Heart Disease Prediction with Naive Bayes Classifier

Yiting Wang, Yubo Zhang

Novermber 27, 2024

1 Introduction

1.1 Heart Disease Prediction

Heart disease, also known as cardiovascular disease (CVD), is a term referring to a variety of heart and blood vesicular problems [1]. As a leading cause of death globally and accounting for 1 in 5 deaths in the U.S, heart disease has a significant impact on populations worldwide [2]. According to (CDC), heart disease cost about $252.2 billion from 2019 to 2020 for health care services, medicines, and lost productivity due to death [3]. Therefore, a reliable prediction method on easily accessible predictors for heart disease is essential for public health management. Benefits of accurately predicting heart disease includes allowing the potential patients to gain early intervention, reducing mortality rate, providing insights into targeted treatment plans by revealing leading predictors for heart disease, and arousing awareness in heart disease risk factors among the general population [4]. In this study, we aim to develop and evaluate a predictive model for heart disease based on Naive Bayes algorithm, leveraging accessible predictors to enhance early intervention.

1.2 Naive Bayes Classifier

Bayesian inference has a wide range of applications in the field of machine learning, including cutting-edge models such as variational autoencoders to estimate the feature vector for image generation and Bayesian optimization for hyperparameter tuning. For this project, we chose to explore the Naive Bayes classifier due to its computational efficiency in terms of both memory and time. The model's assumption about feature independence also reduce the risk of overfitting to complexities in the dataset, and makes it robust with a smaller data size. The Bayes classifier also makes a good use of the concept of updating the posterior using the prior and likelihood.

Given the feature vector X as the input, we want to predict the class variable y, which in the context of heart disease prediction is either having the disease or not. To achieve this with a machine learning approach, the Bayes theorem gives us the posterior distribution $P(y \mid X)$ using the prior distribution $P(y)$ and the conditional probability of observing the training values

given the prior.

$$P(y \mid X) = \frac{P(y)P(X \mid y)}{P(X)} \tag{1}$$

By applying the assumption that all feature variables are independent, we can rewrite the conditional probability of the input to a product of that of all independent features

$$P(y \mid \mathbf{x}) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(\mathbf{x})} \tag{2}$$

Since the denominator is constant a set of feature vector input, we can just take the numerator

$$P(y \mid \mathbf{x}) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y) \tag{3}$$

Now, given training data (X, y) we can calculate the prior $P(y)$ computed as the proportion of samples belonging to each class.

$$P(y) = \frac{\text{Number of samples in class } y}{\text{Total number of samples}} \tag{4}$$

We can also calculate the conditional probability (likelihood) of each feature given the class outcome $P(x_i \mid y)$ and the approach will depend on the type of input features.

Then, the decision rule of the Bayes Classifier takes the most likely output (binary yes or no in our case of predicting heart disease) y_{pred} that maximizes the posterior probability.

$$y_{\text{pred}} = \arg\max_{y} P(y \mid \mathbf{x}) = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y) \tag{5}$$

Here, we can conclude the general discussion of Naive Bayes Classifiers. However, we have yet to specify the approaches to calculate the conditional probability of features for different types of input features. In our project, we will explore the Gaussian Naive Bayes Classifier, the Bernoulli Naive Bayes Classifier, and the Categorical Naive Bayes Classifier.

The Gaussian Bayesian model accepts continuous type variable and assumes a Gaussian distribution for the conditional probability of features.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{6}$$

Where, during training, for each class, the model calculates the mean μ_y and standard deviation σ_y for each feature, and then use these values to calculate the likelihood of features.

The categorical and Bernoulli Naive Bayesian Classifiers handles categorical and binary values respectively.

During training, the prior $P(y)$ is computed as the proportion of samples belonging to each class, and the likelihood is computed as the probability of each feature value occurring in the feature set for each class.

