Machine Learning with a Heart

Caleb McCurdy

University of San Diego

Master of Science, Applied Data Science

ADS 501 – Foundation of Data Science

Section 01

17 October 2022

|  |
| --- |
| **Business Understanding** |

**Background**

Heart disease is the leading cause of death for both men and women in the United States (CDC, 2021). The most common form of heart disease is coronary artery disease (CAD). According to the Centers for Disease Control and Prevention, over 20 million adults in the United States have CAD. This occurs when cholesterol, fat, and other substances stick to the walls lining blood vessels. Plaque build-up, often over the course of many years, can block the blood flow and lead to a heart attack or failure (Beckerman, 2015). According to WebMD, heart disease is usually first diagnosed through an electrocardiogram (EKG) which records electrical activity in the heart (Understanding Heart Disease). Now that we have discussed what heart disease is, who is at risk? The short answer is anyone. However, we can look at certain risk factors and historical data to get a better understanding of who heart disease is more likely to affect. First, a person with high blood pressure, high blood cholesterol, or someone who smokes are among the leading candidates for CAD and other heart diseases (CDC, 2021). Heart disease also is slightly more likely to occur in males than females, 21.6% and 19.5% respectively, of all deaths in the United States in 2020 (CDC Wonder, 2022). As we can see from these large numbers, heart disease is an issue that affects a lot of people, and one that needs special attention.

Our goal is to provide organizations such as the “U.S. Department of Health and Human Services”, specifically the Centers for Disease Control and Prevention, with a model that can be used to predict whether or not a patient has heart disease. Within the CDC there are subcommittees which focus on a certain area of expertise. The main group for our focus is the National Center for Chronic Disease Prevention and Health Promotion (NCCDPHP). This information can then be shared amongst health professionals and physicians in hospitals to save lives and prevent heart disease from going unnoticed. Because our model will not be perfect, we do not suggest ignoring testing for patients with a low predicted probability of disease present, however we intend for the model to be used as a way to flag patients that we strongly recommend get tested.

There are current models that help predict the outcome for patients with the most common form of heart disease coronary artery disease. Logistic and Cox regression methods as well as Markov models exist to suggest treatment plans to physicians (Amiri & Kelishadi, 2012). While risk predictive models have the ability to help overall patient health, they are reliant on the knowledge of an already existing heart disease. Because our goal is to help detect heart disease, however, we will look beyond these models. The best current model, according to research, uses Chi-Square and principal component analysis with random forests (Gárate-Escamila et al., 2020). There was a 98.7% accuracy for the Cleveland Dataset used in this project, although there have been no mentions on the CDC website of use of this model.

DrivenData is an internal sponsor for this project with “cutting-edge practices in data science and crowdsourcing to some of the world’s biggest social challenges and the organizations taking them on” (DrivenData, n.d.). DrivenData provides the problem associated with the dataset shared through the UCI Machine Learning Repository (Statlog Data Set, n.d.).

**Figure 1**

Table

Description automatically generated with low confidence

*Note*. From *CDC Organization Chart*, by Center for Disease Control and Prevention in 2021. The current CDC Director is Rochelle P. Walensky, MD, MPH.

**Figure 2**

Diagram

Description automatically generated

*Note*. From *Chronic Disease Center Organization Profile*, by National Center for Chronic Disease Prevention and Health Promotion in August 2022. Karen Hacker is the Director of CDC’s National Center for Chronic Disease Prevention and Health Promotion (NCCDPHP). The main key individuals for this predictive modeling project are Janet S. Wright, MD, FACC and Mattie Gilliam. Dr. Wright serves as the Director in the Division for Heart Disease and Stroke Prevention. Ms. Gilliam is the Deputy Director of the Division for Heart Disease and Stroke Prevention.

**Business objectives and success criteria**

The objective of the United States Government, more specifically the CDC’s National Center for Chronic Disease Prevention and Health Promotion, is to predict the presence or absence of heart disease within a patient to increase primary prevention. The goal of the CDC is not to provide information such as how long the patient has had heart disease, how severe it is, or any other details outside of whether a patient has heart disease. However, there are still questions that should be considered, both in the construction of the model as well as the deployment of it. These include:

* How long will it take to develop a suitable model? Will this be a sufficient for the CDC?
* How much will it cost to treat patients with positive predictions, both in finances and resources? Will more electrocardiograms need to be produced?
* Should the model be deployed on all people in the United States, or just people that show symptoms? Further, what demographics should be prioritized?
* How often should patients have their data re-run through the model to enhance preventative measures? Every year, every 5 years, etc.?
* How will the CDC use the model to predict heart disease as it pertains to governmental privacy laws? Will they need to ask each individual before doing so?
* After deployment within the United States, how should the model be deployed internationally? Will this affect the resources available domestically? How might foreign policies prevent affect the deployment?

