

Exploratory Data Analysis and Cardiovascular Diseases Risk Prediction

การวิเคราะห์ข้อมูลและทำนายโรคหัวใจด้วยวิธี การเรียนรู้ของเครื่อง





5 Stage Of Data Science Process

- Ask an interesting question
- Get the data
- Explore the date
- Model the data
- Communicate and visualize the results



ที่มาและความสำคัญของปัญหา

โรคหัวใจและหลอดเลือด (cardiovascular diseases) เป็นกลุ่มโรคที่เกิดกับระบบหัวใจและหลอดเลือดซึ่งเป็นสาเหตุการเสียชีวิตลำดับต้นของคนไทย ผู้จัดทำจึงสนใจวิเคราะห์ข้อมูลเชิงสำรวจและการทนายความเสียงการเป็นโรคหัวใจและหลอดเลือด



วัตถุประสงค์

- เพื่อวิเคราะห์ข้อมูลเชิงสำรวจโดยวิธี Exploratory data analysis (EDA)
- เพื่อทำการเป็นโรคหัวใจโดยการสร้าง Machine Learning Model ด้วยวิธี Logistic Regression, Decision Tree และ Random Forest และวัดประสิทธิภาพของโมเดลทั้ง 3 แบบด้วย Confusion Matrix, AUC-ROC Curve, Validation Curve และ Learning Curve

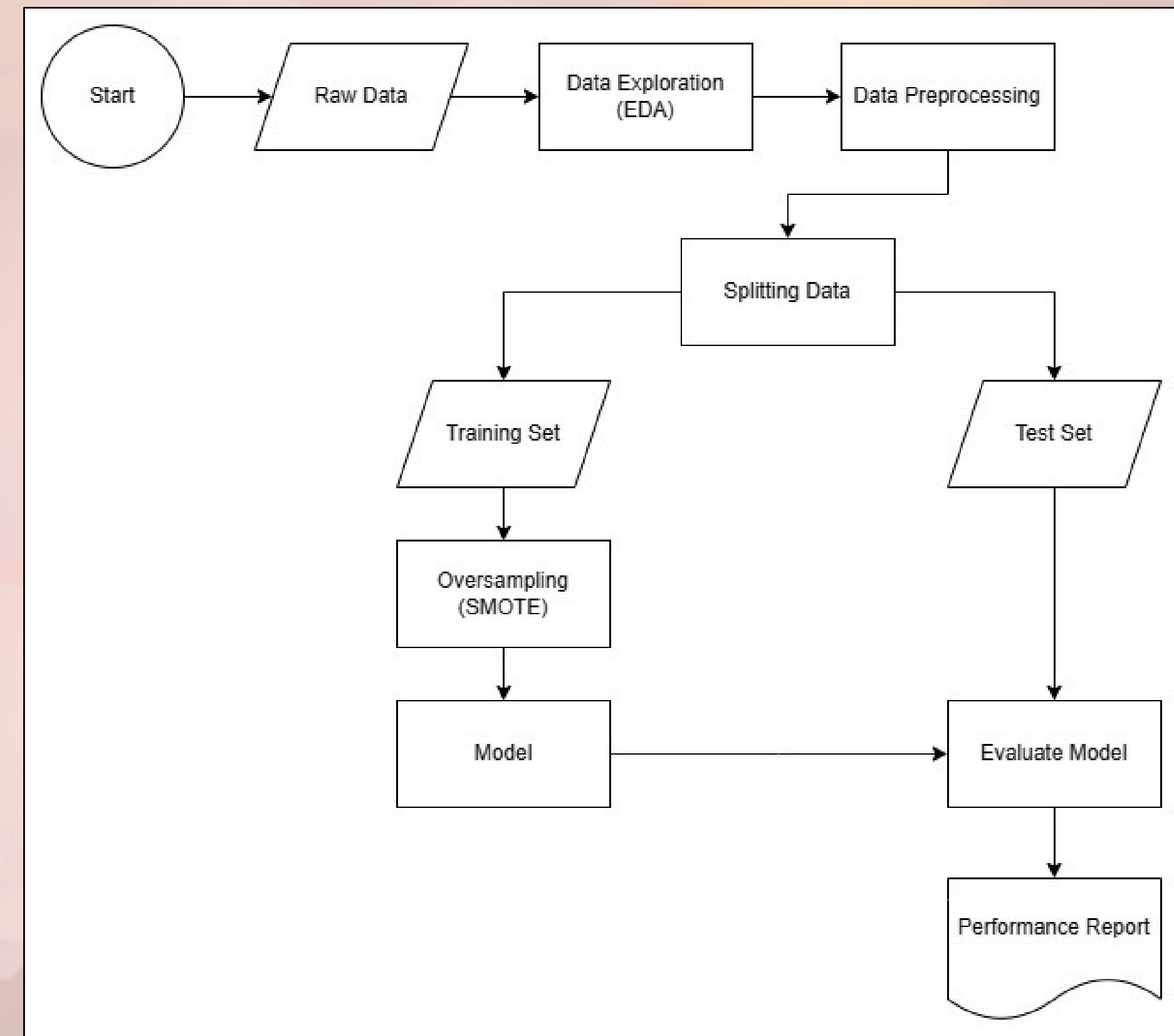
ที่มาของข้อมูล

ข้อมูลที่ใช้ในการวิเคราะห์ข้อมูลเชิงสำรวจและทำนายการเป็นโรคหัวใจนี้นำมาจากเว็บไซต์ <https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset>

ลักษณะของฐานข้อมูล (Dataset) มีทั้งหมด 308855 แถว 19 คอลัม्स โดยแบ่งดังนี้

ตัวแปร (Variables)	คำอธิบาย (Definition)	ประเภทของข้อมูล (Data Types)
General Health	ระดับสุขภาพโดยทั่วไป	Ordinal
Checkup	ช่วงเวลาครั้งล่าสุดที่ตรวจสุขภาพ	Ordinal
Age (Category)	ช่วงอายุ	Nominal
Sex	เพศ	Nominal
Exercise	ออกกำลังกายหรือไม่	Nominal
Heart Disease	เป็นโรคเกี่ยวกับหัวใจและหลอดหรือไม่	Nominal
Skin Cancer	เป็นโรคมะเร็งผิวนังหรือไม่	Nominal
Other Cancer	เป็นโรคอื่นๆที่เกี่ยวข้องกับมะเร็งหรือไม่	Nominal
Depression	เป็นโรคซึมเศร้าหรือไม่	Nominal
Diabetes	เป็นโรคเบาหวานหรือไม่	Nominal
Arthritis	เป็นโรคข้อเสื่อมหรือไม่	Nominal
Smoking History	มีประวัติการสูบบุหรี่หรือไม่	Nominal
Height	ส่วนสูง	Continuous
Weight	น้ำหนัก	Continuous
BMI	ดัชนีมวลกาย	Continuous
Alcohol consumption	ปริมาณการดื่มแอลกอฮอล์ (ก) / เดือน	Continuous
Fruit consumption	ปริมาณการกินผลไม้ (ก) / เดือน	Continuous
Green vegetable consumption	ปริมาณการกินผักใบเขียว (ก) / เดือน	Continuous
Fried potato consumption	ปริมาณการกินมันฝรั่งทอด (ก) / เดือน	Continuous

กระบวนการในการดำเนินงาน



ตรวจสอบข้อมูลที่หายไป (Check Missing Value)

```
[ ] df.isnull().sum()
```

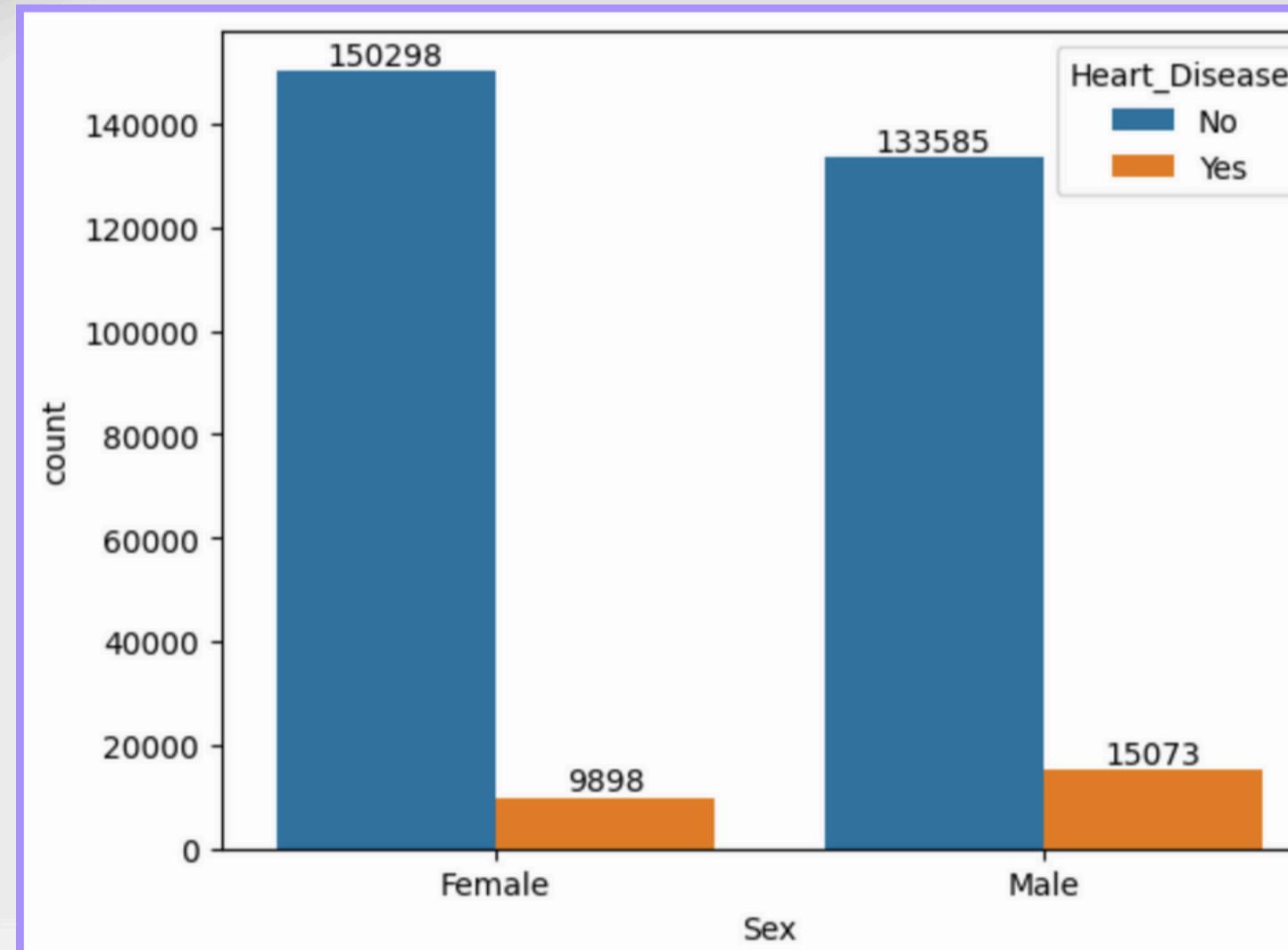
General_Health	0
Checkup	0
Exercise	0
Heart_Disease	0
Skin_Cancer	0
Other_Cancer	0
Depression	0
Diabetes	0
Arthritis	0
Sex	0
Age_Category	0
Height_(cm)	0
Weight_(kg)	0
BMI	0
Smoking_History	0
Alcohol_Consumption	0
Fruit_Consumption	0
Green_Vegetables_Consumption	0
FriedPotato_Consumption	0
dtype: int64	

ปรากฏว่าไม่มีข้อมูลในตัวแปรใดขาดหายไป แสดงว่าสามารถทำ EDA ต่อไปได้

A close-up photograph of a doctor's torso. The doctor is wearing a white medical coat over a light blue dress shirt and a dark blue patterned tie. A blue stethoscope hangs around their neck. The background is a plain, light color.

Exploratory Data Analysis (EDA)

ความสัมพันธ์ของความถี่ในการเป็นโรคหัวใจโดยแบ่งตามเพศ

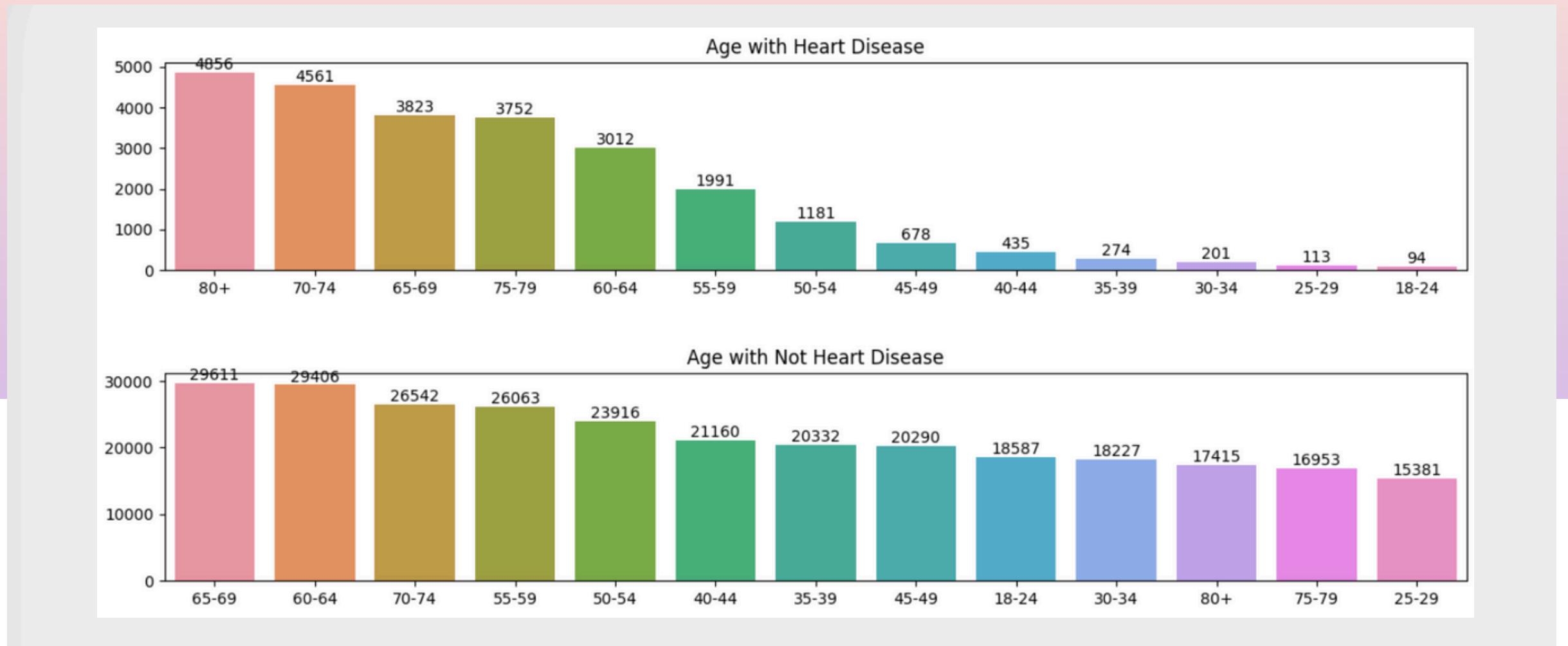


กราฟแท่งแสดงถึงความสัมพันธ์ของความถี่ในการเป็นโรคหัวใจในแต่ละเพศ

สรุปได้ว่าคนที่ไม่ได้เป็นโรคหัวใจในเพศหญิงมีความถี่มากกว่าคนที่ไม่ได้เป็นโรคหัวใจในเพศชาย ในขณะที่คนที่เป็นโรคหัวใจในเพศชายมีความถี่มากกว่าคนที่เป็นโรคหัวใจในเพศหญิง



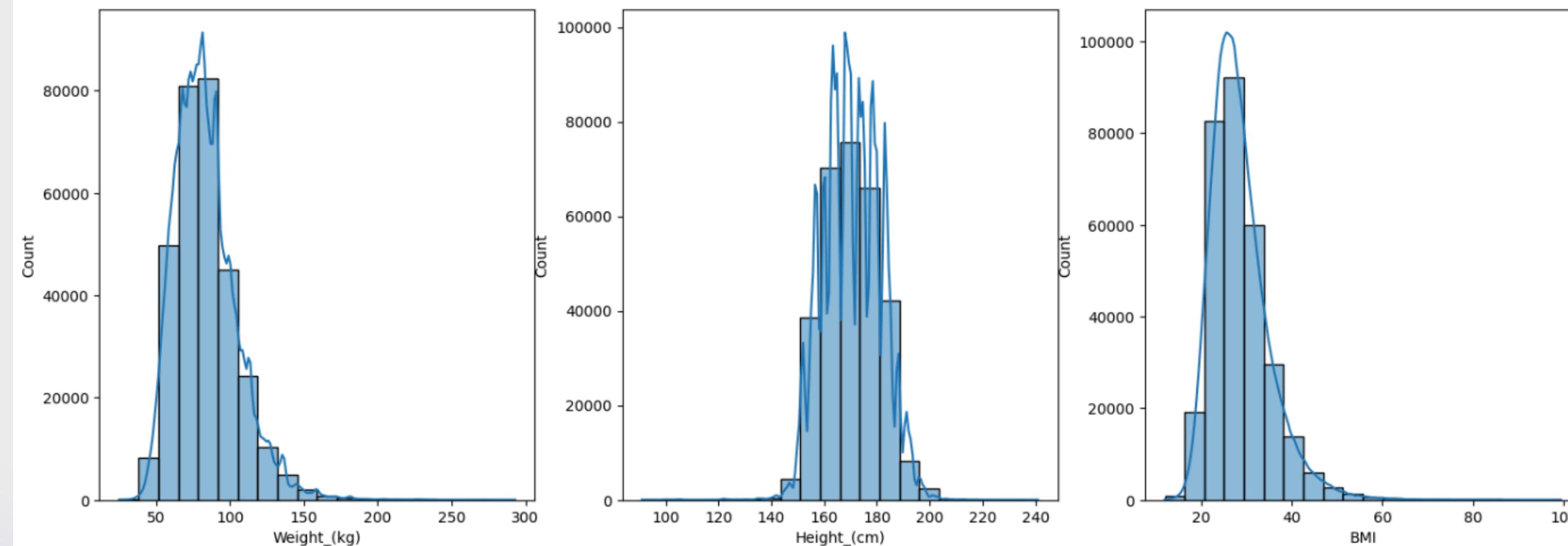
ความสัมพันธ์ของความถี่ในการเป็นโรคหัวใจในแต่ละช่วงอายุ



กราฟแท่งแสดงถึงความสัมพันธ์ของความถี่ในการเป็นโรคหัวใจในแต่ละช่วงอายุ

สรุปได้ว่าคนที่ไม่ได้เป็นโรคหัวใจส่วนใหญ่จะอายุช่วง 65-69 รองลงมาคือ อายุช่วง 60-64 และคนที่เป็นโรคหัวใจส่วนใหญ่จะมีอายุช่วง 80+ รองลงมา คืออายุช่วง 70-74

