Predicting and Reducing Risk for Cardiovascular Disease

Erica Smith

December 2019

Abstract

The data used for this analysis is Svetlana Ulianova's "Cardiovascular Disease" dataset found on Kaggle which contains 11 features and 1 target for 70k patients. A classification model was constructed to determine if cardiovascular disease status can be predicted from select clinical and lifestyle data. A cluster analysis was performed to identify commonalities within groups of patients with cardiovascular disease in order to design risk reduction programs targeted to their specific risk factors. The classification model accurately predicted 72% of patients with cardiovascular disease. The cluster analysis found 5 clusters of patients with differentiated needs including smoking cessation, weight loss, and stress reduction.

Motivation

Each year, 1 in every 4 Americans dies from cardiovascular events including heart attack and stroke; millions more will experience a significant decrease in quality of life due to the effects of coronary artery disease and hypertension (CDC, 2019a). Risk for cardiovascular disease includes a combination of clinical factors as well as lifestyle factors such as high cholesterol, sedentary lifestyle, stress, and smoking. When designing outreach and risk reduction programs, it is important to consider that the most successful programs address the *specific risks* of a particular population. For example, one program may address lifestyle factors like smoking and inactivity while another may address genetic high cholesterol.

The purpose of this analysis is to develop a model that can aid in identification of those at risk for cardiovascular disease and classify them into smaller groups with common risk factors. The model output can be used to channel patients to programs to address their specific health promotion needs.

Dataset(s)

The dataset used for this analysis is titled “Cardiovascular Disease” and may be obtained from Kaggle. It is comprised of 70k records containing

- 11 features (age, gender, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol use, and physical activity
- 1 target (cardiovascular disease)

No information is provided about the source or timeframe of the data.

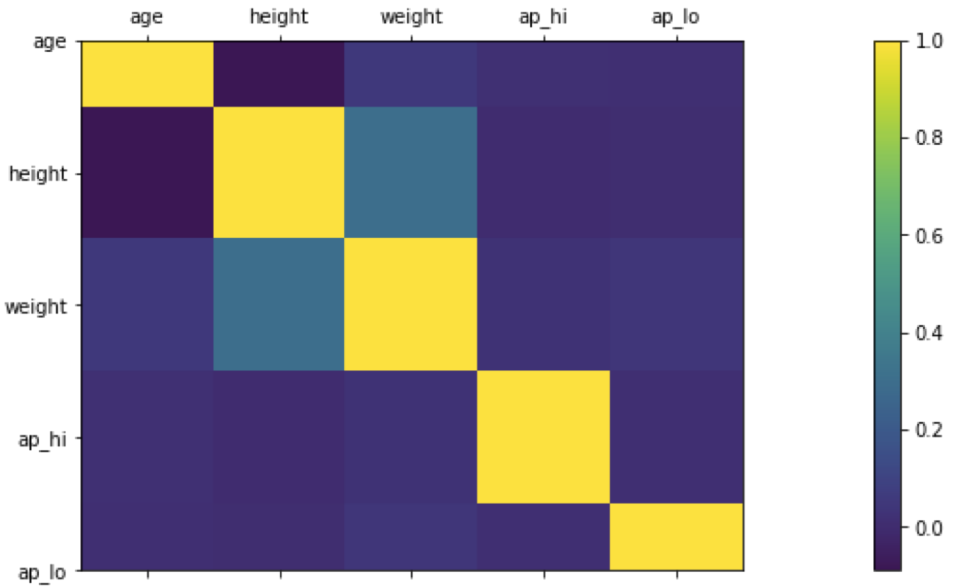
Data Preparation and Cleaning

After reading in the data, it was analyzed for missing and illogical values. No missing data was detected. The dataset did contain 412 records that were filtered out due to one of the following issues (since they are considered illogical given that the minimum age of a patient in the dataset is 29 years):

- Height less than 48 inches (4 feet)
- Weight less than 90 pounds
- Systolic blood pressure less than 50mmHg
- Diastolic blood pressure less than 30mmHg

Data Preparation and Cleaning

A correlation matrix was created to look for and address highly-correlated features. It appears that height and weight are strongly related, which makes sense logically. Therefore, a new variable for BMI was created to normalize the body mass (Weight in pounds x 703 ÷ (Height in Inches²)) (CDC, 2019b)



Data Preparation and Cleaning

Blood pressure features were separated in the dataset. Although the correlation matrix does not show them being highly correlated, systolic (ap_hi) and diastolic (ap_lo) blood pressure readings are usually interpreted together in the clinical setting. To normalize this, the mean arterial pressure (MAP) was calculated for each patient. MAP reflects the blood pressure during a full cardiac cycle and is calculated as ((2 * diastolic reading) + systolic reading) / 3. (Bonsall, 2011)

After cleaning and normalizing the data, the remaining features were age, gender, cholesterol, glucose, smoking status, alcohol use, activity level, BMI, and MAP.

