# Regression Analysis for Heart Disease Prediction

Sona Desai[1], Jamie Guo[1], Krishna Srivatsa[1], and Akhil Ganesan[1]

[1]Georgia Institute of Technology

## Contents

# 1 Objectives & Motivation

In this project, our primary goals are to determine which bioclinical factors are strongly correlated with the contraction of heart disease. Ultimately, we aim to rank these factors and develop potential models to determine a patient's theoretical probability of contracting heart disease given their health data.

To achieve this, we performed initial research reviewing the vast literature covering this disease. Then, we gathered data that included sources across numerous hospitals detailing patients with data on the biomedical factors we believed could be significant based on our literature review. Then, we performed an initial description of the data's features focusing on each feature's spread and central tendency. We also considered the data quality and what data processed we would have to perform before continuing with more advanced statistics.

Finally, after laying the groundwork, we determined the specific statistical methodology we would use to achieve our end goals. To rank the clinical factors by their significance to heart disease, we would consider the Pearson's correlation coefficient for continuous variables and the chi squared test for categorical variables. Then, to consider the significance of interaction between features, we would use an n-way ANOVA test.

After characterizing the features, to develop a comprehensive predictive algorithm, we first plan to apply linear regression to determine the relationship between the continuous features across our data. Then, we take this further and aim to develop a 3-stage machine-learning pipeline for accurate heart disease prediction that includes data preprocessing (converting categorical variables to pseudo-continuous ones using label encoding, normalizing each of the features for balanced learning, etc.), applying principal component analysis (PCA) for feature reduction (based on the results from the aforementioned feature characterization), & finally applying a machine learning model to return a categorical output.

# 2 Clinical Background & Introduction

Within the United States, cardiovascular disease takes more than 2,200 lives per day, making it a pertinent problem within the country to address. It is the leading cause of death in the US, surpassing cancerous diseases (Institute of Medicine, 2011). Further, It has even been confirmed as a "global burden," claiming one third of deaths in the world (Deaton et al., 2011). There is even an additional socioeconomic component within this global issue, with 80% of deaths being from less developed countries (Teo & Rafiq, 2021).

As heart disease becomes a growing concern for the public both inside and outside the United States, understanding the risk factors that lead to this disease is key. Current research on heart disease risk factors is substantial, including smoking, hypertension, and diabetes (Teo & Rafiq, 2021). Additional risk factors include dyslipidemia (lipid imbalance- such as high cholesterol) and obesity (Adhikary et al., 2022). Better understanding which of these risk factors or combinations of them that can cause a higher risk of heart disease is pertinent to tackling this global issue.

**Problem Statement**: Patients require a means for predicting their risk for heart disease due to the increasing prevalence of this global burden.

**Project Deliverable**: This final report will explain the process of the investigation of 11 different parameters that are potential or confirmed risk factors for heart disease and how they were interrelated to develop a model that can assist as a predictive measure for heart disease in a patient. These 11 factors used were age, sex, chest pain type, resting blood pressure, fasting blood sugar, cholesterol, resting ECG, the maximum heart rate, exercise-induced angina, the Oldpeak ECG measurement (ST depression), and the slope of the peak exercise ST segment in the ECG of the patient. Utilizing these 11 factors alongside biostatistics analysis, an ML model was developed as a predictive measure for heart disease.

**Literature Critique: "Prediction of Heart Disease using Multiple Linear Regression Model"**

This study is very applicable to our dataset and investigations. In analyzing a way to predict a patient's chance of heart disease, the researchers within this paper developed a multiple linear regression model to predict a patient's risk of heart disease based on Sex, Chest Pain Type, Fasting Blood Sugar, Resting ECG, the slope of the peak exercise ST segment in the ECG of the patient, ST depression, maximum heart rate, and age. These risk factors are very similar to the ones this proposal is analyzing in its dataset. This study heavily guided the choice in utilizing a linear regression analysis within this proposal to identify relationships between heart disease and various risk factors in the dataset. The study utilized C# and an SQL server to develop the predictive linear regression model, which can potentially provide a comparison for the predictive machine learning model of heart disease deaths to be developed in the semester. They also utilized the Pearson Correlation Coefficient to determine if the regression model was fitted to the data efficiently, supporting the proposal's adoption of the Pearson Correlation Coefficient as a determinant of correlation between risk factors and heart disease. Additionally, this study can be compared with the results of the project within this course to determine if the developed model is also an accurate predictor of heart disease in a patient due to this paper also utilizing similar risk factors.

**Literature Critique: "Gender and CVD- Does It Really Matters?"**

This study shows that gender plays a role in cardiovascular disease (CVD) susceptibility. Despite women generally exhibiting a lower prevalence of CVD compared to men, numerous studies indicate that following an acute cardiovascular event, women face a significantly higher risk of mortality and have a poorer prognosis. These gender discrepancies in CVD epidemiology, pathophysiology, and therapeutic approaches stem from variations in gene expression from the sex chromosomes and subsequent differences in sex hormones. In Western societies, ischemic heart disease typically presents 7-10 years later in women compared to men, with men being more predisposed to ST-elevated myocardial infarction (STEMI) or non-STEMI by a ratio of around 3-4 to 1. This comprehensive review synthesizes gender differences across a spectrum of cardiovascular conditions, encompassing coronary artery disease, heart failure, left bundle branch block (LBBB), atrial fibrillation, as well as examining the impact of medications and risk factors.

