Investigating Heart Disease Diagnosis Correlations Using Exploratory Data Analysis

Colette Lund

School of Mathematical and Natural Sciences, Arizona State University

Introduction

Heart disease is the leading cause of death for men and women in the United States (CDC). Early diagnosis and prevention play critical roles in reducing the negative effects of heart disease. Combining the growing availability of medical datasets with the powerful analytical tools within the field of data science offers insights that can improve disease understanding and diagnosis. The Heart Disease Databases, hosted on the UCI Machine Learning Repository, is a resource ready for use in data science. Containing detailed patient records, it serves as an invaluable dataset for exploring the relationships between clinical parameters (such as age, cholesterol, and blood pressure) and heart disease diagnoses.

This project will perform an investigation into the statistical relationships between patient variables and heart disease diagnosis using exploratory data analysis (EDA) and statistical tests. The analysis of the database will aim to find correlations between multiple patient variables, such as age, cholesterol levels, and blood pressure, and the presence of a heart disease diagnosis. Using the results of the analysis to create a discussion on the significance of certain clinical parameters and heart disease. By focusing on these methods, this study seeks to provide interpretable insights into the dataset and highlight potential predictive signs for heart disease.

Related Works

Foundational Work by Detrano et al. (1989)

The study "International application of a new probability algorithm for the diagnosis of coronary artery disease" by Detrano et al. (1989) represents one of the earliest efforts to apply statistical models to the Cleveland Heart Disease Database. The researchers developed algorithms, particularly logistic regression models, to predict the presence of coronary artery disease based on patient features such as age, cholesterol levels, chest pain type, and exercise-induced angina. Their work demonstrated the potential of using structured patient data to create reliable predictive models for coronary artery disease, achieving substantial accuracy with simple, interpretable methods. This foundational study not only validated the Cleveland dataset as a valuable resource for research but also set a precedent for subsequent investigations into feature correlations and predictive modeling.

Machine Learning Approaches by Alizadehsani et al. (2019)

In "Machine learning-based coronary artery disease diagnosis: A comprehensive review," Alizadehsani et al. (2019) reviewed the application of machine learning models to diagnose coronary artery disease, with a focus on datasets like the Cleveland Heart Disease Database. Their work emphasized the superior performance of machine learning models, such as support vector machines (SVMs), decision trees, random forests, and neural networks, in terms of predictive accuracy compared to traditional statistical methods. By leveraging these models, researchers have achieved diagnostic accuracy rates exceeding 90%. However, the study also highlighted the challenges of model interpretability, which is often critical in clinical settings. Alizadehsani et al.'s work underscores the importance of striking a balance between accuracy and explainability, a challenge that this project addresses by focusing on interpretable statistical methods rather than complex black-box algorithms.

Correlation Analysis by Amal et al. (2020)

The study "Inter-correlation of risk factors among heart patients" by Amal et al. (2020) explored the relationships among various clinical risk factors, including age, weight, sex, diabetes, and blood pressure, in patients with heart disease. Their analysis identified significant inter-correlations, such as the relationship between age and blood pressure or the increased prevalence of diabetes among older patients. By using correlation coefficients and statistical significance testing, Amal et al. demonstrated the value of examining these relationships to enhance our understanding of heart disease risk factors. While their study provided valuable insights into the interconnectedness of these variables, it was limited in scope as it did not include lifestyle factors such as smoking or family history.

Proposed Methodology

To achieve the goal of identifying correlations, the project will follow an approach that combines exploratory data analysis (EDA) and statistical testing in Python. This project uses the heart disease datasets located in UC Irvine Machine Learning Repository. There are four databases within this repository, data from Cleveland, Hungary, Switzerland, and VA Long Beach. Each of these datasets have information about each patient, this includes their age, sex, chest pain type, resting blood pressure (mm Hg), cholesterol(mg/dl), fasting blood sugar, resting electrocardiographic results, maximum heart rate, exercise induced angina, and the diagnosis of heart disease. There were four other variables that were apart of these databases, but these other categories had significant amounts of missing data and were later cast away during the cleaning process.

The project uses the Python libraries pandas and os to load the data into python and data manipulation. Combining the multiple datasets located in UCI will include a process of cleaning each dataset and identifying missing values as well as fixing each dataset together. The exploratory data analysis will use Python libraries matplotlib, numpy, seaborn and sklearn to build distributions, scatterplots, boxplots and correlation matrixes which will highlight relationships between variables. Summary statistics (mean, median, standard deviation) for each variable will provide insight and an overall understanding of the dataset. To perform a statistical analysis of the dataset, pearson correlation coefficients will be calculated to measure correlation between variables, using the Python library scipy. Finally, performing a logistic regression analysis will help to understand how certain variables impact the probability of the presence of heart disease.

Experiment Discussion

Experiment Setup

The setup of this project involves multiple stages, including data preparation, exploratory data analysis (EDA), and the application of statistical and machine learning techniques. To start the project will pull from 4 datasets in the UCI Machine Learning Repository, data from Cleveland, Hungary, Switzerland, and VA Long Beach. Since the datasets are already formatted, the data can be loaded into Python using pandas library and were then combined to create one large dataset with over 900 rows of data. Then the issue of missing data points was identified and addressed. Multiple categorical columns were missing significant amounts of data and thus were dropped from the datasets. Then rows with excessive missing values were removed as well, resulting in one clean data set with 744 rows of data. In order for easier manipulation in later modeling, a second dataset was created where each of the values were normalized to ensure a consistent scale.

