Heart Disease Prediction

Group members: Wenhe Chen, Jiaqi Guo, Sybil Kong, Zeya Ling, Yutong Wei

1. Introduction

Heart disease is a major health concern affecting millions of people globally. In this report, we will analyze a dataset of heart disease patients using K-NN and linear regression algorithms, with the goal of identifying key predictors of heart disease. We find that sex is the most crucial factor and females are more likely to develop heart disease than males.

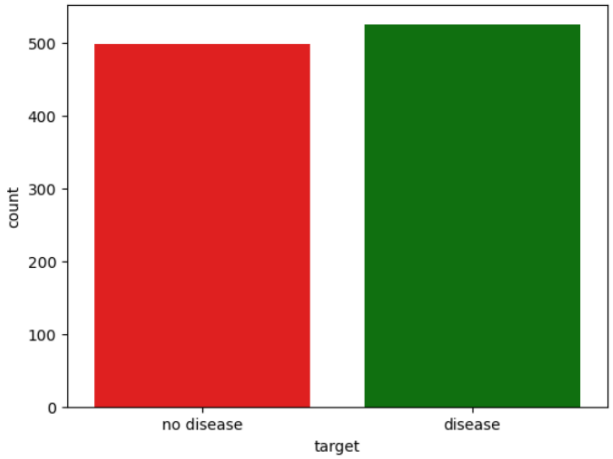
2. Data Exploration

2.1 Dataset and Cleaning

Our dataset (https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset) consists of 14 variables and 1026 rows. We preprocessed and cleaned a dataset by checking for missing values, assessing target variable balance, performing feature scaling and normalization, and identifying and removing highly correlated variables to acquire a high-quality dataset.

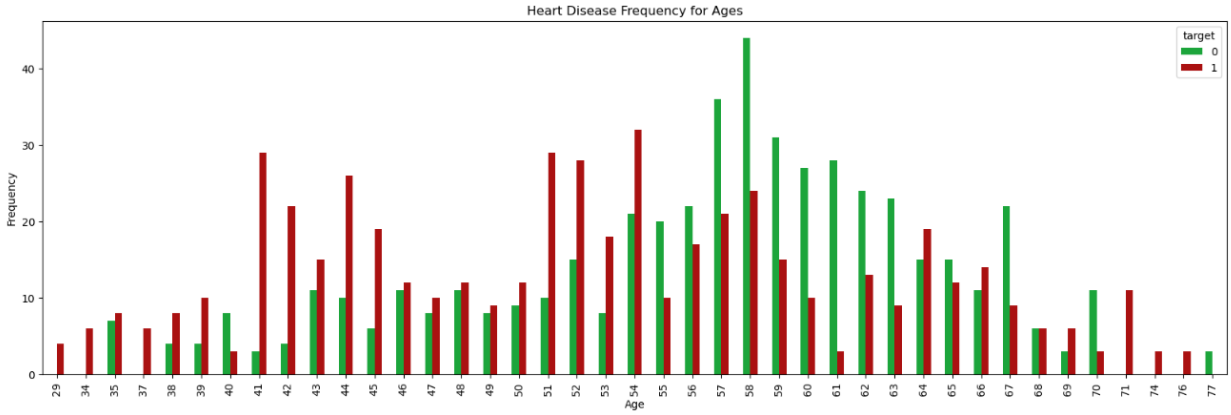
2.2 Exploratory Analysis

For data exploration, we calculated the percentage of 0 and 1 in response (0=no heart disease, 1=having heart disease) and drew the corresponding plot (Figure 1.). We obtained that the ratio of patients with heart disease to those without heart disease is 0.5132: 0.4868. Hence, our dataset is a balanced dataset.



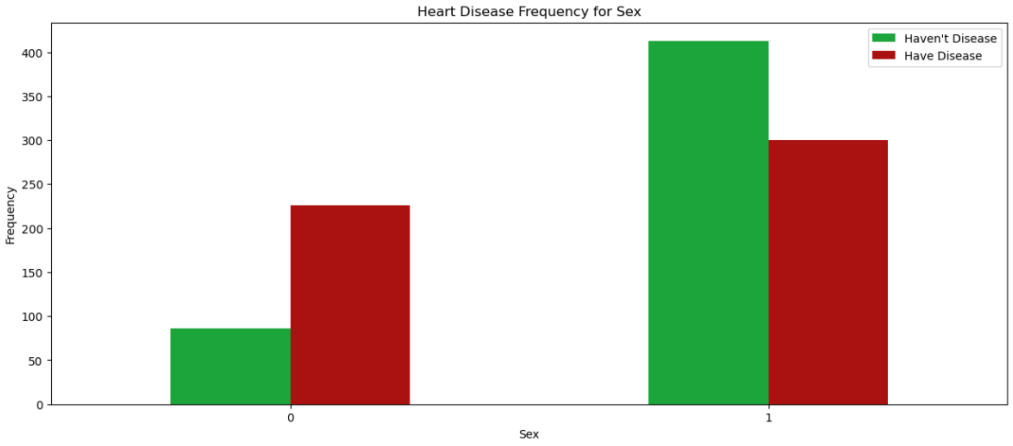
(Figure 1. Plot for the frequency of 0 (red) and 1 (green) in y)

We also explored the distribution of the relationship between the occurrence of heart disease and age. From the following plot (Figure 2.) we find that in different age groups, the ratio of 0 to 1 changes.

