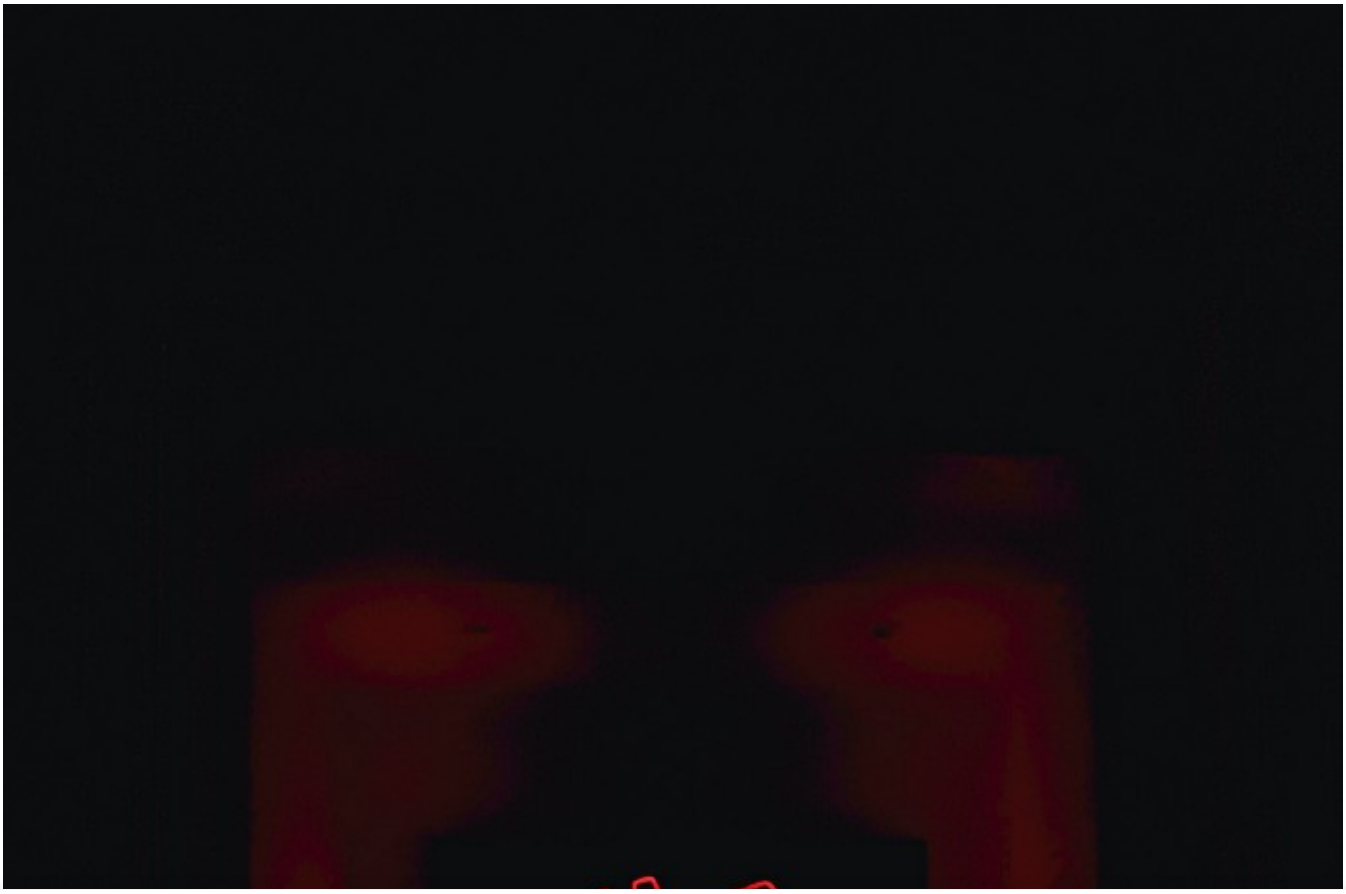
Natassha Selvaraj Follow

May 25, 2020 · 9 min read ★ · Listen

The Framingham Heart Study: Decision Trees

Using decision trees to predict the 10 year risk of developing Coronary Heart Disease

What is the Framingham Heart Study?



