Heart Failure Prediction Analysis

Students:
Ishan Ojha - 216474868
Gurvir Boparai - 217797457

Course:
ITEC 4305 M

Professor:
Soroush Sheikh Gargar

April 27, 2024

Contents

1 Introduction

Heart failure is a growing concern, impacting a vast number of individuals. Early detection is crucial for effective management and improved patient outcomes. This project delves into the application of machine learning to predict heart failure risk based on clinical data. The project employs a data-driven approach, utilizing various techniques to build and evaluate the model. Following data preprocessing to ensure quality, exploratory data analysis is conducted to identify key features. Several machine learning algorithms, including Support Vector Machines (SVM), Random Forest, K-Nearest Neighbors (KNN), and Regression, are explored for their suitability in capturing the complex relationships within the data. Models will be trained and evaluated on its ability to accurately predict heart failure risk for new patients.

1.1 Dataset Description

This dataset, available on Kaggle, contains information on patients and their risk of heart failure. It combines data from five separate sources, resulting in a comprehensive collection of features for heart failure analysis. The list of features available in the dataset are the following:

- **Age** - age of the patient [years]
- **Sex** - sex of the patient [M: Male, F: Female]
- **ChestPainType** - chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- **RestingBP** - resting blood pressure [mm Hg]
- **Cholesterol** - serum cholesterol [mm/dl]
- **FastingBS** - fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]

- **RestingECG** - resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
- **MaxHR** - maximum heart rate achieved [Numeric value between 60 and 202]
- **ExerciseAngina** - exercise-induced angina [Y: Yes, N: No]
- **Oldpeak** - oldpeak = ST [Numeric value measured in depression]
- **ST_Slope** - the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
- **HeartDisease** - output class [1: heart disease, 0: Normal]

1.2 Used Tools

For this assignment several tools have been used:

- **Python** - a general-purpose programming language widely used in machine learning. In particular, the libraries that have been used are the following:
    - **Pandas** - a data manipulation and analysis library used for working with structured data.
    - **Matplotlib** - a plotting library used for creating visualizations and graphs.
    - **Numpy** - a library used for performing mathematical and logical operations on multi-dimensional arrays and matrices.
    - **Seaborn** - a visualization library based on Matplotlib, used for creating attractive and informative statistical graphics.
    - **Scikit-learn** - a machine learning library that provides tools for data preprocessing, classification, regression, clustering, and more.

