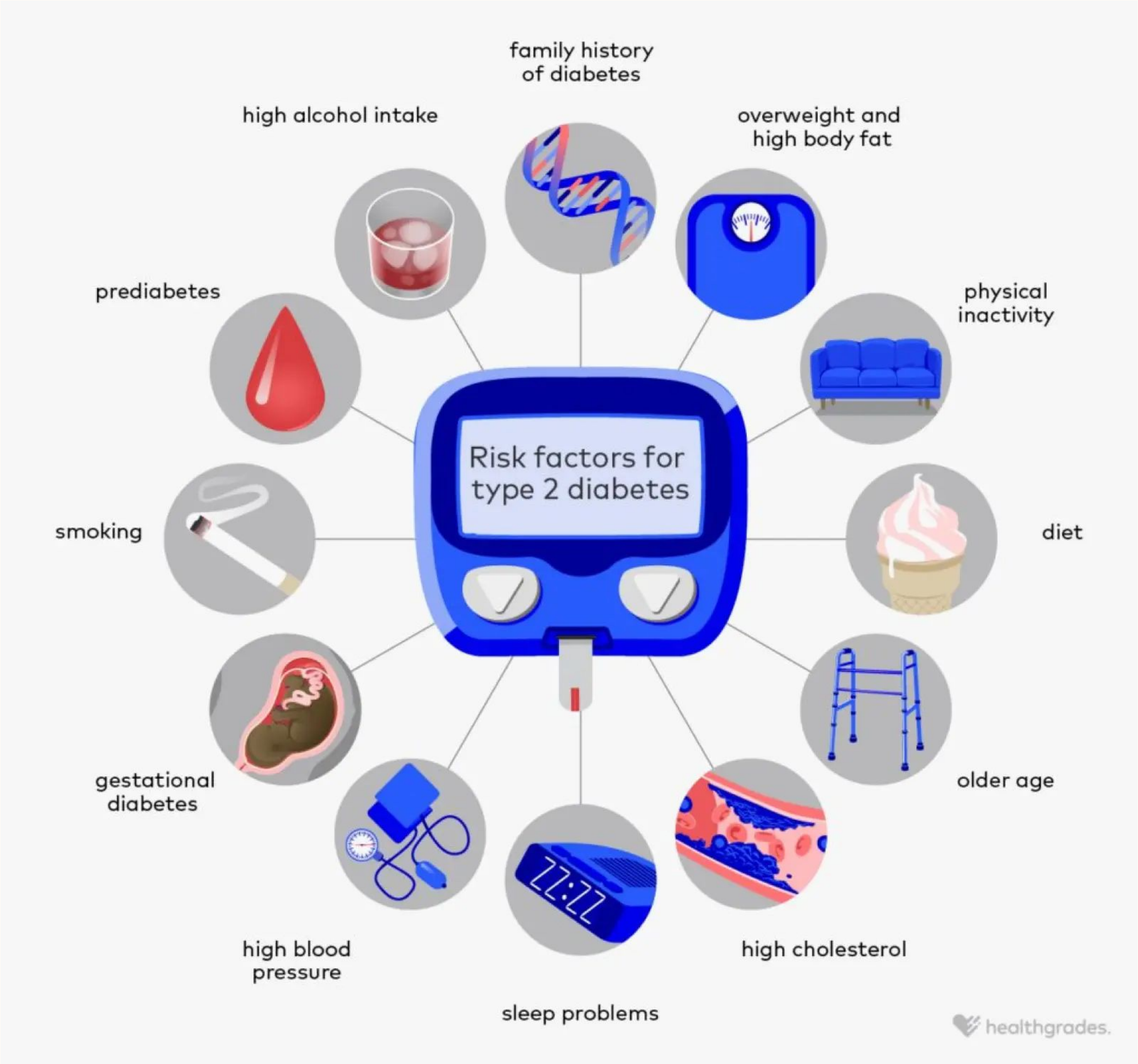


Source: [Risk Factors for Type 2 Diabetes: A Guide]

Predicting Diabetes Risk from Smoking Dataset

A Machine Learning Approach

Presenter: Sina, Xiaoqiao (Pamela), Farwa

Course: SCS_3253

Project Goal & Data

Goal

Predict whether an individual is at risk for diabetes given dataset of health features.

