Heart Attack Analysis and Prediction

XUE, Bingxin

18 December 2023

Abstract

This experiment aims to accurately detect potential heart disease patients by applying predictive models. The dataset for this experiment originates from Kaggle and includes detailed information about participants' physical characteristics. Various preprocessing methods and multiple models were employed to test their effectiveness. Through rigorous experimentation and result comparisons, it was determined that achieving the highest accuracy involves one-hot encoding of categorical data, followed by prediction using the Linear Discriminant Analysis (LDA) model, resulting in an impressive accuracy rate of 98.36%.

1 Introduction

Heart disease stands as a grave health concern, with accurate prediction playing a pivotal role in early intervention and treatment. Extensive research data emphasizes the importance of forecasting heart disease to identify high-risk individuals and implement preventive measures, ultimately reducing the incidence of heart-related ailments and mitigating associated adverse consequences.

2 Exploratory Data Analysis (EDA)

Dataset contains various factors, including:

Feature	Meaning
age	age in years
sex	sex (1 = male; 0 = female)
cp	chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 0 = asymptomatic)
trestbps	resting blood pressure (in mm Hg on admission to the hospital)
chol	serum cholestoral in mg/dl
fbs	fasting blood sugar ¿ 120 mg/dl (1 = true; 0 = false)
restecg	resting electrocardiographic results (0 = hypertrophy; 1 = normal; 2 = having ST-T wave abnormality)
thalach	maximum heart rate achieved
exang	exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slp	the slope of the peak exercise ST segment (0 = downsloping; 1 = flat; 2 = upsloping)
caa	number of major vessels (0-3) colored by flourosopy
thall	1 = fixed defect; 2 = normal; 3 = reversable defect
output	the predicted attribute - diagnosis of heart disease (angiographic disease status)

Data View

```
print(df.shape)
pd.set_option('display.float_format', '{:.2f}'.format)
df.describe()
```

0.0s

(303, 14)

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
count	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00
mean	54.37	0.68	0.97	131.62	246.26	0.15	0.53	149.65	0.33	1.04	1.40	0.73	2.31	0.54
std	9.08	0.47	1.03	17.54	51.83	0.36	0.53	22.91	0.47	1.16	0.62	1.02	0.61	0.50
min	29.00	0.00	0.00	94.00	126.00	0.00	0.00	71.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	47.50	0.00	0.00	120.00	211.00	0.00	0.00	133.50	0.00	0.00	1.00	0.00	2.00	0.00
50%	55.00	1.00	1.00	130.00	240.00	0.00	1.00	153.00	0.00	0.80	1.00	0.00	2.00	1.00
75%	61.00	1.00	2.00	140.00	274.50	0.00	1.00	166.00	1.00	1.60	2.00	1.00	3.00	1.00
max	77.00	1.00	3.00	200.00	564.00	1.00	2.00	202.00	1.00	6.20	2.00	4.00	3.00	1.00

Figure 1: Data Describe (Appearance of Abnormal Values)

```
df[df['caa']==4]
```

0.0s

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
92	52	1	2	138	223	0	1	169	0	0.00	2	4	2	1
158	58	1	1	125	220	0	1	144	0	0.40	1	4	3	1
163	38	1	2	138	175	0	1	173	0	0.00	2	4	2	1
164	38	1	2	138	175	0	1	173	0	0.00	2	4	2	1
251	43	1	0	132	247	1	0	143	1	0.10	1	4	3	0

Figure 2: Duplicated Data

Correlation

```
For variable sex, Cramer's V: 0.275437 and p value: 1e-06
For variable exng, Cramer's V: 0.433668 and p value: 0.0
For variable caa, Cramer's V: 0.48266 and p value: 0.0
For variable cp, Cramer's V: 0.510413 and p value: 0.0
For variable fbs, Cramer's V: 0.0 and p value: 0.625416
For variable restecg, Cramer's V: 0.162926 and p value: 0.006661
For variable slp, Cramer's V: 0.388156 and p value: 0.0
For variable thall, Cramer's V: 0.522014 and p value: 0.0
```

