Multivariate Prediction and Risk Stratification for Cardiovascular Diseases

ORIE 5741 Data Analysis Project

Peilin Li, Guohan Gao, Jingwen Liang

Introduction

Cardiovascular diseases (CVDs) remain the foremost cause of mortality worldwide, claiming approximately 17.9 million lives annually. This alarming figure represents 31% of all global deaths, underscoring the pervasive impact of these diseases. Heart attacks and strokes are the most fatal manifestations of cardiovascular diseases, accounting for 80% of all CVD-related deaths. Alarmingly, one-third of these deaths occur prematurely in individuals under 70 years of age, highlighting the critical need for effective preventive measures and interventions.

Individuals at high risk for cardiovascular diseases—owing to factors such as hypertension, diabetes, hyperlipidemia, or pre-existing heart conditions—benefit immensely from advanced diagnostic tools. Machine learning models stand at the forefront of this technological revolution, offering new avenues for early detection and personalized treatment strategies. In our project, the team used machine learning models to analyze complex datasets to identify patterns and predict risks, thereby facilitating timely and targeted interventions. By trying different machine learning models, the team hopes to identify the optimal model for Cardiovascular Diseases prediction. Such advancements not only promise to enhance clinical outcomes but also pave the way for a proactive approach in managing cardiovascular health, ultimately reducing the global burden of these life-threatening diseases

Problem Statement

There are 11 potential features or variables that may be associated with Cardiovascular Diseases in the dataset. Which of these features are the most relevant factors of Cardiovascular Diseases? Based on this data, is it possible to construct a predictive machine learning model that can accurately assess an individual's risk of developing Cardiovascular Diseases?

Therefore, the primary goal of the project is to:

1. **Pinpoint the most relevant/risky factors of Cardiovascular diseases**
2. **Find accurate predictive models that contributes to a proactive and patient-centric approach in managing cardiovascular health**

Potential Values of the Project

1.Enhancing Early Detection and Management

The initiative facilitates early detection and proactive management of heart conditions through the identification and analysis of critical predictive features. This approach is intended to enable more effective interventions, improve patient survival rates, and ultimately reduce the progression of heart failure.

2.Reduction in Healthcare Costs

Early detection and effective management of heart conditions can substantially reduce long-term healthcare costs. By identifying heart failure risks earlier, the project supports interventions that prevent conditions from worsening, thereby decreasing the need for more extensive and costly treatments.

3.Commercial and Technological Advancement

The project provides commercial enterprises with opportunities to enhance their technological stature and competitive edge. Investing in advanced machine learning models not only aligns with public health objectives but also offers direct commercial benefits such as reduced healthcare insurance payouts and improved overall employee health, strengthening the business's market position.

Dataset

In this project, the team utilized the Heart Disease Dataset Fedesoriano (Soriano, n.d.). This dataset comprises data for 918 patients, each characterized by 11 distinct features along with a binary target variable that denotes the presence or absence of heart disease. The dataset is meticulously curated, free from any missing entries, and encompasses three types of data: numerical, categorical, and binary. This comprehensive and clean dataset ensures a robust foundation for our analytical models.

Here is an overview of the dataset:

Table 1

Dataset Overview

Attribute	Description	Data Type
Age	Age of the patient [years]	Numerical
Sex	Sex of the patient [M: Male, F: Female]	Categorical
ChestPainType	Chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]	Categorical
RestingBP	Resting blood pressure [mm Hg]	Numerical
Cholesterol	Serum cholesterol [mm/dl]	Numerical
FastingBS	Fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]	Binary
RestingECG	Resting electrocardiogram results [Normal, ST: ST-T wave abnormality, LVH: left ventricular hypertrophy]	Categorical