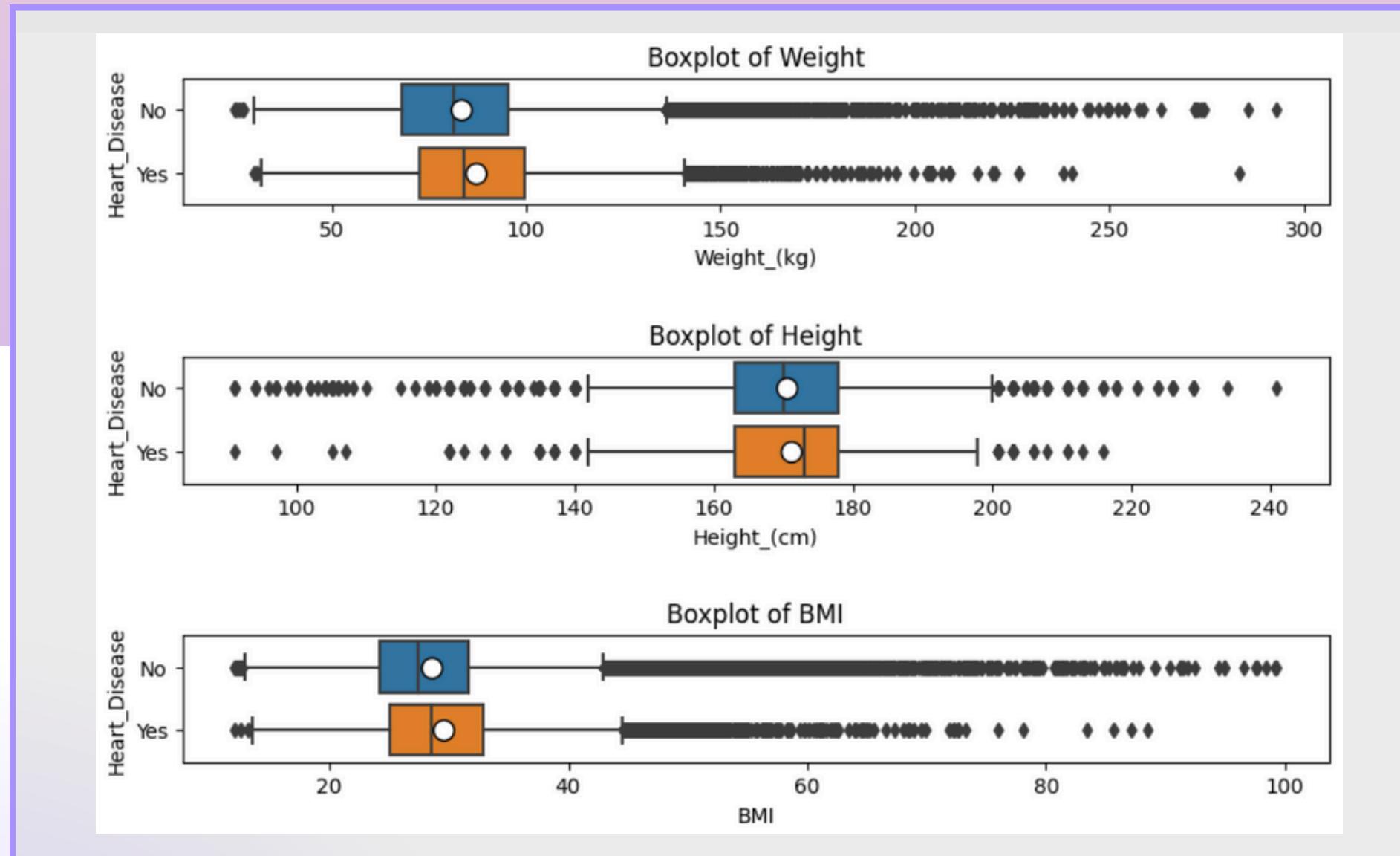
การแจกแจงของค่าน้ำหนัก, ส่วนสูง และ BMI ของคนที่ตอบแบบสอบถามทั้งหมด



กราฟ Histogram ของน้ำหนัก, ส่วนสูง และค่า BMI ของผู้ตอบแบบสอบถามทั้งหมด

สรุปได้ว่าน้ำหนักดูมีแนวโน้มการแจกแจงแบบเบี้ยว ส่วนสูงดูมีแนวโน้มว่ามีการแจกแจงแบบเส้นโค้งปกติ และค่า BMI ดูมีแนวโน้มการแจกแจงแบบเบี้ยวเหมือนกับน้ำหนัก

Boxplot ของน้ำหนัก, ส่วนสูง และ BMI ของคนที่ตอบแบบสอบถามทั้งหมด



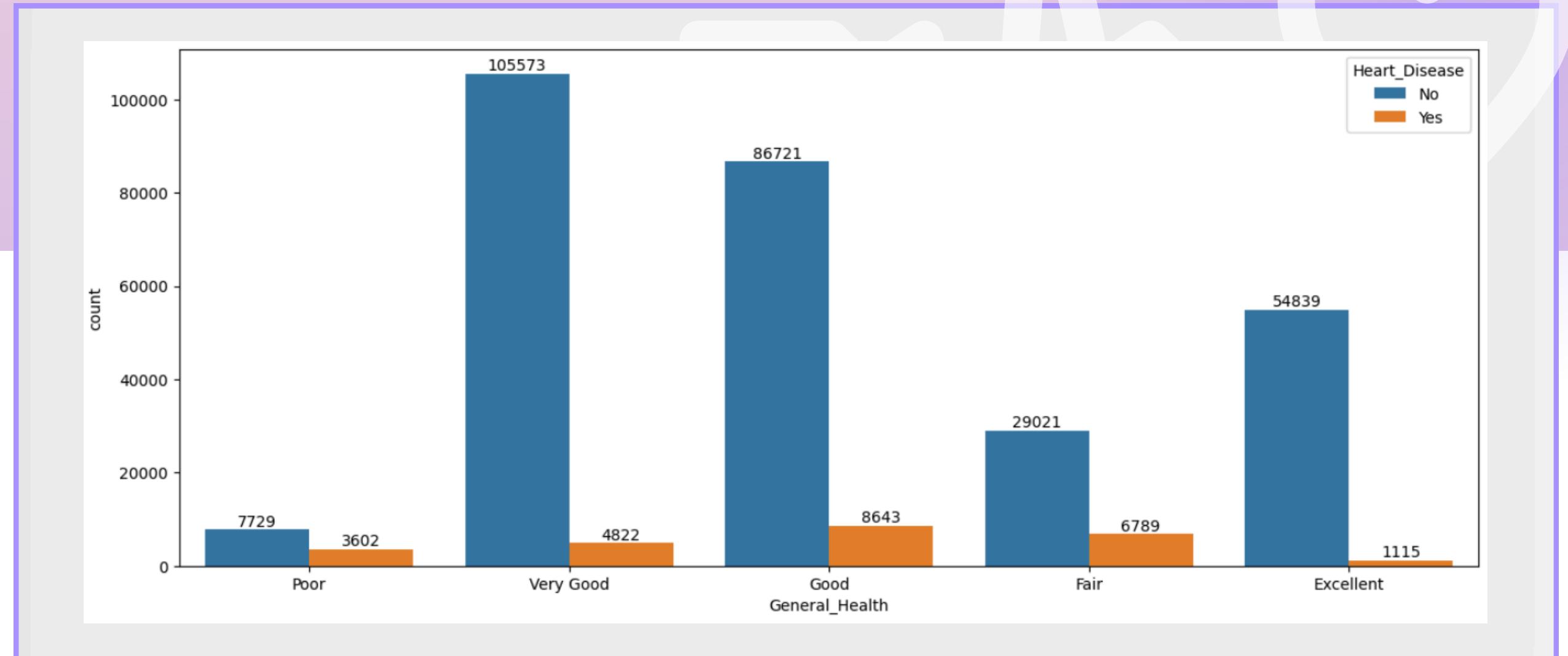
Variables	Mean	Median	Mode
Weight (kg)	83.59	81.65	90.72
Height (cm)	170.61	170.00	168.00
BMI	28.62	27.44	26.63

ค่าเฉลี่ย มัธยฐาน และฐานนิยม ของน้ำหนัก ส่วนสูงและ BMI

Pixel heart icon: Boxplot ของน้ำหนัก, ส่วนสูง และค่า BMI ของผู้ตอบแบบสอบถามทั้งหมด

สรุปได้ว่า ค่าเฉลี่ยของน้ำหนักประมาณ 84 kg ส่วนสูงประมาณ 170 cm และ BMI ประมาณ 28.62 และ BMI มีค่านอกเกณฑ์ (Outlier) ที่น่าจะมากที่สุดรองลงมาคือ น้ำหนักและส่วนสูง

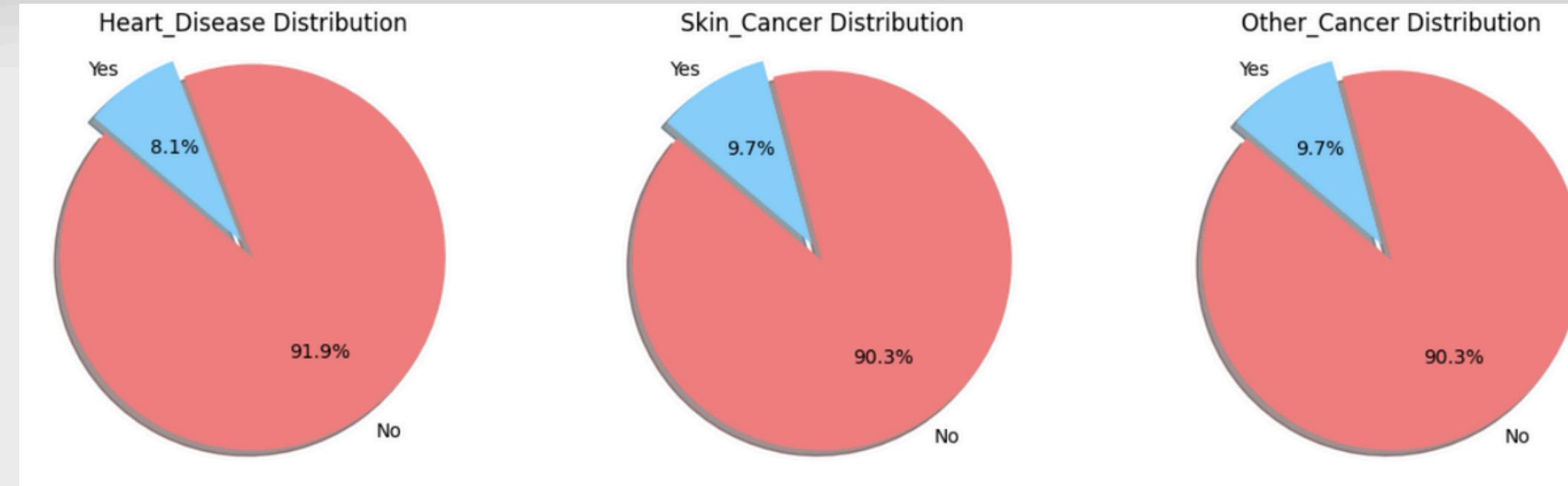
ความสัมพันธ์ของความถี่ในการเป็นโรคหัวใจในแต่ช่วงระดับสุขภาพ



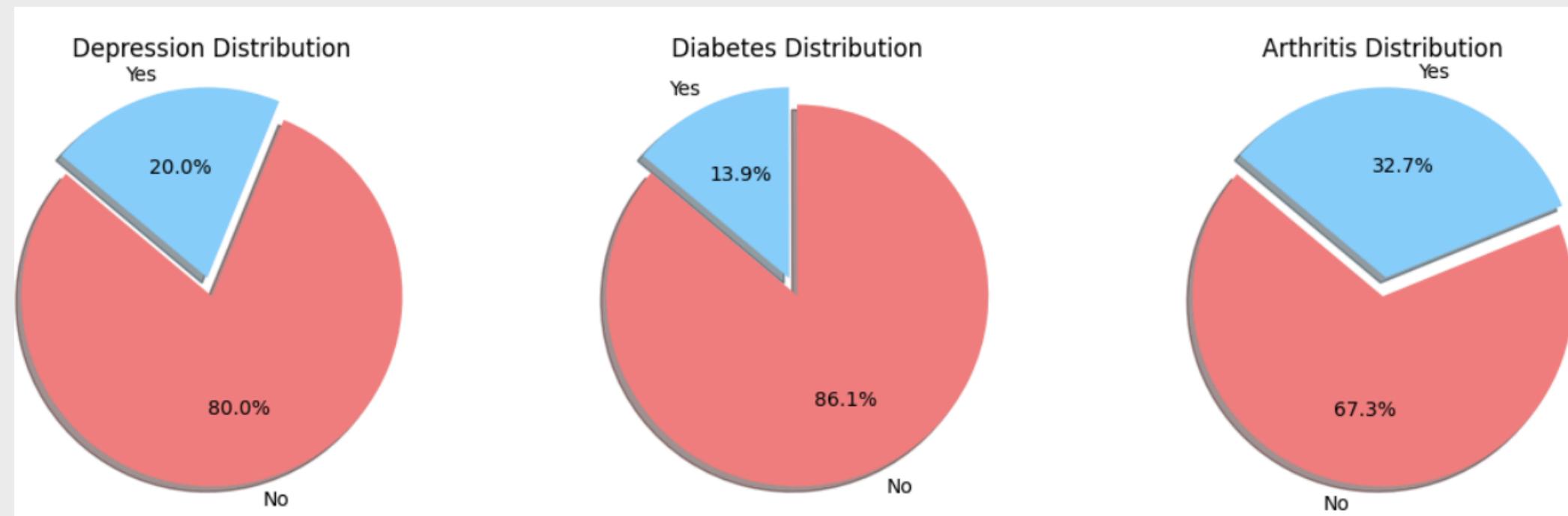
กราฟแท่งแสดงความสัมพันธ์ของความถี่ในการเป็นโรคหัวใจในแต่ช่วงระดับสุขภาพ

สรุปได้ว่าคนที่เป็นโรคหัวใจ ส่วนใหญ่มีระดับสุขภาพอยู่ที่ดี (Good) รองลงมาคืออยู่ในระดับปานกลาง (Fair) และคนที่ไม่ได้เป็นโรคหัวใจ ส่วนใหญ่มีระดับสุขภาพอยู่ที่ดีมาก (Very Good) รองลงมาคืออยู่ในระดับดี (Good)

สัดส่วนของคนที่เป็นและไม่เป็นโรคหัวใจ โรคมะเร็งผิวหนัง โรคมะเร็งอื่นๆ โรคซึมเศร้า โรคเบาหวาน และโรคข้ออักเสบ