2 Method

2.1 Choosing the Dataset

We selected The Heart Disease dataset on Kaggle [5] consisting of 1025 patients. The data was from a 1988 collection, consisting from databases from Cleveland, Hungary, Switzerland, and Long Beach V. Out of the 76 attributes, 14 attributes including "target" (the presence of heart disease) were used in publications, so we will use these 14 attributes graphed below to apply the Naive Bayes Classifier on. The specific structure of the dataset is depicted in Table 1.

Feature Name	Description	Data Type
age	Age of the person in years	Continuous
sex	Sex of the person (1 = male, 0 = female)	Binary
cp	Chest pain type (4 values: 0, 1, 2, 3)	Categorical
trestbps	Resting blood pressure (in mm Hg)	Continuous
chol	Serum cholesterol in mg/dl	Continuous
fbs	Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)	Binary
restecg	Resting electrocardiographic results (0, 1, 2)	Categorical
thalach	Maximum heart rate achieved (in bpm)	Continuous
exang	Exercise induced angina (1 = yes, 0 = no)	Categorical
oldpeak	ST depression induced by exercise relative to rest	Continuous
slope	The slope of the peak exercise ST segment (3 values: 0, 1, 2)	Categorical
ca	Number of major vessels colored by fluoroscopy (0-3)	Categorical
thal	Thalassemia (0 = normal, 1 = fixed defect, 2 = reversible defect)	Categorical
target	Whether the person has heart disease (1 = yes, 0 = no)	Binary

Table 1: Feature Descriptions and Data Types

2.2 Data Distribution

To explore the dataset, we plotted the histogram for each feature as well as the predication target.

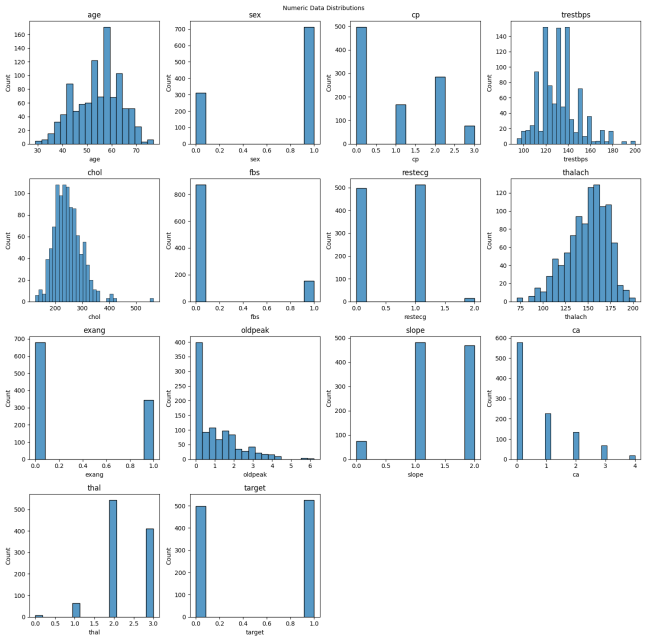

Figure 1: The distribution of 13 features and the target (presence of heart disease)

From the data distribution Figure 1, we can see the the age of the patients are slightly skewed to the right. The sex are predominantly male with a 7 to 3 ratio, which corresponds to the fact that men are more likely to experience heart diseases than women earlier in life. We also explored another dataset on Kaggle [6] and find that this imbalance between sex is also true.

To verify that Gaussian Bayes Classifier is suitable for continuous types in this dataset, we can observe that for features such as cholesterol (chol) and maximum heart rate (thalach), the distribution is approximately follows a standard distribution. For ST depression induced by exercise relative to rest (oldpeak), the distribution is skewed to the left, so this leaves further question to whether our classifier could be further improved with another approach of calculating the feature likelihood.

2.3 Data Covariance

To see whether the Naive assumption that all features are independent in the dataset is true, we plotted the covariance matrix and set the diagonal values to zero and observe the relationship between each feature in Figure 2.

