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

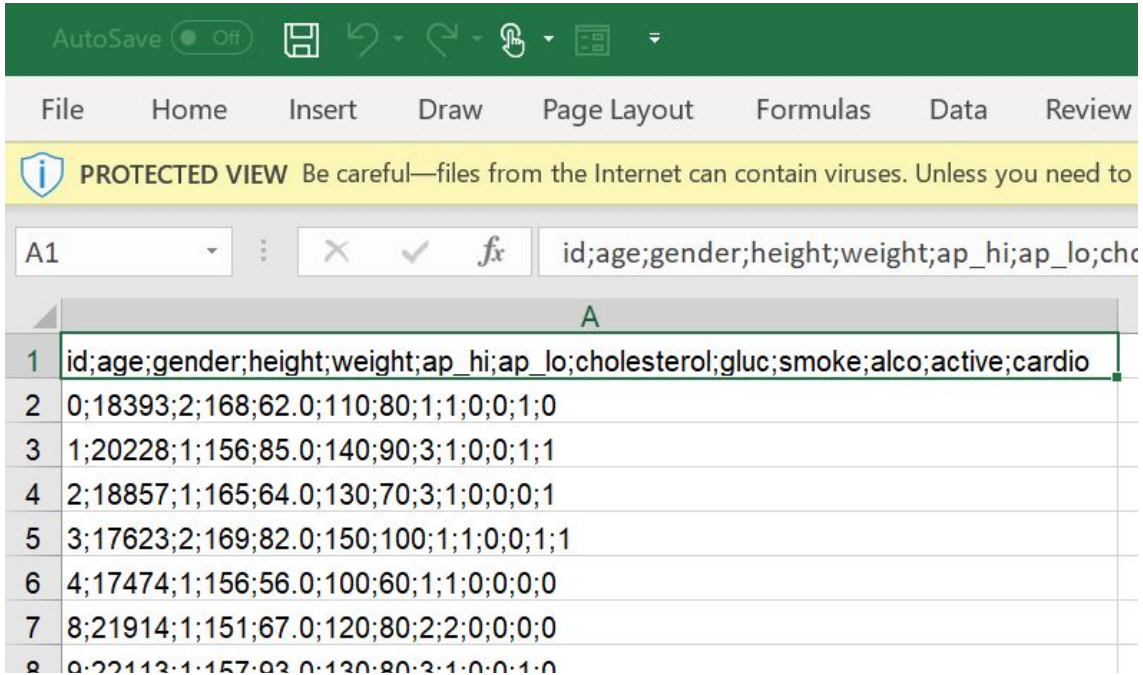
- Chance of getting cardiovascular disease
- Correlations of potential causes
- Prediction

Audience: Clinics & Hospitals

Dataset: Data Cleaning

Raw Dataset:

