CUNY Bernard M. Baruch College

Data Mining for Business Analytics CIS 3920

Which health metrics and tests correlate the most to heart disease?

By
Dorwin Liang
And
David Huang

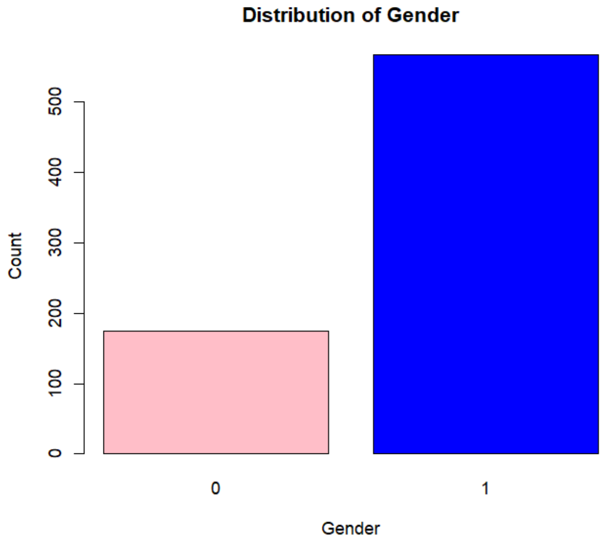
Team Name : Team Heart for All

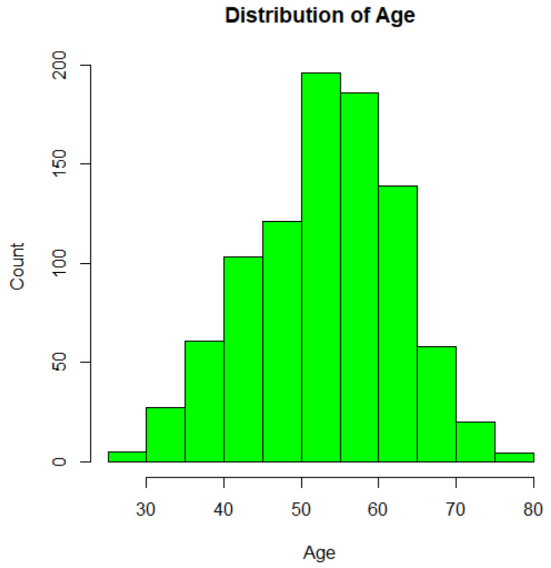
Dataset : UCI Heart Disease Data
Data Source : https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data/
Data Description :

- **Number of Obvservations : 920**
- **Number of Variables : 16**
- **Variable Descriptions**
  - **id** (Unique id for each patient)
  - **age** (Age of the patient in years)
  - **Dataset** (place of study)
  - **sex** (Male/Female)
  - **cp** chest pain type ([typical angina, atypical angina, non-anginal, asymptomatic])
  - **trestbps** resting blood pressure (resting blood pressure (in mm Hg on admission to the hospital))
  - **chol** (serum cholesterol in mg/dl)
  - **fbs** (if fasting blood sugar > 120 mg/dl True/ False)
  - **restecg** (resting electrocardiographic results)
    - Values: [normal, stt abnormality, lv hypertrophy]
  - **thalach:** maximum heart rate achieved
  - **exang:** exercise-induced angina (True/ False)
  - **oldpeak:** ST depression induced by exercise relative to rest
  - **slope:** the slope of the peak exercise ST segment
  - **ca:** number of major vessels (0-3) colored by fluoroscopy
  - **thal:** [normal; fixed defect; reversible defect]
  - **num:** the predicted attribute (0-4). target [0=no heart disease; 1-4 = heart disease present]
- **Target Variable :** Num
  - Indicates the presence or absence of heart diseases
  - 0 indicates a negative diagnosis with no heart disease.
  - 1-4 indicates a positive diagnosis of heart disease
- **What is (are) the question(s) you intend to address?**
  - Which health metrics and tests correlate the most to heart disease?
- **Method to use in this analysis**: Logistic regression analysis
- **Who is your target audience?**
  - Public Health Officials.
    - Insights can guide public health initiatives, campaigns, or policy decisions related to heart health for the public so they can see what correlates the most to heart disease and inform the public.

