



MACHINE LEARNING PROJECT

Red Wine Quality Prediction

WITH SUPPORT VECTOR MACHINE (SVM)

MADE BY
PORNPAILIN KAMONSANTISUK 6310401068



OUTLINE

- ★ INTRODUCTION
- ★ DATASET
- ★ TRAINING MODEL
- ★ FUTURE PREDICTION
- ★ SUMMARY



INTRODUCTION

เนื่องจากในปัจจุบันการดื่มไวน์แดงเพื่อเข้าสังคมได้รับความนิยมมากขึ้นอุตสาหกรรมไวน์แดงจึงมีการขยายตัวเพิ่มขึ้น ในขณะเดียวกันการรับรองคุณภาพของไวน์แดงยังไม่มีมาตรฐานที่ชัดเจน นอกเหนือจากการทดสอบโดยการซิมของบุษย์

ดังนั้นการทดสอบทางเคมีกายภาพ คำนึงถึงองค์ประกอบต่างๆ เช่น ความเป็นกรด ระดับค่า PH น้ำตาล และคุณสมบัติทางเคมีอื่นๆ เป็นองค์ประกอบที่สำคัญในการประเมินและรับรองคุณภาพของไวน์แดง



PROBLEM & GOAL

PROBLEM: เปรียบเทียบ การคำนวณคุณภาพของไวน์แดง ระหว่างองค์ประกอบทางเคมีกายภาพทั้งหมด และ 2 ปัจจัยที่ส่งผลต่อคุณภาพของไวน์แดง

GOAL: คำนวณคุณภาพของไวน์แดงโดยกำหนดค่าคุณภาพของไวน์แดง ดังต่อไปนี้

QUALITY ≤ 5 : BAD WINE

QUALITY > 5 : GOOD WINE

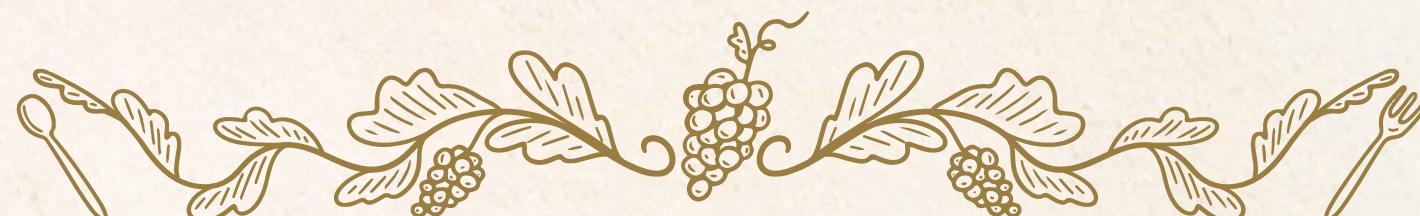
EXPECTED ACCURACY: $\geq 80\%$



DATASET

ข้อมูลชุดนี้เป็นข้อมูลเกี่ยวกับคุณภาพของไวน์และ “VINHO VERDE” จากประเทศโปรตุเกส อธิบายเกี่ยวกับปริมาณของสารเคมีต่างๆที่อยู่ในไวน์และคุณภาพของไวน์และมีทั้งหมด 1599 rows 12 columns

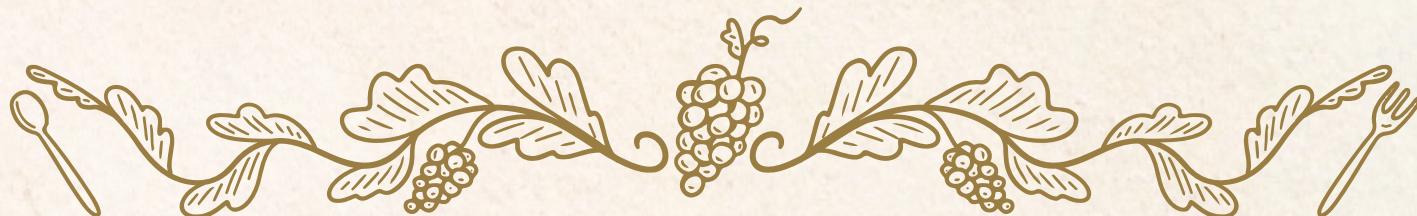
Fixed acidity	Volatile acidity	Citric acid	Residual sugar	Chlorides	Free sulfur dioxide	Total sulfur dioxide	Density	pH	Sulphates	Alcohol	Quality
7.4	0.7	0	1.9	0.08	11	34	1	3.51	0.56	9.4	5
7.8	0.88	0	2.6	0.1	25	67	1	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.09	15	54	1	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.08	17	60	1	3.16	0.58	9.8	6
7.4	0.7	0	1.9	0.08	11	34	1	3.51	0.56	9.4	5
7.4	0.66	0	1.8	0.08	13	40	1	3.51	0.56	9.4	5
7.9	0.6	0.06	1.6	0.07	15	59	1	3.3	0.46	9.4	5
7.3	0.65	0	1.2	0.07	15	21	0.99	3.39	0.47	10	7
7.8	0.58	0.02	2	0.07	9	18	1	3.36	0.57	9.5	7
7.5	0.5	0.36	6.1	0.07	17	102	1	3.35	0.8	10.5	5



DATASET

Input variables (based on physicochemical tests):