Figure 3: Cramer's V

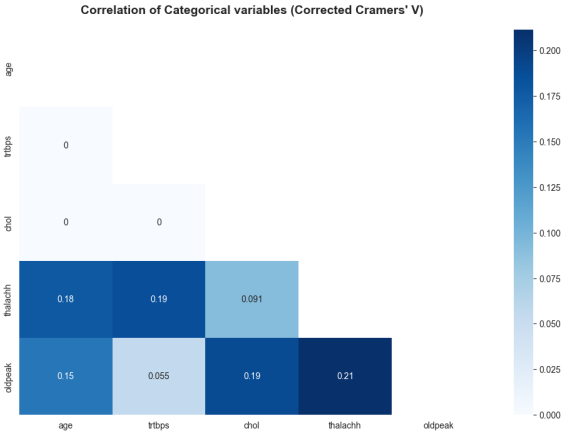

Figure 4: Correlation Coefficient Heatmap

Distribution

Figures are plotted in the appendix.

Association Rule Mining

	antecedents	consequents	support	confidence	lift
70	(fbs_0)	(trtbps_group_low)	0.85	1.00	1.00
69	(trtbps_group_low)	(fbs_0)	0.85	0.85	1.00
81	(sex_1)	(trtbps_group_low)	0.68	1.00	1.00
53	(exng_0)	(trtbps_group_low)	0.67	1.00	1.00
15	(caa_0)	(trtbps_group_low)	0.58	1.00	1.00
668	(exng_0, fbs_0)	(trtbps_group_low)	0.58	1.00	1.00
47	(exng_0)	(fbs_0)	0.58	0.86	1.01
667	(exng_0, trtbps_group_low)	(fbs_0)	0.58	0.86	1.01
669	(exng_0)	(trtbps_group_low, fbs_0)	0.58	0.86	1.01
806	(sex_1, fbs_0)	(trtbps_group_low)	0.57	1.00	1.00

Figure 5: Strong Correlation

	antecedents	consequents	support	confidence	lift
8292	(caa_0, age_group_young, trtbps_group_low, exn...	(output)	0.11	1.00	1.84
6737	(trtbps_group_low, cp_2, thall_2, sex_0)	(output)	0.10	1.00	1.84
10095	(fbs_0, restecg_1, sex_0, exng_0, thall_2)	(output)	0.12	1.00	1.84
4386	(caa_0, exng_0, thall_2, age_group_young)	(output)	0.11	1.00	1.84
2919	(cp_2, thall_2, sex_0)	(output)	0.10	1.00	1.84
1431	(caa_0, thall_2, age_group_young)	(output)	0.12	1.00	1.84
4439	(caa_0, thall_2, fbs_0, age_group_young)	(output)	0.12	1.00	1.84
11019	(caa_0, fbs_0, age_group_young, trtbps_group_l...	(output)	0.11	1.00	1.84
8314	(caa_0, fbs_0, age_group_young, trtbps_group_l...	(output)	0.12	1.00	1.84
8245	(caa_0, fbs_0, age_group_young, exng_0, thall_2)	(output)	0.11	1.00	1.84

Figure 6: Weak Correlation

Dataset

The experiment involved 303 participants, predominantly in the age group of 50-60 years, with the majority being male.

Data Integrity

No missing values were observed. An examination revealed that two participants had identical data. Given the highly unlikely occurrence of complete data duplication for two participants, one of the duplicate records was deemed redundant.

Outliers

The 'caa' variable ranges from 0 to 3, while 'thall' ranges from 1 to 3, suggesting potential outliers or anomalies in the dataset.

Correlation Analysis

Correlation analysis revealed relationships between certain features.
Positive correlations were observed between features such as 'cp' and 'slp,' while a negative correlation was noted with 'caa.'
From the Association Rule Mining section, it is evident that there is a significant correlation among some variables, with support exceeding 80% and confidence reaching 100%. However, the association between features and the outcome is comparatively weaker.

Special Findings

Based on the distribution plots, it is evident that the probability of developing heart disease is higher among females than males. Additionally, individuals diagnosed with heart disease tend to have a lower average age compared to those without heart disease.

3 Methodology

Given the potential biases introduced by discrete values, outliers, and high correlations between variables, corresponding measures were taken in the experiment.

Handling Missing Values

Remove Duplicate Values and Outlier Handling (V1): Eliminating redundant data and transforming outliers into missing values and imputing using methods such as KNN.

