ML | 01418496

Cardiovascular Disease Prediction using ML

THANAKORN WONGSANIT

Author

Thanakorn Wongsanit
6210407960
STUDENT

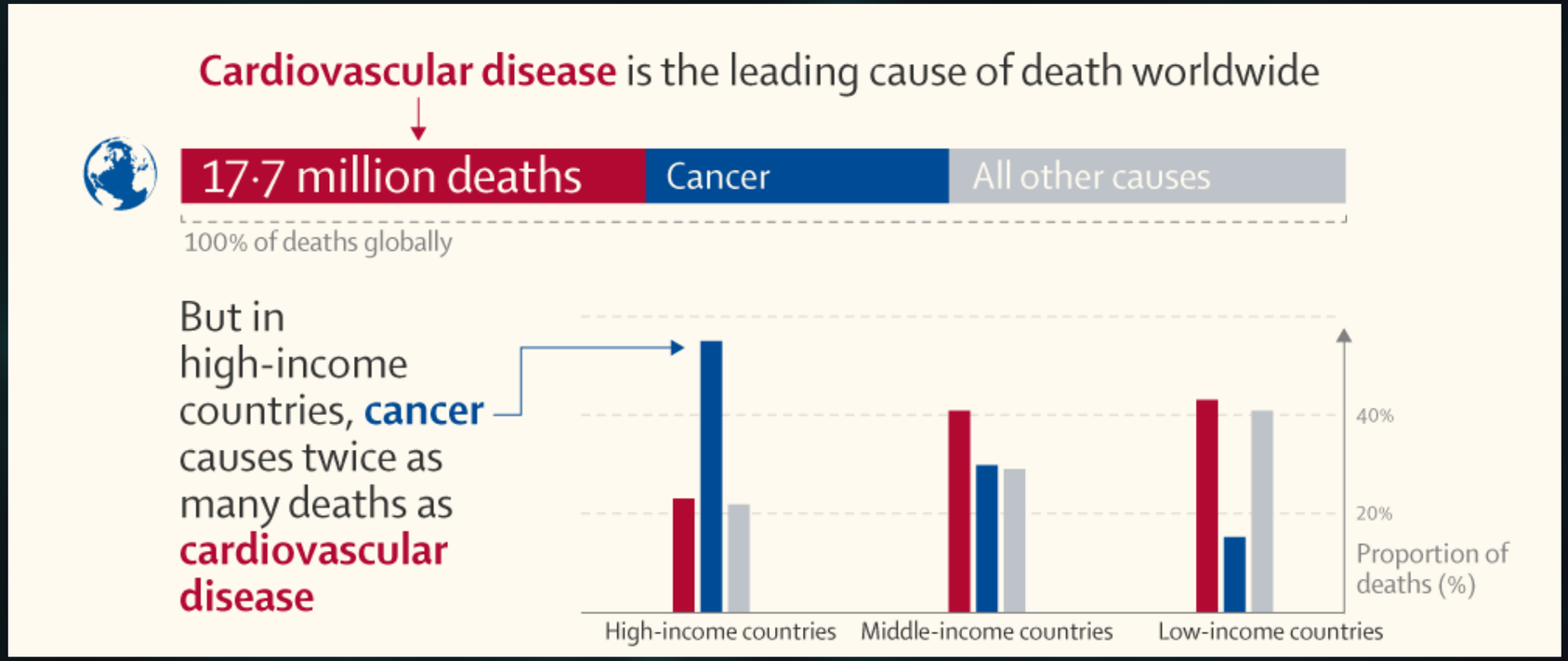
A third-year student in Computer Science at Kasetsart University

Cardiovascular disease

Leading cause of death worldwide

The most important behavioural risk factors

- unhealthy diet
- physical inactivity
- smoking
- alcohol intake
- obesity



important risk factors

Data

แหล่งข้อมูล

kaggle™

cvd_dataset.csv
ข้อมูลชุดนี้มีข้อมูลทั้งหมด 70000 แถว
และ 12 คอลัมน์ ซึ่งประกอบไปด้วยข้อมูลต่างๆที่เกี่ยวข้องกับผู้ป่วย
โดยข้อมูลสามารถแบ่งได้เป็น
3 ประเภท ได้แก่

- Objective: factual information

- Examination: results of medical examination

- Subjective: information given by the patient

Data

id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
85587	18919	1	154	61.0	95	60	1	1	0	0	1	0
72814	21170	1	164	69.0	120	80	1	1	0	0	1	1
17373	19666	1	155	107.0	123	86	1	1	0	0	0	1
76892	22033	2	181	82.0	120	60	1	2	0	0	0	1
33652	23555	1	165	77.0	120	80	3	3	0	0	1	1
96370	18743	2	185	57.0	120	80	1	1	0	0	0	0
54045	21657	1	174	135.0	140	90	1	1	0	0	0	1
92588	16163	1	166	83.0	130	80	2	2	1	0	1	0
44621	15991	2	169	80.0	130	80	1	1	0	0	1	0
94097	23219	1	157	80.0	140	90	3	2	0	0	1	1

