Heart Disease, You Can Run But You Can't Hide!

An examination of the famous Cleveland Clinic Heart Disease Data Using Data Mining

Christian Dobish
Blake Tindol
Leonard Lasek

Final Project
IST 707
Norma Grubb, Tuesday 8pm

Introduction

History

For many decades, Dr. Ancel Keys championed the theory that dietary cholesterol and saturated fats were the primary culprits behind heart disease. His Seven Countries Study in the mid-20th century argued that populations consuming higher amounts of saturated fats had more incidences of coronary heart disease. This led to widespread dietary recommendations to limit saturated fat and cholesterol intake. The problem with this study was that it was completely bogus (full of all sorts of bias, poor data collection, confounders and didn't even consider smoking!) yet remained the law of the land for over 50+ years.

In more recent years, Dr. Robert Lustig, pediatric endocrinologist from UCSF, and other researchers have presented compelling evidence that links the consumption of added sugars, particularly fructose, to metabolic diseases, the most prevalent being heart disease. According to this contemporary viewpoint, excessive sugar intake leads to insulin resistance, increased fat production in the liver, and inflammation, all of which are major risk factors for heart disease[1].

Business Case

While our understanding of root causes of heart disease continues to evolve, it remains the leading cause of death globally, and early detection is our best defense. Because disease symptoms can sometimes be elusive, it's a tall order to rely on human faculties to spot patterns in multidimensional datasets. It's a great opportunity to leverage the power of machine learning data mining techniques to help with early detection. Data mining offers a great opportunity to unveil previously undetected patterns to help aide in the early identification of heart disease. This not only benefits the patient's personal health outcomes but also makes the health care system more efficient and better suited to handle this unfortunate epidemic of heart disease.

Literature Review

The literature supporting the work on cardiovascular disease is vast. The Cleveland dataset we have chosen has been very popular and used numerous times since it was originally collected. Although our models to be used for comparison do not have significant differences from the existing methods, it is still meaningful to observe their adaptability and performance in the real-world setting.

In research by Shah et al. (2020)[2], they focused on creating a predictive model for cardiovascular disease using machine learning techniques on the Cleveland heart disease dataset, which comprises 1025 records with 14 features each. Several supervised classification algorithms were tested, such as naive Bayes, decision tree, random forest, and kNN. Among these, the kNN method proved most accurate, achieving a 90.8% success rate.

In another paper by Acharya et al. (2017)[3] also built predictive supervised models using the Cleveland Heart Disease dataset. The system uses a decision tree classifier to predict the presence or absence of heart disease based on 14 attributes. Interestingly, the model was evaluated using

[1] Teicholz, N. (2015). The Big Fat Surprise: Why Butter, Meat and Cheese Belong in a Healthy Diet. Simon & Schuster
[2] https://link.springer.com/article/10.1007/s42979-020-00365-y
[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5863635/

10-fold cross-validation, and it was shown to have an accuracy of 84.2%. The authors conclude that the system is an effective tool for predicting heart disease, and it can be used to identify patients who are at risk for the disease.