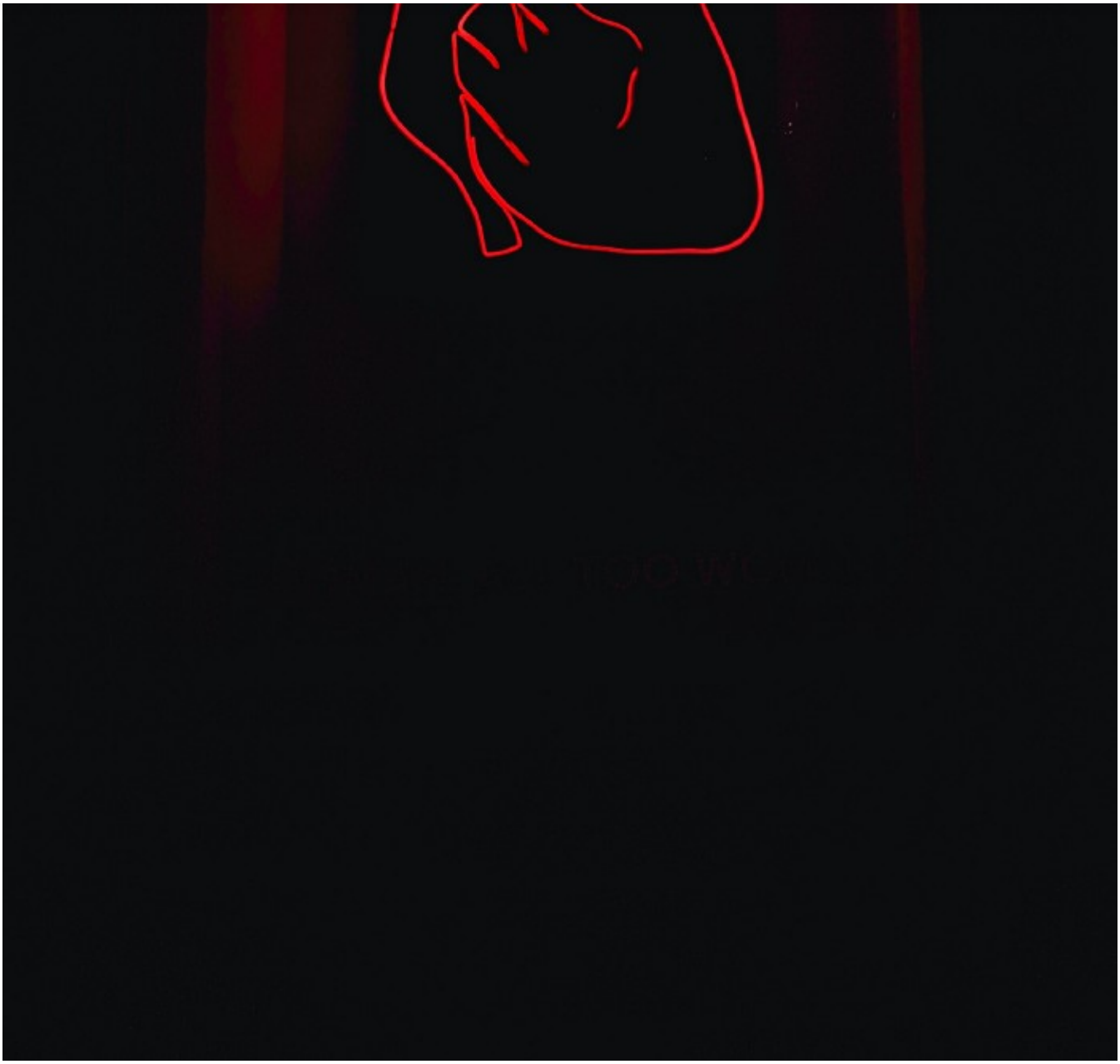


Photo by Alexandru Acea on Unsplash

The Framingham Heart Study was a turning point in identifying the risk factors of heart disease, and is one of the most important epidemiological studies conducted.

A lot of our present understanding of cardiovascular disease can be attributed to this study.

The Framingham Heart Study: Origin

The origin of the FHS can be attributed to the premature death of US President Franklin D. Roosevelt in the year 1945

Photo by Library of Congress on Unsplash

The President had extremely high blood pressure, which at that time was not considered a big deal.

Before his presidency, his blood pressure was 140/100mmHg, which is considered high according to today's standards. One year before his death, the president's blood pressure had shot up to 210/120mmHg, which today is considered a hypertensive crisis.

At that time, his personal physician was a specialist in EENT (Eyes, Ears, Nose, and

The President died of a massive cerebral hemorrhage in the year 1945, and his blood pressure was 300/190mmHg on that day. The death of President Roosevelt paints a picture of our understanding of heart disease in the mid 20th century.

There was a huge increase in deaths from Coronary Heart Disease, or CHD in the 1930's, 1940's, and 1950's. In the 1940's, heart disease was the number one cause of death among Americans.

A Solution?

To better understand heart disease and the measures that could be taken to combat it, the Framingham Heart Study (FHS) was established in the late 1940's.

It was a joint project of Boston University and the National Heart, Lung, and Blood Institute (NHLBI).

A large cohort of initially healthy patients between the age group 30 and 59 in the city of Framingham, Massachusetts were tracked for a period of 20 years, to better understand cardiovascular disease. The study was conducted with an initial cohort of 5209 patients.

The aim of the study was to enroll people free of the disease, and see who developed the disease in the next 20 years.

How was this done?

Every two years, the participants would have to report to a testing center, where an examination was conducted. The patients were examined and their health information was updated.

They were also given questionnaires to fill up, in which they updated behavioral information, such as exercise or smoking habits.

The data collected from this study allowed for a better understanding of the risk factors of heart disease. Medical interventions then took place based on the findings of the FHS.


- Smoking was found to increase the risk of CHD (1960)
- Cholesterol and high BP increased the risk of CHD (1961)
- Physical activity decreased the risk of CHD (1967)
- High levels of HDL cholesterol was found to increase the risk of CHD (1988)
- The lifetime risk of developing CHD was higher in men than in women (1999)
- Obesity is a risk factor for heart failure (2002)

Impact and Further Studies

There were many revolutionary breakthroughs in our knowledge of cardiovascular disease due to the FHS, and many medical interventions have taken place since then to prevent and decrease the risk of heart disease.

