Predicting Heart Disease Severity using Machine Learning

By Yashas Raman, Raghav Sriram
Group 14

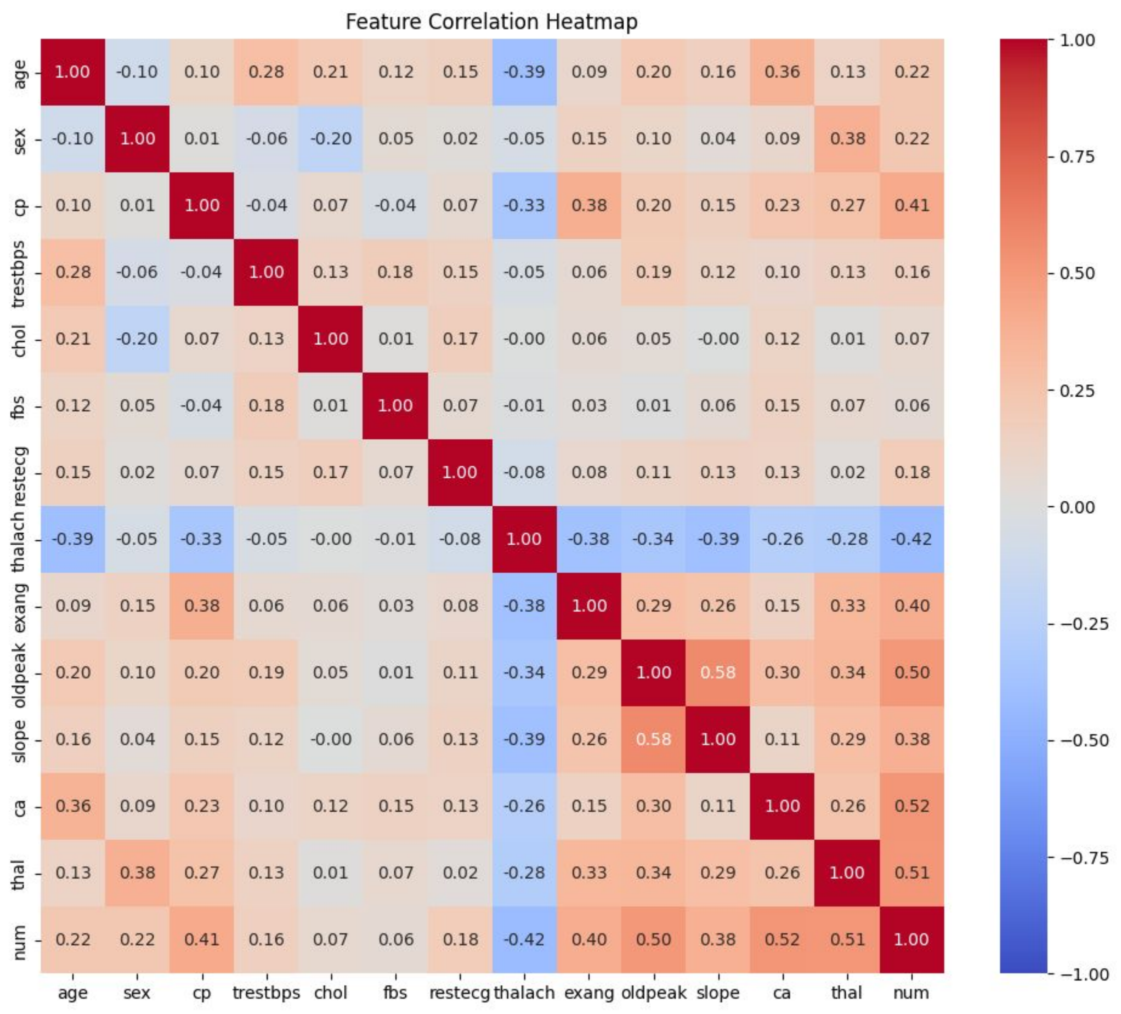
Why Predict Heart Disease?