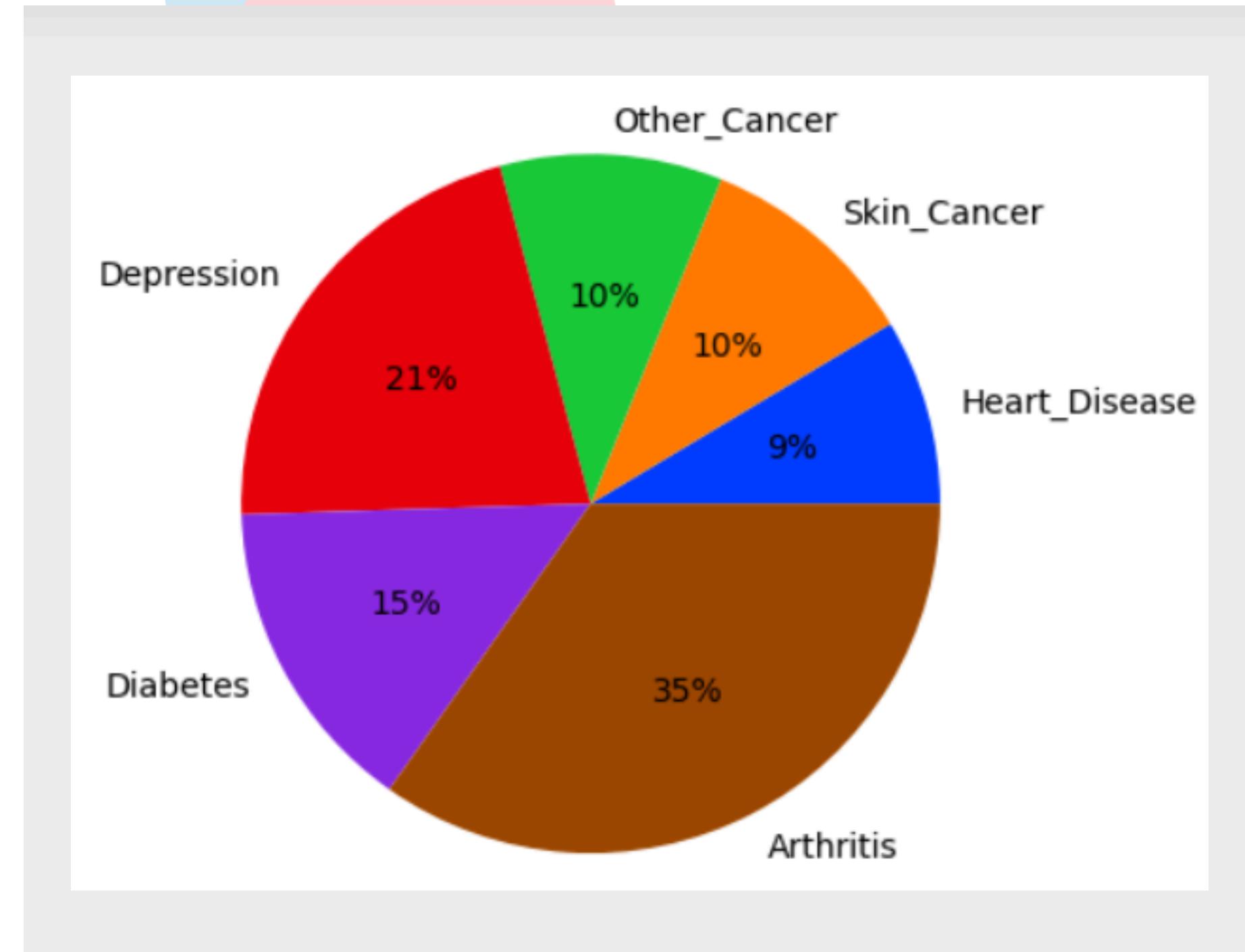


Pie chart ของคนที่เป็นโรคหัวใจ โรคมะเร็งผิวหนัง โรคมะเร็งอื่นๆ

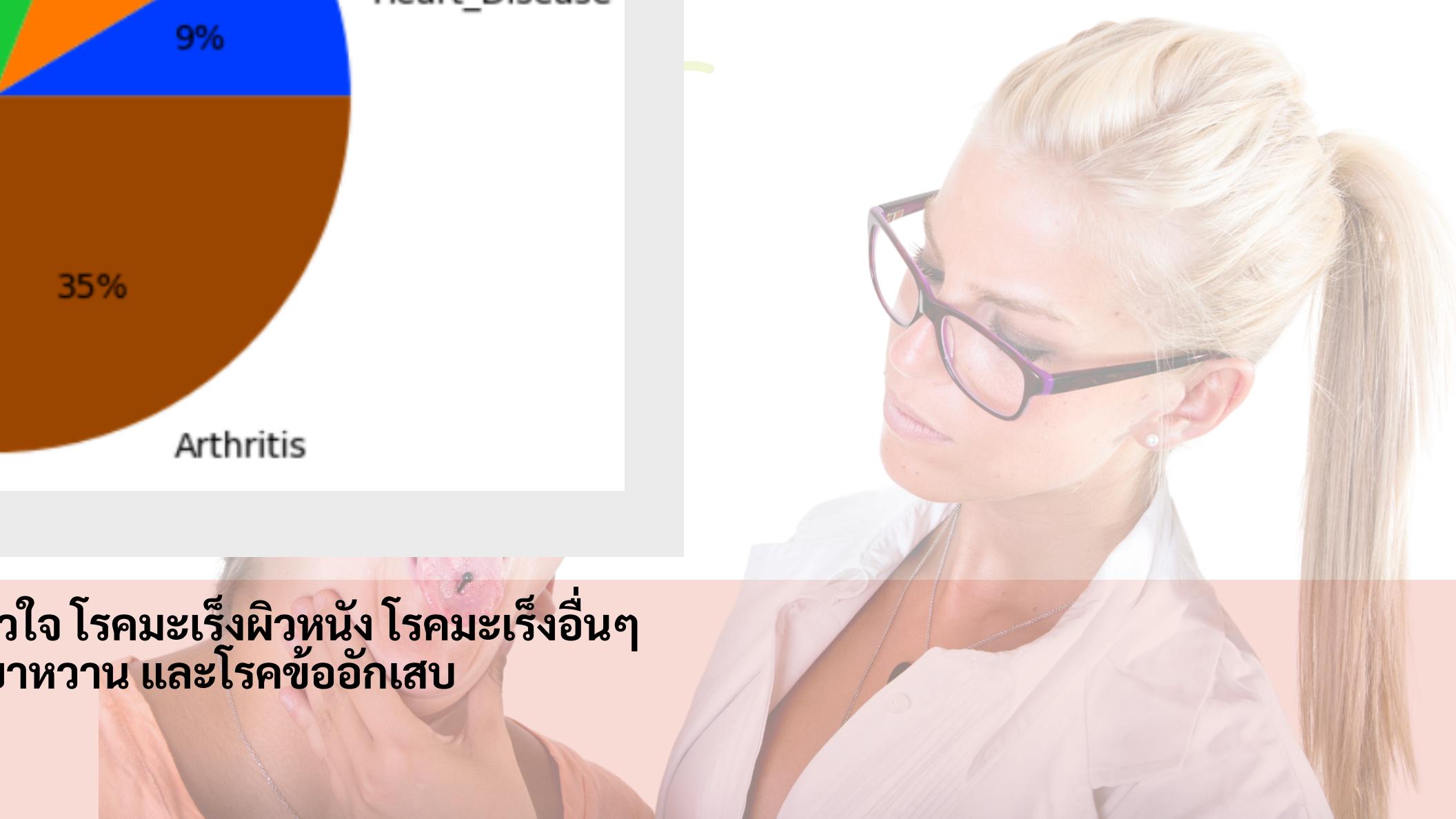


Pie chart ของคนที่เป็นโรคซึมเศร้า โรคเบาหวาน โรคข้ออักเสบ

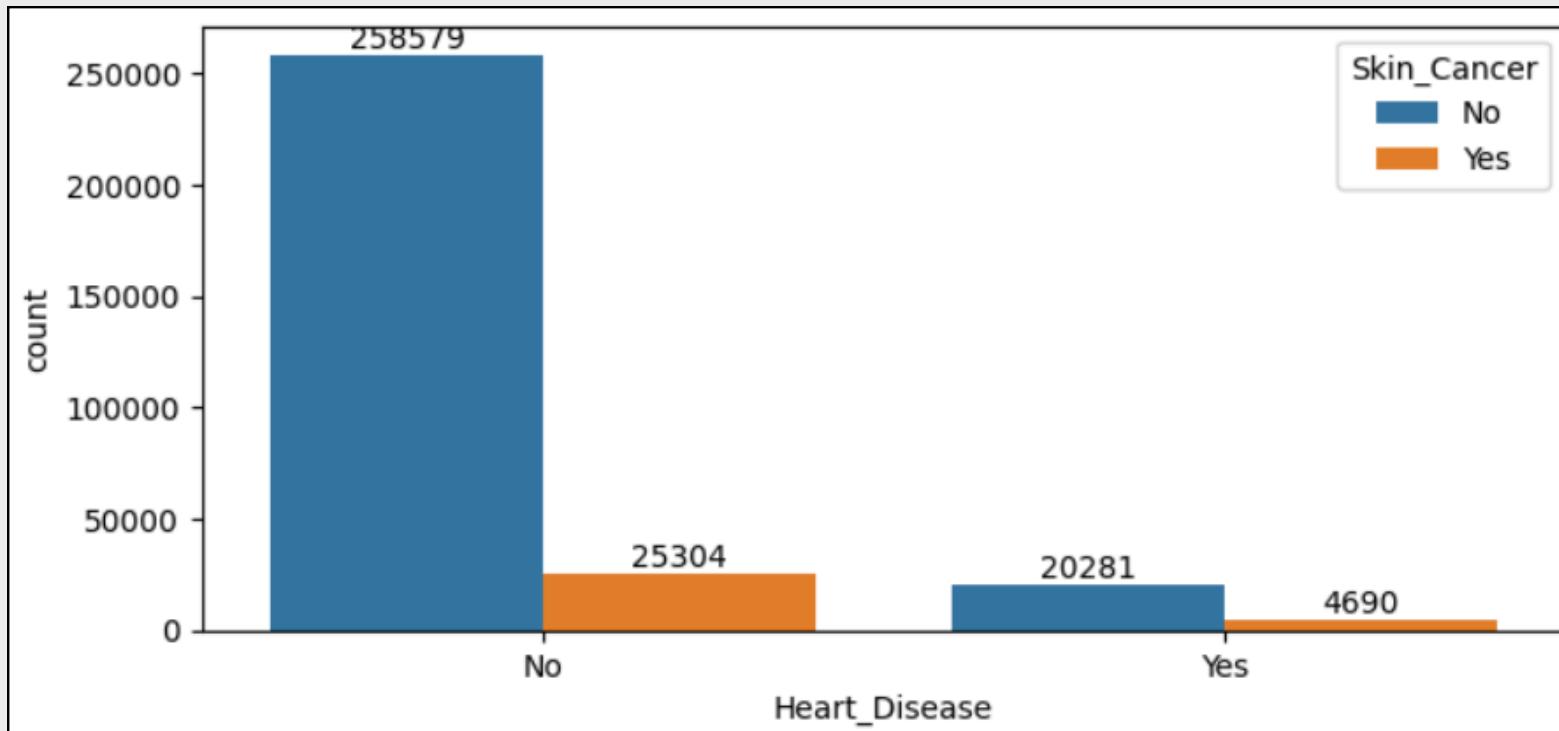
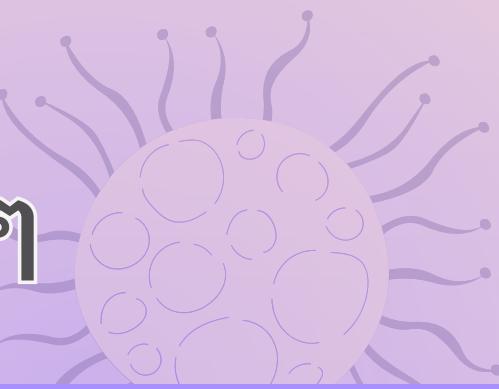
สัดส่วนของคนที่เป็นโรคหัวใจ โรคมะเร็งผิวหนัง โรคมะเร็งอื่นๆ โรคซึมเศร้า โรคเบาหวาน และโรคข้ออักเสบ



Pie chart ของคนที่เป็นโรคหัวใจ โรคมะเร็งผิวหนัง โรคมะเร็งอื่นๆ
โรคซึมเศร้า โรคเบาหวาน และโรคข้ออักเสบ

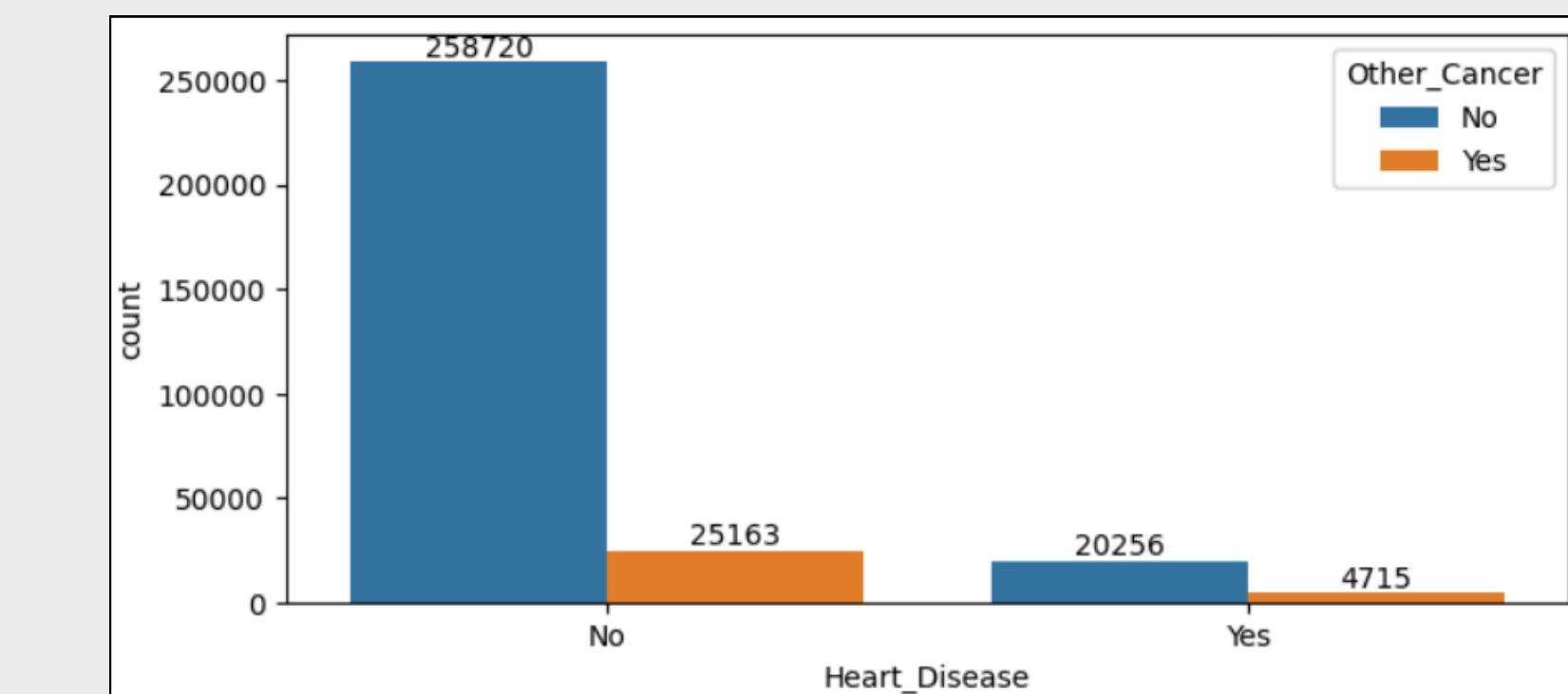


ความสัมพันธ์ของความถี่คนที่เป็นโรคหัวใจ กับโรคมะเร็งผิวหนัง และโรคมะเร็งอื่นๆ



กราฟแท่งแสดงความสัมพันธ์ของความถี่คนที่เป็นโรคหัวใจ กับโรคมะเร็งผิวหนัง

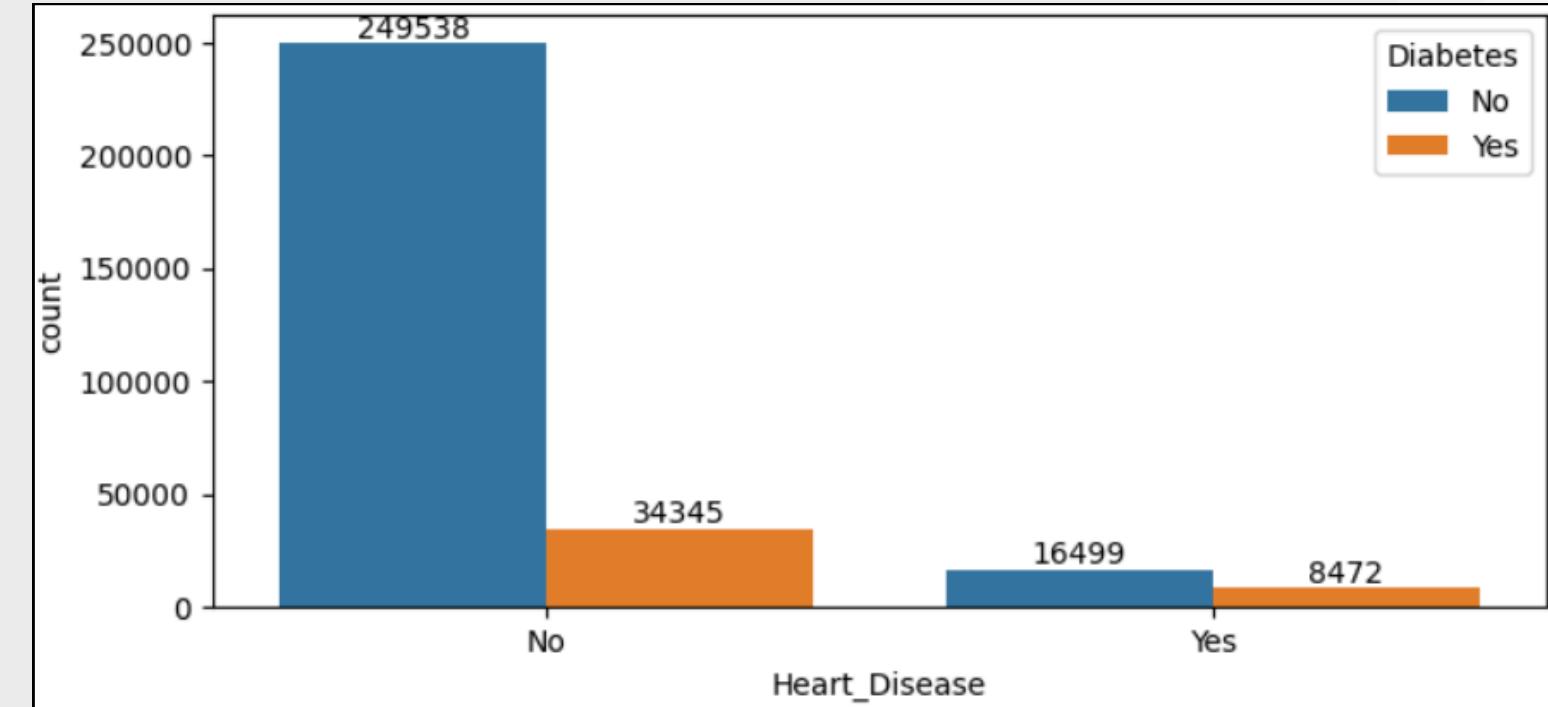
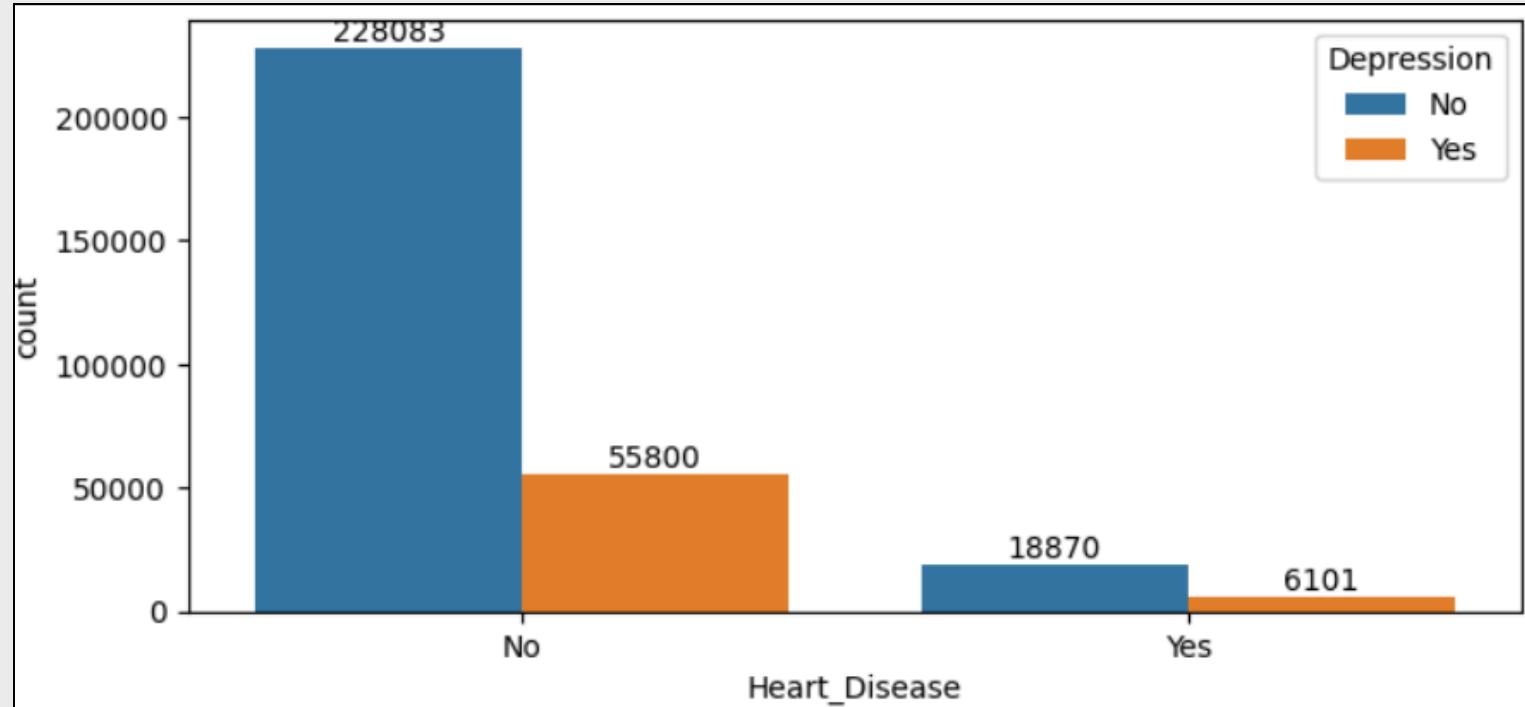
สรุปได้ว่าคนที่ไม่เป็นโรคมะเร็งผิวหนังและไม่เป็นโรคหัวใจมีความถี่สูงที่สุด รองลงมาคือเป็นโรคมะเร็งผิวหนังแต่ไม่เป็นโรคหัวใจ



กราฟแท่งแสดงความสัมพันธ์ของความถี่คนที่เป็นโรคหัวใจ กับโรคมะเร็งอื่นๆ

สรุปได้ว่าคนที่ไม่เป็นโรคมะเร็งอื่นๆและไม่เป็นโรคหัวใจมีความถี่สูงที่สุด รองลงมาคือเป็นโรคมะเร็งอื่นๆแต่ไม่เป็นโรคหัวใจ

ความสัมพันธ์ของความถี่คนที่เป็นโรคหัวใจ โรคซึมเศร้า และโรคเบาหวาน



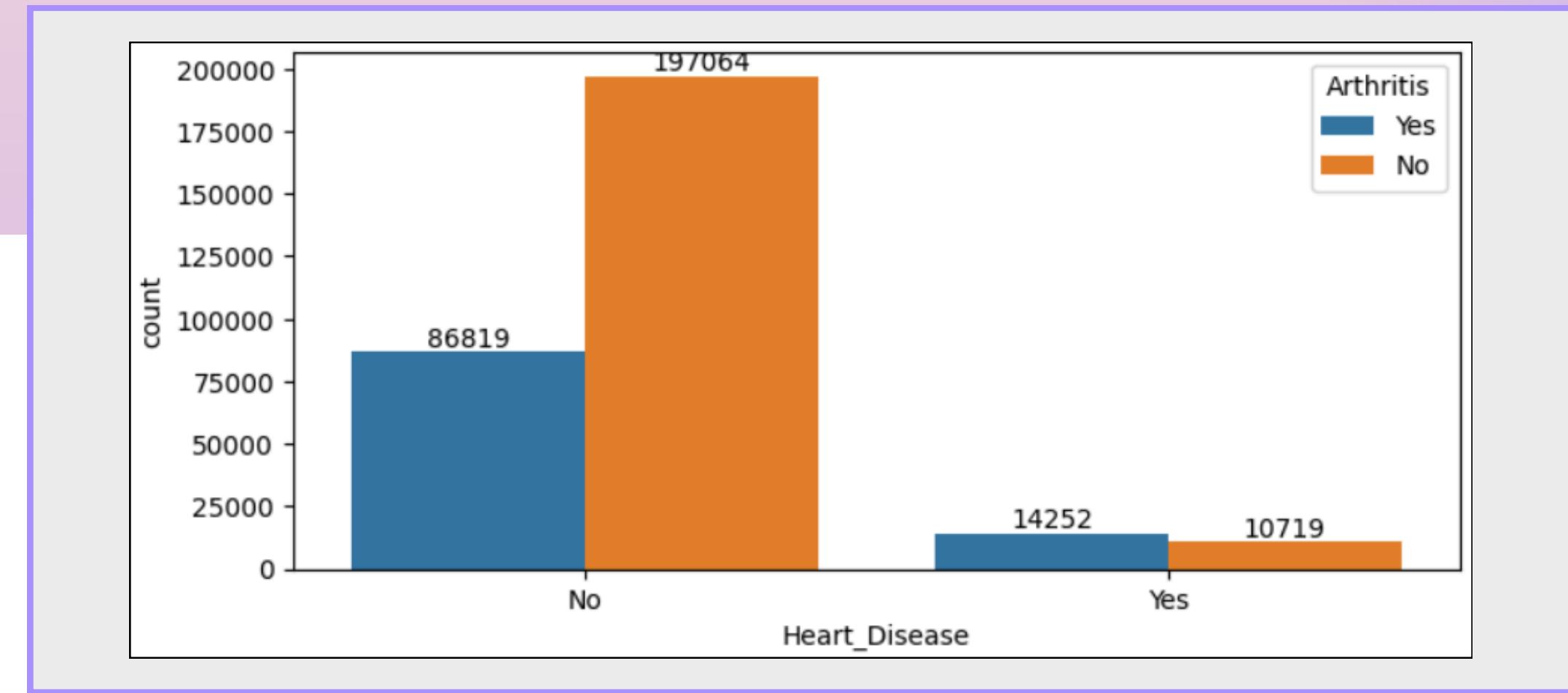
กราฟแท่งแสดงความสัมพันธ์ของความถี่คนที่เป็นโรคหัวใจ กับโรคซึมเศร้า

สรุปได้ว่าคนที่ไม่เป็นโรคซึมเศร้าและไม่เป็นโรคหัวใจมีความถี่สูงที่สุด รองลงมาคือเป็นโรคซึมเศร้าแต่ไม่เป็นโรคหัวใจ

กราฟแท่งแสดงความสัมพันธ์ของความถี่คนที่เป็นโรคหัวใจ กับโรคเบาหวาน

สรุปได้ว่าคนที่ไม่เป็นโรคเบาหวานและไม่เป็นโรคหัวใจมีความถี่สูงที่สุด รองลงมาคือเป็นโรคเบาหวานแต่ไม่เป็นโรคหัวใจ

