Heart Health

UNDERSTANDING HEART HEALTH AND CORONARY HEART DISEASE RISK

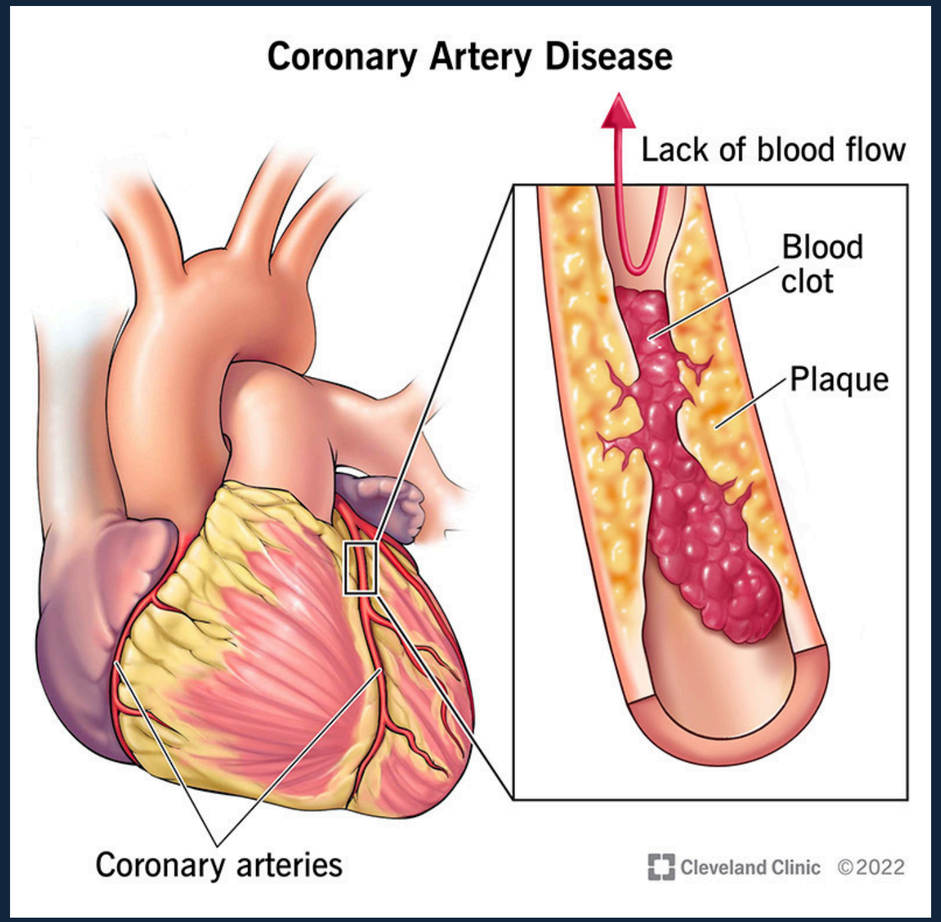
PROJECT 4 - GROUP 5

AAYUSHI, EMILY, MAGGIE, TYLI

Agenda

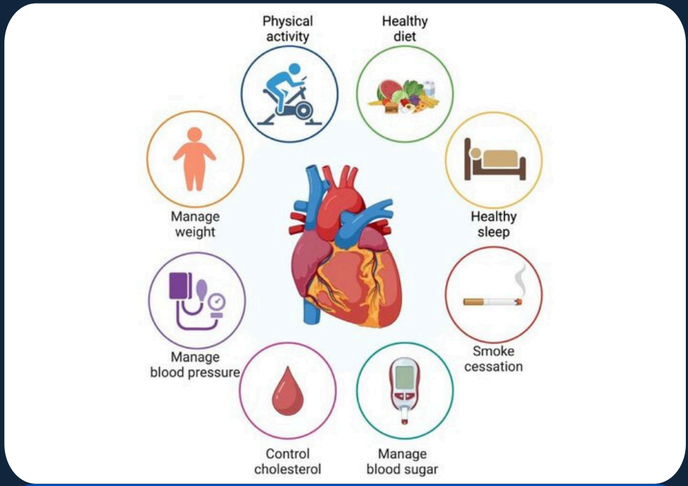
1. Topic overview
2. Data source & objective of the analysis
3. Data exploration/cleaning
4. Machine Learning Models
5. Flask & HTML - Output

