Predicting Cardiovascular Disease

UT DATA SCIENCE BOOT CAMP
FINAL PROJECT

Steve Manz
Adam Skelton
Shannon Dang
Xiao Meng

CVD is the leading death globally.

About 647,000 Americans die from heart disease each year—that's 1 in every 4 deaths.

Heart disease cost the United States around $219 billion a year from 2014 to 2015

Cardiovascular disease (CVD) is a class of disease that involves the heart or blood vessel.

Cardiovascular Disease may be preventable.

Knowing this our group became interested in predicting whether a person is at risk of cardiovascular disease based on individual demographic and health status data such as, age, height, weight, blood pressure, cholesterol levels etc.

Question

What health factors are correlated with Cardiovascular Disease?

How can we help people prevent Cardiovascular Disease?

What health factors are correlated with Cardiovascular Disease?

How can we help people prevent Cardiovascular Disease?

Objective

- Create a machine learning model for predicting the probability of presence of CVD given a new user’s demographic and health information.
- Develop a web application where users input personal data to get a probability of having CVD.

Create a machine learning model for predicting the probability of presence of CVD given new user’s demographic and health information.

Develop a web application where users input personal data to get a probability of having CVD.

Technologies

Database

- QuickDBD for Entity Relationship Diagrams (ERD)
- Postgres for creating a database
- pgAdmin4 for working with the imported data
- SQLAlchemy

Machine Learning and Data Exploration

- Python Libraries:
  - numpy
  - pandas
  - sklearn
  - matplotlib
  - seaborn
- Plotly.js

Web Development

- HTML Styles
  - Bootstrap
- JavaScript Library:
  - D3.js

Data Source

- Cardiovascular Disease Dataset
  - https://www.kaggle.com/sulianova/cardiovascular-disease-dataset
- Data description

  Three types of input features:

  - Objective: factual information;
  - Examination: results of medical examination;
  - Subjective: information given by the patient.

Here is where we got our data set.
There are three types of input features. The objective features that had factual information about the subject.
The examination features that came as a result of a medical evaluation.
Finally the subjective information about a subjects lifestyle choices.

Data Description

Identification Feature

Feature	Type
ID	Integer

Target Variable

Feature	Type	
Presence or Absence of Cardiovascular Disease	Binary	0 - Absence 1 - Presence

Examination Feature

Feature	Type	
Systolic Blood Pressure	Integer	labeled as ap_hi (mm Hg)
Diastolic Blood Pressure	Integer	labeled as ap_lo (mm Hg)
Cholesterol	Categorical	1 - Normal 2 - Above Normal 3 - Well Above Normal
Glucose	Categorial	1 - Normal 2 - Above Normal 3 - Well Above Normal

Objective Feature

Feature	Type	
Gender	Binary	1- Female 2 - Male
Age	Integer	years
Height	Integer	cm
Weight	Integer	kg

Subjective Feature

Feature	Type	
Smoking	Binary	0 - Does Not 1 - Smokes
Alcohol intake	Binary	0 - Does Not 1 - Drinks
Physical activity	Binary	0 - Does Not 1 - Exercises

All data entries were tied together with a unique ID.
Our target variable was the the presence of CVD which was expressed by a 0 for absence or 1 for the presence of CVD.
The examination features systolic and Diastolic Blood pressure and categorical data represented by 1,2, or 3.
The objective features were things like gender, age in years, height in cm, and weight in kg.
The subjective features where the subject was asked if they smoked, consumed alcohol and were physically active.

mm Hg = millimeters of mercury

Data Cleaning

- Changed age from days into years by dividing by 365
- Dropped 24 rows with duplicate values