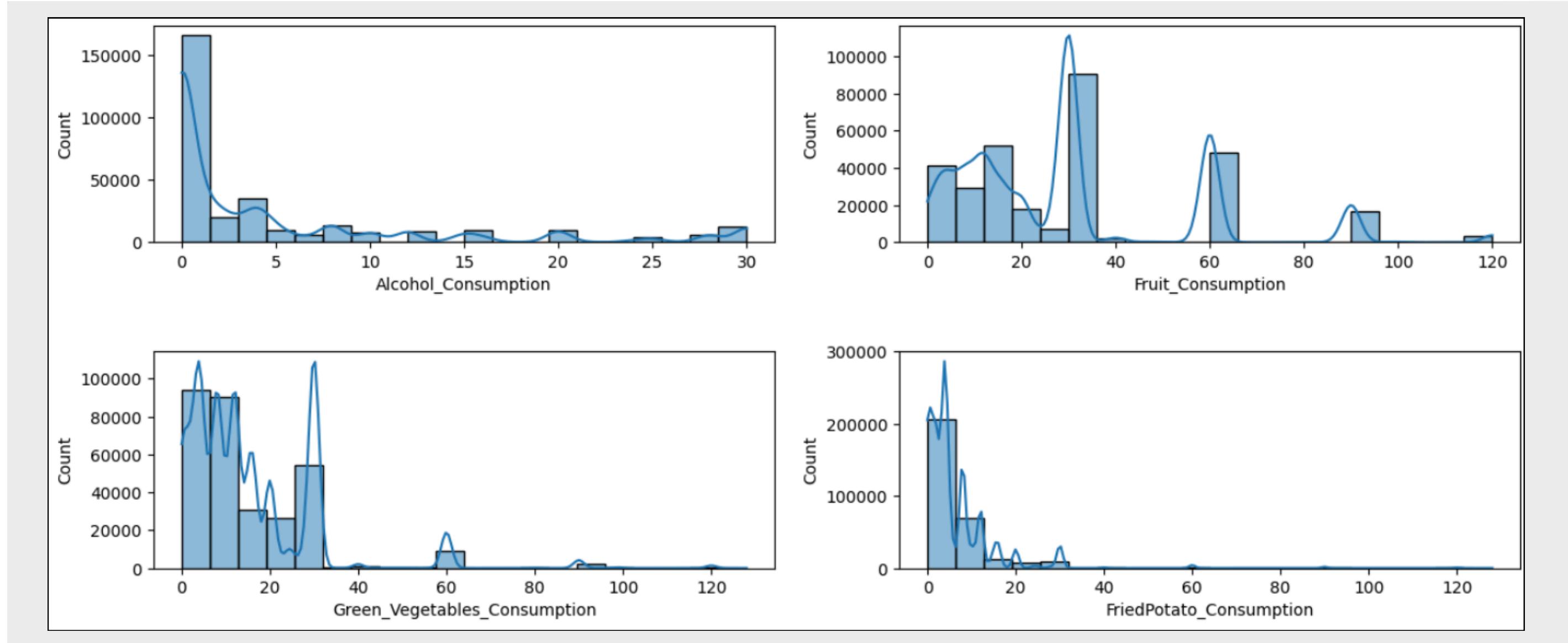
ความสัมพันธ์ของความถี่คนที่เป็นโรคหัวใจกับโรคข้อเสื่อม



กราฟแท่งแสดงความสัมพันธ์ของความถี่คนที่เป็นโรคหัวใจกับโรคข้อเสื่อม

สรุปได้ว่าคนที่ไม่เป็นโรคข้อเสื่อมและไม่เป็นโรคหัวใจมีความถี่สูงที่สุด รองลงมาคือเป็นโรคข้อเสื่อมแต่ไม่เป็นโรคหัวใจ

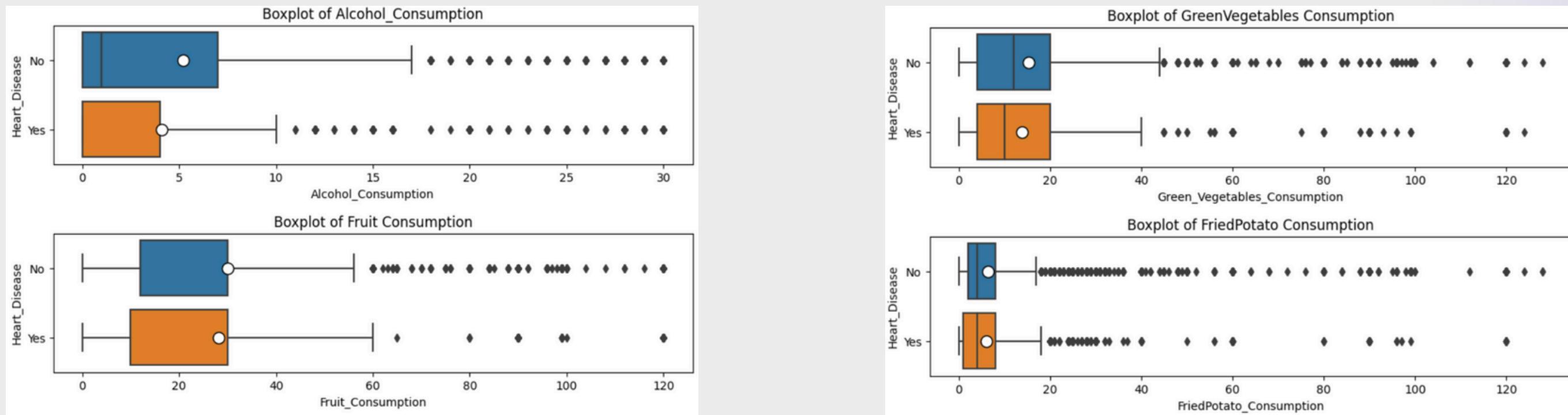
การแจกแจงของปริมาณการดื่มแอลกอฮอล์ การรับประทานผลไม้ การรับประทานผัก ใบเขียว และการรับประทานมันฝรั่งทอด (French fried) ของคนที่ตอบแบบทั้งหมด



กราฟ Histogram ของปริมาณการดื่มแอลกอฮอล์ การรับประทานผลไม้ ผักใบเขียวและมันฝรั่งทอด

สรุปได้ว่าปริมาณการดื่มแอลกอฮอล์ มีการแจกแจงแนวโน้มไปทางเบื้องขวา ปริมาณการรับประทานผลไม้ดูมีแนวโน้มการแจกแจงที่ปกติ ปริมาณการรับประทานผักใบเขียวมีการแจกแจงค่อนไปทางเบื้องขวา และปริมาณการรับประทานมันฝรั่งทอดมีการแจกแจงที่เบื้องขวาเช่นกัน

Boxplot ของปริมาณการดื่มแอลกอฮอล์ การรับประทานผลไม้ การรับประทานผักใบเขียว และการรับประทานมันฝรั่งทอด (French fried) ของคนที่ตอบแบบสอบถามทั้งหมด



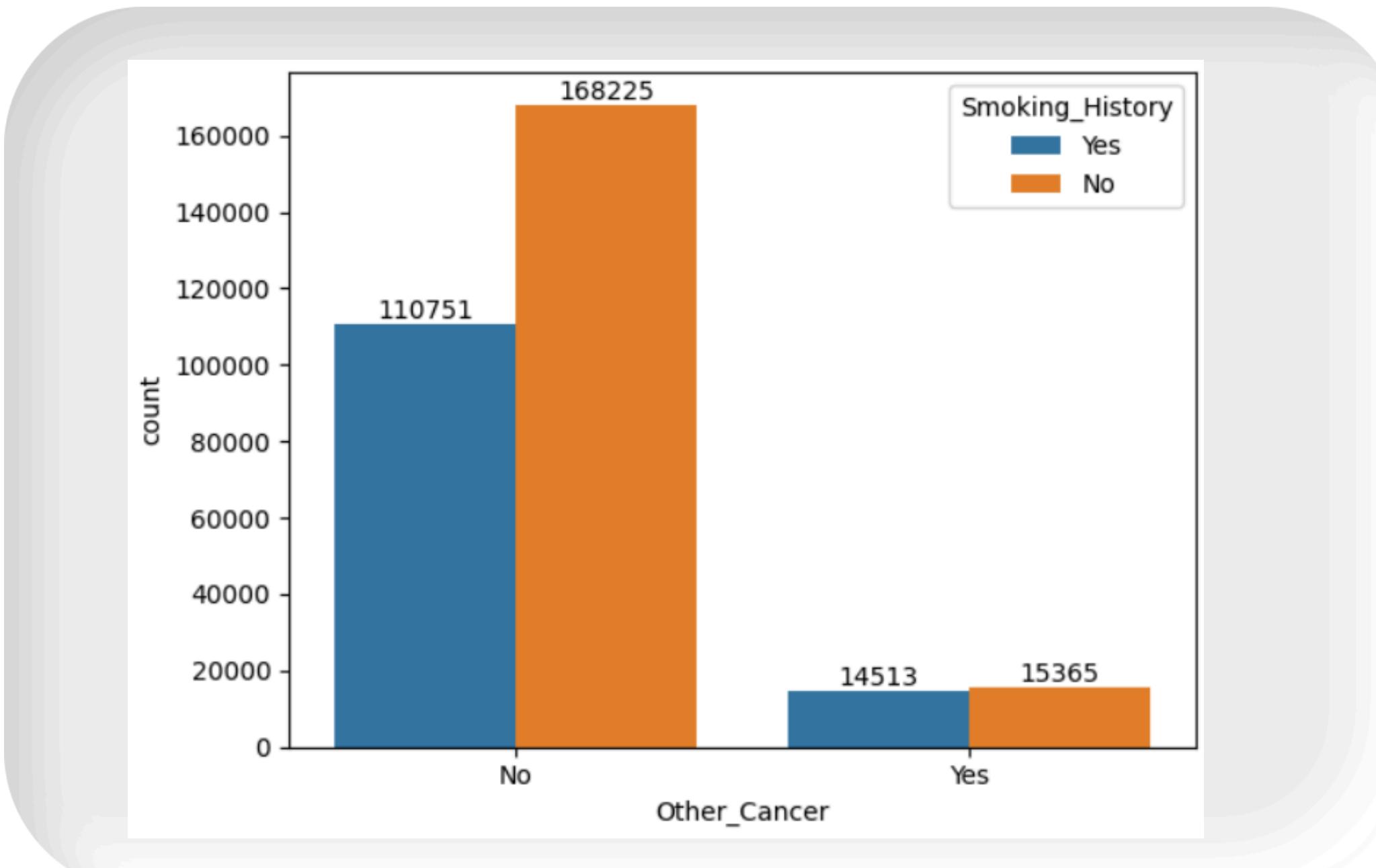
Boxplot ของปริมาณการดื่มแอลกอฮอล์ การรับประทานผลไม้ ผักใบเขียวและมันฝรั่งทอด

Variables	Mean	Median	Mode
Alcohol Consumption (g / month)	5.09	1.00	0.00
Fruit Consumption (g / month)	29.83	30.00	30.00
Green Vegetables Consumption (g / month)	15.11	12.00	30.00
Fried Potato Consumption (g / month)	6.29	4.00	4.00

สรุปได้ว่าปริมาณการดื่มแอลกอฮอล์มีค่าเฉลี่ยประมาณ 5 g/month
ปริมาณการรับประทานผลไม้มีค่าเฉลี่ยประมาณ 30 g/month
ปริมาณการรับประทานผักใบเขียวมีค่าเฉลี่ยประมาณ 15 g/month
ปริมาณการรับประทานมันฝรั่งทอดมีค่าเฉลี่ยประมาณ 6 g/month
และปริมาณการรับประทานมันฝรั่งทอดดูมีค่านอกเกณฑ์ (Outlier)
มากที่สุด

ค่าเฉลี่ย มารยฐาน และฐานนิยม ของปริมาณการดื่มแอลกอฮอล์ การรับประทานผลไม้ การรับประทานผักใบเขียว และการรับประทานมันฝรั่งทอด (French fried) ของผู้ตอบแบบสอบถามทั้งหมด

ความถี่ของการเป็นโรคมะเร็งอื่นๆ กับประวัติการสูบบุหรี่

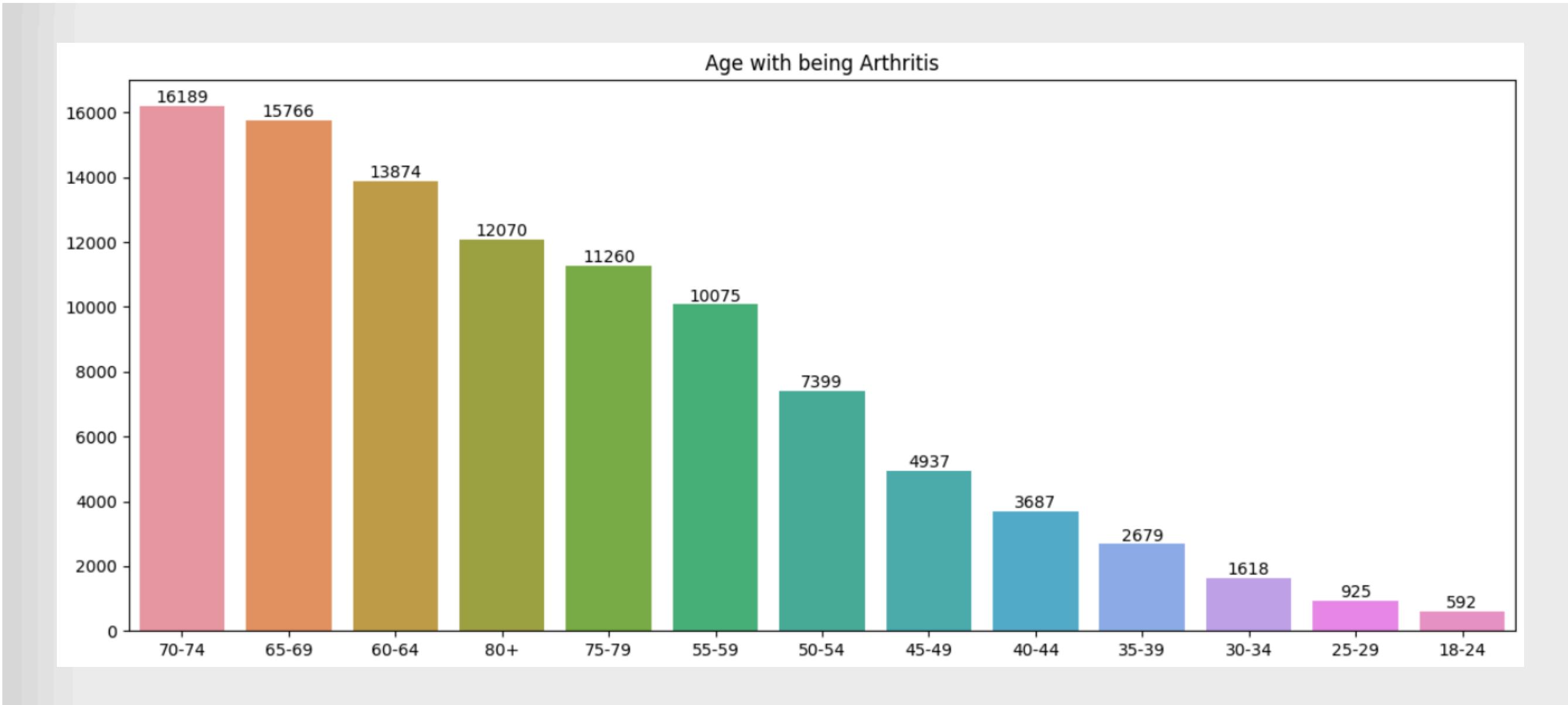


กราฟแท่งแสดงความถี่ของการเป็นโรคมะเร็งอื่นๆ โดยแบ่งตามการสูบบุหรี่

สรุปได้ว่าคนที่ไม่ได้สูบบุหรี่ และไม่ได้เป็นโรคมะเร็งอื่นๆ มีความถี่มากที่สุด รองลงมาคือสูบบุหรี่แต่ไม่ได้เป็นโรคมะเร็งอื่นๆ



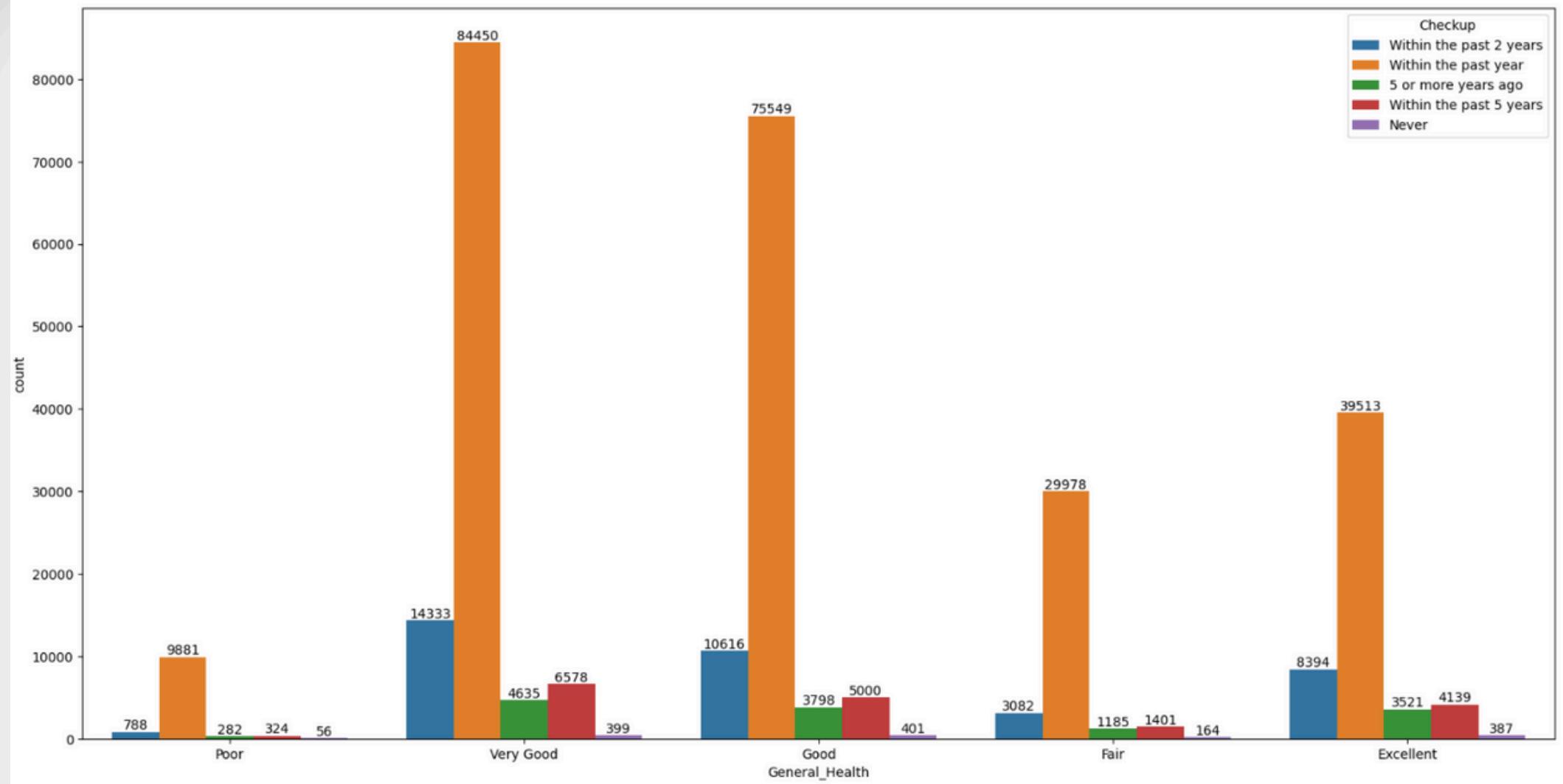
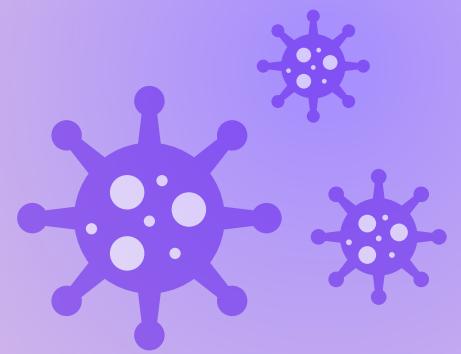
ความถี่ของการเป็นโรคข้อเสื่อมโดยแบ่งตามช่วงอายุ



กราฟแท่งแสดงความถี่ของการเป็นโรคข้อเสื่อมโดยแบ่งตามช่วงอายุ

สรุปได้ว่าช่วงอายุ 70-74 ปี มีความถี่ของการเป็นโรคข้อเสื่อมมากที่สุด
รองลงมา คือ ช่วงอายุ 65-69 ปี