Dataset

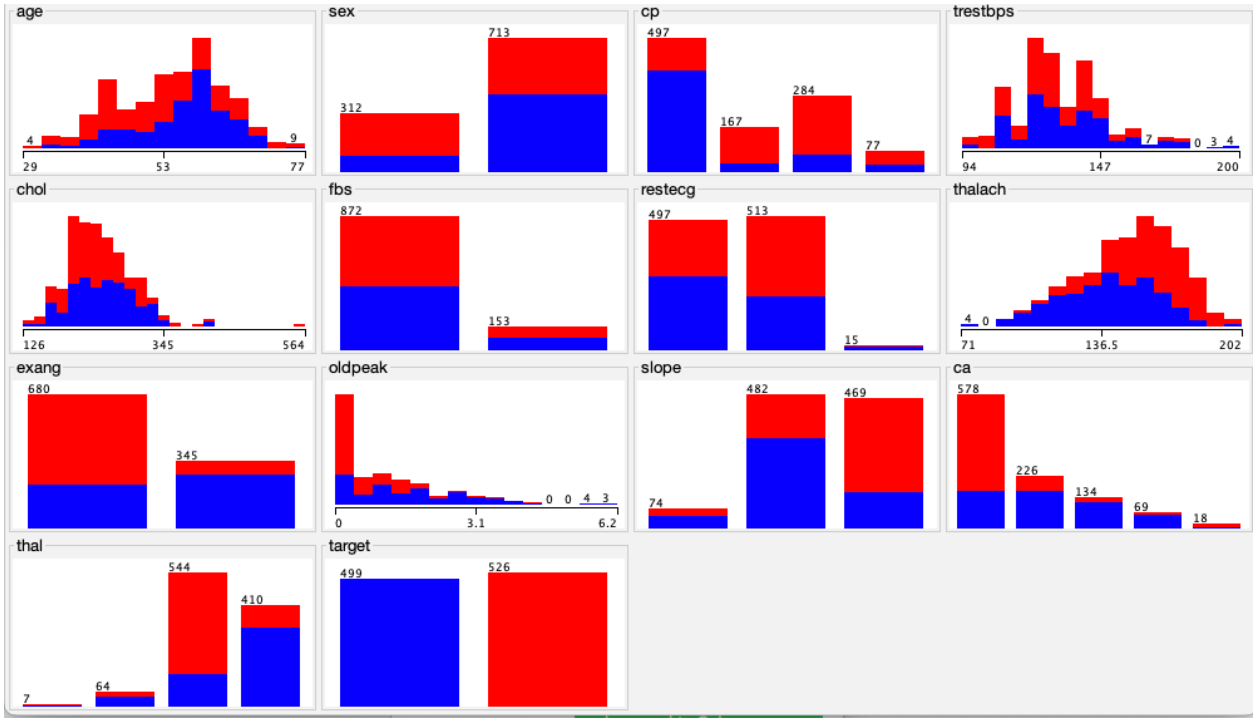
The Cleveland Clinic Foundation heart disease dataset, commonly referred to as the "Cleveland dataset," is one of the most well-known datasets related to heart disease research. It originated from a study conducted at the Cleveland Clinic Foundation in 1988 and has since been widely used in various machine learning and data analysis projects. The original dataset contains over 70 attributes, but the data set we are examining is pared down to 14 attributes. There are 1025 unique records to consider. This paring down of the dataset has become customary over the years as these 14 attributes contribute the most information toward prediction.

Below is a summary of the attributes with corresponding data types.

Parameters	Description	Values
Age	Age in years	Continuous
Sex	Male or female	Nominal 1 = male 2 = female
Thestbps	Resting blood pressure	Continuous (units: mmHg)
Cp	Chest pain type	Nominal 1 = typical type 1 2 = typical angina 3 = non-angina pain 4 = asymptomatic
Chol	Serum cholesterol	Continuous (units: mm/dL)
Fbs	Fasting blood sugar	Nominal 1 >= 120 mg/dL 0 <= 120 mg/dL
Restecg	Resting electrographic results	Nominal 0 = normal 1 = having ST-T wave abnormal 2 = left ventricular hypertrophy
Thalach	Maximum heart rate achieved	Continuous
Old peak	ST depression induced by exercise relative to rest	Continuous
Exang	Exercise induced angina	Nominal 0 = non 1 = yes
Ca	Number of major vessels colored by fluoroscopy	Nominal 0 – 3 range
Slope	Slope of the peak exercise ST segment	Nominal 1 = upsloping 2 = flat 3 = downsloping

thal	Defect type	Nominal 3 = normal 6 = fixed 7 = reversable defect
Num (class var.)	Diagnosis of heart disease	Nominal 0 = no disease 1 = disease

Below, we visually display the data to better understand which fields may provide the most valuable insights for our predictive models. The first thing to note is that the data is fairly well balanced such that the target class variable – i.e. heart disease present in the patient - is split about 50/50. However, there seems to be a clear imbalance in the quantity of males to females. For males the disease/no-disease split is even but for females it favors more females having disease than not. This could play a role in our models. In our pursuit of building predictive models, those attributes which have the most contrast between the target will provide the most information in the models. What we mean is that if every person with diseases had 'thal' defect of type 7 and no person with disease had type 7, then this would provide a lot of information to the model. If 'thal' were evenly split between disease and no disease then it would be no better than a coin flip as to whether it was informative or not and thus not very informative to a model.



