

Hosted by Drivendata

# Predicting Heart Disease

<https://www.drivendata.org/competitions/54/machine-learning-with-a-heart/page/107/>



Heart disease is the number one cause of death worldwide, so if you're looking to use data science for good you've come to the right place. To learn how to prevent heart disease we must first learn to reliably detect it.

Our dataset is from a study of heart disease that has been open to the public for many years. The study collects various measurements on patient health and cardiovascular statistics, and of course makes patient identities anonymous.

## About:

Preventing heart disease is important. Good data-driven systems for predicting heart disease can improve the entire research and prevention process, making sure that more people can live healthy lives.

In the United States, the Centers for Disease Control and Prevention is a good resource for information about heart disease. According to their website:

- ② About 610,000 people die of heart disease in the United States every year—that's 1 in every 4 deaths.
- ② Heart disease is the leading cause of death for both men and women. More than half of the deaths due to heart disease in 2009 were in men.
- ② Coronary heart disease (CHD) is the most common type of heart disease, killing over 370,000 people annually.
- ② Every year about 735,000 Americans have a heart attack. Of these, 525,000 are a first heart attack and 210,000 happen in people who have already had a heart attack.
- ② Heart disease is the leading cause of death for people of most ethnicities in the United States, including African Americans, Hispanics, and whites. For American Indians or Alaska Natives and Asians or Pacific Islanders, heart disease is second only to cancer.

## Problem description

Your goal is to predict the binary class `heart_disease_present`, which represents whether or not a patient has heart disease:

0 represents no heart disease present

1 represents heart disease present

# Dataset

---

There are 14 columns in the dataset, where the `patient_id` column is a unique and random identifier. The remaining 13 features are described in the section below.

`slope_of_peak_exercise_st_segment` (type: int): the slope of the peak exercise **ST segment**, an electrocardiography read out indicating quality of blood flow to the heart

`thal` (type: categorical): results of **thallium stress test** measuring blood flow to the heart, with possible values `normal`, `fixed_defect`, `reversible_defect`

`resting_blood_pressure` (type: int): resting blood pressure

`chest_pain_type` (type: int): chest pain type (4 values)

`num_major_vessels` (type: int): number of major vessels (0-3) colored by flourosopy

`fasting_blood_sugar_gt_120_mg_per_dl` (type: binary): fasting blood sugar > 120 mg/dl

`resting_ekg_results` (type: int): resting electrocardiographic results (values 0,1,2)

`serum_cholesterol_mg_per_dl` (type: int): serum cholestoral in mg/dl

`oldpeak_eq_st_depression` (type: float): oldpeak = **ST depression** induced by exercise relative to rest, a measure of abnormality in electrocardiograms

`sex` (type: binary): `0`: female, `1`: male

`age` (type: int): age in years

`max_heart_rate_achieved` (type: int): maximum heart rate achieved (beats per minute)

`exercise_induced_angina` (type: binary): exercise-induced chest pain (`0`: False, `1`: True)



## Feature data example

---

Here's an example of one of the rows in the dataset so that you can see the kinds of values you might expect in the dataset. Some are binary, some are integers, some are floats, and some are categorical. There are no missing values.

field	value
slope_of_peak_exercise_st_segment	2
thal	normal
resting_blood_pressure	125
chest_pain_type	3
num_major_vessels	0
fasting_blood_sugar_gt_120_mg_per_dl	1
resting_ekg_results	2
serum_cholesterol_mg_per_dl	245
oldpeak_eq_st_depression	2.4
sex	1
age	51
max_heart_rate_achieved	166
exercise_induced_angina	0