- Heart disease is a leading cause of death in humans
- Improve detection methods
- Project Goal: Detect presence and severity of heart disease from clinical features
- Tasks:
  - Binary classification: disease or no disease
  - Multiclass classification: severity (0–4 scale)

Research Methods

Dataset & Preprocessing

- UCI Heart Disease dataset (303 patients)
- 13 Features
  - Age, sex, chest pain type, cholesterol, max heart rate, etc.
- Preprocessing:
  - Missing value imputation
  - Feature scaling



Feature Correlation Heatmap
age
sex
cp
trestbps
chol
fbs
restecg
thalach
exang
oldpeak
slope
ca
thal
num
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

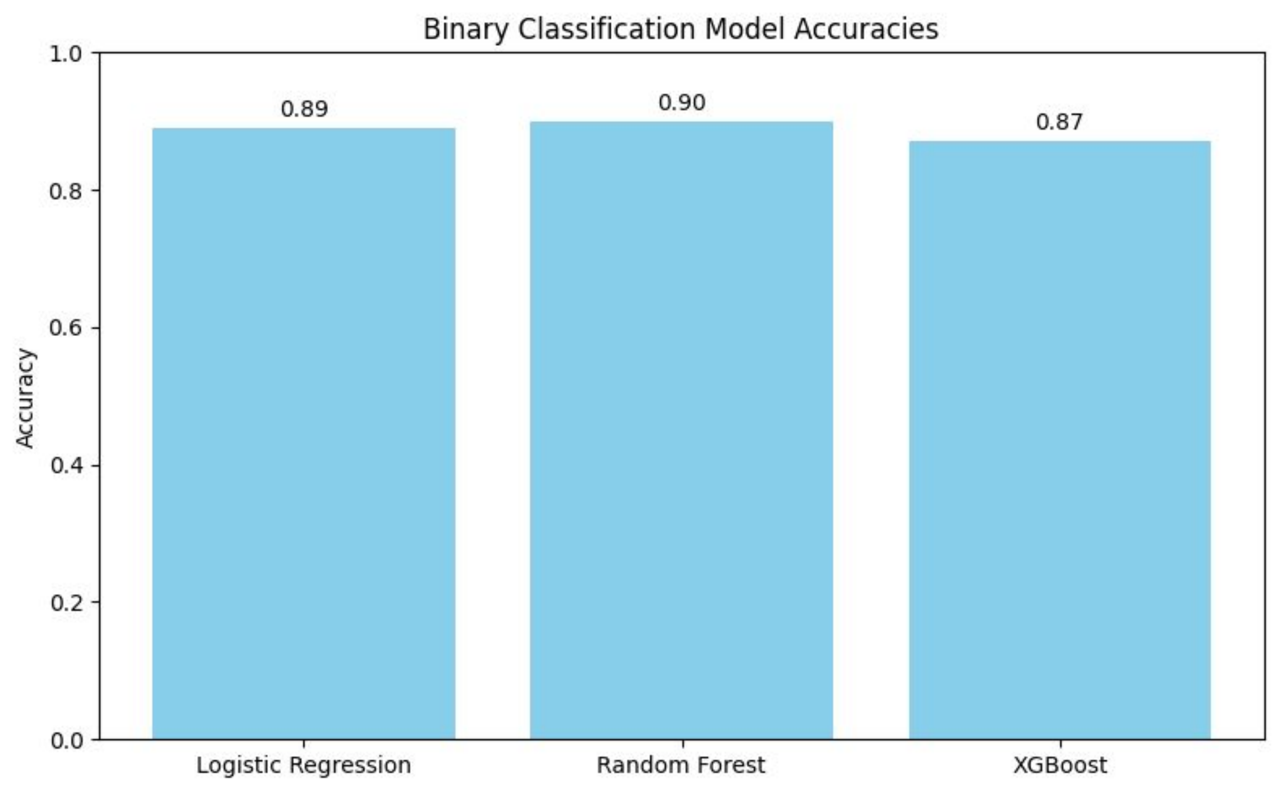
Model Development and Evaluation

- Used an 80/20 train-test split to train and evaluate our models
- Models Used:
  - Logistic Regression
  - Random Forest
  - XGBoost
  - Applied to both binary and multiclass tasks