Dataset

15k initial rows. 23 features used.

could not use test.csv dataset from kaggle

Target Variable

0: No Risk (FBS < 100 mg/dL). 1: At Risk (100-125 mg/dL).

2: Prediabetic (FBS >125 mg/dL).

Source: Good news for those with type 2 diabetes: Healthy lifestyle matters (Harvard Health)

Data Exploration

Feature variables:

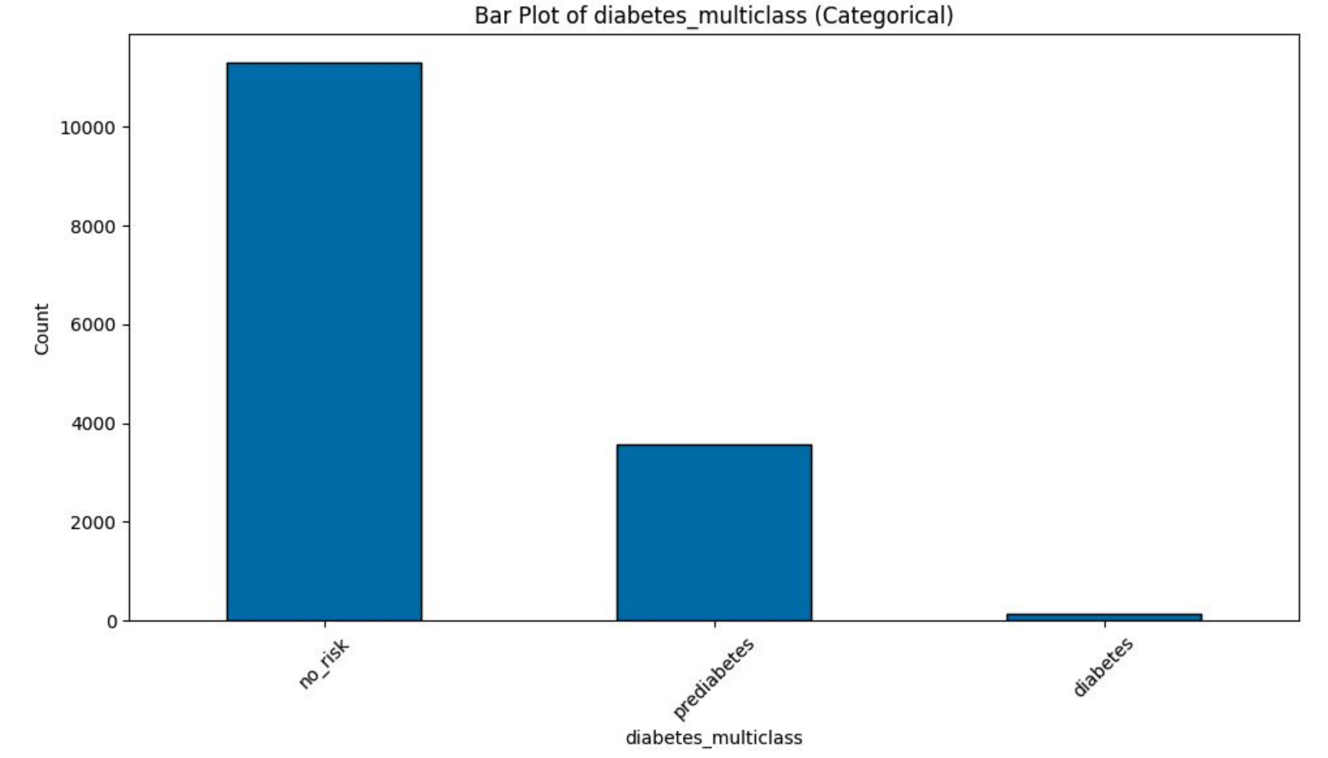
- 4 categorical features
- 18 numerical features