These questions will be dependent on the federal budget and resources, staff availability, hospital compliance, medical protection laws, foreign policies, and patient willingness. The CDC must act in accordance with the laws in place, both foreign and domestic, and determine how much personnel it can assign to this task. Further, it is important that the model doesn’t overlook the ethics aspect or have too many false negatives so that the hospitals can deploy it without fear of losing customers and lives.

If the project can be completed while addressing the former concerns, there will be a potential large boost in health among concerned patients. This, ultimately, is the goal of the CDC and is expected to be a benefit on their end. For the individual medical companies that use the model in practice, they can expect to increase profits by having the ability to help patients identify potential disease. The ability to identify potential diseases will provide increased ratings of the governmental agency, both domestically and internationally.

Our project will have business success criteria such as improving the rate of heart disease detection, increasing patient sing-ups for testing by 10% as awareness spreads, and decreasing the cost for hospitals as they can more accurately determine which patients need electrocardiograms. Hospitals, as well as health organizations can determine the success based on these criteria. To understand what success would look like internationally, there should be standards set by public health experts within the CDC to evaluate the results. If our project can achieve these business-level baselines, as well as others set by the CDC, then we will be able to consider it successful by business standards.

**Inventory of resources**

Hardware used in this process will be an always available MacBook Air from the consultant. Programming languages that may assist in the development of our model are RStudio (in R) and Jupyter Notebook (in Python), with limited knowledge. Excel spreadsheets are used for data storage in a .csv file. Electrocardiogram may be used to help test the accuracy of our predictive model.

External personnel in this process include doctors to diagnose heart disease, computer scientists, programmers, and data scientists. These personnel will be able to create a satisfactory model as well as test patients to evaluate the accuracy of our model.

Data is provided from the Cleveland Heart Disease Database consisting of UCI Machine Learning Repository under the “Statlog (Heart) Data Set”. This can be found at <http://archive.ics.uci.edu/ml/datasets/statlog+(heart)> with the heart.dat and heart.doc files being the ones of interest. An in-depth list of the features is provided by the DrivenData API at <https://www.drivendata.org/competitions/54/machine-learning-with-a-heart/page/109/#features_list>.

**Requirements, assumptions, constraints, and RESOLVEDD Strategy**

There are several requirements that need to be met for this project to run successfully. The target group is the Centers for Disease Control and Prevention (CDC) as well as hospitals that help identify and combat heart disease for its patients. For this project, we need it to be completed as quickly as possible without overlooking any steps that could compromise its effectiveness. Ideally, the project would be finished in 8 weeks, and the model deployed soon after if the results are satisfactory. The model should be compared against new data collected to ensure that there is not overfitting done within the training dataset that affects the productivity of the model. The model needs to have a false negative rate of less than 5% for it to meet the accuracy requirement. The project needs to include a well-designed plan for the deployment of the model once it is completed. The CDC as well as hospitals should have easy access to the model, not only on first deployment but also with new updates. The project should include a reevaluation every two years to ensure its dependability. Between the reevaluations, the organization should continue to collect new data and use it to better train the model. In order for the project to meet all requirements, it should be easily reproduced through the same processes and obtaining the same resulting model. Otherwise, there is an issue with the reliability of the project. The government will need to decide on the security requirements for this model as patients misusing it could obtain incorrect information and assumptions about their health. Because of medical privacy laws such as HIPPA, there are legal requirements that the project must meet in data collection. One example of these restrictions is that the patients whose information is being collected have full knowledge about the use of it. The project begins with data covering only patients from the United States, however, if the government can obtain clearance to collect data and share the resulting model internationally, then the legal restrictions of those countries need to be considered. The project needs to have a transparent timeline of when the data was collected and report new updates so that they can be used immediately for the best possible chance of combatting heart disease in its early stages.

This project includes several assumptions across all stages of its development. This project assumes equally likely presence of heart disease for people in all states and countries. That is, geographical location does not affect the probability of heart disease within a patient. Another assumption is that the data which was collected in 1988 applies equally to the heart disease in the present. Also, we are assuming that the data collected over 30 years ago is as accurate as data collected now. There is an assumption on the availability of data as we are going to need more data collected in the future to allow for better models to make predictions. Another assumption that we are making as this project progresses further is that we will be able to collect data in foreign countries; that is, that there will not be issues with data governance and privacy. This will allow for our model to help lives everywhere rather than just domestically. Because of the descriptive features, we are assuming that certain medical tests and devices such as electrocardiograms are readily available or affordable for everyone. Additionally, we are assuming that we have the financial resources to treat patients with a positive heart disease presence. Finally, we have the assumption that organizations will not use the prediction as definite result as some test patients may be near the threshold.

Now that we have discussed requirements and assumptions, we want to also go over potential constraints for this project. The Government will place a timeline on the project completion, likely around 8-10 weeks. They will also decide how much funding this project will receive. There are legal constraints in place for international data collection and information sharing. Further, domestic laws such as HIPPA laws limits the amount of data that can be collected as participants have to give informed consent. The safe harbor method of privacy laws also “requires covered entities or business associates to remove all 18 identifiers of PHI from data in order to ensure that the data cannot be traced back to one person” (HealthITSecurity, 2021). There may be some resource constraints of equipment in certain places that does not allow for patients to get certain tests. There is no password required to access the dataset and it has been converted to a CSV file which is accessible in most programming languages. All information about the data is available, and so there are no constraints regarding accessibility.

