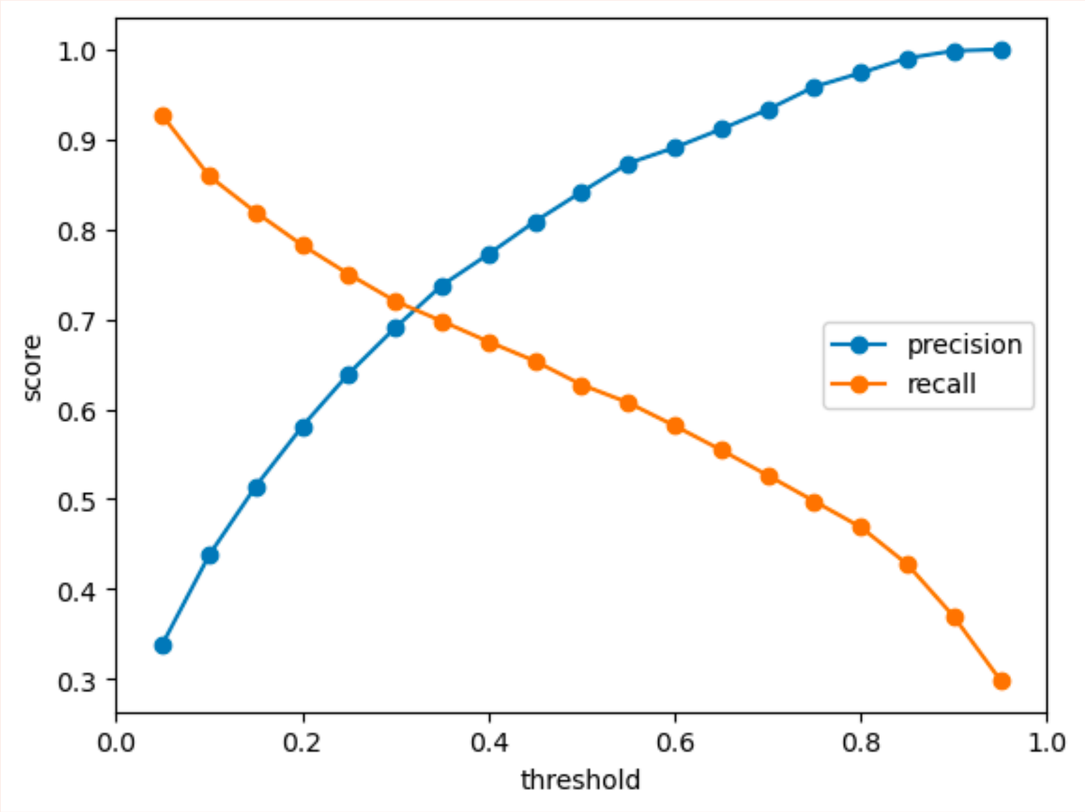


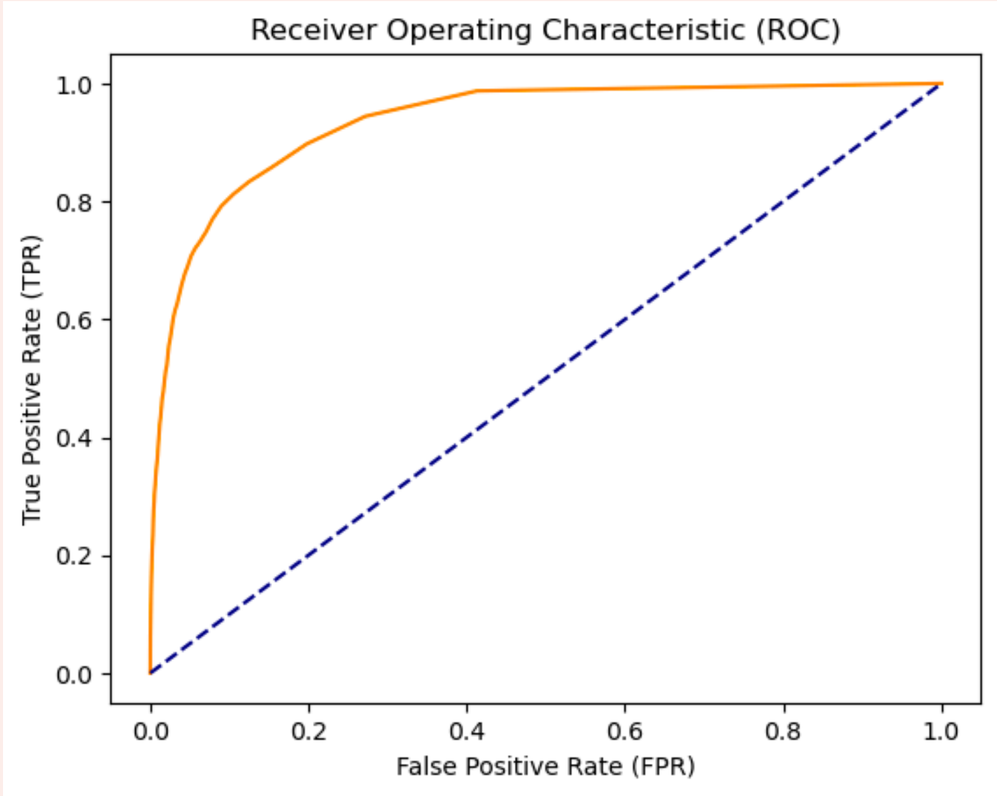


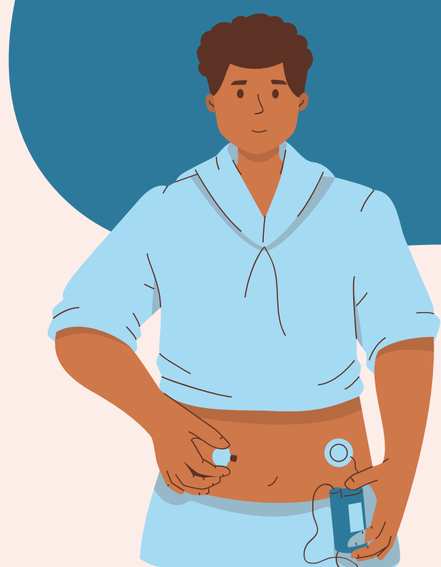
1.My recall for 1 is low, indicates that I am missing some positive cases.
2.My support for class 0 is 91%, while for class 1 it is 9%, indicating that my data is imbalanced.

PRECISION-RECALL TRADE-OFF AND ROC CURVE

The ROC curve shows that my model has a good overall performance with a high TPR and low FPR.