**Literature Critique: "Aging-associated cardiovascular changes and their relationship to heart failure"**

This paper analyzes the impacts of aging on cardiovascular processes that result in the observed increased rates of heart failure. Specifically, 5 categories are identified/analyzed: structural cardiac changes increasing the load on the heart, functional cardiac changes diminishing the heart's ability to maintain cardiac homeostasis, cardiovascular repair mechanisms declining in efficiency, increase in cardiovascular disease prevalence rates, & the impact of the deterioration of other organ systems, which interact with the cardiovascular system causing body degradation (this isn't analyzed in the study). Overall, the paper quantifies the impact these categories contribute to increased susceptibility to cardiac failure at higher ages, while focusing on plausible biological mechanisms to explain these impacts. One commonly-used tool in this study was a family of regression analyses for specific biological metrics to accumulate evidence for plausible biological mechanisms that explain the observed impacts causing heart failure.

**Literature Critique: "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms"**

In this study, the researchers focused on improving heart disease prediction accuracy and precision by analyzing and evaluating pre-existing models created using various classification algorithms and feature selection techniques. This study was performed due to the prevalence of various models used to predict heart disease, and their accuracy can significantly impact public health outcomes. Their study used the Cleveland heart disease dataset and applied ten feature selection techniques, including the ANOVA test.

The use of ANOVA in this paper is directly applicable to our proposed use in this project. In their research study, ANOVA was used to create a subset of relevant clinical factors by determining the statistical significance of each factor's contribution to the variation in disease outcomes. Specifically, ANOVA analyzes the variance within groups of patients with similar features to identify factors that have a significant impact on the prediction of disease outcome. ANOVA works by simplifying the data to a subset of important features, which results in a reduced training time and a more accurate predictive performance.

**Motivation for Biostatistics Analysis**: Based on the literature critiques conducted, and knowledge gained from BMED 2400, four components of statistical testing were utilized to supplement the ML model development and meet the project deliverables. The first analysis utilized Linear Regression with the Pearson Correlation Coefficient to complete the first two project deliverables. These deliverables involved finding a relationship with numerical and continuous datasets with multiple values, allowing linear regression to be appropriate. Additionally, to find the strength of the relationship, PCC was appropriate to supplement the linear regression analysis to see how strong the correlation was between factors for deliverable 1) and 2). For the next analysis, the Chi Squared Test was used for the third project deliverable. It was used specifically due to the presence of the categorical variable of chest pain type. Because of this it was appropriate to use due to being able to describe statistical significance with categorical inputs/outputs. For the final analysis, the ANOVA test was used. This was utilized to rank the risk factors to also assist with the ML model development. It is appropriate for this goal due to being able to assess the level with which independent variables can influence the dependent variables. These four components were integrated to supplement the ML model development alongside a Principal Component Analysis (PCA) and neural network.

# 3 Data Description

## 3.1 Features

- **Age**: Age of the patient (years); generally older patients may have other health complications that result in increased susceptibility to heart failure (i.e. blood plaque buildup)

| Patient | Mean | Median | Mode | Min | Max | S | $S^2$ |
|---|---|---|---|---|---|---|---|
| Non-Diseased | 50.55 | 51 | 54 | 28 | 76 | 9.44 | 89.21 |
| Diseased | 55.90 | 57 | 58 | 31 | 77 | 8.73 | 76.16 |
| Overall | 53.51 | 54 | 54 | 28 | 77 | 9.43 | 88.97 |

- **Sex**: Sex of the patient (M: male or F: female); research indicates due to differences in sex and metabolic physiology, females are less likely to have heart disease, but if they contract it they have a lower survival rate

| Patient | M | F | Total |
|---|---|---|---|
| Non-Diseased | 267 | 143 | 410 |
| Diseased | 458 | 50 | 508 |
| Total | 725 | 193 | 918 |

- **Chest Pain**: Type of chest pain (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic); one major symptom of heart failure is perceived chest pain, which can be subcategorized into whether it's anginal (caused by reduced blood flow to the heart) or not and typical or atypical (simple discomfort v.s. a burning/stabbing pain)

| Patient | ATA | NAP | ASY | TA | Total |
|---|---|---|---|---|---|
| Non-Diseased | 149 | 131 | 104 | 26 | 410 |
| Diseased | 24 | 72 | 392 | 20 | 508 |
| Total | 173 | 203 | 496 | 46 | 918 |

- **Resting Blood Pressure** (BP): Resting blood pressure (mm Hg); significantly high blood pressure (called hypertension) is associated with heart disease/failure as prolonged high blood pressure can cause extra wear on cardiac muscles resulting in heart failure. During heart failure however, the lack of pumping blood will result in significantly lower blood pressure than normal.

| Patient | Mean | Median | Mode | Min | Max | S | $S^2$ |
|---|---|---|---|---|---|---|---|
| Non-Diseased | 130.18 | 130 | 120 | 80 | 190 | 16.50 | 272.24 |
| Diseased | 134.19 | 132 | 140 | 0 | 200 | 19.83 | 393.18 |
| Overall | 132.40 | 130 | 120 | 0 | 200 | 18.51 | 342.77 |

