Heart Disease Prediction

Group 5: Vicky Jin; Mario Ma; Yunze Wang; Jiaqi Wu

1. Introduction to the Task and Dataset

The global impact of cardiovascular diseases remains paramount, with millions of fatalities every year. Our project leverages a Kaggle dataset to predict the individual's previous heart problems. The dataset comprised 8,763 samples with 25 variables, with the dependent variable indicating the presence of heart disease. Independent variables included age, sex, cholesterol, blood pressure, and others. This report will detail our data-driven approach, from preprocessing through to machine learning analysis and conclude with our findings on heart disease predictors.

---

2. Data Preprocessing and Exploration

To prepare our dataset, we ensured there were no missing values and normalized the data prior to analysis. Key variables like age, blood pressure, and exercise hours were binned to categorize continuous data effectively. We further refined the data by converting blood pressure readings to pulse pressure and eliminating non-analytical variables like Patient ID. Initial exploration via distribution analysis indicated a fair balance among the variables, though some skew was noted. One-hot encoding transformed categorical data, and Pearson and ANOVA tests examined correlations.

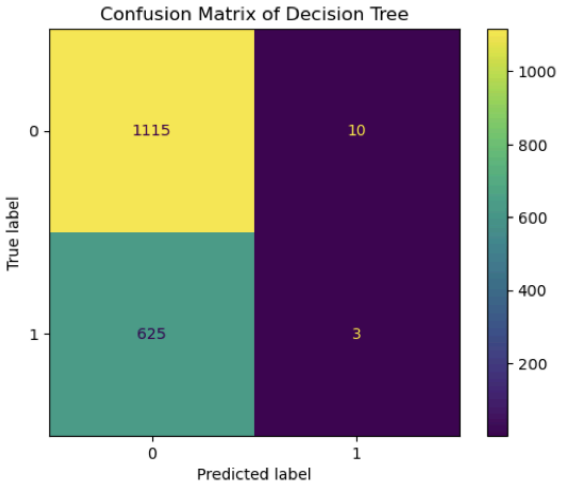
---

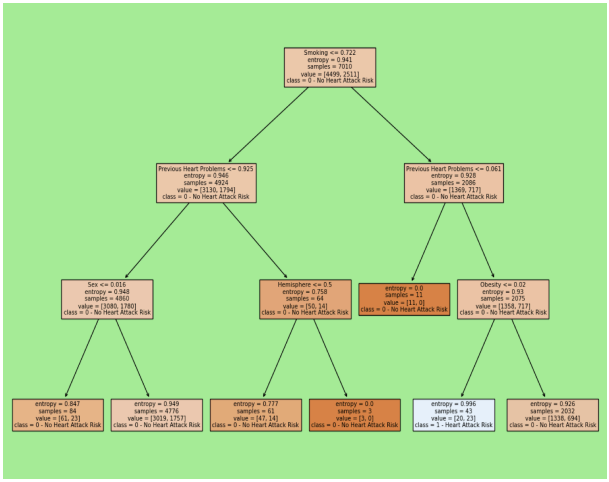
3. Method: Model Selection and Optimization

Analysis revealed a fairly balanced distribution across numerical and categorical variables, though challenges arose in correlating individual factors with heart disease due to possible non-linearity.

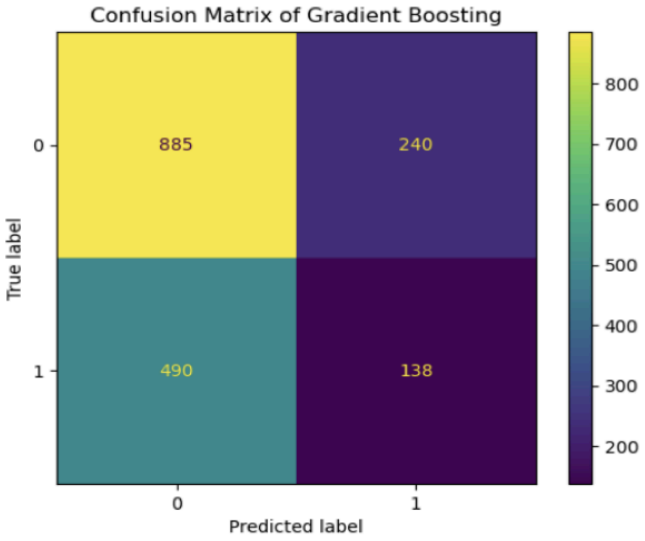
Multiple models were employed:

- Naive Bayes classification yielded a 64% accuracy but failed to recognize negative samples.
- Logistic regression mirrored the accuracy of Naive Bayes with similar limitations.
- Decision tree classifiers managed to identify some positive cases without overall accuracy improvements.