- fixed acidity - ความเป็นกรดคงที่ บ่งบอกถึงกรดอันตรีย์ที่ไม่ระเหยในไวน์ ส่วนใหญ่มาจากการองุ่น
- volatile acidity - ความเป็นกรดระเหย บ่งบอกถึงกรดอันตรีย์ที่ระเหยได้ เกิดขึ้นระหว่างการหมัก
- citric acid - กรดซิตริก กรดอันตรีย์ที่พบในองุ่น
- residual sugar - น้ำตาลตากด้าง คือ น้ำตาลที่ไม่ได้ถูกย่อยสลายระหว่างการหมัก
- chlorides - คลอไรด์ เกลือแร่ ส่งผลต่อรสเด็ด และความเป็นกรด



DATASET

Input variables (based on physicochemical tests):

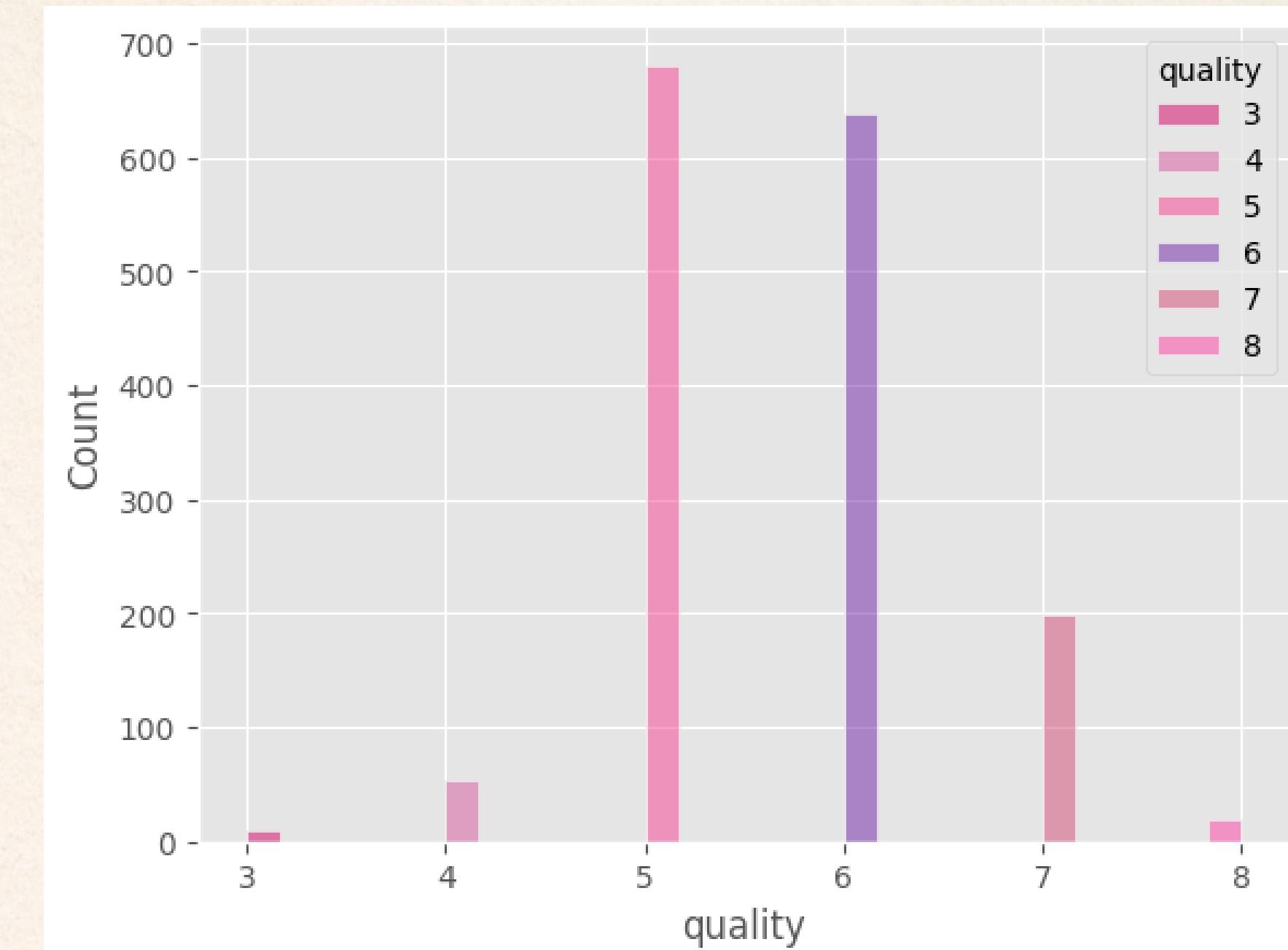
- free sulfur dioxide - ซัลเฟอร์ไดออกไซด์อิสระ สารต้านอนุมูลอิสระที่เติมลงในไวน์ ช่วยป้องกันการเสื่อมสภาพ
- total sulfur dioxide - ซัลเฟอร์ไดออกไซด์อิสระ + ซัลเฟอร์ไดออกไซด์ที่จับกับสารอื่น
- density - ความหนาแน่น ไวน์อาจมีความหนาแน่นน้อยกว่าหรือมากกว่าน้ำได้
- pH - ความเป็นกรดหรือด่าง
- sulphates - ผลผลอยได้จากน้ำตาลในไวน์ที่หมักโดยยีสต์ ให้เป็นแอลกอฮอล์
- alcohol - แอลกอฮอล์ ผลิตภัณฑ์จากการหมักน้ำตาล



