Missing Data

CS1090A Introduction to Data Science

Pavlos Protopapas, Kevin Rader, and Chris Gumb

Photo: Eleonore Wen
Jiu Zhai Gou, China

Outline

- Motivation
- Random Forest
- Variable Importance
- **Missing Data (again)**
- Class Imbalance
- Tree building algorithms

Missing Data: Example

Consider the below real-world dataset used to predict the presence of heart disease.

Weight (kg)	Height (cm)	Chest Congested?	Good Blood Circulation?	Arteries Blocked?	Heart Disease
58.3	125.3	No	No	No	No
65.7	193.2	Yes	No	Yes	Yes
NaN	112	No	Yes	No	No
112	165.7	No	No	NaN	Yes
45	135	No	No	No	No
40	120	No	No	No	Yes

Missing Data: Example

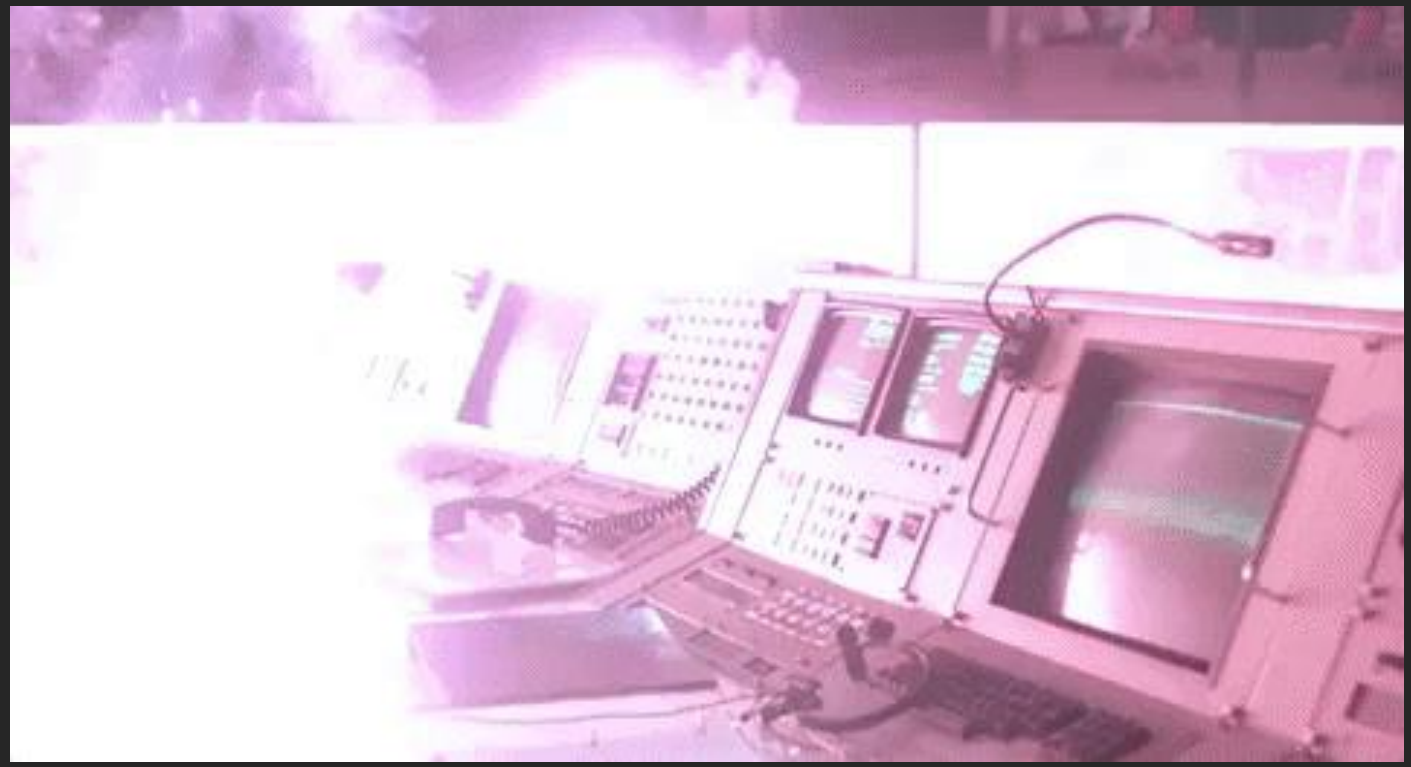
Real-world datasets most often have **missing values**.