Coronary Heart Disease (CHD)



- CHD (AKA Coronary Artery Disease) is a condition where the coronary arteries become narrowed or blocked due to the buildup of atherosclerotic plaques, leading to reduced blood flow to the heart muscle.
- **CHD is the leading cause of death worldwide, responsible for millions of deaths annually.**

Objective of the Analysis



To predict the risk of developing coronary heart disease (CHD) in the next 10 years based on various health and demographic factors.

To identify key factors contributing to heart disease risk.

To develop a predictive model for healthcare professionals.

Data Source

Heart Disease Dataset

Predictive Factors and Risk Assessment for Coronary Heart Disease (CHD)
Open and ethically-sourced data found on Kaggle.

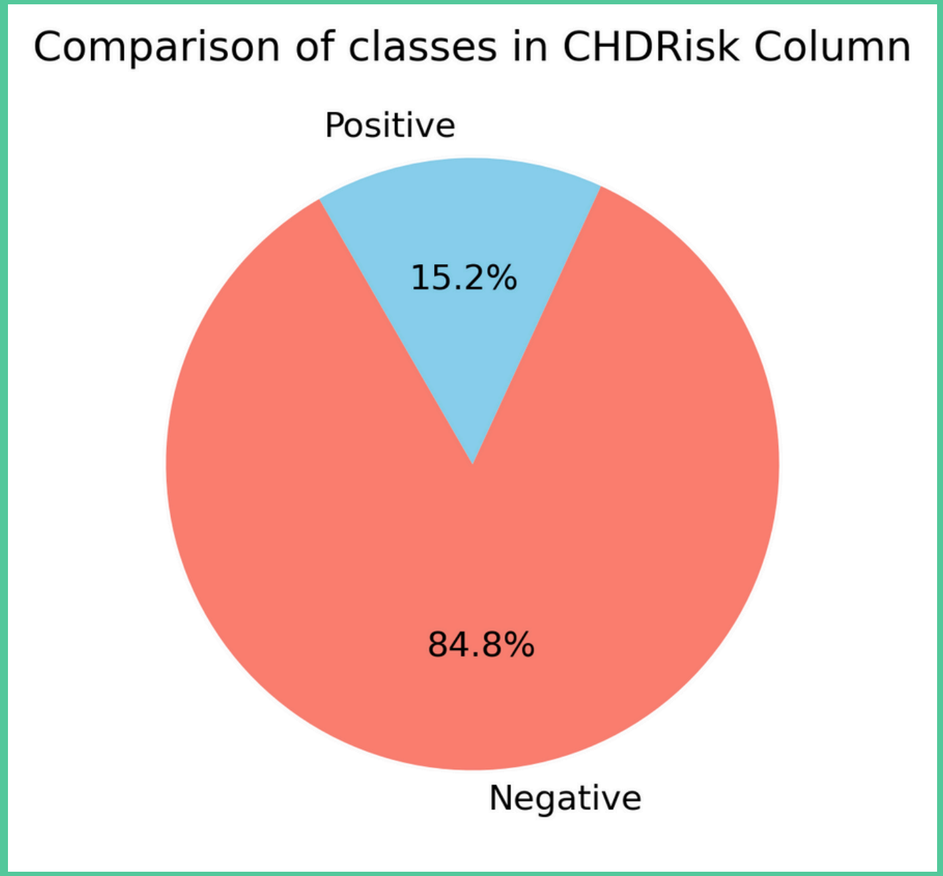
https://www.kaggle.com/datasets/mahdifaour/heart-disease-dataset/code