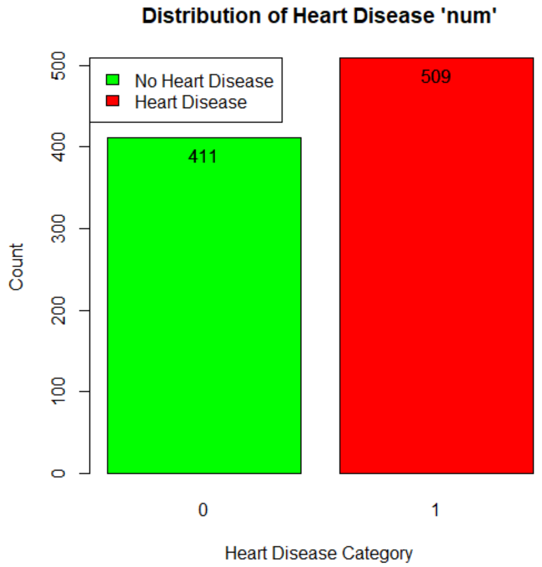
Progress Report

So far, our group has finalized the dataset and familiarized ourselves with the data. We collaborated on R Studio to conduct exploratory data analysis. Before anything, since we focused on logistic regression, we transformed the target variable into a binomial category to indicate the presence of heart disease or no presence. First, we decided to explore the dataset by looking at its dimensions and the statistical summary of each column. Through this, we found interesting facts about the dataset, such as the data being skewed heavily towards males with 726 records for males and only 194 records for females.



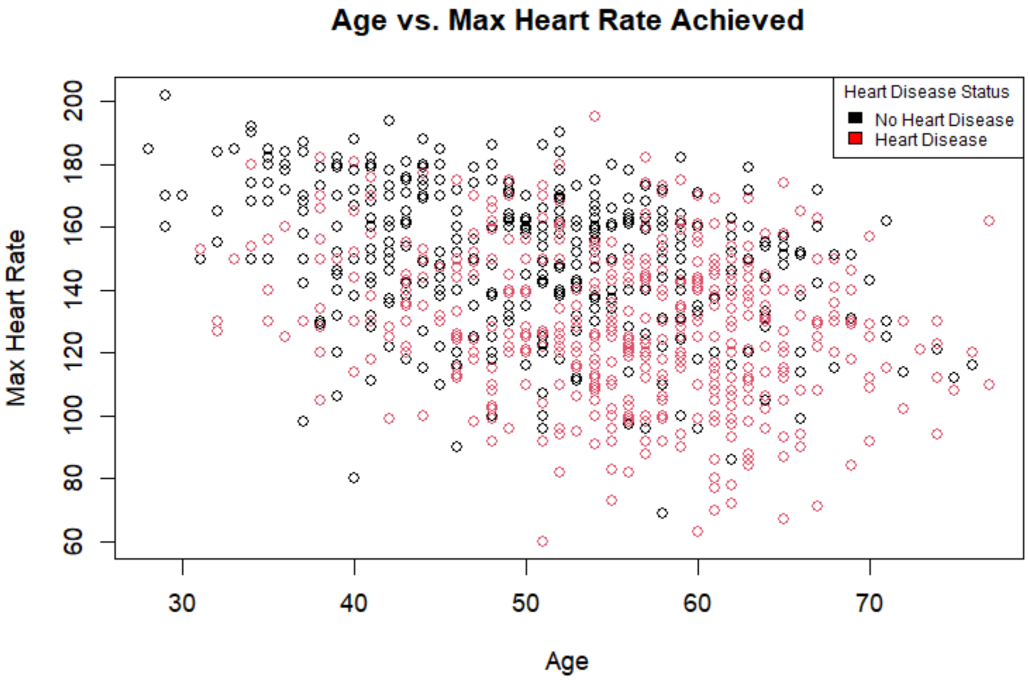


The age distribution in this dataset is slightly right-skewed with the mode being 50-55 year olds. The age spread in this dataset is 28 to 77 years old.