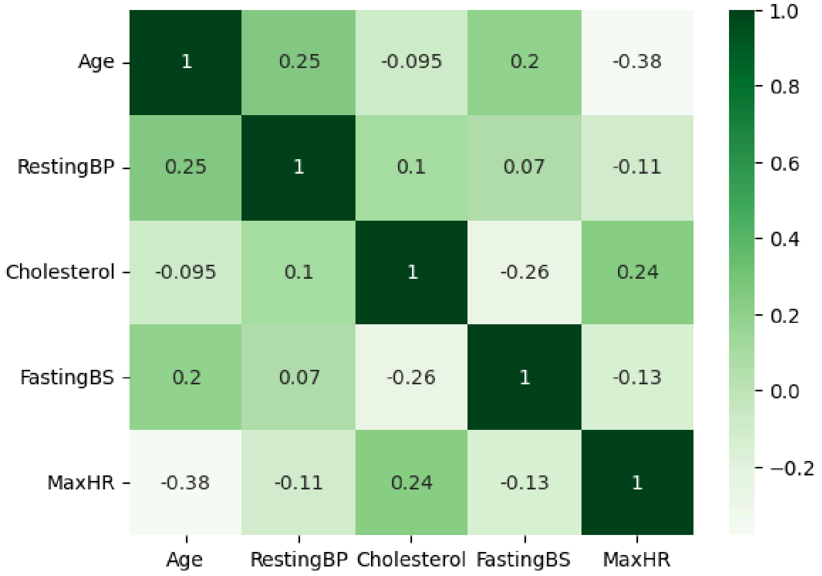
MaxHR	Maximum heart rate achieved [Numeric value between 60 and 202]	Numerical
ExerciseAngina	Exercise-induced angina [Y: Yes, N: No]	Binary
Oldpeak	ST depression induced by exercise relative to rest [Numeric value]	Numerical
ST_Slope	Slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]	Categorical
HeartDisease	**Output class [1: heart disease, 0: Normal]**	**Binary**

Exploratory Data Analysis(EDA)

To evaluate dataset quality, an EDA was conducted to explore the relationships and distributions among the features. The correlation matrix(shown in Figure 1), focusing on the numerical features—Age, Resting Blood Pressure (Resting BP), Cholesterol, Fasting Blood Sugar (Fasting BS), and Maximum Heart Rate (Max HR), revealed low correlations, indicating weak linear relationships between these features. For example, the highest positive correlation, a modest 0.25 between Age and Resting BP, suggested only a slight increase in blood pressure with age, while a negative correlation of -0.38 between Age and Max HR suggested a decline in maximum heart rate as individuals get older. Furthermore, the low correlations between other pairs of features, such as Resting BP and Cholesterol (0.1), Cholesterol and Fasting BS (-0.26), and Fasting BS and Max HR (-0.13), underscored the complex and multifaceted nature of cardiovascular health indicators.

Figure 1

Features Correlation Matrix



Numerical features distribution revealed that most ages centered around the 60s, which showed that the sample was mostly composed of middle-aged or elderly people(see Appendix A, Figure A1). Oldpeak values, representing heart stress scores, clustered near zero, suggesting minimal cardiac stress in many subjects(see Appendix A, Figure A5). Regarding Cholesterol, while most levels fell within the normal range, a significant peak at higher levels was observed, reflecting variance in dietary or genetic factors among the population(see Appendix A, Figure A3). Resting Blood Pressure typically peaked between 120 and 140 mmHg, aligning with the regular range for adults(see Appendix A, Figure A2). Furthermore, the distribution of Maximum Heart Rate was bimodal, illustrating varied fitness levels across the study population(see Appendix A, Figure A4).

In terms of the distribution of categorical and binary features, male patients outnumbered females(see Appendix A, Figure A7), and the majority reported no symptoms of chest pain(see Appendix A, Figure A10). Most participants had normal Fasting Blood Sugar levels(see Appendix A, Figure A9), and less than half reported exercise-induced chest pain, indicating these were not key risk factors in the sample(see Appendix A, Figure A8). While most ECG results were normal, some indicated potential heart issues. The most common ST Slope observed was flat, which could signify a higher risk of heart disease and underscore the need for detailed heart evaluations(see Appendix A, Figure A6).

In the Heart Disease Dataset, there was a slight imbalance between the number of patients with and without heart disease, with 508 patients having heart disease and 410 patients being classified as normal. This disparity could potentially impact the accuracy of prediction results. To address this issue and enhance the fairness and accuracy of predictions across different patient groups, the Synthetic Minority Over-sampling Technique(SMOTE) was employed(Chawla, Bowyer, Hall, & Kegelmeyer, 2002). SMOTE generated new synthetic samples from the minority class ('Normal' in this case) to balance the dataset. This ensured that the models would be trained on a more balanced representation of both classes. This approach not only supports better model performance but also contributes to more equitable health outcomes by improving the model's ability to accurately predict heart disease across diverse patient profiles.

Methodology

In the data preprocessing phase, it was important to conduct an integrity check to ensure that there were no missing values in the dataset. Subsequently, categorical features such as sex and resting ECG, which cannot be directly used in the algorithms, were converted into numerical features. This conversion was accomplished by one-hot encoding, which transformed these categorical variables into binary data so that the models could interpret them. As a result, the feature space expanded from 11 features to 21 features. Finally, the dataset was divided into training and testing sets to evaluate the performance of the model, with 80% of the data allocated for training and 20% reserved for testing.

