

# CorHealth

Group: Python Newbie

By Jingwen Zheng, Lanlan Zhang, Sinn Munn Siow,  
Jiehe Huang, Mengjie Yang, Xi Li

# Agenda

---

- Business Problem
- Dataset Understanding & Data Cleaning
- Feature 1 - Percentage of People with and without Cardiovascular Disease
- Feature 2 - Attribute Correlation to Cardiovascular Disease
- Feature 3 - Key Attributes Analysis
- Feature 4 - Prediction
- Execution



# Business Problem



- Chance of getting cardiovascular disease
- Correlations of potential causes
- Prediction

**Audience:** Clinics & Hospitals

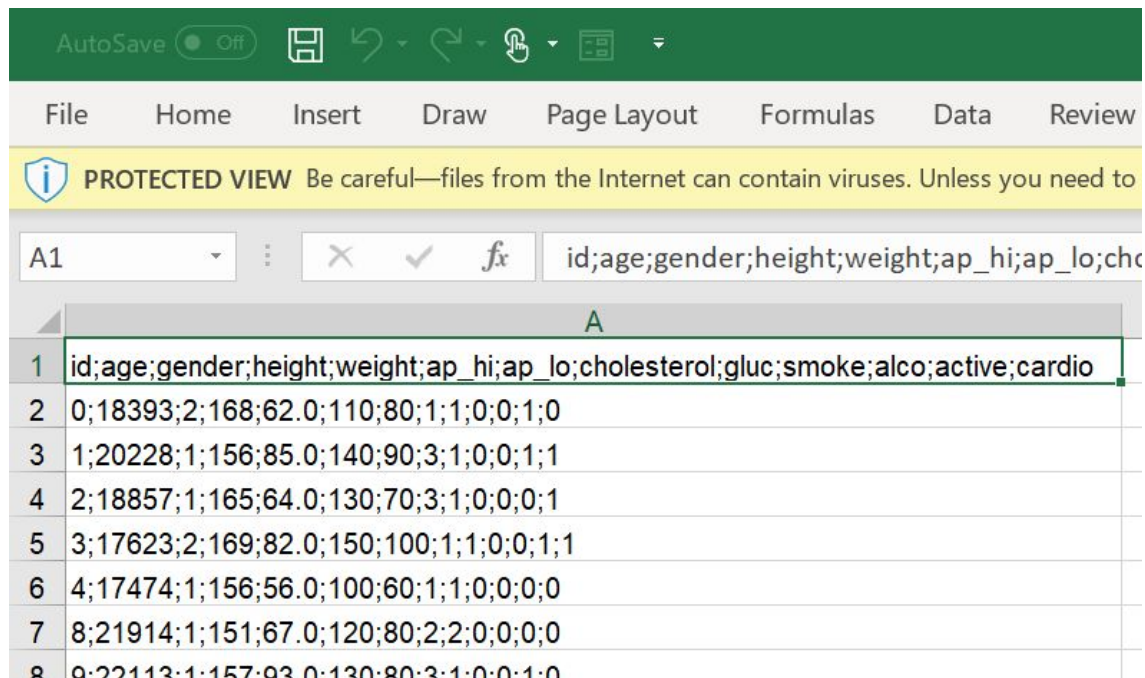


# Dataset: Data Cleaning

— — —

## Raw Dataset:

- 13 Attributes
- 70,000 instances



AutoSave Off

File Home Insert Draw Page Layout Formulas Data Review

**PROTECTED VIEW** Be careful—files from the Internet can contain viruses. Unless you need to

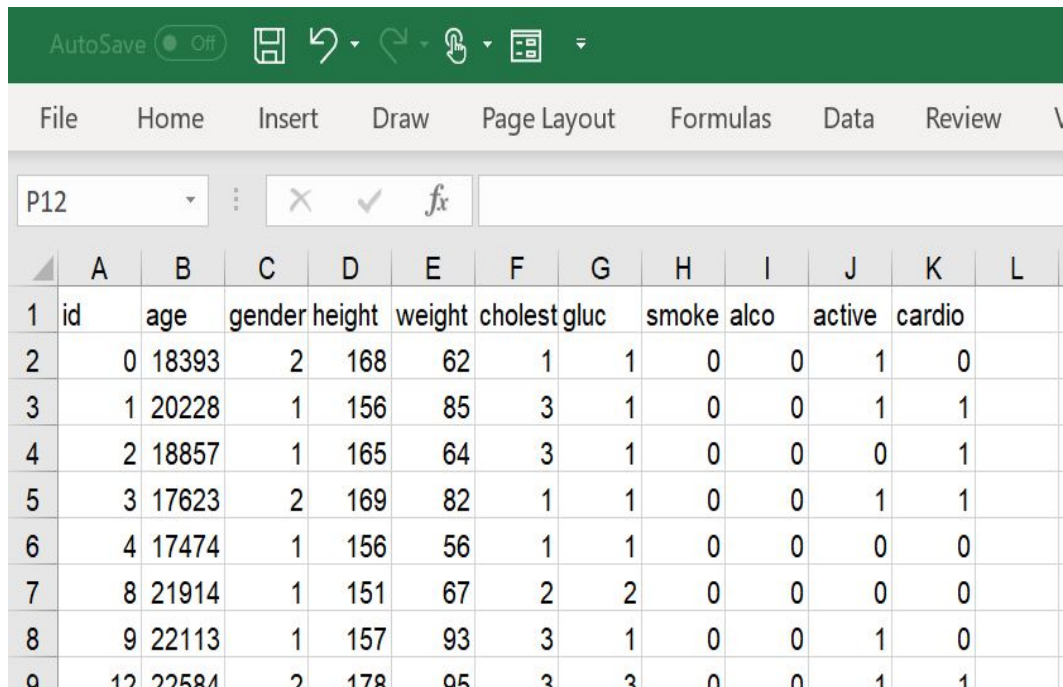
A1

	A
1	id;age;gender;height;weight;ap_hi;ap_lo;cholesterol;gluc;smoke;alco;active;cardio
2	0;18393;2;168;62.0;110;80;1;1;0;0;1;0
3	1;20228;1;156;85.0;140;90;3;1;0;0;1;1
4	2;18857;1;165;64.0;130;70;3;1;0;0;0;1
5	3;17623;2;169;82.0;150;100;1;1;0;0;1;1
6	4;17474;1;156;56.0;100;60;1;1;0;0;0;0
7	8;21914;1;151;67.0;120;80;2;2;0;0;0;0
8	0;22113;1;157;63.0;130;80;2;1;0;0;1;0

# Data Cleaning: Result

## Cleaned Data

- 10 applicable attributes
