PREDICTING DIABETES

By Jacob Wilson, Sravya Murala

INTRODUCTION

Diabetes affects 800+ million people globally, with rising prevalence in developing nations. Early detection is critical for disease management, yet traditional diagnostic methods rely on invasive lab tests. This project leverages machine learning to enable screening using accessible demographic and lifestyle factors.

PROBLEM STATEMENT

- Can machine learning predict diabetes presence using only demographic and lifestyle data?
- We compare two approaches: binary classification (Diabetes vs. No Diabetes) and multi-class classification (five diabetes types).
- Which approach achieves better predictive performance for clinical screening?

DATASET

100,000
Records

30+
Features

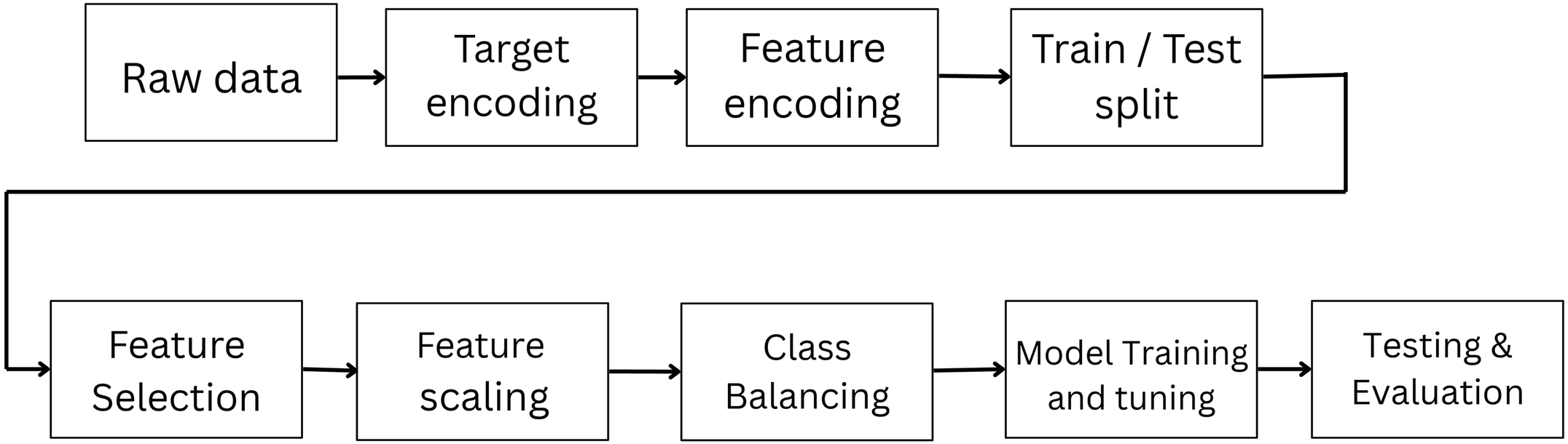
Demographic Data

Clinical Measurements

Lifestyle Habits

Family History

APPROACH


Raw data
Target encoding
Feature encoding
Train / Test split
Feature Selection
Feature scaling
Class Balancing
Model Training and tuning
Testing & Evaluation

EXECUTION

- **Mutual Information:** Non-parametric, captures non-linear relationships
- **Random Forest Importance:** Gini-based mean decrease in impurity ranking
- **Voting Consensus:** Intersection of top 15 from both methods
- **Final Feature Space:** 39 → 10 features (74% dimensionality reduction)
- **Benefits:** Prevents overfitting, improves interpretability, reduces computational cost