Along with the constraints listed above, project designers need to also consider ethical concerns during the completion of the project. In order to maintain a practice that attempts to avoid any ethical dilemmas, we will use the RESOLVEDD Strategy outlined in Figure 3 below (Kemell & Vakkuri, 2019).

**Figure 3**

*RESOLVEDD Strategy*

Diagram

Description automatically generated

*Note*. From “Implementing AI Ethics in Practice: An Empirical Evaluation of the RESOLVEDD Strategy”, by Vakkuri and Kemell (2019), *International Conference on Software Business*.

***Review* –** The main purpose of this project is to help accurately predict which patients have heart disease based on a number of factors. Heart disease is extremely prevalent as the “leading cause of death for men, women, and people of most racial and ethnic groups” (CDC, 2021). Over 7% of adults aged 20 or older in the United States have coronary artery disease, the most common form of heart disease (CDC, 2021). Because of these unfortunate statistics, we feel the need to develop a model that will allow us to allow for more patients with heart disease to live healthier lives, including detecting heart disease in those who otherwise may not have discovered it.

***Estimate* –** While there do not appear to be many ethical concerns with the prediction of heart disease, one main conflict is that false negative can impact a patients’ likelihood of receiving care. That is, if the model predicts an absence of heart disease when it is truly present, a patient may not seek the necessary help.

***Solutions* –** One possible solution to the problem is to lower the threshold in the confusion matrix to decrease the number of negative predictions. Another solution is to emphasize that our model is used as a complementary tool for heart disease diagnosis, and the predictions are not taken as definite.

***Outcomes* –** A probable outcome of the first problem is that there are more positive predictions, and fewer negative predictions. Because of this, the number of false positives will increase, but these are less harmful than the false negatives we are decreasing. A consequence of this action will be that more patients will believe they have heart disease when they do not. A hopeful outcome for the second solution is that patients and physicians do not solely rely on the model, especially in the case of a negative result, but instead seek further testing. If not, there will be a consequence of over-relying on the model with human lives at stake.

***Likely Impact* –** The likely impact of the first main solution is the subsequent decisions that may follow predictions. There will be fewer patients who ignore the possibility of heart disease which is good for saving lives. However, this is likely to include an increase in financial resources needed as more patients who do not have heart disease will require further testing and care. The likely impact of the second main solution is that the model will not be used as the only tool for diagnosis. If a patient has a positive prediction, then extra testing is urgent. With a negative prediction, doctors should still perform tests if they have any suspicion of heart disease within a patient. When used as a complementary tool, the model can provide an immense amount of help for diagnosing without the risk of causing a patient to not receive the necessary help.

***Values* –** Each solution has values that it upholds and violates. From the first solution, we are upholding the values that human lives are most important, and thus lowering the false negative rate (FNR). However, this solution also violates values by unnecessarily installing fear into patients that they have heart disease as the false positive rate (FPR) increases. From the second solution, we are upholding the value that many people share of trusting human perspectives over machines. Also, we are not accidentally misinforming patients without a second opinion. However, given our recommendation, patients may feel unable to fully place their trust in the model as we are not relying on it solely ourselves.

***Evaluate* –** Both of the proposed solutions for this project appear to provide positive value. Using the first main solution, we can correctly identify more patients with heart disease and provide them with the necessary care. From the second solution, we are able to allow for the model to be a complementary tool. Overall, these are both effective solutions for our project.

***Decide* –** While these solutions are not mutually exclusive, the best thing to do from a business perspective is to lower the threshold of the confusion matrix. This will allow for more patients to be correctly identified of having heart disease, and more resources can be allocated to fight it. Recommendation for the model to be a complementary tool is important but minimizing the patients with false negative predictions takes precedence.

***Defend* –** As the false negative rate decreases with our main solution, the number of false positive predictions will increase. This will cause some unease amongst many people which the CDC will have to answer questions for. Along with the mental side, there will also be more financial resources poured into equipment to test for heart disease. This will cause some stir as Government spending always does. Despite these objections, identifying disease and allowing more people to gain care when needed trumps all those drawbacks. We are able to allow for more correct positive predictions, which makes up for the additional false positives. As for the increase in resources needed with more positive predictions, this is something that our government is going to have to decide for allocation amounts. However, we believe that all needed funding should be given as heart disease is a major issue affecting health. Thus, our main solution is an effective one for this project.