Quality	Counts
3	10
4	53
5	681
6	638
7	199
8	18

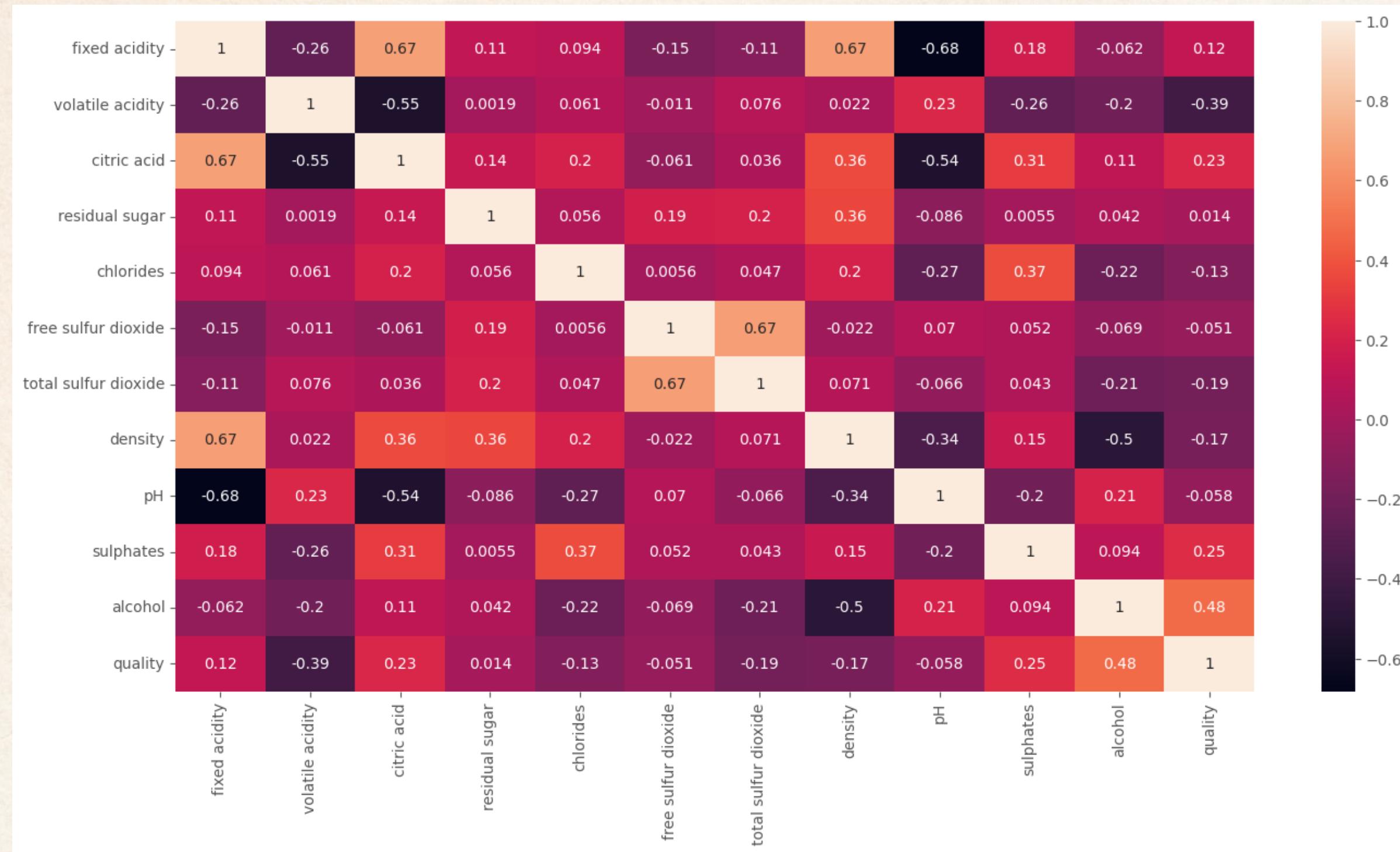
Output variable (based on sensory data):

- quality
- score between 0 and 10



DATA ANALYSIS

Compute pairwise correlation of columns



quality	1.00
alcohol	0.48
sulphates	0.25
citric acid	0.23
fixed acidity	0.12
residual sugar	0.014
free sulfur dioxide	-0.051
pH	-0.058
chlorides	-0.13
density	-0.17
total sulfur dioxide	-0.19
volatile acidity	-0.39

IMPORT LIBRARIES

```
import numpy as np  
import pandas as pd  
import seaborn as sns  
import matplotlib.pyplot as plt  
from sklearn import preprocessing  
plt.style.use("ggplot") #using style ggplot  
from sklearn.metrics import accuracy_score, confusion_matrix
```