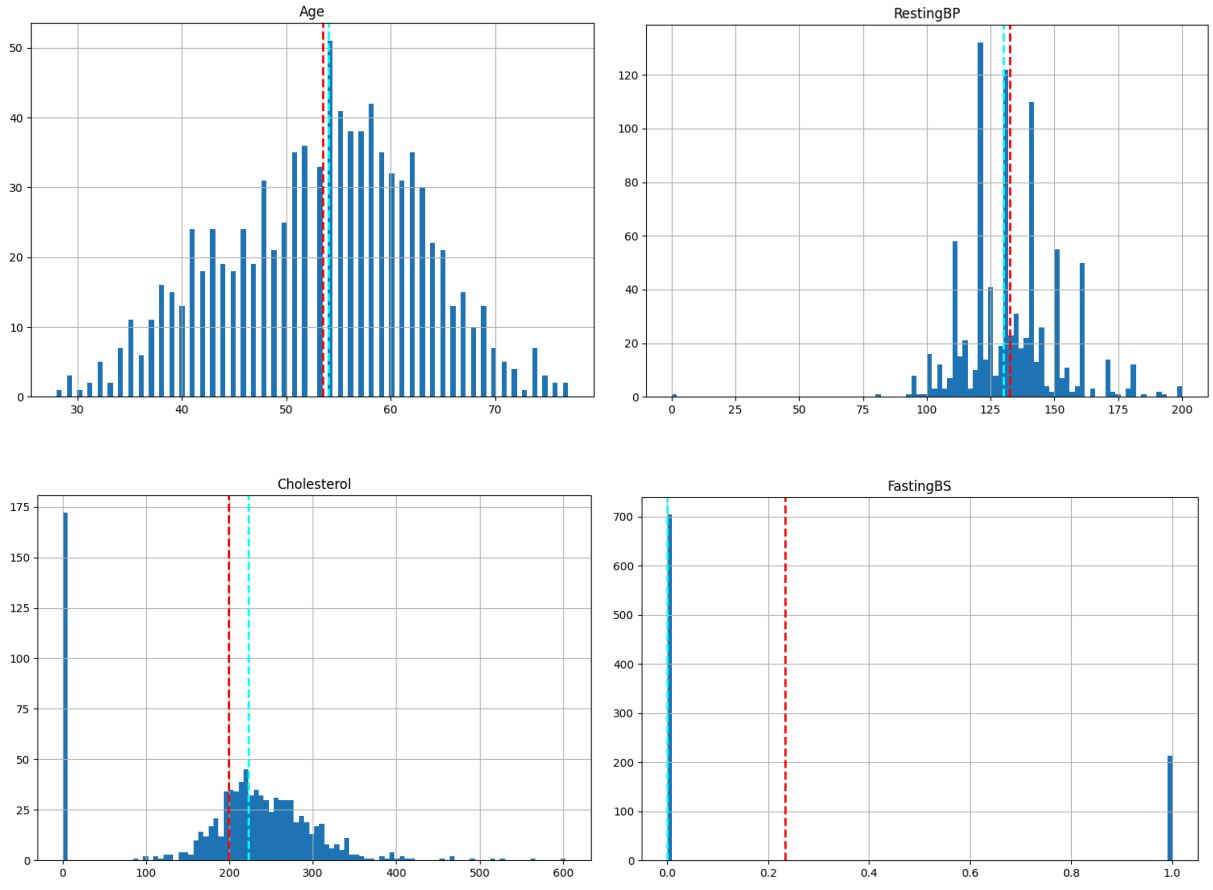
- **Kaggle** - a platform for data scientists and machine learning practitioners. It provides access to datasets, code notebooks, and a community of data science professionals to help participants improve their skills and solve complex problems.
- **Jupyter Notebook** - an open-source web application that allows users to create and share documents containing live code, equations, visualizations, and narrative text.

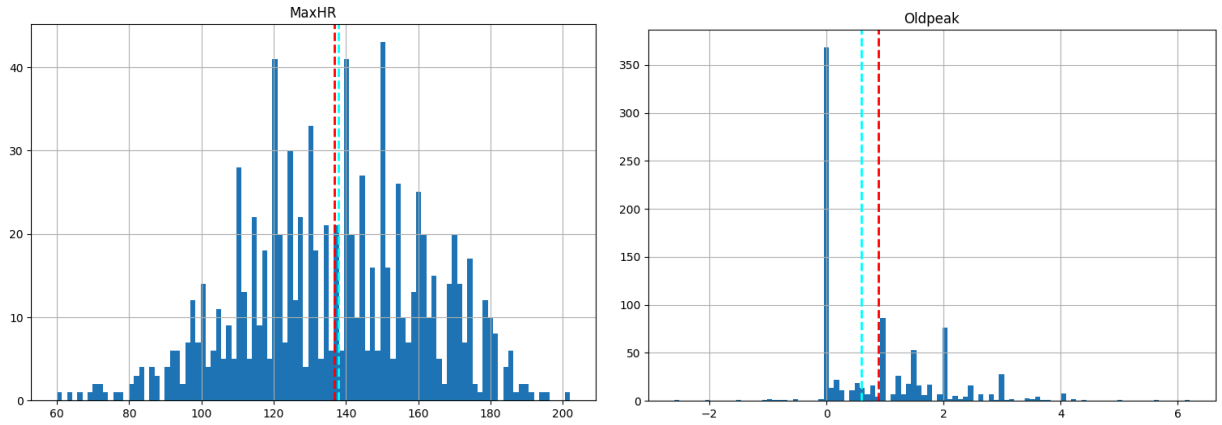
2 Data Exploration

In this phase we go through the process of examining, displaying, and evaluating our dataset in order to find patterns, trends, and connections that may be helpful in constructing predictive models.