**Risks and contingencies**

As is the case with any predictive model, there are some risks that need to be considered. This model has the potential to predict disease, ultimately saving human lives. However, if we are not careful then we can run into issues of funding, biases, resources, and harmfully incorrect predictions. Despite the risks, there are some contingencies that can help to allow for our model to exist and help detect disease. The risks and contingencies identified are displayed in the following table:

**Table 1**

*Risks and Contingencies Table*

|  |  |
| --- | --- |
| Risk | Contingency |
| Other countries complete a predictive model first. | Better predicting heart disease is a global health issue that should be collaborative, not a competitive one. If another country is attempting to build a similar model, our government should work with them rather than against them. |
| There may not be enough equipment to run the increase in necessary electrocardiograms. | The U.S. government can provide a larger budget for public health that will help cover production and shipment costs. |
| U.S. government may not fund the project development if results do not show a large increase in heart disease identification. | We can rely on health experts to ensure that our model is efficient and satisfactory. Further, we can choose only the necessary features to ensure more control. |
| False negatives could cause patients to not get tested and the care needed | We will make sure to adjust the model to provide extra caution on not allowing for too many false negatives. As false negatives are more harmful than false positives, we will weigh these as such within the model. Also, we stress that a negative prediction does not disqualify a patient from getting an EKG test. |
| The model may favor certain demographics over others. | We will take extra care and consideration for possible biases and discrimination that our model could be furthering. We want to ensure that our model predicts well for all people. |
| How might privacy laws such as HIPPA affect the ability to collect and use data? | Personal data privacy laws need to be thoroughly investigated as more availability to data is preferrable, but we can still gain meaningful knowledge from data acquired with informed consent. |

**Terminology**

No prior glossary of terminology was provided by DrivenData.

**Business Glossary**:

Cholesterol: A substance made in the liver that is a structural component of cell membranes and helps regulate cell function.

Cox Regression: A model that tests how risk factors affect the time of a disease occurring.

Electrocardiogram (EKG): Recording of the heart’s electrical activity.

Heart Attack: A medical emergency, sometimes fatal, in which blood flow is blocked causing sections of the heart muscle to not receive enough oxygen.

Logistic Regression: A model that tests how the odds of a disease being present are affected by a risk factor.

**Data Science Glossary**:

Accuracy: How often the predictions match the true answer. This can be found by dividing the total number of correct predictions by the total number of predictions.

Analytical Base Table (ABT): Table used for building models and predicting the outcome of an instance.

API: Application program interface.

Chi-Squared Test: A test producing a statistic to evaluate how well a model fits the data.

CRISP-DM: A methodology for predictive analytics that is a “hierarchical process model, consisting of sets of tasks described at four levels of abstraction (from general to specific): phase, generic task, specialized task, and process instance”. (Chapman et al., p. 6)

Excel: Microsoft spreadsheet to host data tables with the ability to perform some computation.

Exploratory Data Analysis: “Looking at data to discover relationships not previously detected.” (Two Crows Consulting, 2022)

Feature: A measurable piece of data input that can be used in predictive modeling

Jupyter Notebook: Python-based environment to perform exploratory data analysis.

Markov Model: A model describing the probability of each possible event occurring as a result of the previous state.

Predictive Modeling: A process used to predict futures events and outcomes by analyzing patters in given data

Principal Component Analysis: Process that allows for the summarization of large datasets into smaller tables for easier visualization and understanding.

RStudio: R-based environment to perform exploratory data analysis

**Data mining goals and success criteria**

The intended output of this project is a predictive model that can accurately place patients into binary categories of having the presence of heart disease or absence of heart disease. Thus, the goal of the data mining process is to allow the CDC and health organizations predict heart disease presence based on a set of features. Using the dataset at our disposal, we want to train the model through the ε0 bootstrap process which is the preferred process for datasets of our size. Using historical data, we can predict future existences of heart disease by allowing our model to compare the data values of test patients against the former patients for each feature. Because we already have our dataset at our disposal, our predictive model should not take more than two weeks to build, evaluate, and adjust. Over time, as new records become available, we will want to readjust our model to have the best fit. At the beginning, we should incorporate the new data as every 500-1,000 records become available. This is because of the small size of our starting dataset. Once our continuous dataset grows significantly larger, we can start to spread out the time in between model reconstruction. We do not need to revisit the model often to check whether the current societal state affects the presence of heart disease because our features are not based around that. This process should be done quickly so that we are not wasting time predicting heart disease presence.

A successful model in our case will be able to accurately predict which patients have heart disease, and thus should proceed with further testing. Because we want to minimize the number of false negatives, we can lower our threshold of the confusion matrix; it is more harmful to incorrectly predict a patient does not have heart disease when they do than vise-versa. Based on the size of our dataset and the number of ε0 bootstrap iterations, a false negative rate of less than 5% would make the model a success. As we move forward, we can use more historical data to improve the accuracy. Ideally the FNR, false negative rate, would be 0%. However, to achieve this we may have lowered the threshold too low, and thus increase the false positive rate substantially. While false negatives are more severe than false positives, too high of a false positive rate would signal an insufficient model that merely is predicting patients to have heart disease at a much higher rate than the societal truth. Further, keeping the false positive rate from getting too large will allow for the easier deployment of our model as there is less likely to be an EKG test shortage or funding issue. Because we are using a ε0 bootstrap process, the average of the individual performance measures of the models is how we will be able to determine the overall performance of the model. Success criteria for this model also involves the accessibility of it. The model should be easily used and shared in Jupyter Notebook.

