CP322 Group 9 Machine Learning Presentation:

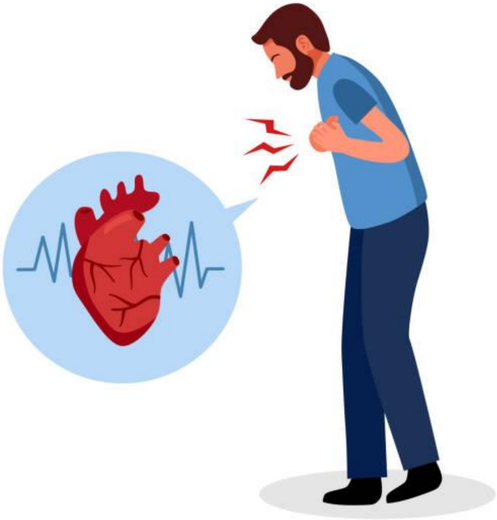
Heart Disease Risk Prediction

By: Simran Badwal, Charnel Dolon, Marc Niven Kumar, Brandon Pham, Erin Israt Urbi

Introduction

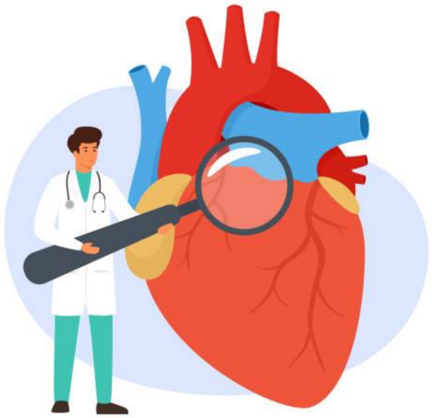
- **Overview/Context**
  - o The objective of this project is to successfully predict heart disease risk using machine learning methods based on a set of health indicators
  - o Globally, the leading cause of mortality is heart disease. The ability to successfully predict heart disease risk can save many lives

- **What questions are being addressed?**
  - o What are the most important indicators of heart disease?
  - o Which machine learning model performs best?
  - o How accurately can we predict the likelihood of heart disease using machine learning models?



Introduction - Related Works

- Previous works have shown the use of machine learning algorithms to predict the risk of heart disease:
  - Naïve Bayes
  - Support Vector Machine
  - Decision Tree
  - Artificial Neural Networks



Some Examples of Related Works:

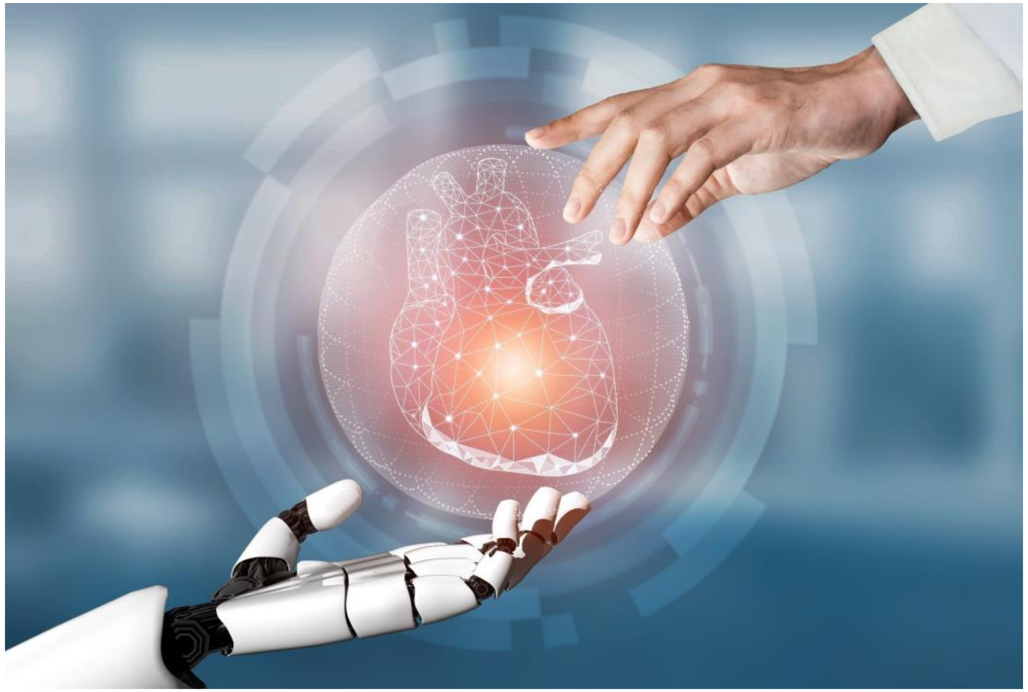
- Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M. W., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Computers in biology and medicine, 136, 104672. https://doi.org/10.1016/j.compbiomed.2021.104672
- Ngufor, C., Hossain, A., Ali, S. & Alqudah, A. Machine learning algorithms for heart disease prediction: a survey. Int. J. Comput. Sci. Inform. Secur. 14 (2), 7–29 (2016).
- Yang, M., Wang, X., Li, F. & Wu, J. A machine learning approach to identify risk factors for coronary heart disease: a big data analysis. Comput. Methods Programs Biomed. 127, 262–270 (2016).

Solution/Methods

The project shows the implementation of multiple machine learning models to predict the likelihood of heart disease.

The adopted methods used include:

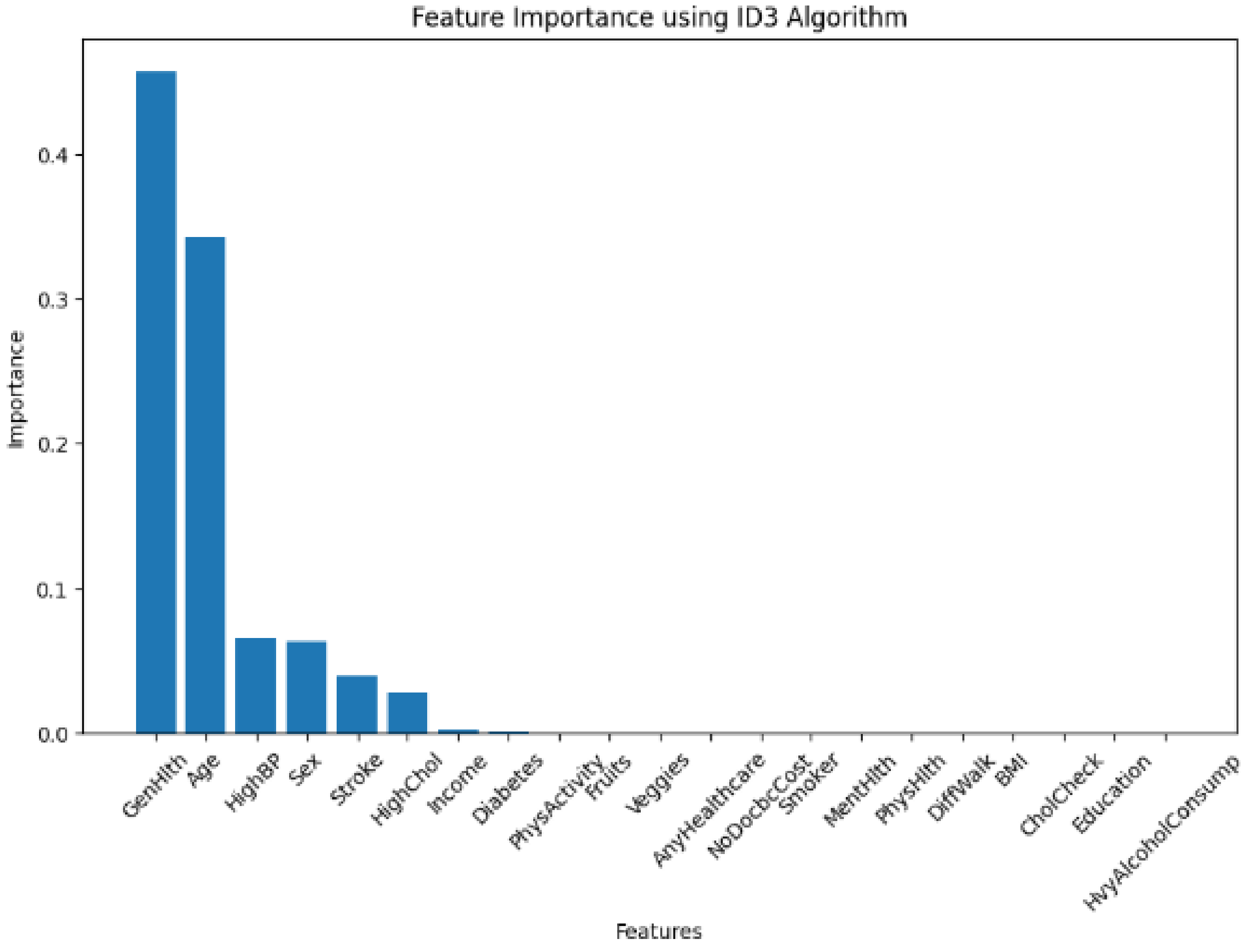
- Decision Tree
- Naive Bayes
- K- Nearest Neighbours (KNN)
- Logistic Regression



Data and Experiments

- **Dataset**
  - Contains health indicators of heart disease
  - Features include attributes like blood pressure, cholesterol levels, BMI, physical activity, smoking status, and general health rating
  - Target Feature: HeartDiseaseorAttack
- **Preprocessing**
  - Missing Feature Check
  - Split dataset into training and testing sets of 70% and 30% respectively
  - Down-sampling majority class

Feature Engineering: ID3 Algorithm



Feature Importances:

	Feature	Importance
13	GenHlth	0.457190
18	Age	0.342746
0	HighBP	0.065804
17	Sex	0.063346
5	Stroke	0.039752
1	HighChol	0.028378
20	Income	0.002326
6	Diabetes	0.000457
7	PhysActivity	0.000000
8	Fruits	0.000000
9	Veggies	0.000000
11	AnyHealthcare	0.000000
12	NoDocbcCost	0.000000
4	Smoker	0.000000
14	MentHlth	0.000000
15	PhysHlth	0.000000
16	DiffWalk	0.000000
3	BMI	0.000000
2	CholCheck	0.000000
19	Education	0.000000
10	HvyAlcoholConsump	0.000000

Decision Tree

- Easy to implement
- Works in both classification and regression tasks

Evaluation

Accuracy: 0.76

Precision: 0.78

Recall: 0.79

F1-Score: 0.76

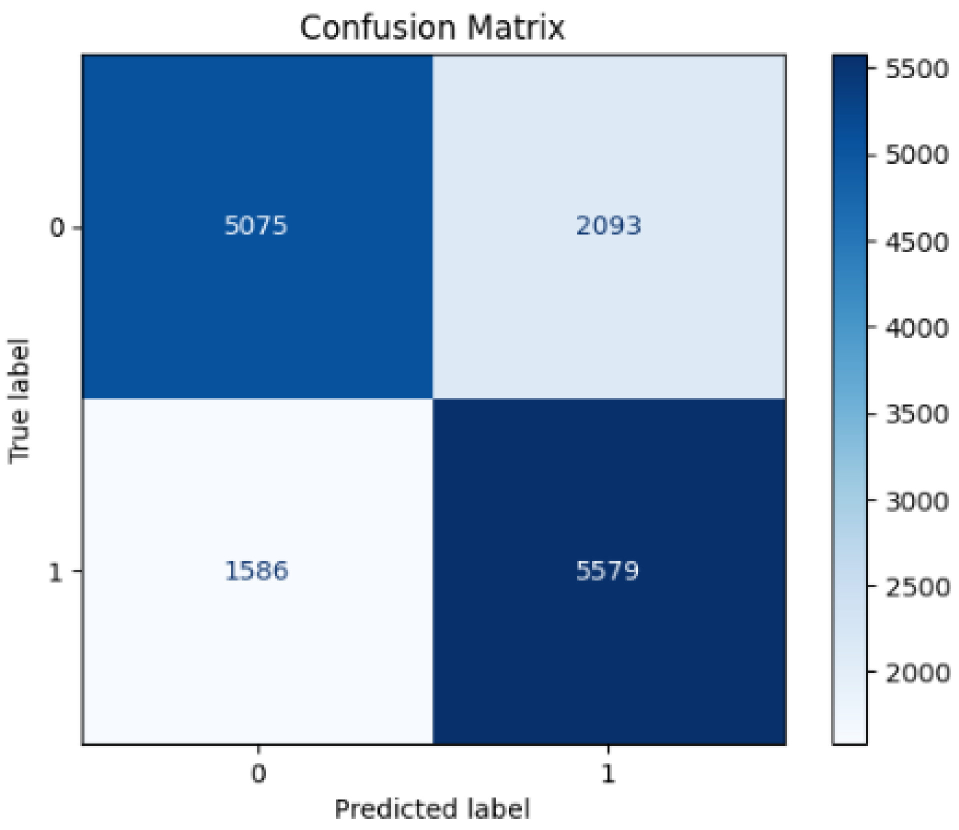
ROC-AUC: 0.82

Observations

- ROC Curve of 82%
- High True Positive/ Negative rates

Potential Causes

- Sensitive to unbalanced datasets
- Larger dataset may improve evaluation accuracies.