1.Load red wine dataset

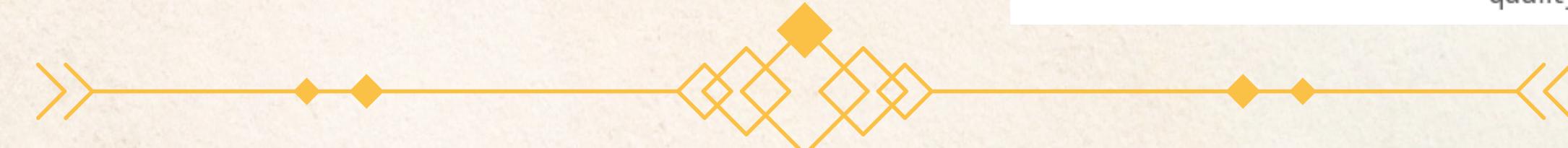
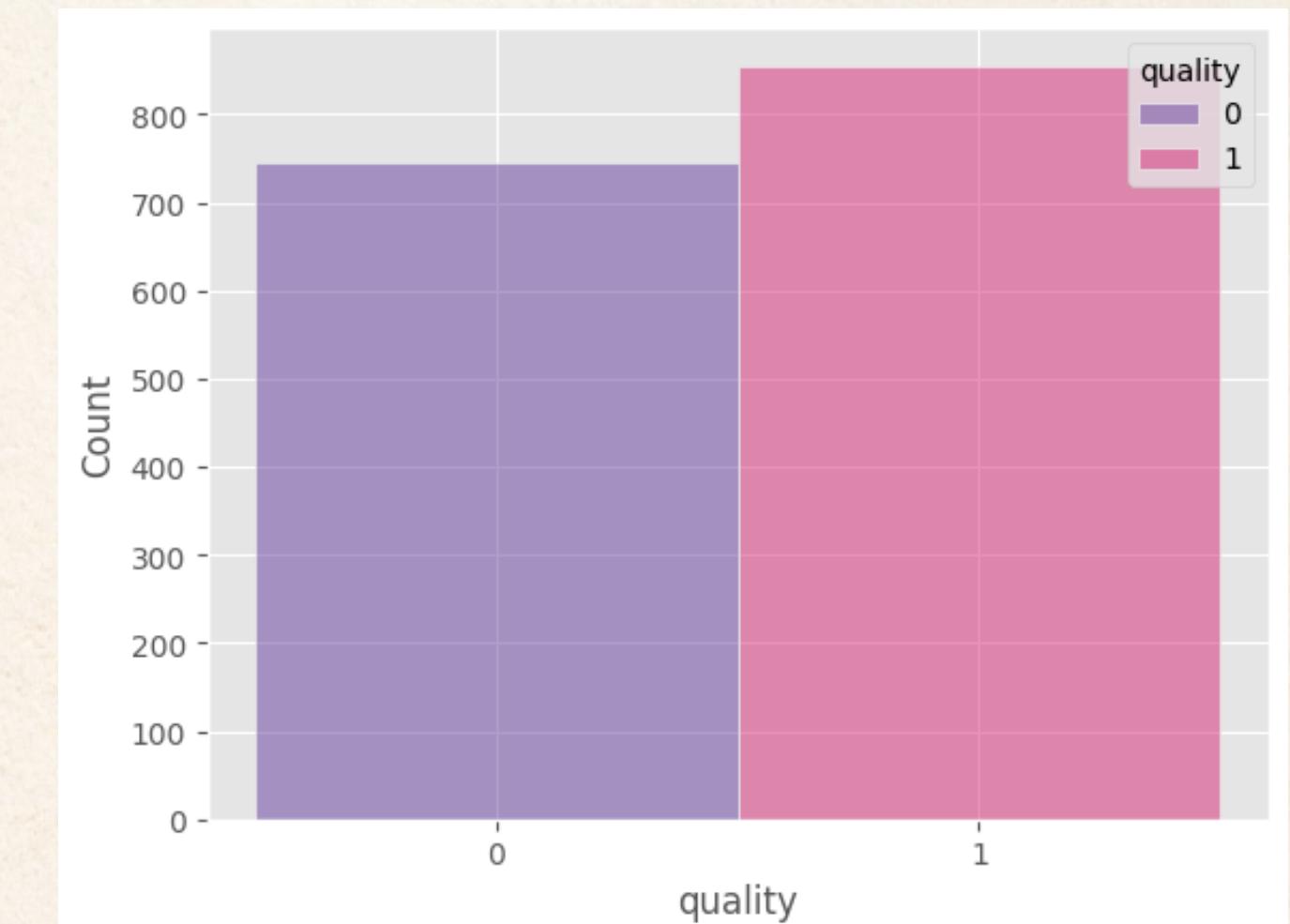
2.Preprocessing Data for binary classification

แบ่งคุณภาพของไวน์ (quality) เป็นสองกลุ่ม ดีอ "0" และ "1" โดยใช้วิธี
การกำหนดขอบเขต (binning) หรือการแบ่งช่วงของค่าคุณภาพ