Data Exploration

Target Variable

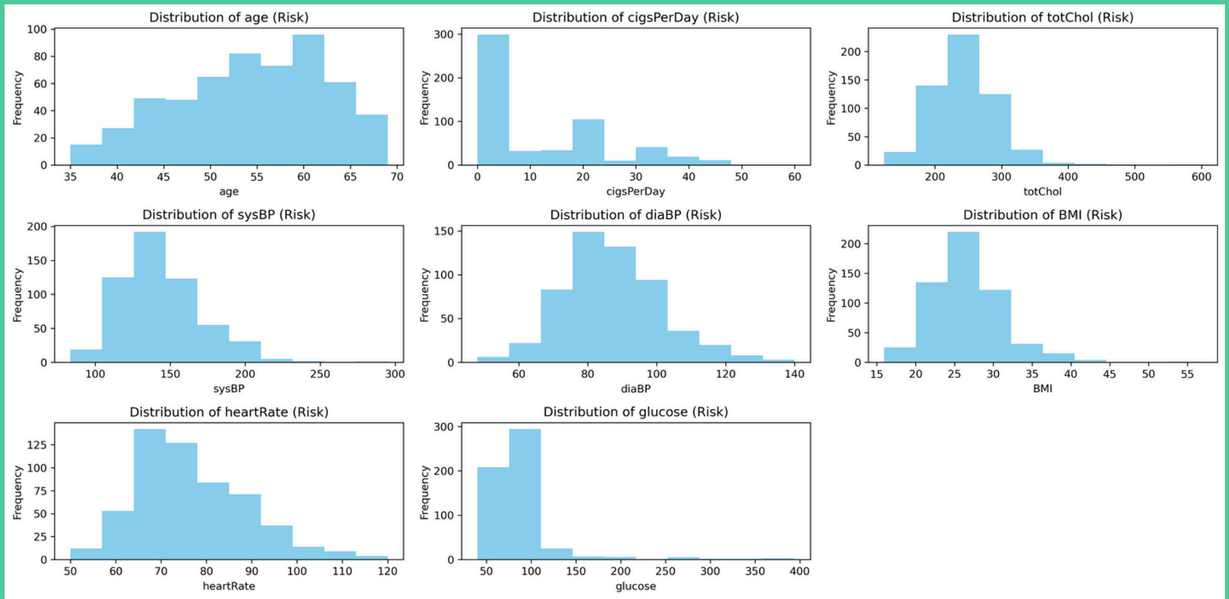
Target variable: CHD Risk

- Risk of developing CHD in the next 10 years
- Binary value (0 = not at risk; negative, 1 = at risk; positive)
- Notable class imbalance was found in our target variable