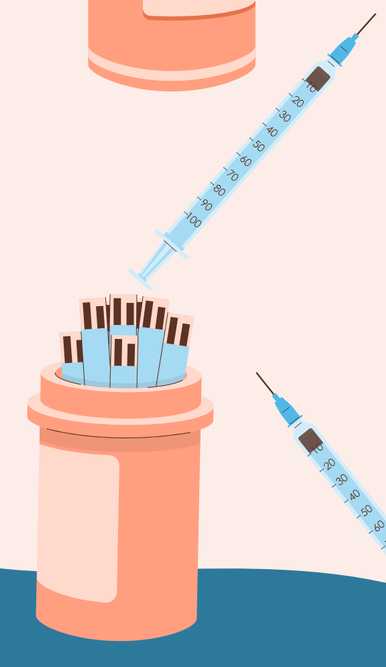


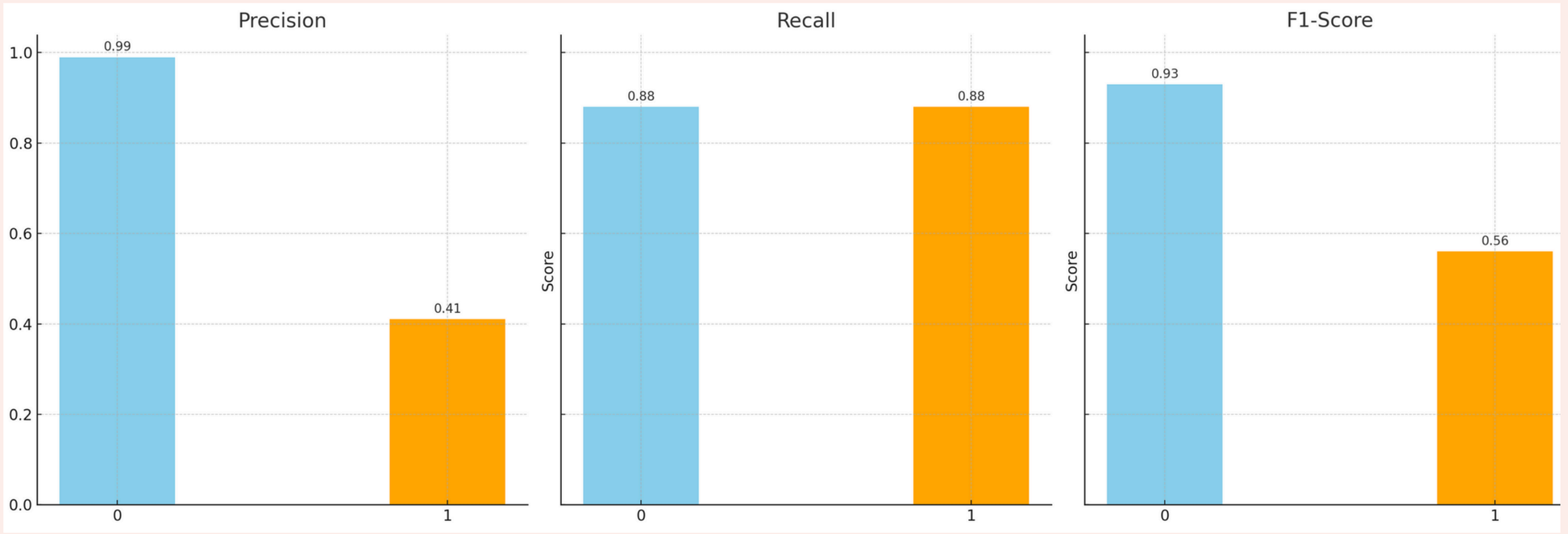




UPSAMPLING, DOWNSAMPLING AND SMOTE