Table 1: Description of Variables in dataset

Variable Name	Description	Units
age	The age of the patient.	Years
sex	The sex of the patient.	1 = male; 0 = female
cp	The type of chest pain in a patient.	Ranges from 1-4 depending on type of pain
trestbps	Resting Blood Pressure.	Mm Hg
chol	Measure of total amount of cholesterol in a patient's blood.	Mg/dl
fbs	Fasting blood sugar of a patient.	> 120 mg/dl (1 = true; 0=false)
restecg	Resting electrocardiographic results of a patient.	0 = normal; 1=presence of abnormality
thalach	Maximum heart rate achieved in patient.	BPM
exang	Exercise induced angina.	1=yes; 0=no
target	Diagnosis of heart disease in patient.	0=no (<50% diameter narrowing) 1=yes (>50% diameter narrowing)

After these datasets were made in python, Pearson correlation tests were conducted to evaluate the statistical significance of relationships between variables and the diagnosis of heart disease. Relationships with statistically significant correlations, p-value < 0.05, were flagged for further investigation and the Pearson coefficient was noted for the strength of these relationships. The variables which showed a dependent relationship with the diagnosis of heart disease were then further analyzed using scatterplots, boxplots and distributions. Scatterplots helped to explore relationships between variables and their association with heart disease diagnosis. Boxplots were used to examine the distribution of data across categories. The distribution plots of key features were used to identifiers and assess normality.

Finally, K-Means Clustering was applied to the variables that were found to have the strongest correlation with a heart disease diagnosis. Providing insight into the natural groupings within the dataset. Then SVM was used to classify a patient's heart disease diagnosis based on their clinical parameters. Each of these tools are used to analyze the data and result with significant conclusions made regarding the data and the significance of variables in a patient and their diagnosis of heart disease.

Experiment Results

The question for this research project is which variables of a patient show a statistically significant correlation with a heart disease diagnosis. The hypothesis was that the following variables: cholesterol levels, age, chest pain, and blood pressure, will show a statistically significant association with a heart disease diagnosis.

After the Pearson correlation test was performed the results showed that each of the variable's test had a p-value < 0.05, thus each variable has a significant relationship with the diagnosis of heart disease. Next the correlation coefficient showed the strength of these relationships. The strongest of these is the relationship between exercise induced angina and the diagnosis of heart disease as well as the relationship between chest pain and the diagnosis of heart disease. The relationships of age and maximum heart rate and the diagnosis of heart disease are also moderately strong, thus this project will further analyze these correlations.

Table 2: Results of Pearson Correlation Test

Variable	Correlation Coefficient	P-value
age	0.35452	0.0
sex	0.25011	0.0
cp	0.39668	0.0
trestbps	0.15628	2e-05
chol	-0.19775	0.0
fbs	0.13086	0.00035
restecg	0.16588	1e-05
thalach	-0.35759	0.0
exang	0.4264	0.0

Figure 1: Correlation Matrix between Variables

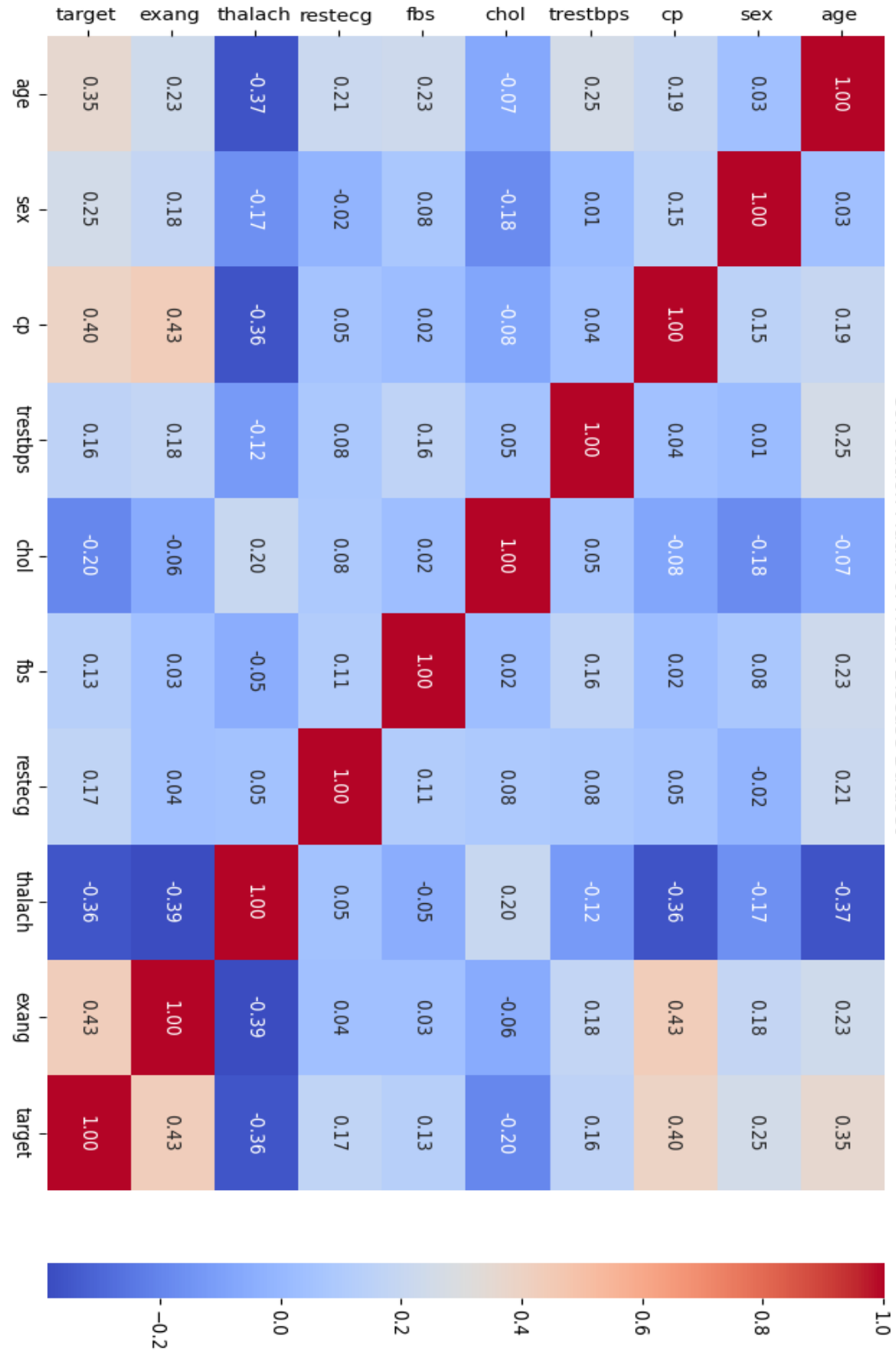
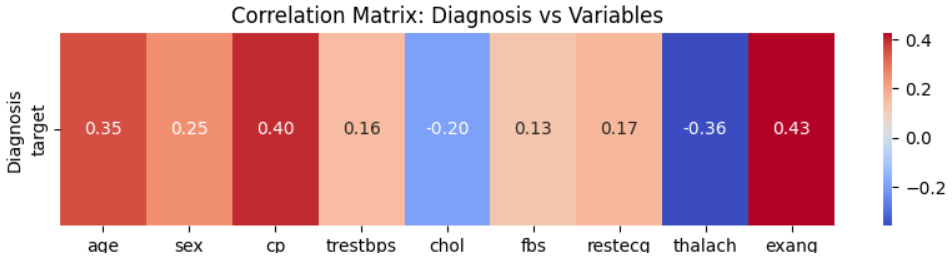