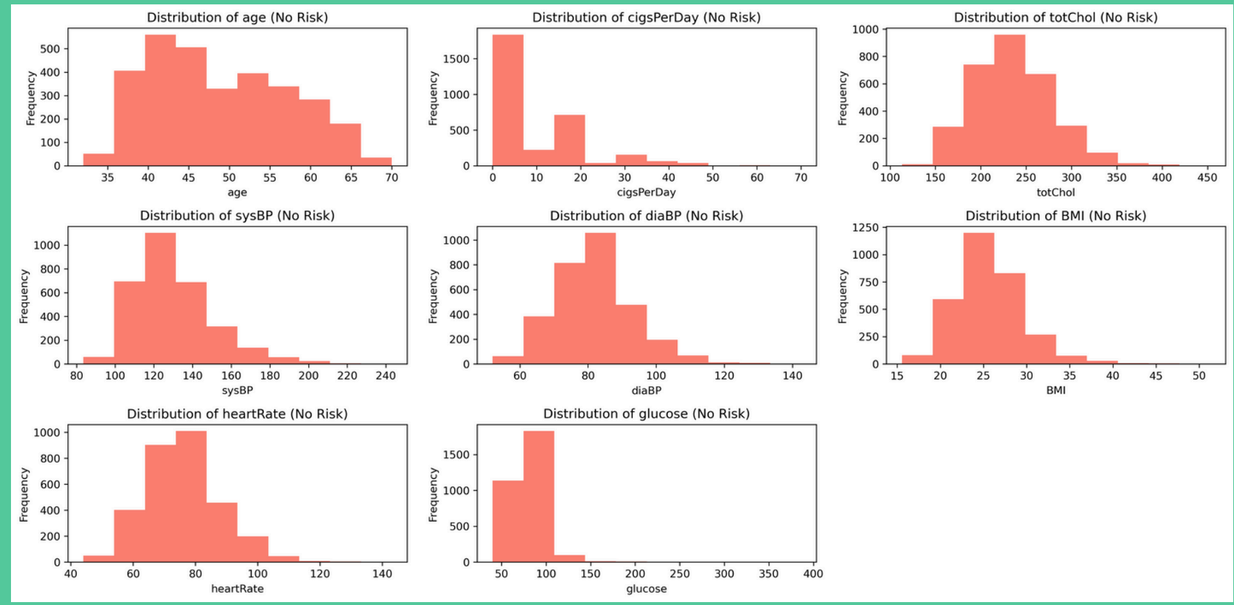


Frequency Distributions:
Continuous Variables

At Risk Distributions



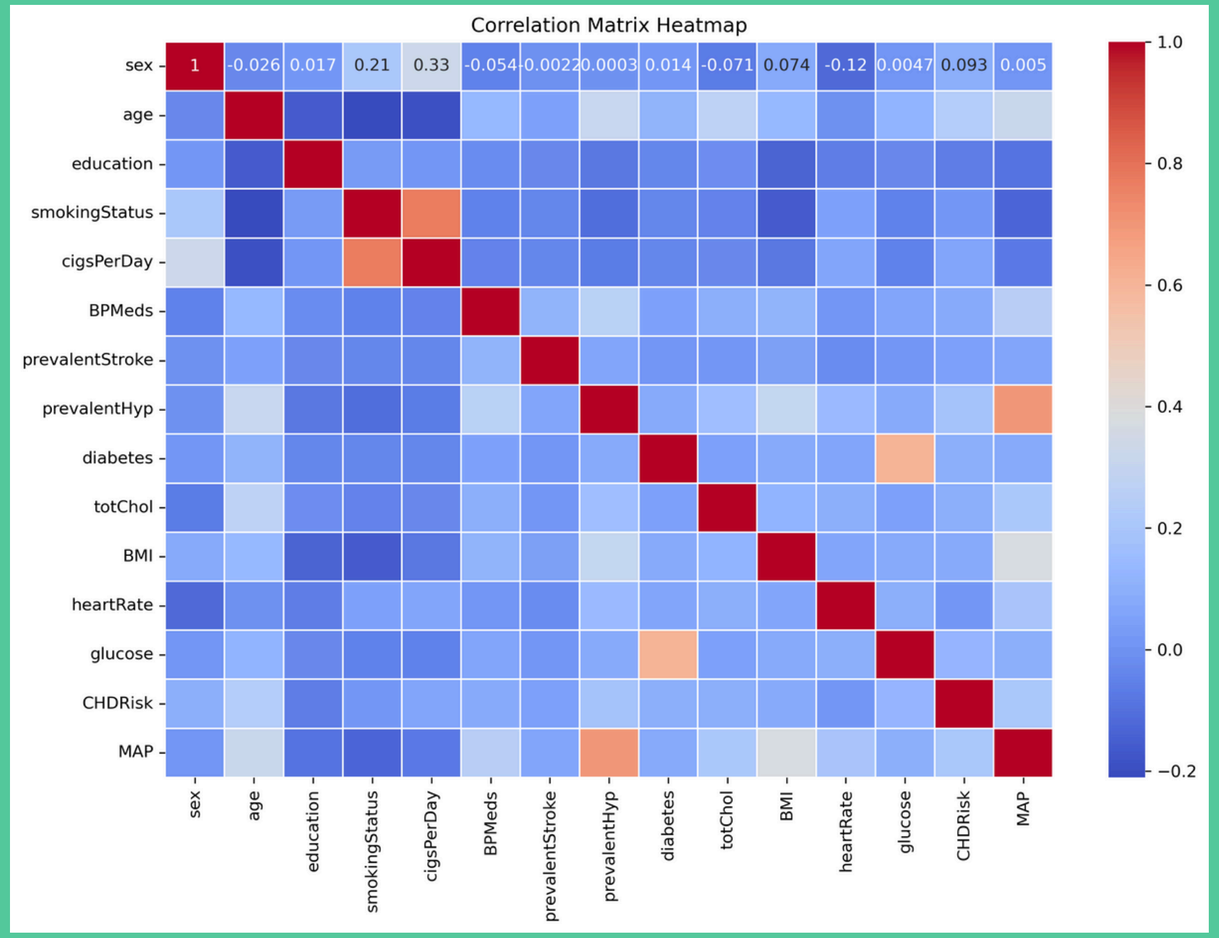
No Risk Distributions



Noticeable differences in distributions in between classes in age, diastolic BP and heart rate, highlight the potential importance of these features for for our ML models

Feature Correlations

- There are several features within our dataset that showed high correlation to other variables
- We attempt to address this problem in our ML model optimization



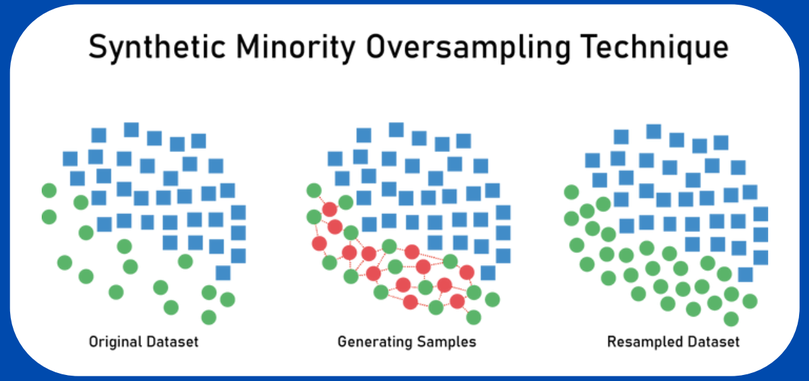
Machine Learning Models & Model Optimization

Machine Learning Models

Baseline Models

1. Logistic Regression: selected due to its simplicity, interpretability, and effectiveness in binary classification tasks.
2. Neural Network: selected for its capability to capture intricate patterns in the data and ability to handle nonlinear relationships.
3. Random Forest: chosen due to its ability to handle complex relationships and feature importance assessment.

To address the class imbalance in our target variable, **Synthetic Minority Over-Sampling Technique (SMOTE)** was used to balance the classes for model training.