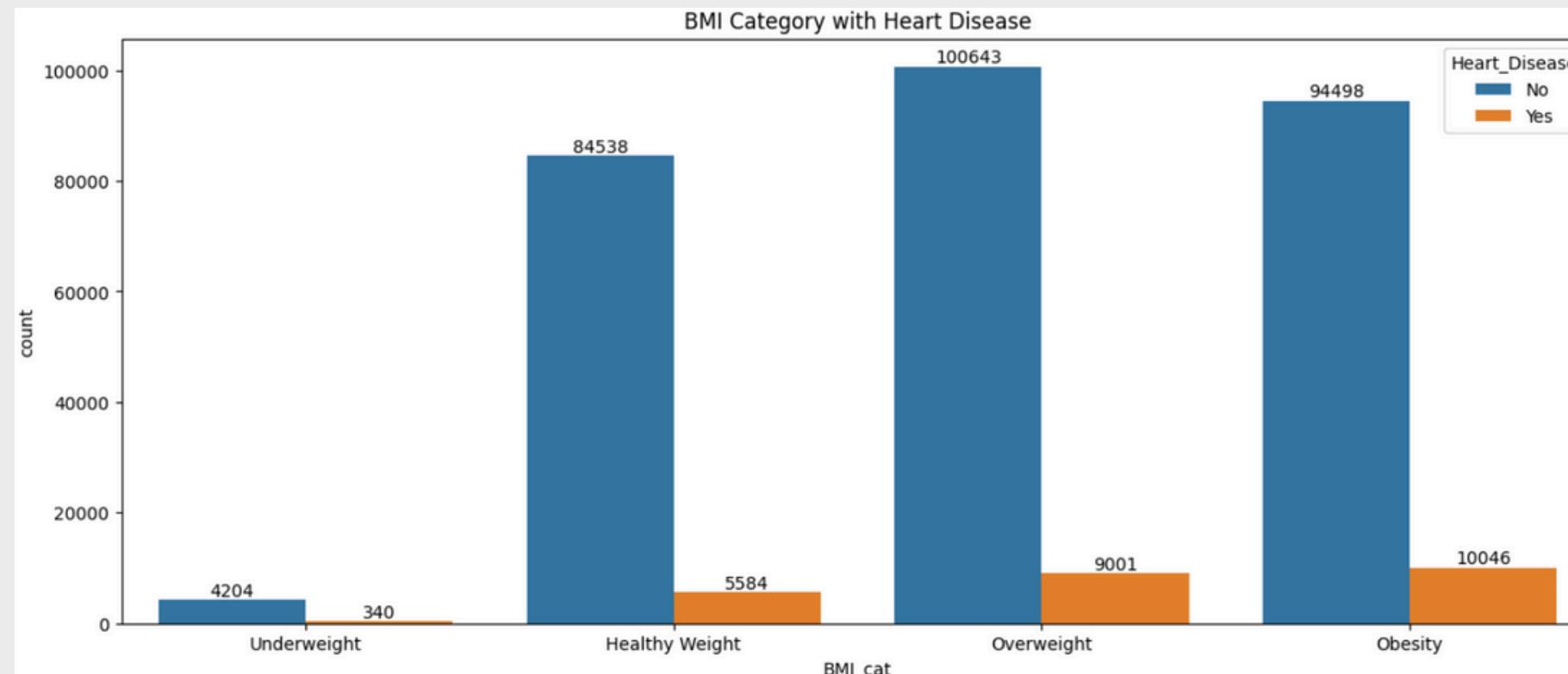
ความสัมพันธ์ระหว่างครั้งล่าสุดที่ผู้ตอบแบบสอบถามไปตรวจสุขภาพ กับระดับสุขภาพโดยทั่วไป



สรุปได้ว่าผู้ที่ไปตรวจสุขภาพเมื่อ 2 ปีที่แล้ว ส่วนใหญ่จะอยู่ในระดับสุขภาพดีมาก ผู้ที่ไปตรวจสุขภาพภายในปีที่แล้ว ส่วนใหญ่จะอยู่ในระดับสุขภาพดีมาก ผู้ที่ไปตรวจสุขภาพ 5 ปีที่แล้วหรือมากกว่านั้น ส่วนใหญ่จะอยู่ในระดับสุขภาพดีมาก ผู้ที่ไปตรวจสุขภาพภายใน 5 ปีที่แล้ว ส่วนใหญ่จะอยู่ในระดับสุขภาพดีมาก และผู้ที่ไม่เคยไปตรวจสุขภาพเลย ส่วนใหญ่จะอยู่ในระดับสุขภาพดี

กราฟแท่งแสดงถึงความถี่ของระดับสุขภาพต่างๆโดยแบ่งตามช่วงเวลาที่ไปตรวจสุขภาพครั้งล่าสุด

ผู้ตอบแบบสอบถามที่เป็นโรคหัวใจส่วนใหญ่อยู่ในระดับ BMI อยู่ในเกณฑ์ไหน?

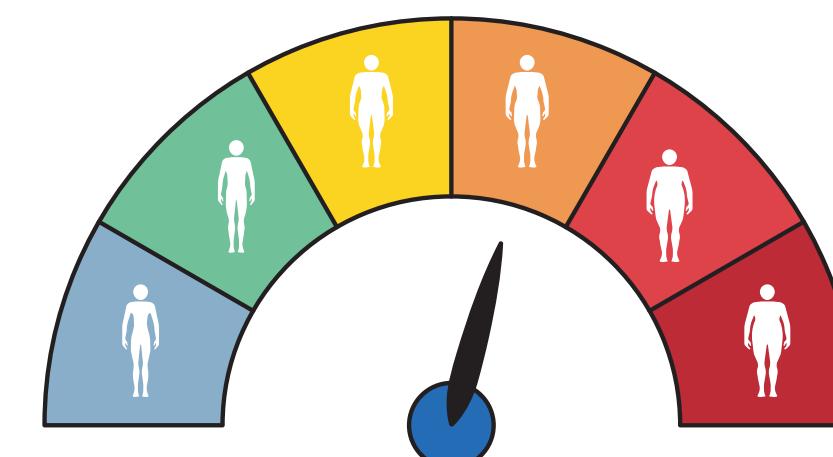


Classification	BMI Target
Underweight	Below 18.5
Healthy	Between 18.5 – 24.9
Overweight	25 – 29.9
Obese	Over 30

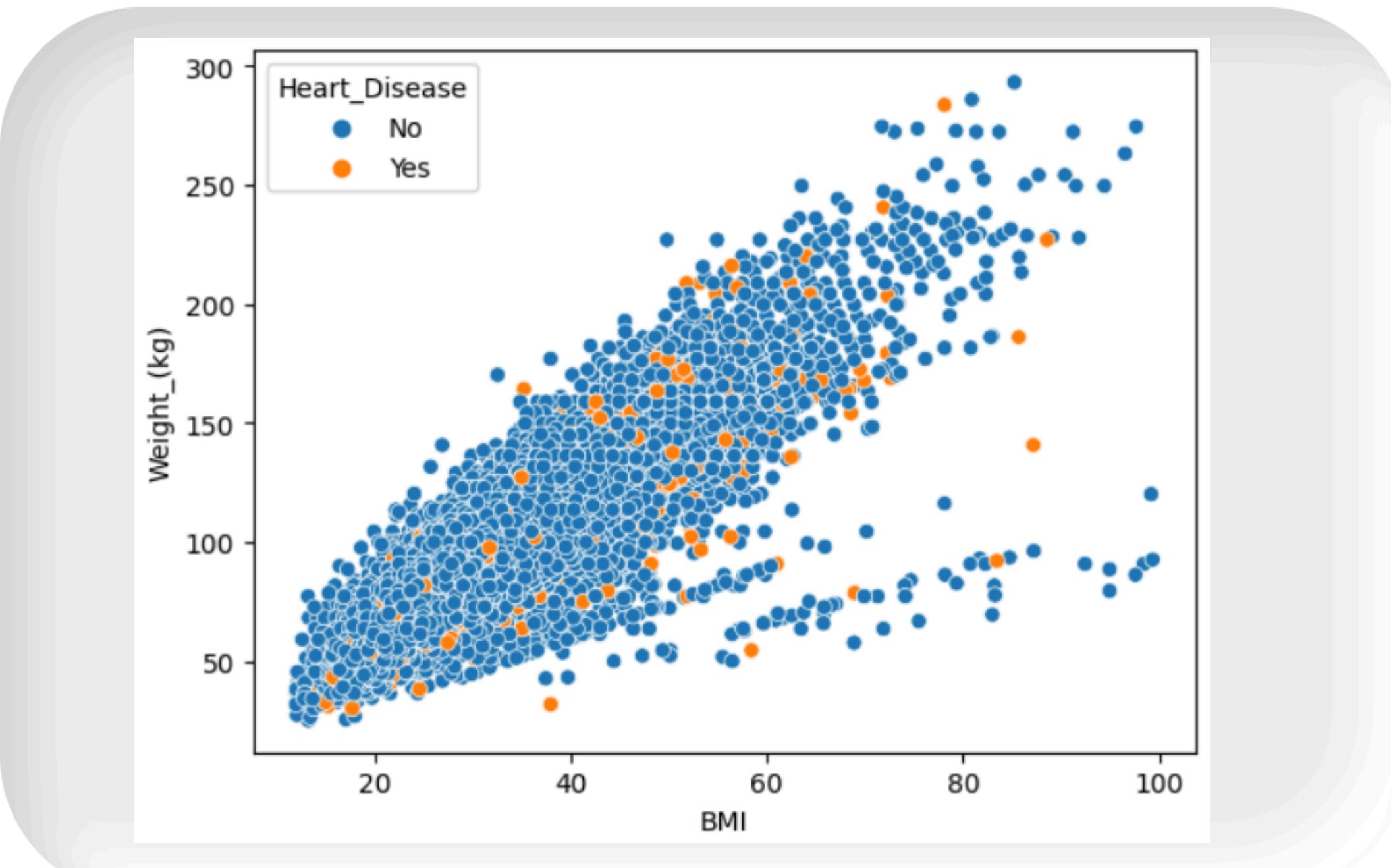
ตารางค่า BMI แต่ละช่วงระดับ
(อ้างอิงจาก : <https://diabetesmyway.nhs.uk>)

กราฟแท่งแสดงถึงความสัมพันธ์ระหว่างช่วงระดับ BMI กับการเป็นโรคหัวใจ

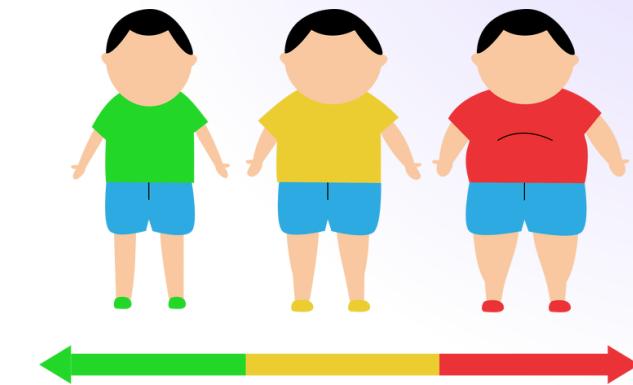
สรุปได้ว่าคนที่เป็นโรคหัวใจส่วนใหญ่จะมีค่า BMI อยู่ในช่วงของการเป็นโรคอ้วน (Obesity) มากที่สุด ส่วนคนที่ไม่ได้เป็นโรคหัวใจส่วนใหญ่จะมีค่า BMI ในช่วงน้ำหนักเกิน (Overweight) มากที่สุด



ค่า BMI และน้ำหนักของผู้ป่วยเป็นโรคและไม่เป็นโรคหัวใจมีความสัมพันธ์กันอย่างไร

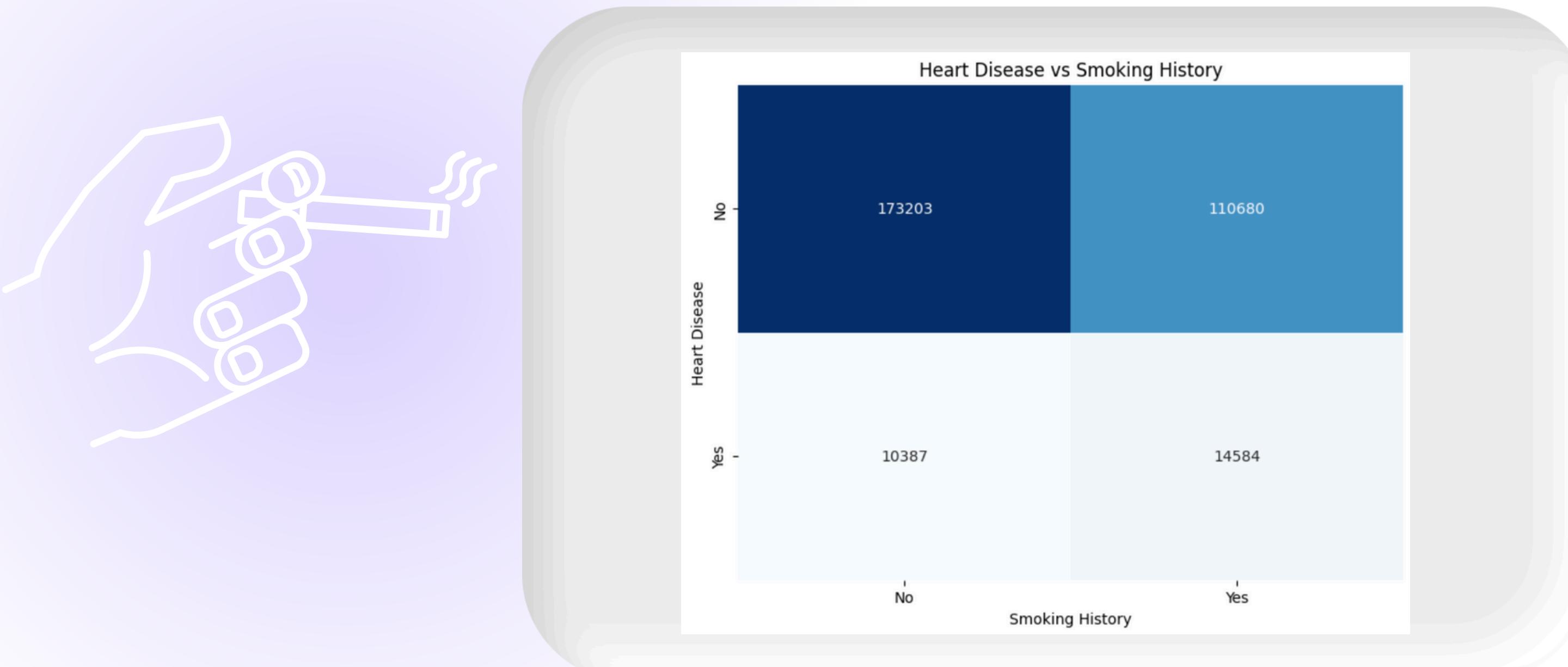


กราฟ Scatter plot แสดงความสัมพันธ์ระหว่างค่า BMI และน้ำหนัก



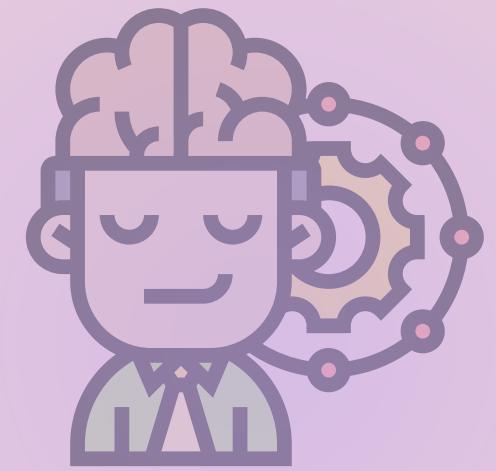
สรุปได้ว่าความสัมพันธ์ระหว่างค่า BMI และน้ำหนักมีความสัมพันธ์เชิงเส้นกันในทางบวก (เมื่อค่า BMI เพิ่ม น้ำหนักก็มีแนวโน้มที่จะเพิ่มขึ้นด้วย)

ความสัมพันธ์ของผู้ป่วยที่เป็นโรคหัวใจและประวัติการสูบบุหรี่



แผนภาพ Heatmap แสดงถึงความสัมพันธ์ของผู้ป่วยที่เป็นโรคหัวใจและประวัติการสูบบุหรี่

สรุปได้ว่าผู้ที่ไม่ได้สูบบุหรี่และไม่ได้เป็นโรคหัวใจมีความถี่มากที่สุด รองลงมาคือ ผู้ที่สูบบุหรี่และไม่ได้เป็นโรคหัวใจ



Model the data



เตรียมข้อมูล (Data Preprocessing)

Ordinal Encoder

	General_Health	Checkup
0	Poor	Within the past 2 years
1	Very Good	Within the past year
2	Very Good	Within the past year
3	Poor	Within the past year
4	Good	Within the past year

Encoding

	General_Health	Checkup
0	0.0	1.0
1	3.0	0.0
2	3.0	0.0
3	0.0	0.0
4	2.0	0.0

ตัวอย่างโปรแกรม

```
# ดูแล้วที่ต้องใช้ Ordinal Encoder ได้แก่ General_Health, Checkup
from sklearn.preprocessing import OrdinalEncoder
gh_ranks = ['Poor','Fair','Good','Very Good','Excellent']
od_encoder = OrdinalEncoder(categories = [gh_ranks])
df['General_Health'] = od_encoder.fit_transform(df[['General_Health']])
checkup_ranks = ['Within the past year','Within the past 2 years',
                 'Within the past 5 years','5 or more years ago','Never']
od_encoder2 = OrdinalEncoder(categories = [checkup_ranks])
df['Checkup'] = od_encoder2.fit_transform(df[['Checkup']])
```

เตรียมข้อมูล (Data Preprocessing)

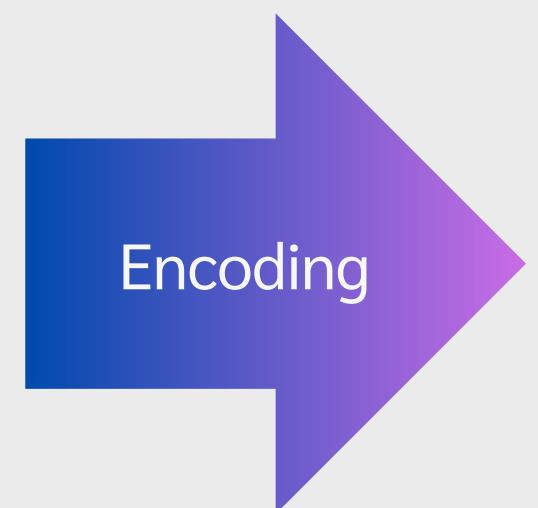
One-Hot Encoder

Encoding

เตรียมข้อมูล (Data Preprocessing)

Mapping

	Age_Category	Diabetes
0	70-74	No
1	70-74	Yes
2	60-64	Yes
3	75-79	Yes
4	80+	No



	Age_Category	Diabetes
0	72	0
1	72	1
2	62	1
3	77	1
4	80	0

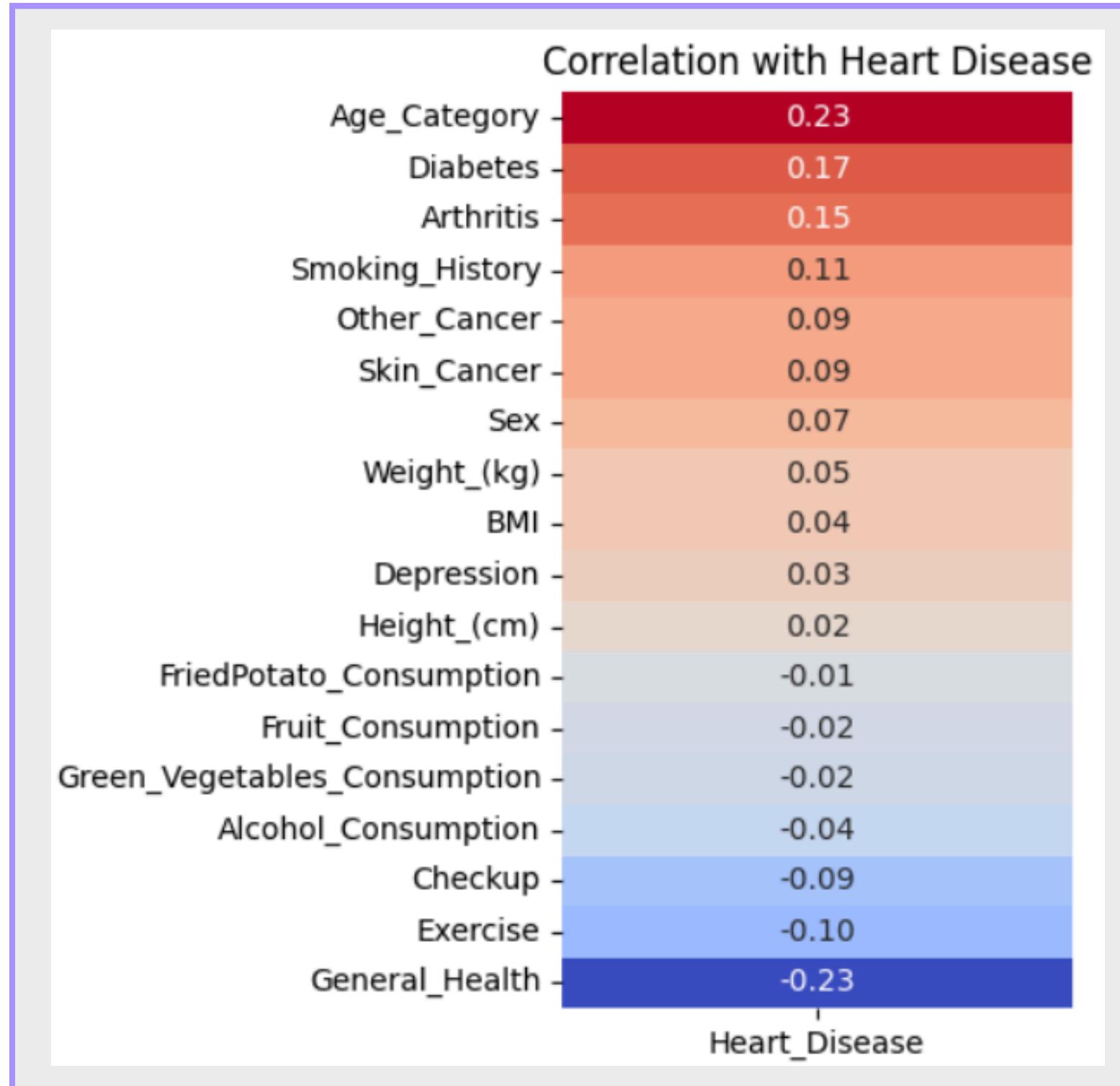
ตัวอย่างโปรแกรม

```
#Age_Category mapping
age_mapping = {'18-24' : 22, '25-29' : 27, '30-34' : 32,
                '35-39' : 37, '40-44' : 42, '45-49' : 47,
                '50-54' : 52, '55-59' : 57, '60-64' : 62,
                '65-69' : 67, '70-74' : 72, '75-79' : 77,
                '80+' : 80,
}
df['Age_Category'] = df['Age_Category'].map(age_mapping)

diabetes_mapping = {'Yes' : 1, 'No' : 0,
                    'No, pre-diabetes or borderline diabetes' : 0,
                    'Yes, but female told only during pregnancy' : 1
}
df['Diabetes'] = df['Diabetes'].map(diabetes_mapping)
```