Naïve Bayes

- Simple and efficient for classification problems
- Well-suited for datasets with categorical variables

Evaluation

Accuracy: 0.72

Precision: 0.74

Recall: 0.68

F1-Score: 0.71

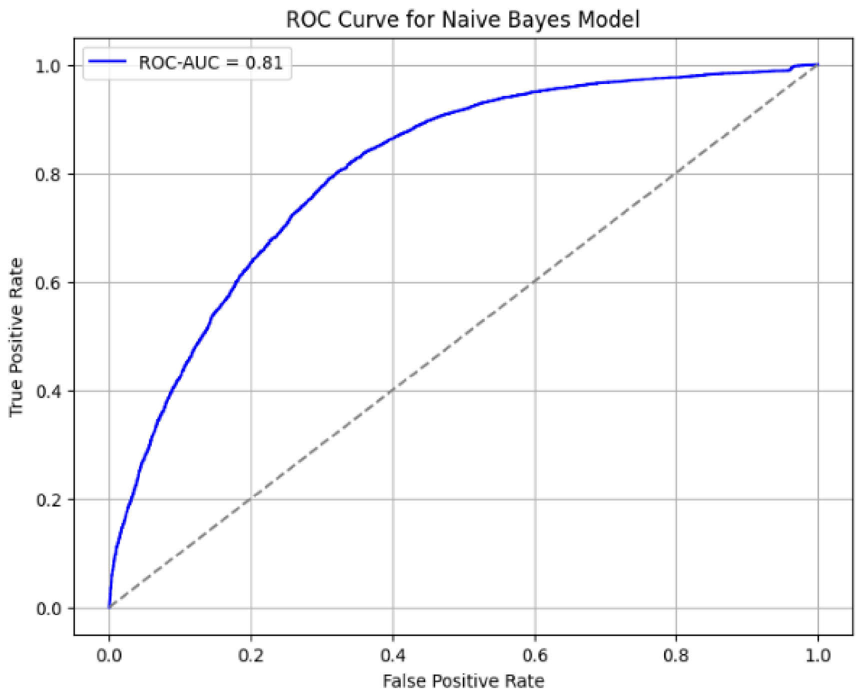
ROC-AUC: 0.81

Observations

- 68% of the actual heart disease cases were correctly identified

Potential Causes

- Dependent features
    - PhysActivity and BMI
    - Smoker and HighBP
- Threshold value



Regression

- Fast to train and works well even with relatively large datasets.
- Prevents overfitting and helps the model generalize better.

Evaluation

Accuracy: 0.74

Precision: 0.72

Recall: 0.77

F1-Score: 0.75

ROC-AUC: 0.81

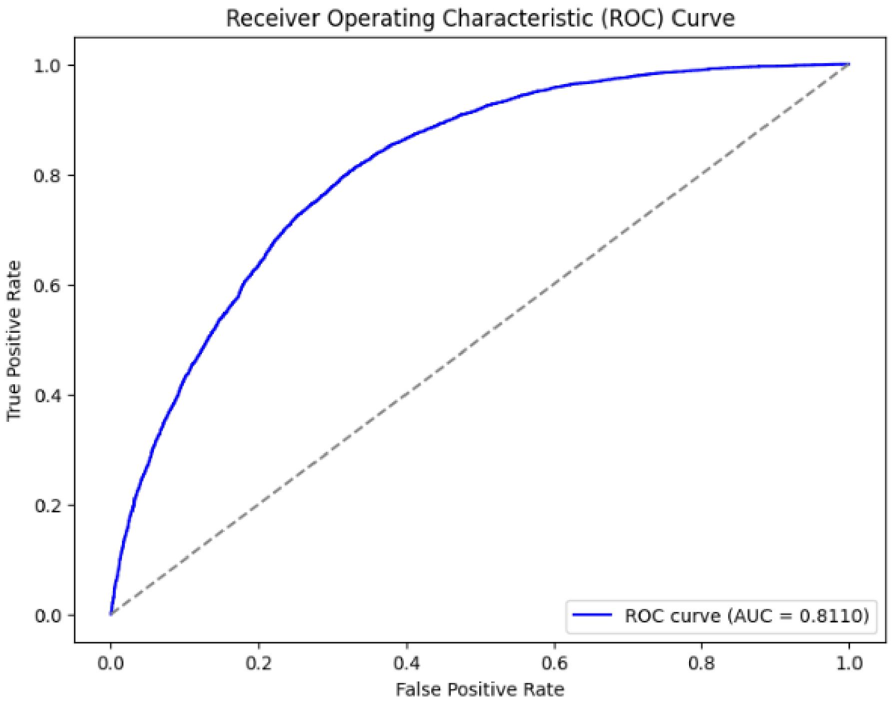
Observation:

•**Low False Positive Rate: Not too many** non-heart disease cases are misclassified as heart disease.

•**High Recall (77%): 77**% of actual heart disease cases are correctly identified.

Potential Causes

- **Class Imbalance**: If one class dominates, the model may struggle to balance precision and recall.