This project used three machine learning models to predict heart disease risk. One key model was the Support Vector Machine(SVM). SVM is a common supervised learning method for classification and regression(Cortes & Vapnik, 1995). To improve SVM's performance, dimension reduction was implemented by Principal Component Analysis(PCA)(Jolliffe, 2002). This helped find the most important features and make the models simpler. SMOTE(Chawla, Bowyer, Hall, & Kegelmeyer, 2002) was also used with SVM to fix the class imbalance issue. Different kernel functions, such as linear, radial basis function(RBF), polynomial, and sigmoid, were applied to see how they affected the models' performance.

Logistic Regression, a statistical model commonly used for binary classification problems, was also implemented(Hosmer, Lemeshow,& Sturdivant, 2013). The Logistic Regression model was trained and evaluated with SMOTE(Chawla, Bowyer, Hall, & Kegelmeyer, 2002) to ensure a balanced learning process. Furthermore, the application of regularization techniques in Logistic Regression was explored to mitigate overfitting and improve the model's generalization ability.

Moreover, the Random Forest algorithm, an ensemble learning method that combines multiple decision trees to make robust predictions, was employed(Breiman, 2001). Known for its ability to handle high-dimensional data and capture complex relationships between features, the Random Forest model was trained with SMOTE(Chawla, Bowyer, Hall, & Kegelmeyer, 2002) to address the class imbalance and its performance in predicting cardiovascular disease risk was assessed.

By applying these machine learning techniques, the project aimed to develop more accurate and reliable models for predicting and stratifying the risk of cardiovascular diseases. Using SMOTE(Chawla, Bowyer, Hall, & Kegelmeyer, 2002), dimension reduction(Jolliffe, 2002 for PCA), regularization(Hastie, Tibshirani, & Friedman, 2009), and ensemble methods such as Random Forest(Breiman, 2001) allowed the challenges posed by class imbalance, high-dimensional data, and potential overfitting to be addressed. Through comprehensive evaluation and comparison of these models, it was possible to identify an effective strategy for forecasting the likelihood of cardiovascular disease risk, providing valuable insights for clinical decision-making and patients' care and management.

Results

In this project, the team used accuracy(Equation (1)), precision(Equation (2)), recall(Equation 3), and Mean Squared Error(MSE)(Equation 4) as the model performance evaluation metrics.

$$Accuracy = \frac{Number\ of\ Correct\ Prediction}{Total\ Number\ of\ Prediction} \tag{1}$$

$$Presicion = \frac{True\ Positives}{Treu\ Positive + False\ Positives} \tag{2}$$

$$Presicion = \frac{True\ Positives}{Treu\ Positive + False\ Negatives} \tag{3}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \tag{4}$$

The results of each model are shown as below(the greatest rates for accuracy, precision, recall, and the smallest score in MSE are bolded):

Table 2

Models Evaluation Part. 1

	SVM (Linear Kernel)	SVM (Linear Kernel) with PCA	SVM with SMOTE(Linear Kernel)	SVM with SMOTE(RBF Kernel)	SVM with SMOTE(Polynomial Kernel)	SVM with SMOTE(Sigmoid Kernel)
Accuracy	76.6%	71.7%	76.4%	65.1%	69.8%	46.2%
Precision	74.7%	66.4%	77.1%	**81.8%**	82.8%	44.2%
Recall	75.6%	80.2%	72.6%	35.3%	47.1%	45.1%
MSE	0.234	0.283	0.236	0.349	0.302	0.538

Table 3

Models Evaluation Part.2

	Logistic Regression	Logistic Regression with SMOTE	Random Forest(100 estimator)	Random Forest with SMOTE(100 estimator)	Random Forest with SMOTE (200 estimators)
Accuracy	77.7%	75.5%	79.4%	**82.1%**	**82.1%**
Precision	77.1%	77.8%	77.9%	80.8%	80.8%
Recall	74.4%	68.6%	77.9%	**82.4%**	**82.4%**
MSE	0.223	0.245	0.207	**0.179**	**0.179**

According to the results listed above, the Random Forest(Breiman, 2001) with SMOTE(Chawla, Bowyer, Hall, & Kegelmeyer, 2002) has the best overall performance among all the models since it has the greatest accuracy and recall rates and the smallest MSE.