Weight (kg)	Height (cm)	Chest Congested?	Good Blood Circulation?	Arteries Blocked?	Heart Disease
58.3	125.3	No	No		
65.7	193.2	Yes	No		
NaN	112	No	Yes		No
112	165.7	No	No	NaN	Yes
45	135	No	No	No	No
40	120	No	No	No	Yes

Missing values can occur due to a variety of reasons!

Maybe someone forgot to fill out the response sheet properly.

...perhaps a device malfunction..



...or maybe it was done purposely!

...or maybe it was done purposely!

BUT, HOW DO WE DEAL WITH MISSING DATA?

Missing Data: Example

There are many ways to deal with this missing data by performing imputation which we have learned.

However, in decision trees, we can handle missing values implicitly, which are called **surrogate splits**!

Introducing Surrogate Splits!

The basic idea is that during training, we find alternative splits, or "surrogate splits", that can be used during prediction.

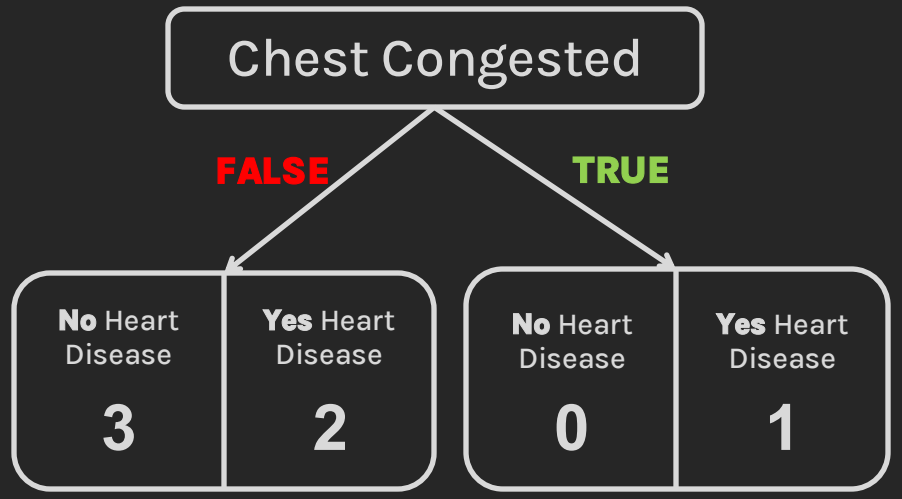
Missing Data: Decision Trees

The basic idea is that during training, we find alternative splits, or "surrogate splits", that can be used during prediction.

Missing Data: Decision Trees

As a first step, we do our usual tree thing.
Let's start with the predictor "Chest Congested" and split.

Weight (kg)	Height (cm)	Chest Congested?	Good Blood Circulation?	Arteries Blocked?	Heart Disease
58.3	125.3	No	No	No	No
65.7	193.2	Yes	No	Yes	Yes
75.2	112	No	Yes	No	No
112	165.7	No	No	Yes	Yes
45	135	No	No	No	No
40	120	No	No	No	Yes