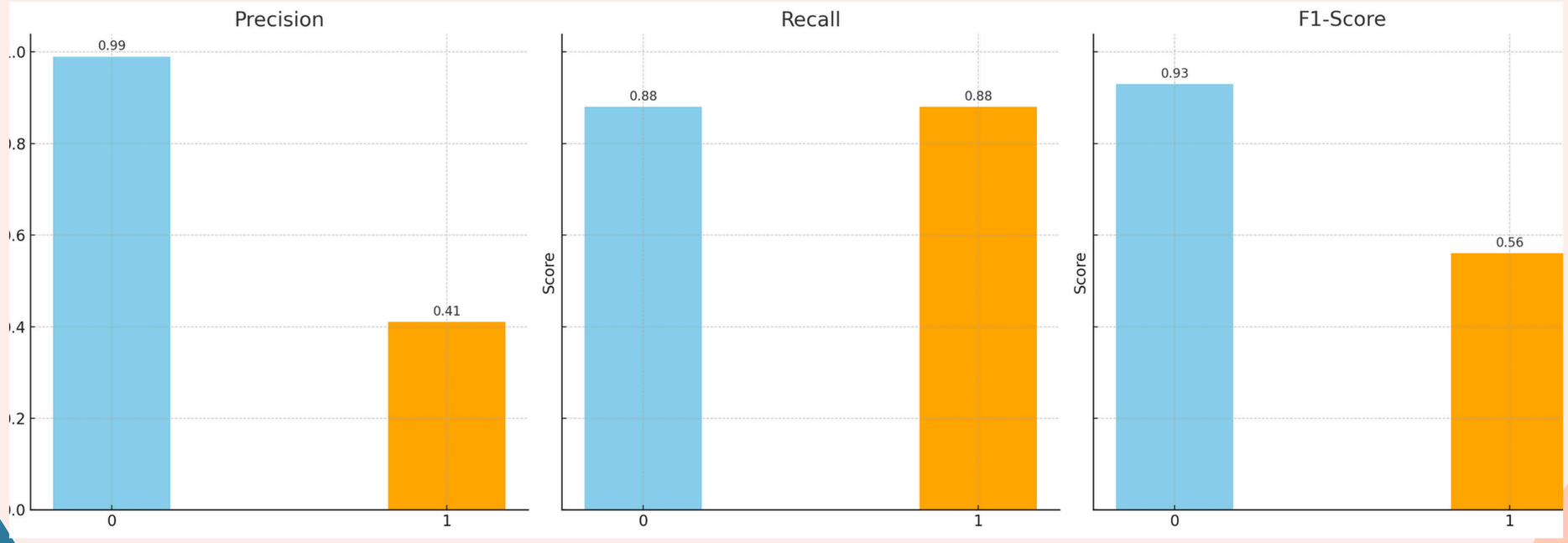
UPSAMPLING





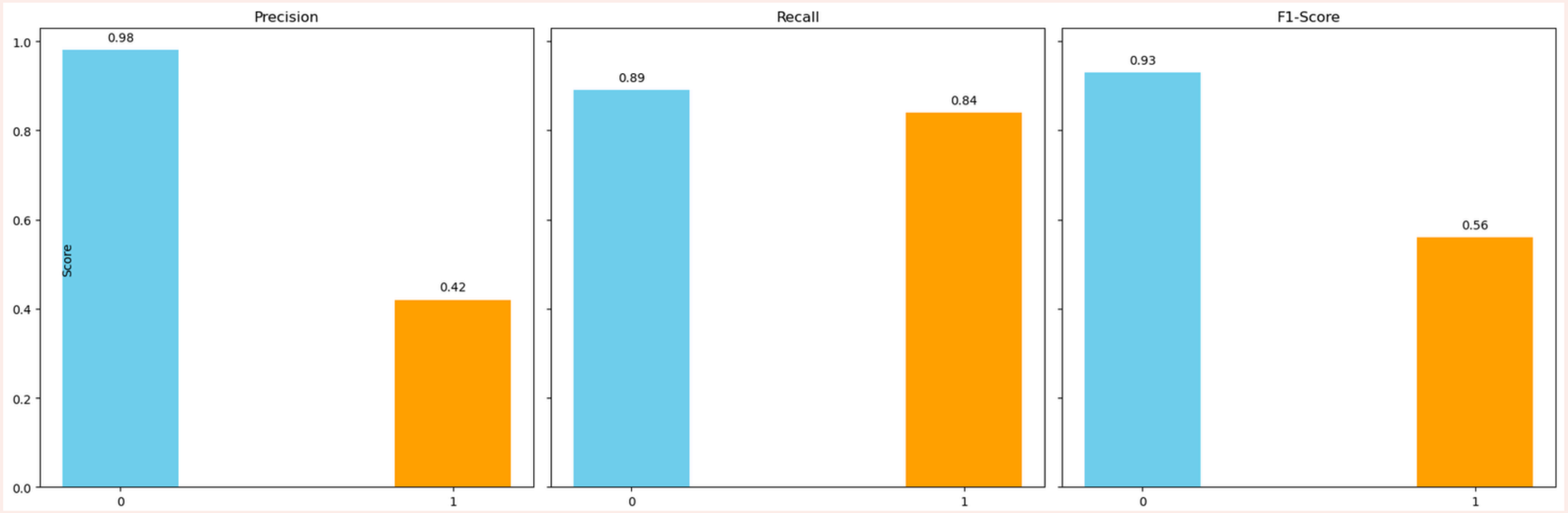
My support is balanced as 0 is 17562 and 1 is 1688