For the SVM models, the SVM model using the linear kernel function has 76.6% accuracy and 74.7% precision, which means it has better performance among other SVM models in terms of these two metrics. Integrating PCA() with SVM and reducing the feature dimensionality to 9, on the contrary, reduces accuracy and precision rates to 71.7% and 66.4% respectively. However, the recall rate for the SVM with PCA has an improvement, which means it has better performance in identifying true positive rates. Comparing SVM with different kernels, the linear kernel has the highest precision of 77.1% and a relatively high accuracy rate of 76.4%. The RBF function and Polynomial kernels, have the higher precision rate at certain points but they have much lower recall rate and increase MSE, especially the RBF kernel with a recall rate of only 35.3%. In this project, as the team applied SMOTE(Chawla, Bowyer, Hall, & Kegelmeyer, 2002) to the dataset and the data sample became balanced, as the kernel function changes, the model has different performance respectively. The performance of each model with different kernel functions ranking from the best to the worst performance is: Linear Kernel, RBF Kernel, Polynomial Kernel, and Sigmoid Kernel. Surprisingly, even though the SVM using the

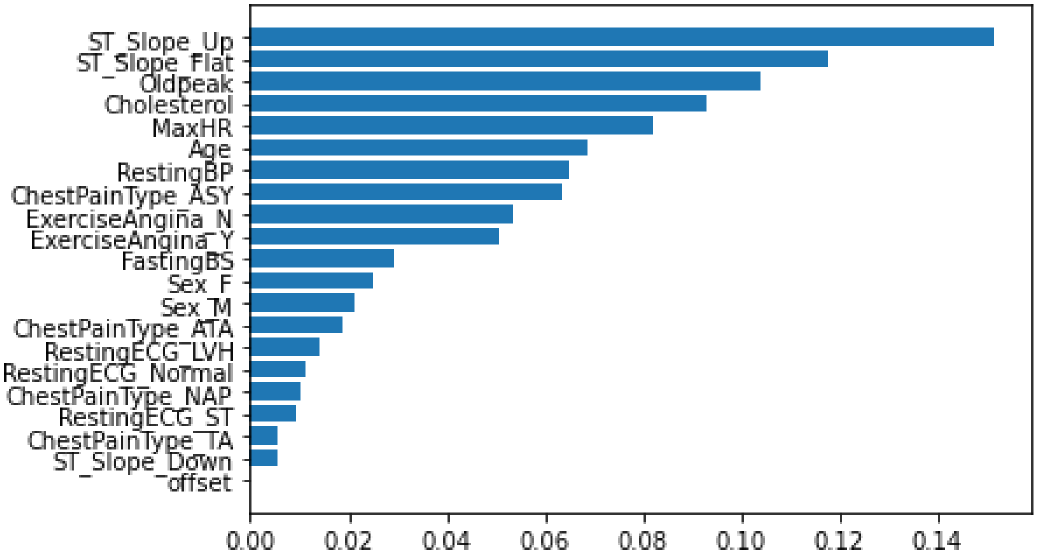
Polynomial Kernel has relatively low accuracy and recall rate, it has the highest precision rate of 82.8%. The Logistic Regression model has better overall performance compared to the SVM and its performance remained stable before and after applying SMOTE(Chawla, Bowyer, Hall, & Kegelmeyer, 2002). The Random Forest model(Breiman, 2001) has the best overall performance among the other models as it has the highest accuracy rate of 82.1%, precision rate of 80.8%, and the highest recall rate of 82.4%. While the team tried to increase the number of estimators to improve the performance, it did not gain significant improvements, which is also reasonable since it indicates that it has reached an optimal level of complexity with the current model.

While the team was trying to eliminate the effect of data imbalance by applying SMOTE(Chawla, Bowyer, Hall, & Kegelmeyer, 2002) to the dataset, according to the results above, the team did not observe a consistent enhancement in performance for most of the models. Especially, SVM models and the Logistic Regression did not benefit a lot from SMOTE(Chawla, Bowyer, Hall, & Kegelmeyer, 2002) in terms of accuracy rate, which remained under 77%. However, SMOTE(Chawla, Bowyer, Hall, & Kegelmeyer, 2002) notably improved the performance of Random Forest models(Breiman, 2001), pushing its accuracy up to 82.1%.