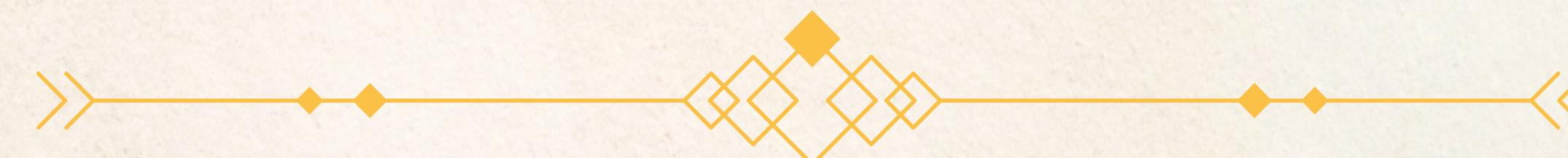
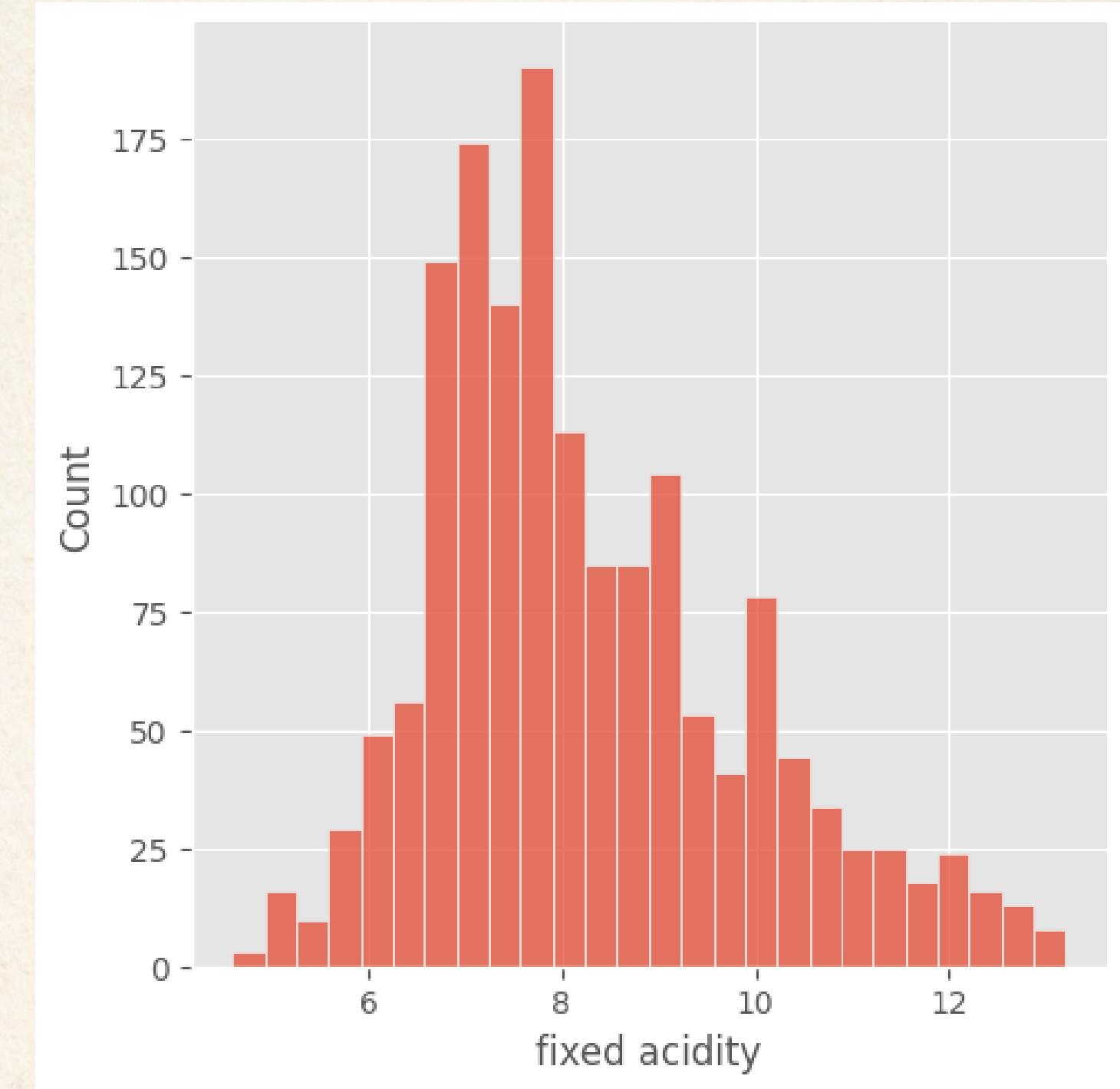
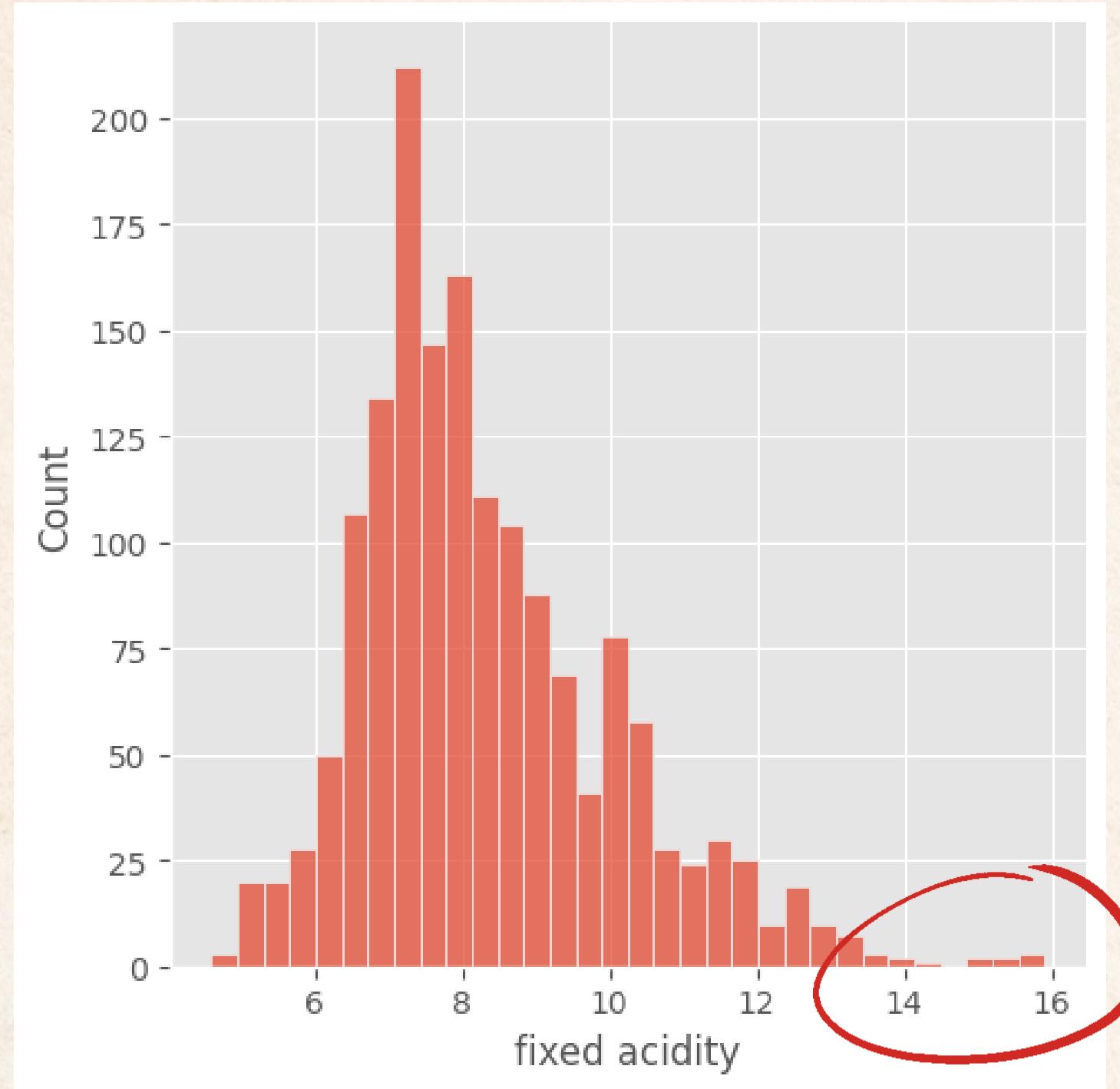
กำหนดให้