Research Question(s)

This analysis focuses on two research questions:

- Can we use clinical and behavioral data predict cardiovascular disease?
- How can we cluster patients with cardiovascular disease into small groups with common features in order to design programs to address their specific risk factors?

Research Question 1: Methods

The first research question requires a classification analysis. This is because we are using data to predict whether a patient has cardiovascular disease and we have a known response which allows us to assess the model's accuracy.

In this analysis, the features age, gender, cholesterol, glucose, smoking, alcohol use, activity level, BMI, and MAP were used to try to predict cardiovascular disease. Two thirds of the data were used to train the model and the last third was held out for testing. A random seed was used for establishing the test data tuples.

Research Question 1: Findings

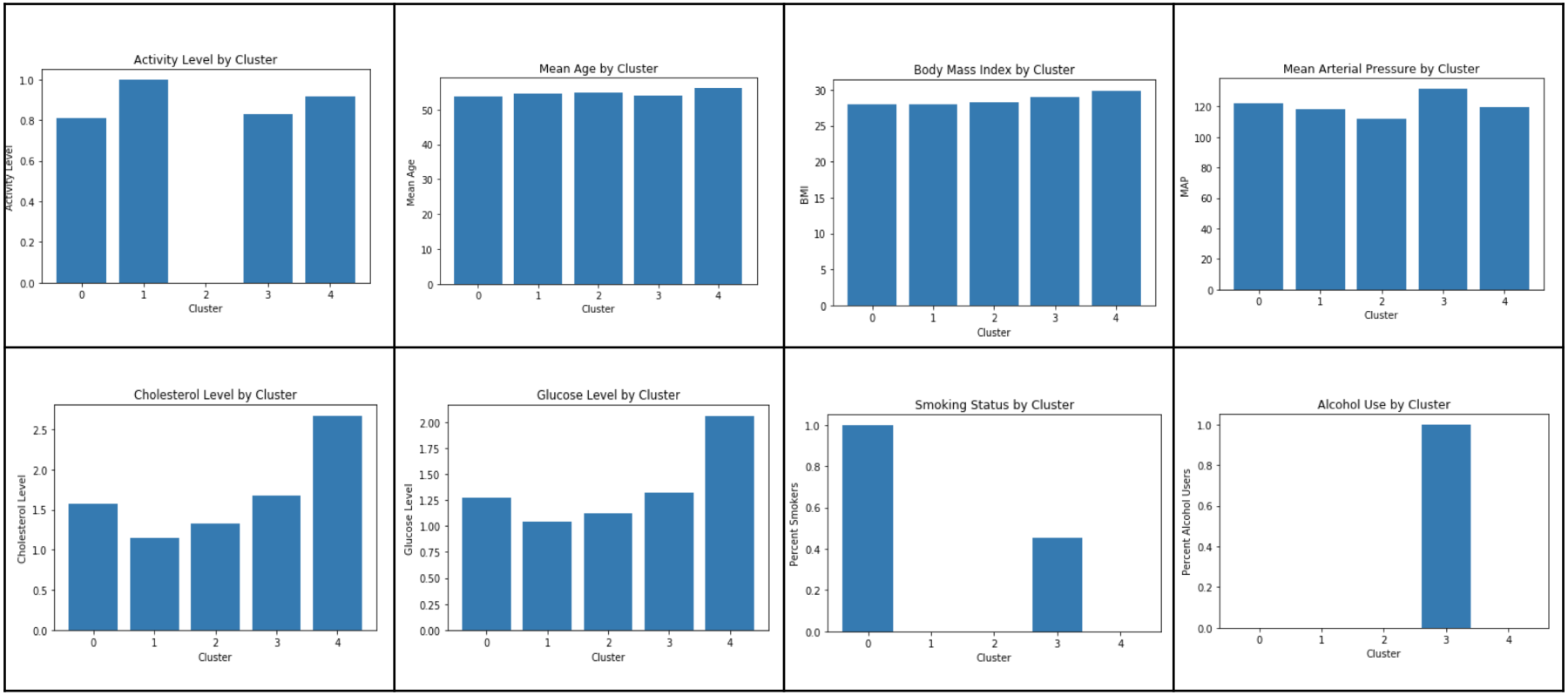
The model accurately predicted 72% of patients with cardiovascular disease. The stability of the model was tested by resetting the random seed several times; each time the prediction remained between 72.0% and 72.9%.