- Split data into columns by attribute



The screenshot shows a Microsoft Excel spreadsheet with a green ribbon at the top. The ribbon includes the 'AutoSave' button (set to 'Off'), a 'Save' icon, and several undo/redo icons. The main menu bar includes 'File', 'Home', 'Insert', 'Draw', 'Page Layout', 'Formulas', 'Data', 'Review', and 'View'. The active cell is P12. The spreadsheet contains a table with 12 columns (A-L) and 10 rows of data. The columns are labeled: 'id', 'age', 'gender', 'height', 'weight', 'cholest', 'gluc', 'smoke', 'alco', 'active', and 'cardio'. The data is as follows:

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	age	gender	height	weight	cholest	gluc	smoke	alco	active	cardio	
2	0	18393	2	168	62	1	1	0	0	1	0	
3	1	20228	1	156	85	3	1	0	0	1	1	
4	2	18857	1	165	64	3	1	0	0	0	1	
5	3	17623	2	169	82	1	1	0	0	1	1	
6	4	17474	1	156	56	1	1	0	0	0	0	
7	8	21914	1	151	67	2	2	0	0	0	0	
8	9	22113	1	157	93	3	1	0	0	1	0	
9	12	22584	2	178	85	3	3	0	0	1	1	

# Menu / Interface

— — —

## Menu

```
=====
Feature 1: Heart Disease Distribution
Feature 2: Attribute Correlation to Cardiovascular Disease
Feature 3: Key Attributes Analysis
Feature 4: Prediction
```

## Options

```
=====
If you want Feature 1, input 1
If you want Feature 2, input 2
If you want Feature 3, input 3
If you want Feature 4, input 4
If you do not want to continue, input exit
```