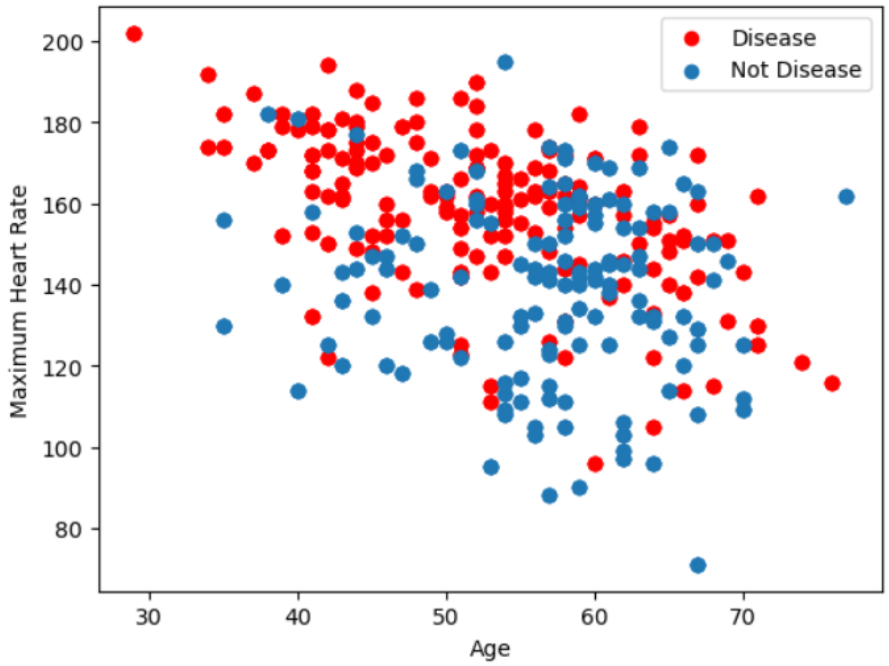


(Figure 2.)

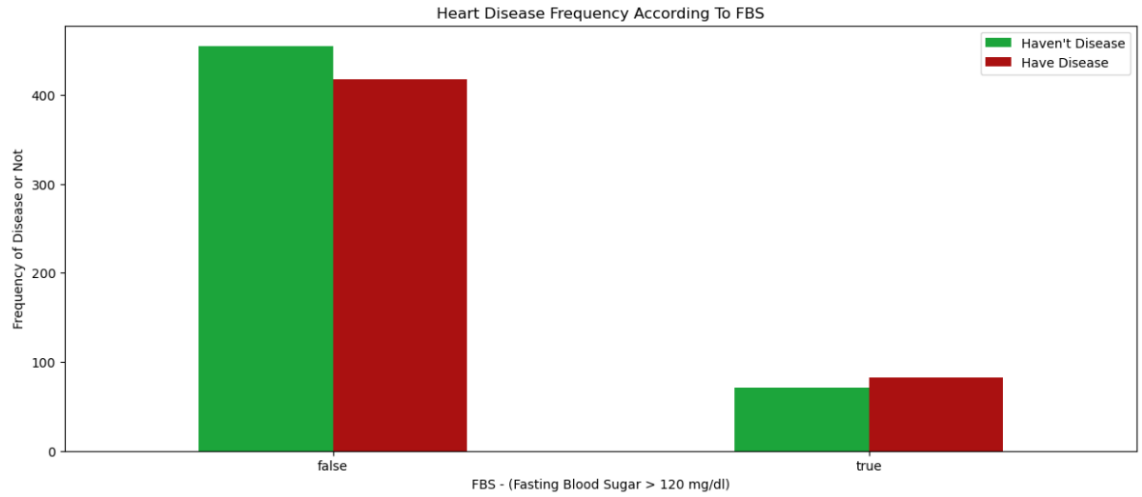
To gain an initial understanding of the relationship between variables and the target, we draw plots for Heart Disease Frequency with Sex (Figure 3.), Maximum Heart Rate (Figure 4.), and FBS (Figure 5.):



(Figure 3.)



(Figure 4.)



(Figure 5.)