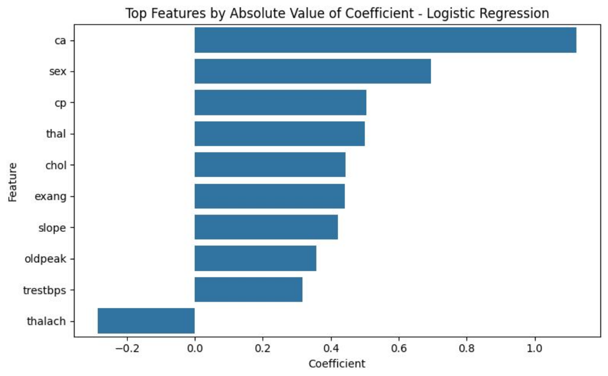
```
"Logistic Regression": LogisticRegression(max_iter=1000),
"Random Forest": RandomForestClassifier(n_estimators=100, random_state=42),
"XGBoost": XGBClassifier(eval_metric='logloss')
```

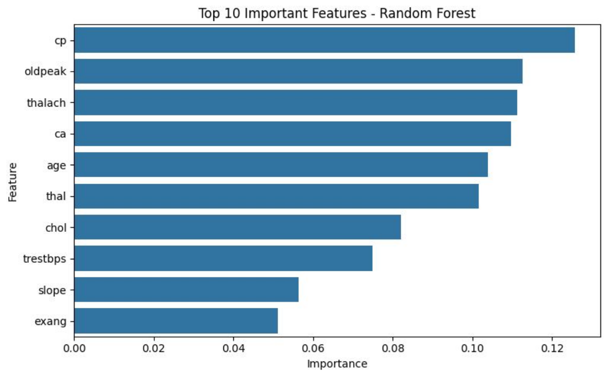
Results and Analysis

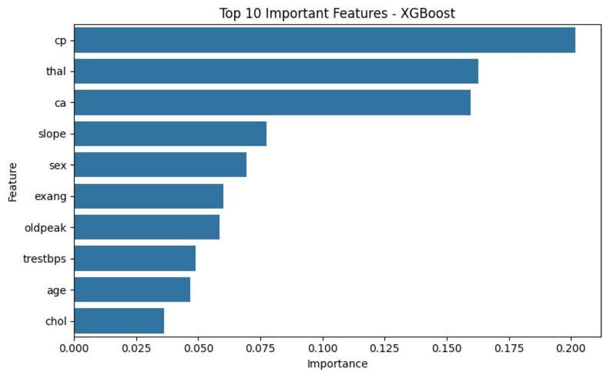
Binary Classification Results