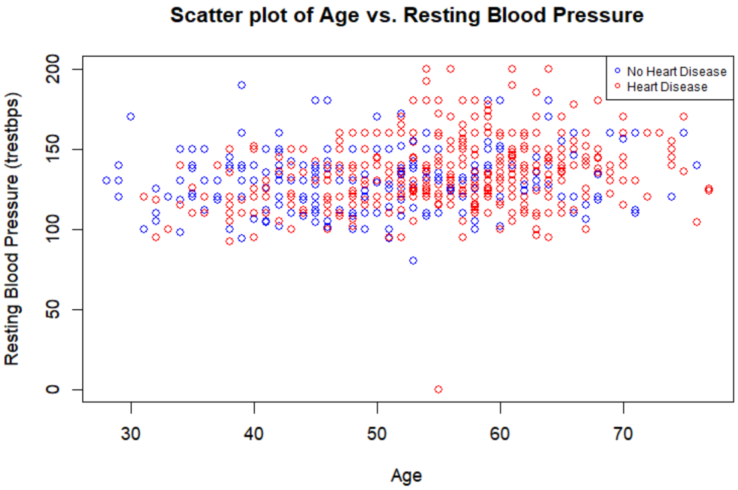


The distribution of records of individuals with heart disease is fairly equal with 508 records of individuals with heart disease and 411 records of individuals without heart disease.

Next, we decided to explore some correlations between different variables. One of the correlations we decided to explore was between age and max heart rate. It appears that there is a general trend where the maximum heart rate decreases as age increases.



Next, we decided to explore the correlation between age and resting blood pressure. The scatter plot showed no signs of a trend between these two data points but does show signs of heart disease being more prevalent as age increases.



Our plan for the remaining task is to explore the usage of logistic regression analysis and learn how to implement it into our dataset to develop multiple models. After that, we will use methods to evaluate each model's performance to reach our final model.

Data Preprocessing

Taking a look at the dataset, we realized that 3 columns had way too much missing data, which we couldn't fill in or remove the rows for, so we dropped the columns. The three columns are 'ca', 'thal', and 'slope' which represent the number of major vessels, heart defects, and the slope of the peak exercise ST segment, respectively. Next, we also removed the 'id' column as it would interfere with the accuracy of the model given that the column was just an index of each row. For the remaining 12 columns, we needed to remove rows with missing values in the variables that had missing values in them. To check which columns had missing data, we created a filter on Excel and spotted 7 columns that had missing data. In R, we created a list of these columns and used a built-in function to automatically remove rows where there are missing values. This resulted in a total of 741 rows remaining in our dataset.

For the next step, we converted categorical variables into a format that the models could interpret better using one-hot encoding, which turns a categorical column into separate columns with 0 or 1 indicating its presence or not. To do this, we researched and found out that there was a library called fastDummies that could help make this process more efficient and get rid of human error.

Afterwards, we ran into a problem with the logistic regression model we used later on where it wouldn't run because of a multicollinearity issue, so we resolved this by getting rid of columns that were giving us this issue.

We then converted the rest of the categorical variables which were 'sex', 'fbs', and 'exang' into the binary format they needed to be in for the models to interpret. In the last step of the data preprocessing stage, we scaled the numerical columns to standardize them. We did this for the six numerical columns in this dataset using the scale function in R.