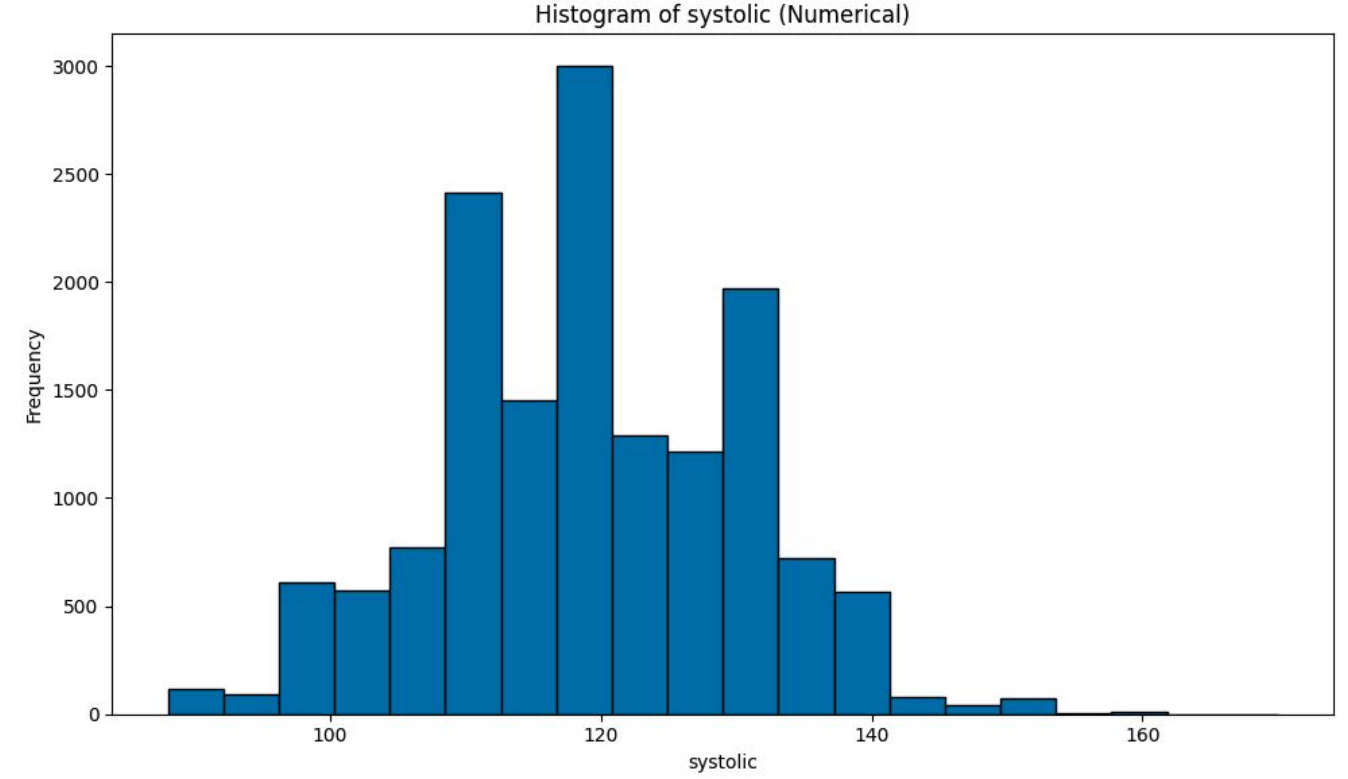
Labels:

- diabetes_multiclass

Observations:

- No missing value
- Some numerical features are skewly distributed. No apparent outliers found (Ex. Systolic blood pressure)





Age (5-year gap)
Height (cm)
Weight (kg)
Waist circumference (cm)
Eyesight (left)
Eyesight (right)
Hearing (left)
Hearing (right)
Systolic blood pressure
Diastolic blood pressure (relaxation)
Total Cholesterol
Triglyceride
HDL cholesterol
LDL cholesterol
Hemoglobin
Urine protein
Serum creatinine
AST (glutamic oxaloacetic transaminase)
ALT (glutamic oxaloacetic transaminase)
GTP (γ-GTP)
Dental caries
Smoking status