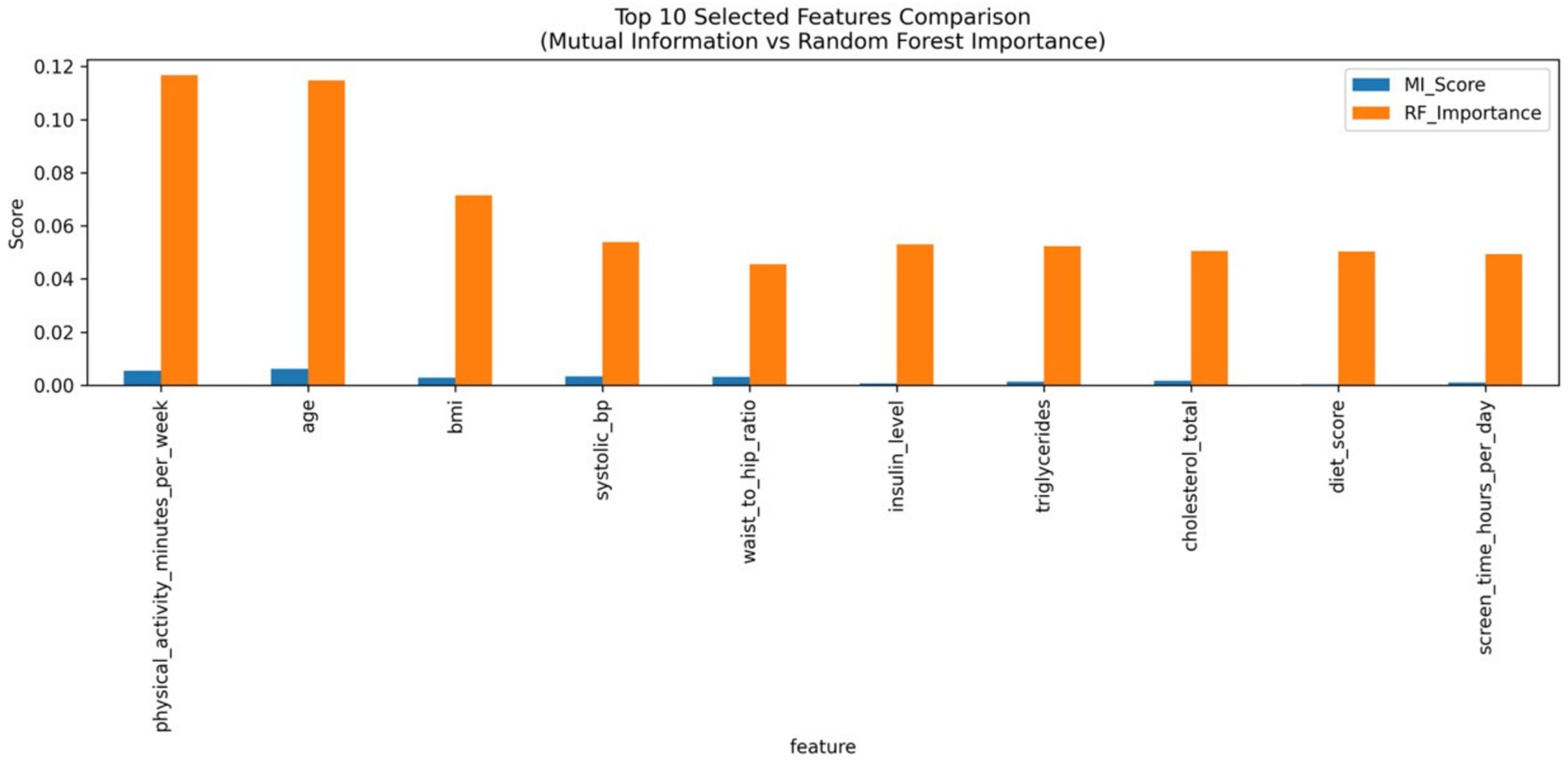


Fig: Mutual Information vs. Random Forest Importance

- Model Training + Testing:
  - Random Forest Classifier
  - XGBoost Classifier
  - CatBoost Classifier

- Hyperparameter Tuning:
  - Cross-validation
  - Random Forest and XGBoost tuned with optimal parameters