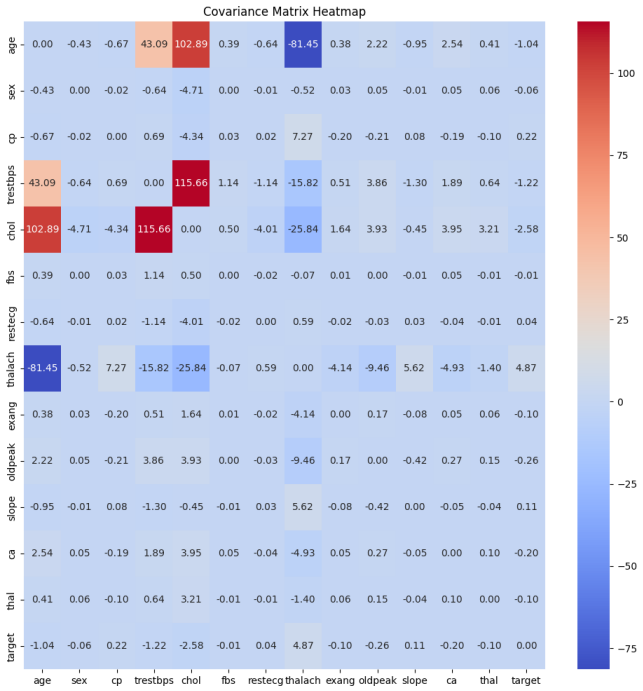

Figure 2: The covariance matrix showing the relationships between 13 features and the target (presence of heart disease)

The strongest relationship exists between cholesterol level (chol) and resting blood pressure (trestbps). They have a strong positive relationship. Other strong relationships includes cholesterol and age, thalach and age, trestbps and age.

To address how these strongly related features affect the performance of model due to their contradiction on the assumption that features are independent, we removed each related features one by one and did not observe significant performance increase.

2.4 Splitting Data

We randomly selected 70% of the dataset as the training data, and 30% of it as the test data. Since training for Bayes Classifier requires no hyper-parameter tuning, validation set is not required.

Furthermore, we spitted the features in training and testing input sets into three groups: continuous, binary, categorical. We want to train four different Bayes classifiers using different input type and thus different approach for calculating the feature likelihood. We used Gaussian Naive Bayes Classifier for continuous types, Bernoulli Naive Bayes Classifier for binary features, and Categorical Naive Bayes Classifier for categorical types.

We also designed a model that uses the equally weighted average of the log probability of each of the models to a combined model.

2.5 Model Training and Prediction

We utilized the sci-kit learning library which implements the Naive Bayes Classifiers discussed in the first section of this report. We used the training set with the appropriate feature types for each Naive Bayes Classifier, trained each model separately, and evaluated the accuracy of each model and their precision, recall, and F1-score. Furthermore, we used the average of the log likelihood of each model to predict with a combined model. This is a form of model aggregation. The implementation is shown in Figure 3

```python
# Split into train and test sets
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=42)

# Fit separate Naive Bayes models
gnb = GaussianNB()
bnb = BernoulliNB()
cnb = CategoricalNB()

# Separate features by type
continuous_features = ['age', 'trestbps', 'chol', 'thalach', 'oldpeak']
binary_features = ['sex', 'fbs']
categorical_features = ['cp', 'restecg', 'exang', 'slope', 'ca', 'thal']

X_train_continuous = X_train[continuous_features]
X_train_binary = X_train[binary_features]
X_train_categorical = X_train[categorical_features]

# Fit models to respective data
gnb.fit(X_train_continuous, y_train)
bnb.fit(X_train_binary, y_train)
cnb.fit(X_train_categorical, y_train)
```

```python
# Combine log probabilities for predictions
def predict_combined(X_continuous, X_binary, X_categorical):
    # Log probabilities from each model
    log_prob_gnb = gnb.predict_log_proba(X_continuous)
    log_prob_bnb = bnb.predict_log_proba(X_binary)
    log_prob_cnb = cnb.predict_log_proba(X_categorical)

    # Combine log probabilities
    combined_log_prob = log_prob_gnb + log_prob_bnb + log_prob_cnb

    # print(combined_log_prob)

    # Predict the class with the highest combined probability
    return np.argmax(combined_log_prob, axis=1)

X_test_continuous = X_test[continuous_features]
X_test_binary = X_test[binary_features]
X_test_categorical = X_test[categorical_features]

y_pred = predict_combined(
    X_test_continuous,
    X_test_binary,
    X_test_categorical
)

# Evaluate the model
acc_combined = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))
```