Taking a closer look at the plots there is appears to be a strong bias for heart disease with those presenting with higher maximum heart rates. This would make sense as it might indicate the instance of an actual heart attack or heart disease. Also, a person without heart disease we would assume is a healthier individual and thus would have a lower heart rate and the chart bears this out. We see similar skew for 'ca' as well as 'chol'. We will keep an eye out for the role these attributes might play in our predictive models.

Model Summary

We built and compared the performance of four machine learning models: Random Forest, Naive Bayes, kNN and Support Vector Machines. We also explored feature importance to gain insights into the factors contributing to heart disease. We chose these models for various reasons. First, we found they were used most heavily in the literature and would like to try to match or beat the results achieved in those models. Second, with a binary categorical output class variable, the data is set up in such a way that these models are most appropriate.

Naïve Bayes Model
The Naive Bayes algorithm is a probabilistic data mining method based on Bayes' theorem with an assumption of independence among predictors. The core of naïve bayes is based off an equivalence between two conditional probabilities[4] which are inverses of each other – namely, the prior probabilities and the predicted probabilities. The naïve Bayes algorithm will iteratively go through the records and determine all the prior probabilities. These prior probabilities are then updated as it iterates through the records. The algorithm is termed "naive" because it assumes all input variables are independent of each other, which is rarely the case in real-world scenarios.

K-Nearest Neighbors (kNN)
kNN, is another type of classification algorithm where a new untrained record is assigned to the class that is most common among its k nearest neighbors. Because kNN relies on keeping the training data and performs calculations at prediction time, it's sometimes referred to as a "lazy learner." (Weka even stores it in the "lazy" folder) This contrasts with "eager learners" like decision trees, which build a model from the training data and then discard the training data when making predictions. The one main drawback from the kNN method is that because it holds on to the training data for prediction, it can be computationally expensive, especially as the data set being analyzed increases in size and dimension.

Support Vector Machine (SVM)
SVM aims to find the best hyperplane that separates different classes in a dataset such as male vs female, for example. In a two-dimensional space, this hyperplane is simply a line. In higher dimensions, it becomes a plane or some hyper-dimensional object.

The "best" hyperplane is the one that maximizes the margin between the two classes. This is important because the wider margin will generalize better to test data, help mitigate overfitting and be less sensitive to outliers and noise impacting the model. The margin is defined as the distance between the hyperplane and the closest data point from either class. The **support vectors** are the data points that lie closest to the decision boundary (or hyperplane). The closer these data points are to the hyperplane, the harder it is to determine which side of the boundary they belong to.

Random Forest
The random forest classification algorithm is closely related to the decision tree model. A decision tree is a network where internal nodes represent features, branches represent decision rules, and

[4] Bayes Theorem: P(A | B) = P(B | A) * P(A) / P(B)
P(A | B) = Probability of event A given B is true, predicted probability
P(B | A) = Probability of event B given A is true, prior probability

each leaf node[5] represents a class label. A random forest expands on this design by building a collection ("forest") of decision trees. Rather than relying on a single decision tree, it aggregates the results from multiple decision trees to give a final prediction.

Recall that in the decision tree model it relies on the statistic called information gain[6] ("IG"). For each attribute in the dataset, the algorithm calculates the IG that would result from splitting the data based on that attribute. The attribute with the highest IG is selected as the root node. This process is then repeated recursively for each subset of the data created by the split, to create the subsequent nodes of the tree.

In contrast, each tree in a random forest is trained on a random subset of the data by sampling with replacement. This method is known as bootstrapping. While the model also relies on the concept of information gain, it is only considering a random subset of the attributes and not all attributes like the decision tree when selecting the node. This adds two layers of randomness to the model. The random forest will perform this process for a specified number of trees which is a key parameter in the model. As a result of this randomness and tree collection, the random forest model is less prone to overfitting like the decision tree. By averaging the predictions of many trees, the model often achieves higher predictive performance and is less sensitive to noise and outliers.