BINARY CLASSIFICATION RESULTS

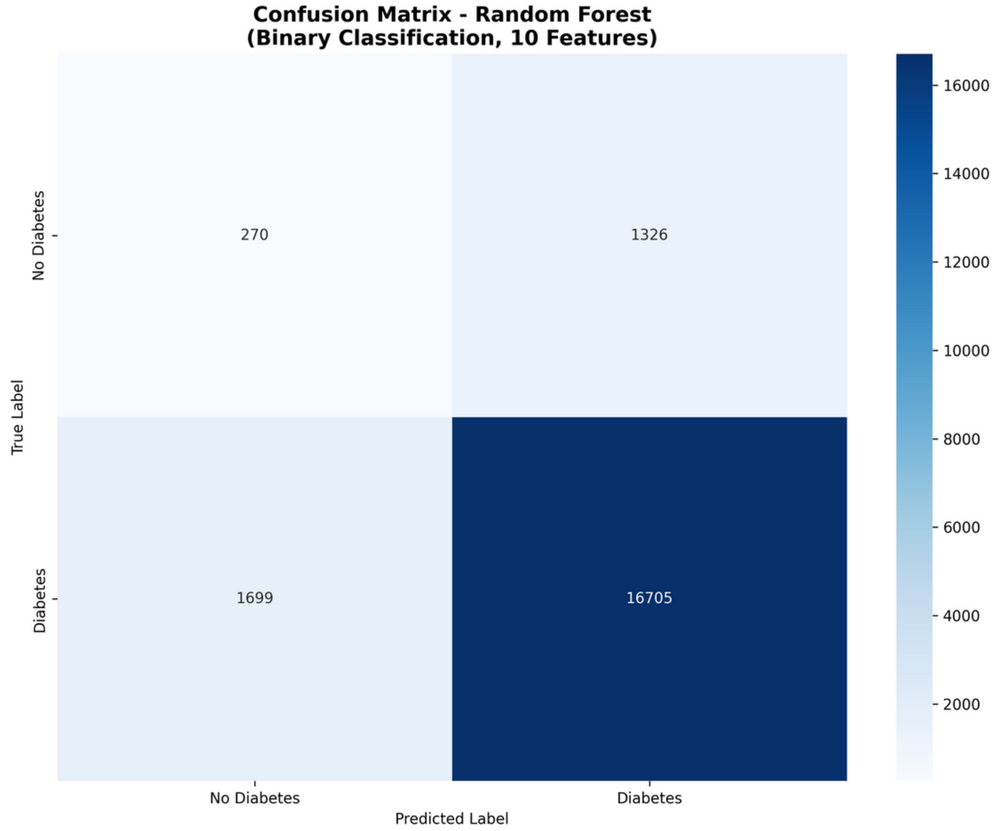