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

⇩

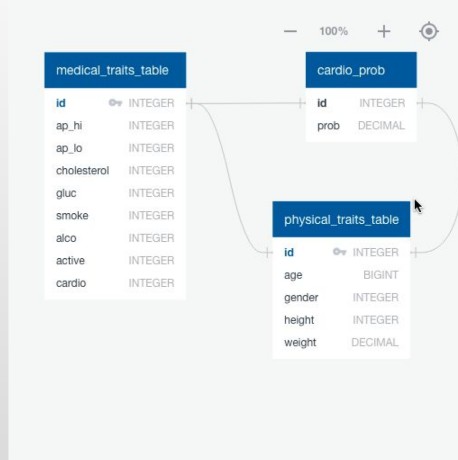
	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	50	2	168	62.0	110	80	1	1	0	0	1	0
1	1	55	1	156	85.0	140	90	3	1	0	0	1	1
2	2	51	1	165	64.0	130	70	3	1	0	0	0	1
3	3	48	2	169	82.0	150	100	1	1	0	0	1	1
4	4	47	1	156	56.0	100	60	1	1	0	0	0	0

Source: Kaggle

We began with 70,000 patients with 11 features (ID excluded), including one target variable, “cardio”.
We cleaned our data in order to run it through a machine learning model. We turned “age” from days into years, dropped “id”, and dropped 24 duplicate values, so in the end we have 11 features and 69,976 patients.

Database Usage

- Local pgAdmin4
- Joined two tables on the ID key from the medical and physical traits tables
- Joined the probability of CVD on the ID key for our final deployed database



Database storage was fairly simple for this project. We used a local pgAdmin4 database to house our data. We used sqlAlchemy to access the database in our program files.
We used two joins to get our final table that was used in our project. We joined the medical traits table to the physical traits table on ID primary key. Then we joined that Cardio table to the Cardio Probabilities table for our final table.

Correctable and Uncorrectable Factors

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	50	2	168	62.0	110	80	1	1	0	0	1	0
1	1	55	1	156	85.0	140	90	3	1	0	0	1	1
2	2	51	1	165	64.0	130	70	3	1	0	0	0	1
3	3	48	2	169	82.0	150	100	1	1	0	0	1	1
4	4	47	1	156	56.0	100	60	1	1	0	0	0	0

Correctable Factors:
- Weight
- Systolic blood pressure (ap_hi)
- Diastolic blood pressure (ap_lo)
- Cholesterol
- Smoke
- Alcohol
- Active
- Glucose*

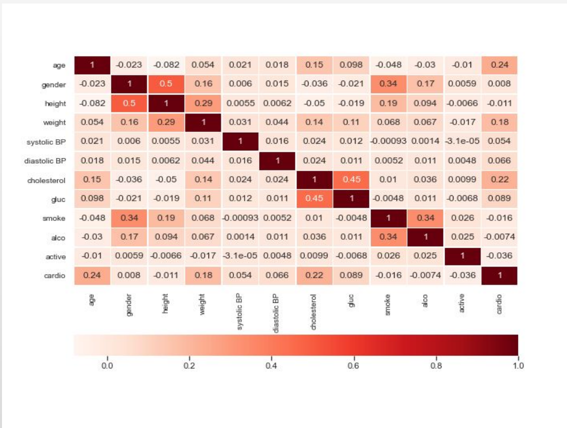
Uncorrectable Factors:
- Gender
- Height
- Glucose*

Before we jump in we want to address that our project gives you some insight of how to prevent CVD but is not a comprehensive study of what you should do to prevent it. If you think you may be at risk of CVD, please seek a medical professional.

Now let’s start, to help us answer the question of how to prevent CVD, it’ useful to separate our features into correctable and uncorrectable factors through behavior.