Machine Learning Models

Baseline Results

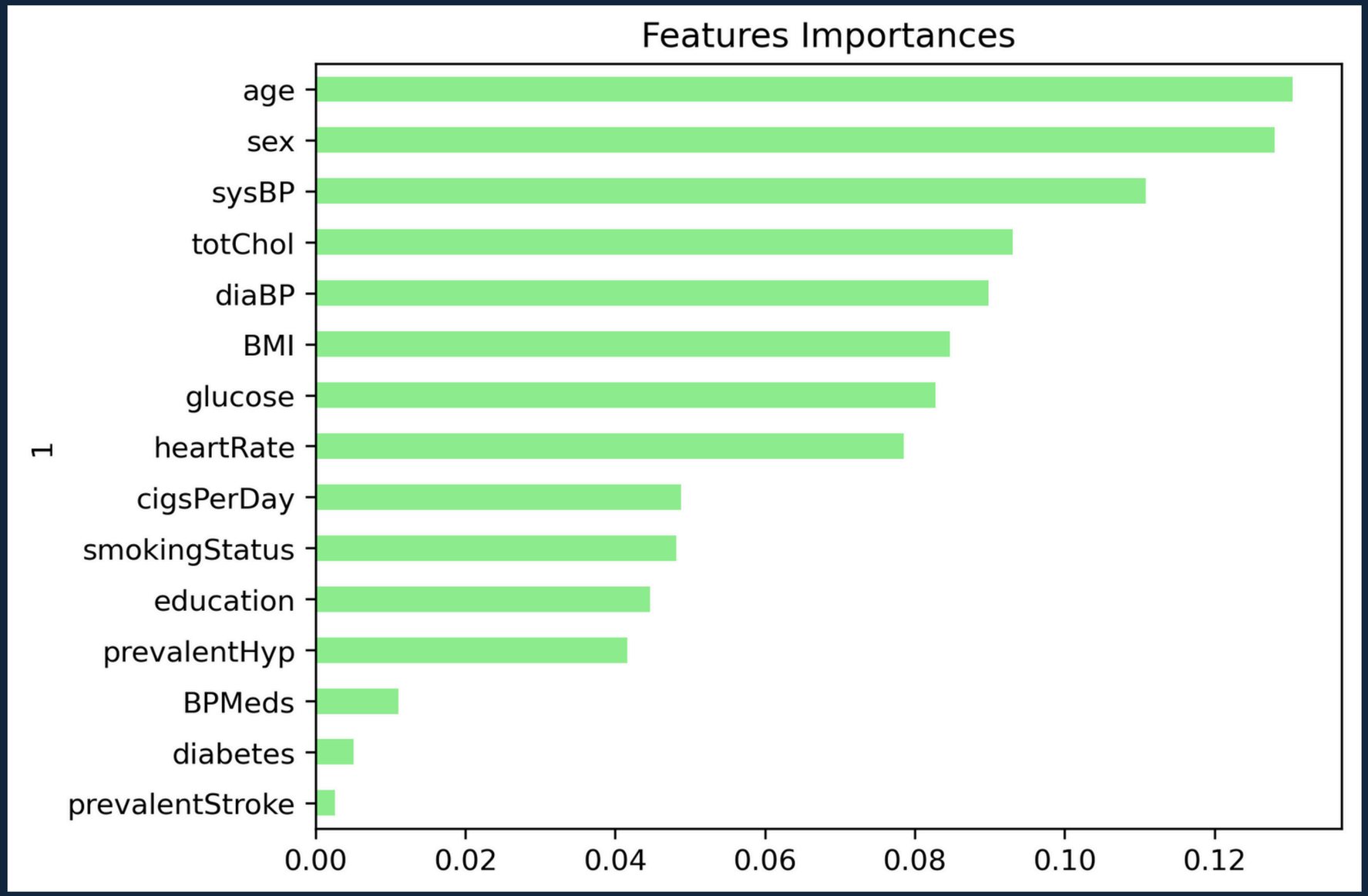
Model **accuracy** and **recall** (for the minority class) were metrics of particular interest for our baseline models, as our goal is to accurately classify those at risk of developing CHD

Comparison of Baseline Model Performances

Model	Accuracy	Recall (Class 1)
Logistic Regression	0.6626	0.65
Neural Network	0.7055	0.3551
Random Forest	0.7989	0.25

Machine Learning Models

Random Forest Model



Machine Learning Models

Model Optimization Attempts

1. Dropping features based on low Principle Component Analysis (PCA) loadings
    a. Sex, blood pressure medications, previous stroke, and diabetes were identified as having low loadings from our PCA analysis