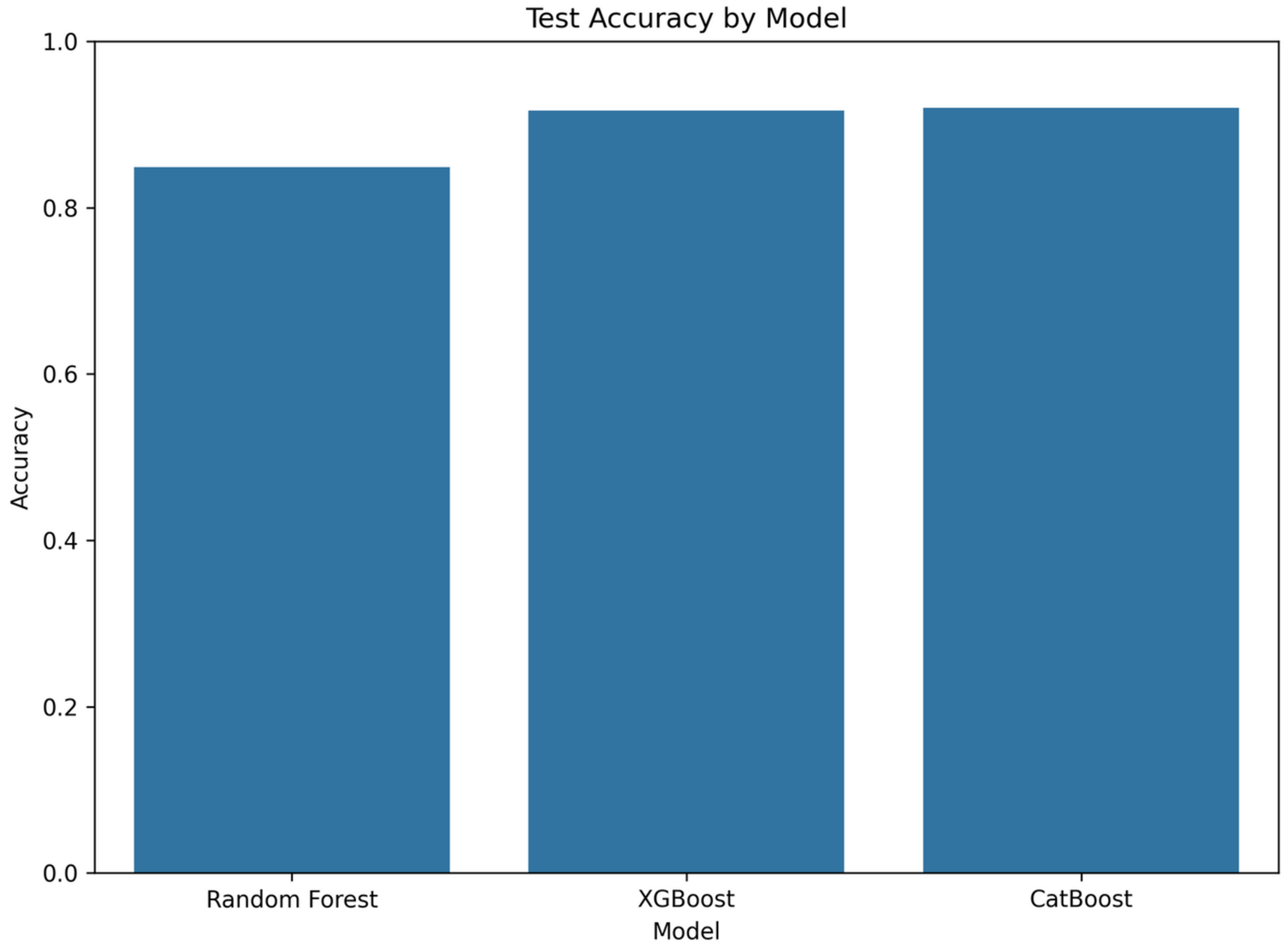
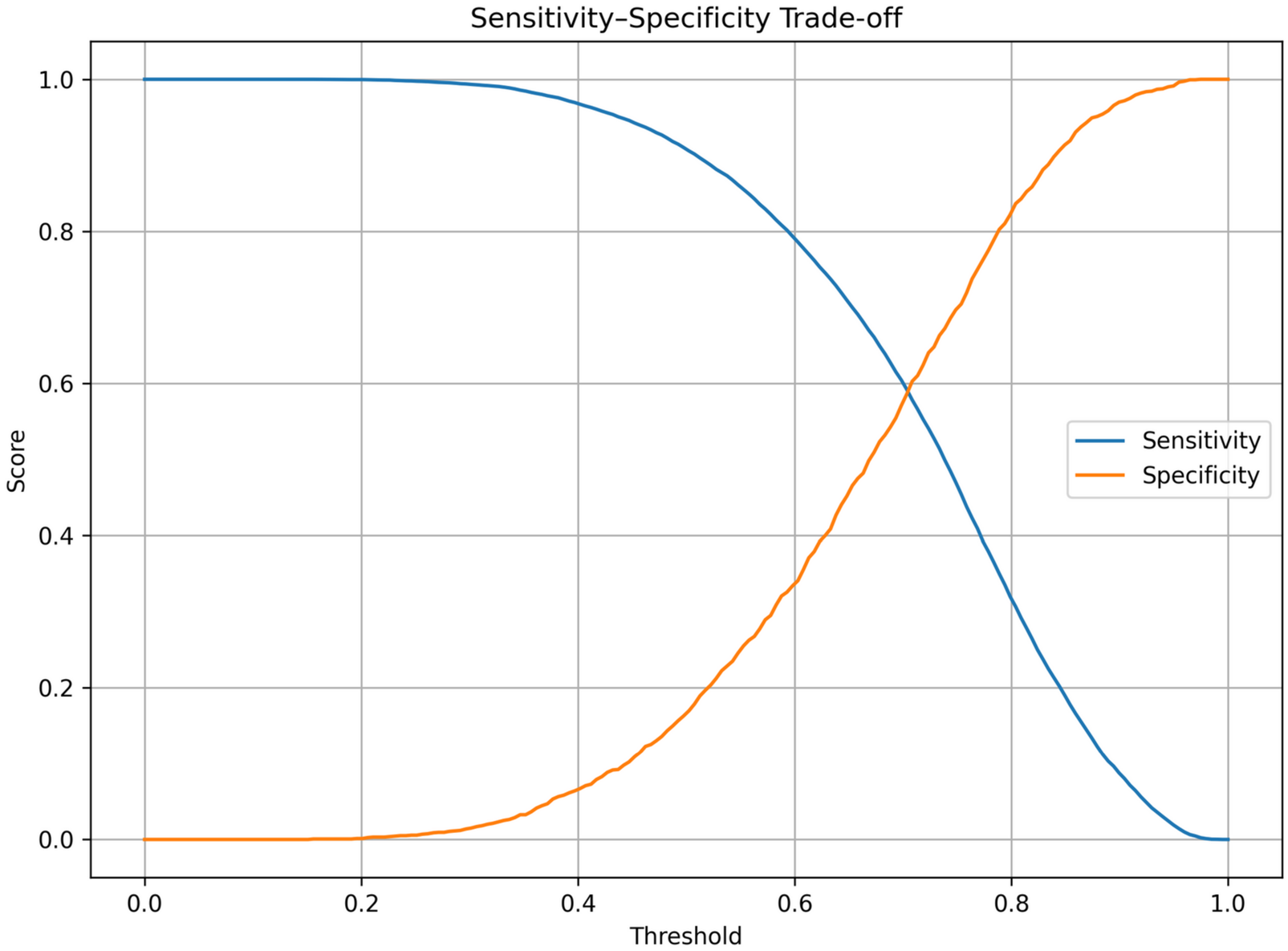