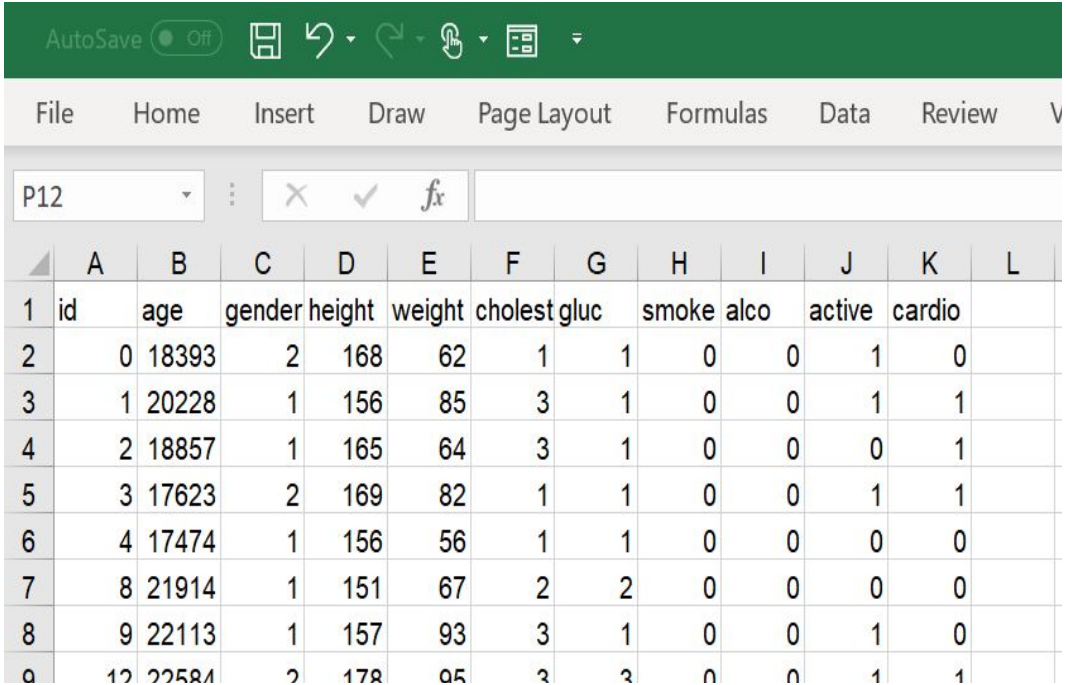
- 13 Attributes
- 70,000 instances



Data Cleaning: Result

Cleaned Data

- 10 applicable attributes
- Split data into columns by attribute



	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	95	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 Analaysis
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
```

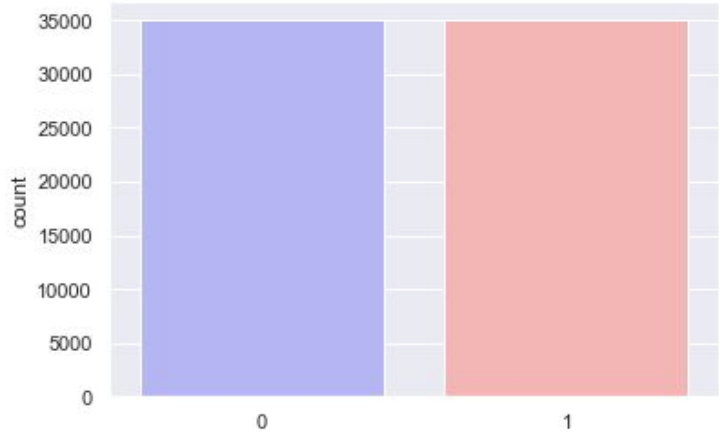
Feature 1

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%
```



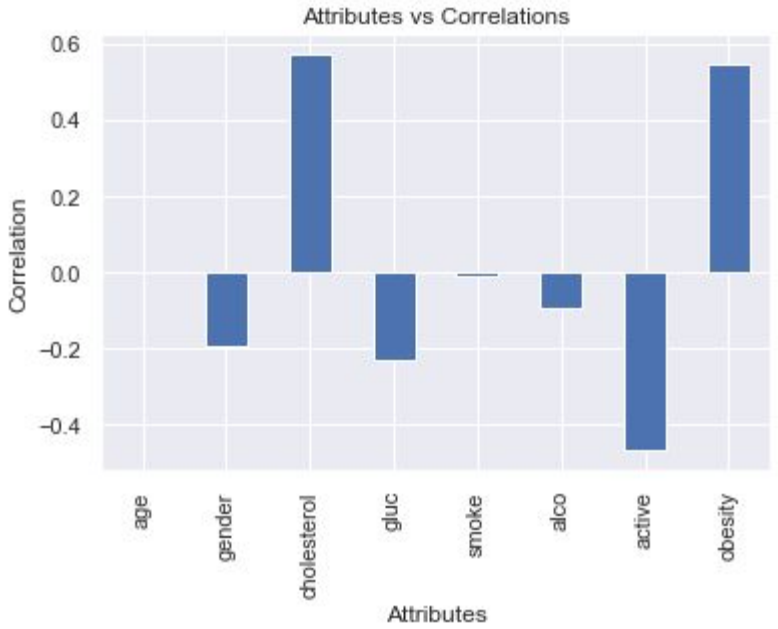
Feature 2

Attribute Correlation:

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

  Cholesterol
  Gender
  Active

Feature 2



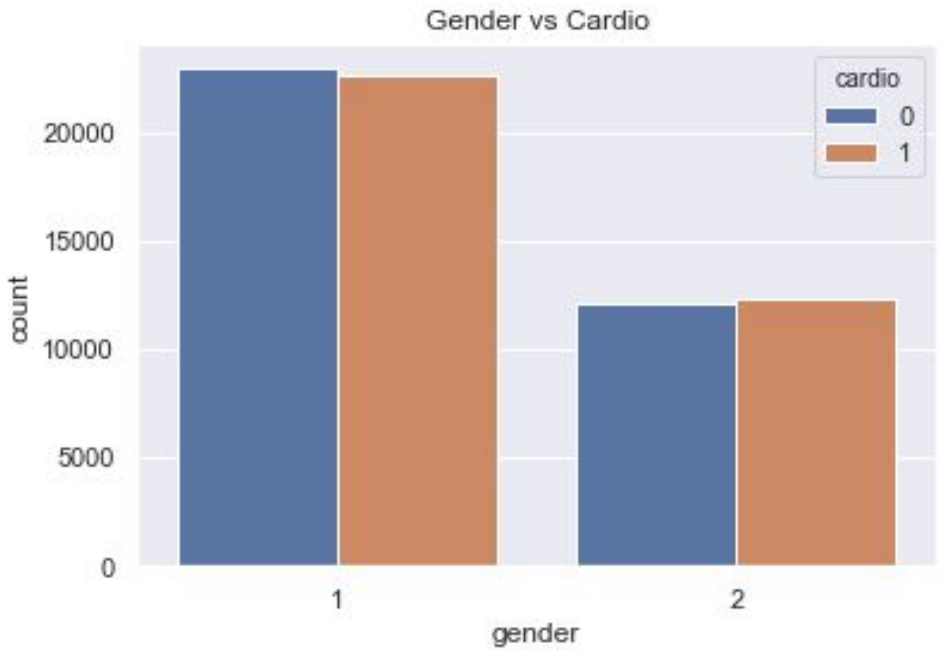
	Attribute	Coefficient
0	age	0.001346
1	gender	−0.195377
2	cholesterol	0.570289
3	gluc	−0.234476
4	smoke	−0.010327
5	alco	−0.097524
6	active	−0.468912
7	obesity	0.547765

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.