Factors Correlated with Cardiovascular Disease (CVD)

Correlation Matrix

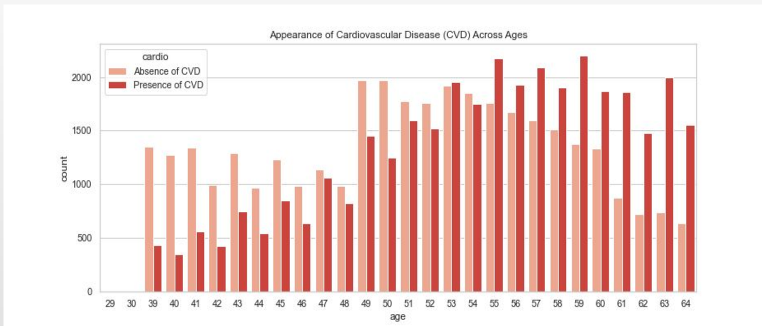


Features Correlated with CVD

Relatively Strong Positive Correlation	• Age (r = 0.24) • Weight (r = 0.18) • Cholesterol (r = 0.22)
Weak Positive Correlation	• Systolic Blood Pressure (r = 0.054) • Diastolic Blood Pressure (r = 0.066) • Glucose (r = 0.089)
Weak Negative Correlation	• Height (r = - 0.011) • Smoke (r = - 0.016) • Active (r = - 0.036)
Almost No Correlation	• Gender (r = 0.008) • Alcohol (r = - 0.0074)

On the left is the correlation matrix with all the features. The correlation coefficients, r, is in each box and the darker red boxes indicates features with a strong positive correlation. The very bottom row, labeled cardio, is our target variable, whether you have CVD or not. We can separate these correlation coefficient into four categories seen on the right.

Exploratory Data Analysis (EDA)
Age

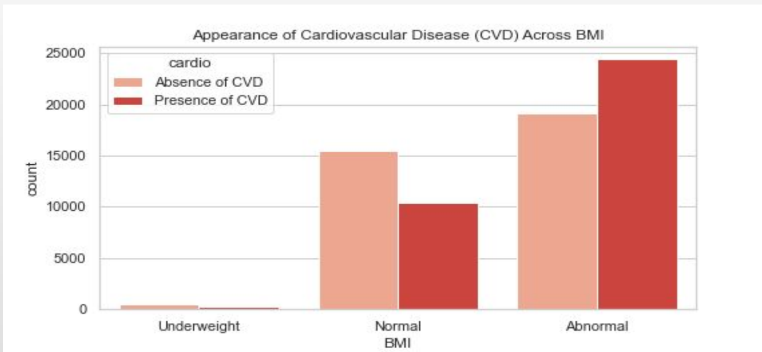


- The higher the age, the higher the percentage of people having CVD.
- Starting at age 55 a higher percentage of people have CVD.

In the couple slide of after this, you will see similar graphs to this one where the red bar indicates Presence of CVD and the Light Orange bar indicates Absence of CVD.

Looking at this chart of CVD Across Age, the red bar surpasses the orange bar starting at age 55 indicating that the higher the age, the higher the percentage of people having CVD.

Exploratory Data Analysis (EDA)
Body Mass Index (BMI)



BMI	Range
Underweight	< 18.5
Normal	18.5 - 25
Abnormal	> 25

Source: National Heart and Lung Institute
BMI = weight(kg) /[height(m)]^2

Higher percentage of people with abnormal BMI have CVD.

Patients with abnormal BMI are more at risk of cardiovascular disease than patients with normal or underweight BMI. We can see this by looking at the abnormal category where patients with cardiovascular disease surpasses patients with our cardiovascular disease.