Figure 2: Correlation Matrix between Variables and Diagnosis



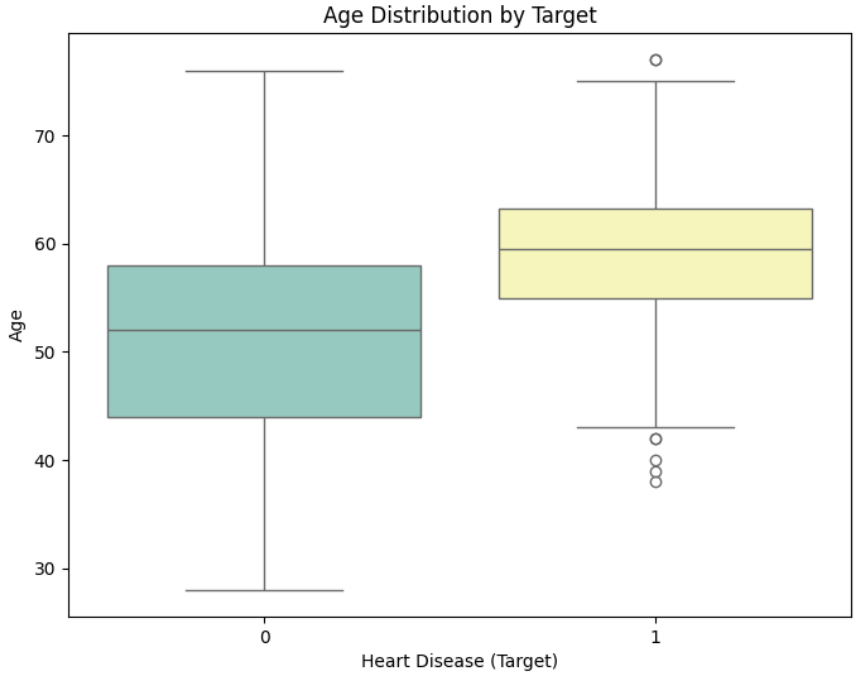


Figure 3: Boxplots of Age and Diagnosis

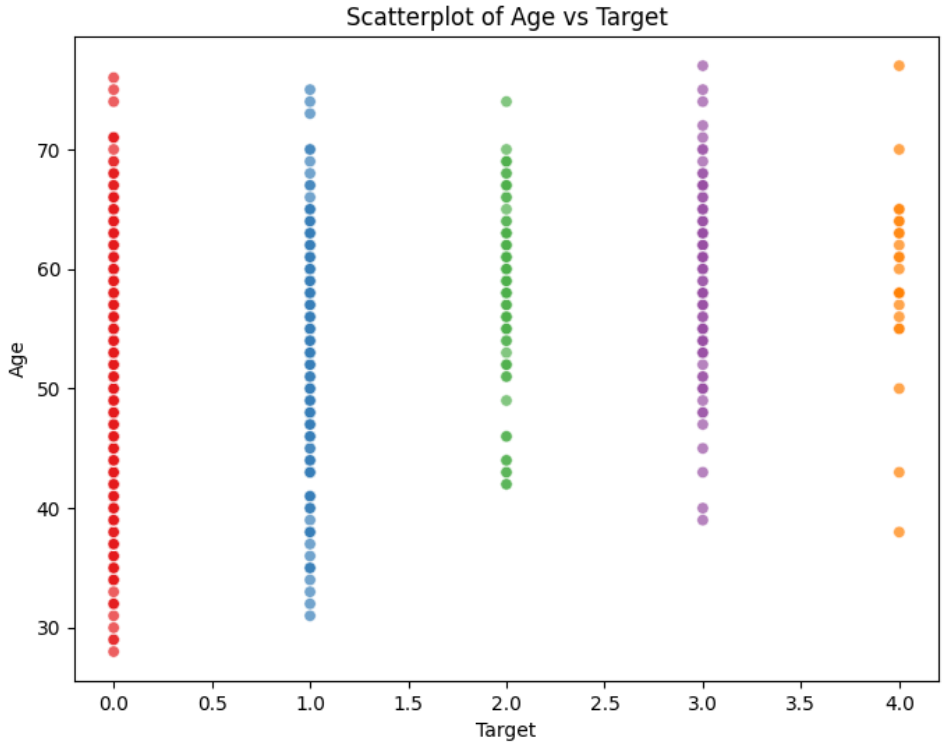


Figure 4: Scatterplot of Age by Diagnosis

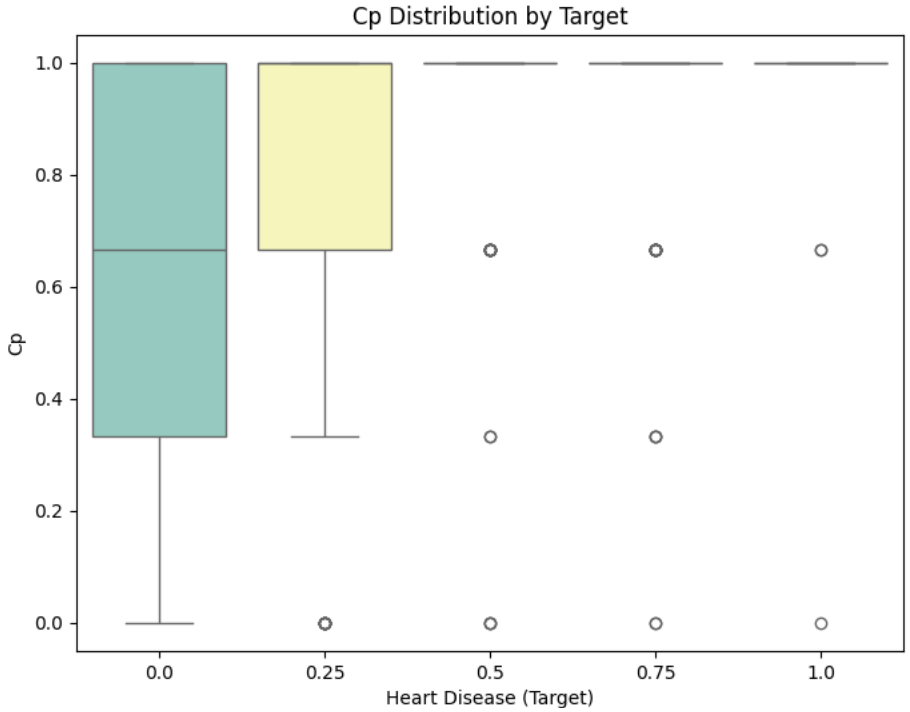


Figure 5: Boxplots of Chest pain and Diagnosis

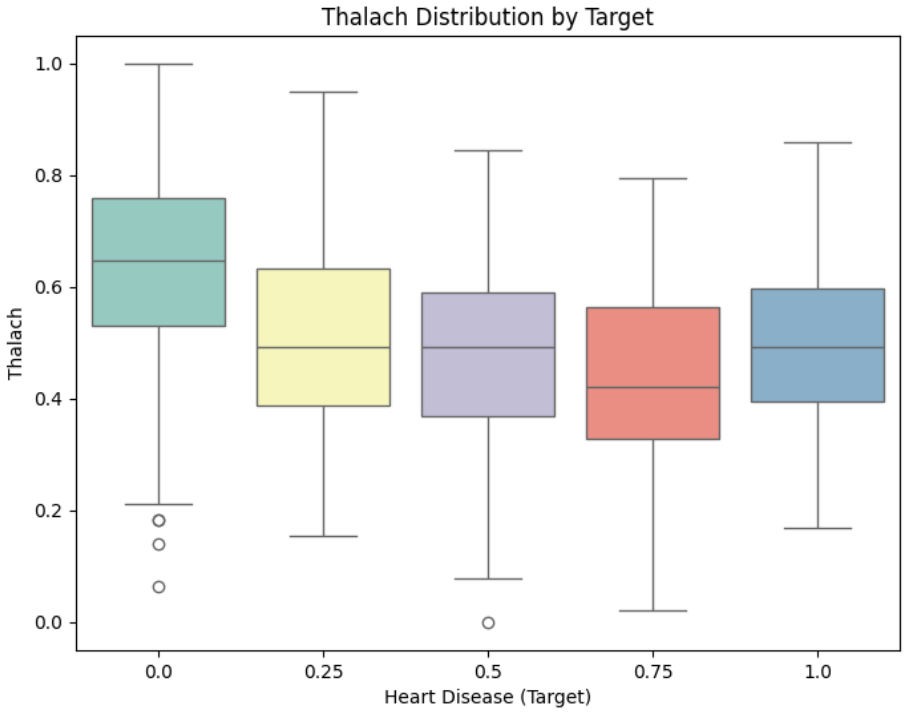


Figure 6: Boxplots of Maximum Heart Rate and Diagnosis

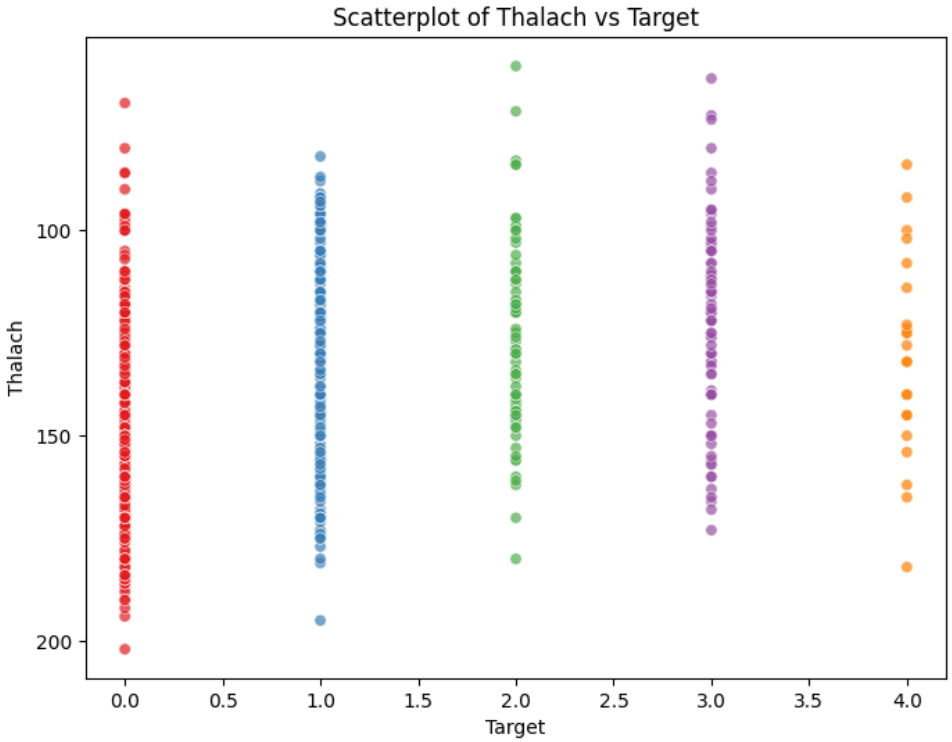


Figure 7: Scatterplot of Maximum Heart Rate by Diagnosis

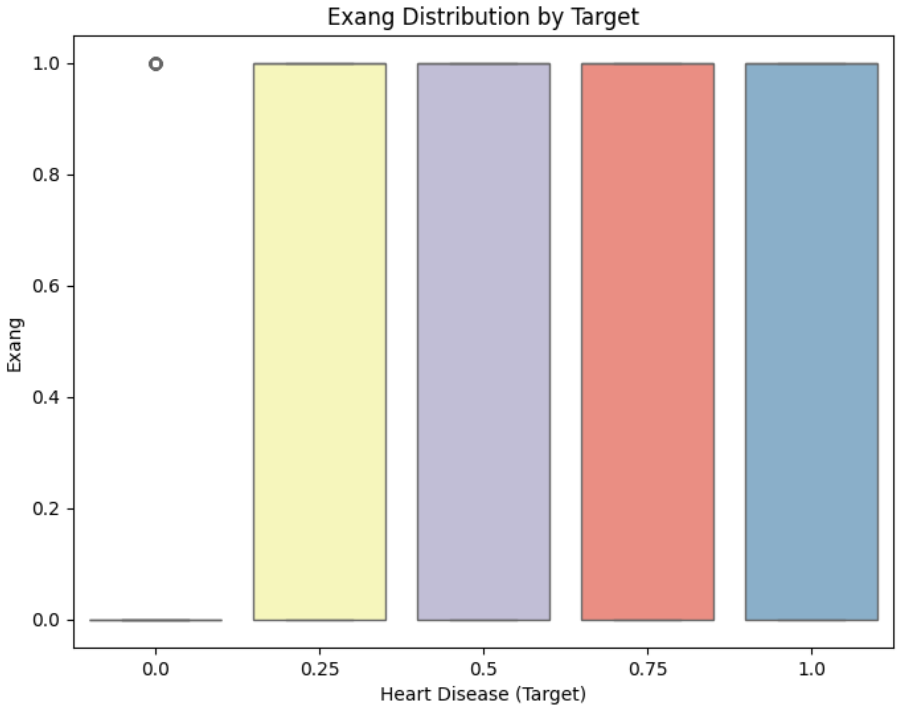


Figure 8: Boxplots of Exercise Induced Angina by Diagnosis

The correlation matrix, visualized using a heatmap, provided an overview of all variable relationships. This visualization confirmed that the strongest correlations with heart disease diagnosis corresponded to age, chest pain type, maximum heart rate, and exercise-induced angina, consistent with the statistical findings. To further investigate these relationships, scatterplots and boxplots were generated.

A notable result from the plots was the boxplot of age and the diagnosis of heart disease showed us that the presence of heart disease was in patients older than 40 years old with an average of 60 years of age, which is significant because the data had ages from 30 years to 80 years of age, this result showed that there is a higher probability to be diagnosed with heart disease at an older age. The boxplot of chest pain type against the diagnosis of heart disease shows that when there is a diagnosis of heart disease (≥ 0.5) patients on average presented with the highest level of chest pain other than some outliers. The scatterplot and boxplot of maximum heart rate against the diagnosis of heart disease indicated that there is some correlation between the variables. When the diagnosis of heart disease was present (≥ 0.5) patients had a lower maximum heart rate on average.

The results of the correlation analysis support the hypothesis that both demographic factors and clinical features are critical in predicting heart disease. While some variables, such as

cholesterol and resting blood pressure, were found to have weaker correlations with heart disease, their inclusion in the overall model remains justified due to their statistically significant relationships.