Outlier Handling

Standardization (V2): Standardizing numerical features to reduce the impact of outliers on model performance.

Categorical Data Processing

One-Hot Encoding (V3): Transforming categorical variables into one-hot encoding.

Continuous Data Processing

Discretization (V5): Discretizing continuous data to make it more suitable for model training.

Others

Combination of Standardization and One-Hot Encoding (V4): Obtaining comprehensive features by combining standardization and one-hot encoding.

Conducting hypothesis tests based on the two identified observations, the results indicate that the probability of females developing heart disease is not equal to that of males. Additionally, it is observed that the probability of developing heart disease is related to age. Combining these findings with the insights from the distribution plots, it is confirmed that the likelihood of heart disease among females is indeed higher than that among males, and the group with heart disease tends to have a younger average age compared to those without heart disease.

Model Selection

As the primary objective of this experiment is to infer whether participants are likely to develop heart disease based on their body information, it falls under a classification problem. Therefore, various classification models were employed for testing, including:, 'BernoulliNB', , 'KNN', , 'GradientBoosting', 'XGBoost', 'Adaboost'.

Model
Support Vector Machine
Linear Regression
Logistic Regression
Ridge Regression
Linear Discriminant Analysis
K-Nearest Neighbors

Model
Decision Tree
Random Forest
Bernoulli Naive Bayes
Gaussian Naive Bayes
GradientBoosting
XGBoost
Adaboost

4 Experiments

Evaluation

Given the distinct evaluation criteria for different models – such as using MSE for linear models and accuracy for regression models – both metrics were employed to assess model performance in the experiments.

Model scores served as statistical indicators to comprehensively gauge the performance variations across different methods. If a model exhibited the smallest MSE or the highest accuracy among all models, a score of one was assigned.

The first round of testing methodology primarily focused on comparing the impact of data preprocessing, considering variations before and after processing:

Version	**Methods**
V0	Original Data
V1	Remove Duplicate + Using KNN to Handling Abnormal Value
V2	Standardization
V3	One-Hot Encoding
V4	V2 + V3
V5	Discretization
V6	PCA

For the V1 example:

```
SVM                 MSE: 0.2951          Accuracy: 0.7049

Logistic            MSE: 0.0492          Accuracy: 0.9508

Linear              MSE: 0.0736          Accuracy: 0.918

Ridge               MSE: 0.0737          Accuracy: 0.918

BernoulliNB         MSE: 0.1148          Accuracy: 0.8852

LDA                 MSE: 0.082           Accuracy: 0.918

KNN                 MSE: 0.2131          Accuracy: 0.7869

DecisionTree        MSE: 0.1311          Accuracy: 0.8689

GaussianNB          MSE: 0.1311          Accuracy: 0.8689

RandomForest        MSE: 0.082           Accuracy: 0.918

GradientBoosting    MSE: 0.1311          Accuracy: 0.8689

XGBoost             MSE: 0.1148          Accuracy: 0.8852

Adaboost            MSE: 0.1475          Accuracy: 0.8525

Hightest Accurary: 0.9508                model: Logistic
Lowest MSE: 0.0492                       model: Logistic
Model Score: [0. 2. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
```

Figure 7: V1

We can see that the regression model performs best for both accuracy and mse.

The First Round Result

Figure 8: The First Round Result

5 Discussion

1. The results reveal that LDA performs the best, followed by logistic regression, both of which are linear classifiers. This suggests that the dataset may exhibit significant linear separability in the feature space.

2. Given the superior performance of the V3, subsequent experiments will build upon the V3 model. Attempts will be made to enhance model accuracy through the adjustment of hyperparameters.

3. The processing of discrete values and outliers led to a decrease in model accuracy, indicating a potential flaw in the handling of discrete values and outliers. The following experiments will try alternative methods for addressing these issues.

4. Despite the application of PCA, accuracy remained the same. Considering the dataset's scale, this case could be attributed to the data's inherent unsuitability for dimensionality reduction. Given that the original dimensionality of the data is not high, the correlation between features, while present, might not be strong enough to impact model performance significantly through dimensionality reduction.