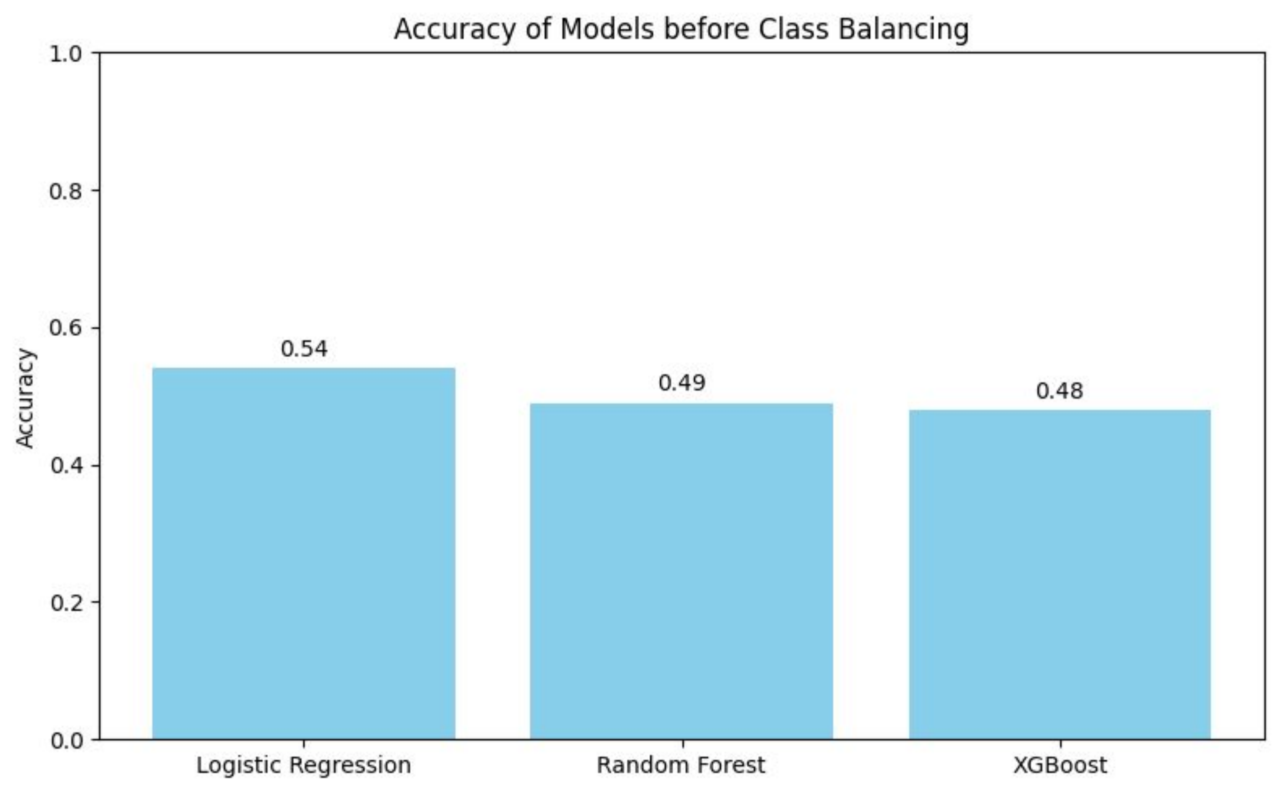
Most Relevant Features



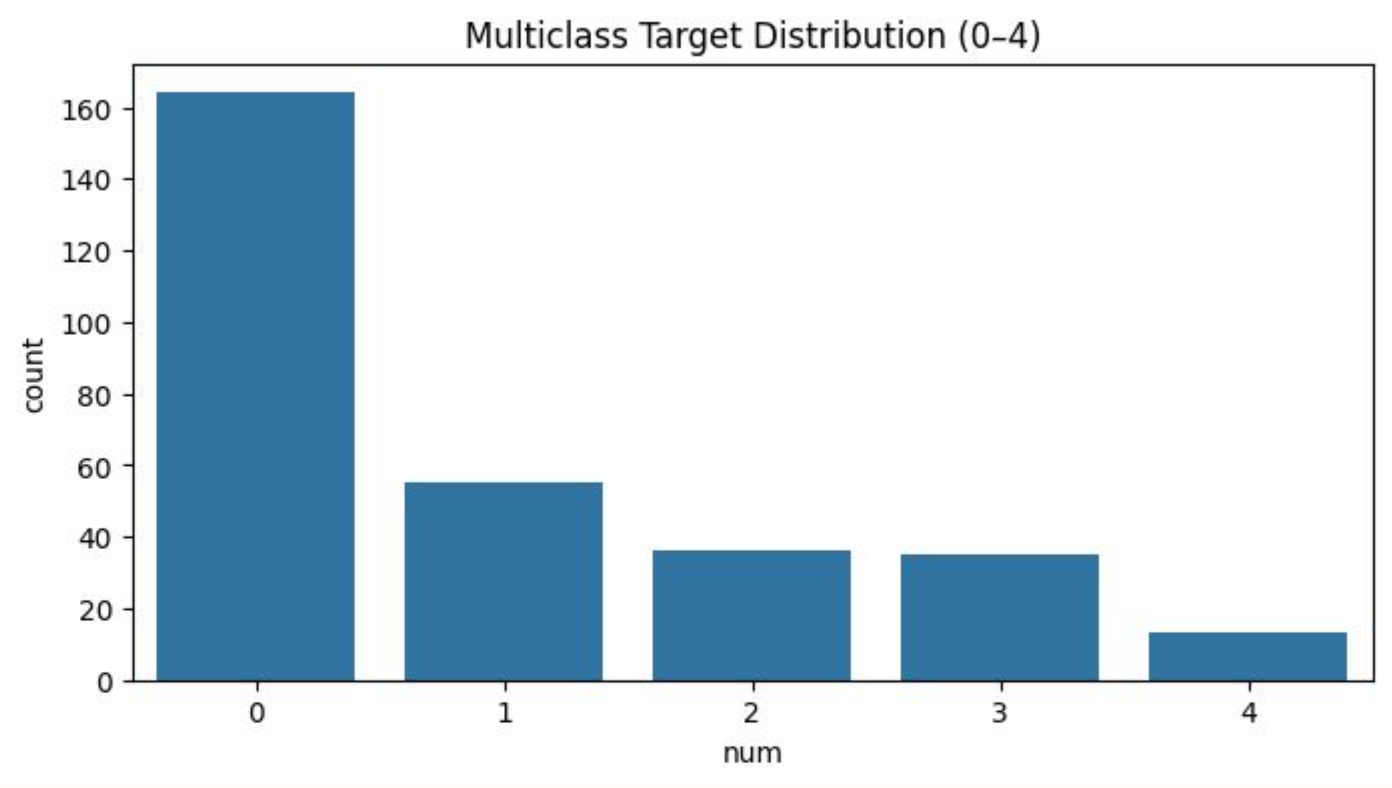




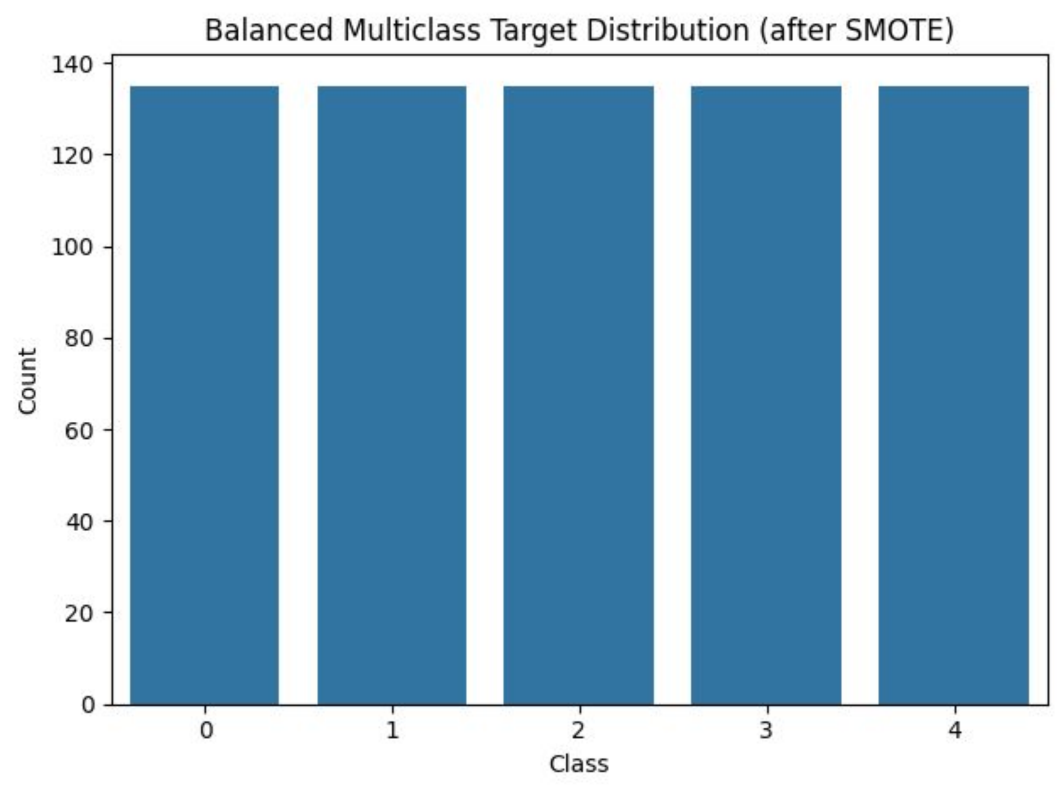
Results for Multiclass Prediction



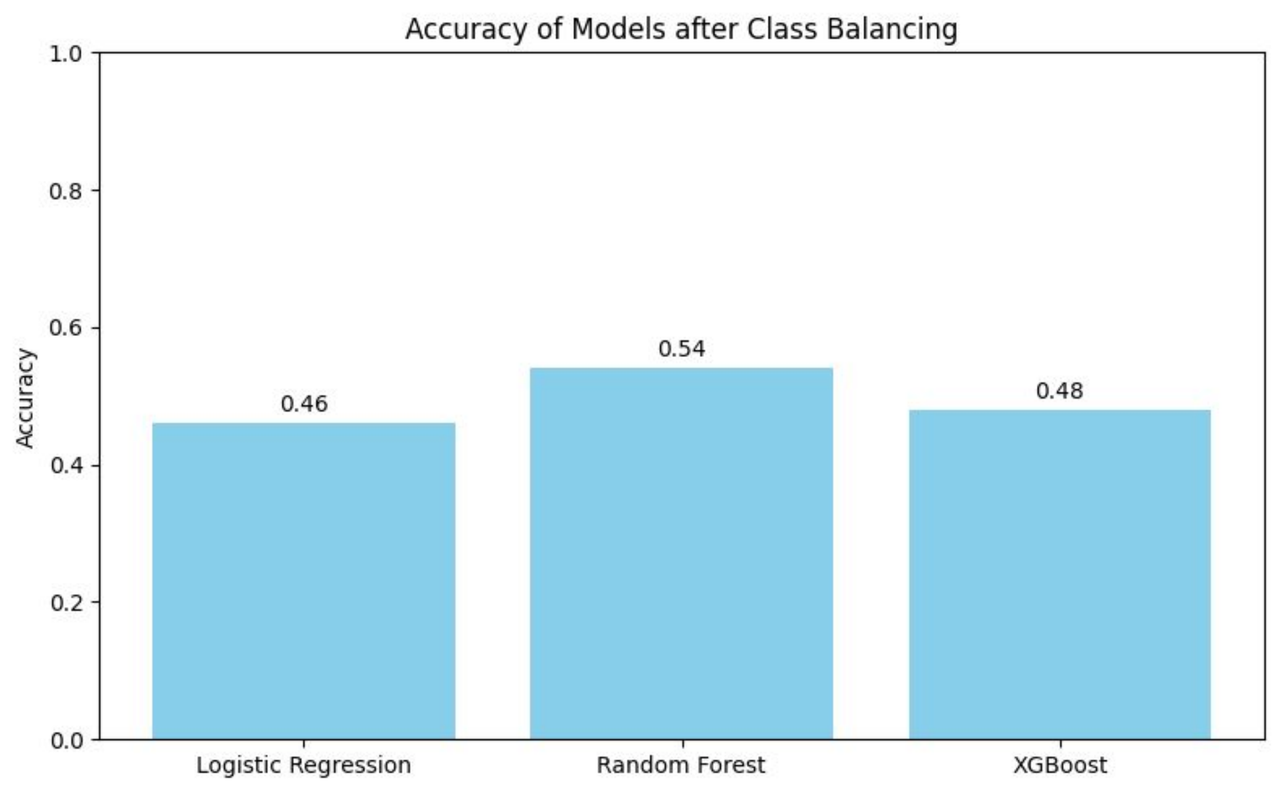
Unbalanced Multiclass Dataset



Balancing Dataset via Oversampling



Multiclass results after Class Balancing



Analysis

- All three models performed relatively similarly across binary and multiclass prediction
- Random Forest slightly improved after class balancing while Logistic Regression got slightly worse
    - Overall, class balancing did not help model performance
- All models performed very well on binary while significantly worse on multiclass

Discussion and Reflection

What We Learned

- Machine learning can be very useful in detecting presence of heart disease
- Severe class imbalance likely harmed performance
- Multiclass prediction looks to be a difficult problem in general
- Relatively small sample size could have limited model generalization

Future Directions

- Potentially use other models
  - SVM
  - Neural Networks
- Attempt methods on more balanced dataset
- Apply to other diseases