- **Cholesterol**: Concentration of cholesterol in the blood (mm/dl); this includes readings of both LDL (low-density lipoproteins, makes up the majority of blood cholesterol and is a major risk factor for molecule buildup that blocks blood vessels causing heart failure) & HDL (high-density lipoproteins, generally referred to as the "good cholesterol" but compose the minority fraction)

| Patient | Mean | Median | Mode | Min | Max | S | $S^2$ |
|---|---|---|---|---|---|---|---|
| Non-Diseased | 238.77 | 231.5 | 240 | 85 | 564 | 55.39 | 3068.56 |
| Diseased | 251.06 | 246 | 282 | 100 | 603 | 62.46 | 3901.59 |
| Overall | 244.64 | 237 | 254 | 85 | 603 | 59.15 | 3499.14 |

- **Fasting Blood Sugar**: Categorical fasting blood sugar (1 if $>$ 120 mg/dl, otherwise 0); high blood sugar levels (correlated with other conditions like diabetes) can result in blood vessel damage over time, which accumulate to a greater risk of heart disease.

| Patient | 0 | 1 |
|---|---|---|
| Non-Diseased | 366 | 44 |
| Diseased | 338 | 170 |
| Total | 704 | 215 |

- **Resting ECG**: Resting electrocardiogram (ECG) 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); an abnormal ECG can be either LVH (indicates the heart's wall is thickening) or ST (indicates the coronary artery is blocked), both of which cause impede blood flow and increase the risk to heart failure

| Patient | Normal | ST | LVH |
|---|---|---|---|
| Non-Diseased | 267 | 61 | 82 |
| Diseased | 285 | 117 | 106 |
| Total | 552 | 178 | 188 |

- **Max Heart Rate**: Achieved max heart rate (between 60 and 202); failure to achieve the close to the estimated maximum heart rate for a patient's age group can indicate reduced cardiac ability and susceptibility to heart diseases

| Patient | Mean | Median | Mode | Min | Max | S | $S^2$ |
|---|---|---|---|---|---|---|---|
| Non-Diseased | 148.15 | 150 | 150 | 69 | 202 | 23.29 | 542.33 |
| Diseased | 127.66 | 126 | 120 | 60 | 195 | 23.39 | 546.95 |
| Overall | 136.81 | 138 | 150 | 60 | 202 | 25.46 | 648.23 |

- **Exercise Angina**: Exercise-induced angina/chest pain (Y or N); if exercise induced angina, it could indicate the clogging of coronary arteries (likely by cholesterol), which is a high risk factor for further heart diseases

| Patient | N | Y |
|---|---|---|
| Non-Diseased | 355 | 55 |
| Diseased | 192 | 316 |
| Total | 547 | 371 |

- **Old Peak**: ST Depression; depressions in an ECG (levels where the ECG falls below the baseline reading) can indicate blockage of blood flow to the heart, which could result in heart disease

| Patient | Mean | Median | Mode | Min | Max | S | $S^2$ |
|---|---|---|---|---|---|---|---|
| Non-Diseased | 0.41 | 0 | 0 | -1 | 4.2 | 0.70 | 0.49 |
| Diseased | 1.27 | 1.2 | 0 | -3 | 6.2 | 1.15 | 1.33 |
| Overall | 0.89 | 0.6 | 0 | -3 | 6.2 | 1.07 | 1.14 |

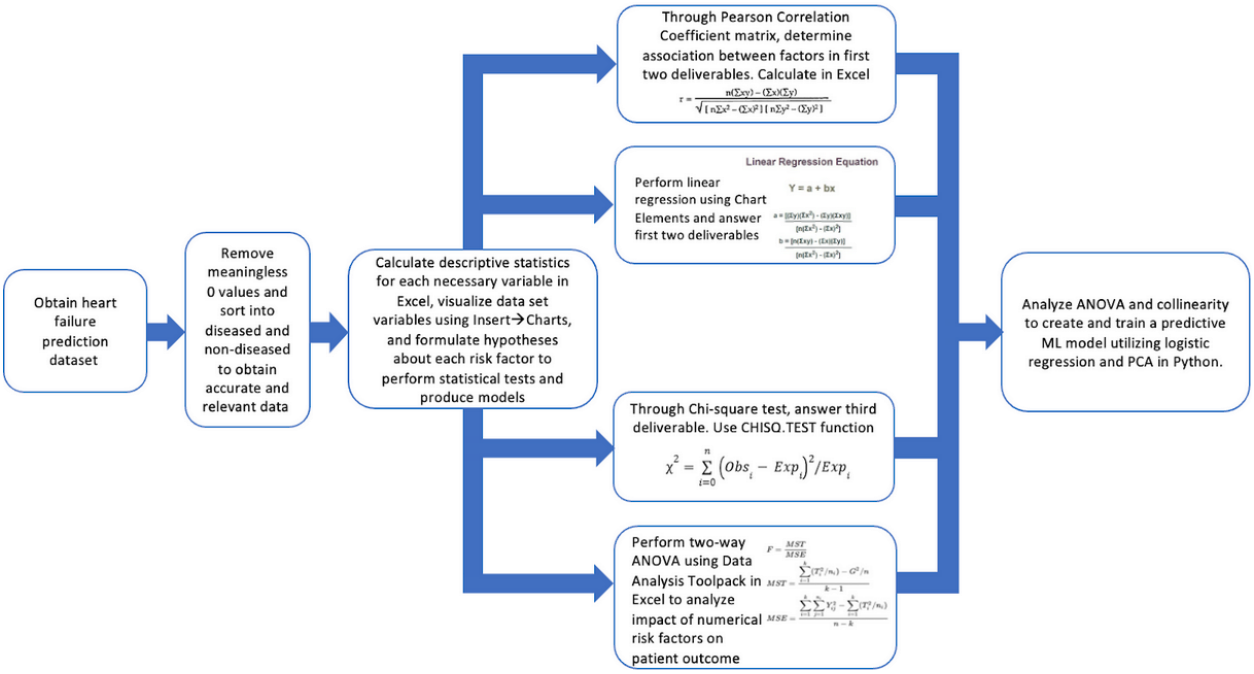
- **ST Slope**: Slope of the peak exercise ST segment (Up/Upsloping, Flat, or Down/Downsloping); the slope of the ST segments in ECGs in recent research has been used to predict restricted blood vessels, which correlate with heart disease conditions