Data mining Methods

The first and most ideal model we chose was logistic regression because it fit the question we intended to address perfectly. We wanted to see which variables greatly contribute to the presence of heart disease and logistic regression is a model that helps predict the probability of a binary outcome, and in addition, provides us with coefficients for each variable to see how much it contributes, as well as how statistically significant it is. Its interpretability is an advantage as it would allow our target audience, public health officials, to fully understand the impact of each variable. For all of our models, we used all of the variables available in our processed dataset. Specifically for logistic regression, the process of one-hot encoding of categorical variables was crucial for ensuring the model performed as expected because logistic regression is designed to work with numerical values.

KNN is the second model we chose to use because it can predict patterns and clusters in the dataset. We wanted to use a model that had a classification approach and dealt with the data's proximity to one another. To use KNN, we included scaling to standardize numerical columns so that all variables contributed equally to the distance algorithm that KNN relies on. The model runs ten times for the number of neighbors in the model, which was up to 10. Then the dataset is divided into five folds randomly with one fold being used as the test set, and the remaining four folds as the training set. The model then has a matrix that contains the accuracy for each combination of neighbors and folds for model performance analysis.

For our third model, we chose to use Naive Bayes to test if it would outperform the other two models. The different algorithms would help give a different perspective on the dataset by showing how related the variables are to one another. To create a Naive Bayes model, we imported a package that we found through research that contains a function for the Naive Bayes classifier. The Naive Bayes function will be the same as the other models in the way that all variables will be used in the model.

Calculating and comparing model performances

To see the performance of the three different models, we used 5-fold cross-validation to calculate the average accuracy of the models. We also extracted the confusion matrix of the last folds of each model to calculate its respective recall, precision, and accuracy.

	Average	**Last Fold**		
Model Type	5-fold Cross-validation Accuracy	Recall	Precision	Accuracy
Logistic Regression	0.8077	0.8169	0.7733	0.8101
KNN	0.9100	0.8947	0.9067	0.9051
Naive Bayes	0.7854	0.8636	0.7600	0.8291

Figure 1: Table of the average 5-fold cross-validation accuracy as well as the recall, precision, and accuracy of the last fold for each of the three models.

The table (Figure 1) above shows the average accuracy of each model as well as the recall, precision, and accuracy of the last fold of each model. By comparing the performance of the three models, we can see that KNN had the best accuracy in predicting the presence of heart disease with 91% accuracy. Logistic regression and Naive Bayes had about the same in terms of average accuracy at 80.77% and 78.54% respectively.

Looking at the last fold of each model, KNN also beat out the other two models in terms of precision and recall. in the scenario where a model is built to predict the presence of heart disease, it is important for accuracy, recall, and precision to be high. Accuracy reflects the overall correctness of the model which is important. A high recall rate means that false negatives are minimized which is extremely important because not correctly identifying a patient with heart disease can lead to serious consequences. Having high precision is also beneficial to minimize unnecessary interventions with patients who do not have early signs of heart disease.

The best model for predicting the presence of heart disease

Although KNN has the highest accuracy, recall, and precision, the model does not directly give us the answer we were looking for. Logistic regression remains the best model in predicting the presence of heart disease because it directly answers the question we were looking for. The ability to estimate the effect size of each variable is crucial in a task like identifying what contributes most to heart disease. While KNN has a higher predictive accuracy, logistic regression provides valuable insights into the specifics. With an average accuracy of 80.77%, it is still a high-performing model that is much more accurate compared to a random classifier.

The summary of the logistic regression (Figure 2) allows us to view the coefficient of each variable and whether it is statistically significant or not. Through this summary, we were able to see that four statistically significant variables had high correlations with heart disease. These four variables are the gender of the person (sex), the presence of exercise-induced chest pain (exang), ST depression induced by exercise relative to rest (oldpeak), and chest pain that had no symptoms (cp_asymptomatic).