K-Nearest Neighbor (knn)

Evaluation

Accuracy: 0.7433

Precision: 0.7272

Recall: 0.7786

F1-Score: 0.7520

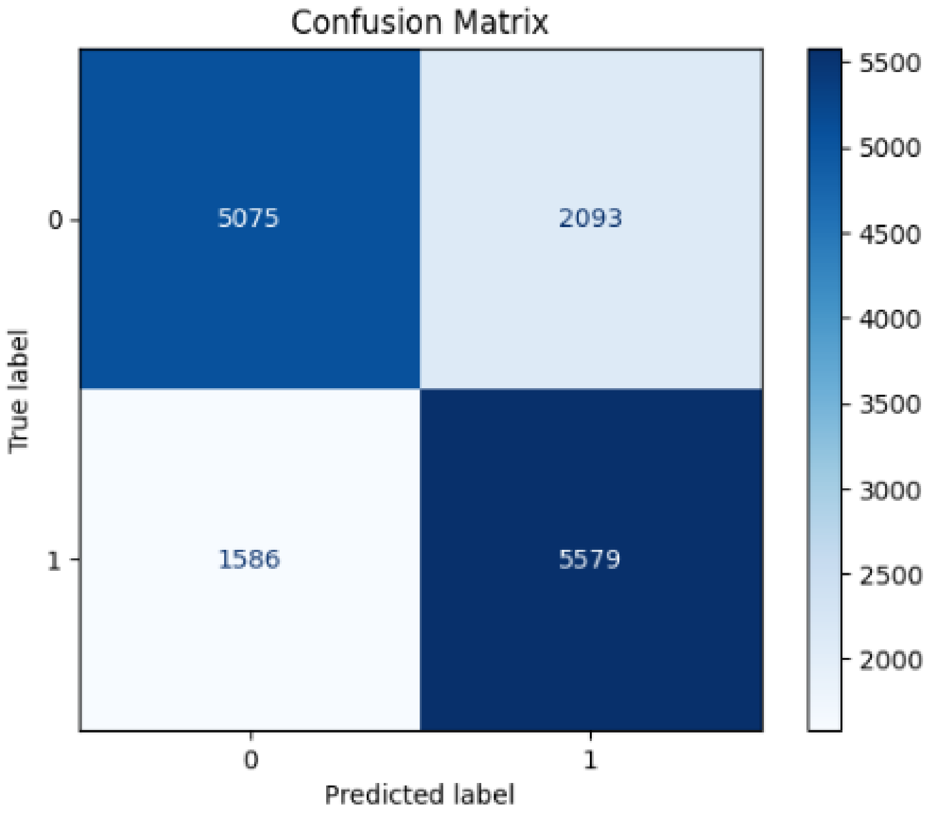
ROC-AUC: 0.6101

Observation:

- Average Accuracy as compared to other models
- A very low ROC Curve Score of 61%

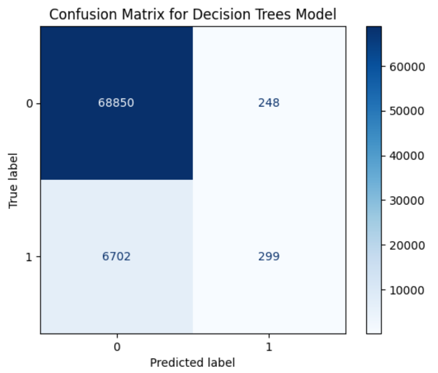
Potential Issues:

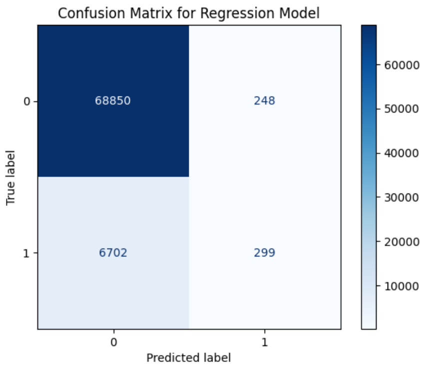
- Probability Estimation Issues
- Sensitivity to Scaling and Parameters

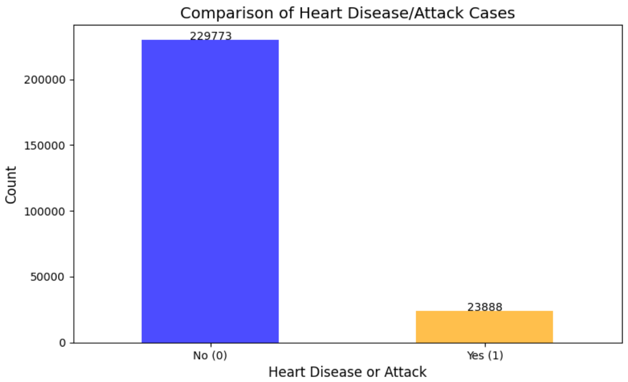


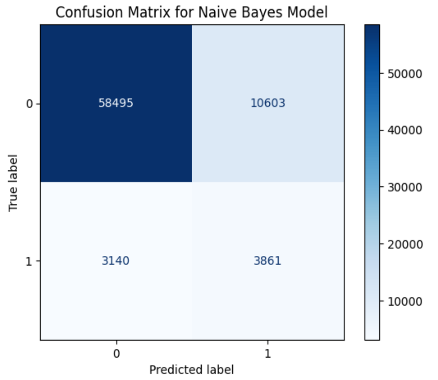
Original vs Balancing the Dataset

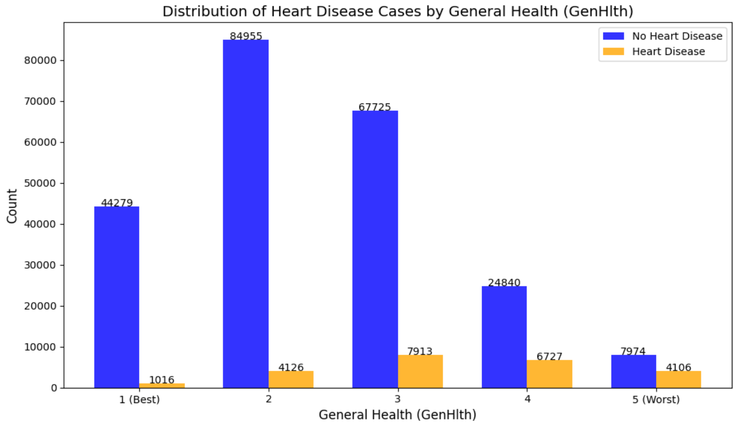
- Original Data had issues with accuracy for detecting a Heart Attack due the inbalance of the set, with the skew of non-Heart Diseased individuals to Heart Diseased individuals being a ratio of 11:1
- Downsampling is a common data processing technique that addresses imbalances in a dataset by removing data from the majority class such that it matches the size of the minority class. (https://www.ibm.com/topics/downsampling)
- Downsampling increased the model's accuracy in terms of detecting Heart Disease by 50% on average, while sacrificing the accuracy for detecting non-Heart Disease individuals by 20% on average.