Around 20 years after the original cohort, a second study was started. This study involved the offspring of the first cohort and their spouses, and took place in 1971.

In the year 2002, the third generation cohort started, who were the grandchildren of the original cohort. The study is ongoing, and has expanded to take in various other risk factors such as family history, social network analysis, and genetic information.

Overall, there have been around 2400 studies written using data from the FHS, and this number continues to grow each year.

The FHS has contributed greatly in reducing mortality rates associated with CHD, and has corrected many clinical misconceptions about heart disease. We know so much more about reducing the risk of heart disease, treatment, and improving quality of life due to the FHS.

Decision Trees to Predict 10 Year CHD Risk

Now, we will attempt to build a decision tree classifier in Python that can predict the ten year risk of a patient developing CHD, given certain risk factors.


First, I read the data set into a data frame using the Pandas library.

```
# Creating the data frame
framingham = pd.read_csv('framingham.csv')
framingham.head()
```

Checking the head of the data frame:

male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes
1	39	4.0	0	0.0	0.0	0	0	0
0	46	2.0	0	0.0	0.0	0	0	0
1	48	1.0	1	20.0	0.0	0	0	0
0	61	3.0	1	30.0	0.0	0	1	0
0	46	3.0	1	23.0	0.0	0	0	0

Image by author

The data frame consists of 16 variables; 15 independent variables, or risk factors, and 1 dependent variable.

Risk Factors/Independent Variables

1. **Demographic:**

- Male : A value of 1 indicates that the participant is male, and 0 indicates they are female.
- Age: The age of the participant
- Education Level: 1-High School, 2-High School Diploma/GED, 3-College, 4-Degree

2. Behavioral:

- currentSmoker: 1- The participant is a current smoker, 0- participant does not smoke currently

- BPMeds: Amount of BP medication the participant is on
- prevalentStroke: 0- no prevalence of stroke, 1-has had occurrences of stroke
- prevalentHyp: 0-no prevalence of hypertension, 1-prevalence of hypertension
- diabetes: 0-no diabetes, 1-has diabetes

4. Risk factors from first physical examination:

- totChol: Total cholesterol
- sysBP: Systolic blood pressure
- diaBP: Diastolic blood pressure
- BMI: Body Mass Index
- Heart Rate: Heart rate in bpm
- Glucose: Glucose level (mg/dL)

Exploratory Data Visualization

Before creating the model, I will perform some exploratory data visualization, to get some insights on the data.

As mentioned above, there are many risk factors such as smoking and high cholesterol levels that were found by the Framingham Heart Study to increase 10 year risk of CHD.

I will take a look at some of these risk factors, and see if I can find these relationships in this dataset. This will be done using the Seaborn library.

First, I will create a count plot of the variable ‘education,’ in order to get a better understanding on the level of education of the participants.

```
sns.countplot(x='education',data=framingham)
```

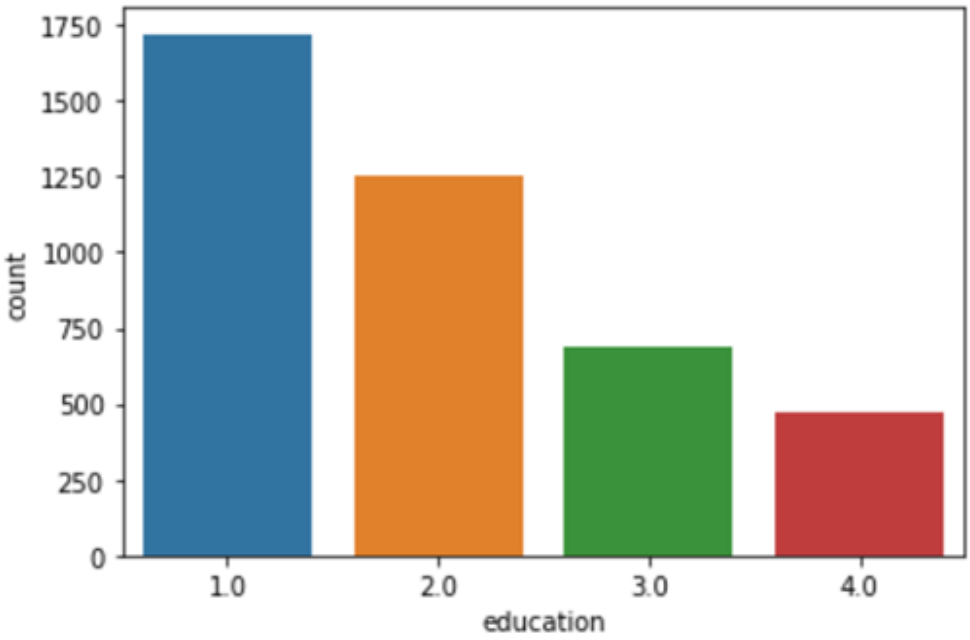



Image by author

Most participants seem to have some form of high school education. A fewer number of them have a diploma or went to college, and very few have a degree.