Fig: Confusion Matrix- Binary classification


Test Accuracy by Model
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
Random Forest
XGBoost
CatBoost
Model


Sensitivity–Specificity Trade-off
Score
Threshold
Sensitivity
Specificity
0.0
0.2
0.4
0.6
0.8
1.0

MULTI-CLASS CLASSIFICATION RESULTS

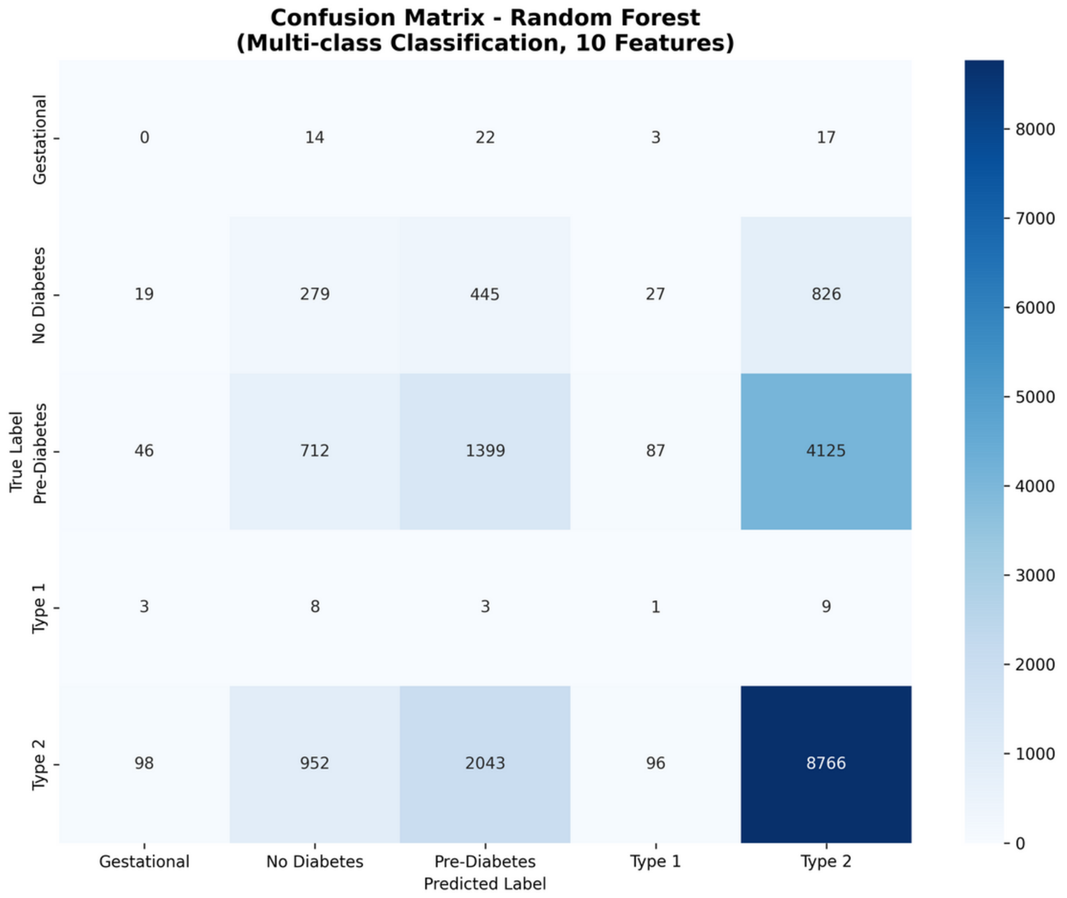
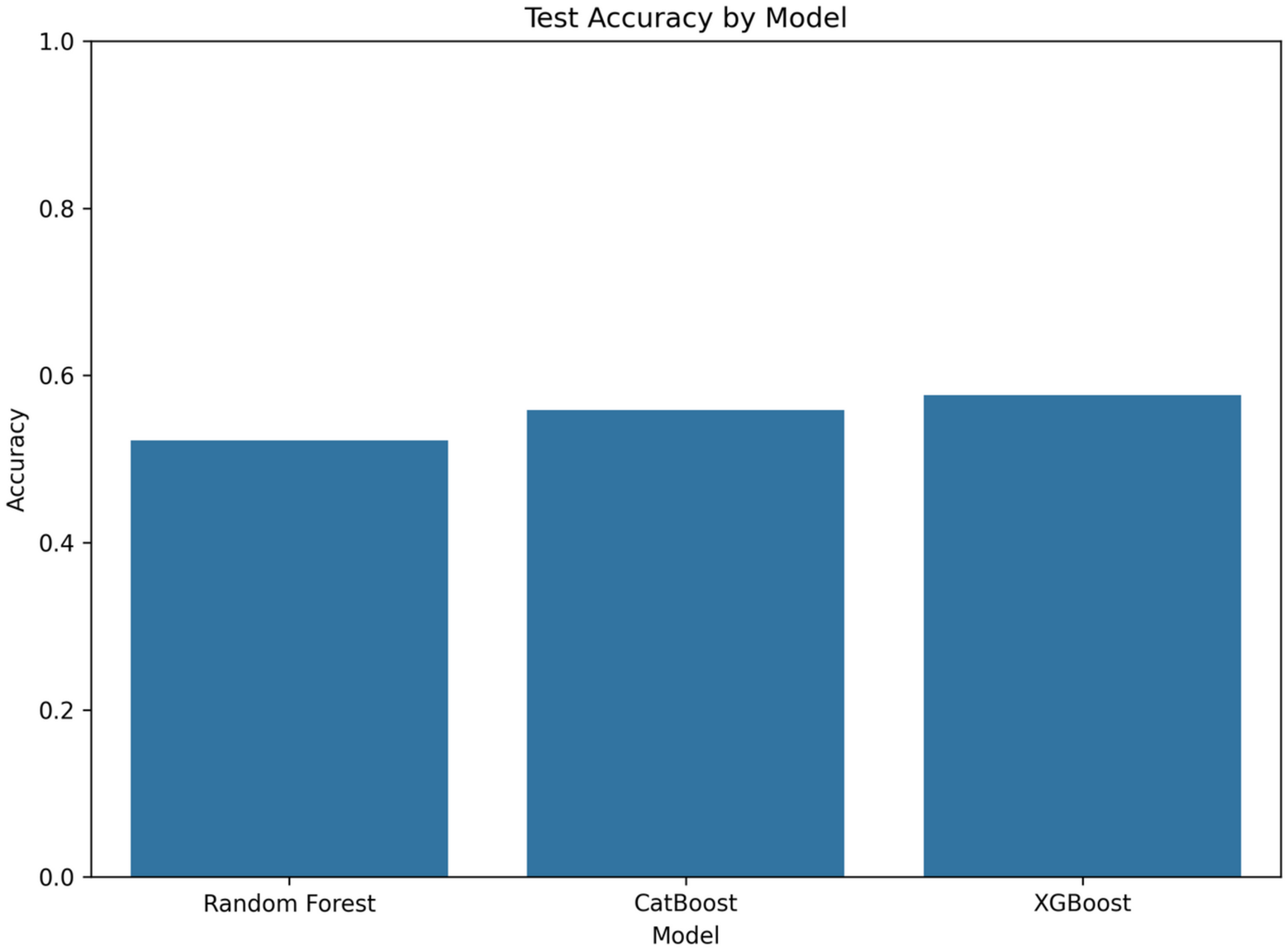
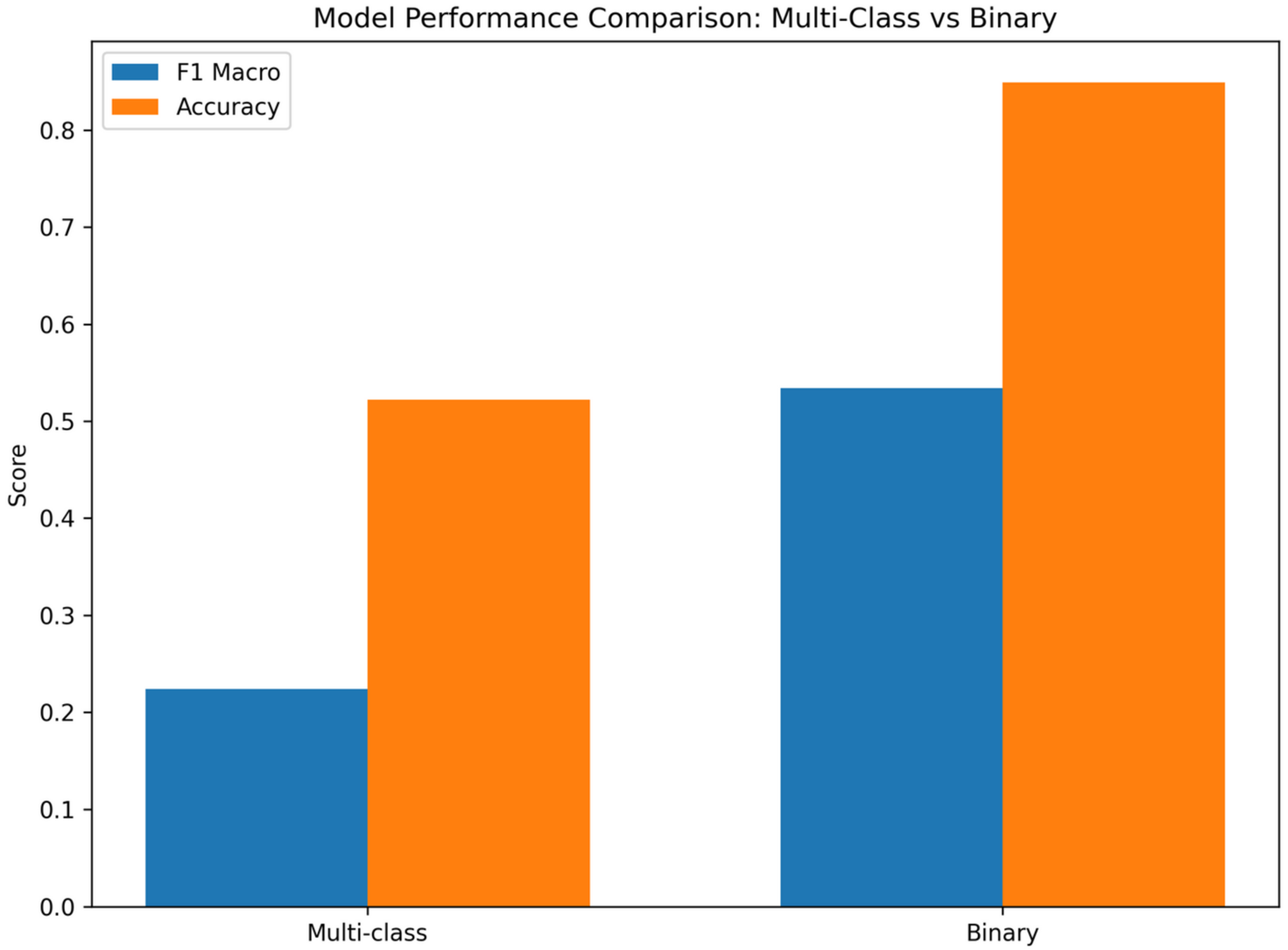