- Random Forest showed a training accuracy of 69% but did not improve test accuracy.
- Support Vector Machine (SVM) results still not well.
- Finally, after trying Gradient Boosting, we got another result. The model can recognize both positive and negative samples better and the training accuracy increased to 76%. However, the test Accuracy is only 58%



	Gradient boosting	score_GB
0	Accuracy train	0.76
1	Accuracy test	0.58
2	Precision train	0.79
3	Precision test	0.37
4	Recall train	0.46
5	Recall test	0.22
6	f1 score train	0.58
7	f1 score test	0.27
8	ROC AUC score train	0.69
9	ROC AUC score test	0.50

---

4. Summary

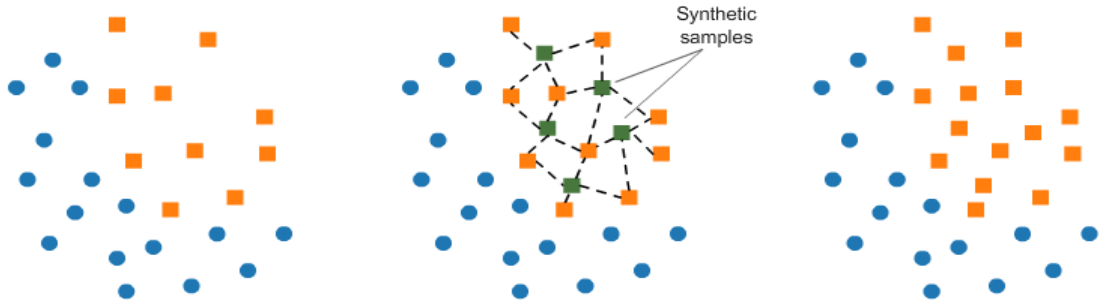
The challenge of linking specific variables to the occurrence of heart problems was evident, possibly due to the non-linear and complex nature of the data. Our efforts to understand the influencing factors did not yield a strong correlation, with most variables displaying weak links to heart problems. This underlines the complex interplay of factors involved in cardiovascular conditions and the need for sophisticated modeling techniques.

	Metric	Multinomial Naive Bayes	Decision Tree	Random Forest	Gradient Boosting
0	Accuracy train	0.64	0.64	0.69	0.76
1	Accuracy test	0.64	0.64	0.64	0.58
2	Precision train	0.00	0.53	1.00	0.79
3	Precision test	0.00	0.23	0.33	0.37
4	Recall train	0.00	0.01	0.15	0.46
5	Recall test	0.00	0.00	0.00	0.22
6	f1 score train	0.00	0.02	0.26	0.58
7	f1 score test	0.00	0.01	0.00	0.27
8	ROC AUC score train	0.50	0.50	0.57	0.69
9	ROC AUC score test	0.50	0.50	0.50	0.50