Exploratory Data Analysis (EDA)
Blood Pressure

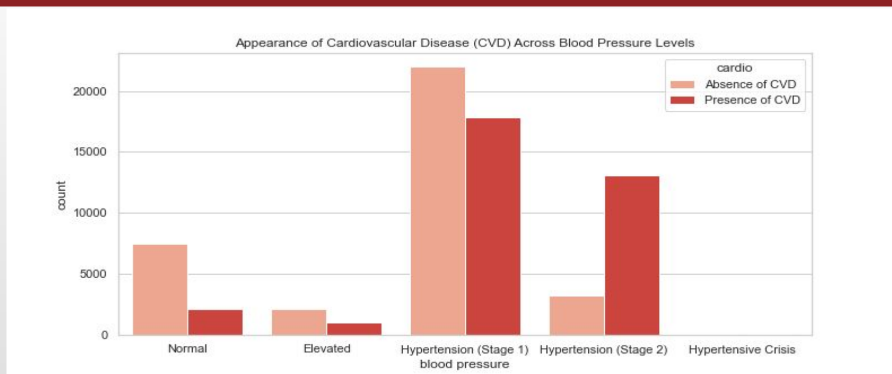
Blood Pressure Stages

Blood Pressure Category	Systolic mm Hg (upper #)		Diastolic mm Hg (lower #)
Normal	less than 120	and	less than 80
Elevated	120-129	and	less than 80
High Blood Pressure (Hypertension) Stage 1	130-139	or	80-89
High Blood Pressure (Hypertension) Stage 2	140 or higher	or	90 or higher
Hypertensive Crisis (Seek Emergency Care)	higher than 180	and/or	higher than 120

Source: American Heart Association

Our data has systolic and diastolic blood pressure, so we decided to translate those numerical data to blood pressure categories

Exploratory Data Analysis (EDA)
Blood Pressure



The higher the blood pressure, the higher the percentage of people have CVD. A person with Hypertension (Stage 2) has a much higher chance of having CVD than those even in Hypertension (Stage 1).

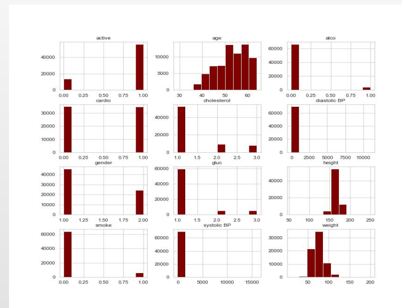
In this graph, it appears that there aren't any patients with Hypertensive Crisis, but there are 96 patients and they all have CVD.
Patients who develop hypertension are more at risk of cardiovascular disease. We can see this by looking at Hypertension (Stage 1) where there is an increasing number of patients who do have cardiovascular disease are and it is further confirmed when looking at Hypertension (Stage 2) where patients with cardiovascular disease excesses patients with our cardiovascular disease.

Data Preprocessing & Feature Reduction

Features	Ranking from RFE
Age	1
Gender	3
Height	1
Weight	1
ap_hi	1
ap_lo	1
cholesterol	1
gluc	2
smoke	5
alco	6
active	4

Feature Selection/Engineering

- Correlation Matrix
  - Features Correlated with CVD:
    - age, height, weight, ap_hi, ap_lo, cholesterol, gluc
- Recursive Feature Elimination (RFE)
  - Selected Features:
    - age, height, weight, ap_hi, ap_lo, cholesterol



Used StandardScaler to standardized our data

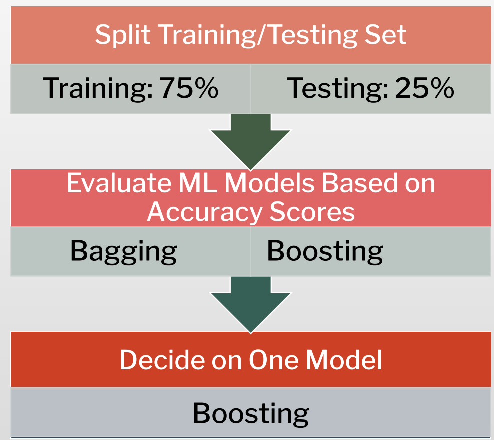
Split the dataset into training (75%) and testing data (25%)

Recursive Feature Elimination selects features by recursively considering smaller and smaller sets of features. First, it trains on the initial set of features. Then, the least important features are pruned from current set of features. That process is recursively repeated on the pruned set until the desired number of features in this case 6 is eventually reached.
Then we plotted our features on a histogram to see whether scaling is necessary, and yes it is as you can see the top middle histogram is left skewed. And then we split our dataset into 75% training and 25% testing.