แผนภาพค่าสัมประสิทธิ์สหสัมพันธ์ ของตัวแปรต่างๆต่อโรคหัวใจ

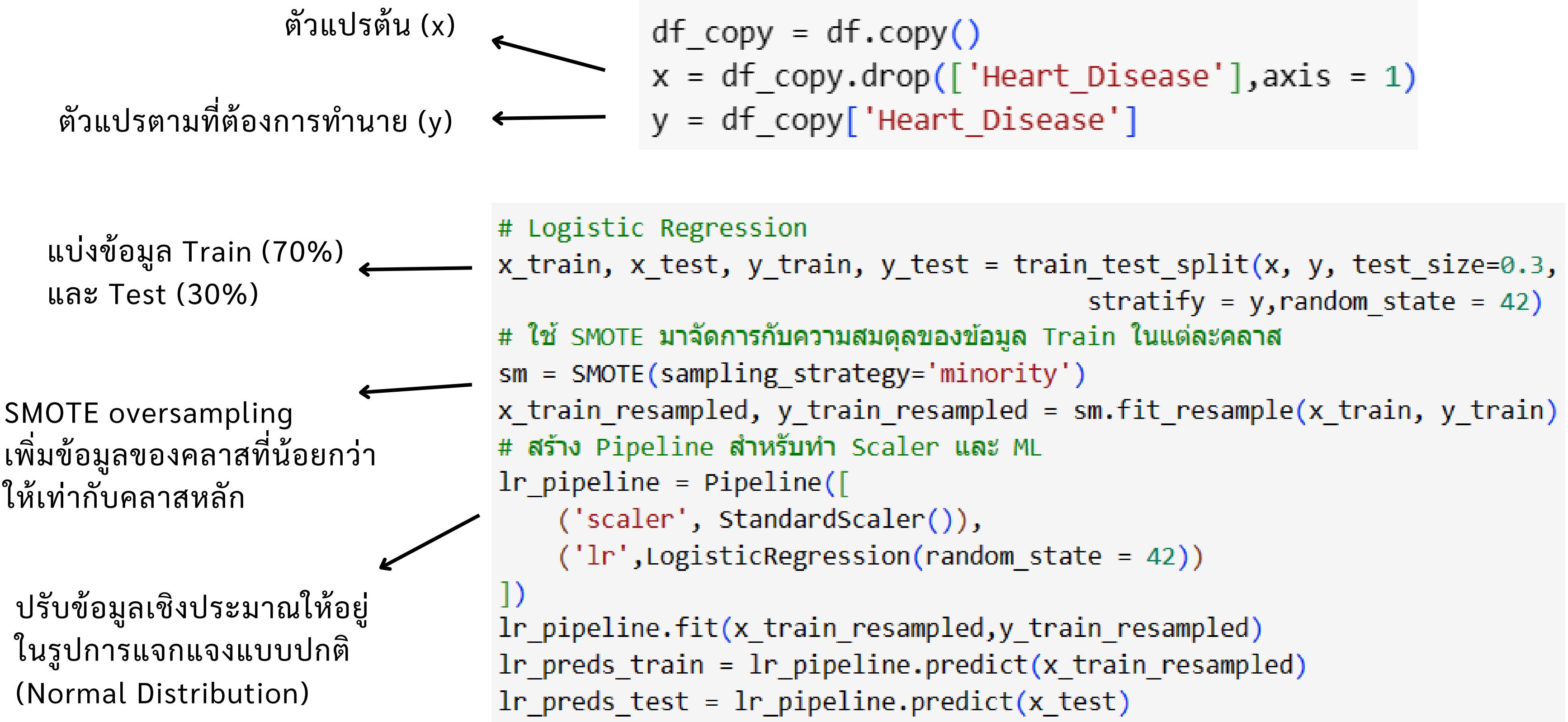
Correlation coefficient with Heart Disease



พบว่าตัวแปรที่มีความสัมพันธ์กับโรคหัวใจมากที่สุดคือ อายุ โดยมีค่าสหสัมพันธ์เท่ากับ 0.23 ซึ่งแสดงว่าเมื่ออายุมากขึ้น โอกาสที่จะเป็นโรคหัวใจก็จะเพิ่มขึ้นตามไปด้วย รองลงมาคือ การเป็นโรคเบาหวาน โดยมีค่าสหสัมพันธ์เท่ากับ 0.17 ซึ่ง แสดงว่าผู้ที่เป็นโรคเบาหวานมีความเสี่ยงที่จะเป็นโรคหัวใจมากกว่าผู้ที่ไม่ได้เป็นโรคเบาหวาน

ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient)

Logistic Regression



Logistic Regression

ผลลัพธ์ Classification Report ของชุดข้อมูลทดสอบ

===== Classification Report (Test data) =====

	precision	recall	f1-score	support
0.0	0.97	0.74	0.84	85166
1.0	0.20	0.75	0.32	7491
accuracy			0.74	92657
macro avg	0.59	0.75	0.58	92657
weighted avg	0.91	0.74	0.80	92657

Model	Accuracy	Precision	Sensitivity (Recall)	F1-score
Class 0 (ไม่เป็นโรคหัวใจ)	0.74	0.97	0.74	0.84
Class 1 (เป็นโรคหัวใจ)		0.20	0.75	0.32

- Accuracy: โมเดลมีประสิทธิภาพในการ Predict โดยรวมค่อนข้างดี
- Precision: โมเดลมีความแม่นยำในการ Predict ในกรณีที่คนไม่เป็นโรคหัวใจได้ค่อนข้างดี แต่มีความแม่นยำในการ Predict ในกรณีที่คนเป็นโรคหัวใจได้ไม่ดี
- Recall: โมเดลมีความสามารถในการตรวจจับคนที่เป็นและไม่เป็นโรคหัวใจจริงๆ ได้ดี
- F1-Score: โมเดลมีความสมดุลระหว่าง Precision และ Recall ใน Class 0 แต่มีความไม่สมดุลใน Class 1

สูตรการคำนวณ Accuracy, Precision, Recall และ F1-Score

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

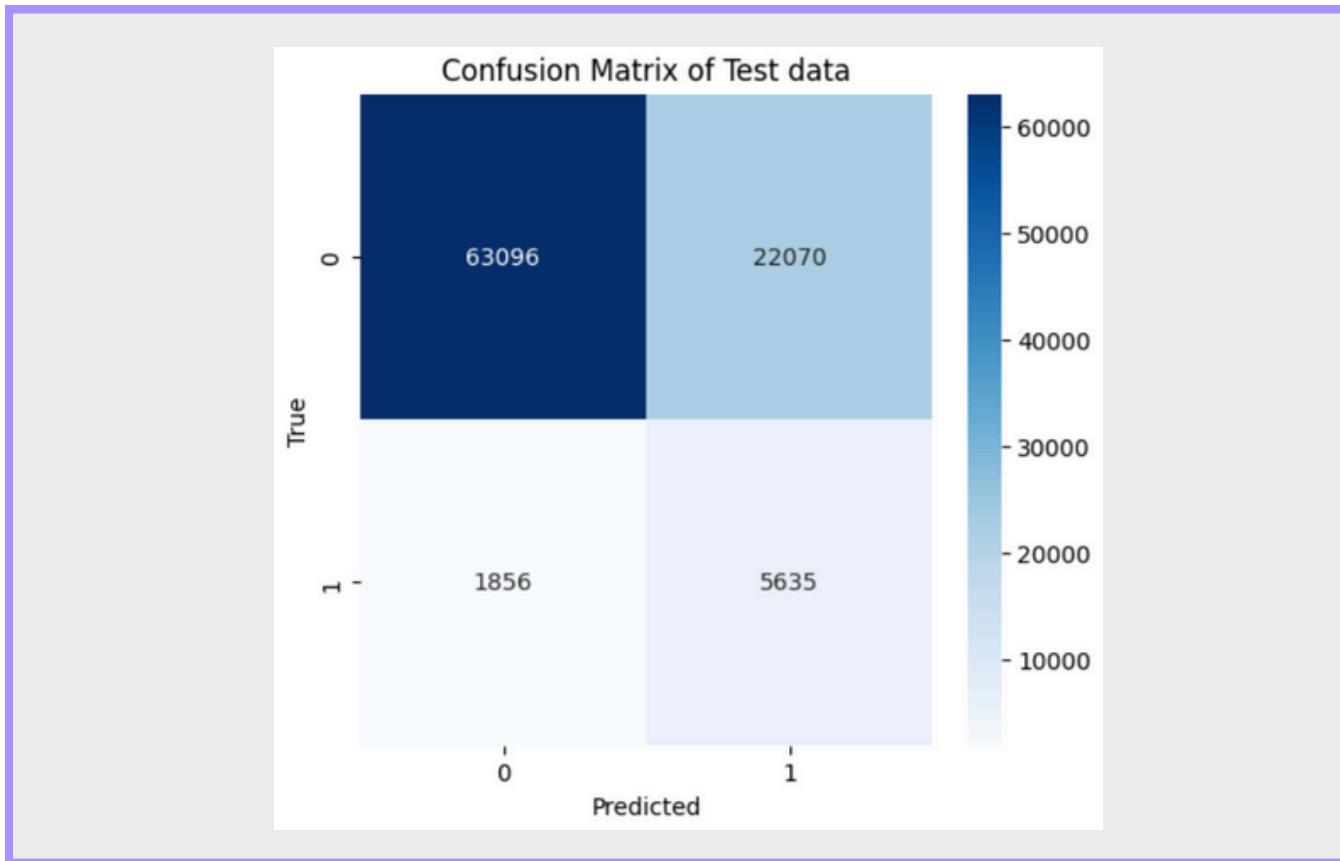
		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

ด้วยปัจจุบัน Confusion Matrix ขนาด 2x2

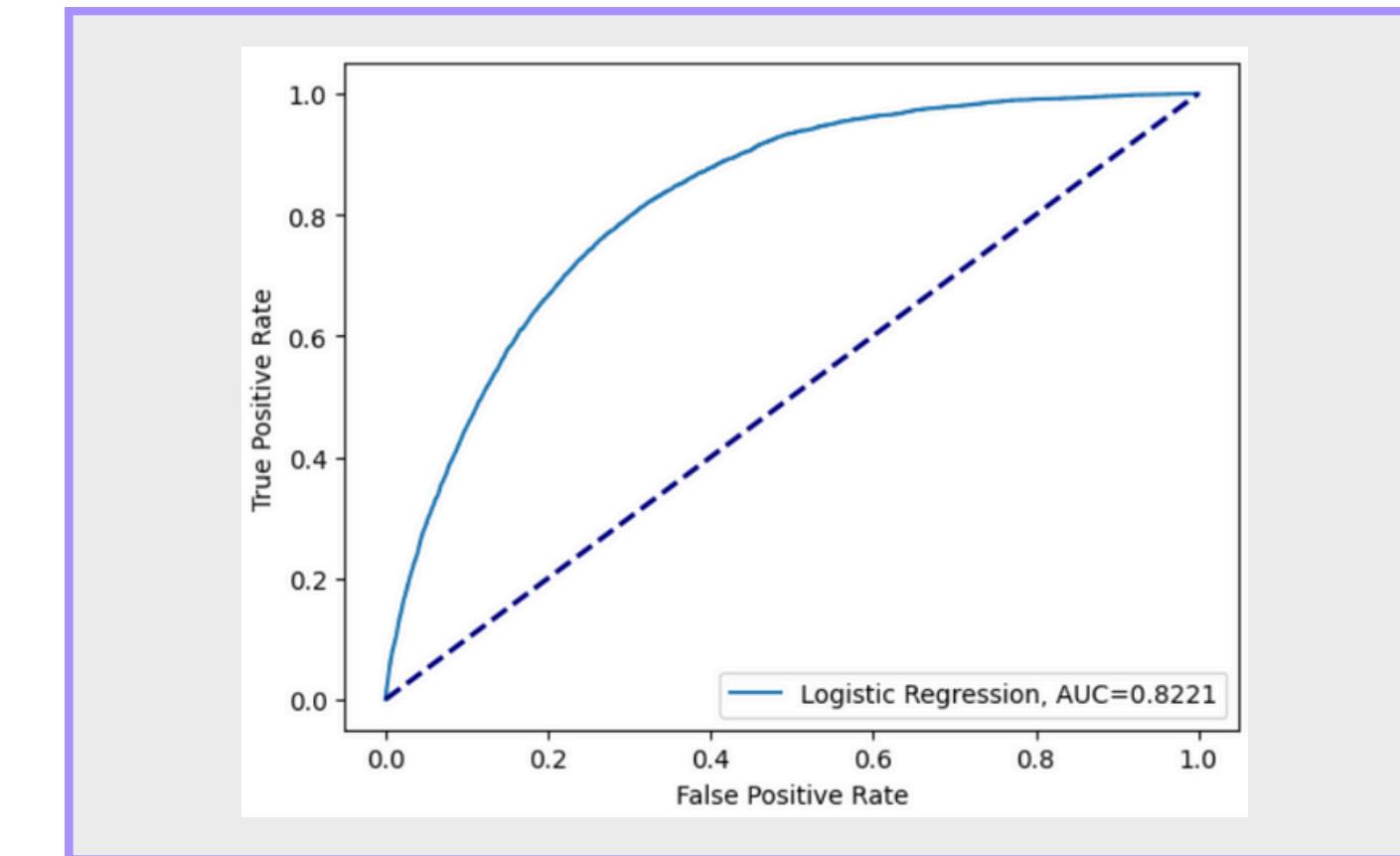
ที่มา : <https://www.fticonsulting.com/de-de/france/insights/articles/machine-learning-model-metrics-trust-them>

Logistic Regression

ผลลัพธ์ของ Confusion Matrix และ AUC-ROC Curve



ผลลัพธ์ Confusion Matrix ของชุดข้อมูลทดสอบ

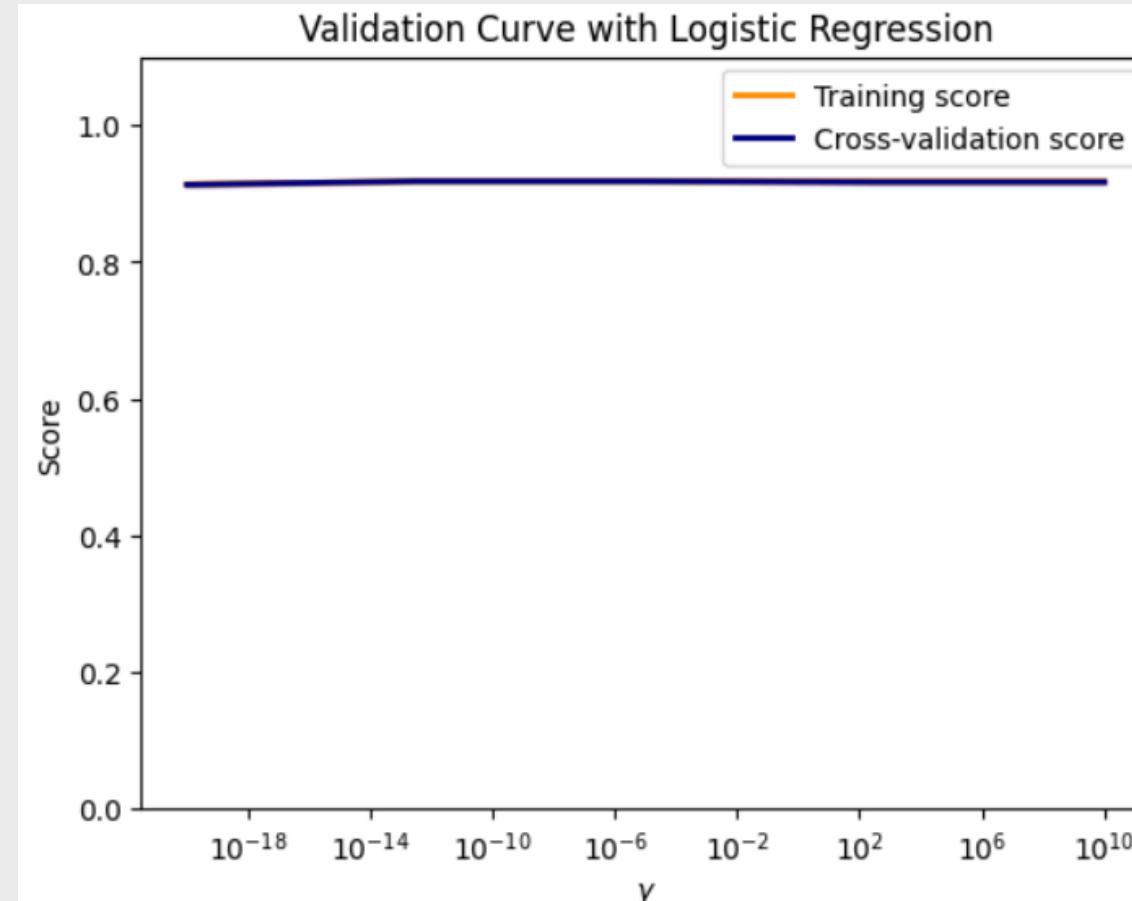


ผลลัพธ์ของ AUC-ROC Curve

True Negative (TN) : โมเดลทำนายว่าไม่เป็นโรคหัวใจซึ่งผลลัพธ์จริงคือไม่เป็นโรคหัวใจ โดยทำนายได้ 63096 คน
False Negative (FN) : โมเดลทำนายว่าไม่เป็นโรคหัวใจซึ่งผลลัพธ์จริงคือเป็นโรคหัวใจ โดยทำนายได้ 1856 คน
False Positive (FP) : โมเดลทำนายว่าเป็นโรคหัวใจซึ่งผลลัพธ์จริงคือไม่เป็นโรคหัวใจ โดยทำนายได้ 22070 คน
True Positive (TP) : โมเดลทำนายว่าเป็นโรคหัวใจซึ่งผลลัพธ์จริงคือเป็นโรคหัวใจ โดยทำนายได้ 5635 คน
เส้น ROC นั้นวัดสัดส่วนของ Sensitivity (Recall) และ (1-Specificity) โดยที่ค่า Specificity หาได้จาก $\frac{TN}{(TN + FP)}$ ซึ่งจากการมีค่า AUC เท่ากับ 82.21% ซึ่งแสดงว่าโมเดล มีประสิทธิภาพโดยรวมระดับดี