| Patient | Up | Flat | Down |
|---|---|---|---|
| Non-Diseased | 317 | 79 | 14 |
| Diseased | 78 | 381 | 49 |
| Total | 395 | 460 | 63 |

## 3.2 Data Quality

Within this dataset, there is only one significant shortcoming in the data quality: some entries in the cholesterol column were likely unrecorded and currently have a strand of inputs of 0 (which were removed before the cholesterol data was analyzed). This may result in inaccurate measurements using cholesterol as a metric to predict heart disease likelihood. Besides this, in some statistics there are a few outliers (i.e. the minimum "max heart rate" is 60, the minimum "resting BP" is 0 due to one input missing actual data, etc.) which can be accounted for with better data inputting and cleanup processes (i.e. removing the data points with incomplete data in the categories of analysis).

# 4 Statistical Analysis

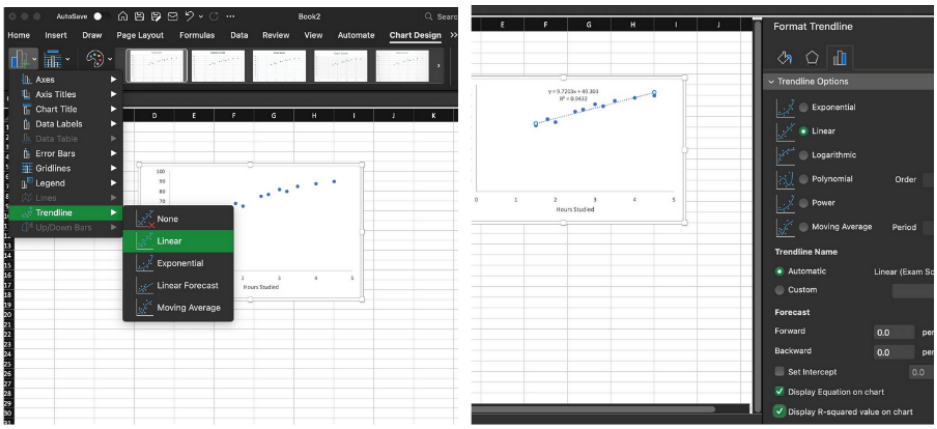



## 4.1 Linear Regression & Pearson Correlation Coefficient

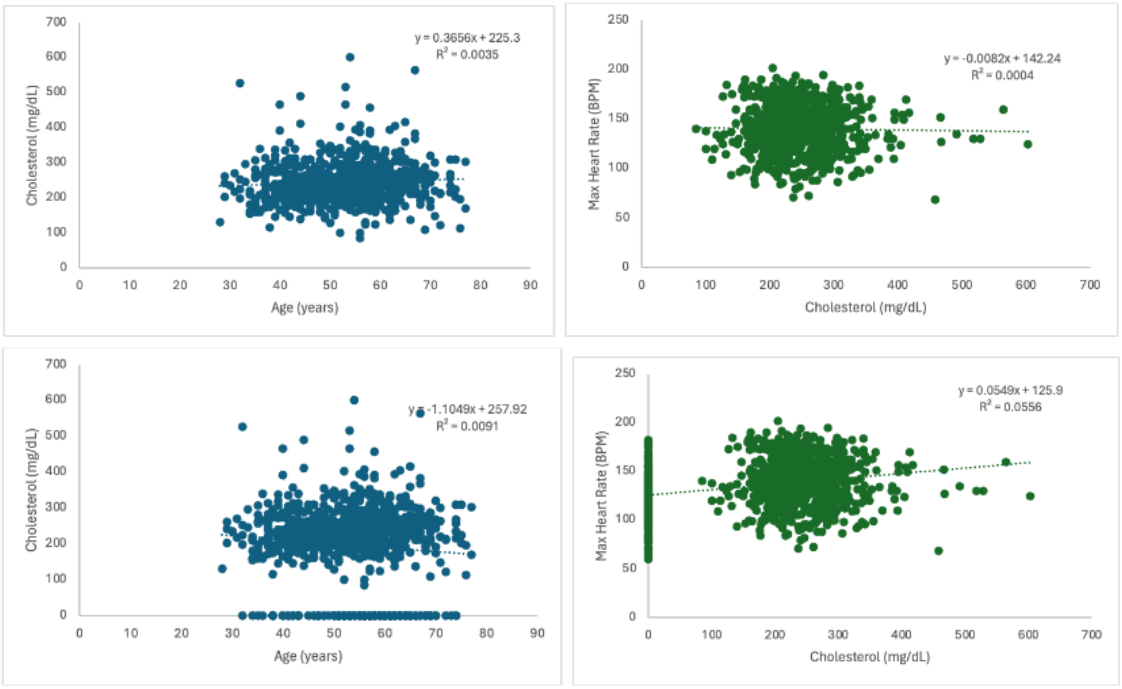
The linear regression method was used to address the statistical analysis deliverables within the project: (1) Does age affect cholesterol levels and why? (2) Does cholesterol affect maximum heart rate? These continuous data sets are numerical values that can be compared with a linear regression statistical analysis utilizing previous research into linear regression from the knowledge inquiry. Specifically, hypotheses developed for this portion of the project were the following:

**Null**: Age does not have any effect on cholesterol levels. Cholesterol does not affect maximum heart rate.
**Alternative**: As age increases, cholesterol increase. As cholesterol increases, maximum heart rate increases.

To determine if there is a relationship, the Pearson Correlation Coefficient (PCC) was used to determine if there is a strong correlation between risk factors. If the PCC value is above $R = 0.7$ (a strong correlation value), the null hypothesis will be rejected and the alternative hypothesis will be accepted, indicating a strong positive correlation. To analyze (1) and (2), Excel was used utilizing the "Chart Elements" function to perform a linear regression. First the independent and dependent variable data sets were pasted into an Excel sheet. Then the charts function was utilized to generate a Scatter Plot, followed by selecting "Insert" then "Add Chart Element", and finally "Trendline" and "Linear." This will generate a trendline in the scatter plot. Utilizing the Trendline Options function, one can display the $R^2$ value and the linear regression equation as well. This analysis through Excel allowed for the statistical analysis of whether age is correlated with cholesterol and whether cholesterol is correlated with maximum heart rate. The figure is included below for reference to the pathway for performing this statistical analysis using Excel.




Regarding final results, the first graph pair below shows the results of the linear regression analysis without outliers and then the second graph pair below shows these results with outliers.




Based on the R value for both deliverables in both conditions, the null hypothesis was accepted both with and without outliers for both deliverables (due to $R < 0.7$ for both conditions in both deliverables). The R value was found by simply taking the square root of the $R^2$ value as well as utilizing the PEARSON function in Excel for each comparison. This was a surprise to the team as it was hypothesized that there would be a strong positive correlation between age and cholesterol as well as cholesterol and maximum heart rate. Because of these results, it was concluded that age does not have a significant effect on cholesterol and cholesterol does not have a significant effect on maximum heart rate. Future work involves utilizing data sets from the same location/ethnicity to determine if variability could be due to differing populations in one dataset. This would also control for other extraneous variables allowing for more of the results to be focused on the specific relationship between the two variables.

## 4.2 Chi Squared Test

The chi squared test can address the remaining statistical analysis deliverables: (3) How does Chest Pain Type vary among heart disease vs non-heart disease patients? To use this test, categorical data in contingency tables (see the section on data features) will be used as an "observed" table, while an "expected" table is developed using the population proportions assuming the null hypothesis: there is no variation between all the categories being analyzed. Then, these tables are used to calculate a chi squared value using the formula below, which is compared with a chi squared table and significance value/degree of freedom to determine if variation is statistically significant (i.e. the categories are not correlated).

$$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp} \tag{1}$$

This calculation can be done in excel using the CHISQ.TEST function, inputting the actual range (the contingency table from the data features) and the expected range (which is calculated using the total proportions for each category in the table). However, we performed this calculation step-by-step in the below spreadsheet (analyzing the dependence of chest pain and heart disease):

| **Observed** | ATA | NAP | ASY | TA | |
|---|---|---|---|---|---|
| 0 | 149 | 131 | 104 | 26 | 410 |
| 1 | 24 | 72 | 392 | 20 | 508 |
| | 173 | 203 | 496 | 46 | 918 |
| | | | | | |
| **Expected** | ATA | NAP | ASY | TA | |
| 0 | 77.2658 | 90.66449 | 221.5251 | 20.54466 | 410 |
| 1 | 95.7342 | 112.3355 | 274.4749 | 25.45534 | 508 |
| | 173 | 203 | 496 | 46 | 918 |
| | | | | | |
| **(Obs-Exp)^2/Exp** | 66.59863 | 17.94477 | 62.35023 | 1.448586 | |
| | 53.75086 | 14.48298 | 50.32204 | 1.169134 | |
| | | | | | |
| **X^2** | 268.0672 | | | | |
| **P-value** | 8.08E-58 | | | | |

The p-value ($8.08 \times 10^{-58}$) being less than 0.05 (5% significance) indicates a rejection of the null hypothesis & acceptance of the alternative hypothesis that chest pain varies significantly with heart disease (higher ASY & lower ATA, NAP, & TA in heart diseased patients in particular).