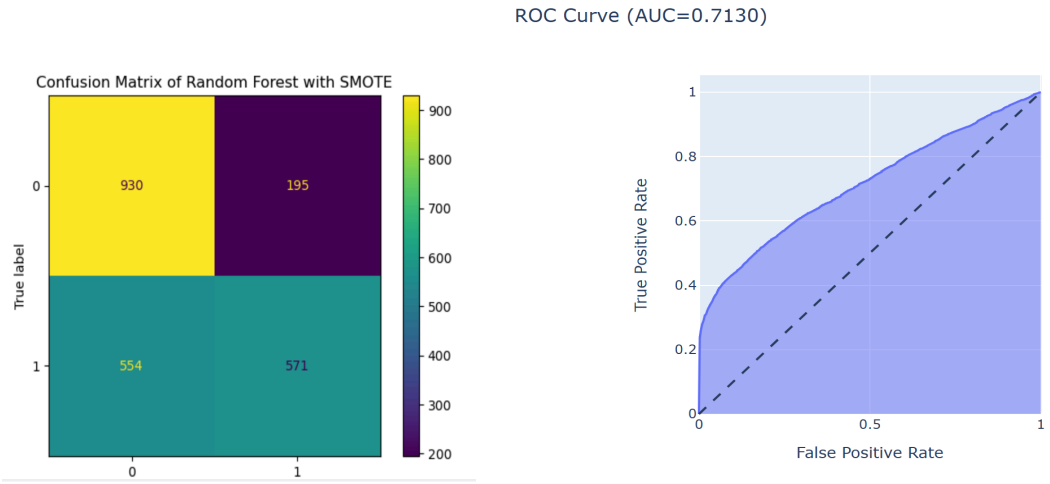
4.1 Adjustment

To address the imbalance in our dataset's dependent variable, we explored a new balanced dataset using the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE synthesizes new samples from the minority class by identifying their nearest neighbors and generating instances along the line segments connecting them. This approach not only increases the size of the minority class but also broadens the decision region for that class.

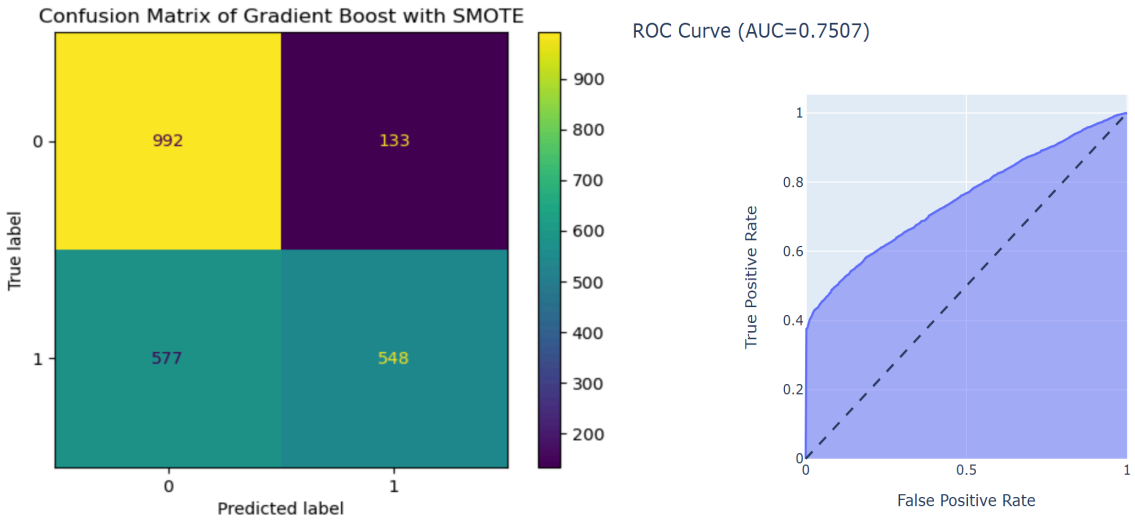
Applying machine learning models directly to the original, imbalanced dataset proved ineffective. By balancing the dataset using SMOTE and refining our models with grid search and Optuna, we enhanced model robustness. We assessed model effectiveness using the AUC metric, aiming to improve the reliability of our heart disease predictions by correcting the initial imbalance.



After adjusting the samples, the decision tree model improved. Moreover, the predictive accuracy of the random forest model exceeded 64%, reaching 66%, and the AUC value also significantly increased. This indicates that sample adjustment has a positive effect on improving prediction outcomes.



Boost class algorithms also showed better performance, maintaining accuracy for negative samples while improving positive predictions, achieving an overall accuracy of 70%. Additionally, they demonstrated a strong AUC value of approximately 0.72.



---

5. Conclusion

Regarding the adjustment, the use of various machine learning algorithms to predict heart disease yielded varying outcomes. While the Gradient Boosting model distinguished better between positive and negative events, its performance on test data indicated overfitting, a typical problem in machine learning that must be addressed in future iterations.

Previous heart problem prediction accuracy declined as a result of weaker correlations and the complexity of factors contributing to heart disease. This emphasizes the need for more robust models or ensemble methods that can better capture the underlying patterns in the data.

Subsequent research could concentrate on feature selection and engineering to better understand the influence of each element on cardiovascular disease. Furthermore, newer techniques such as deep learning or more extensive hyperparameter tuning may improve model performance.

---

6. Contribution

	Proposal	Coding	Presentation	Report
Vicky Jin	1	1	1	1
Mario Ma	1	1	1	1
Yunze Wang	1	1	1	1
Jiaqi Wu	1	1	1	1

We all participated and worked hard in this project so we rated everyone as 1.