quality 2 ถึง 5 = "0" คุณภาพที่ไม่ดี

quality 5 ถึง 8 = "1" คุณภาพที่ดี



3. Select & Deal with the Outliers



4. Select features and target Variables for All features prediction

```
data_X = red_wine.drop('quality', axis=1)
```

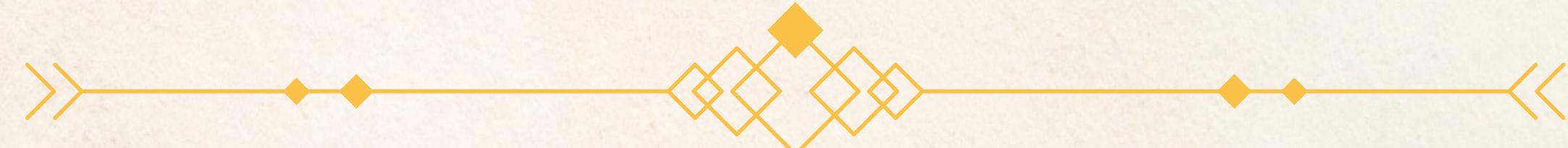
- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol

```
data_y = red_wine['quality']
```

- quality

```
print(data_X.shape)  
print(data_y.shape)
```

(1410, 11)
(1410,)



4. Select features and target Variables

for selected features : alcohol & fixed acidity

```
data_X = red_wine[['alcohol', 'fixed acidity']]
```

- alcohol
- fixed acidity

```
data_y = red_wine['quality']
```

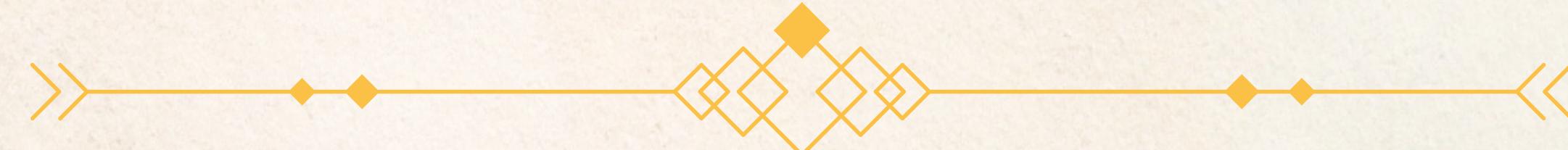
- quality

```
print(data_X.shape)
```

```
print(data_y.shape)
```

```
(1410, 2)
```

```
(1410,)
```



5. Split data set to training & test set

```
def train_test_split(X,y,test_size):  
    test_size = test_size  
    train_size = 1 - float(test_size)  
    total_rows = red_wine.shape[0]  
    split = int(total_rows * (train_size))  
    X_train = data_X[0:split]  
    X_test = data_X[split:]  
    y_train = data_y[0:split]  
    y_test = data_y[split:]  
    return X_train, X_test, y_train, y_test
```