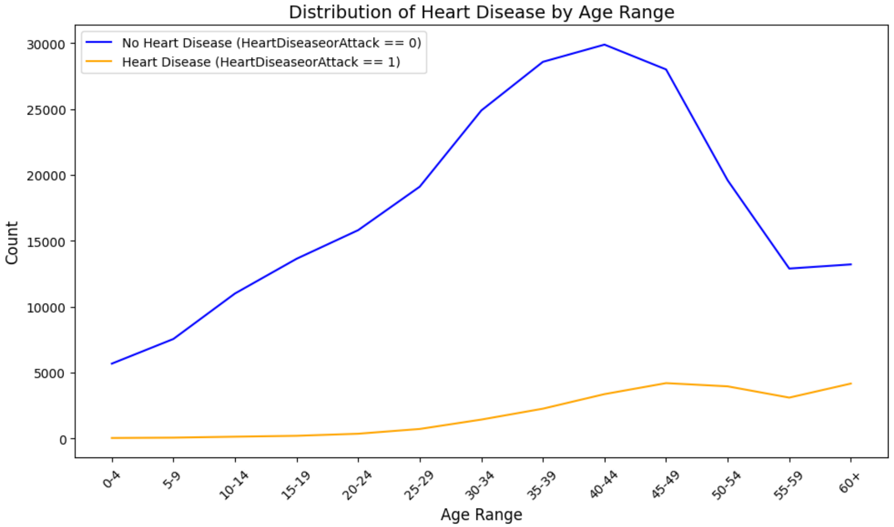


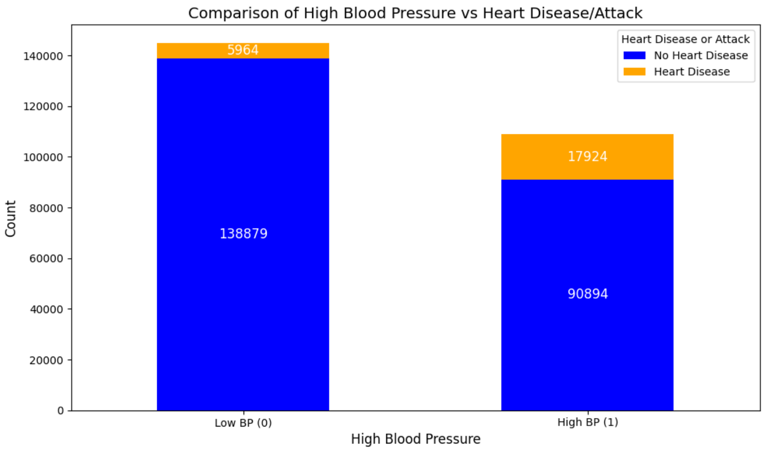






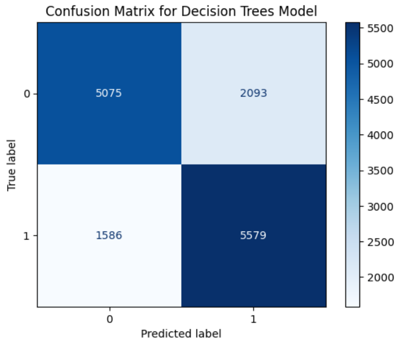


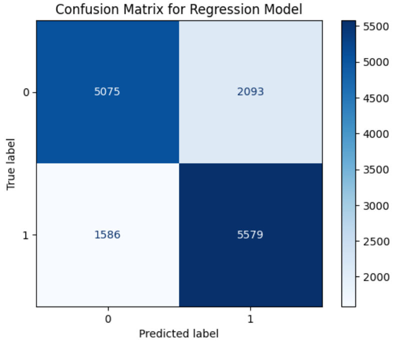


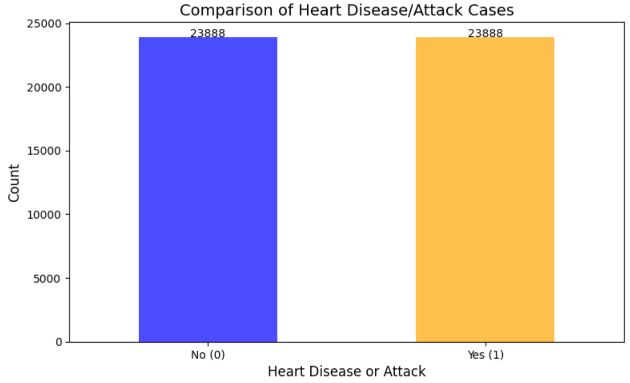


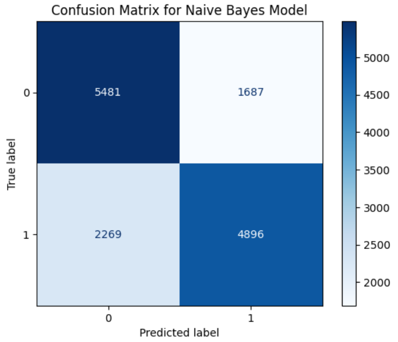
Original vs Balancing the Dataset

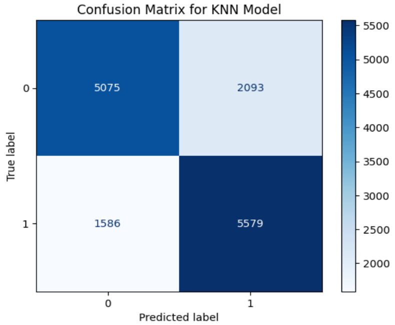
- Original Data had issues with accuracy for detecting a Heart Attack due the inbalance of the set, with the skew of non-Heart Diseased individuals to Heart Diseased individuals being a ratio of 11:1
- Downsampling is a common data processing technique that addresses imbalances in a dataset by removing data from the majority class such that it matches the size of the minority class. (https://www.ibm.com/topics/downsampling)
- Downsampling increased the model's accuracy in terms of detecting Heart Disease by 50% on average, while sacrificing the accuracy for detecting non-Heart Disease individuals by 20% on average.