**Project plan/ Order of tasks**

This project is estimated to take at least 37 days. Phase One which is about obtaining and preparing the data is expected to take 5 days. Phase Two which focuses on connecting the data mining goals to the business goals is expected to take 12 days. Finally, Phase Three which covers creating the predictive model has a 20-day estimate. In the Gantt Chart, we can see a more thorough breakdown of each phase and their expected time consumption. The Gantt Chart is below:

**Chart, timeline

Description automatically generated**

|  |
| --- |
| **Data Understanding** |

**Initial data collection report**

The dataset accessed through UCI’s Machine Learning Repository includes a data file which was converted into a .csv file in Excel. There is also a word processing document with descriptions of the dataset, also accessible through the UCI Machine Learning Repository. Because the data comes from a 1988 collection, the predictions have the opportunity to be outdated. However, these attributes still affect patients, and so we feel comfortable using the recorded data.

There are thirteen attributes in the dataset, previously extracted from a larger data set of 75 attributes that help to define a patient. Among these are age, sex, serum cholesterol in mg/dl, resting electrocardiographic results, and number of major vessels colored by fluoroscopy. The attributes have varying types, namely: real, ordered, binary, and nominal. The variable to be predicted, or the target feature, is the absence (1) or presence (2) of heart disease. We will consider all attributes in our model to provide the most accurate predictions of the presence of a heart disease. There are no missing values in the dataset across the 270 observations; thus, we will not have to fill in any missing cells. The file was loaded into Jupyter Notebook, in Python, to prepare the data further and perform analysis.

**Data description report**

**File: house.dat** converted to **house.csv**

As stated in the above section, our data is collected from a study on heart disease and made public through the UCI Machine Learning Repository. The public dataset contains a data file which we converted into a csv file in Excel as well as a document outlining the list of features. The dataset contains 270 records of 14 features, thirteen descriptive features and one target feature. The thirteen descriptive features are broken up into real, ordinal, binary, and nominal attribute types.

There are six descriptive features in the dataset that are real attributes. The real attributes in this dataset consist of age, resting blood pressure, serum cholesterol, maximum heart rate achieved, oldpeak, and major vessels colored. Age (type: int) refers to the age of the patient in years. Resting blood pressure (type: int) refers to the patients’ resting systolic blood pressure in mmHg. Serum cholesterol (type: int) refers to the patients’ serum cholesterol level in mg/dl. Maximum heart rate (type: int) refers to the patients’ maximum heart rate level achieved in beats per minute. Oldpeak (type: float) refers to ST depression which is an abnormality finding on an electrocardiogram. Major vessels colored (type: int) refers to the number of major vessels colored by fluoroscopy, with values ranging between 0 and 3. Summary statistics for each of the real attributes can be found in Table 2 below. We can use the minimum, maximum, means, standard deviations, medians, and quartiles to get a better understanding of the distributions for each feature. The age feature has the range of 29 to 77 years which shows that we only have data for adults, and the mean age of over 54 suggests an older sample population. Because of the large ranges and variances for some of the attributes, we want to ensure that we gather more data as we move forward to obtain a more accurate understanding of the population distributions. Figure 4 shows the histograms for most of the features of real attribute, also adding to our distribution knowledge. As we can see, resting blood pressure and oldpeak have right skewed distributions, maximum heart rate has a left skewed distribution, while age and serum cholesterol levels have somewhat normal distributions.

**Table 2**

*Descriptive Statistics*

Table

Description automatically generated

There is only one ordinal attribute in the dataset which is slope of the peak exercise ST segment. This value is of type integer and is an “electrocardiography read out indicating quality of blood flow to the heart” (DrivenData, n.d.).

There are three descriptive features in our dataset that are binary attributes. These binary attributes are sex, fasting blood sugar, and exercise-induced angina. Sex (type: binary) refers to the patients’ sex; 1 for male, 0 for female. Fasting blood sugar (type: binary) refers to whether the patients’ fasting blood sugar is greater than 120 mg/dl; a 1 denotes it is greater, a 0 denotes it is not greater than the threshold. Exercise-induced angina (type: binary) is essentially whether the patient has exercise-induced chest pain; a 1 denotes there is exercised-induced chest pain, a 0 denotes none recorded. The target feature is also a binary value, with values of 1 and 2 rather than 0 and 1. If the predicted value is a 1, then this indicates the absence of heart disease within that patient. If the predicted value is a 2, however, this indicates the presence of heart disease within that patient. Figure 5 displays the bar graphs for each of the binary attributes that are descriptive features. Figure 6 displays the bar graph for the number of patients where heart disease is present or absent. The bar graphs give us a conceptual picture of relative counts for each of the binary features. For sex, there are 87 females and 187 males in the dataset. 40 of the patients in the dataset have fasting blood sugar levels over 120 mg/dl, while 230 do not. 181 of the patients do not have exercise-induced angina, leaving 89 with it. We can also achieve the number of patients in our training set for heart disease: presence in 120 patients, absence for 150 patients.