2.1 Histograms

Histograms are commonly used to analyze the distribution of numerical data variables. By visualizing the distribution of the data, a histogram can provide insights into the central tendency, variability, and shape of the data. Histograms include identifying outliers, identifying skewed distributions, identifying clusters or subgroups within the data, and identifying the presence of any patterns or trends.

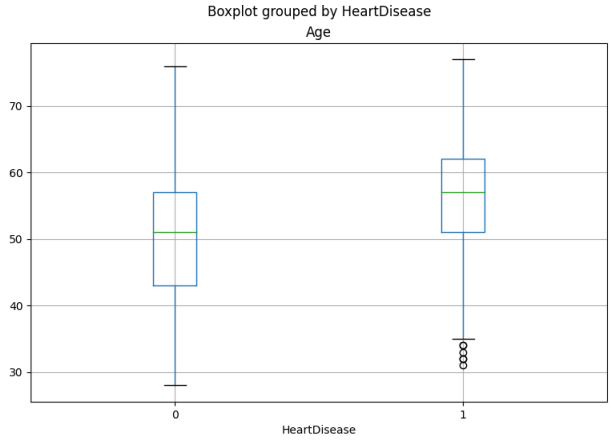


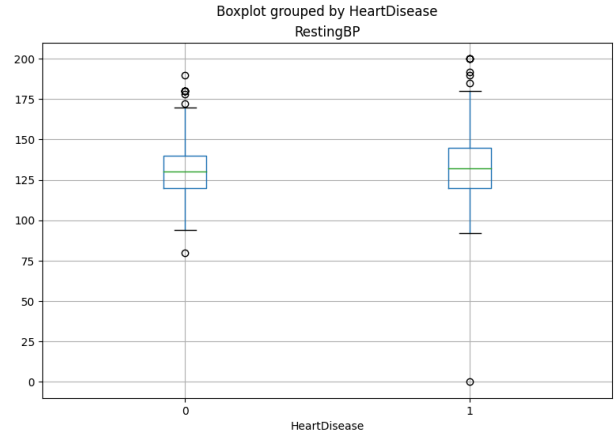


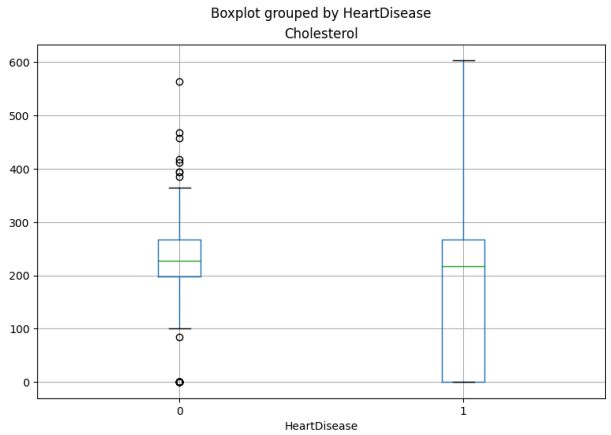
The histograms of age, resting blood pressure, cholesterol, fasting blood sugar, maximum heart rate, and oldpeak values reveal some key insights. The data shows that most individuals are in their 50s and 60s. For resting blood pressure, there is variability in readings, with some individuals having extremely low (~0 mmHg) or high (~200 mmHg) values. Similarly, cholesterol levels range from very low (~0 mg/dL) to very high (~600 mg/dL). Maximum heart rate and oldpeak values also show variability, with some individuals having rates and values outside the typical range. This highlights the importance of further investigation and potential interventions.