For exploring the relationships among the most influential features—age, chest pain type, maximum heart rate, and exercise induced angina—K-Means clustering was applied to group patients based on these variables. Additionally, a Support Vector Machine (SVM) was utilized to define a decision boundary for heart disease classification. When K-Means clustering was applied to these categories two significant clustered are identified. Patients in the first cluster tend to have no heart disease diagnosis and the patients within the second cluster tend to be diagnosed with heart disease.

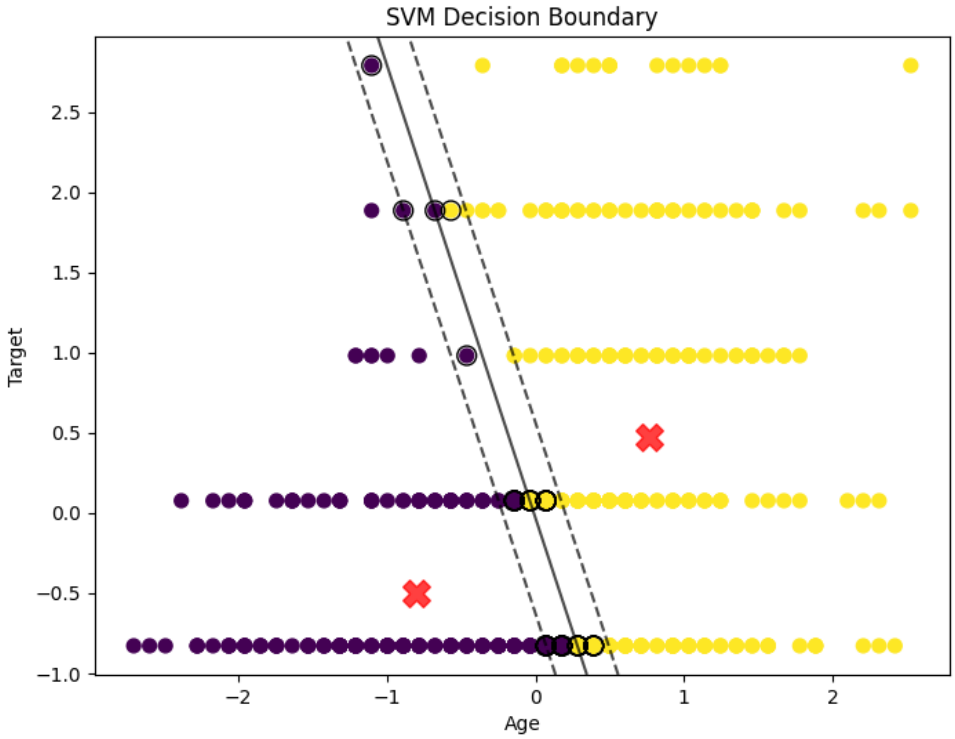


Figure 9: K-Means Clustering and SVM Decision Boundary with Age vs Diagnosis

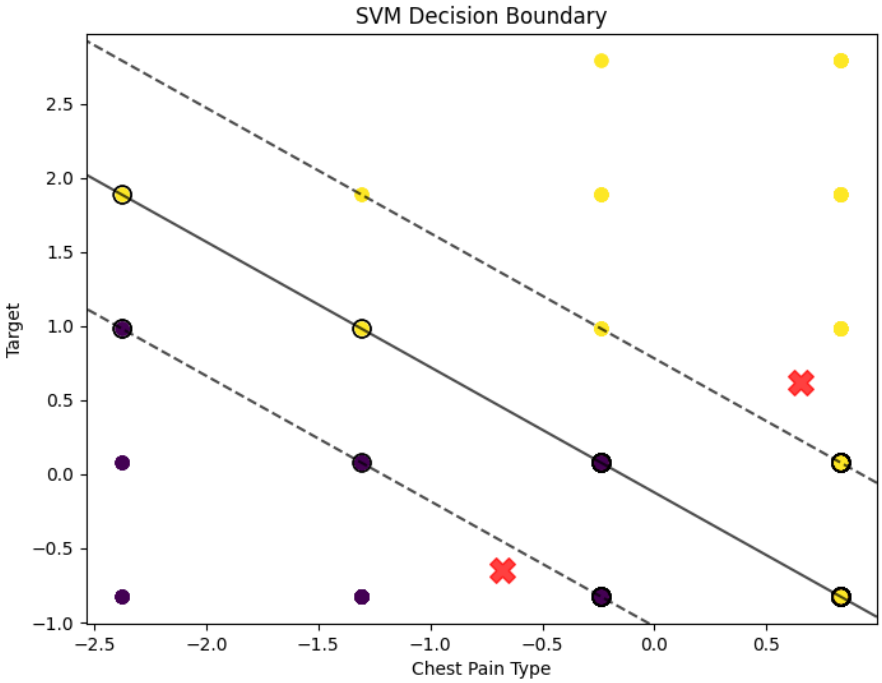


Figure 10: K-Means Clustering and SVM Decision Boundary with Chest Pain Type vs Diagnosis