Figure 3: Data splitting and model training (left), model aggregation (right), implemented in Python

3 Result

We have trained a Guassian Native Bayes Classifier using continuous features, a Bernoulli Naive Bayes Classifier using binary features, a Categorical Naive Bayes Classifier using categorical features (not including the binary features), and a aggregated model with the previous three models, and their prediction accuracy is depicted in the following Figure 4. The Categorical model and the combined model are in a tie for the best accuracy, with an accuracy rate of 85.7% and 84.1%. Interestingly, the added predictors in combined variables do not improve the model performance in prediction accuracy. We will discuss possible explanations and improvements for that in the discussion section.

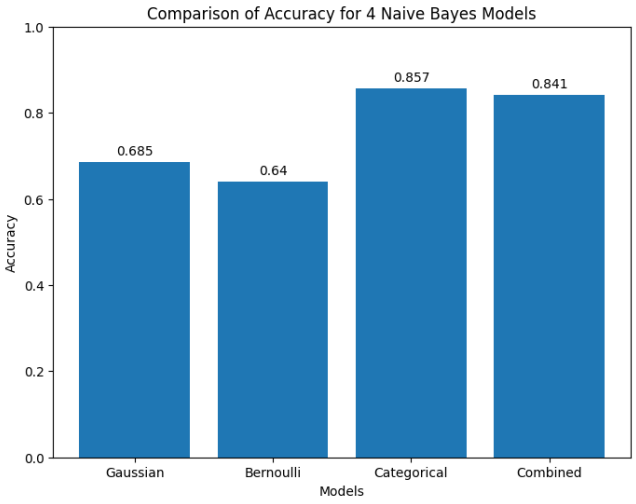

Figure 4: Accuracy for 4 Naive Bayes Classifiers

For our combined model, the precision, recall, and F1-score are plotted below in Figure 5

Accuracy: 0.8409090909090909

Classification Report:

	precision	recall	f1-score	support
0	0.89	0.79	0.84	159
1	0.80	0.90	0.85	149
accuracy			0.84	308
macro avg	0.85	0.84	0.84	308
weighted avg	0.85	0.84	0.84	308

Figure 5: The classification report for the combined model

4 Discussion

Naive Bayes models have high interpretability, allowing for easy understanding of how features influence the classification. Taking the best predicting model, the Categorical model as an example (Figure 6), we can understand how each value of a categorical feature contributes to the classification of whether the individual has heart disease or not by looking at the distribution of yes/no heart disease. In the distribution for chest pain feature (feature "cp", upper left), we can see that when a person has no chest pain (cp=0), they are more likely to be predicted as without heart disease. Any level of chest pain (cp= 1, 2, 3) would contribute to increased probability of being predicted as having heart disease.