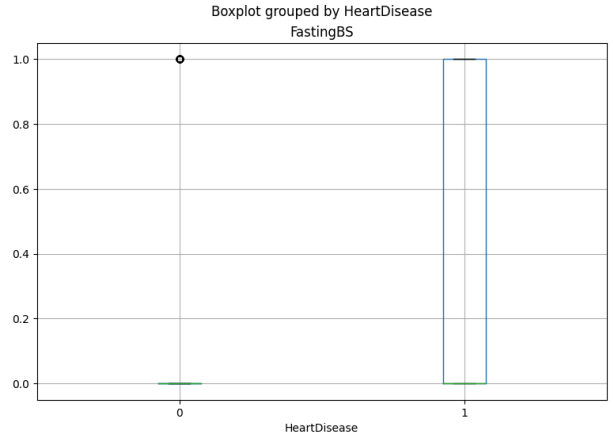
2.2 Box Plots

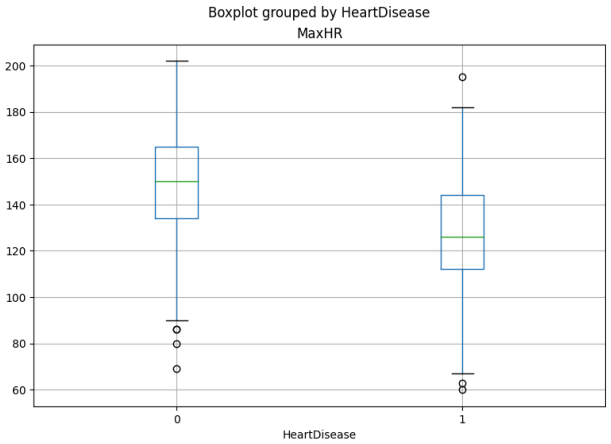
Box plots provide a visual summary of a dataset's distribution. They display the median, quartiles, and outliers of a dataset in a concise and easy-to-understand manner. They help detect potential outliers and understand the spread and central tendency of the data.