Research Question 2: Methods

The second research question is a cluster analysis because we are using data to partition patients into smaller groups based on commonalities in their risk factors. Since we do not have data to compare to the model's output, this unsupervised learning must be evaluated to ensure it is clinically logical.

For this analysis, the prepped data was filtered to only include patients with known cardiovascular disease (n = 34795). The features age, gender, cholesterol, glucose, smoking, alcohol use, activity level, BMI, and MAP were used in the analysis. Because the features were of mixed types (binary, scaled, continuous), they were normalized using the StandardScaler method. Initially, the analysis was set up to create 10 clusters (k = 10). However, the results were too diverse and the cluster size was reduced to 5.

Research Question 2: Findings



Research Question 2: Findings

Cluster 0	100% of patients in this cluster are smokers. They could benefit from a risk reduction program aimed at smoking cessation
Cluster 1	The majority (52%) of patients fall into this cluster. On average, they have normal blood pressure, BMI, cholesterol, glucose levels and activity levels, and they don't drink or smoke.
Cluster 2	100% of patients in this cluster have very low activity levels. Their other clinical and lifestyle variables are within normal limits. They could benefit from a risk reduction program aimed at increasing physical activity.
Cluster 3	Patients in this cluster have high blood pressure, tend to be smokers and drink alcohol. Since tobacco and alcohol are often used by people under stress, these patients could benefit from a stress reduction program.
Cluster 4	Patients in this cluster have an elevated body mass index and high cholesterol and glucose levels. Clinically, the combination of these factors is known as metabolic syndrome. These patients could benefit from a program aimed at weight reduction through healthy eating.

Limitations

There are several limitations associated with this analysis:

- While the features used in this analysis are common to many cardiovascular diseases, some diseases have different clinical manifestations. For example, arrhythmias can cause low blood pressure while coronary artery disease can cause high blood pressure. The type of cardiovascular disease is not given in the dataset, which may introduce bias when interpreting results.
- The lab value features are scaled, not continuous. A continuous value could help better differentiate patients.
- The lifestyle variables are binary. The model could potentially be improved by providing continuous data instead of binary; for example number of minutes of exercise or packs of cigarettes smoked per day.

Conclusions

This analysis demonstrated that a simple model consisting of a few clinical and lifestyle variables can do a fairly good job of predicting cardiovascular disease. A screening tool could be developed to check for these features and quickly detect and manage risk in a doctor's office.

The analysis also produced some key insights into how to reduce risk in a few clusters of patients with cardiovascular disease. Although the majority of patients fell into a poorly differentiated cluster (cluster 1), the remaining clusters are very well differentiated and lend themselves to programs to address very specific risk factors.

Acknowledgements

Thank you to Svetlana Ulianova for posting the Cardiovascular Disease dataset on Kaggle.

The statistics on heart disease mortality and morbidity in the US were obtained from the Centers for Disease Control and Prevention's Heart Disease Facts Web page at https://www.cdc.gov/heartdisease/facts.htm.

References

Bonsall, L. (2011). Calculating the mean arterial pressure (MAP). Retrieved December 10, 2019 from https://www.nursingcenter.com/ncblog/december-2011/calculating-the-map

Centers for Disease Control and Prevention. (2019a). Heart disease facts. Retrieved December 11, 2019 from https://www.cdc.gov/heartdisease/facts.htm

Centers for Disease Control and Prevention. (2019.b). Retrieved December 10, 2019 from https://www.cdc.gov/healthyweight/assessing/bmi/childrens_bmi/childrens_bmi_formula.html

Ulianova, S. (2018). Cardiovascular disease dataset. Retrieved December 8, 2019 from https://www.kaggle.com/sulianova/cardiovascular-disease-dataset