Challenges

- Significant class imbalance. 'Prediabetic' minority class (~0.92%) [138/15,000]
- Not removing Fasting Blood Sugar from training data (rule-based approach)
- SMOTE on entire dataset before splitting to training/testing
- Dataset was not intended for diabetes classification
- Feature Engineering did not provide any benefit to accuracy (Ex. BMI)

Methodology

Preprocessing

- SMOTE
- Feature Engineering (BMI calculation, height to weight ratio)
- Dropping unnecessary columns (ID, Fasting Blood Sugar)
- Handling Missing Data
- Label Encoding
- Feature Scaling (StandardScaler)
- Splitting training data into train/test (Kaggle had separate train/test)

Models Evaluated

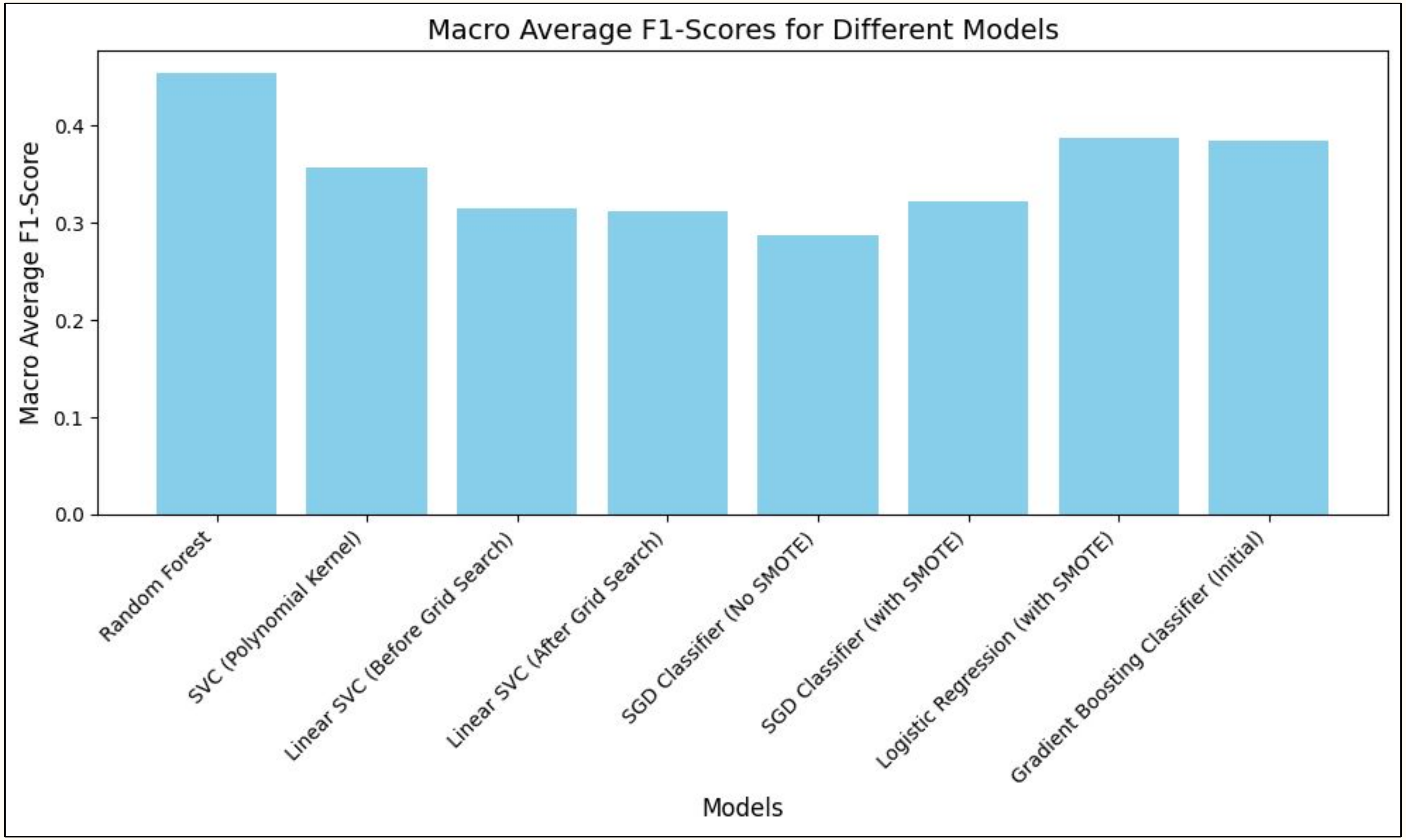
- Logistic Regression (LR)
- Support Vector Classifier (SVC)
- Random Forest (RF)
- XGBoost (XGB) - Often for imbalance dataset
- SGD Classifier
- Gradient Boost Classifier