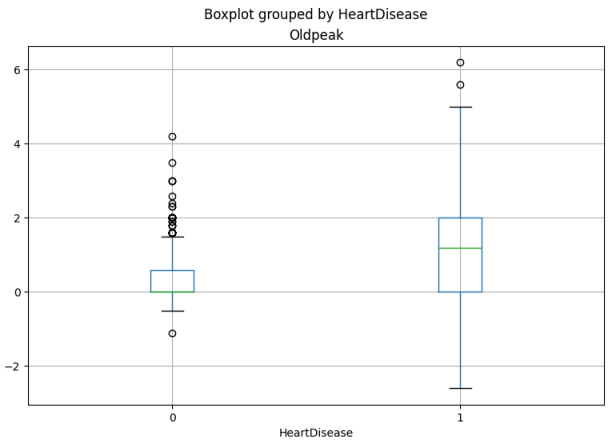








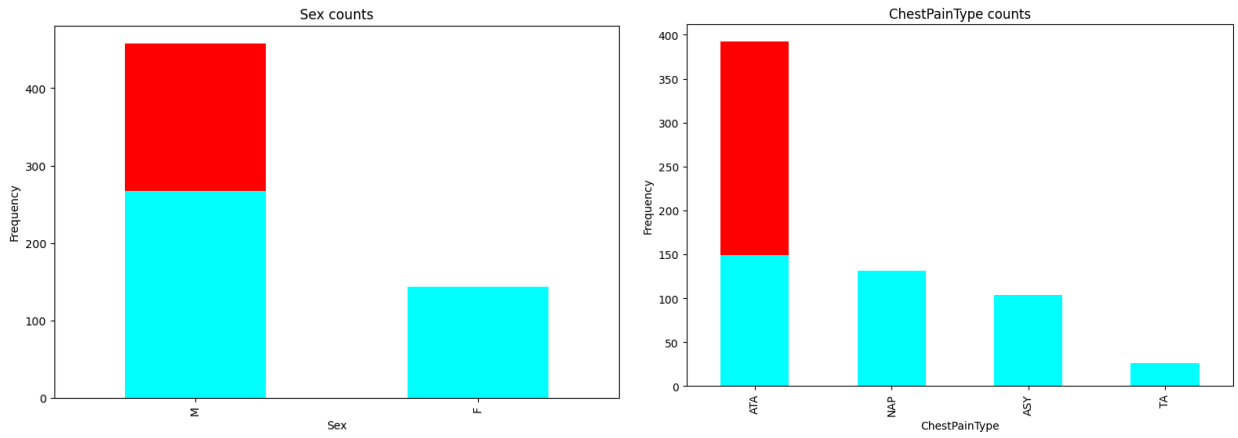


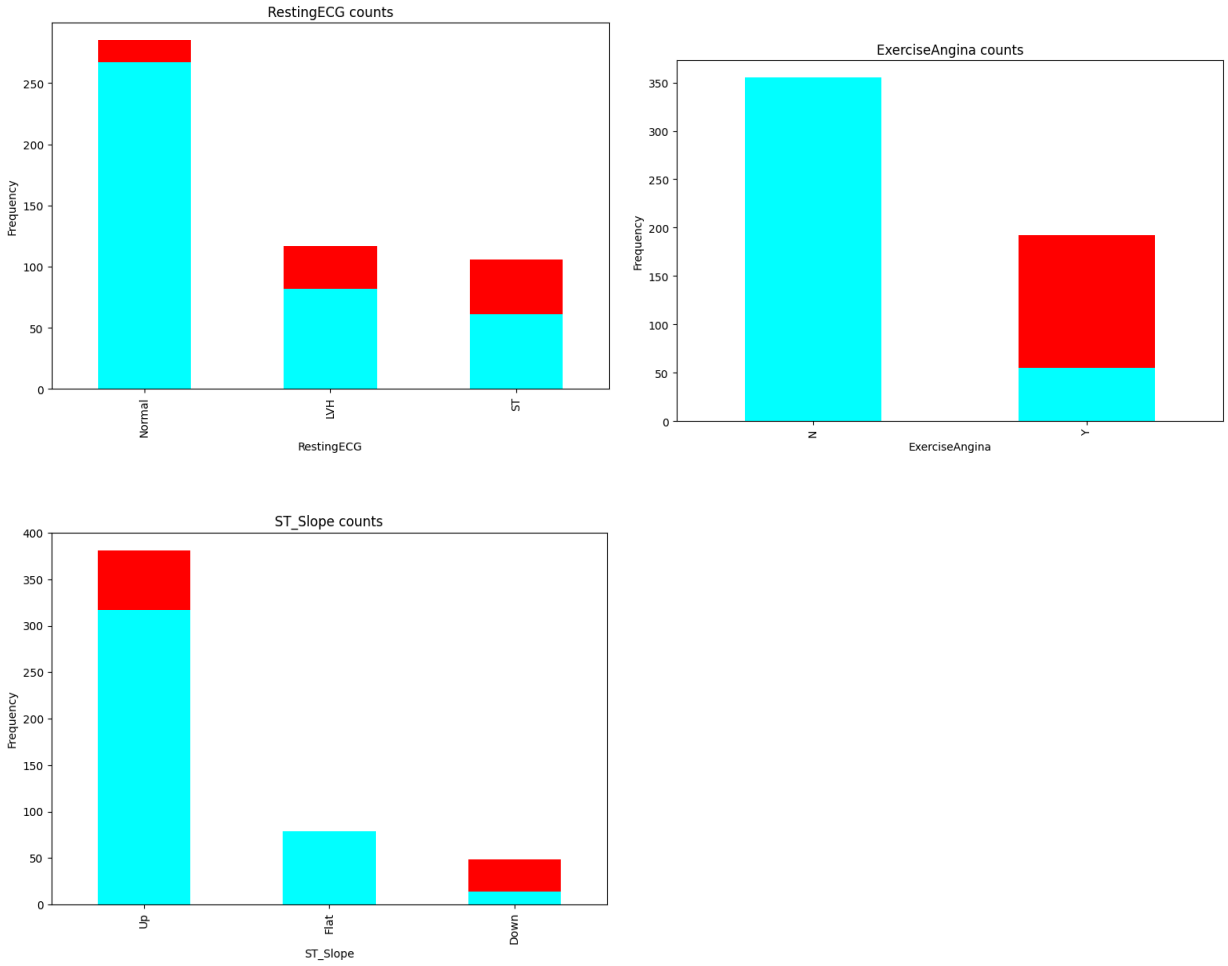


We have grouped our boxplots by heart disease, 0 meaning normal and 1 meaning heart disease. Further delving into these boxplots, it is evident that in the context of normal, the median age is around 50 years with a range of 25 to 70. Resting blood pressure typically hovers around ~125 mmHg and cholesterol levels average at ~220 mg/dL with maximum heart rate of ~165 bpm. In contrast, individuals with heart disease tend to be older, with a median age of around ~57 years. Resting blood pressure rises is about the same, cholesterol range levels increase with maximum heart rate decreasing to around ~125 bpm. Visually, it is obvious to see that there are a lot of outliers for these numerical features which leads us to use IQR detection and removal in our data cleaning step.

2.3 Stacked Bar Charts

Stacked bar charts display multiple variables in a single bar, with each variable represented by a different color within the bar. It helps show the distribution of a categorical variable within different groups