## 4.3 ANOVA Testing

ANOVA Test is used to determine "how much" independent variables had an effect on dependent variables. In our case, the independent variables are the risk factors associated with heart disease, and the dependent variables are whether or not the patient has heart disease. The risk factors that we will observe are Age, Sex, Cholesterol Levels, Max Heart Rate, and Blood Pressure. We will use the results of the ANOVA test to rank the importance of each of the risk factors on the probability of a patient having health disease or not. An ANOVA Test does this by comparing the ratio of the variation across groups to the variance within groups to determine level of importance. Using these numbers, we are able to calculate the F-statistics, which from there, we can calculate the p-value. The formula that we use for ANOVA Testing is listed bellow:

$$F = MST/MSE \tag{2}$$

$$MST = \frac{\sum_{i=1}^{k}(T_i^2/n_i) - G^2/n}{k-1} \tag{3}$$

$$MSE = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}^2) - \sum_{i=1}^{k}(T_i^2/n_i)}{n-k} \tag{4}$$

The easiest way to conduct the ANOVA Test is through Excel. First we will have to clean up the data. After cleaning out the data, divide up and separate out each of the risk factors. After getting the data points for the specific risk factor, separate them into two groups, with and without heart disease. Using the "Data Analysis", select "Anova: Singe Factor". Select your data points for the two groups and Excel will return a table of the calculated values. The table will include the variance of the group and the F-statistics. It will also calculate the p-value. The results of all of the tests are shown below.

Anova: Age

SUMMARY

| Groups | Count | Sum | Average | Variance |
|---|---|---|---|---|
| No Heart Disease | 410 | 20726 | 50.55122 | 89.206417 |
| Heart Disease | 508 | 28397 | 55.899606 | 76.161499 |

ANOVA

| Source of Variation | SS | df | MS | F | P-value | F crit |
|---|---|---|---|---|---|---|
| Between Groups | 6490.0868 | 1 | 6490.0868 | 79.160779 | 3.008E-18 | 3.8516306 |
| Within Groups | 75099.304 | 916 | 81.98614 | | | |
| Total | 81589.391 | 917 | | | | |

Anova: Blood Pressure

SUMMARY

| Groups | Count | Sum | Average | Variance |
|---|---|---|---|---|
| No Heart Disease | 410 | 53374 | 130.18049 | 272.23629 |
| Heart Disease | 508 | 68166 | 134.18504 | 393.17674 |

ANOVA

| Source of Variation | SS | df | MS | F | P-value | F crit |
|---|---|---|---|---|---|---|
| Between Groups | 3638.4186 | 1 | 3638.4186 | 10.727228 | 0.0010953 | 3.8516306 |
| Within Groups | 310685.25 | 916 | 339.17604 | | | |
| Total | 314323.67 | 917 | | | | |

Anova: Cholesteral Levels

SUMMARY

| Groups | Count | Sum | Average | Variance |
|---|---|---|---|---|
| No Heart Disease | 390 | 93120 | 238.76923 | 3068.5636 |
| Heart Disease | 356 | 89378 | 251.0618 | 3901.5905 |

ANOVA

| Source of Variation | SS | df | MS | F | P-value | F crit |
|---|---|---|---|---|---|---|
| Between Groups | 28122.955 | 1 | 28122.955 | 8.1138508 | 0.0045141 | 3.8539879 |
| Within Groups | 2578735.9 | 744 | 3466.0428 | | | |
| Total | 2606858.8 | 745 | | | | |

Anova: Max Heart Rate

SUMMARY

| Groups | Count | Sum | Average | Variance |
|---|---|---|---|---|
| No Heart Disease | 410 | 60742 | 148.15122 | 542.33404 |
| Heart Disease | 508 | 64849 | 127.65551 | 546.94815 |

ANOVA

| Source of Variation | SS | df | MS | F | P-value | F crit |
|---|---|---|---|---|---|---|
| Between Groups | 95308.3 | 1 | 95308.3 | 174.91359 | 1.1378E-36 | 3.8516306 |
| Within Groups | 499117.34 | 916 | 544.88792 | | | |
| Total | 594425.64 | 917 | | | | |

Anova: Sex

SUMMARY

| Groups | Count | Sum | Average | Variance |
|---|---|---|---|---|
| No Heart Disease | 410 | 553 | 1.3487805 | 0.227688 |
| Heart Disease | 508 | 558 | 1.0984252 | 0.0889127 |

ANOVA

| Source of Variation | SS | df | MS | F | P-value | F crit |
|---|---|---|---|---|---|---|
| Between Groups | 14.220617 | 1 | 14.220617 | 94.253184 | 2.822E-21 | 3.8516306 |
| Within Groups | 138.20313 | 916 | 0.1508768 | | | |
| Total | 152.42375 | 917 | | | | |

To rank the importance of each of the risk factors, we will look at the p-values of each factor. The smaller the p-value, the more it contributes to the cause of heart disease. After looking to the data, we were able to conclude a ranking as follows:

1. Max HR
2. Sex
3. Age
4. Blood Pressure
5. Cholesterol Levels

Max Heart Rate seemed to be the biggest risk factor, while Cholesterol Levels looks to have the least impact. This matched with our Literature Critique, as the paper "Gender and CVD- Does It Really Matters?"

states that gender plays a big role in the development of heart disease. Women are a lot more susceptible than men. This is consistent with our ranking, as Sex ranks second on contributing to heart disease.