Fig: Confusion Matrix- Mutli classification


Test Accuracy by Model
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
Random Forest
CatBoost
XGBoost
Model

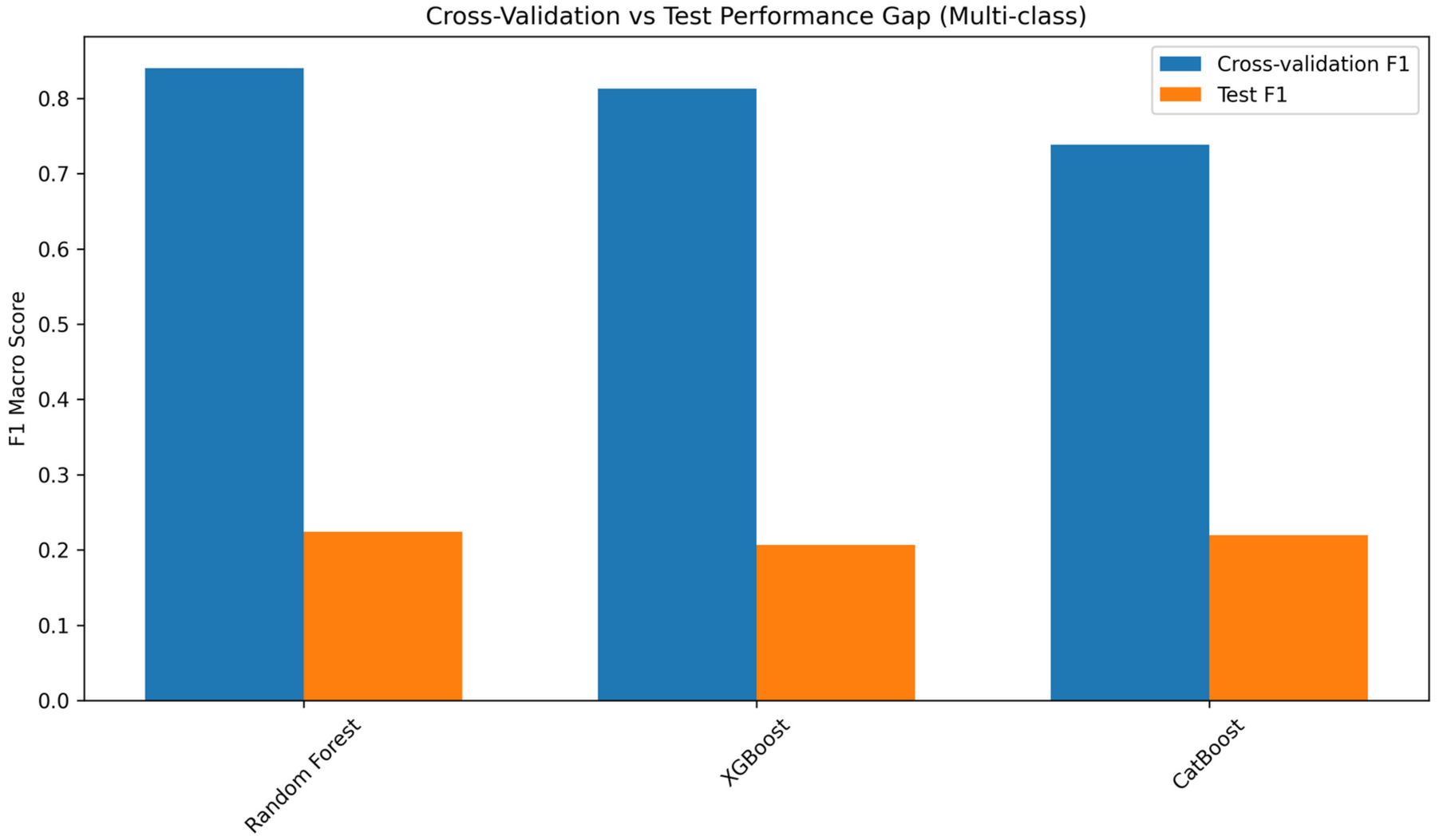

Model Performance Comparison: Multi-Class vs Binary
F1 Macro
Accuracy
Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Multi-class
Binary

KEY FINDINGS

1. Binary classification is clinically viable but not recommended
   - 84.88% accuracy, 91% sensitivity
   - Suitable for population screening

2. Multi-class exceeds data capability
   - 52.23% accuracy is insufficient
   - Demographic features can't differentiate types

3. Problem formulation matters most
   - 32% accuracy gap = feature limitation, not algorithm failure
   - No algorithm can overcome insufficient data


Cross-Validation vs Test Performance Gap (Multi-class)
Cross-validation F1
Test F1
F1 Macro Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Random Forest
XGBoost
CatBoost

CONCLUSION

Binary classification achieved 84.88% accuracy and 91% sensitivity, suitable for population screening. The performance gap reflects problem difficulty, and demographic features identify diabetes presence but cannot differentiate types without biomarkers. The model functions as a first-stage screening tool requiring confirmatory testing.

THANK YOU