or categories. By stacking the bars on top of each other, it provides a clear visual representation of how the different variables contribute to the overall total, making it easier to identify patterns or trends in the data.

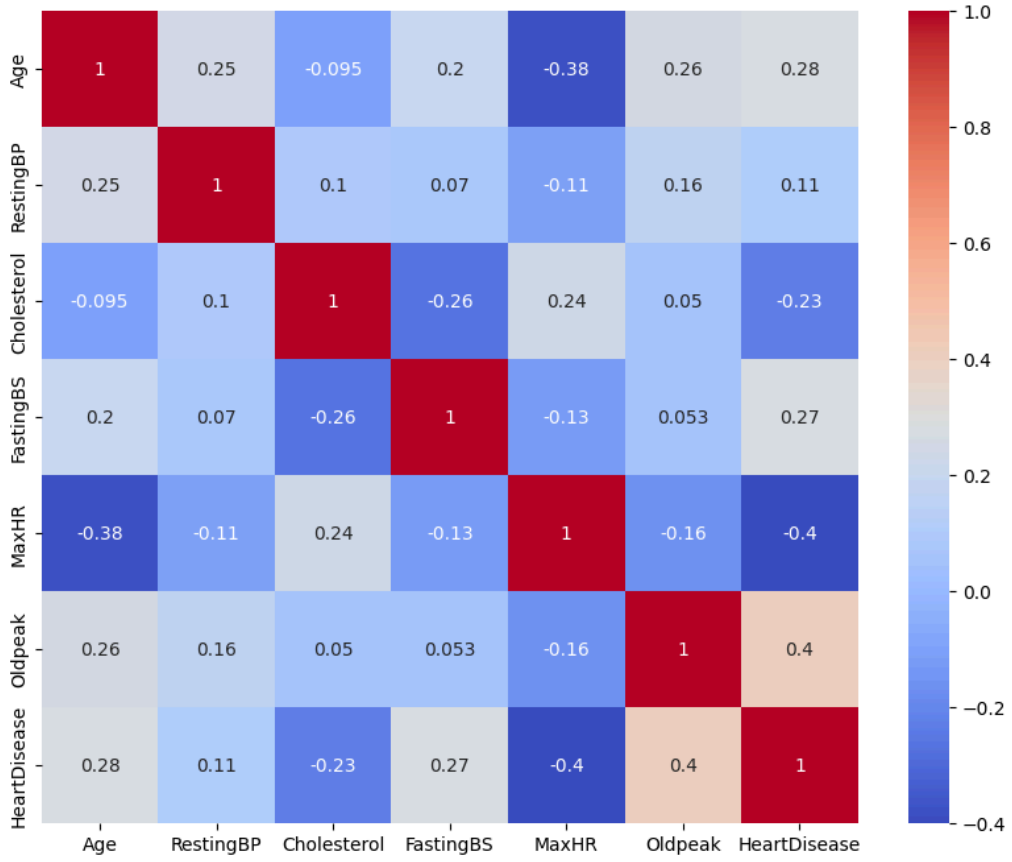




The stacked bar charts help us visualize the categorical variables and their counts. The Sex chart indicates that the number of males with heart disease is higher than females, as is the number of males without heart disease. The Chest Pain Type chart shows that atypical angina (ATA) is the most common type of chest pain among individuals with heart disease, as well as among those without heart disease. In the RestingECG chart, having ST is most prevalent among those with heart disease, while a normal ECG is most common among those without heart disease. The ExerciseAngina chart reveals that a positive exercise test is most common among individuals with heart disease, while a negative exercise test is most common among those without heart disease. Lastly, the ST_Slope chart indicates that having ‘Up’ or ‘Down’ is more common among individuals with heart disease.