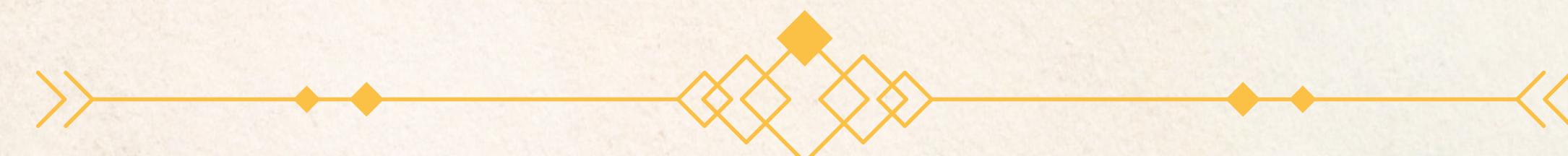
ກໍານົດໃຫ້ test size = 0.2

X_train: 1128

X_test: 282

y_train: 1128

y_test : 282



TRAINING MODEL

กำหนด parameter ให้ SVM ดังต่อไปนี้

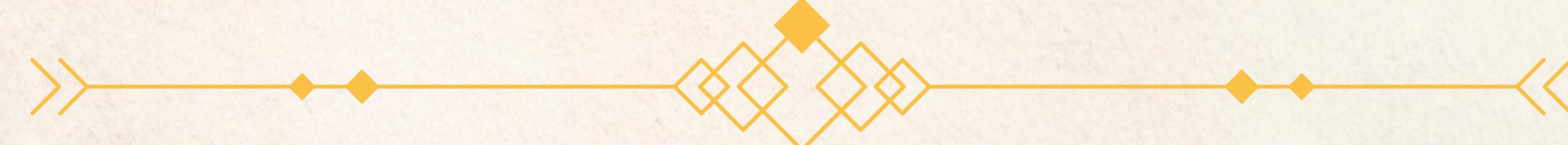
- learning_rate = 0.01
- lambda_param=0.0001
- n_iterations=1000
- initial w = 0
- b = 1000

for All features prediction

Accuracy Score: 71.28 %

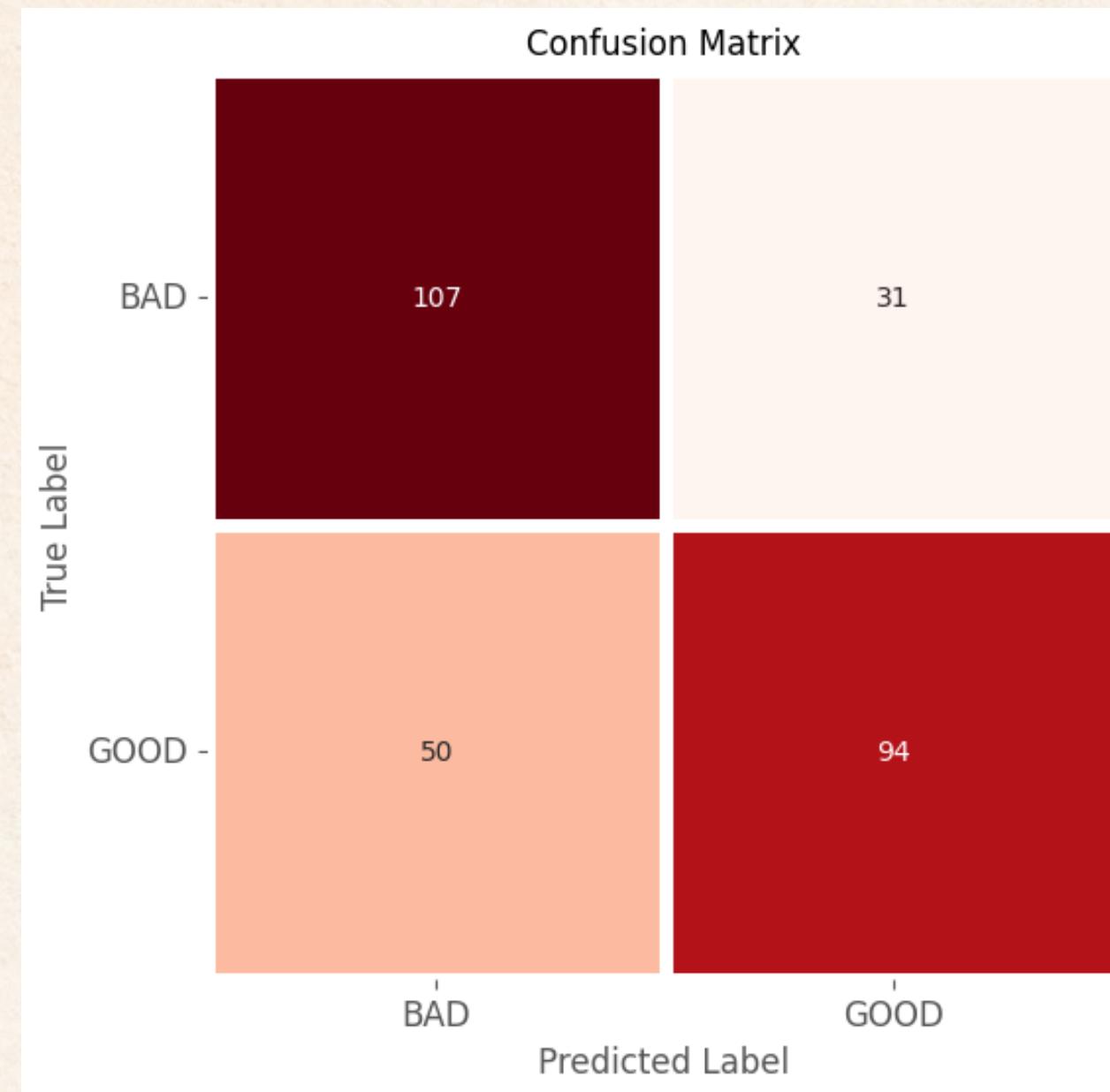
**for selected features :
alcohol & fixed acidity**