2. Dropping binary values with high correlation to other measurements within our dataset
    a. Dropping smoking status (due to high correlation with cigarettes per day)
    b. Dropping hypertension (due to high correlation to blood pressure measurements)
    c. Dropping diabetes (due to high correlation to blood glucose)

3. Dropping features of lowest importance (from RF evaluation)
    a. Diabetes, blood pressure medication, and previous stroke

4. Mathematically combining systolic and diastolic blood pressure to produce Mean Arterial Pressure (MAP) and dropping individual BP measurements

$$MAP = DP + 1/3(SP - DP)$$

Machine Learning Models

Model Optimization Results

Neural Network Optimization Results

Optimization	Accuracy	Loss	Recall (Class 1)
Optimization 1	0.7033	1.4461	0.2681
Optimization 2	0.7319	1.6564	0.2464
Optimization 3	0.7319	1.5531	0.2971
Optimization 4	0.7275	1.5690	0.2464
Baseline	0.7055	1.2111	0.3551

Random Forest Optimization Results

Model	Accuracy	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1 Score (Class 0)	F1 Score (Class 1)
Optimization 1	78.9%	87%	29%	88%	27%	88%	28%
Optimization 2	80.22%	87%	31%	90%	27%	89%	27%
Optimization 3 (Best Model)	81.98%	87%	37%	92%	26%	89%	27%
Optimization 4	80.99%	87%	33%	91%	24%	89%	28%
Baseline Model	79.8%	87%	30%	90%	25%	88%	27%

Best Model

Analysis & Conclusion

Best Performing Model

RF Model Optimization 3 (dropping features of low importance) exhibited the best performance

Metric	Accuracy	Precision (Class 0)	Recall (Class 0)	F1-score (Class 0)	Precision (Class 1)	Recall (Class 1)	F1-score (Class 1)
Value	0.8198	0.87	0.92	0.90	0.37	0.26	0.31

Why is this the best model?

- **Accuracy:** Exhibits the **highest accuracy** among all attempts, indicating superior overall predictive performance.
- **Precision:** Showcases the **highest precision** for class 1 (at risk), implying fewer false positives.
- **Not Overfit:** Accuracy **scores across k-fold validation** (k=5) were consistent, demonstrating model is not overfit.
- **Overall Performance:** Considering accuracy, precision, and requiring the **lowest number of features** to train, optimization 3 emerges as the most well-rounded and effective model for CHD classification prediction.