Now, I will look into some risk factors. In the year 1960, smoking was said to increase the risk of CHD. I will try to visualize this relationship.

```
sns.catplot(x='TenYearCHD',y='cigsPerDay',kind='bar',data=framingham
)
```

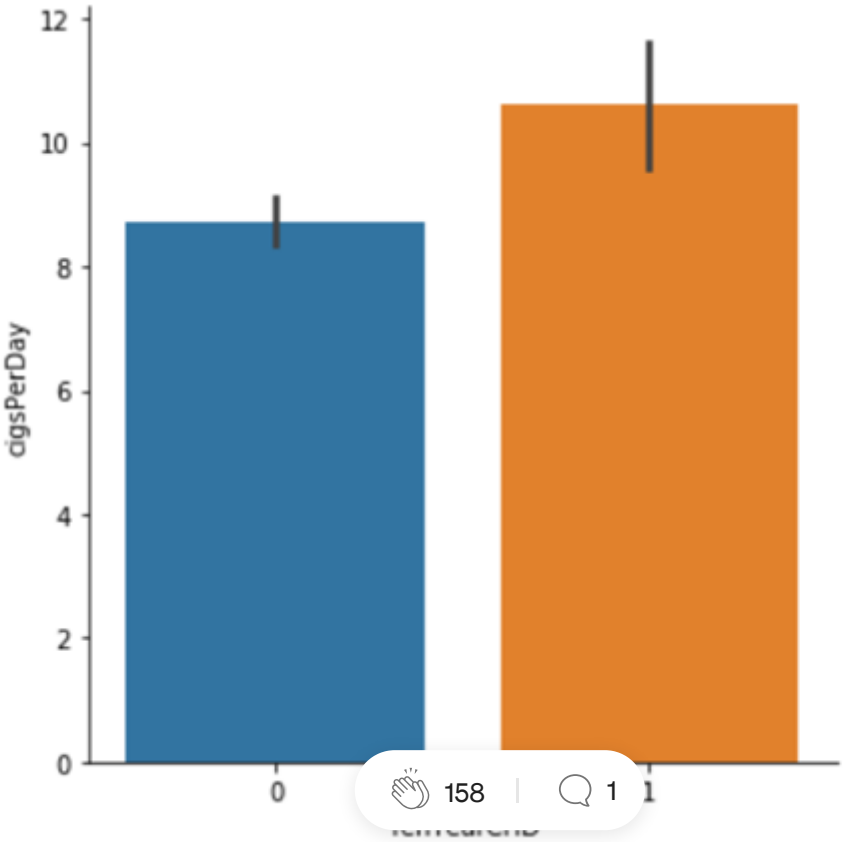


Image by author

It can be seen that patients who have a ten year CHD risk smoke more cigarettes per day than those who do not.

Now, I will try to see if there is a relationship between age and the ten year risk of CHD. I will sort this by the category smoker.

```
sns.boxplot(x='TenYearCHD',y='age',hue='currentSmoker',data=framingh
am)
plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)
```

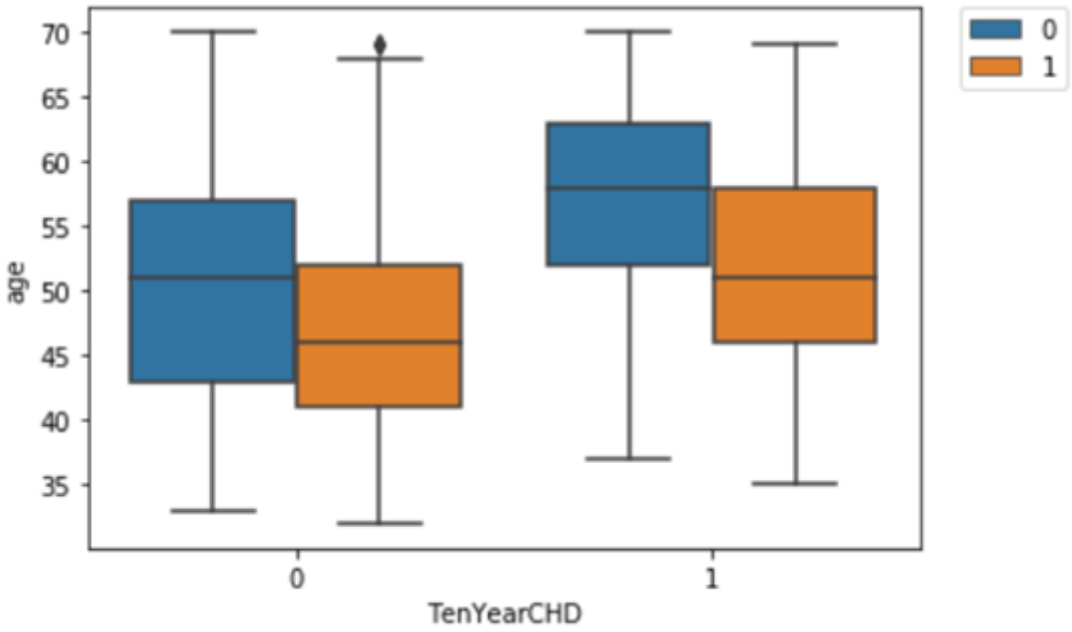


Image by author

From here, it can be seen that older patients are more likely to develop CHD. Smokers seem to be at a larger risk of developing CHD at a younger age, as compared to non-smokers.