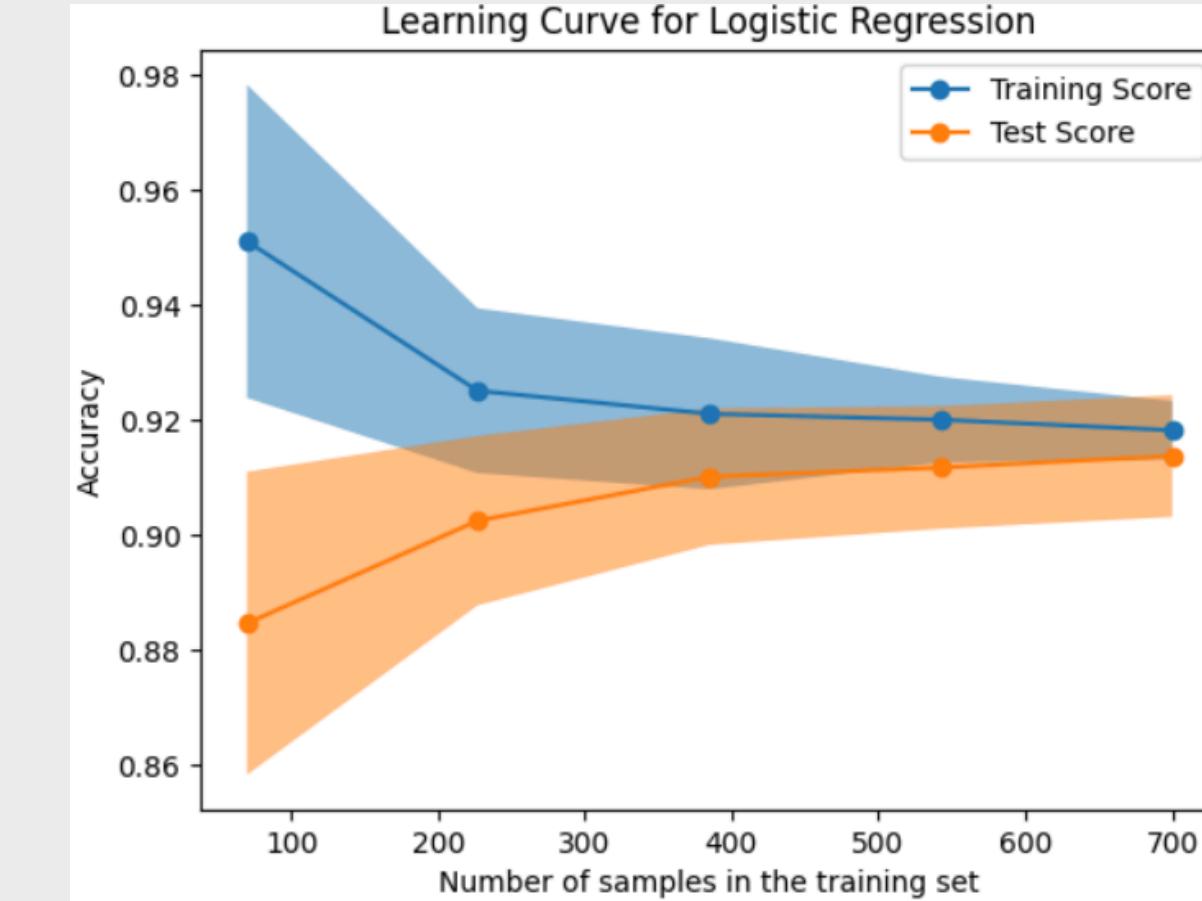
Logistic Regression

Validation Curve และ Learning Curve



ผลลัพธ์ Validation Curve

แกน x คือ ค่า C (inverse of regularization strength)
โดยที่ถ้าค่า C มากจะทำให้โมเดลซับซ้อนน้อย และถ้าค่า C
น้อยจะทำให้โมเดลซับซ้อนมาก และแกน Y คือ Score



ผลลัพธ์ Learning Curve

แกน x คือ จำนวนข้อมูลฝึกสอนที่ใช้ และแกน y คือ Accuracy ซึ่งแสดงให้เห็นว่าช่วงที่ดีที่สุดคือ เส้นกราฟทั้งสองจะลู่เข้าหากัน

Decision Tree

ปรับข้อมูลเชิงประมาณให้อยู่ในรูปการแจกแจงแบบปกติ (Normal Distribution)

นำไป Predict กับข้อมูล Test

```
# Decision Tree
# สร้าง Pipeline สำหรับทำ Scaler และ ML
dt_pipeline = Pipeline([
    ('scaler', StandardScaler()),
    ('dt', DecisionTreeClassifier(random_state = 42))
])
dt_pipeline.fit(x_train_resampled,y_train_resampled)
dt_preds_train = dt_pipeline.predict(x_train_resampled)
dt_preds_test = dt_pipeline.predict(x_test)
```

Decision Tree

ผลลัพธ์ Classification Report ของชุดข้อมูลทดสอบ

===== Classification Report (Test data) =====

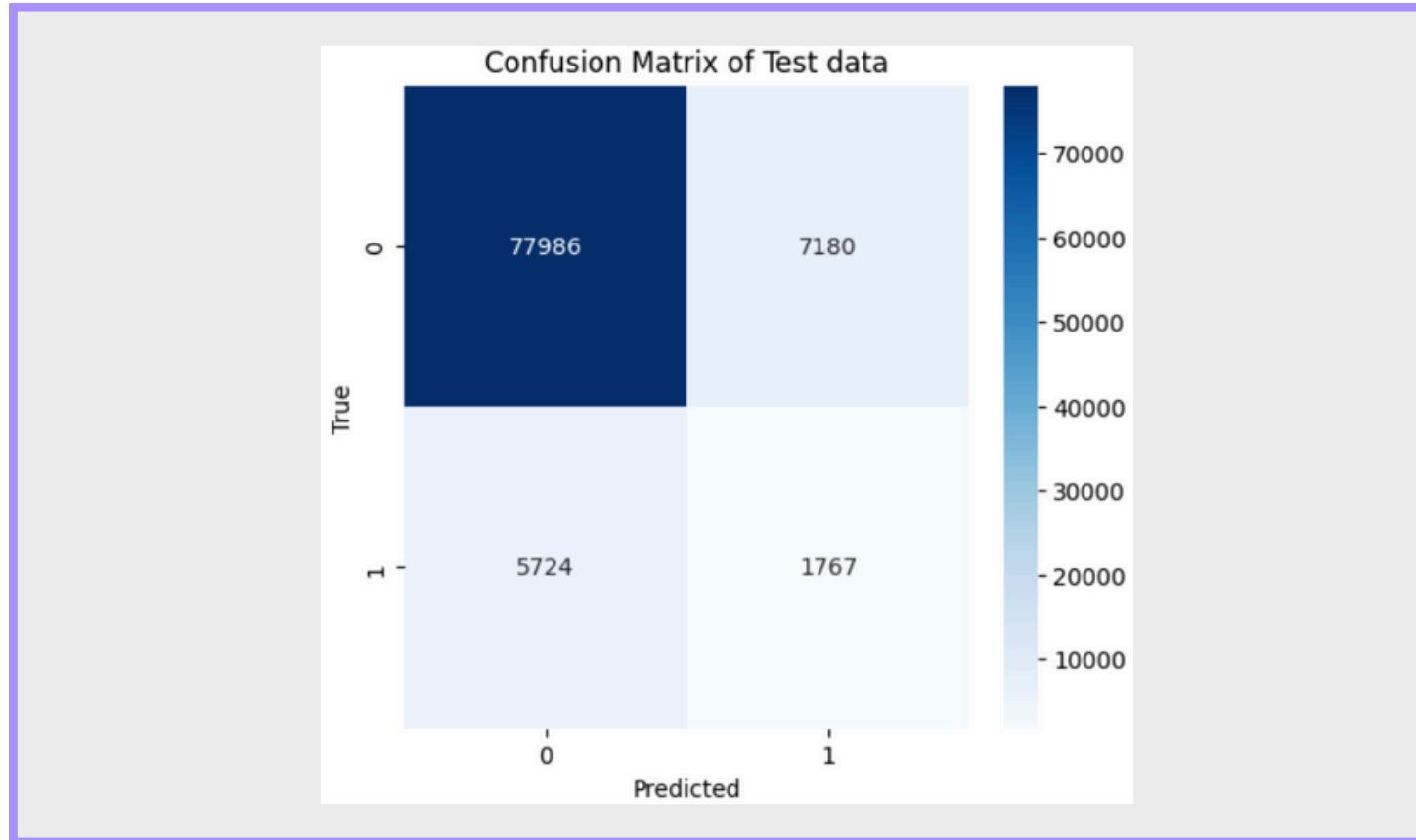
	precision	recall	f1-score	support
0.0	0.93	0.92	0.92	85166
1.0	0.20	0.24	0.21	7491
accuracy			0.86	92657
macro avg	0.56	0.58	0.57	92657
weighted avg	0.87	0.86	0.87	92657

Model	Accuracy	Precision	Sensitivity (Recall)	F1-score
Class 0 (ไม่เป็นโรคหัวใจ)	0.86	0.93	0.92	0.92
Class 1 (เป็นโรคหัวใจ)		0.20	0.24	0.21

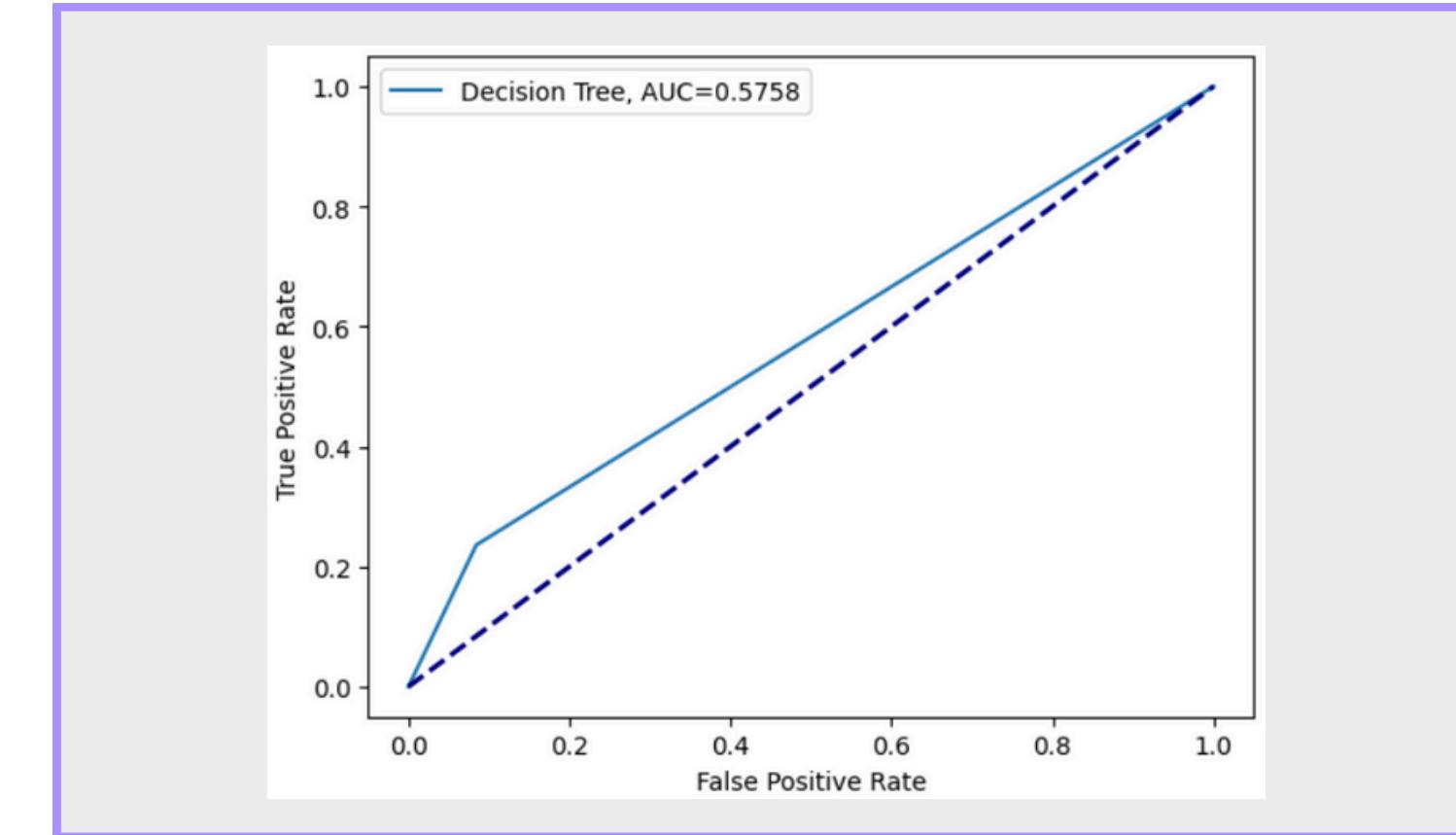
- Accuracy: โมเดลมีประสิทธิภาพในการ Predict โดยรวมค่อนข้างดี
- Precision: โมเดลมีความแม่นยำในการ Predict ในกรณีที่คนไม่เป็นโรคหัวใจได้ค่อนข้างดี แต่มีความแม่นยำในการ Predict ในกรณีที่คนเป็นโรคหัวใจได้ไม่ดี
- Recall: โมเดลมีความสามารถในการตรวจจับคนที่ไม่เป็นโรคหัวใจจริงๆ ได้ดี แต่ตรวจจับคนเป็นโรคหัวใจไม่ดี
- F1-Score: โมเดลมีความสมดุลระหว่าง Precision และ Recall ใน Class 0 แต่มีความไม่สมดุลใน Class 1

Decision Tree

ผลลัพธ์ของ Confusion Matrix และ AUC-ROC Curve



ผลลัพธ์ Confusion Matrix ของชุดข้อมูลทดสอบ



ผลลัพธ์ของ AUC-ROC Curve

True Negative (TN) : โมเดลทำนายว่าไม่เป็นโรคหัวใจซึ่งผลลัพธ์จริงคือไม่เป็นโรคหัวใจ โดยทำนายได้ 77986 คน

False Negative (FN) : โมเดลทำนายว่าไม่เป็นโรคหัวใจซึ่งผลลัพธ์จริงคือเป็นโรคหัวใจ โดยทำนายได้ 5724 คน

False Positive (FP) : โมเดลทำนายว่าเป็นโรคหัวใจซึ่งผลลัพธ์จริงคือไม่เป็นโรคหัวใจ โดยทำนายได้ 7180 คน

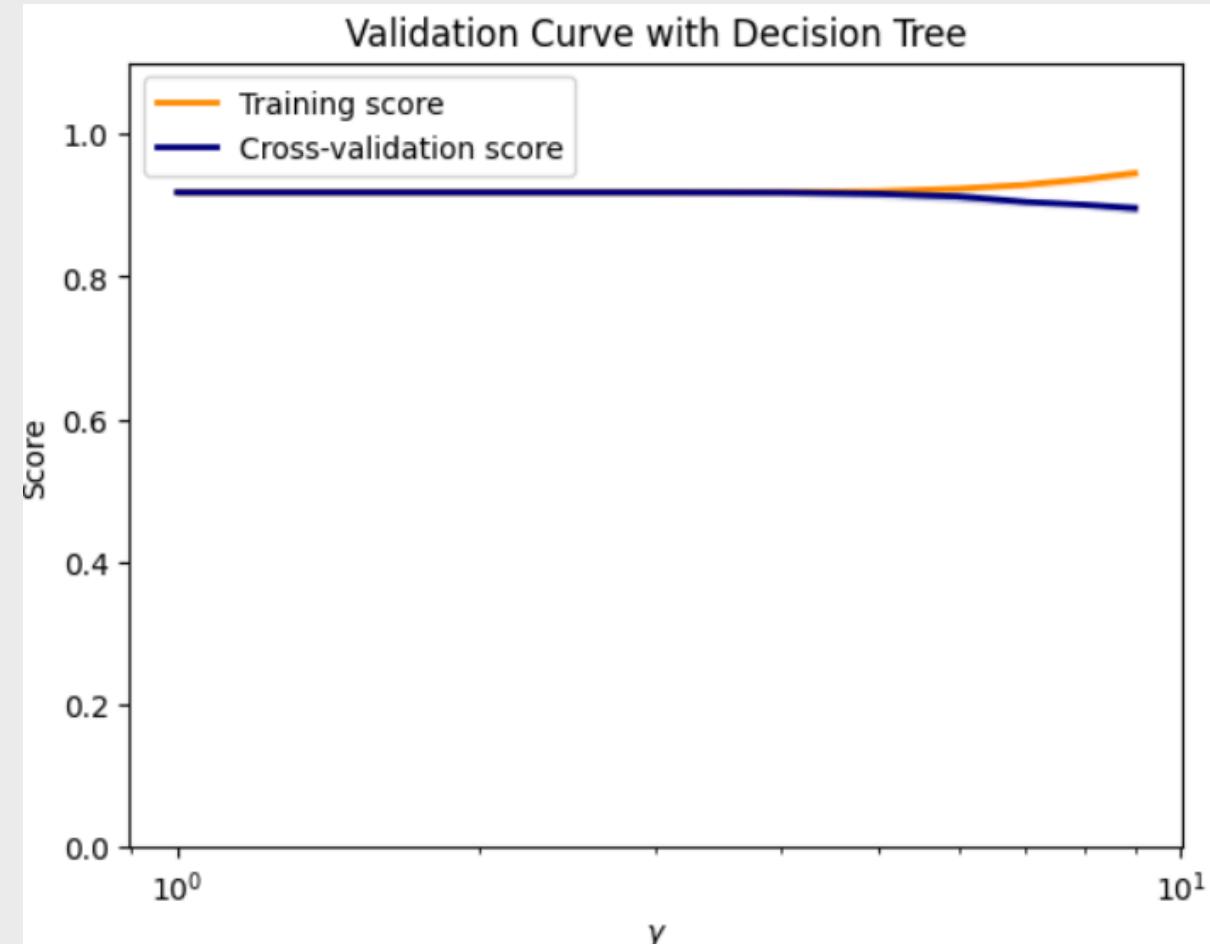
True Positive (TP) : โมเดลทำนายว่าเป็นโรคหัวใจซึ่งผลลัพธ์จริงคือเป็นโรคหัวใจ โดยทำนายได้ 1767 คน

เส้น ROC นั้นวัดสัดส่วนของ Sensitivity (Recall) และ (1-Specificity) โดยที่ค่า Specificity หาได้จาก

$TN / (TN + FP)$ ซึ่งจากการมีค่า AUC เท่ากับ 57.58% ซึ่งแสดงว่าโมเดลมีประสิทธิภาพโดยรวมระดับปานกลาง

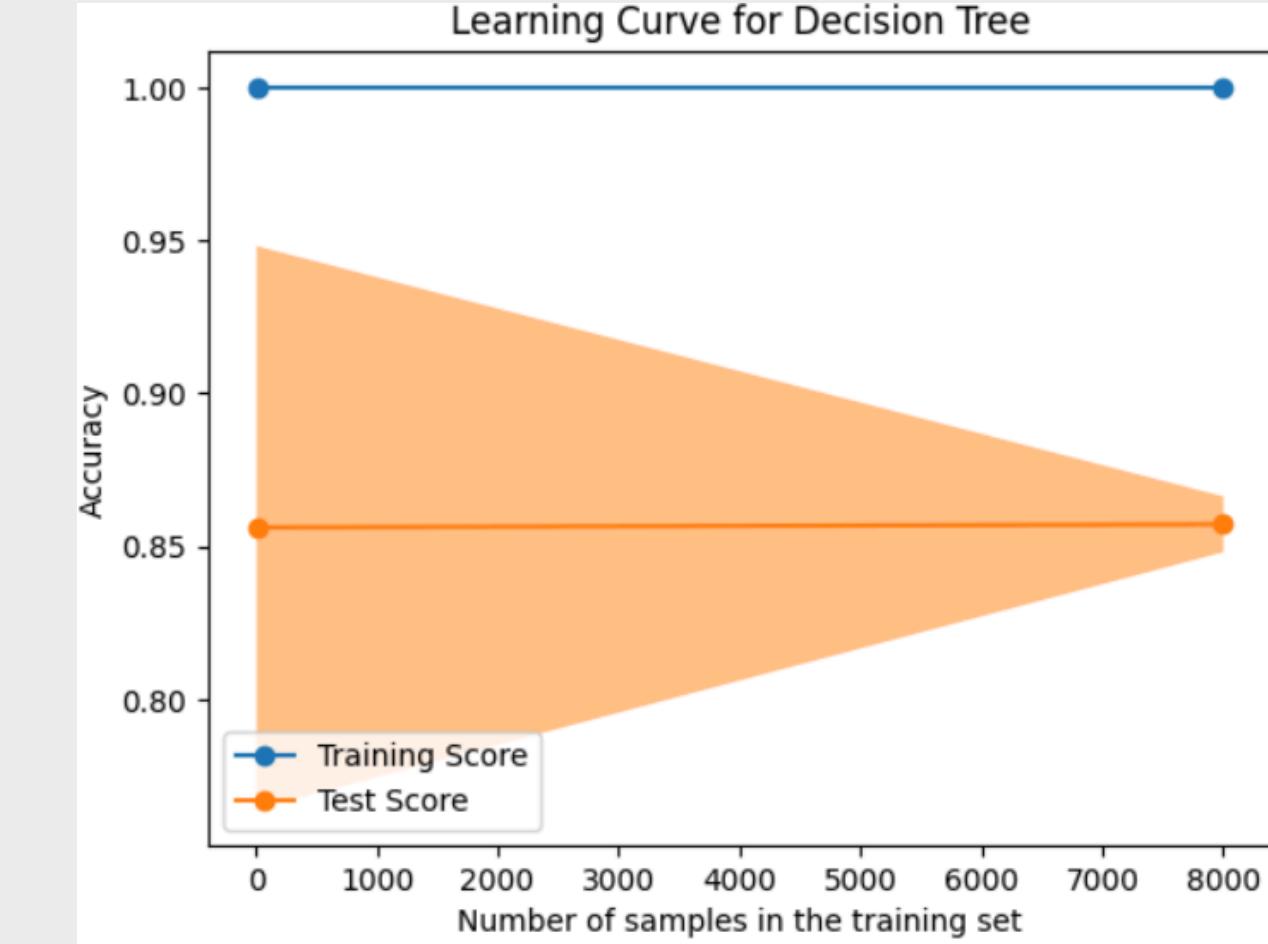
Decision Tree

Validation Curve และ Learning Curve



ผลลัพธ์ Validation Curve

แกน x คือ ค่า max depth ตัวควบคุมว่าจะให้ต้นไม้ของเรามีความลึกกี่ชั้นยิ่งค่า depth มากโมเดลก็จะซับซ้อนยิ่งขึ้นและแกน Y คือ Score