Additionally, according to the result of random forest(Breiman, 2001), the team were able to draw a bar chart with the features sorted from the most important to the least importance(as shown in Figure 2). According to the bar chart, the most influential feature is *ST_Slope_Up*. This indicates that the upward slope of the ST segment in an ECG reading is highly important in predicting the outcome. *ST_Slope_Flat* is also quite significant, but slightly less so than an upward slope. Therefore, it indicates that different behaviors in the ST segment are critical indicators for the model, reflecting variations in heart function. Then the *Oldpeak* ranks next, and *Cholesterol*, *Maximum Heart Rate*, and *Age* as important but less influential compared to the ST segment features. Each of the above features plays a role in our predictive capabilities but to a lesser extent.

Figure 2

Features Importance



Discussion

Limitation and Future Direction

1.Sample Size

The dataset only comprises 918 observations, which might not be sufficient to capture all the variability in heart failure outcomes across different populations. One of the direct solutions would be to increase the number and diversity of patient records to enhance the generalizability and robustness of the model. Besides, including data from a wider range of geographical locations (better if international) and demographics can help understand how different populations are affected by cardiovascular diseases.

2. Feature Scope

The dataset only includes primarily clinical attributes, potentially omitting other predictors like lifestyle factors, genetic predispositions, or socioeconomic status. The lack of other potential attributes could affect the generalization of the models. Therefore, it would be helpful to consider incorporating more features or non-traditional data types like patient lifestyle information, genetic markers, or long-term monitoring data if possible in the future.

3.Static Model

The current model doesn't account for changes over time in patient condition. To improve, if follow-up time is recorded, we can use time-series analysis or survival analysis techniques to model the timing of events and not just their occurrence.

Weapon of Math Destruction and Mitigation Strategies

1.Are outcomes hard to measure?

Measuring outcomes in this project can be complex due to the multifaceted nature of cardiovascular diseases and the variability in how individuals respond to treatments suggested by predictive models. However, short-term indicators and longitudinal studies may provide more immediate and observable metrics to gauge the effectiveness and accuracy of the predictions.

2.Could its prediction harm anyone?

The predictions from the machine learning models could potentially harm individuals through overdiagnosis, leading to unnecessary treatment, or underdiagnosis, missing crucial preventive measures. Therefore, it is important for us to ensure that the model predictions are used as a support tool alongside traditional clinical assessments to mitigate these risks.

3.Could it provide a feedback loop?

There's a risk of creating feedback loops where predictive models might perpetuate biases or lead to uneven resource allocation. To avoid this, it's essential for us to regularly update the models with diverse data sets and maintain broad criteria for care to ensure equitable treatment and prevention strategies across different populations.

References

Breiman,L.(2001).Random Forests. *Machine Learning,45*(1), 5-32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *The Journal of Artificial Intelligence Research, 16*, 321-357. https://doi.org/10.1613/jair.953

Cortes,C.,& Vapnik,V.(1995).*Support-Vector NetworksMachine Learning, 20*(3),273-297

Hastie, T., Tibshirani, R., & Friedman, J. (2009).*TheElements ofStatistical Learning: Data Mining, Inference,and Prediction.*Springer.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R.X.(2013). *Applied Logistic Regression* (3rd ed.). Wiley.

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer.

Soriano, F. (2021). *Heart Failure Prediction* [Data set]. Kaggle. Retrieved Month Day, Year, from https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data