Next, I will take a look at the relationship between age, prevalent stroke, and the ten year risk of developing CHD.

```
sns.boxplot(x='TenYearCHD',y='age',hue='prevalentStroke',data=framin
gham)
plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)
```

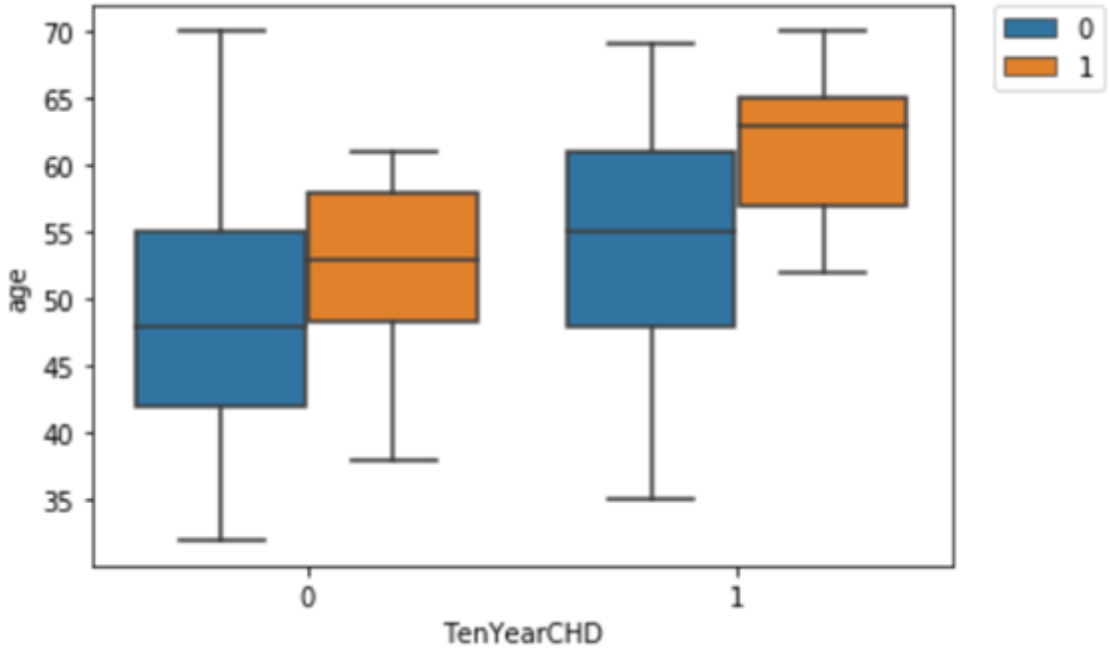


Image by author

It looks as though strokes are more prevalent in participants of an older age group.

Now, I will take a look at the variables age, diabetes, and ten year risk of developing CHD.

```
sns.boxplot(x='TenYearCHD',y='age',hue='diabetes',data=framingham)
plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)
```

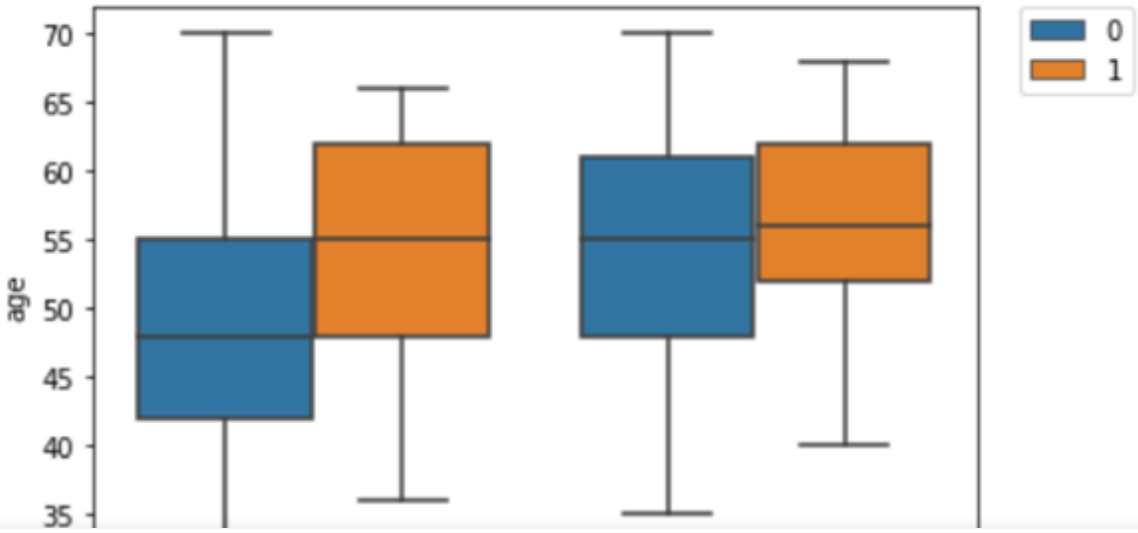


Image by author

Again, it looks as though an older participant is more likely to have diabetes than a younger one.