Model	# F1-Score (weighted)	# Accuracy
XGBoost	0.328571	0.680769
Random Forest	0.370988	0.686538
SVC (RBF Kernel)	0.162983	0.657692
SVC (Polynomial Kernel)	0.365219	0.692308
Linear SVC	0.336727	0.684615
SGD Classifier (No SMOTE)	0.337317	0.682692
SGD Classifier (with SMOTE)	0.34966	0.676923
Logistic Regression (with SMOTE)	0.348107	0.678846
Gradient Boosting Classifier (Initial)	0.368829	0.688462
Gradient Boosting Classifier (with GridSearchCV &	0.369256	0.688462

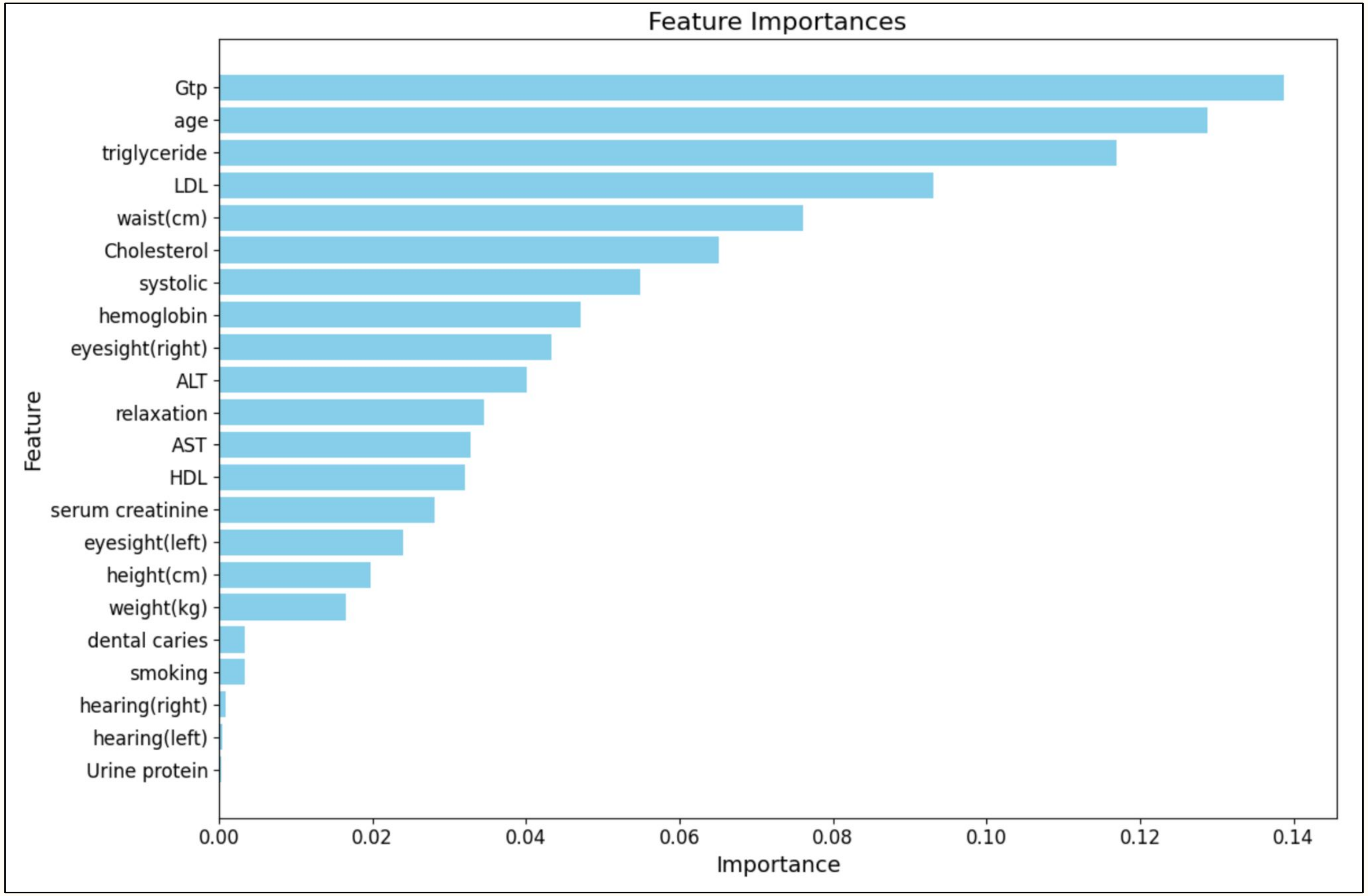
Imbalance Handling

- Auto Class Weighting

Macro Average Visualized



Feature Importance - Random Forest



Our Findings:

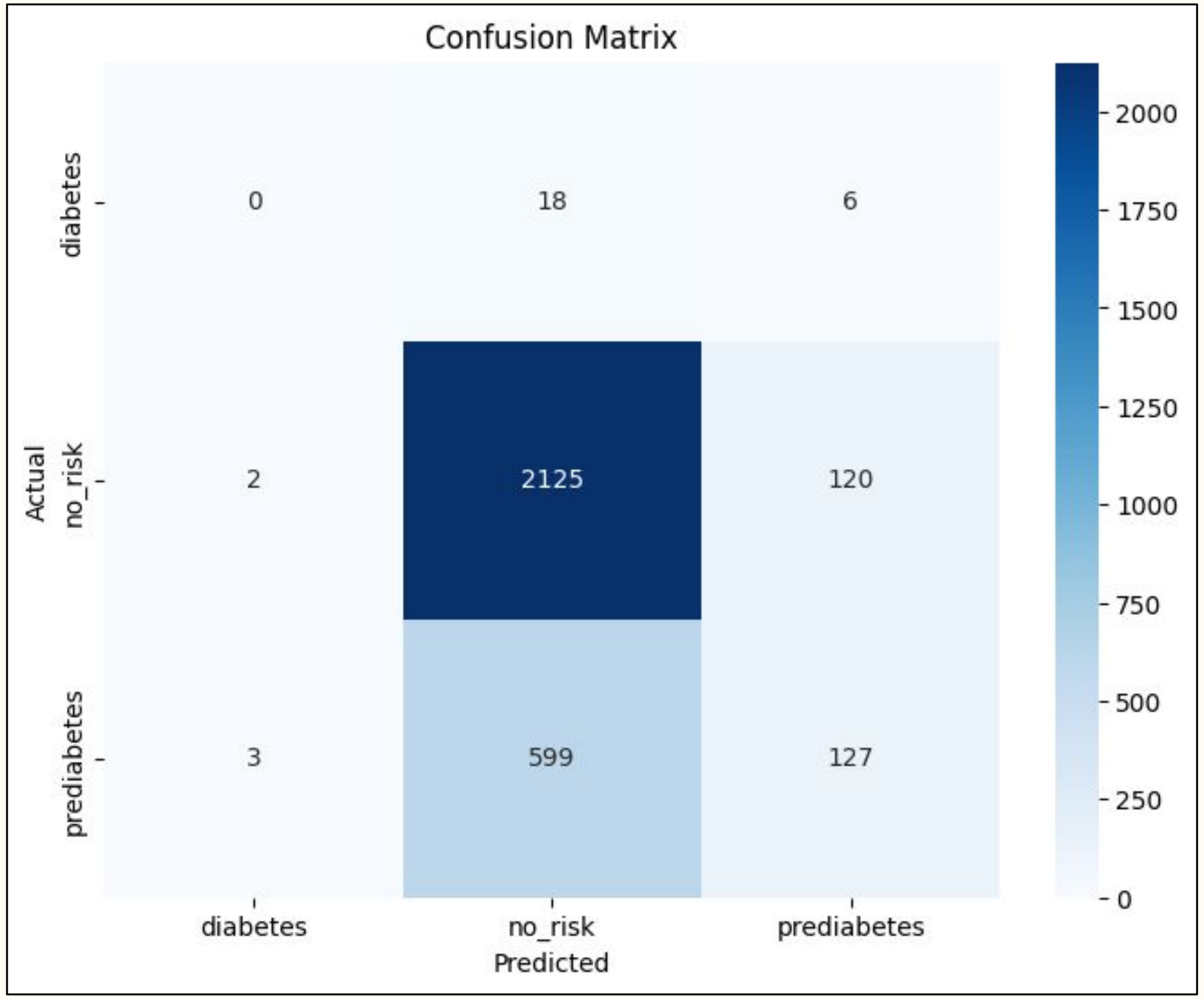
GTP stands for **Gamma-glutamyl transferase (GGT)**. It's an enzyme that's primarily found in the liver but is also present in other organs like the kidneys and **pancreas**.