6 The Second Round Test

Methods

Version	Methods
V7	Interquartile Range (IQR) Method: Discrete values were transformed to the upper or lower quartile of the respective feature.
V8	Remove Duplicate + Using mode to Handling Abnormal Value
V9	Remove Duplicate + Using Random Forest to Handling Abnormal Value

The Second Round Result

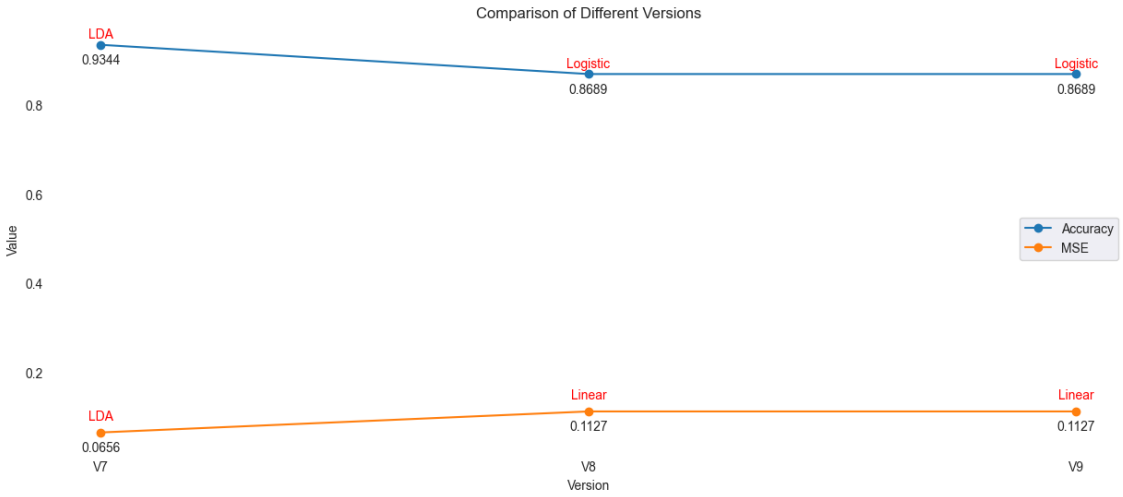

Figure 9: The Second Round Result

Other Methods and Analysis

Based on the results of the initial testing, attempting hyperparameter tuning on the best-performing model did not lead to an improvement in accuracy. This could be attributed to several factors:

Performance Plateau: Given that the LDA model already achieved high accuracy in the first round of testing, further hyperparameter tuning might not have significantly improved performance. The model may have reached a plateau in its ability to capture patterns in the data.

Data Characteristics: The nature of the dataset might limit the effectiveness of hyperparameter tuning. Some datasets inherently possess complexity or noise, making it challenging for a model to adapt to additional patterns through simple hyperparameter adjustments.

Insufficient Samples: The model might struggle to learn sufficient patterns if the training data is relatively small. In such cases, hyperparameter tuning could be constrained, and its impact may be limited.

7 Conclusion

Based on the presented results, the outcomes of processing discrete values and handling outliers do not surpass the performance of using the raw, unprocessed data. Therefore, the optimal approach involves transforming categorical data using one-hot encoding and subsequently employing the LDA model for prediction.

References

Jay. (2020). *Description of Features*. Kaggle. https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/discussion/234843

Naman manchanda. (2020). *Heart Attack - EDA + Prediction (90% Accuracy)*. Kaggle. https://www.kaggle.com/code/namanmanchanda/heart-attack-eda-prediction-90-accuracy

Shubhankar. (2021). *Heart Attack-(Complete EDA—Acc-93%—Auc-98%—F1-93%)*. Kaggle. https://www.kaggle.com/code/shubhankar095/heart-attack-complete-eda-acc-93-auc-98-f1-93