DOWNSAMPLING



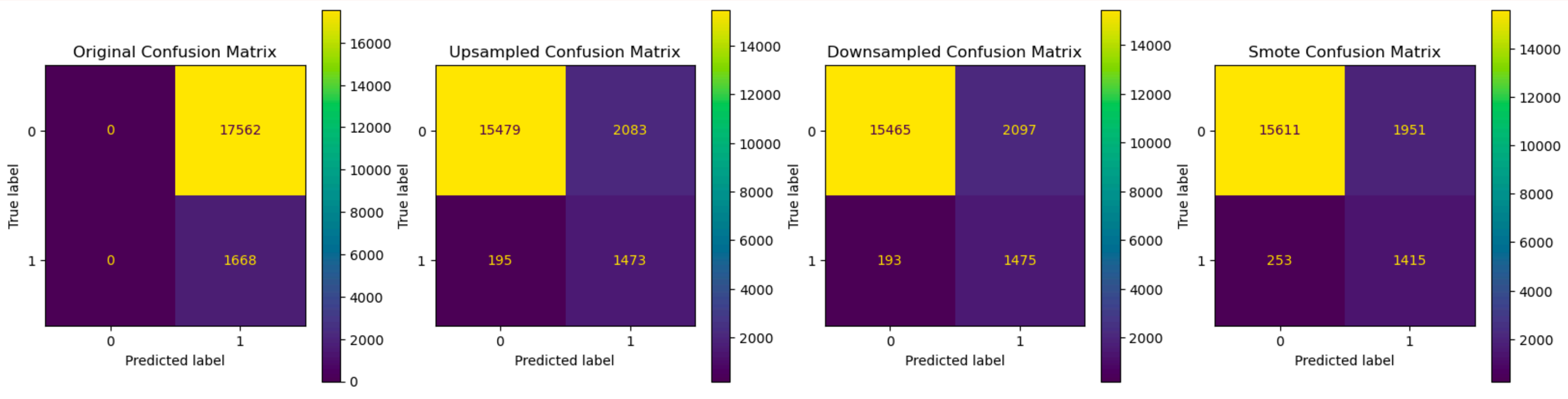
My support is balanced as 0 is 17534 and 1 is 1696

SMOTE



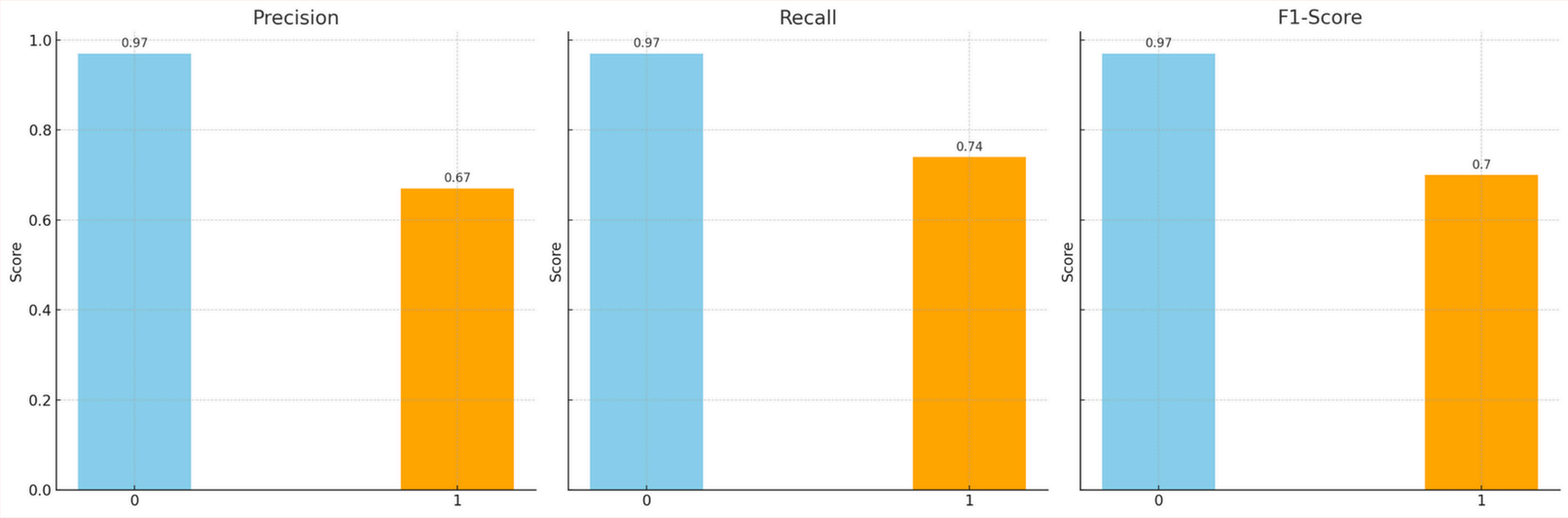
My support is balanced as 0 is 17439 and 1 is 17627