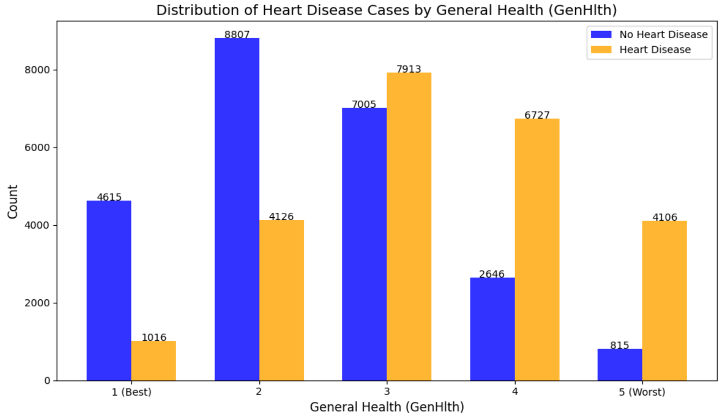


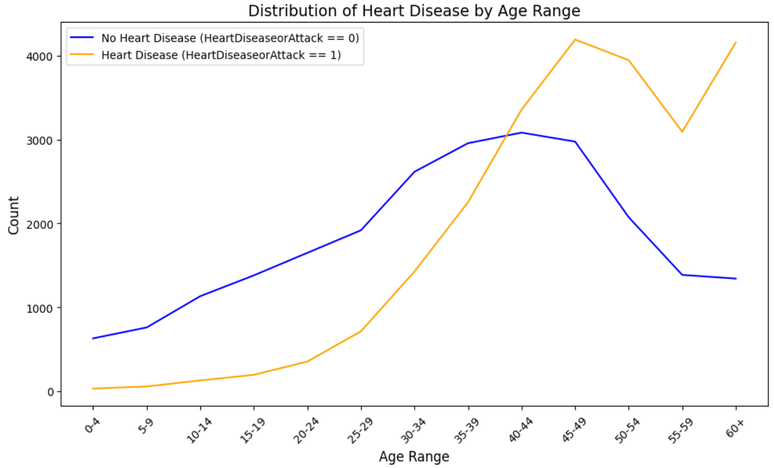


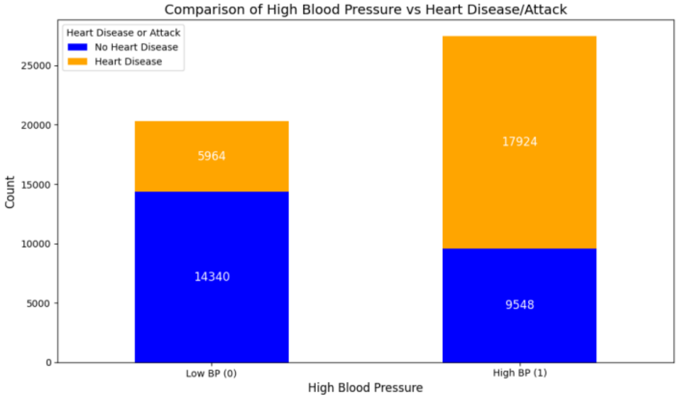