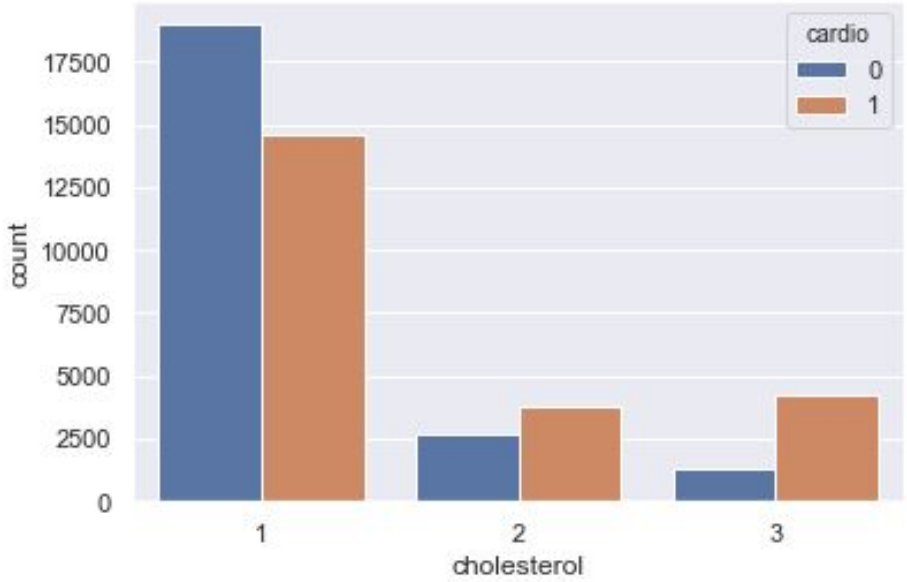
Features 3 Result- Gender

Feature 3

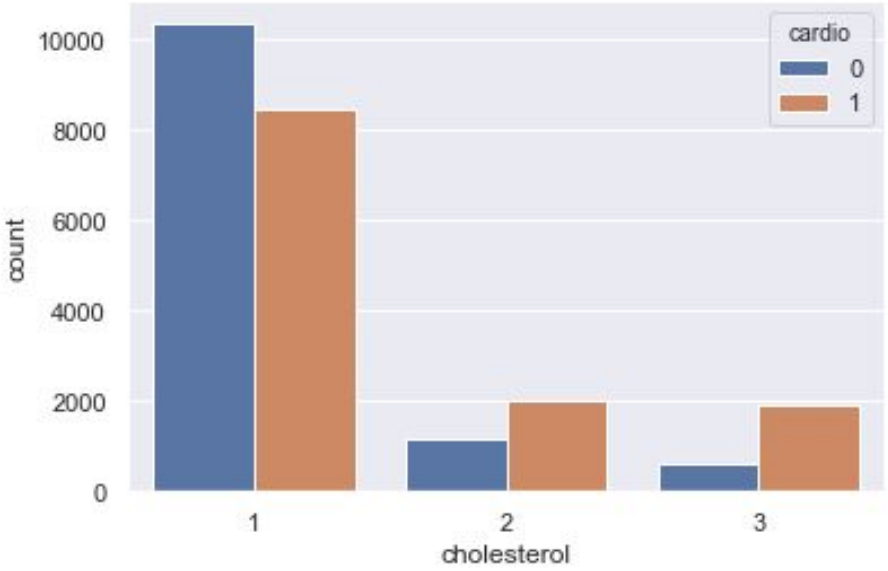


Features 3 Result - Cholesterol Level

Cholesterol vs Cardio: Female

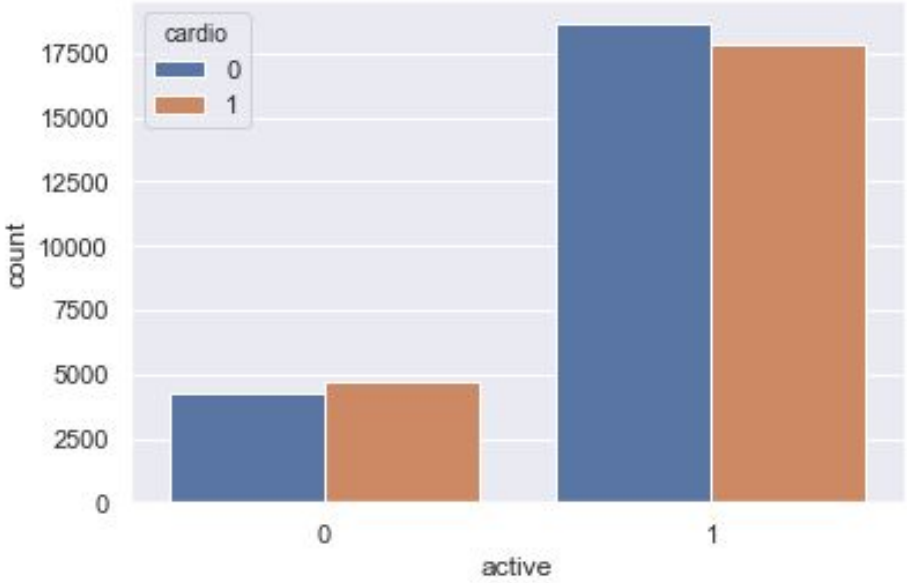


Cholesterol vs Cardio: Male

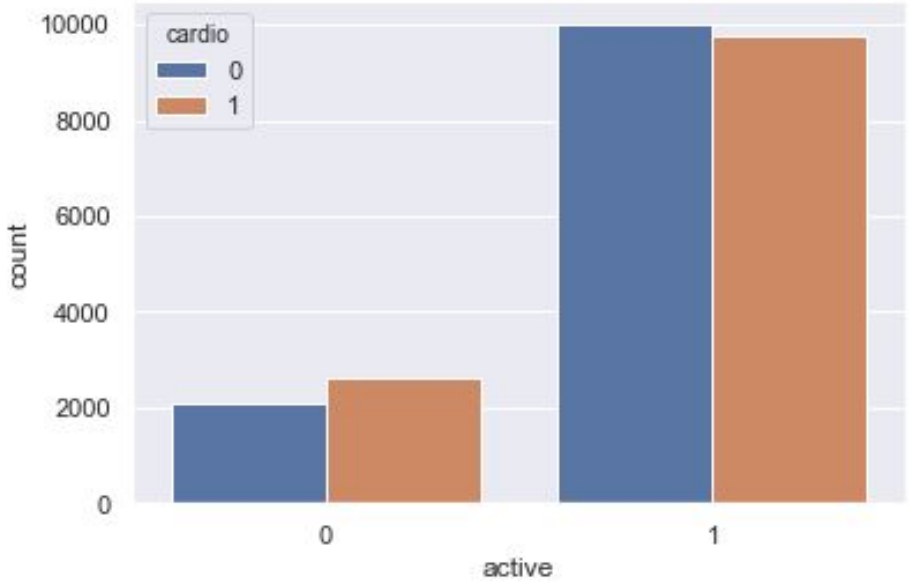


Features 3 Result - Active Level

Active vs Cardio: Female



Active vs Cardio: Male

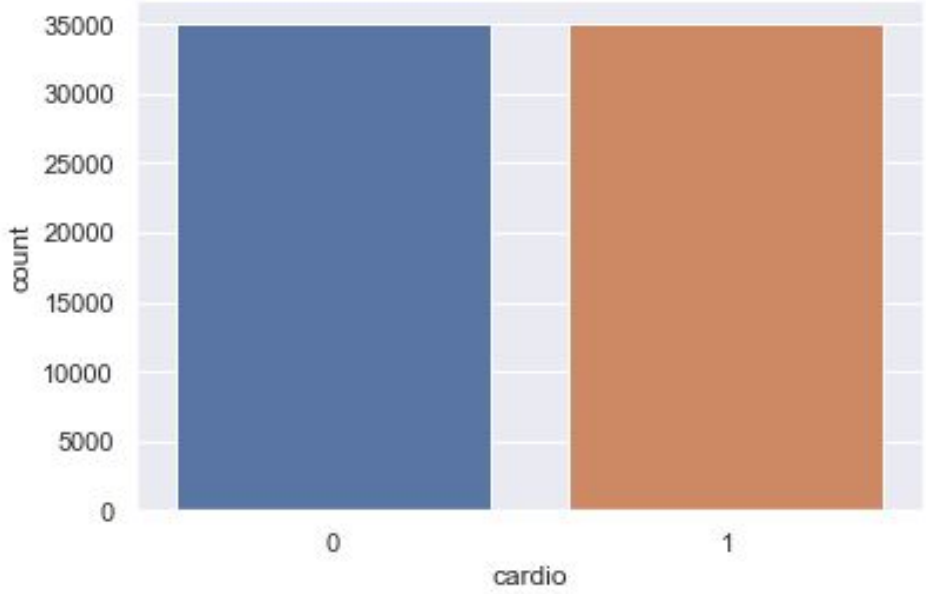


Feature 4 - Prediction