2.4 Correlation Matrix

Correlation matrices assist in analyzing the relationship between variables within a dataset. They provide a numerical value that indicates how closely related two variables are to each other. It is important to note that correlation does not imply causation. Just because two variables are correlated does not mean that one variable causes the other.



The correlation matrix has been plotted visualizing it with a heatmap: the legend tells that the red colors show high and positive correlation, while the blue ones high and negative; otherwise the colors closest to white show low correlation. We find that there are no strong positive or negative relationships present between any features in this dataset. This is fine because as mentioned before, correlation does not imply causation.

3 Data Preprocessing

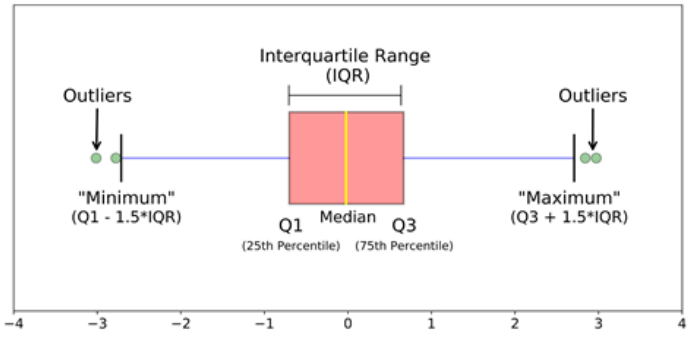
Data preprocessing is a crucial step where the raw data is transformed and cleaned to make it usable for training a model. This process involves tasks such as handling missing values and removing outliers. By preprocessing the data, we ensure that our algorithms can effectively learn patterns and make more accurate predictions. Proper data preprocessing can significantly improve the performance and efficiency of a model.

3.1 Data Cleaning

The first step was to detect and deal with any missing values. While looking at our data, it was evident that we had no NULL values as shown below.

```
Age              0
Sex              0
ChestPainType    0
RestingBP        0
Cholesterol      0
FastingBS        0
RestingECG       0
MaxHR            0
ExerciseAngina   0
Oldpeak          0
ST_Slope         0
HeartDisease     0
dtype: int64
```

Then we moved our focus to handle outliers that were observed during our data exploration step. Interquartile range (IQR) was chosen to detect and treat the outliers.



This method, as shown visually above, is a method used to identify and eliminate extreme values, or outliers, from a dataset based on the Interquartile Range (IQR). The IQR is calculated as the difference between the third quartile (Q3) and the first quartile (Q1) of the dataset. Outliers are then identified as data points that fall below *Q1 - 1.5(IQR)* or above *Q3 + 1.5(IQR)*. These outliers can skew data analysis and modeling results, so removing them can help improve the accuracy and reliability of our models. Using this technique we removed 330 outliers, bringing our total count from 918 to 588. Once this operation was complete, preprocessing of the data had been conducted.

4 Data Analysis

Data analysis involves examining and interpreting large datasets to uncover valuable insights, patterns, and trends. We have employed classification algorithms to determine which one performs best with respect to the dataset under consideration in order to predict the target variable. The cleaned dataset has been appropriately split into two subsets with specified proportions before each of these techniques is applied: the test set is used to evaluate the model's performance, and the training set is used to develop and train the model.

4.1 Support Vector Machines (SVM)

SVMs separate data into distinct classes. SVMs can be characterized as a multidimensional space, in this case where each data point represents a patient (with features like age, blood pressure, etc.). SVM finds the optimal dividing line that maximizes the distance between the high-risk and low-risk patient groups. This clear separation allows SVM to accurately classify new patients based on which side of the hyperplane they fall on. In this study, SVM achieved impressive results with a training accuracy of 90% and a test accuracy of 89.83%. Importantly, it balanced precision (correctly identifying high-risk patients) and recall (not missing high-risk patients) by accurately classifying nearly 89.6% of patients overall.