Finally, there are three descriptive features of the nominal type. These attributes are chest pain, resting EKG, and thal. Chest pain (type: int) has four possible values, 1-4, for the four types of chest pain. Resting EKG (type: int) refers to the resting electrocardiographic results with possible values of 0, 1, or 2. Thal (type: categorical) refers to “results of thallium stress test measuring blood flow to the heart” (DrivenData, n.d.). This attribute has possible values of normal, 3; fixed defect, 6; and reversible defect, 7. Figure 7 displays the bar graphs for each nominal feature. Just as in the binary feature section, the bar graphs give us a conceptual picture of relative counts for each of the nominal features. For chest pain, there are 20 patients with type 1, 42 patients with type 2, 79 patients with type 3, and 129 patients with type 4. For resting electrocardiographic results, 131 patients have a value of 0, 2 patients have a value of 1, and 137 patients have of value of 2. For the results of thallium stress tests measuring blood flow, 152 patients fall under the normal category, 14 have fixed defect, and 104 have reversible defect.

By looking at Table 3 and Figure 8, we can see the correlations between each of the attributes. In Table 3, we have the correlation statistic which shows the exact strength. Figure 8 uses a heat map to give an overview of how strong each correlation is. For the heat map, the lighter the pixelated cell, the higher positive correlation. Thus, we can use this heat map to quickly find relative correlation strengths. Based on Table 3, we can deduce that the strongest correlation to heart disease is the result of thallium stress tests. Thal and heart disease have a correlation coefficient of 0.525 which is moderate. The strongest positive correlation between two descriptive features is with slope of the peak exercise ST segment and oldpeak. This has an even stronger correlation coefficient of 0.610, although this value is also moderate. Most of the correlations between features are weak, as suggested by the low correlation coefficients.

**Data exploration report**

The target feature for this project is the presence, or absence, of heart disease. Thus, data exploration should be taken on how other features relate to this target feature. Our null hypothesis will be that our descriptive features will not be able to provide valuable insight into predicting the target feature, and thus should not be incorporated in the final model. Our alternative hypothesis is that our descriptive features indicate correlation and insight into a prediction of our target feature. First, we started off by exploring the relationship between sex and heart disease presence as this is a distinction made by the CDC for heart disease (CDC Wonder, 2022). Figure 9 displays a bar chart of this data. The horizontal axis shows the sex of the patients, while the vertical axis shows the number of patients in each category. The coloring of the bars differentiates patients with the absence (blue) or presence (orange) of heart disease. This allows us to obtain an understanding for the relative existence of heart disease among each sex in our dataset. As we can see, over half of the patients in our data set that are males have heart disease; 100 present, 83 absent. For females, however, there are over three times as many patients without heart disease as there are with it; 20 present, 67 absent. Because of the start distribution differences, this gives us an idea of what to explore further in our model despite a 0.298 correlation as described in Table 3. The CDC may want to explore how sex affects the probability of heart disease further.

Other features were compared against heart disease as well to further explore their relationship to the target feature. We followed the same process as for sex and heart disease to gain a relational understanding between the number of major vessels colored by fluoroscopy and heart disease, occurrence of exercise-induced angina and heart disease, and finally, the results of thallium tests and heart disease. These can be seen in Figure 10, Figure 11, and Figure 12, respectively. These features were chosen due to their relative correlation with heart disease. In Figure 10, we see that patients in the dataset are more likely to not have heart disease if none of their major vessels are colored by fluoroscopy, 120 to 40. However, the patients with any of the major vessels colored by fluoroscopy are more likely to have the presence of heart disease. This can be used as a quick measure to help determine a patients’ heart disease status. As the number of major vessels colored increases, the percentage of those patients with heart disease also increases. This helps explain the 0.455 correlation coefficient found in Table 3 between number of major vessels colored and the presence of heart disease. In Figure 11, we notice that a patients’ chance of heart disease presence is largely different depending on the existence of exercise-induced angina for them. Patients in this dataset who have exercise-induced angina are almost three times as likely to have heart disease than to not. Among the subpopulation of patients without exercise-induced angina, the number with heart disease is less than half of those without it. This is a large spread, justifying the 0.419 correlation coefficient. Figure 12 is the one that will be able to show us the most as it pertains to predicting heart disease. This is because Thal is the most correlated feature with the target feature with a coefficient of 0.525. Using Figure 12, patients with normal thallium stress tests appear to be much more likely to be heart disease-free; 119 to 33. Patients with a thallium stress test result showing a reversible defect experience the opposite phenomenon; that is, higher chance of having heart disease than not. The value counts for this subpopulation in our dataset are 25 absent and 79 present. It is harder to gain knowledge about the relationship between heart disease presence among patients whose thallium stress test results showed a fixed defect. This is because our subpopulation only totals 14 individuals, 8 with heart disease and 6 without it. Despite the need for a larger sample size among one of the subpopulations, we can see clear distinctions in the presence of heart disease for patients with normal results versus those with a reversible defect.