3. Method

3.1 Algorithms Selection

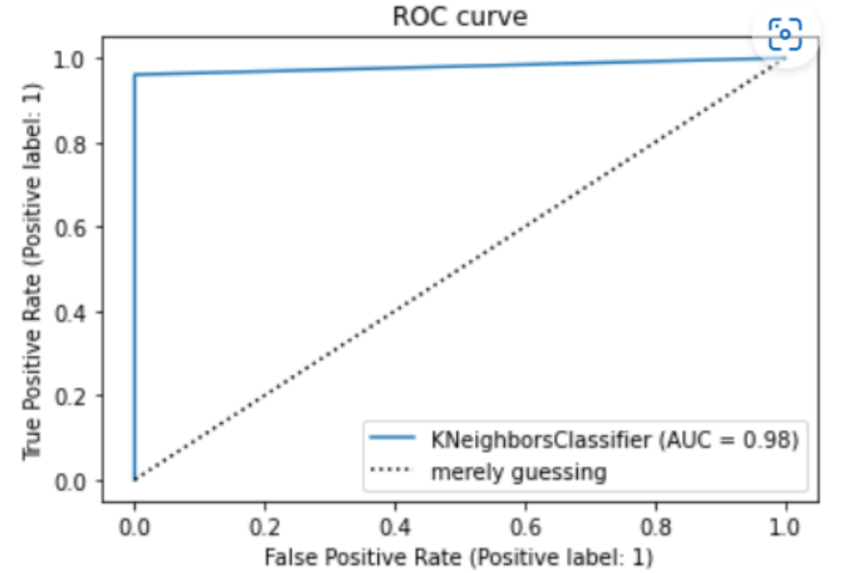
In order to find the best classification algorithm for a binary classification task, four common classifiers - Support Vector Machine, Logistic Regression, K-Nearest Neighbors, and Decision Tree - were evaluated. The hyperparameters for each classifier were optimized using GridSearchCV. After the evaluation, KNN was found to have the highest accuracy score of

0.981, with optimal hyperparameters of n_neighbors=1, p=1, and weights=uniform. Since accuracy was deemed the most important performance metric for this problem, KNN was selected as the preferred classifier.

3.2 Model Evaluation

For the purpose of evaluating the efficacy of the selected KNN classifier, the model was tested on the holdout test set. According to the classification report, the KNN classifier obtained an accuracy of 0.981, precision of 1.0, and recall of 0.961. The area under the ROC curve was 0.98, indicating that the model distinguishes between positive and negative cases exceptionally well. GridSearchCV was used to tune the hyperparameters, ensuring that the model is optimized for efficacy on unseen data. In addition, the model did not exhibit any indications of overfitting because the accuracy during training and testing were comparable.

```
accuracy=0.981, precision=1.0, recall=0.961, auc=0.98
```



3.3 Feature Selection

We applied backward elimination using a p-value approach to select the most significant features affecting heart disease risk. This technique repeatedly eliminated features with the highest p-value above 5% at a time and returned the regression summary with all p-values below 5%. After applying backward elimination, the final model included 11 significant features. Notably, sex played a substantial role in heart disease prediction, emphasizing its importance as a factor in risk assessment.

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.1815	1.159	2.746	0.006	0.910	5.453
sex	-1.8340	0.254	-7.216	0.000	-2.332	-1.336
cp	0.8457	0.098	8.592	0.000	0.653	1.039
trestbps	-0.0193	0.005	-3.553	0.000	-0.030	-0.009
chol	-0.0059	0.002	-2.928	0.003	-0.010	-0.002
restecg	0.4284	0.188	2.277	0.023	0.060	0.797
thalach	0.0251	0.005	4.808	0.000	0.015	0.035
exang	-0.9904	0.223	-4.441	0.000	-1.427	-0.553
oldpeak	-0.5645	0.115	-4.894	0.000	-0.791	-0.338
slope	0.5414	0.188	2.885	0.004	0.174	0.909
ca	-0.7659	0.102	-7.545	0.000	-0.965	-0.567
thal	-0.8768	0.152	-5.750	0.000	-1.176	-0.578

3.4 Sex Feature

We investigated the effect of sex on heart disease incidence by predicting morbidity for each patient using logistic regression. We then compared the predicted morbidity between males and females and found that females had a higher incidence of heart disease than males, when controlling for other biological factors. This suggests that sex is an important predictor of heart disease risk.

4. Conclusion and Future Works

In conclusion, our study aimed to develop a predictive model for heart disease incidence using machine learning techniques. Through data exploration, we identified sex as a significant predictor and found that females may be more prone to heart disease than males. Using KNN and logistic regression, we developed a model with high accuracy and performance metrics. However, our study has some limitations, such as the use of a single dataset and limited sample size. Future research could expand the recruitment pool to include a more diverse group of participants to improve the sample size and generalizability of the results. Overall, our study contributes to the field of heart disease prediction and provides insight into the potential benefits of using machine learning techniques for medical diagnosis.

5. Contributions

Member	Proposal	Coding	Presentation	Report
Wenhe Chen	1	1	1	1
Jiaqi Guo	1	1	1	1
Sybil Kong	1	1	1	1
Zeya Ling	1	1	1	1
Yutong Wei	1	1	1	1