High levels of GTP can indicate liver damage or disease.

Our model rates GTP as of the highest importance.

XGBoost Results

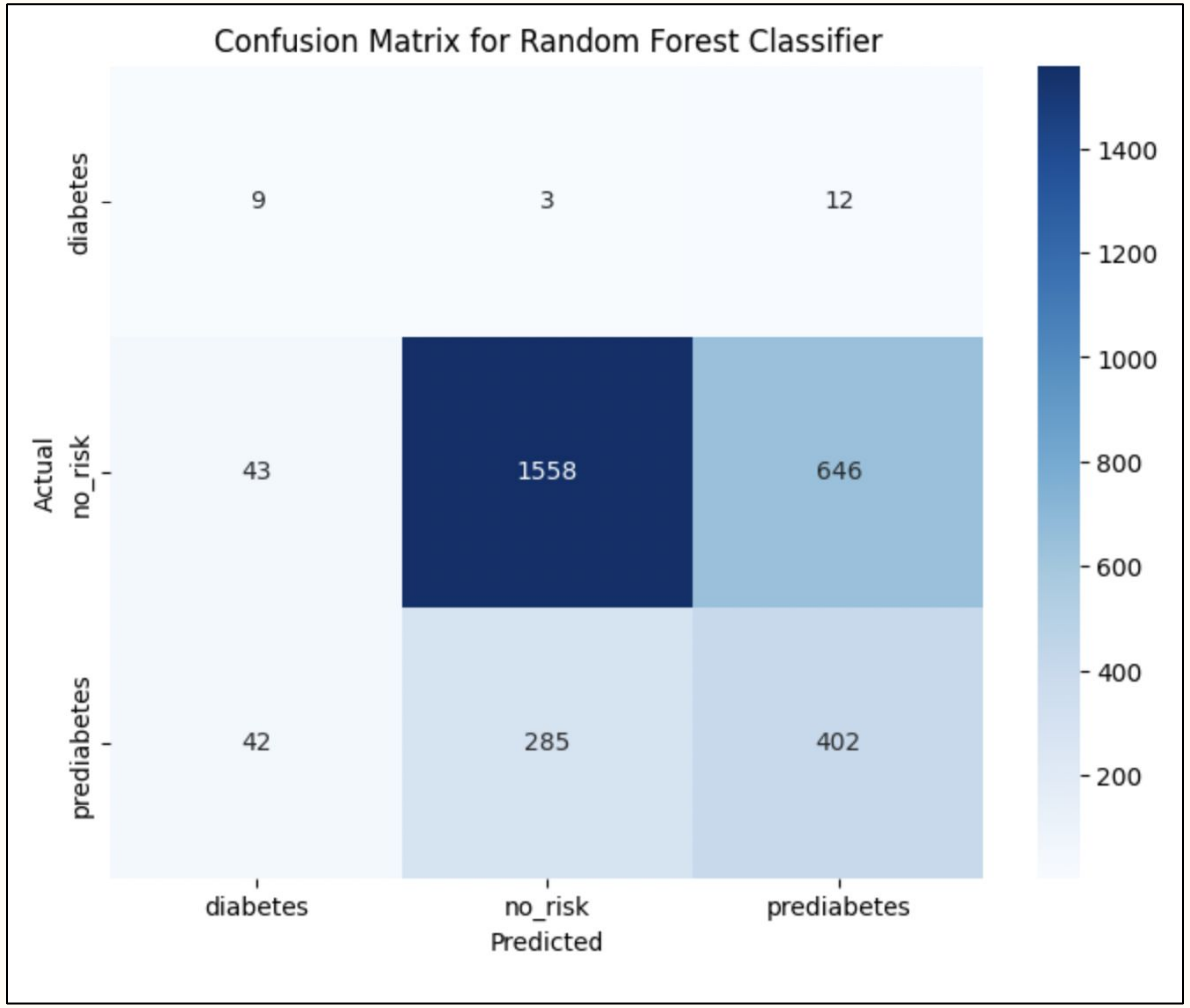
```
Classification Report:

              precision    recall  f1-score   support

    diabetes       0.00      0.00      0.00        24
     no_risk       0.77      0.95      0.85      2247
 prediabetes       0.50      0.17      0.26       729

    accuracy                           0.75      3000
   macro avg       0.43      0.37      0.37      3000
weighted avg       0.70      0.75      0.70      3000


Confusion Matrix:

[[   0   18    6]
 [   2 2125  120]
 [   3  599  127]]
```