Now that we have formed several visualizations with features that have moderate positive relationships to heart disease, we want to do the same with a feature that is negatively correlated to the target feature. The maximum heart rate of a patient, which has a correlation factor of -0.419 with heart disease presence, is the feature that we will use in our next exploratory visualization. In Figure 13, we use a box plot to show the distributions of maximum heart rate for each of the subpopulations of patients with (2) or without (1) heart disease. While there is some overlap, we can use this plot as a complementary tool to help predict the presence or absence of heart disease. Also, it is important to look at a feature with a very low correlation coefficient, near zero. The one that we will use is fasting blood sugar levels. The cutoff used in the dataset is 120 mg/dl which, according to Figure 14, does not do a sufficient job at differentiating between patients with or without heart disease present. Based on the bar graph, it appears that the patients in our dataset with fasting blood sugar levels above and below that threshold have the same probabilities of heart disease. The only noticeable difference is the scales of each bar, due to the size counts of each subpopulation, but the relations appear similar. Because of this, we will reject the null hypothesis that each of the features in our dataset will not do a sufficient job at predicting the presence of heart disease. However, as the model is built, the team will want to consider running tests and only using features that have significant predictive power on the target feature. RStudio and Jupyter Notebook, in R and Python respectively, can be used to perform tests.

**Data quality report**

There are no data quality issues relating to invalid data in our dataset. Using Python, we have discovered that there are no missing values in our dataset. The non-null count is 270 for all of the features. Thus, we do not need to worry about the handling missing data section of a data quality report. In Table 4, there is an analytics base table (ABT) that separates the features into continuous and categorical. For each of the continuous features in Table 4a, we have reported the following values: count, cardinality, minimum, 1st quartile, mean, median, 3rd quartile, maximum, and standard deviation. The missing values percentage was omitted due to their being no missing values in the dataset. There appears to be a good spread and representation for each of the continuous features, although some outliers do exist for the maximum heart rate, resting blood pressure, and serum cholesterol levels. Potential handling strategies for these outliers is shown in Table 5. The clamp transformation ranges are based on the interquartile ranges (IQR). The lower value is 1st quartile minus 1.5 times IQR, and the upper value is 3rd quartile plus 1.5 times IQR. For each of the categorical features in Table 4b, the following values were reported: count, cardinality, mode, mode frequency, mode percentage, 2nd mode, 2nd mode frequency, and 2nd mode percentage. Because the features of sex, fasting blood sugar, and exercise induced angina have cardinalities of 2, the values included in their first and second modes account for 100% of their records. All of the modes for categorical features account for a majority of the records, over 50%, except for chest pain type which has a 47.8% plurality. Between the two feature types in Table 4, there do not appear to be any issues with irregular cardinalities.

The dataset that we are using is very clean due to proper collection and reporting measures. Because this dataset only includes 13 attributes of a larger set of 75, any duplicate and similar meaning attributes have already been removed. There are no spelling or formatting issues due to their being only numerical values reported for binary and categorical features. For example, possible values of the thallium stress test results are 3, 6, and 7 to represent normal, fixed defect, and reversible defect, respectively. The units of measurement are consistent within each feature. The fasting blood sugar levels are measured in mg/dl, serum cholesterol levels are measured in mg/dl, age is measured in years, and maximum heart rate is measure in beats per minute for each record. Although the data appears to be accurate, collected properly, and without noise, the CDC may want to consider referencing another dataset or collecting one of their own. This will help minimize the possibility of falsely reported values affecting the key statistics for the features.

**References**

Amiri, M., & Kelishadi, R. (2012). Comparison of models for predicting outcomes in patients with coronary artery disease focusing on microsimulation. *International Journal of Preventive Medicine*, *3*(8), 522–530. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3429798/

Beckerman, J. (2015, August 28). *What Is Coronary Artery Disease?* WebMD; WebMD. https://www.webmd.com/heart-disease/coronary-artery-disease

*CDC Organization Chart*. (2019). https://www.cdc.gov/about/organization/orgchart.htm

Centers for Disease Control and Prevention, National Center for Health Statistics. About Multiple Cause of Death, 1999–2020. CDC WONDER Online Database website. Atlanta, GA: Centers for Disease Control and Prevention; 2022. Accessed September 9, 2022.

Centers for Disease Control and Prevention. (2021, September 27). *Heart disease Facts*. Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/facts.htm

Chapman, P., Clinton, J., Kerber, R., Khabaza, T. Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0*: Step-by-Step Data Mining Guide. SPSS, Copenhagen.