Appendix A

Numerical Data Distribution

Figure A1

Age Distribution

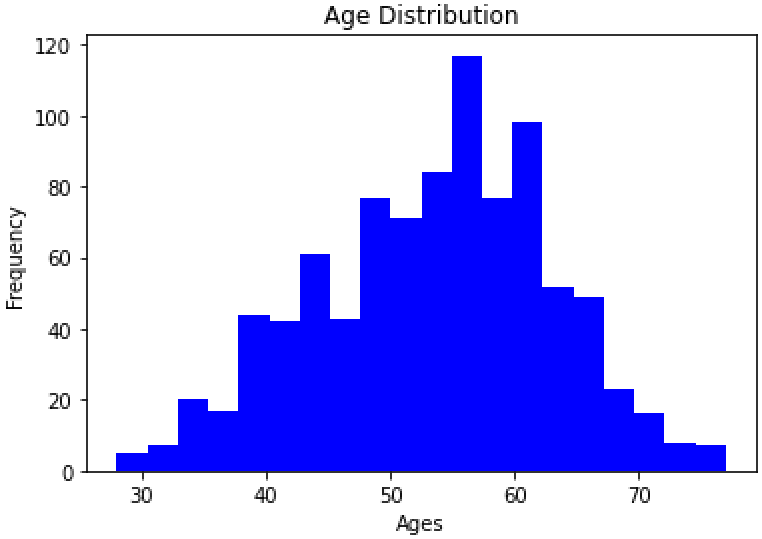


Figure A2

Resting Blood Pressure Distribution

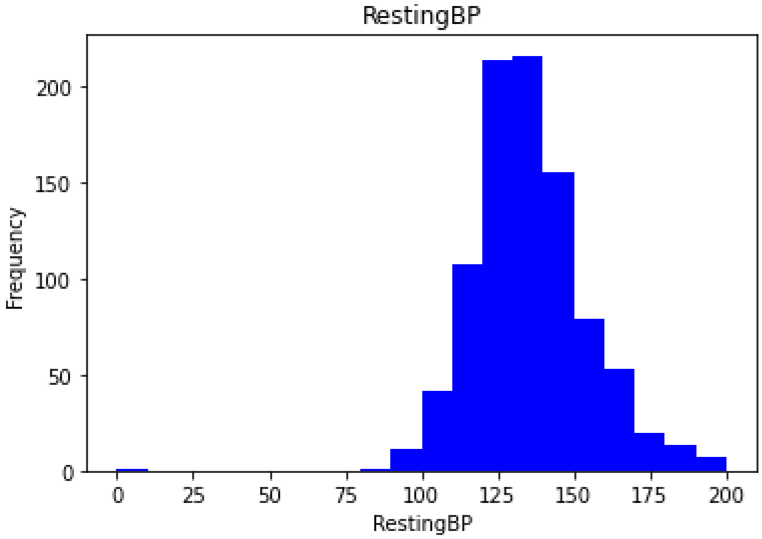


Figure A3

Cholesterol Distribution

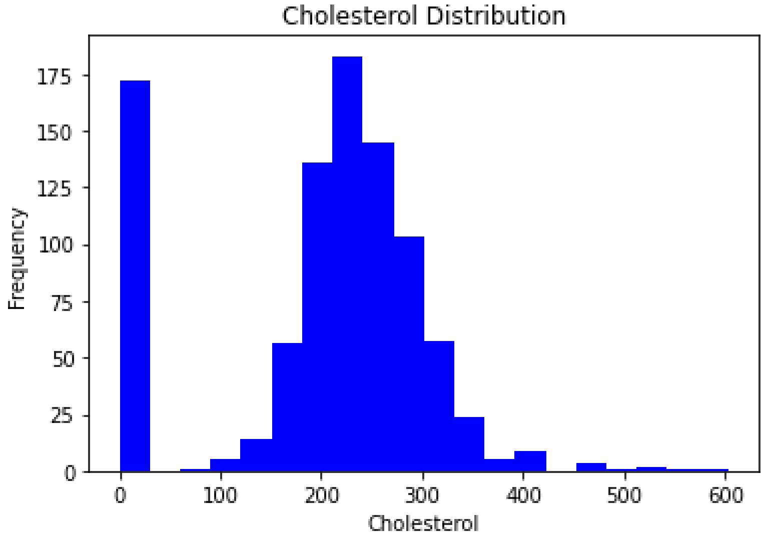


Figure A4

Maximum Heart Rate Distribution

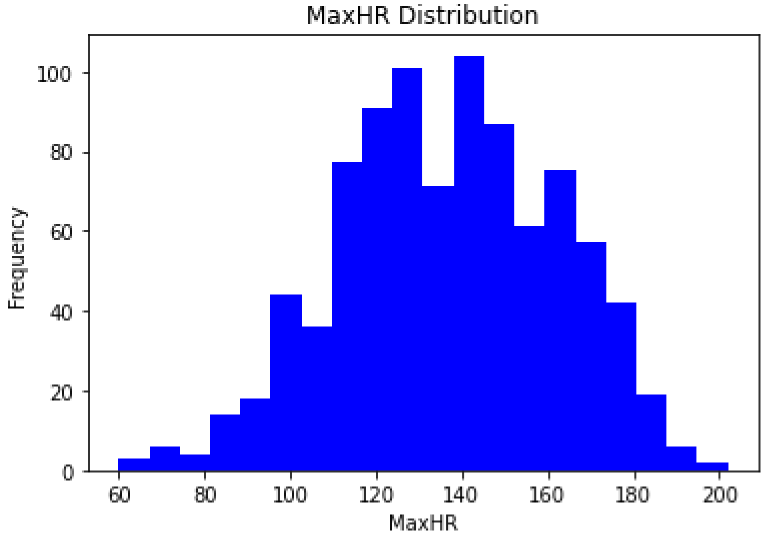
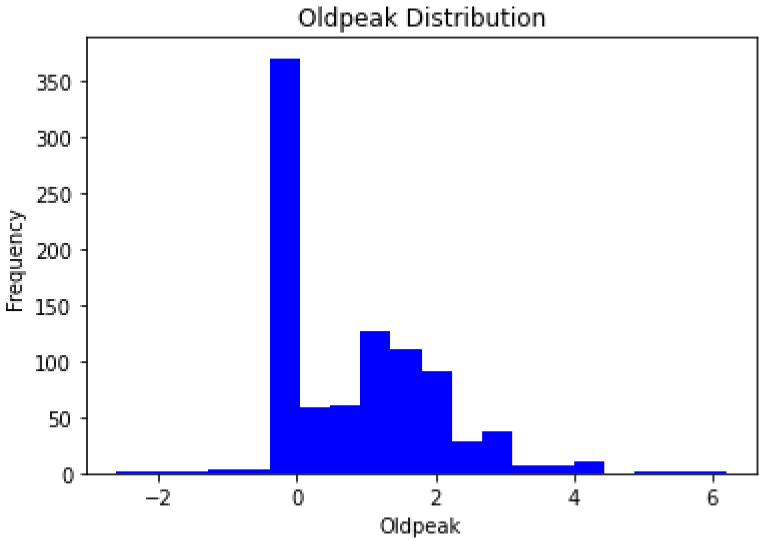