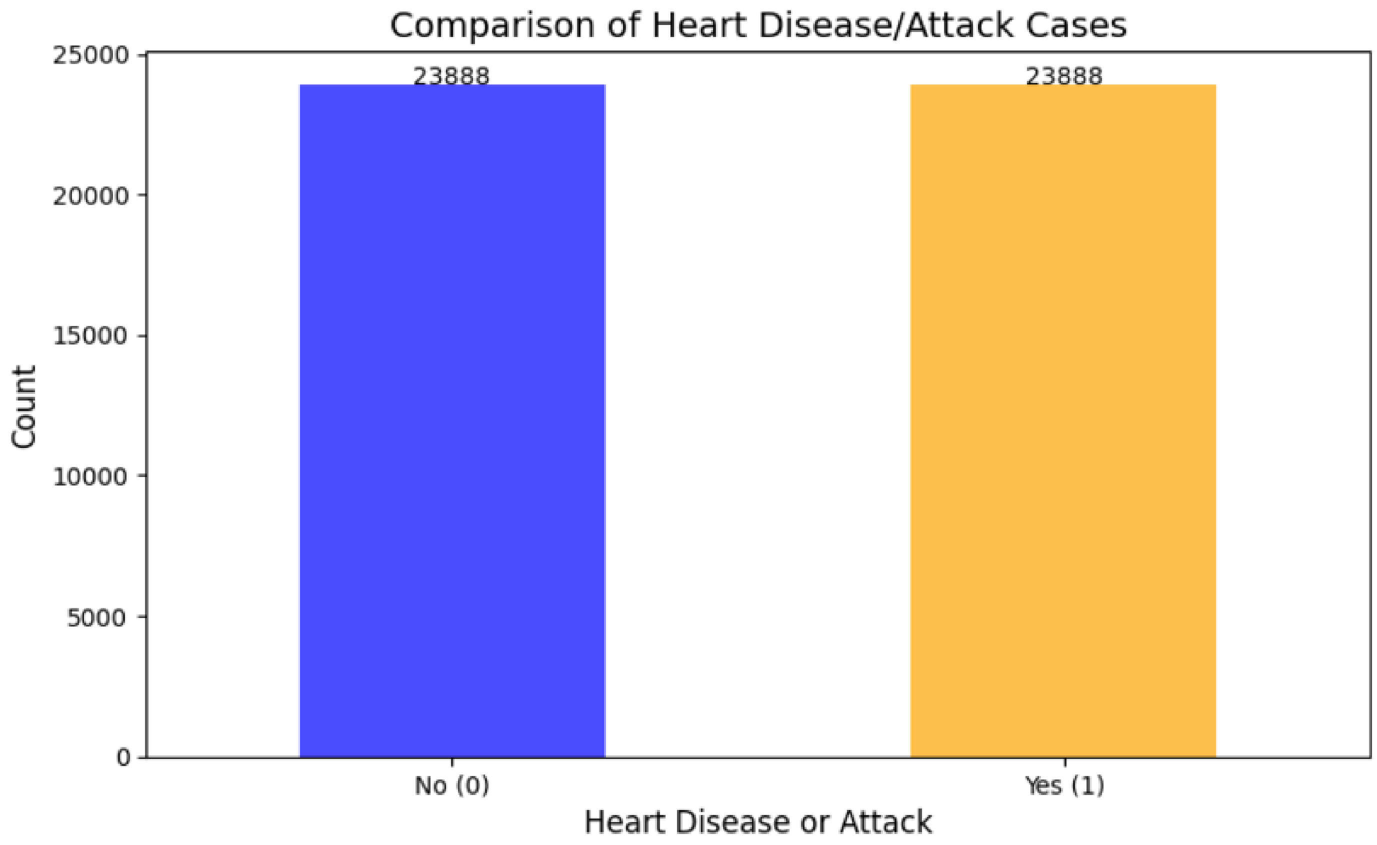




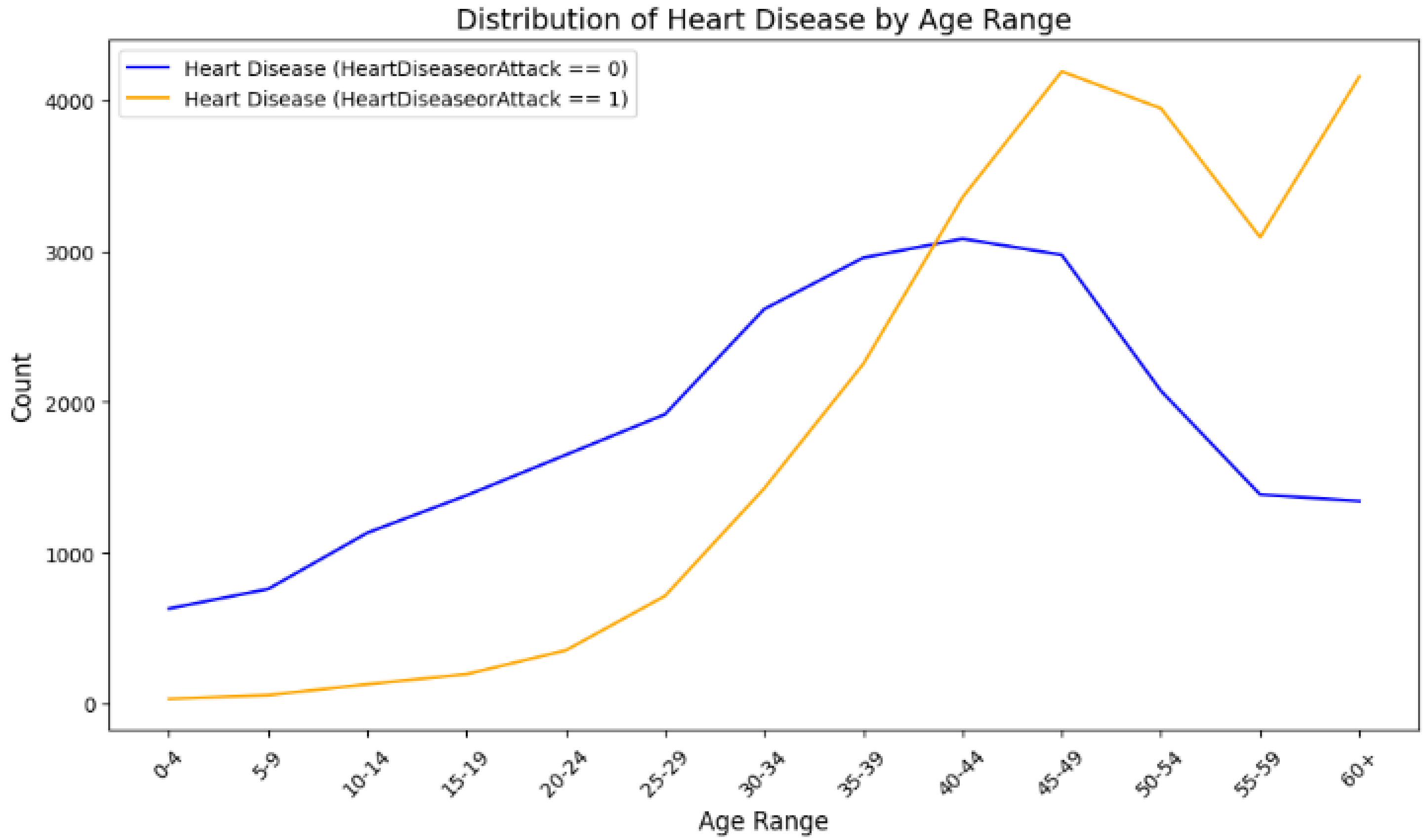




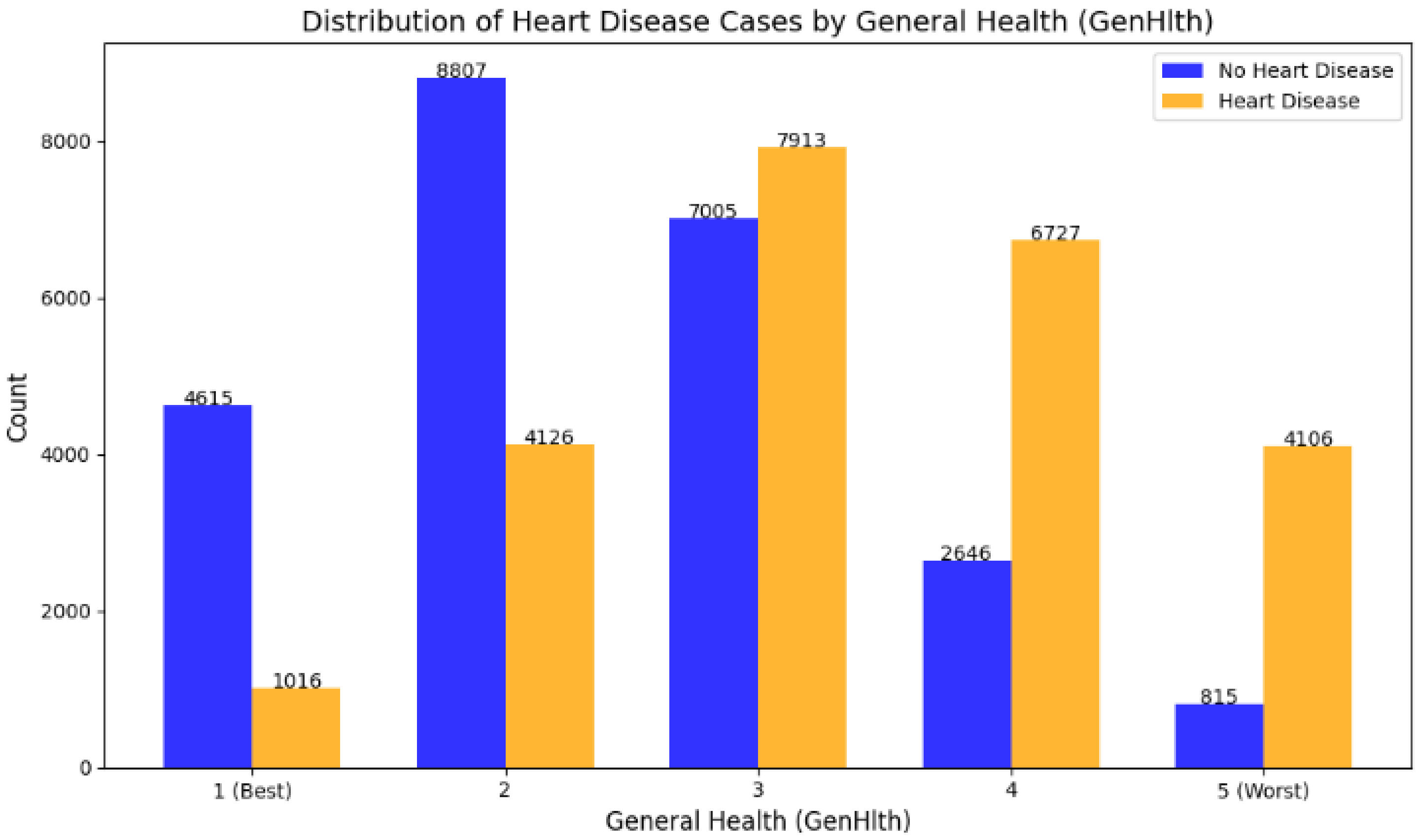
Comparison of Heart Disease/Attack Cases