Now, I will take a look at total cholesterol levels. In the year 1961, an increase in cholesterol levels was found to increase the risk of CHD.

```
sns.boxplot(x='TenYearCHD',y='totChol',data=framingham)
plt.ylim(80)
```

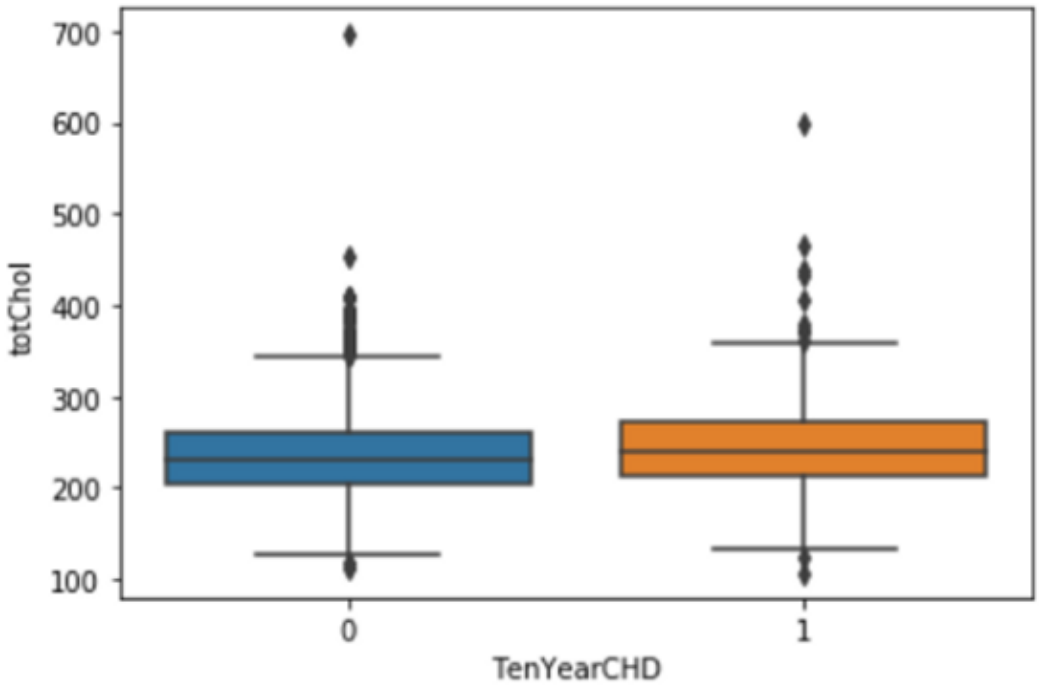


Image by author

Patients with a ten year CHD risk have slightly higher cholesterol levels than patients who don't, though the difference is very small and not significant.

This may be attributed to the fact that the variable 'total cholesterol' includes both LDL and HDL. LDL, or 'bad cholesterol' is said to increase the risk of CHD. HDL, or 'good cholesterol' is said to decrease the risk of CHD.

Next, I will take a look at both systolic and diastolic blood pressure, and visualize their relationship with ten year CHD risk.

```
sns.catplot(x='TenYearCHD',y='sysBP',kind='bar',data=framingham)
```

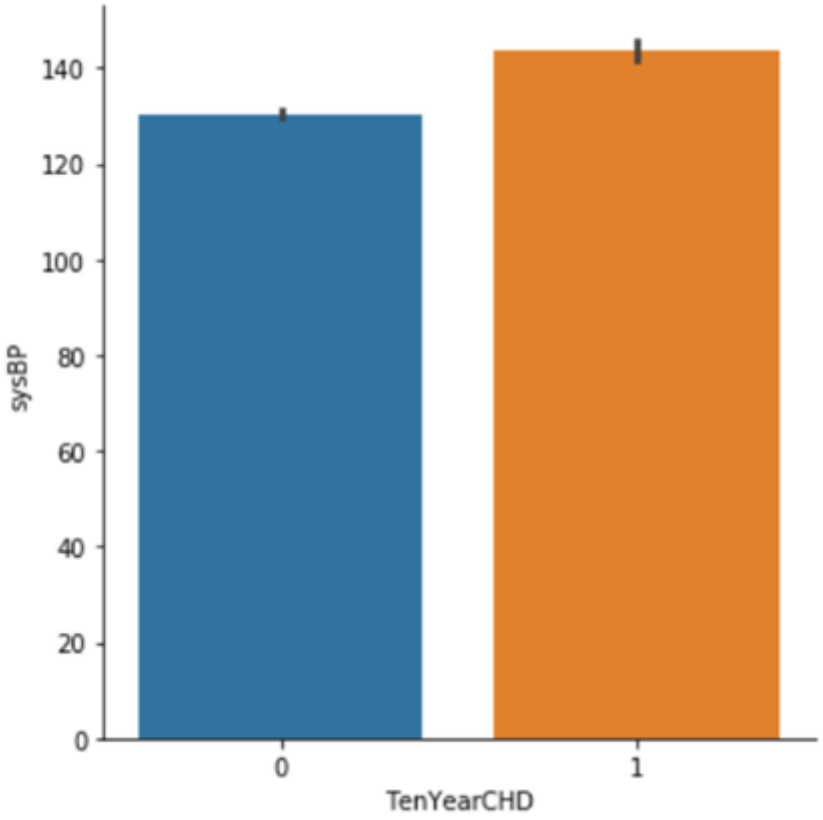


Image by author

```
sns.catplot(x='TenYearCHD',y='diaBP',kind='bar',data=framingham)
```

