Heart Disease Data Science Mini-Project

Introduction

This project explores some classic data science techniques on the Cleveland Heart Disease dataset. My goal was to better understand how common health risk factors, such as blood pressure, cholesterol, blood sugar, and exercise tolerance might relate to heart disease, and to practice the analysis methods I have been learning in online courses.

Data Cleaning and Exploration

First, I cleaned the data to deal with missing values and make sure everything was in the right format. Then, I looked at the main variables and their distributions. Most patients in the dataset are middle-aged or older, and there is a mix of people with and without heart disease.

I checked the relationships between risk factors using a correlation matrix. Surprisingly, most variables were not strongly correlated with each other, except for maximum heart rate (MaxHR), which was moderately lower in patients with heart disease. This matches what I have read about how lower exercise tolerance can be a warning sign for cardiac problems.

Visualization

To dig deeper, I used boxplots, violin plots, and pairplots to compare risk factors across heart disease categories. Again, MaxHR stood out as the most different between groups, while blood pressure, cholesterol, and fasting blood sugar overlapped quite a bit. The visualisations really helped me spot these patterns.

Predictive Modelling

I tried building a simple logistic regression model to predict heart disease using just these four features. After standardising the data and splitting it into train and test sets, the model reached about 77% accuracy with a decent AUC score. MaxHR turned out to be the most influential variable in the model. This was satisfying, as it matched my earlier findings and made sense clinically.

Unsupervised Clustering

To see if there were any “hidden” patient groups, I used K-means clustering with different numbers of clusters. The clusters mostly reflected small differences in risk factors, but didn’t line up well with actual heart disease status. I learned that just looking for natural groupings is not enough when the outcome is complex and affected by many factors.

I selected MaxHR and RestingBP for K-means clustering because they are fundamental clinical indicators in the assessment of heart disease. However, the resulting clusters showed that patients were widely scattered across the plot, without clear separation. This outcome is consistent with clinical understanding: while heart rate and blood pressure are often directly related in situations such as stress or anxiety, they can also be inversely related in cases of dehydration or advanced disease decompensation. As a result, relying solely on these two variables does not capture the full complexity of cardiovascular risk. A clearer distinction may require either more comprehensive datasets or multidimensional analysis that incorporates additional clinical features.

Reflection and Next Steps

This project was my first time putting all these data science steps together on a real healthcare dataset. I learned a lot about cleaning data, visualising patterns, and building simple models. It was also helpful to see the limits of what you can do with a small number of features and basic methods.

I am looking forward to learning more advanced modelling techniques, working with bigger and richer datasets, and understanding more about the medical side as well as the math. I know there is a lot I have not covered yet, but this project made me even more interested in health data science and more confident about applying for the MSc.