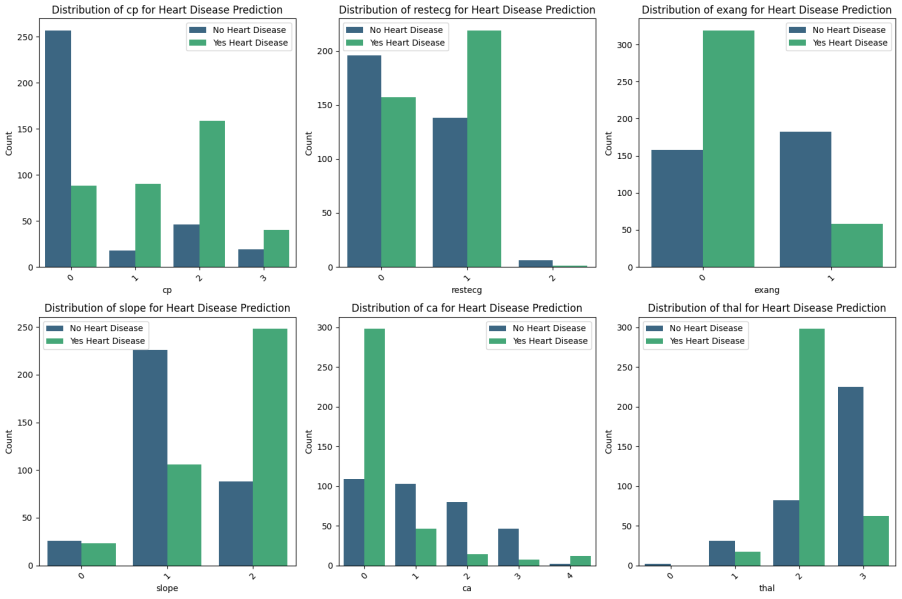

Figure 6: Distribution of yes/no Heart Disease within Categorical features

While using naive Bayes models are good exploratory approaches in studying the relationship between predicting factors and heart disease, this method has limitations. Firstly, the accuracy of our models are not sufficient for accurate predictions of heart disease for large populations for medical services. All models performance in recall are not satisfactory, indicating that the models' lack of capability of identifying true positive cases. False negatives in heart disease predictions can lead to missed opportunities in early intervention, late treatment for the patient and increased mortality. Therefore, increasing recall is an important goal in future research.

Secondly, the current model could not effectively leverage all available predictors. Overall, the Categorical model performed the best, and it seems that including the continuous and binary predictors does not improve the model's ability to accurately predict heart disease. Regarding the current dataset, the continuous variables do not perfectly follow Gaussian distribution, which might have led to biases in predictions and lowered the accuracy in our Gaussian-naive Bayes model. In future studies, we should increase the sample size to ensure Gaussian distribution among the continuous predictors in order to improve model performance.

Third, the Naive Bayes models are oversimplified to summarize the actual relationship between predictors and heart disease. The feature independence assumption is crucial for the Naive Bayes models, but it brings significant restrictions when attempting to include more informative predictors into the dataset. In real-world situations, it is unlikely that features related to heart disease are independent from each other. Therefore, researchers should also explore more advanced classifying methods such as Decision Tree (DT), Random Forest (RF), K-nearest neighbors (KNN), and etc.

Additionally, all features in a Naive Bayes model are considered to be contributing to the

prediction of heart disease equally, which is not likely to represent the reality. While the models shed a light on some potential dos and don'ts for preventing heart disease, further studies are needed for understanding the complex relationship between predictors and heart disease. Causal studies with controlled variables are needed for researchers to gain further insights into the prediction and prevention of heart disease.

5 Conclusion

In this study, we explored the use of the Naive Bayes model in predicting heart disease on 13 easily accessible predictors. Among the four models we tested, the Categorical Naive Bayes model and the combined model reached moderately high accuracy of 85.7% and 84.1%, respectively. In order to improve model accuracy and generalization of modeling outcomes, some future directions are to test the models on larger datasets, include other informative predictors, and use more advanced classifier models. Causal studies are also necessary to understand the relationship between heart disease and risk factors.

References

[1] https://medlineplus.gov/lab-tests/heart-disease-risk-assessment/

[2] https://www.cdc.gov/heart-disease/about/index.html

[3] https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html

[4] El-Sofany, H., Bouallegue, B. & El-Latif, Y.M.A. A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. Sci Rep 14, 23277 (2024). https://doi.org/10.1038/s41598-024-74656-2

[5] https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

[6] https://www.kaggle.com/datasets/asgharalikhan/mortality-rate-heart-patient-pakistan-h

[7] https://www.geeksforgeeks.org/naive-bayes-classifiers

[8] https://scikit-learn.org/1.5/modules/naive_bayes.html