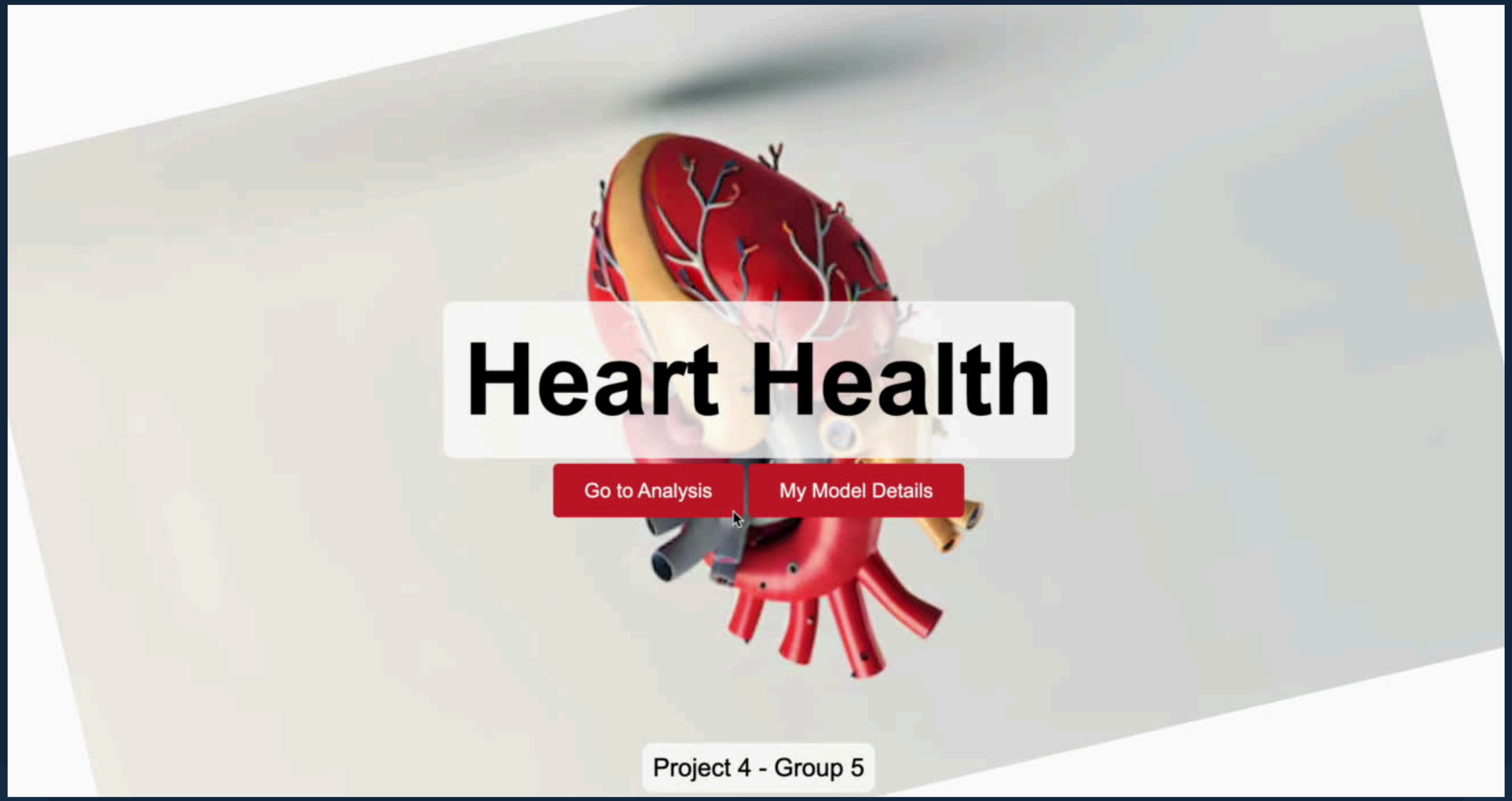
Limitations

- While we were able to produce a model with high accuracy, our **recall scores for the minority class (those at risk of developing CHD) were consistently low**
- Our best model was only able to correctly predict those at risk of developing CHD 26% of the time, leading to a **high false negative rate**
    - This has **significant implications in health data**, as misidentifying those at risk of developing a disease is a serious concern
- Although attempts were made to improve recall scores for the minority class (such as lowering the classification threshold), these attempts negatively affected overall model accuracy
- In order to address this issue, more data must be collected, specifically for those in the at-risk population

Recommendations:

1. Adopt Random Forest Optimization 3 for its highest accuracy (81.98%) and improved minority class precision (37%).
2. Continue to explore and evaluate performance of other ensemble learning models
3. Continue excluding features with low importance to streamline models. Continue trying new optimizations until the outcome is met.
4. Adjust decision thresholds to enhance recall for the minority class.
5. Regularly monitor performance with fresh data, using a combination of metrics for a comprehensive view.
6. Explore additional hyperparameter tuning and advanced optimization techniques for further improvement. Explore other binning techniques for columns like age, BMI, Total Cholesterol etc.

Output


Heart Health
Go to Analysis
My Model Details
Project 4 - Group 5

Consult us for Questions

Thank you..!!