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

Feature 4 - Logistic Regression

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



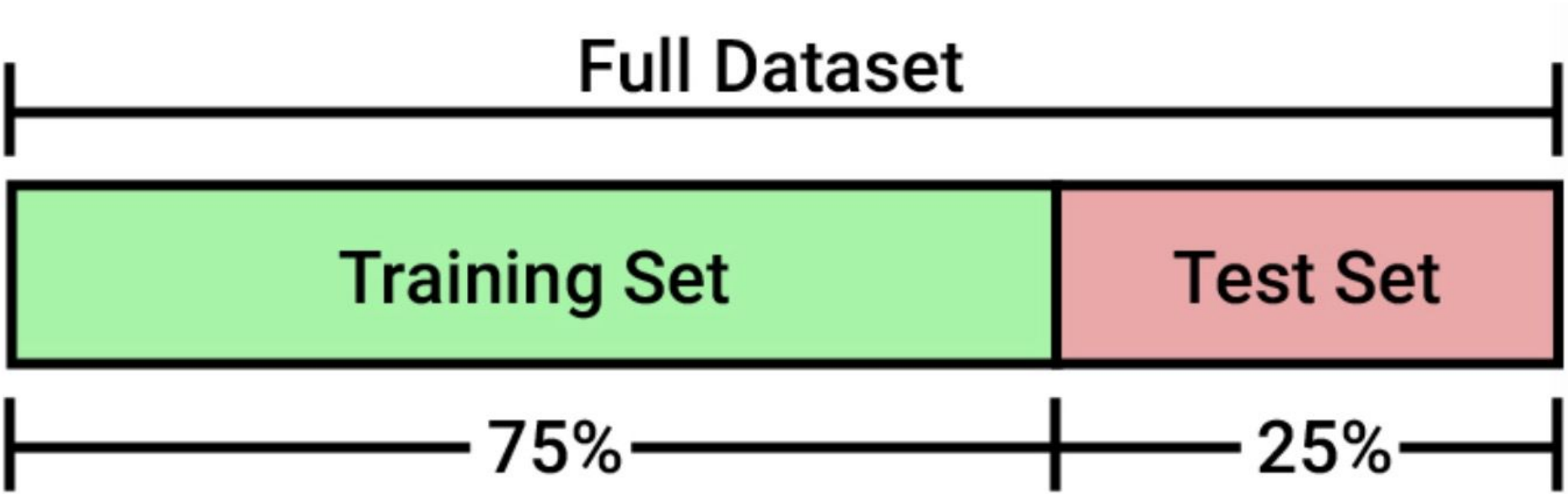

```
from sklearn import metrics
from sklearn.linear_model import LogisticRegression
```

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

Limitation	Solution
Accuracy of the Prediction Model (64%~)	Optimize model - Narrow down the number of key attributes
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?