# Feature 1

— — —

## Heart Disease Distribution:

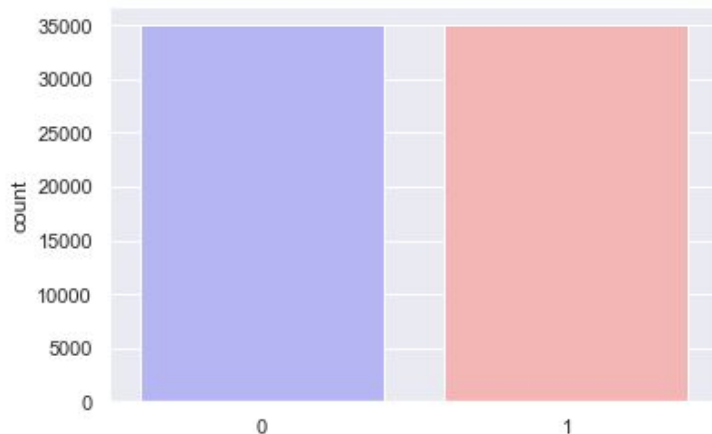
- Target variable is evenly distributed
- Why?

*To better help our model  
to find the patterns*

### Feature 1

% of patients have not had cardio disease: 50.03%

% of patients have had cardio disease: 49.97%



# Feature 2

---

## Attribute Correlation:

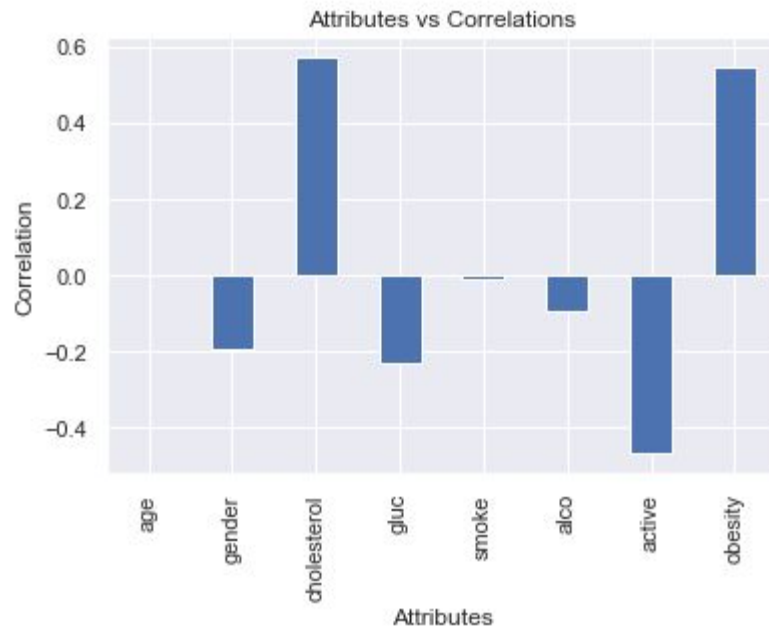
- To see how the attributes weight on disease
- Top 3 driven factors:

*Cholesterol*

*Gender*

*Active*

Feature 2



	Attribute	Coefficient
0	age	0.001346
1	gender	-0.195377
2	cholesterol	0.570289
3	gluc	-0.234476
4	smoke	-0.010327
5	alco	-0.097524
6	active	-0.468912
7	obesity	0.547765