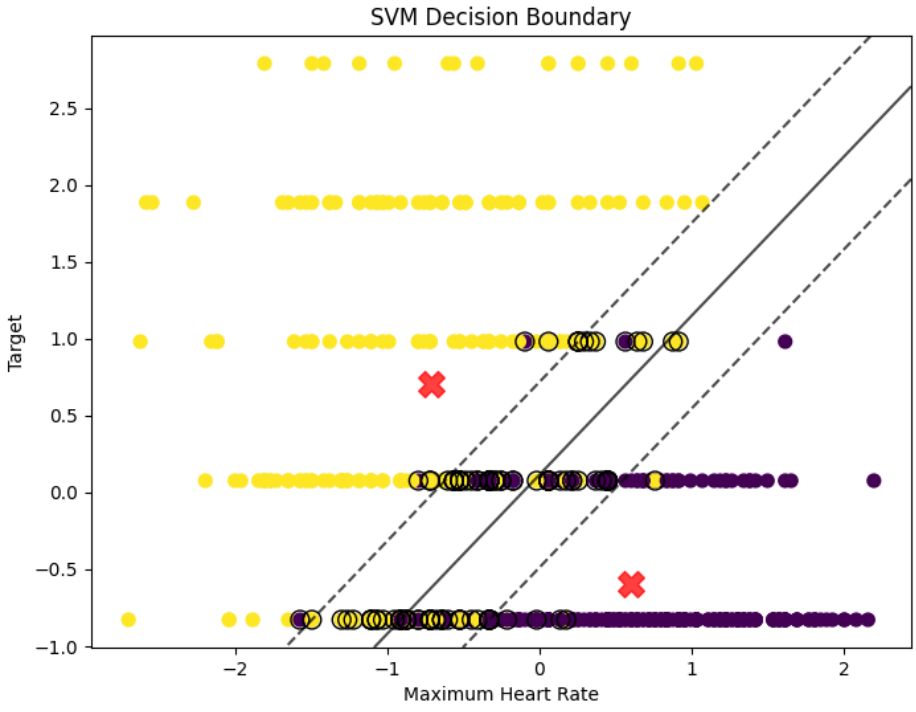


Figure 11: K-Means Clustering and SVM Decision Boundary with Maximum Heart Rate vs Diagnosis

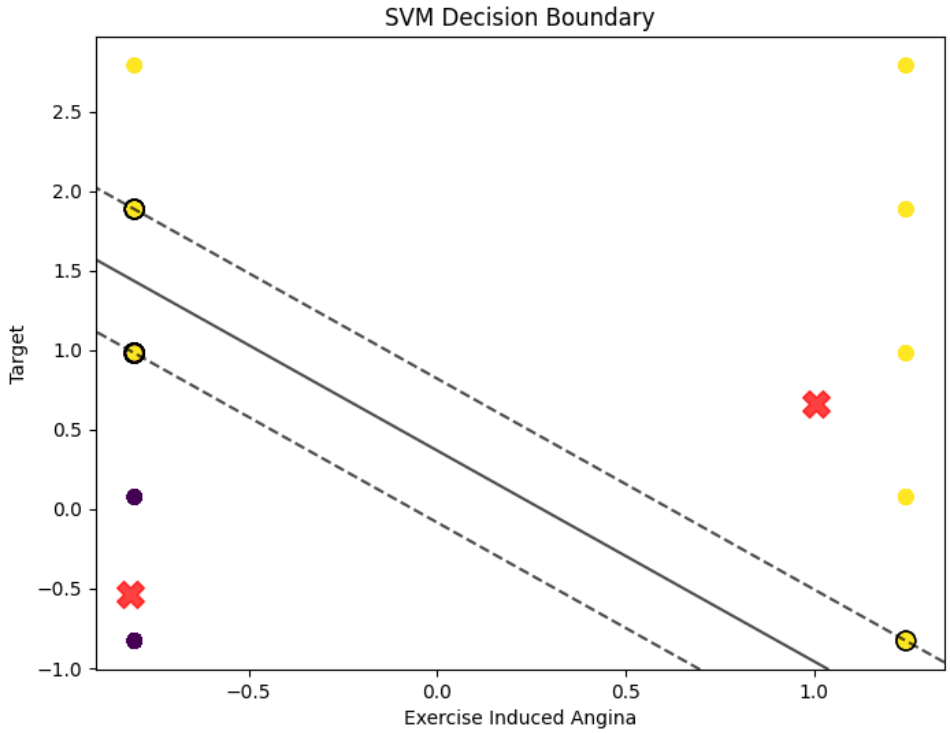


Figure 12: K-Means Clustering and SVM Decision Boundary with Exercised Induced Angina vs Diagnosis

To classify patients into heart disease and non-heart disease groups, a support vector machine (SVM) was trained using the four features with the strongest correlations. The SVM's decision boundary was overlaid on the clustered data to visually represent its classification capability. The SVM boundary effectively separated the cluster associated with higher heart disease prevalence from the cluster with lower prevalence. The SVM model achieved high classification accuracy on all four features and their clustering. Thus, indicating that the four features are not only strongly correlated with the diagnosis of heart disease but also provide the model with strong enough correlations to allow the SVM model to accurately classify patients as having heart disease or not.

Scatterplots with the K-Means clusters and the SVM boundary illustrated the relationship between patient groupings and the model's classification decisions. Patients near the SVM boundary showed mixed predictions, consistent with the moderate correlations observed in our statistical testing.

The combination of K-Means clustering and SVM classification highlights the value of these four features—age, chest pain type, maximum heart rate, and exercise-induced angina—in both understanding patient groupings and developing predictive models. Clustering provided a deeper understanding of the underlying data structure, while the SVM boundary validated the

ability of these features to make accurate predictions. These results suggest that even simple models trained on carefully selected variables can achieve high interpretability and strong performance, making them useful for real-world diagnostic applications.