Appendix

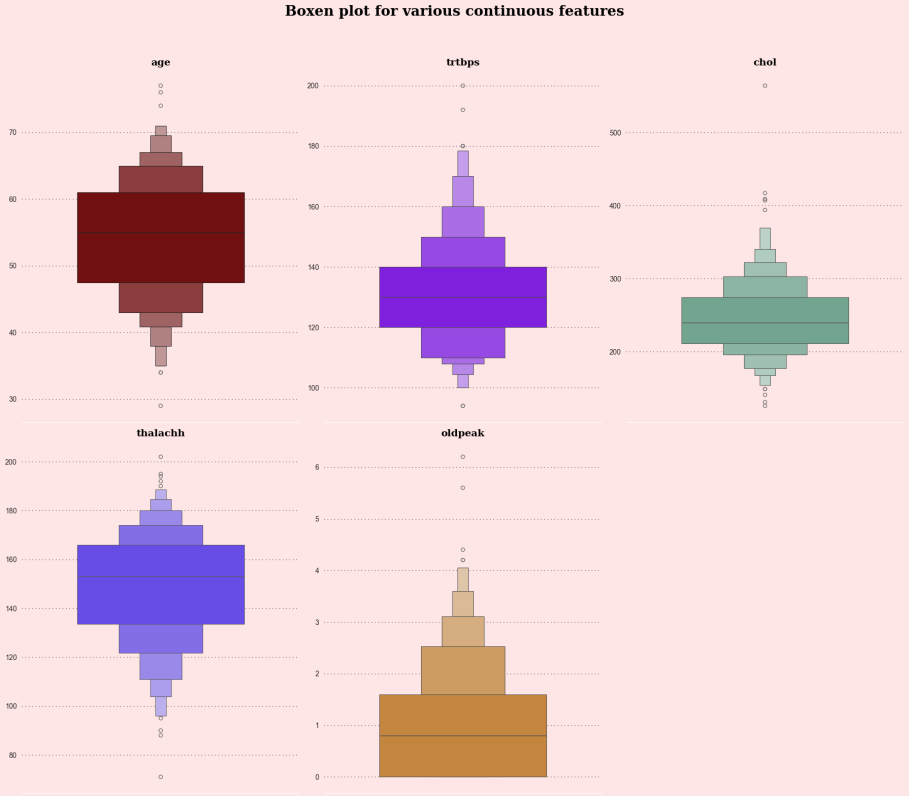

Figure 10: Overall Distribution

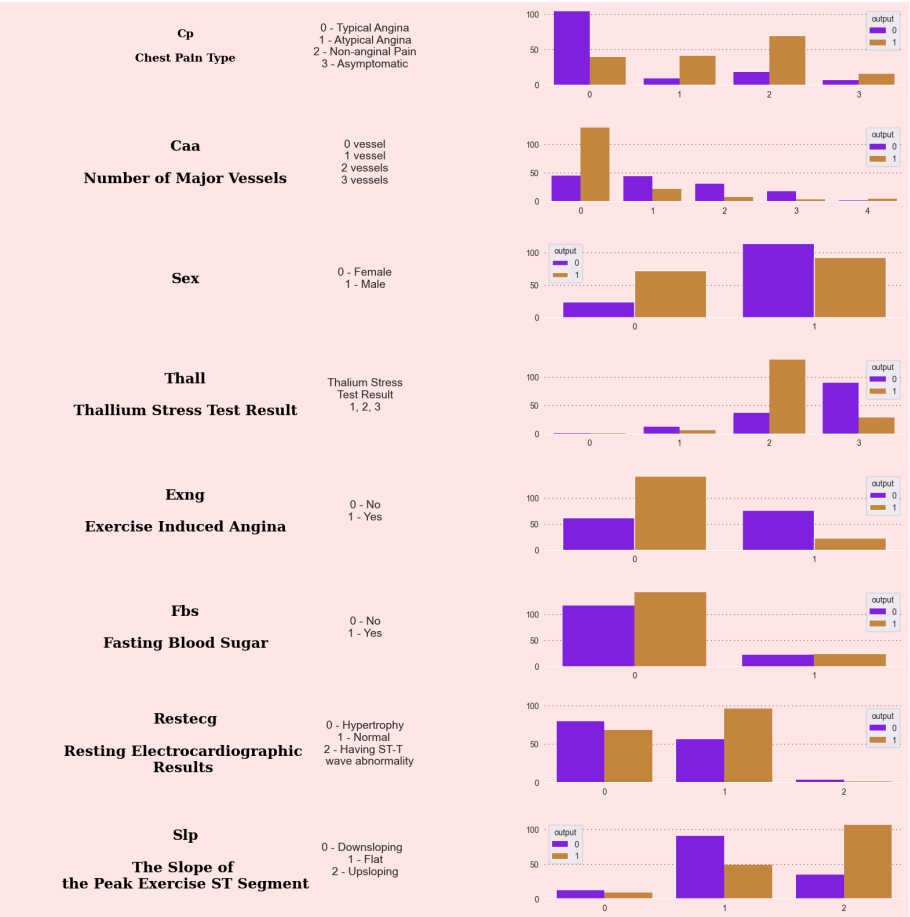