Figure A5

Oldpeak Distribution



Categorical Data Distribution

Figure A6

ST_Slope Distribution

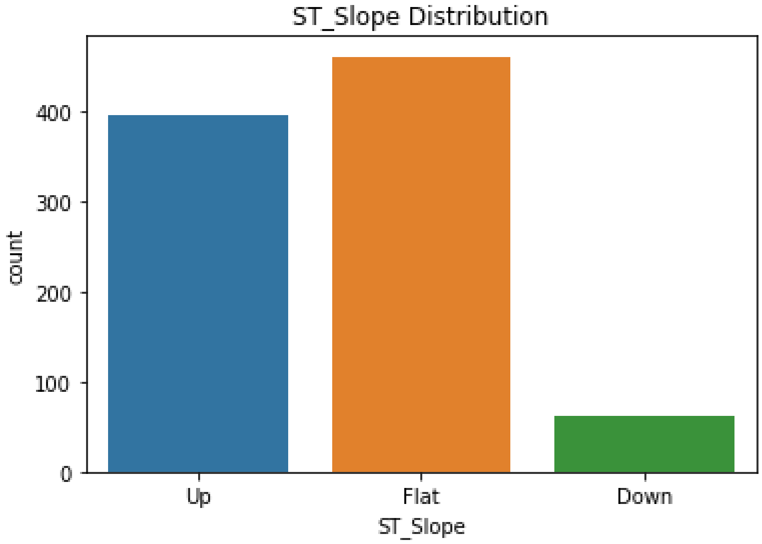


Figure A7

Sex Distribution

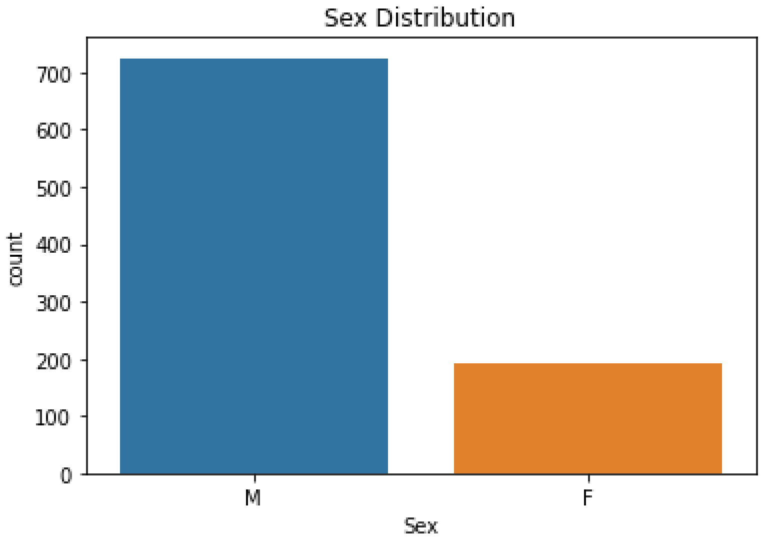


Figure A8

Exercise-induced Angina Distribution

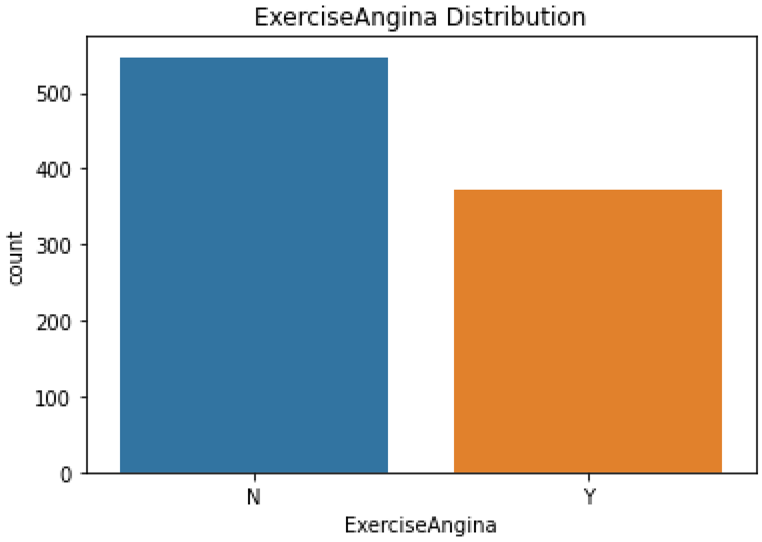


Figure A9

Fasting Blood Sugar Distribution

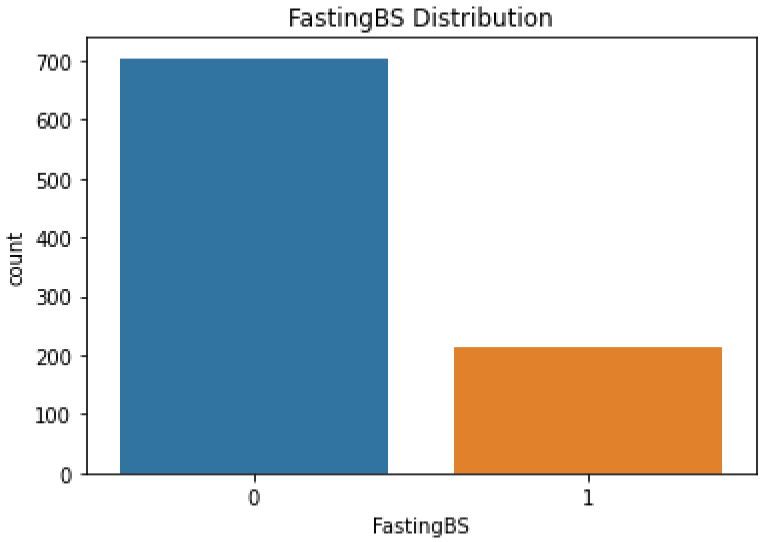