CONFUSIONMATRIXDISPLAY



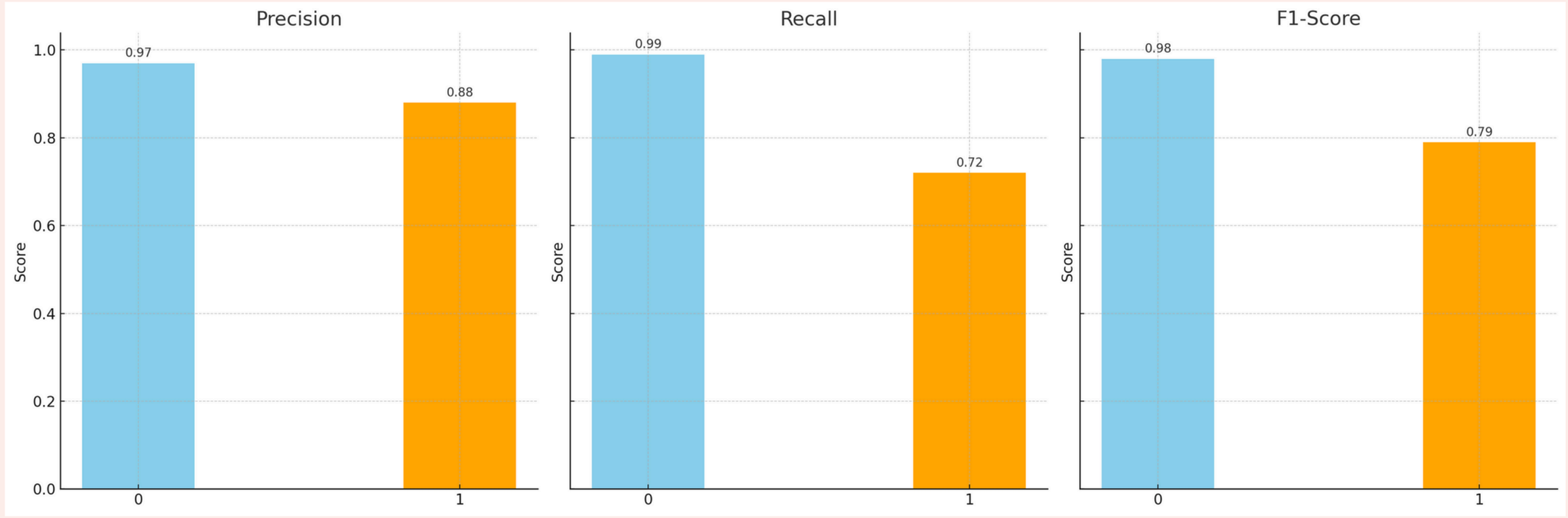
The SMOTE technique seems to be the most effective method for improving the logistic regression model's performance on this dataset, providing a good balance in classification and reducing the number of misclassifications.

Decision Tree with Smote

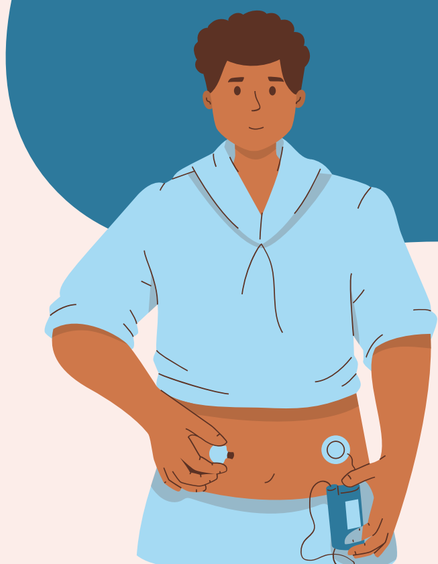


My support is balanced as 0 is 17534 and 1 is 1696

USE GRIDSEARCHCV FOR RANDOM FOREST, ADABOOST AND XGBOOST



XGBoost on the SMOTE dataset yields the best results.



THANK YOU!

Thank you so much for watching my presentation!
Do you have any questions, comments, or suggestions?