4.2 Random Forest

Random Forest uses the power of multiple decision trees to enhance prediction accuracy. Unlike a single decision tree, Random Forest builds a collection of them, each trained on a random subset of features from the dataset. When a new patient needs classification, they’re evaluated by every tree in the forest. The final call is made by the majority vote, reducing the influence of any

single tree and its potential biases. This approach helps combat overfitting, a common issue where a model performs well on training data but struggles with unseen data.

Random Forest achieved a perfect score on the training data (100% accuracy), its performance dipped on unseen test data (88.14%). However, it excelled at identifying high-risk patients with a high recall rate of 82%. This means it effectively captured most at-risk patients. However, its precision (89.13%) was slightly lower than SVM, indicating to us there was a possibility of misclassifying some low-risk patients as high-risk.

4.3 K-Nearest Neighbors (KNN)

KNN makes predictions based on the concentration of clusters. KNN relies on the principle of similarity. For a new patient, KNN identifies the k most similar patients (neighbors) from the training data based on their features (age, blood pressure, etc.). The new patient is then assigned the most common class label (high-risk or low-risk) among these neighbors. The value of k, the number of neighbors considered, can be adjusted to fine-tune the model's accuracy.

KNN's training accuracy (86.81%) was lower than SVM and Random Forest, it achieved a test accuracy on par with SVM (89.83%). Notably, KNN shined in minimizing false positives (identifying low-risk patients correctly) with a precision of 93.18%. This means it effectively avoided misclassifying healthy patients as high-risk. However, its recall (82%) was similar to the other techniques, indicating it might miss some high-risk patients.

4.4 Linear Regression

Linear Regression paints a continuous picture of risk. Unlike the classification techniques discussed earlier, linear regression doesn't categorize patients as high-risk or low-risk. Instead, it

focuses on modeling the relationship between various factors (age, blood pressure, etc.) and the likelihood of developing heart failure.

While it doesn't directly provide classifications for clinical decision-making, linear regression achieved a low mean squared error on both the training (0.1104) and test data (0.0915). Mean squared error signifies how closely the predicted probabilities resemble the actual outcomes. A lower error indicates a better fit of the model to the data, suggesting linear regression effectively captured the underlying trends.

In essence, linear regression offers valuable insights into how different factors influence the risk of heart failure, even though it doesn't provide classifications for immediate clinical use.

5 Conclusion

This report investigates the application of machine learning algorithms to predict heart failure risk using a dataset from Kaggle containing information on 918 patients. The dataset includes features like age, sex, blood pressure, cholesterol, ECG results, etc. Following data cleaning and exploration, four machine learning models were evaluated:

1. Support Vector Machine (SVM): Achieved a training accuracy of 90% and a test accuracy of 89.83%. It precisely classified 89.58% of patients overall.
2. Random Forest: Achieved a perfect training accuracy (100%) but a test accuracy of 88.14%. It had a high recall (82%) but slightly lower precision (89.13%) compared to SVM.
3. K-Nearest Neighbors (KNN): Achieved a training accuracy of 86.81% and a test accuracy of 89.83% (matching SVM). It showed the highest overall precision (93.18%) but a recall of 82%.
4. Regression: This model predicts continuous values (instead of classifying patients as high/low risk). It achieved a low mean squared error on both training and test data (0.1104 and 0.0915 respectively).

In conclusion, SVM and KNN achieved the highest accuracy (89.83%) but differed in precision-recall trade-off. SVM offered a balanced approach (precision: 89.58%, recall: 86%), while KNN prioritized precision (93.18%) at the expense of recall (82%). Random Forest showed a good recall (82%) but slightly lower accuracy (88.14%). The choice of model depends on the specific clinical need. If correctly identifying all high-risk patients is crucial, prioritizing

recall (like with Random Forest) might be preferred. If minimizing false positives is essential, KNN with its high precision could be better.