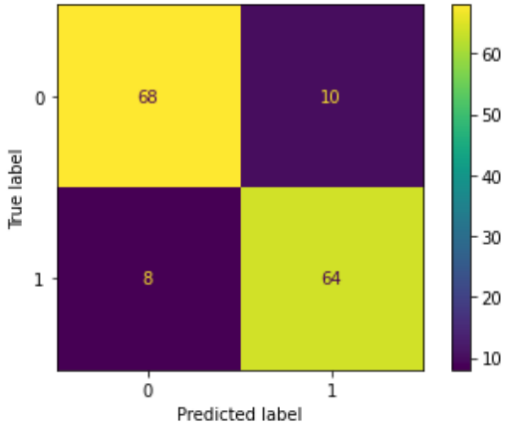
## 4.4 Machine Learning Model

We will be utilizing a 3-stage machine learning pipeline that involves data preprocessing, dimensionality reduction, and ultimately a classifier yielding an accurate predictor.

The data preprocessing first involves cleaning up the dataset (removing data points that are incomplete & accounting for outliers). After, we perform label encoding: using an encoder to transfer categorical data fields to numerical by assigning numbers to each category. This is critical as 6 of the 11 fields the dataset provides are categorical and significantly increase the data available for analysis. Following label encoding for categorical fields, we perform data normalization, which transforms the mean of each field to 0 (and standardizes the variance), assigning a z-score to every other datapoint. By normalizing the data, it prevents the model from assigning reliance to whichever fields have higher natural scales, resulting in enhanced model stability and increased result accuracy. The final step in preprocessing involves the train-test split, for which the dataset was split into the standard 80% for training and 20% for testing (with randomized shuffling for the data allocation).

The next stage in the pipeline involves data dimensionality reduction. This is a critical step as it allows for less-intensive training and more accurate results. For this, we use a principle component analysis (PCA), which applies singular value decomposition (SVD) to determine and eliminate collinear parameters. This unsupervised machine learning algorithm allows us to remove input parameters which are linearly-dependent on other inputs (i.e. collinear in the dataset). The existence of such data will typically hinder a classifier algorithm as collinear data being provided only serves to reinforce that data's influence in the prediction algorithm, which will reduce the convergence accuracy and/or increase time to convergence. As a result, its elimination through PCA will enhance the final results of the predictor. Additionally, the Pearson Correlation Coefficient test can provide information about how these risk factors are related and possible collinear risk factors.

The final stage in the pipeline is the prediction algorithm. The prediction algorithm of choice is a multi-layer perceptron classifier, which is based on a neural network architecture (our specific model uses 2 hidden layers of 10 neurons each, with a logistic activation function). In our case, the independent variable will be all of the signs that the patient has, such as age, gender, blood pressure, and max heart rate. Our dependent variable outcome in this case is whether or not the patient is or is not at risk for heart disease. This model can account for risk factors associated with heart disease and help patients gain valuable information on the status of tier health.



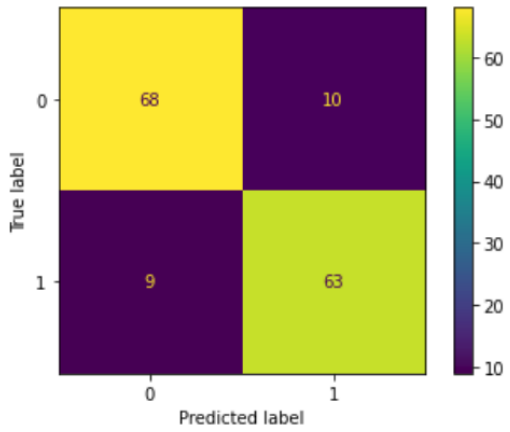

The final model performance results are as follows:

- **Accuracy**: $(68 + 64)/(68 + 64 + 8 + 10) = 88\%$
- **Sensitivity**: $64/(64 + 8) = 0.89$

- **Specificity**: $68/(68+10) = 0.87$

Overall this indicates the model can be a fairly accurate death predictor with respect to heart disease for the given dataset features. The code for this model can be found at this github: https://github.com/akhilganesan/Heart-Disease-Classifier.

To verify this model's results, we retested this data using a simple logistic regression model. The results of this are summarized in the below confusion matrix.




These results are strikingly similar to those of the MLP model other than being slightly inaccurate on this specific dataset (by 1 datapoint). This could be interpreted as an indication that the MLP model hasn't been overfit to the specific dataset and the trends it uses for prediction may translate to the general population, verifying the model for greater wide-scale testing/use.