Figure 11: Categorical Data Distribution

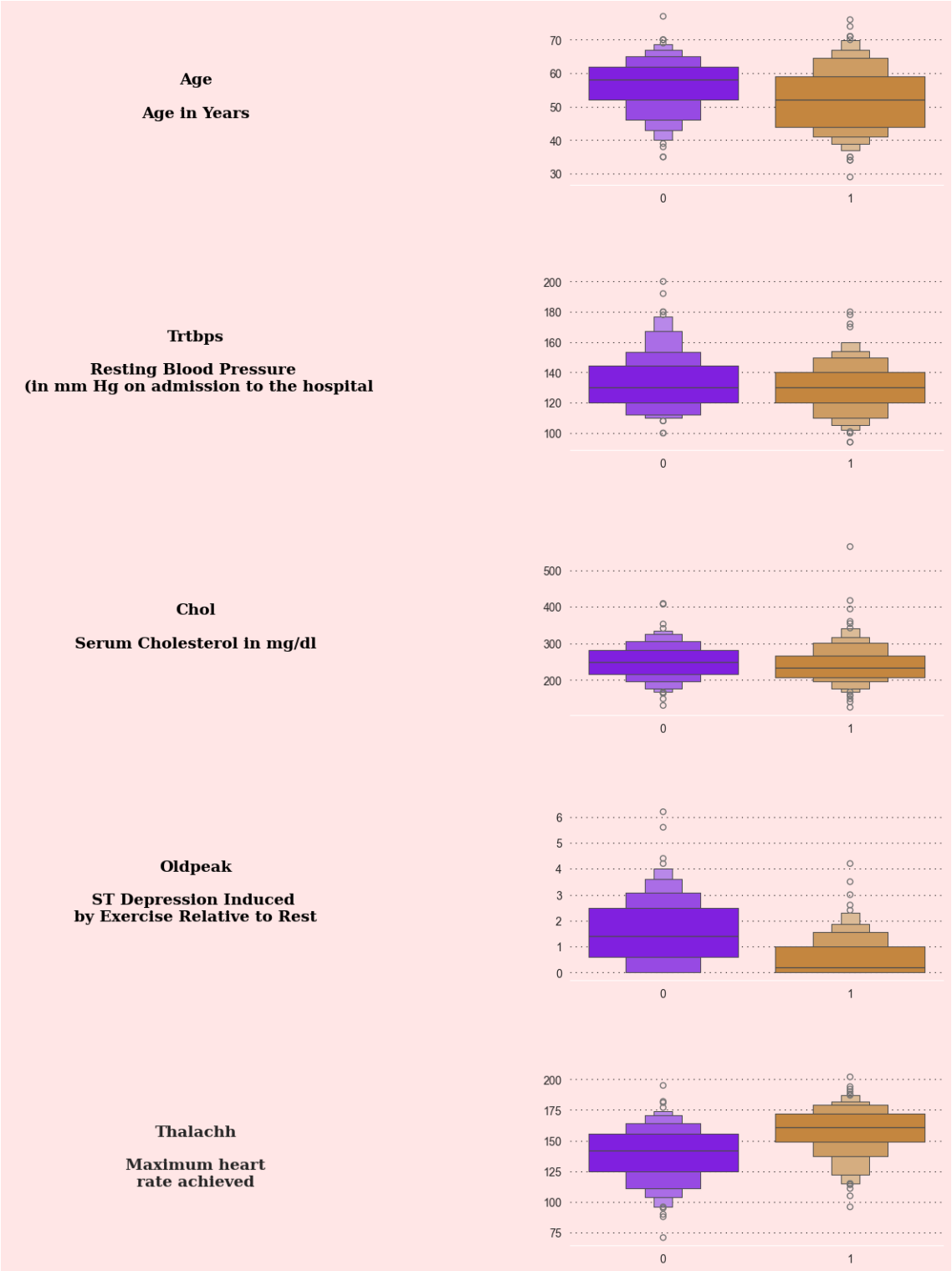

Figure 12: Continuous Data Distribution