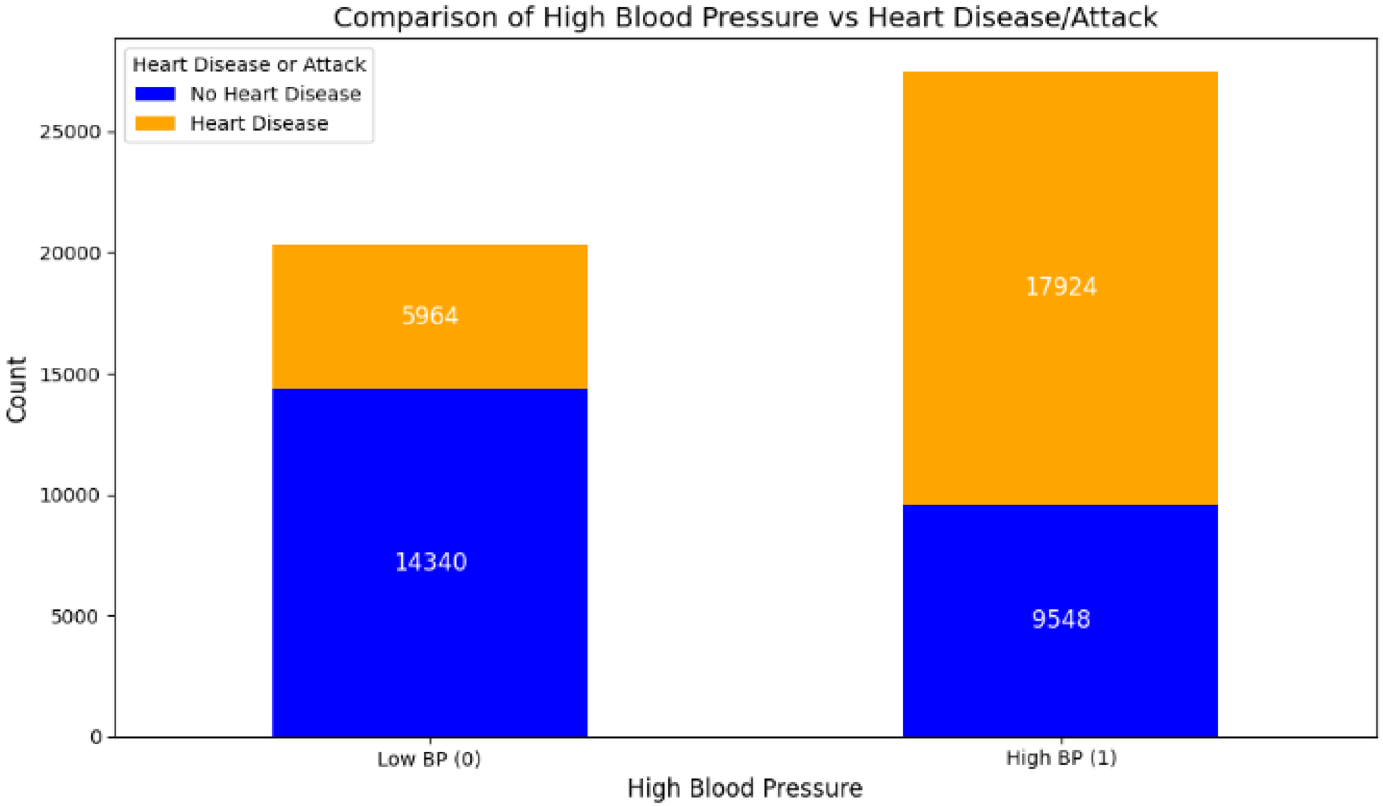
Distribution of Heart Disease by Age Range

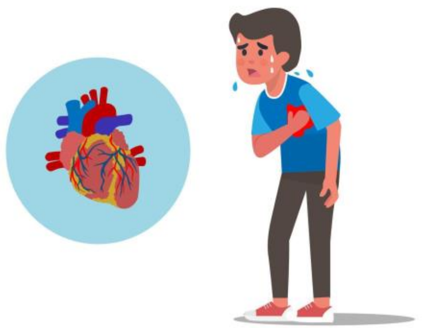


Distribution of Heart Disease Cases by General Health



Comparison of High Blood Pressure vs Heart Disease/Attack





Conclusion

- **What are the most important indicators of heart disease?**
  - High Blood Pressure, GenHlth, Age
- **Which machine learning model performs best?**
  - Decision Tree
    - Highest Accuracy
    - Highest Recall
- **How accurately can we predict the likelihood of heart disease using machine learning models?**
  - 76% accuracy
  - Tool for early detection
  - Future Improvements