วิเคราะห์ข้อมูลเบื้องต้น

```
cvd.describe()
```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc
count	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000
mean	19468.865814	1.349571	164.359229	74.205690	128.817286	96.630414	1.366871	1.226457
std	2467.251667	0.476838	8.210126	14.395757	154.011419	188.472530	0.680250	0.572270
min	10798.000000	1.000000	55.000000	10.000000	-150.000000	-70.000000	1.000000	1.000000
25%	17664.000000	1.000000	159.000000	65.000000	120.000000	80.000000	1.000000	1.000000
50%	19703.000000	1.000000	165.000000	72.000000	120.000000	80.000000	1.000000	1.000000
75%	21327.000000	2.000000	170.000000	82.000000	140.000000	90.000000	2.000000	1.000000
max	23713.000000	2.000000	250.000000	200.000000	16020.000000	11000.000000	3.000000	3.000000

```
cvd.groupby('gender')['height'].mean()
```

```
gender
1    161.355612
2    169.947895
Name: height, dtype: float64
```

จะเห็นว่า 2 มีความสูงเฉลี่ยมากกว่า 1 ดังนั้น 2 แทนเพศชาย

```
cvd['cardio'].value_counts(normalize=True)
```

```
0    0.5003
1    0.4997
Name: cardio, dtype: float64
```

วิเคราะห์ว่าข้อมูลมีคนที่เป็นโรคและไม่เป็นโรคใกล้เคียงกันหรือไม่

คัดกรองข้อมูลและเตรียมข้อมูล

1. **นำข้อมูลแถวที่มีค่าเป็น null ออก**

2. systolic blood pressure **ต้องมีค่ามากกว่า** diastolic blood pressure

3. **นำblood pressure ที่น้อยหรือมากเกินความเป็นจริงที่จะเป็นไปได้**

4. **ภาวะอ้วนนับเป็นหนึ่งในปัจจัยเสี่ยง**

เพิ่มfeature BMI

5. **Preprocessing**

StandardScaler()

6. **ทำการแบ่งกลุ่มของข้อมูล**

เพื่อที่จะทำการ train model และ test model

ML classification algorithms

Binary Classification

01 K-Nearest Neighbors

02 Logistic Regression

03 Support Vector Machine

04 Naive Bayes

train model

K-Fold Cross Validation

ทำให้ Dataset ที่มีการ Bias ของข้อมูล
Train Model ได้แม่นยำมากขึ้น

Search parameter

GridSearchCV ใน Scikit-Learn

Collect train model accuracy

เพื่อใช้ประกอบการวิเคราะห์เลือก model

Models evaluation

เรียง accuracy จาก test score

	Model	Score_train	Score_test
2	Support Vector Machine	73.31	72.87
1	Logistic Regression	72.33	72.01
0	KNN	74.22	71.58
3	Naive Bayes	70.91	71.10

เรียง accuracy จากผลต่างของคะแนนจากการ train และ test

	Model	Score_train	Score_test	Score_diff
3	Naive Bayes	70.91	71.10	0.19
1	Logistic Regression	72.33	72.01	0.32
2	Support Vector Machine	73.31	72.87	0.44
0	KNN	74.22	71.58	2.64

ดังนั้นจึงเลือก Logistic Regression

เรียง accuracy จาก test score

	Model	Score_train	Score_test
2	Support Vector Machine	73.31	72.87
1	Logistic Regression	72.33	72.01
0	KNN	74.22	71.58
3	Naive Bayes	70.91	71.10

เรียง accuracy จากผลต่างของคะแนนจากการ train และ test

	Model	Score_train	Score_test	Score_diff
3	Naive Bayes	70.91	71.10	0.19
1	Logistic Regression	72.33	72.01	0.32
2	Support Vector Machine	73.31	72.87	0.44
0	KNN	74.22	71.58	2.64

Live Demo

Web Application

เพิ่มเติม
เปรียบเทียบ SVM
แบบ linear และ polynomial

เรียง accuracy จาก test score

	Model	Score_train	Score_test
0	แบบ Linear	72.28	72.01
1	แบบ Polynomial	71.67	71.24

เรียง accuracy จากผลต่างของคะแนนจากการ train และ test

	Model	Score_train	Score_test	Score_diff
0	แบบ Linear	72.28	72.01	0.27
1	แบบ Polynomial	71.67	71.24	0.43