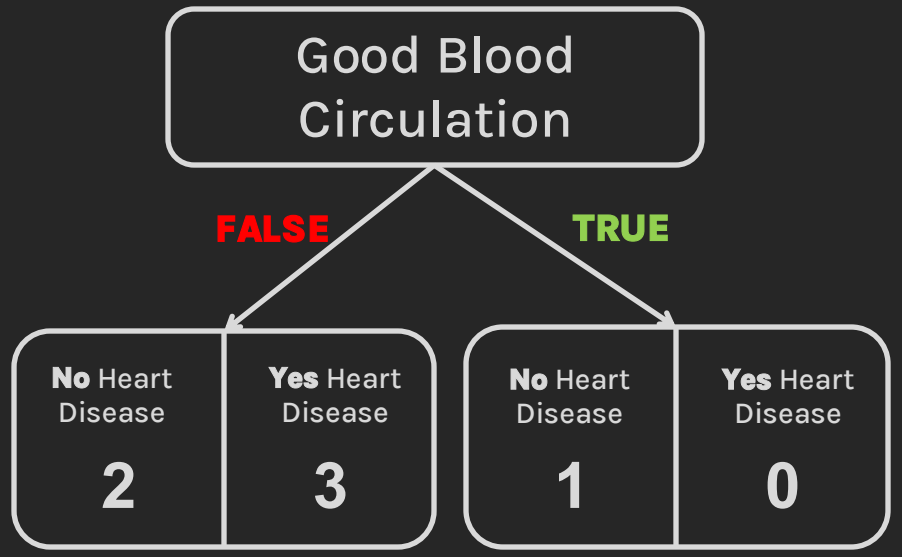
- WHEN "CHEST CONGESTED" IS FALSE, WE HAVE [3 NOs, 2 YESes].
- WHEN "CHEST CONGESTED" IS TRUE, WE HAVE [0 NOs, 1 YES].

We will be counting the number of "yes" or "no" in the response variable after each split.

Missing Data: Decision Trees

Then examine the split for “Good Blood Circulation”.

Weight (kg)	Height (cm)	Chest Congested?	Good Blood Circulation?	Arteries Blocked?	Heart Disease
58.3	125.3	No	No	No	No
65.7	193.2	Yes	No	Yes	Yes
75.2	112	No	Yes	No	No
112	165.7	No	No	Yes	Yes
45	135	No	No	No	No
40	120	No	No	No	Yes

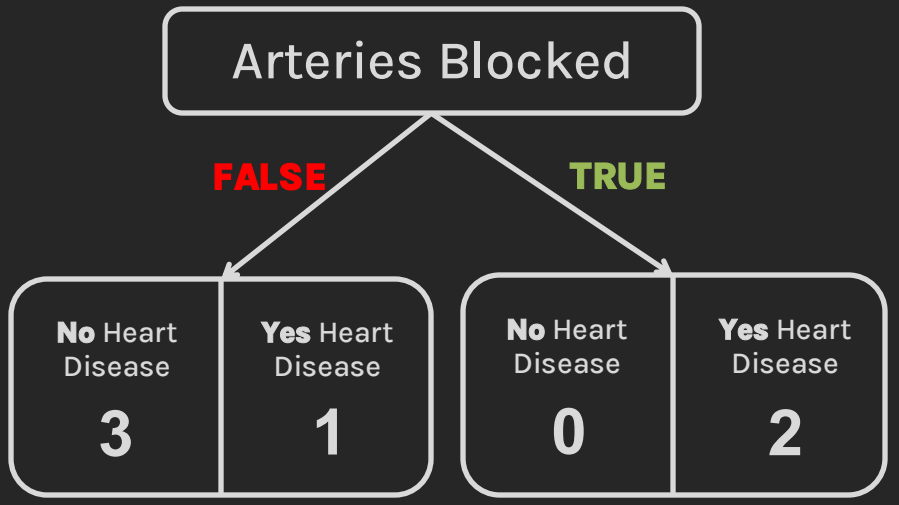


- WHEN "GOOD BLOOD CIRCULATION" IS FALSE, WE HAVE [2 NOs, 3 YESes].
- WHEN "GOOD BLOOD CIRCULATION" IS TRUE, WE HAVE [1 NO, 0 YESes]

Missing Data: Decision Trees

… for “Arteries Blocked”:

Weight (kg)	Height (cm)	Chest Congested?	Good Blood Circulation?	Arteries Blocked?	Heart Disease
58.3	125.3	No	No	No	No
65.7	193.2	Yes	No	Yes	Yes
75.2	112	No	Yes	No	No
112	165.7	No	No	Yes	Yes
45	135	No	No	No	No
40	120	No	No	No	Yes



- WHEN "ARTERIES BLOCKED" IS FALSE, WE HAVE [3 NOs, 1 YES].
- WHEN "ARTERIES BLOCKED" IS TRUE, WE HAVE [0 NOs, 2 YESes].