# Feature 3

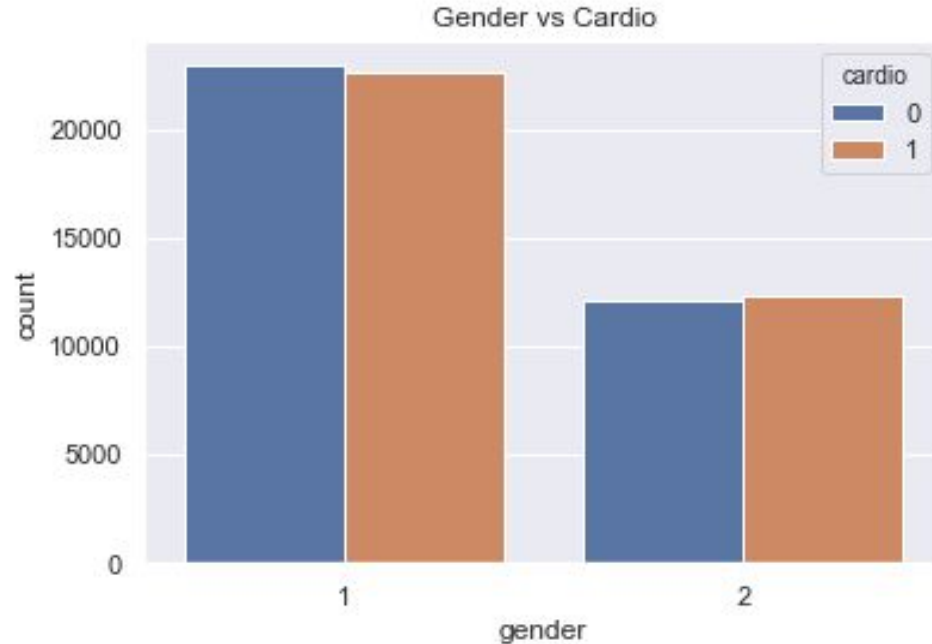
---

Based on the results from feature 2, our model will display the visualization of three key attributes (gender, cholesterol level, and active level) that lead to cardiovascular disease.