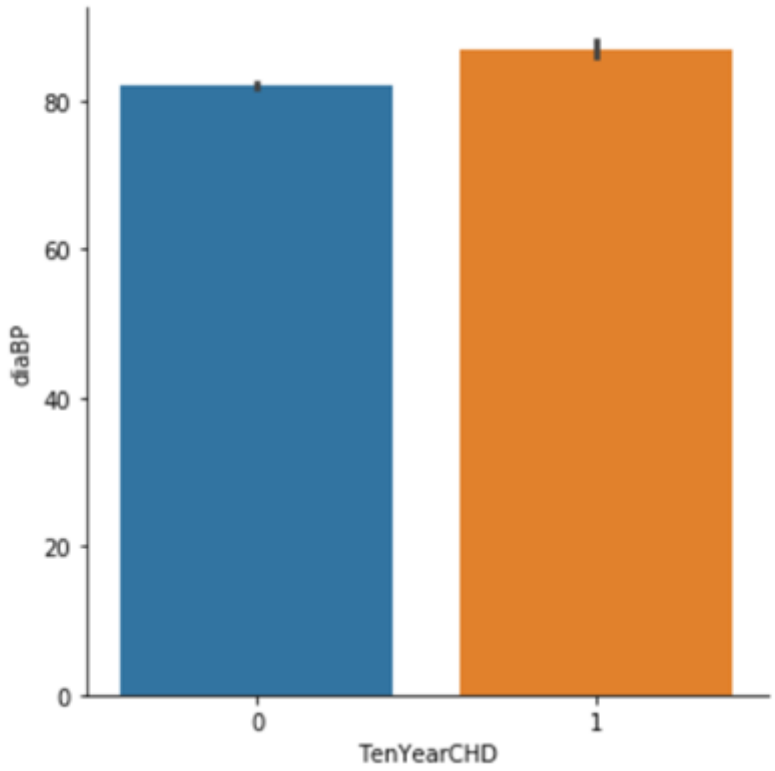


Image by author

Blood pressure does seem to be linked with coronary heart disease. Patients with a ten year CHD risk seem to have higher blood pressure than the ones who don't.

In the year 2002, obesity was found to be a risk factor in developing CHD. I will now take a look at the relationship between BMI and the ten year risk of CHD.

```
sns.catplot(x='TenYearCHD',y='BMI',kind='bar',data=framingham)
```

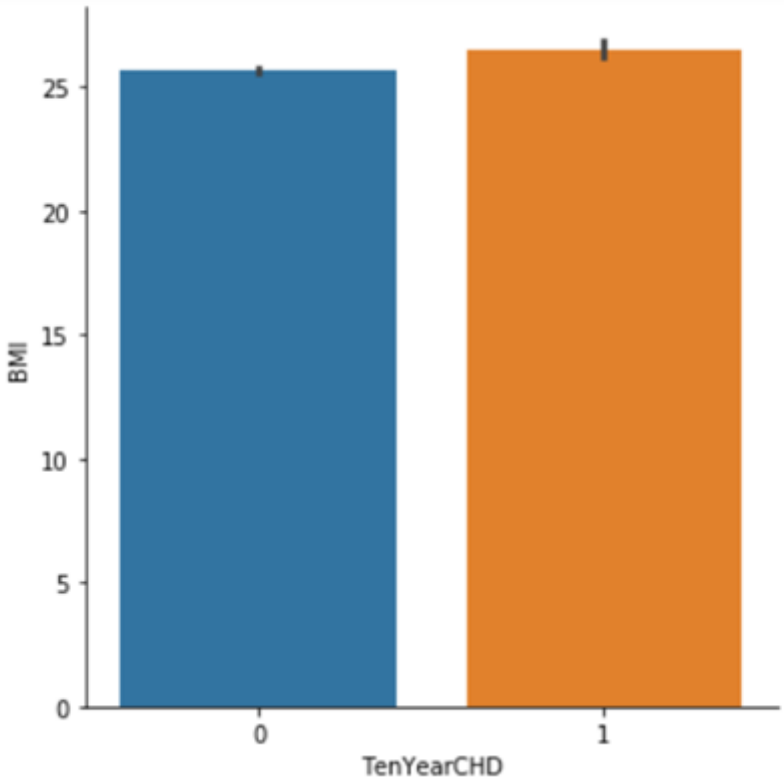


Image by author

Patients with CHD are seen to have slightly higher BMI than patients without.