Machine Learning Models

We tuned and trained six supervised learning models:

Machine Learning (ML) Models	Accuracy Training	Accuracy Testing
K-Nearest Neighbor	69.8%	68.4%
Support Vector Machine (SVM)	71.6%	71.5%
Logistic Regression		71.9%
Decision Tree	73.3%	72.9%
Random Forest	73.1%	73.2%
Gradient Boosting Tree	73.6%	73.4%

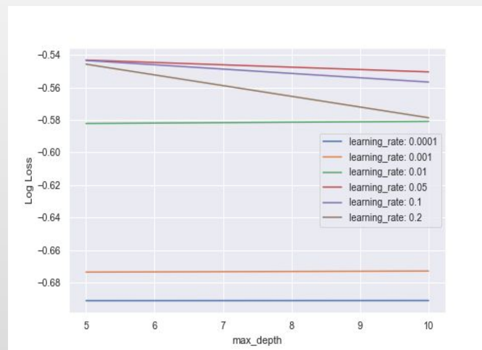


Talk about machine learning model and how we choose the gradient boosting tree.

Gradient Boosting Tree

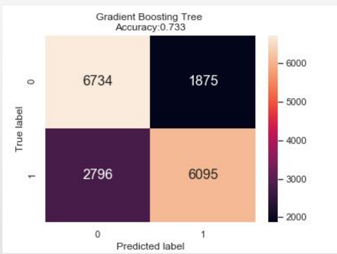
- Tuning parameters to evaluate Gradient Boosting Tree model.
  - Max_depth [5, 8]
  - learning_rate [0.0001, 0.001, 0.01, 0.05, 0.1, 0.2]
- Chosen parameters:
  - max_depth=5
  - learning_rate=0.05

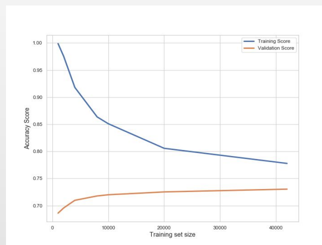
Tune Learning Rate and Max Depth



Details about gradient boosting tree

Gradient Boosting Tree





	precision	recall	f1-score	support
0	0.71	0.78	0.74	8609
1	0.76	0.69	0.72	8891
accuracy			0.73	17500

Accuracy Score

- Accuracy (training): 0.739
- Accuracy (testing): 0.734

Talk about accuracy score.

Results

- Demonstration
- Hosted App

Demo our web app

Summary

Factors correlated/uncorrelated with CVD

Relatively Strong Positive Correlation	• Age (r = 0.24) • Weight (r = 0.18) • Cholesterol (r = 0.22)
Weak Positive Correlation	• Systolic Blood Pressure (r = 0.054) • Diastolic Blood Pressure (r = 0.066) • Glucose (r = 0.089)
Weak Negative Correlation	• Height (r = - 0.011) • Smoke (r = - 0.016) • Active (r = - 0.036)
Almost No Correlation	• Gender (r = 0.008) • Alcohol (r = - 0.0074)

Developed a web app based on Gradient Boosting Tree Model to predict probability of getting CVD given new users’ personal information to inform them the risk and prevent CVD before it becomes serious.

Future Analysis

- Increase model accuracy
  - Collect more variables based on medical diagnostic guidance.
    - e.g. Family history, some related symptoms like fatigue, shortness of breath, Irregular heartbeat etc..
- Add more features in our web app to help users at a high risk of Cardiovascular Disease to make decision on check or treatment.
  - Provide information about Medical check or treatment for high risk users.
  - Identify sub-categories of CVD or distinguish CVD from other diseases if identify the user is at high risk.

References

[1] Center for Disease Control and Prevention. 2020. *Heart Disease Facts | Cdc.Gov.* [online] Available at: <https://www.cdc.gov/heartdisease/facts.htm> [Accessed 29 Aug 2020].

[2] McGillJr, Henry C., et al. “Preventing Heart Disease in the 21st Century.” *Circulation*, 4 Mar. 2008, www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.107.717033.

[3] BMI Range https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmicalc.htm

Questions