- Misleading Accuracy
- Adjusting Classes to only no_risk and diabetes

Random Forest Results

```
Classification Report:

              precision    recall  f1-score   support

    diabetes       0.10      0.38      0.15        24
     no_risk       0.84      0.69      0.76      2247
 prediabetes       0.38      0.55      0.45       729

    accuracy                           0.66      3000
   macro avg       0.44      0.54      0.45      3000
weighted avg       0.73      0.66      0.68      3000


Confusion Matrix:

[[   9    3   12]
 [  43 1558  646]
 [  42  285  402]]
```



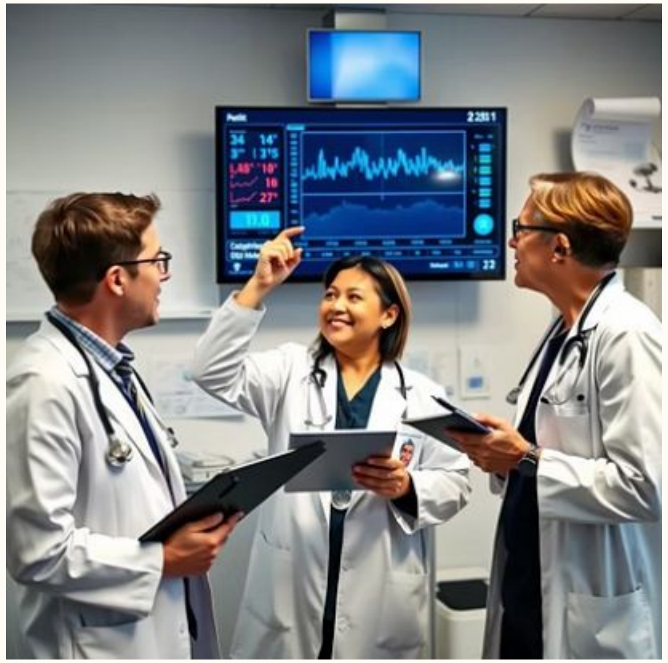
- Auto class weighting (class_weight='balanced')

Conclusion & Next Steps

Next Steps

- Explore other models for imbalanced dataset
- More Rigorous Hyperparameter Tuning (RandomizedSearchCV, Bayesian Optimization)
- A dataset related to diabetes with more balanced classes
- Other methods to correct for imbalance
- Selecting most important features
- More feature engineering

Questions?



References

1. Smoking dataset: https://www.kaggle.com/competitions/binary-smoke-detector
2. Gamma for Presentation template