Conclusion

The primary hypothesis of this study was that age, cholesterol levels, type of chest pain, and resting blood pressure would exhibit significant correlations with heart disease diagnosis and could be used to effectively differentiate between patients with and without heart disease. The analysis supported this hypothesis, with statistical tests confirming that all variables were dependent on the heart disease diagnosis. The statistical tests also showed that the strongest correlations were between age, chest pain type, maximum heart rate, and exercise-induced angina and the patient's diagnosis of heart disease. These four features showed correlation coefficients greater than 0.3, indicating moderate to strong associations. These results highlight the importance of these features in the diagnostic process, validating their relevance in both clinical and research settings.

The exploratory data analysis, including the creation of a correlation matrix and visualizations such as scatterplots and boxplots, reinforced the strength of these variables as predictors of heart disease. The application of K-Means clustering revealed two distinct patient groupings, with one cluster demonstrating a higher likelihood of heart disease due to lower maximum heart rates, higher levels of chest pain, older age, and prevalent exercise-induced angina. The SVM model, trained on the most influential features, further validated the effectiveness of these variables for heart disease classification, achieving high accuracy and creating a clear decision boundary that effectively distinguished between heart disease and non-heart disease patients. These findings not only align with prior research but also provide additional evidence that moderate correlations can be leveraged effectively in clinical and predictive contexts. By focusing on the most influential variables, future machine learning models can be optimized for both performance and interpretability.

These findings contribute to the field of medical data analysis by providing a statistical and machine learning-based approach to identifying key risk factors and developing models that are both interpretable and accurate. The ability to highlight influential features and segment patients into clusters can aid healthcare professionals in assessing risk and making informed decisions about further testing or interventions. This work supports a more targeted approach to diagnostics, potentially improving patient outcomes and optimizing healthcare resources. Future studies could extend this research by incorporating additional datasets or exploring the impact of lifestyle and genetic data to refine predictions. Advanced techniques such as ensemble methods or deep learning models could also be examined for potential improvements in accuracy and robustness. By continuing to explore and validate the relationships between clinical features and heart disease, future research can contribute to more personalized and effective healthcare strategies.

Acknowledgements

This study was made possible by the Cleveland Heart Disease Database, hosted on the UCI Machine Learning Repository. We extend our gratitude to the Cleveland Clinic Foundation for compiling and sharing this valuable dataset, which has facilitated numerous advancements in cardiovascular research.

We would also like to thank the UCI Machine Learning Repository for providing open access to high-quality datasets that support academic and research efforts across the globe.

Additionally, we acknowledge the developers and contributors of Python libraries such as pandas, numpy, matplotlib, os, seaborn, sklearn and scipy, whose tools played a vital role in conducting the analyses in this study.

Finally, we thank our instructors and peers in the Mathematical Methods in Data Science course for their guidance that shaped this project.

Author Contributions

The author solely conceptualized and designed the study, conducted all analyses, interpreted the results, and drafted the manuscript. The author also prepared and executed the data preprocessing, performed the statistical testing, and visualized the results using Python. All aspects of the project, including methodology development, literature review, and manuscript writing, were completed independently by the author.

Data Availability

The Heart Disease Databases are publicly available through the UCI Machine Learning Repository, one of the most widely used platforms for hosting datasets in the field of data science and machine learning. The dataset was originally compiled at the Cleveland Clinic Foundation and is specifically intended for educational and research purposes.

The dataset comprises over 900 records with 14 attributes, including 13 clinical features and one target variable (num) that indicates the diagnosis of heart disease. The clinical features include age and sex of the patient; results from medical tests, such as cholesterol levels, maximum heart rate achieved, and fasting blood sugar levels; and symptoms or diagnostic measures, such as chest pain type and exercise-induced angina.

The dataset is freely accessible online at the following link: Cleveland Heart Disease Database - UCI Machine Learning Repository. The Cleveland Heart Disease Database has been anonymized to protect patient privacy, ensuring compliance with ethical standards for medical data usage. While the dataset is comprehensive, there are some limitations. Certain features of the databases contain missing data points, which may require imputation or exclusion during preprocessing. A relatively small sample size of patients, which may limit generalizability to larger populations. Data collection methods and population demographics from the Cleveland Clinic Foundation may not fully represent global populations, necessitating caution when extrapolating results. This open access to the Heart Disease Databases enables researchers worldwide to contribute to the study of heart disease, making it a vital resource for advancing cardiovascular health research.

References

Alam, M. J., Alnafeesah, A. I., & Saeed, M. (2020). Inter-correlation of risk factors among heart patients. *AIMS public health*, *7*(2), 354–362. https://doi.org/10.3934/publichealth.2020030

Alizadehsani, R., Abdar, M., Roshanzamir, M., Khosravi, A., Kebria, P. M., Khozeimeh, F., Nahavandi, S., Sarrafzadegan, N., & Acharya, U. R. (2019). Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Computers in biology and medicine*, *111*, 103346. https://doi.org/10.1016/j.compbiomed.2019.103346

Center for Disease Control and Prevention. (June 2023). *Heart disease prevalence - health, United States*. https://www.cdc.gov/nchs/hus/topics/heart-disease-prevalence.htm

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H. Guppy, Stella Lee & Victor Froelicher. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease.

The American Journal of Cardiology, *Volume 64* (Issue 5, 1989), pages 304-310, ISSN 0002-9149. https://doi.org/10.1016/0002-9149(89)90524-9.

Van Rossum, G., & Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waskom, M., Botvinnik, Olga, O'Kane, Drew, Hobson, Paul, Lukauskas, Saulius, Gemperline, David C, … Qalieh, Adel. (2017). *mwaskom/seaborn: v0.8.1 (September 2017)*. Zenodo. https://doi.org/10.5281/zenodo.883859