Predictive Model Summary

We built and compared the performance of three machine learning models: Random Forest, Naive Bayes, kNN and Support Vector Machines. The models were evaluated using metrics such as accuracy, precision, recall, and F1-score. The results are as follows:

Model	Accuracy	Correct	Precision	Recall	F1	Notes
Naïve Bayes	81.8%	838	82.0%	78.6%	89.6%	
SVM	89.9%	921	89.9%	89.2%	90.0%	
kNN	83.9%	860	84.0%	83.9%	83.9%	K = 3
Random Forest	98.5%	1010	98.6%	98.5%	98.5%	Default settings, 3-fold cv

For the Random Forest we used cross validation to help mitigate the potential for overfitting. The other models we used a test/train split.

The Random Forest was the best performing predictive model across all metrics, most notably predicting 1010 out of 1025 instances resulting 98.5% **accuracy**. This is an excellent result.

[5] A leaf-node is the terminal leaf of a particular branch.

[6] Information gain is calculated as the difference between the entropy of the original dataset and the weighted sum of the entropies of the subsets resulting from the split.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 487  12 |   a = 0
   3 523 |   b = 1
```

In reference to the above confusion matrix, ‘a’ corresponds to no disease and ‘b’ corresponds to disease being present.

There were 487 non-disease records classified correctly and 3 records which were positive for disease but classified as non-disease. The ratio of 487 / (487 + 3), or about 99.4% is the **precision** metric for the non-disease category. Similarly, the precision for the disease-positive category amounts to 523/(523+12) or about 97.8%. In total, the weighted average precision for the Random Forest model works out to 98.6%. A high precision for a class means that when the model predicts disease or no disease, it's very likely to be correct. If precision is low then it means the model often mistake disease for non-disease and non-disease for disease.

Recall refers to the number of actual instances of disease or no disease that were predicted correctly. Like precision, recall can be calculated for each element of the class variable. To derive the recall, we look across the rows in the confusion matrix. This contrasts with looking to the columns like we did for precision. Thus, the recall for class ‘a’, the non-disease records, is 487 / (487 + 12), or about 97.6%. Similarly, for the disease positive records we calculated a recall of 523/(523 +3) or about 99.4% for a total weighted recall of the model of 98.5%. Recall emphasizes avoiding false negatives.

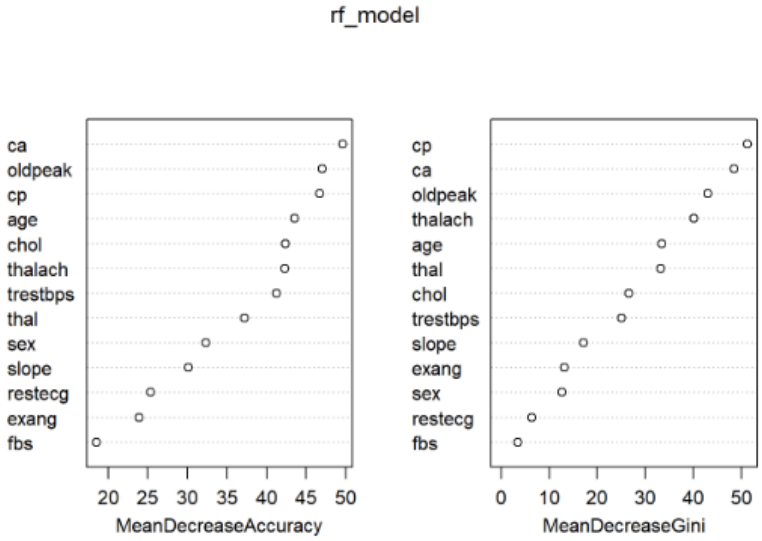
There is a fourth statistic called the F-measure which is a function of precision and recall. The F-measure is a number between 0 and 1, where 1 indicates perfect precision and recall, and 0 indicates neither precision nor recall. It's a way to succinctly represent the balance between precision and recall, giving an overall model performance metric. With an F1 of 98.5% the model is a very strong model.

An important consideration with an accuracy this high is that the model has suffered from overfitting. Overfitting is a phenomenon that is characterized by a model which is built to predict on itself very well but does not model well with new data. As we mentioned before, the random forest has multiple unique characteristics which distinguish it from its cousin decision trees. These features focus on using random selection of attributes and then using averages to build a model. These steps are supposed to mitigate overfitting, but to be sure, we also ran a 3-fold validation as well.

Feature Importance

Using the Random Forest algorithm, we identified the top 5 important features contributing to heart disease prediction. These features are:
1. Chest pain type (cp)
2. Number of major vessels colored by fluoroscopy (ca)
3. ST depression induced by exercise relative to rest (oldpeak)

4. Maximum heart rate achieved (thalach)
5. Age



The top 5 important features contributing to heart disease prediction are consistent with known risk factors and symptoms:

1. Chest pain type (cp): Chest pain is a common symptom of heart disease, and different types can indicate various issues. For example, angina is a symptom of coronary artery disease, making chest pain type a key feature in predicting heart disease.