## 4.5 Final Summary

Our project focused on using statistical analyses to predict the likelihood of heart disease based on various factors by analyzing datasets. We explored the clinical background of heart disease risk factors and applied tests such as ANOVA, Linear Regression, Pearson correlation, and Chi-Squared to investigate associations between factors like age, sex, cholesterol levels, blood pressure, max heart rate, and the presence of heart disease. These initial analyses guided our selection of variables for our final machine learning model, as we were able to create a ranking of these factors. Through our calculations and modeling, we were able to predict the probability of heart disease occurrence for patients based on the risk factors listed above. Our findings were compared to existing literature on heart disease risk factors, providing valuable insights into predicting and understanding heart disease. Based on the linear regression model, there was no correlation found between age and cholesterol or max heart rate and cholesterol. This aligns with the final ranking list formed from the ANOVA test due to the lack of these factors being ordered next to each other showing their independence. The Chi Square results answered the third deliverable in the project to determine if chest pain type differed significantly with heart disease. These tests helped supplement the formation of the final ML heart disease model for this final project.

# 5 Supporting Documents, Codes, and Team

## ML Model Code

https://github.com/akhil-ganesan/Heart-Disease-Classifier

## Literature Paper PDFs

- **"Prediction of Heart Disease using Multiple Linear Regression Model"**: https://www.ijedr.org/papers/IJEDR1704226.pdf
- **"Gender and CVD- Does It Really Matters?"**: https://www.sciencedirect.com/science/article/pii/S014628062300021X?via%3Dihub
- **"Aging-associated cardiovascular changes and their relationship to heart failure"**: https://www.sciencedirect.com/sdfe/pdf/download/eid/1-s2.0-S1551713611001012/first-page-pdf
- **"Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms"**: https://www.hindawi.com/journals/acisc/2021/5581806/

## Oral Presentation

https://docs.google.com/presentation/d/1zgZtDP3KTyGpCCXlKmiJCupWSAQm3YI15k1qmtVu8r4/edit?usp=sharing

## Teamwork in Overall Project

Overall, our team worked very well together on this project. From day 1, our team communicated the goals and objectives of each step of the project with each other. We effectively split up the workload and frequently checked in on one another to review progress. We divided the work based on our prior skillset, so each team member was comfortable with their workload. We met on Zoom often to discuss progress and record our presentations. Our team made sure to submit project components early; this practice worked best as we had ample time to get started on the next component, work on in class assignments, and receive feedback for our submitted assignments. Our team chemistry could have been better had we met in person more often to work on the project. Meeting in person more often would have likely improved the efficiency of the project completion.

## Contributions

- **Sona Desai**: Clinical Background and Introduction; Literature Critique; Results and Conclusion - Linear Regression & Pearson Correlation Coefficient; Supporing Documents, Codes, and Team
- **Jamie Guo**: Literature Critique; Results and Conclusions - ANOVA Test & Final Summary; Supporting Documents, Codes, and Team; Appendix
- **Krishna Srivatsa**: Literature Critique; Results and Conclusion - Pearson Correlation Coefficient; System Flow Diagram; Supporting Documents, Codes, and Team; Teamwork Review
- **Akhil Ganesan**: Objectives & Motivation; Literature Critique; Data Description; Results and Conclusion - Chi Squared Test & ML Model; Supporting Documents, Codes, and Team; Appendix

# 6 Bibliography


1. Adhikary, D., Barman, S., Ranjan, R., & Stone, H. (2022). A Systematic Review of Major Cardiovascular Risk Factors: A Growing Global Health Concern. Cureus, 14(10), e30119. https://doi.org/10.7759/cureus.30119

2. Deaton, C., Froelicher, E. S., Wu, L. H., Ho, C., Shishani, K., & Jaarsma, T. (2011). The global burden of cardiovascular disease. European journal of cardiovascular nursing, 10 Suppl 2, S5–S13.https://doi.org/10.1016/S1474-5151(11)00111-3

3. Fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [Date Retrieved] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

4. Institute of Medicine (US) Committee on a National Surveillance System for Cardiovascular and Select Chronic Diseases. A Nationwide Framework for Surveillance of Cardiovascular and Chronic Lung Diseases. Washington (DC): National Academies Press (US); 2011. 2, Cardiovascular Disease. Available from: https://www.ncbi.nlm.nih.gov/books/NBK83160/

5. Kaushalya Dissanayake, Md Gapar Md Johar, "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms", Applied Computational Intelligence and Soft Computing, vol. 2021, Article ID 5581806, 17 pages, 2021. https://doi.org/10.1155/2021/5581806

6. Prasad, K. P. D. (2017). Prediction of Heart Disease using Multiple Linear Regression Model. IJEDR, 5(4).https://www.ijedr.org/papers/IJEDR1704226.pdf

7. Strait, J. B., & Lakatta, E. G. (2012). Aging-associated cardiovascular changes and their relationship to heart failure. Heart failure clinics, 8(1), 143–164. https://doi.org/10.1016/j.hfc.2011.08.011

8. Suman, S., Pravalika, J., Manjula, P., & Farooq, U. (2023). Gender and CVD- Does It Really Matters?. Current problems in cardiology, 48(5), 101604. https://doi.org/10.1016/j.cpcardiol.2023.101604

9. Teo, K. K., & Rafiq, T. (2021). Cardiovascular Risk Factors and Prevention: A Perspective From Developing Countries. The Canadian journal of cardiology, 37(5), 733–743. https://doi.org/10.1016/j.cjca.2021.02.009