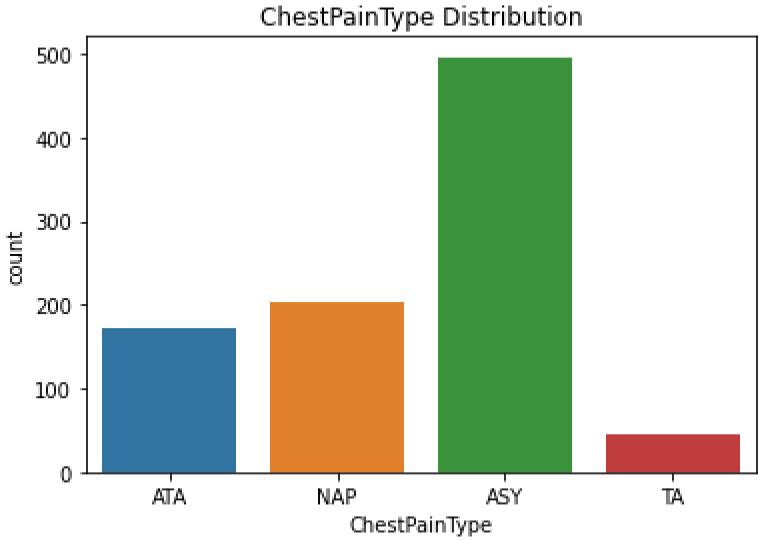


Figure A10

ChestPain Type Distribution



Appendix B

Links

ORIE 5741 Project Github:
https://github.com/Peilin0310/ORIE5741-Project.git

Contribution

Team Members	Contributions
Peilin Li	Code, Results, References
Guohan Gao	Introduction, Problem Statement, Dataset, Discussion
Jingwen Liang	Dataset, Methodology, References