In the year 1961, high BP was said to increase the risk of CHD. Finally, I will be taking a look at BP medications, and see if there is a relationship.

```
sns.catplot(x='TenYearCHD',y='BPMeds',kind='bar',data=framingham)
```

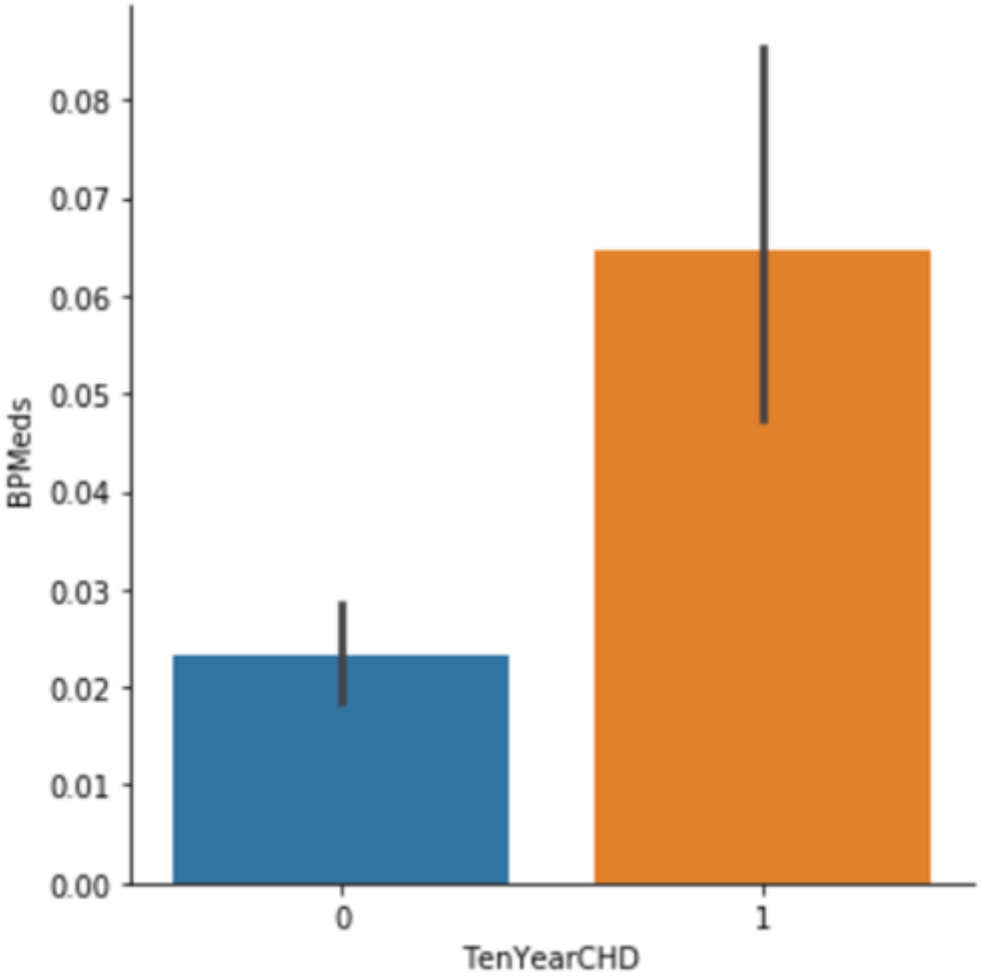


Image by author

A higher dosage of BP medications is associated with a larger ten year CHD risk.

Data Preprocessing

Before fitting the data to the model, I will first need to prepare it.

```
# Checking for null values
framingham.isnull().any()
```

There are a couple of columns with null values in the data frame. There are many methods to treat missing values, such as imputing with the mean, but for the sake of simplicity, I am just going to drop them.

The model that we create to predict the ten year risk of CHD needs to perform better than the baseline.

A baseline model is one that predicts the majority class all the time.

```
framingham['TenYearCHD'].value_counts()
```

```
0    3101
1     557
Name: TenYearCHD, dtype: int64
```

Here, the majority class is 0, or absence of ten year CHD. The baseline accuracy is computed:

```
# Baseline accuracy:
3101/(3101+557)
```

The baseline accuracy is 0.85, and the model has to beat this baseline.

Train-test split

I am going to split the model into two sets, training and testing. I will train on one set, and evaluate on the test set.

```
from sklearn.model_selection import train_test_split
X = framingham.drop('TenYearCHD',axis=1)
y = framingham['TenYearCHD']
X_train, X_test, y_train, y_test =
train_test_split(X,y,test_size=0.35)
```


To overcome this, I did both; oversampling and undersampling. Then, I created a pipeline for a decision tree classifier.

```
oversample = RandomOverSampler(sampling_strategy='minority')
X_over, y_over = oversample.fit_resample(X, y)
X_train, X_test, y_train, y_test =
train_test_split(X_over,y_over,test_size=0.35)

steps = [('under', RandomUnderSampler()), ('model',
DecisionTreeClassifier())]
pipeline = Pipeline(steps=steps)
```

Fitting the Data

```
pipeline.fit(X_train,y_train)
```

Making Predictions on Test Data

```
pipepred = pipeline.predict(X_test)
```

Evaluating the Model

```
from sklearn.metrics import classification_report,accuracy_score
print(classification_report(y_test,pipepred))
```

	precision	recall	f1-score	support
0	0.98	0.81	0.89	1095
1	0.84	0.98	0.90	1076
accuracy			0.90	2171
macro avg	0.91	0.90	0.90	2171
weighted avg	0.91	0.90	0.90	2171

And that's it! We have successfully built a decision tree classifier to predict a patient's 10 year risk of CHD.

The complete code can be found here.

Sources

[1] Bertsimas, D. (2020). The Analytics Edge.

[2] Hajar R. (2016). Framingham Contribution to Cardiovascular Disease.