ผลลัพธ์ Learning Curve

แกน x คือ จำนวนข้อมูลฝึกสอนที่ใช้ และแกน y คือ ค่า Accuracy ซึ่งสามารถสรุปจากโมเดลได้ว่าอาจจะต้องเพิ่มข้อมูลฝึกสอนให้มากขึ้น

Random Forest

ปรับข้อมูลเชิงประมาณให้อยู่ในรูปการแจกแจงแบบปกติ (Normal Distribution)

นำไป Predict กับข้อมูล Test

```
# Random Forest  
# สร้าง Pipeline สำหรับทำ Scaler และ ML  
rf_pipeline = Pipeline([  
    ('scaler', StandardScaler()),  
    ('rf', RandomForestClassifier(random_state = 42))  
])  
rf_pipeline.fit(x_train_resampled,y_train_resampled)  
rf_preds_train = rf_pipeline.predict(x_train_resampled)  
rf_preds_test = rf_pipeline.predict(x_test)
```

Random Forest

ผลลัพธ์ Classification Report ของชุดข้อมูลทดสอบ

===== Classification Report (Test data) =====

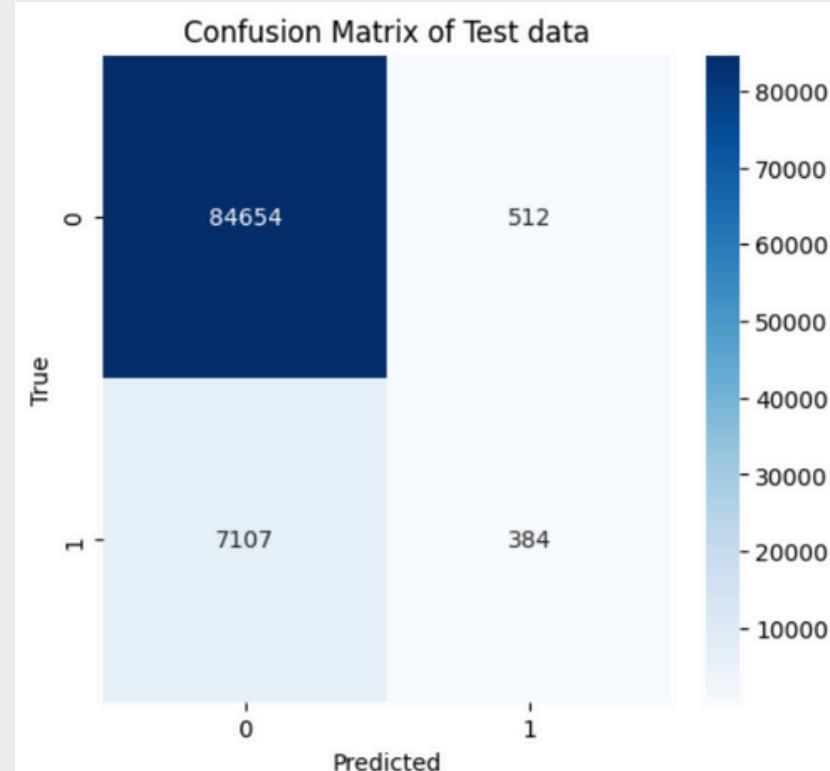
	precision	recall	f1-score	support
0.0	0.97	0.74	0.84	85166
1.0	0.20	0.75	0.32	7491
accuracy			0.74	92657
macro avg	0.59	0.75	0.58	92657
weighted avg	0.91	0.74	0.80	92657

Model	Accuracy	Precision	Sensitivity (Recall)	F1-score
Class 0 (ไม่เป็นโรคหัวใจ)	0.74	0.97	0.74	0.84
Class 1 (เป็นโรคหัวใจ)		0.20	0.75	0.32

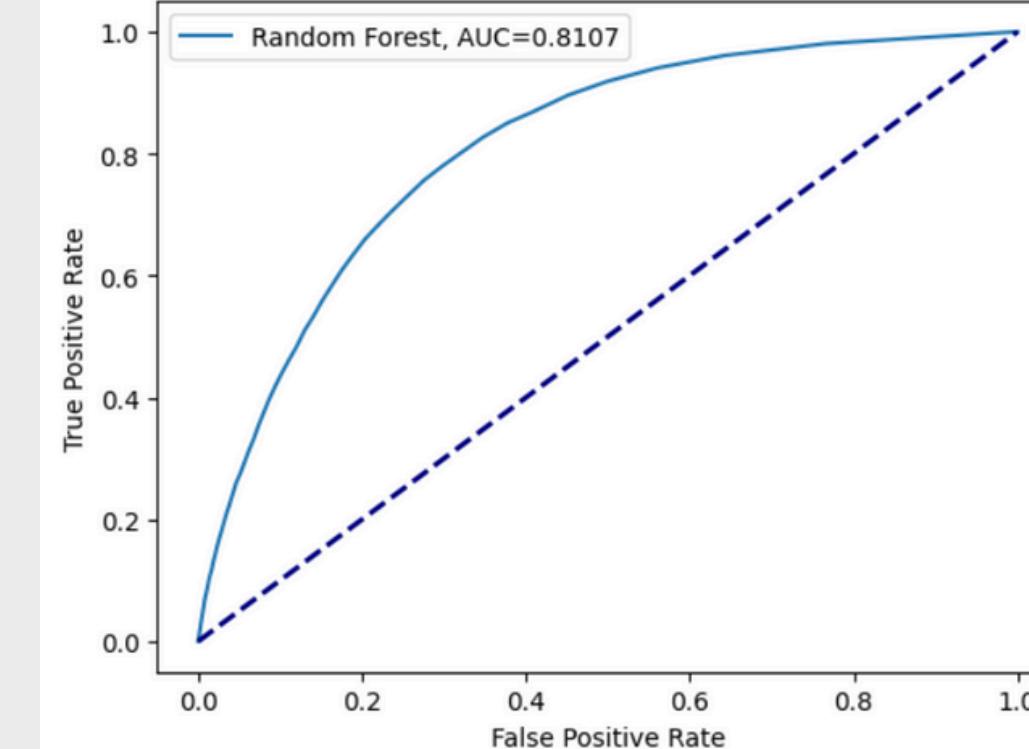
- Accuracy: โมเดลมีประสิทธิภาพในการ Predict โดยรวมค่อนข้างดี
- Precision: โมเดลมีความแม่นยำในการ Predict ในกรณีที่คนไม่เป็นโรคหัวใจได้ค่อนข้างดี แต่มีความแม่นยำในการ Predict ในกรณีที่คนเป็นโรคหัวใจได้ไม่ดี
- Recall: โมเดลมีความสามารถในการตรวจจับคนที่เป็นและไม่เป็นโรคหัวใจจริงๆ ได้ดี
- F1-Score: โมเดลมีความสมดุลระหว่าง Precision และ Recall ใน Class 0 แต่มีความไม่สมดุลใน Class 1

Random Forest

ผลลัพธ์ของ Confusion Matrix และ AUC-ROC Curve



ผลลัพธ์ Confusion Matrix ของชุดข้อมูลทดสอบ

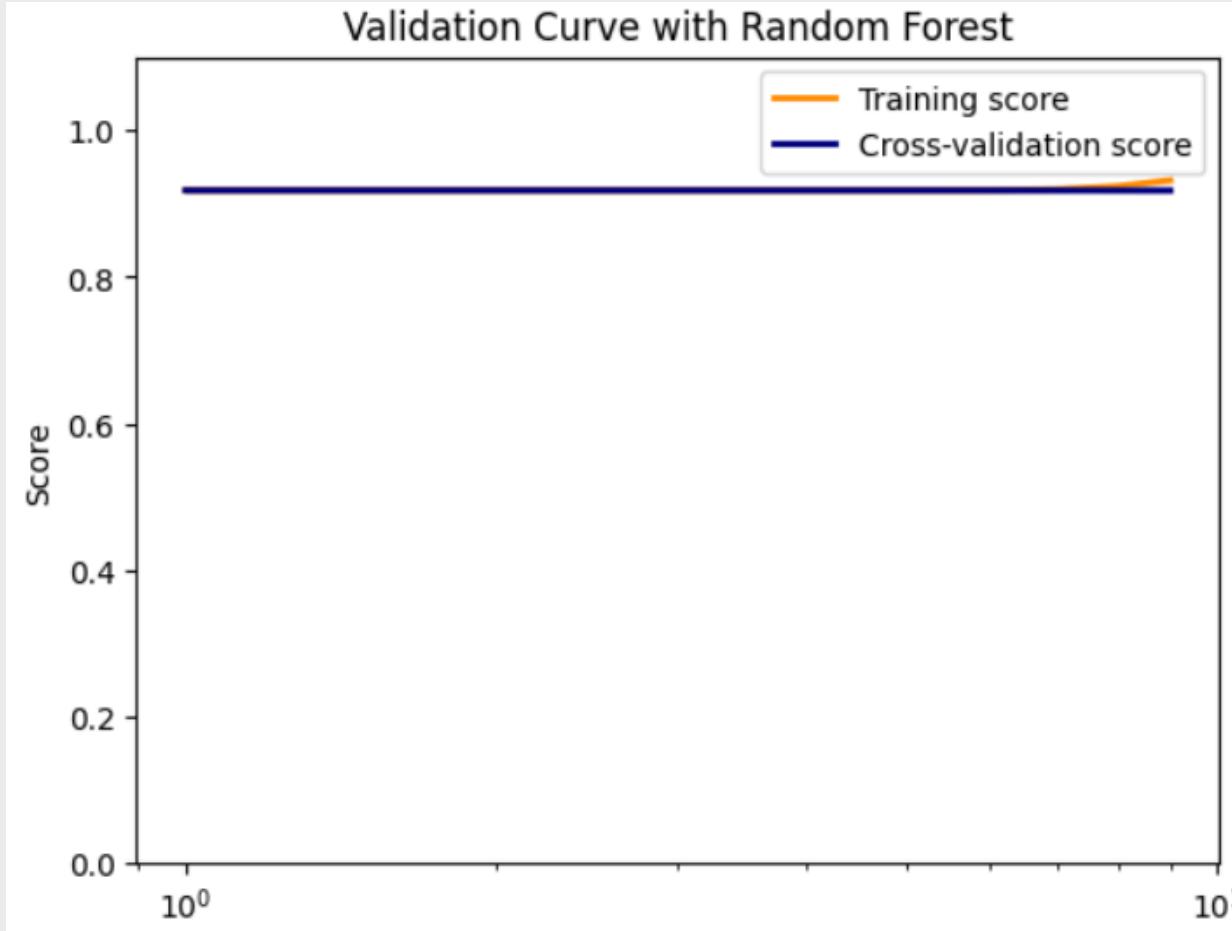


ผลลัพธ์ของ AUC-ROC Curve

True Negative (TN) : โมเดลทำนายว่าไม่เป็นโรคหัวใจซึ่งผลลัพธ์จริงคือไม่เป็นโรคหัวใจ โดยทำนายได้ 84654 คน
False Negative (FN) : โมเดลทำนายว่าไม่เป็นโรคหัวใจซึ่งผลลัพธ์จริงคือเป็นโรคหัวใจ โดยทำนายได้ 7107 คน
False Positive (FP) : โมเดลทำนายว่าเป็นโรคหัวใจซึ่งผลลัพธ์จริงคือไม่เป็นโรคหัวใจ โดยทำนายได้ 512 คน
True Positive (TP) : โมเดลทำนายว่าเป็นโรคหัวใจซึ่งผลลัพธ์จริงคือเป็นโรคหัวใจ โดยทำนายได้ 384 คน
เส้น ROC นั้นวัดสัดส่วนของ Sensitivity (Recall) และ (1-Specificity) โดยที่ค่า Specificity หาได้จาก $TN / (TN + FP)$ ซึ่งจากการมีค่า AUC เท่ากับ 81.07% ซึ่งแสดงว่าโมเดล มีประสิทธิภาพโดยรวมระดับดี

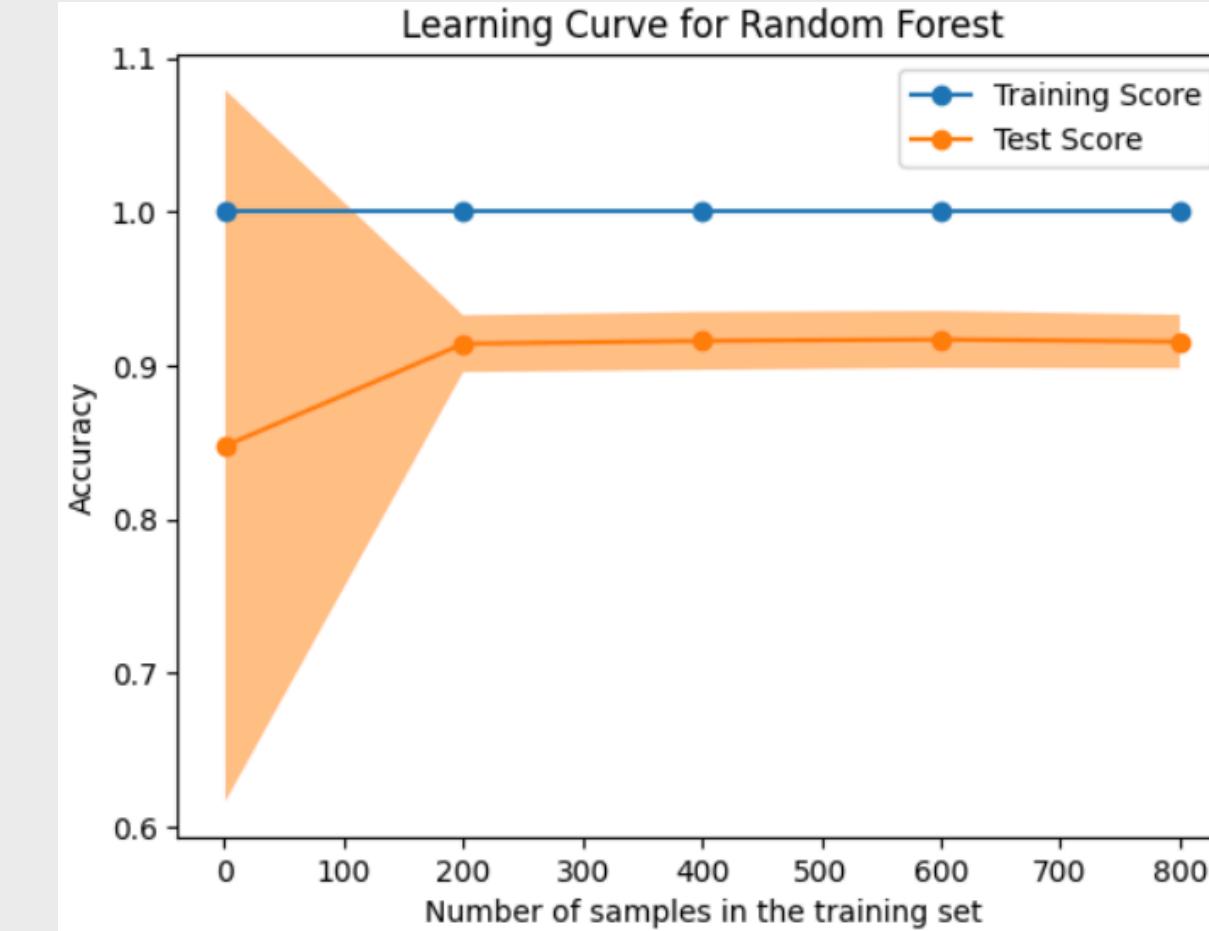
Random Forest

Validation Curve และ Learning Curve



ผลลัพธ์ Validation Curve

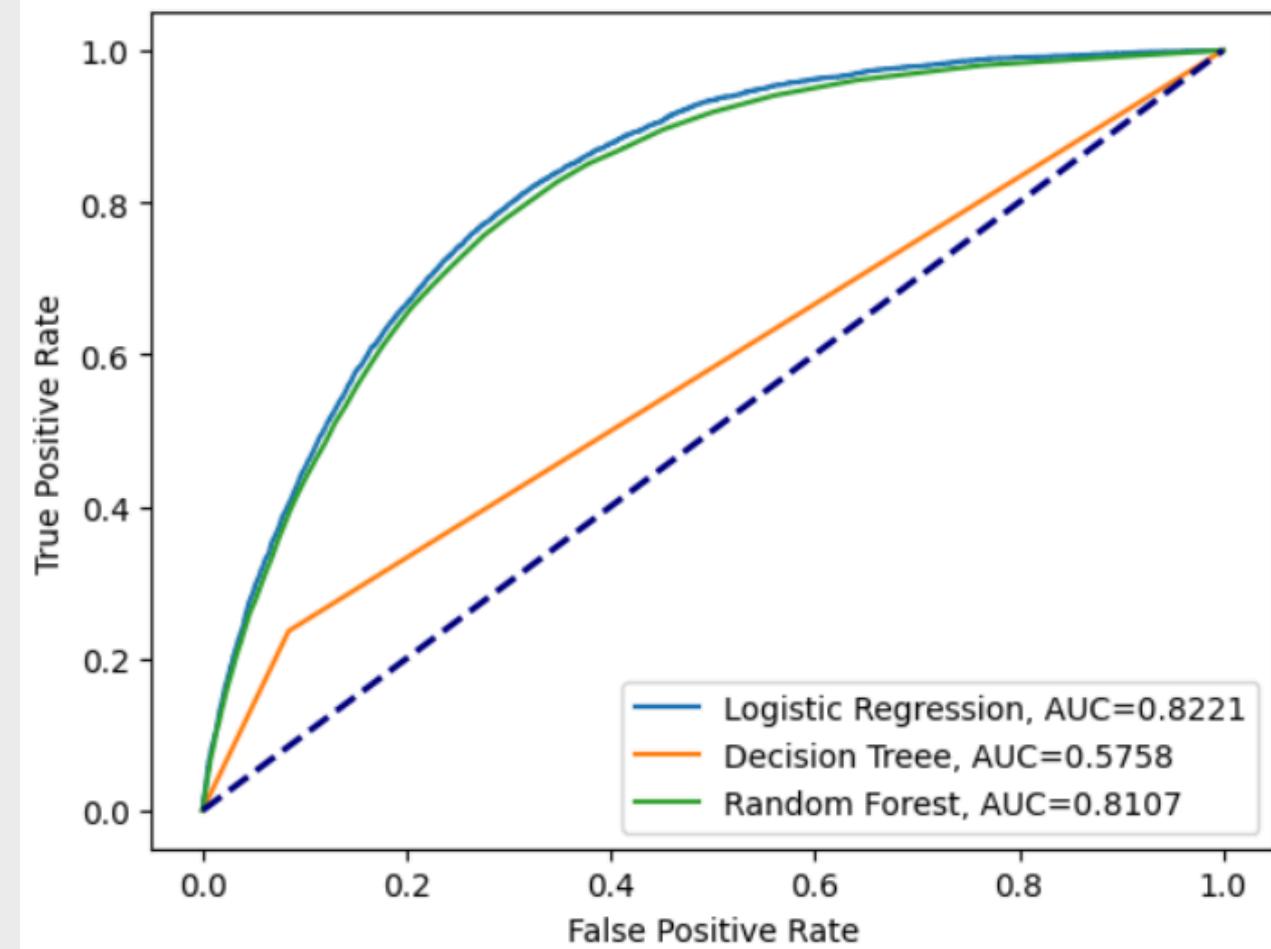
แกน x คือ ค่า max depth ตัวควบคุมว่าจะให้ต้นไม้ของเรามีความลึกกี่ชั้นยิ่งค่า depth มากโมเดลก็จะซับซ้อนยิ่งขึ้นและแกน Y คือ Score



ผลลัพธ์ Learning Curve

แกน x คือ จำนวนข้อมูลฝึกสอนที่ใช้ และแกน y คือค่า Accuracy จากโมเดลสรุปได้ว่าอาจจะต้องเพิ่มข้อมูลฝึกสอนให้มากขึ้น

สรุปผลลัพธ์ทั้งหมด



ผลลัพธ์ AUC-ROC Curve ของทั้ง 3 โมเดล

โมเดล Logistic Regression มีค่า AUC = 82.21% ซึ่งสูงที่สุด รองลงมาคือ Random Forest และ Decision Tree

Model	Accuracy
Logistic Regression	0.74 (74%)
Decision Tree	0.86 (86%)
Random Forest	0.74 (74%)

ผลลัพธ์ค่า Accuracy ของทั้ง 3 โมเดล

ค่า Accuracy ที่แสดงภาพรวมของความแม่นยำในการทำนายปรากฏว่า Decision Tree สูงที่สุด แต่ทั้งนี้อาจจำต้องดูสถิติอื่นช่วยด้วยเช่น Precision, Recall เป็นต้น