2. Number of major vessels colored by fluoroscopy (ca): More affected arteries indicate a higher risk of heart disease due to widespread blood flow issues to the heart.

3. ST depression induced by exercise relative to rest (oldpeak): Greater ST depression during exercise compared to rest can signal reduced blood flow to the heart, indicating heart disease.

4. Maximum heart rate achieved (thalach): A lower maximum heart rate during exercise may suggest a less efficient heart and potential heart disease. People with heart disease may experience symptoms at lower heart rates.

5. Age: The risk of heart disease increases with age due to factors like plaque buildup, age-related heart changes, and increased likelihood of other risk factors.

These top 5 features align with established heart disease risk factors and symptoms, making them logical contributors to the prediction of heart disease in the Random Forest model.

Conclusions

We built and compared the performance of four machine learning models for heart disease prediction using the Cleveland Clinic Foundation heart disease dataset. The Random Forest model

achieved an impressive 98.5% accuracy. This result is stronger than the experiments from the literature review. We also identified the top 5 important features contributing to heart disease prediction. These insights can help medical professionals focus on the most relevant factors when diagnosing and treating patients and improve the early detection of heart disease.

	PREDICTED	
ACTUAL	**True Positive:** No disease, correctly predicted	**False Positive:** No disease, predicted incorrectly
	False Negative: has disease, but predicted incorrectly	**True Negative:** Has disease, predicted correctly

The above table helps illustrate the ways in which we can understand the results of the model in a real word context. We can use the results to help us better understand the implications for healthcare industry and patient outcomes and care, among other things.

The emphasis on false negatives becomes clear when we consider the consequences of missing a positive case. In the context of heart disease, a false negative would mean that an individual who actually has heart disease is incorrectly classified as not having the disease. This type of error could lead to a lack of proper medical attention or treatment. We obviously want to minimize this outcome. On the other hand, a false positive is a situation in which a person does not have disease but it is predicted that they do. The associated cost with a false positive might include a person making some decisions to better their health through diet and exercise. Compared to illness, we find the emphasis on false negatives to be much more important.

With a high precision of 98.6%, the healthcare industry can be confident that most of the patients identified by the model as having heart disease likely do have the disease. This means resources, like specialized doctors and medical equipment, can be allocated more efficiently to patients who need them most. On the other hand, a lower precision would mean the healthcare system might waste resources on patients mistakenly identified as having heart disease. As we noted above, this concern pales in comparison to the risks of a false negative. However, if these costs begin to grow, we will know where to look first to address the issue.

The high recall of 98.5% indicates that the model is excellent at identifying those who have heart disease. A missed diagnosis (false negative) in heart disease can have severe consequences, including death. The high recall ensures most of the patients with heart disease are identified and treated accordingly. However, the slight discrepancy between precision and recall highlights that no model is perfect. There remains a slight risk that a patient with heart disease might not be identified or that a healthy patient might undergo unnecessary tests.

These high accuracy numbers demonstrate the reliability and potential of machine learning in complimenting traditional diagnostic procedures. However, it's also a reminder that while these tools can assist, human oversight remains crucial especially when explaining the outcomes and risks to patients.