Let's compare the split distribution of "***arteries blocked***" with the other two predictors: "***chest congested***" and "***good blood circulation***".

For "arteries blocked":

- When "blocked arteries" is false, we have [3 NOs, 1 YES].
- When "blocked arteries" is true, we have [0 NOs, 2 YESes].

For "chest congested":

- When "chest congested" is false, we have [3 NOs, 2 YESes].
- When "chest congested" is true, we have [0 NOs, 1 YES].
- The difference in the distributions of chest-congested and blocked arteries is 2.

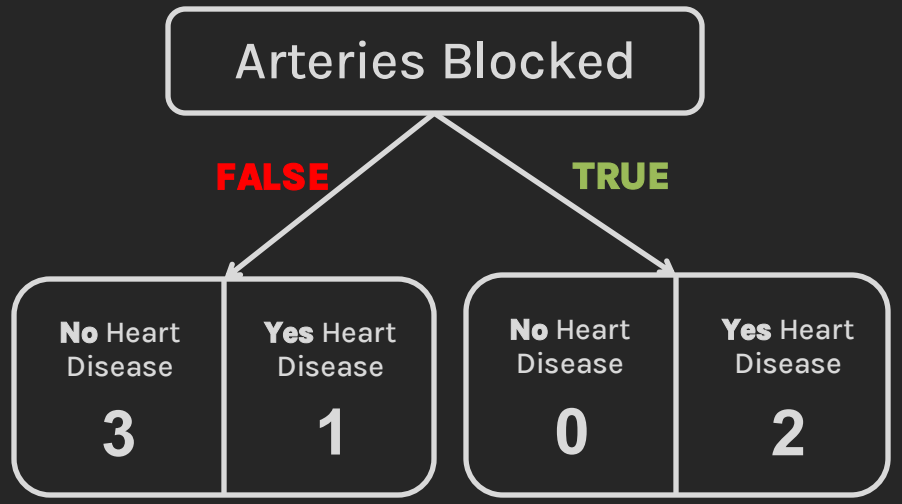
For "good blood circulation":

- When "good blood circulation" is false, we have [2 NOs, 3 YESes].
- When "good blood circulation" is true, we have [1 NO, 0 YESes].
- The difference in the distributions of good blood circulation and blocked arteries is 6.

The difference in distribution indicates the number of flips required to achieve alignment.

Which of these two closely resemble split for ***Arteries Blocked***?

Weight (kg)	Height (cm)	Chest Congested?	Good Blood Circulation?	Arteries Blocked?	Heart Disease
58.3	125.3	No	No	No	No
65.7	193.2	Yes	No	Yes	Yes
NaN	112	No	Yes	No	No
112	165.7	No	No	Yes	Yes
45	135	No	No	No	No
40	120	No	No	No	Yes



As “chest congested” shows the most similar split distribution to ***arteries blocked***, it will be our second choice for a split during prediction if the value of ***arteries blocked*** is missing.

Missing Data: Decision Trees

Now that you understand the intuition behind it, what do you think is the surrogate for ***Weight***?

The logical guess would be height!

Weight (kg)	Height (cm)			teries Blocked?	Heart Disease
58.3	125.3	No	No	No	No
65.7	193.2	Yes	No	Yes	Yes
75.2	112	Yes	Yes	No	No
112	165.7	No	No	Yes	Yes
45	135	No	No	No	No
40	120	No	No	No	Yes

Missing Data: Decision Trees

During training, for every optimal split, we create a **rank** list of **surrogate splits**. This ranking is based on **similarity** between the split distributions of the optimal predictor and every other predictor.

Some Important Points about Surrogate Splits

- Surrogates can help us understand the primary splitter.
- Surrogates perform better when there is multi-collinearity.
- There is no guarantee that useful surrogates can be found!