Variables	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-1.67515	0.60588	-2.765	0.005695
age	0.18821	0.13218	1.424	0.154465
sex	**1.3767**	**0.26861**	**5.125**	**2.97E-07**
trestbps	0.06754	0.11223	0.682	0.495262
chol	-0.00214	0.13591	-0.016	0.987193
fbs	0.52313	0.30412	1.72	0.085413
thalach	-0.24082	0.13488	-1.785	0.074192
exang	**0.9898**	**0.23771**	**4.164**	**3.13E-05**
oldpeak	**0.78088**	**0.12913**	**6.047**	**1.47E-09**
dataset_Cleveland	-0.55814	0.39509	-1.413	0.157744
dataset_Hungary	-0.52124	0.38513	-1.353	0.175925
dataset_Switzerland	4.07426	1.10975	3.671	0.000241
cp_asymptomatic	**1.2814**	**0.44356**	**2.877**	**0.004012**
\`cp_atypical angina\`	-0.50348	0.494	-1.019	0.308112
\`cp_non-anginal\`	-0.15567	0.45835	-0.342	0.732652
\`restecg_lv hypertrophy\`	0.23148	0.27991	0.827	0.408237
\`restecg_st-t abnormality\`	-0.17262	0.34079	-0.507	0.6125

Figure 2: The summary output of our logistic regression model which shows four statistically significant variables with positive correlations to presence of heart disease

Actionable steps for public health officials

From knowing the four statistically significant variables that strongly correlate with the presence of heart disease, we can provide actionable steps for public health officials to inform the public.

Firstly, given that gender is statistically significant and has a positive value, it means that men are more likely to develop heart disease than women. Additionally, the presence of exercise-induced chest pain also correlates highly with the presence of heart disease. To tackle this, public health officials can create and launch targeted campaigns for men who have chest pain while exercising and persuade them to get a heart health screening.

According to the model, it appears that the presence of heart disease is also prevalent in those who have asymptomatic symptoms, meaning they have an underlying medical condition but without knowing it because it does not physically affect them. In this case, ST depression induced by exercise is highly correlated to the presence of heart disease as well as asymptomatic chest pain. These two conditions are not easily noticeable as it does not have a reaction in your body that would have you thinking something is wrong. To solve this issue, public health officials will also have to make another campaign focusing on the entire population, focusing on addressing misconceptions about heart disease, and emphasizing that early signs of heart disease can be asymptomatic.

Conclusion

Using the UCI Heart Disease Dataset and a logistic regression model, we were able to reveal four statistically significant variables that contribute to the presence of heart disease. These variables are gender, exercise-induced chest pain, ST depression, and asymptomatic chest pain. These findings are important and relevant for public health officials as they can create health campaigns targeted towards men with high risk of heart disease as well as health campaigns that inform the general population of asymptomatic heart disease.

However, the results of our models should be taken with a grain of salt given the limitations of the dataset. There may be other variables that have an even higher correlation to the presence of heart disease which were not part of the dataset we used. During the data pre-processing, we also had to remove 179 rows in the dataset which accounted for roughly 19.4% of the data, leaving the model to work with 741 rows of data which is not much data to work with given an industry such as healthcare where precision and accuracy is important and the size of the dataset greatly affects those measures.

References

Meyer, D. (2023, February 1). *Naivebayes: Naive Bayes classifier in E1071: MISC functions of the Department of Statistics, Probability Theory Group (formerly: E1071), Tu Wien*. naiveBayes: Naive Bayes Classifier in e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. https://rdrr.io/rforge/e1071/man/naiveBayes.html

What is one-hot encoding. Deepchecks. (2021, August 5). https://deepchecks.com/glossary/one-hot-encoding/#:~:text=One%2Dhot%20encoding%20in%20machine,categorical%20data%20in%20machine%20learning.

Kaplan, J. (n.d.). *Fast creation of dummy (binary) columns and rows from categorical variables*. Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables fastDummies. https://jacobkap.github.io/fastDummies/

Bhandari, A. (2023, November 9). *Multicollinearity: Causes, effects and detection using VIF (updated 2023)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/

Burns, E., Buttner, R., & Buttner, E. B. and R. (2022, March 16). *The St Segment*. Life in the Fast Lane • LITFL. https://litfl.com/st-segment-ecg-library/