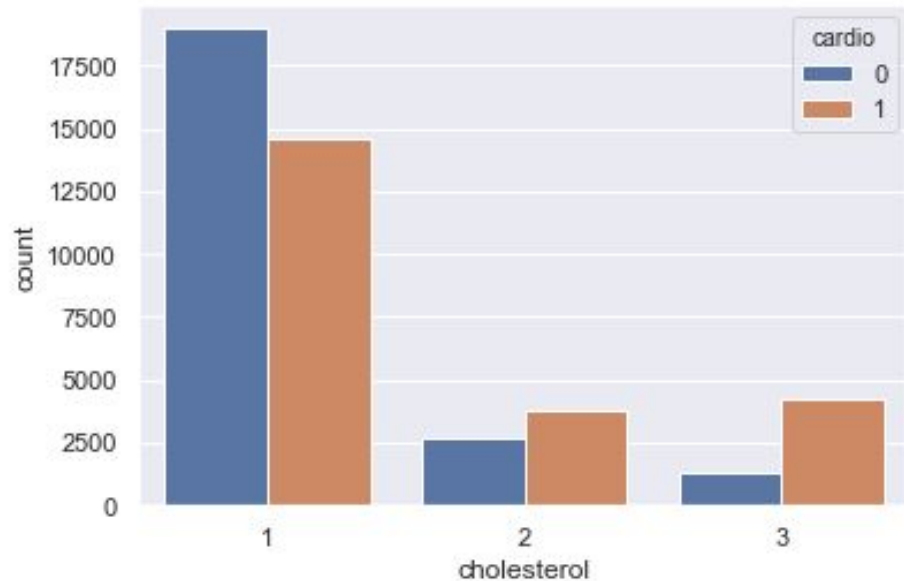
# Features 3 Result- Gender

Feature 3

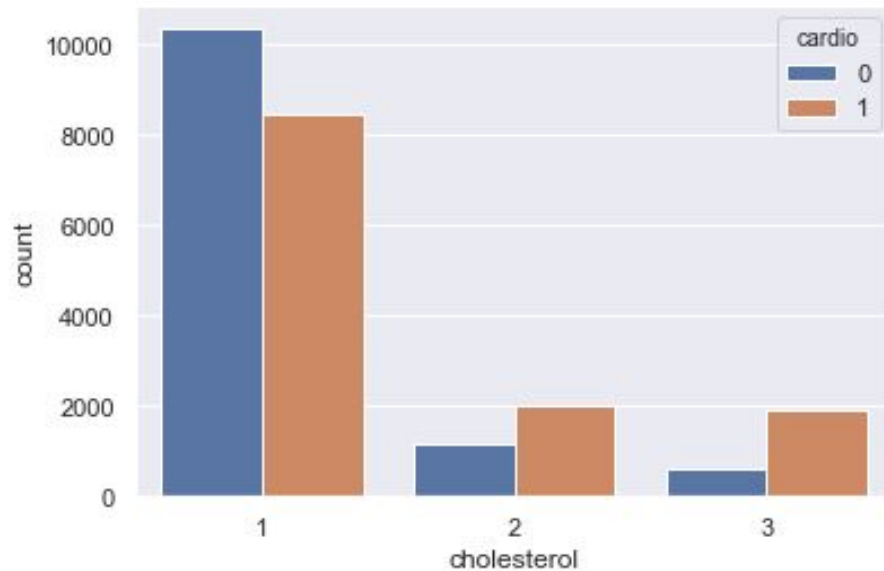


# Features 3 Result - Cholesterol Level

Cholesterol vs Cardio: Female



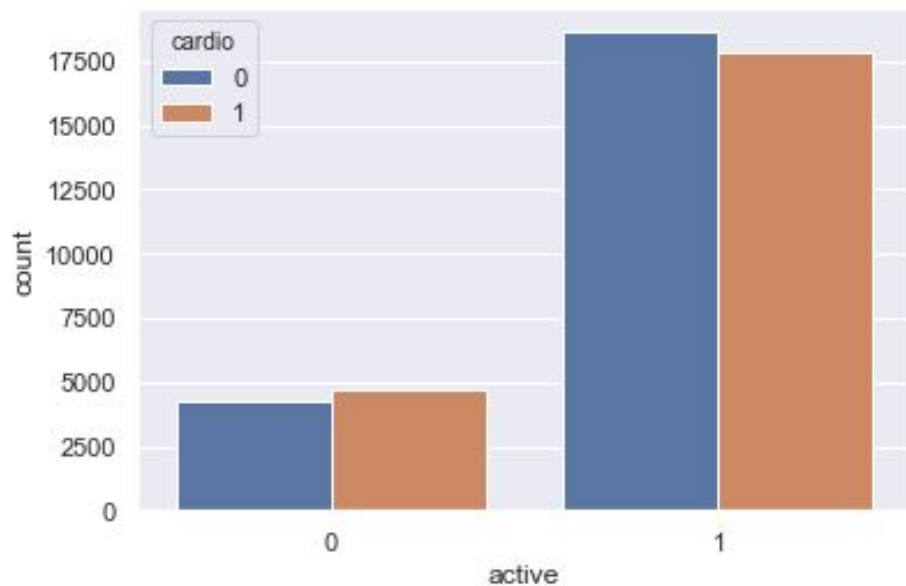
Cholesterol vs Cardio: Male



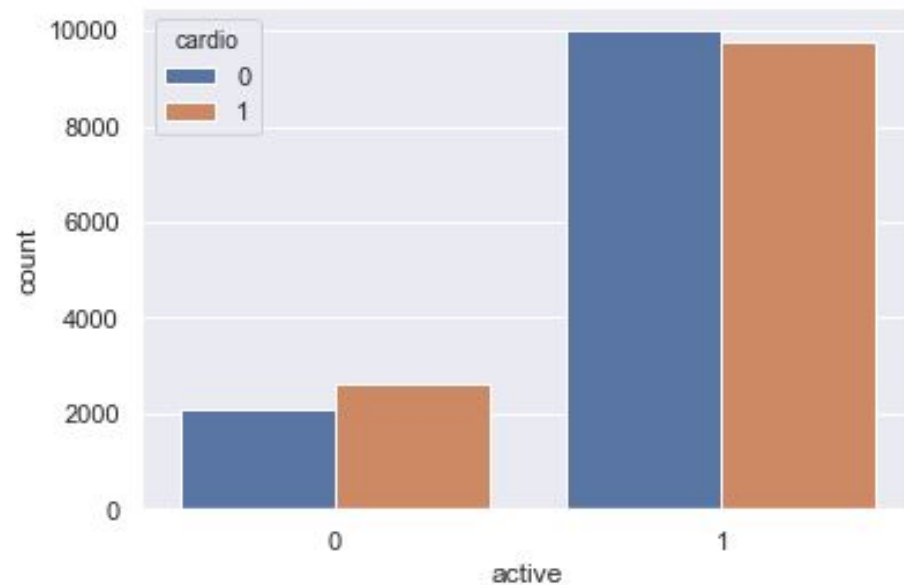
# Features 3 Result - Active Level

— — —

Active vs Cardio: Female



Active vs Cardio: Male



## Feature 4 - Prediction

---

- Predict the chance of whether a person gets cardiovascular disease or not based on user inputs