*Chronic Disease Center Organization Profile | CDC*. (2022, August 30). Www.cdc.gov. https://www.cdc.gov/chronicdisease/center/organization/index.htm

DrivenData. (n.d.). *Warm Up: Machine Learning with a Heart*. DrivenData. Retrieved September 12, 2022, from https://www.drivendata.org/competitions/54/machine-learning-with-a-heart/page/109/#features\_list

Gárate-Escamila, A. K., Hajjam El Hassani, A., & Andrès, E. (2020). Classification models for heart disease prediction using feature selection and PCA. *Informatics in Medicine Unlocked*, *19*, 100330. https://doi.org/10.1016/j.imu.2020.100330

*Glossary of data mining terms > Two Crows Consulting*. (n.d.). Twocrows.com. Retrieved September 12, 2022, from http://twocrows.com/data-mining/dm-glossary/#anchor315003

HealthITSecurity. (2021, October 15). *De-Identification of PHI According to the HIPAA Privacy Rule*. HealthITSecurity. https://healthitsecurity.com/features/de-identification-of-phi-according-to-the-hipaa-privacy-rule

*UCI Machine Learning Repository: Statlog (Heart) Data Set*. (n.d.). Archive.ics.uci.edu. Retrieved September 8, 2022, from http://archive.ics.uci.edu/ml/datasets/statlog+(heart)

‌*Understanding Heart Disease*. (n.d.). WebMD. https://www.webmd.com/heart-disease/understanding-heart-disease-symptoms

Vakkuri, V., & Kemell, K. K. (2019). Implementing AI Ethics in Practice: An Empirical Evaluation of the RESOLVEDD Strategy. In *International Conference on Software Business* (pp. 260-275). Springer, Cham.

‌

**Figure 4**

*Histograms of Features with Real Attributes*

Chart, histogram

Description automatically generated

**Figure 5**

*Bar Graphs of Binary Features*

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

**Figure 6**

*Bar Graph of Presence of Heart Disease*

Chart, bar chart

Description automatically generated

**Figure 7**

*Bar Graphs of Nominal Features*

Chart, bar chart, histogram

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

**Figure 8**

*Correlation Heat Map*

Chart

Description automatically generated

**Figure 9**

**Chart, bar chart

Description automatically generated**

**Figure 10**

Chart, bar chart

Description automatically generated

**Figure 11**

**Chart, bar chart

Description automatically generated**

**Figure 12**

**Chart, waterfall chart

Description automatically generated**

**Figure 13**

**Chart, box and whisker chart

Description automatically generated**

**Figure 14**

**Chart, bar chart

Description automatically generated**

**Table 3**

*Correlations Between Attributes*

Table

Description automatically generated

**Table 4**

*Quality Data Report – ABT*

(a) Continuous Features

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature | Count | Card. | Min | 1st Qrt. | Mean | Median | 3rd Qrt. | Max | Std. Dev. |
| Age | 270 | 41 | 29.0 | 48.0 | 54.4 | 55.0 | 61.0 | 77.0 | 9.1 |
| Resting BP | 270 | 47 | 94.0 | 120.0 | 131.3 | 130.0 | 140.0 | 200.0 | 17.9 |
| Serum Cholesterol | 270 | 144 | 126.0 | 213.0 | 249.7 | 245.0 | 280.0 | 564.0 | 52.7 |
| Max. Heart Rate | 270 | 90 | 71.0 | 133.0 | 149.7 | 153.5 | 166.0 | 202.0 | 23.2 |
| Oldpeak | 270 | 39 | 0.0 | 0.0 | 1.1 | 0.8 | 1.6 | 6.2 | 1.2 |

(b) Categorical Features

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature | Count | Card. | Mode | Mode Freq. | Mode % | 2nd Mode | 2nd Mode Freq. | 2nd Mode % |
| Sex | 270 | 2 | 1 | 183 | 67.8 | 0 | 87 | 32.3 |
| Chest Pain | 270 | 4 | 4 | 129 | 47.8 | 3 | 79 | 29.3 |
| Fasting Blood Sugar | 270 | 2 | 0 | 230 | 85.2 | 1 | 40 | 14.8 |
| Resting EKG | 270 | 3 | 2 | 137 | 50.7 | 0 | 131 | 48.5 |
| Exercised Induced Angina | 270 | 2 | 0 | 181 | 67.0 | 1 | 89 | 33.0 |
| Thal | 270 | 3 | 3 | 152 | 56.3 | 7 | 104 | 38.5 |

**Table 5**

*Quality Data Report – Data Quality Issues*

|  |  |  |
| --- | --- | --- |
| Feature | Data Quality Issue | Potential Handling Strategies |
| Max. Heart Rate | Outliers (low) | Clamp transformation (manual: 83, 216) |
| Resting BP | Outliers (high) | Clamp transformation (manual: 90, 170) |
| Serum Cholesterol | Outliers (high) | Clamp transformation (manual: 112, 381) |