## Predict Heart Disease

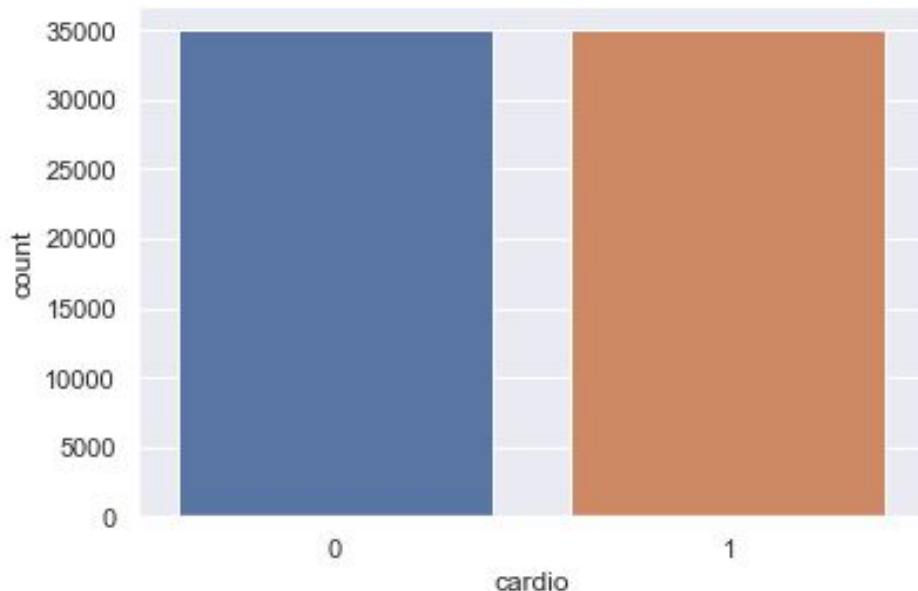


# Feature 4 - Logistic Regression

---

- Machine learning technique used to predict the probability of a categorical dependent variable
- Target variable: cardio

```
2 from sklearn import metrics  
3 from sklearn.linear_model import LogisticRegression
```



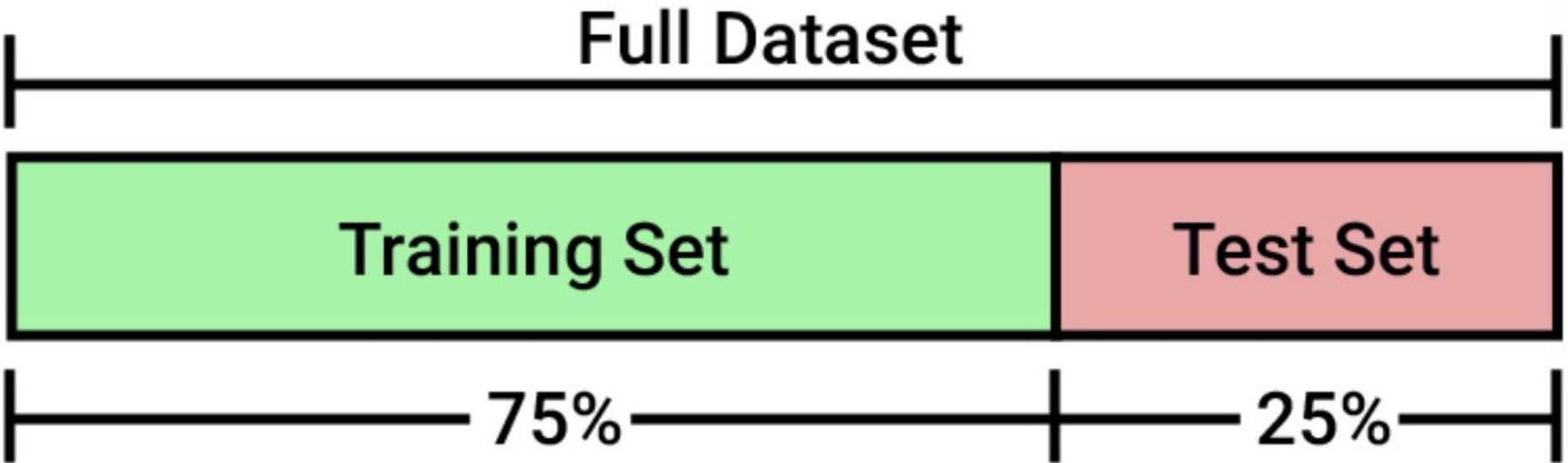
# Feature 4 - Data Training & Testing

---

```
6 from sklearn.model_selection import train_test_split
```

- Training data used to fit the model
- Test data used to validate the model
- Make prediction of variables on the test data instead of overfit or underfit the model

## Feature 4 - Data Training & Testing





# Potential Improvements

— — —



<u>Limitation</u>	<u>Solution</u>
Accuracy of the Prediction Model (64%~)	Optimize model <ul style="list-style-type: none"><li>- Narrow down the number of key attributes</li></ul>
Small dataset	Gather more data from clinics and hospitals
Better interface platform	Using the web application learned in class to optimize the model

**Let's run the code!**

**Thank you!**

**Any question?**