Accuracy Score: 72.70 %

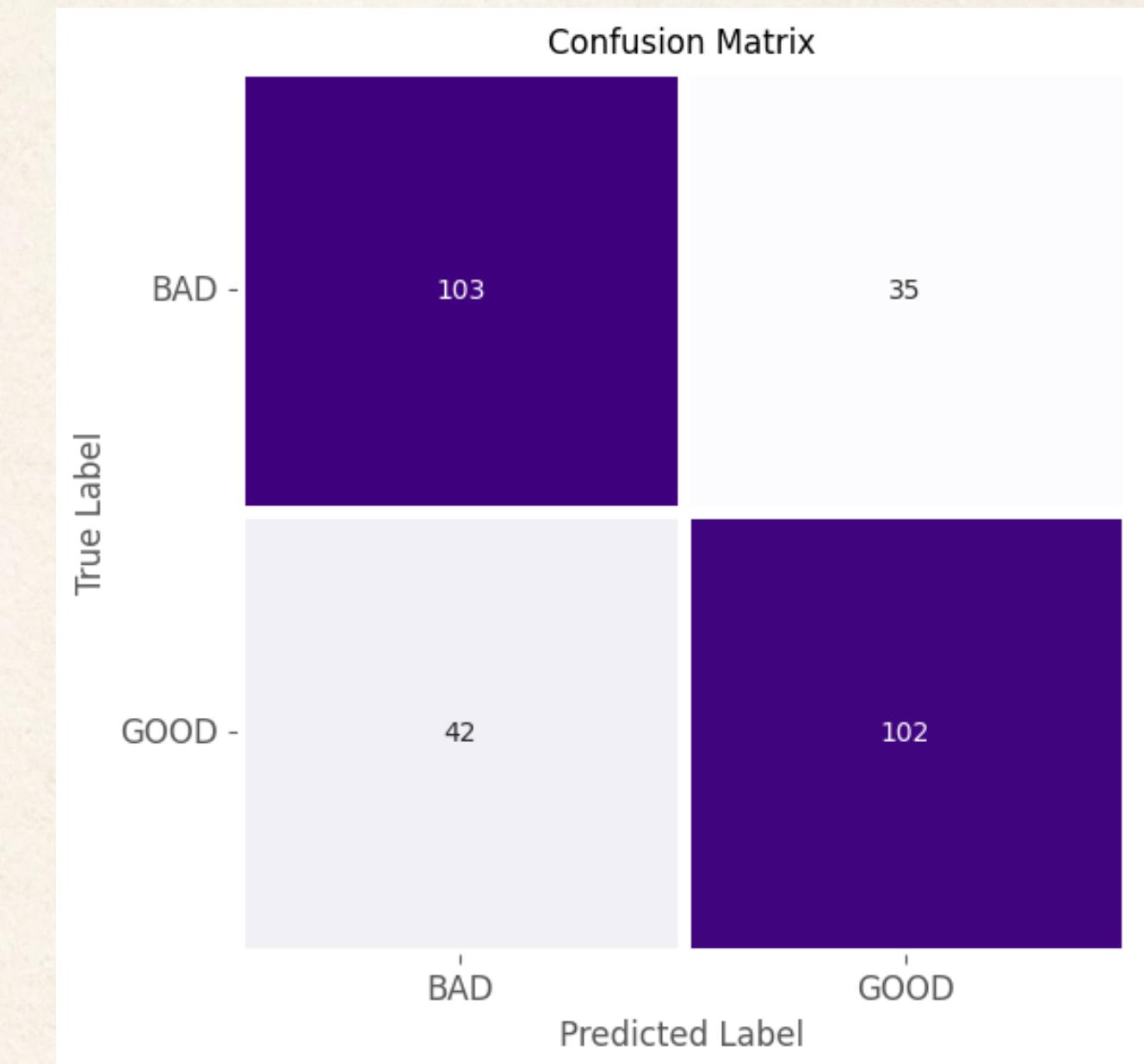


PREDICT USING THE TRAINED SVM MODEL

All features



Selected Features



FUTURE PREDICTIONS

- **Select or create new data points**

```
new_redwine = red_wine.sample(5)
```

- **Extract features for prediction for All Features**

```
new_data_X = new_redwine.drop('quality', axis=1)  
new_data_y = new_redwine[['quality']]
```

- **Extract features for prediction for Selected Features**

```
new_data_X = new_redwine.drop('quality', axis=1)  
new_data_y = new_redwine[['quality']]
```

FUTURE PREDICTIONS

- Make predictions using the trained SVM model

```
svm_y_pred_new = svm_model.predict(new_data_X.to_numpy())
```

- Map predicted labels to meaningful categories if necessary

```
svm_y_pred_mapped = np.where(svm_y_pred_new == 1, "GOOD", "BAD")
```

- Print the predicted quality labels

```
print("Predicted quality of new red wine samples:", svm_y_pred_mapped)
```

- Print the actual quality labels for comparison quality labels

```
print("Actual quality of new red wine samples:" , new_redwine['quality'].values)
```



RESULT OF FUTURE PREDICTIONS

All features

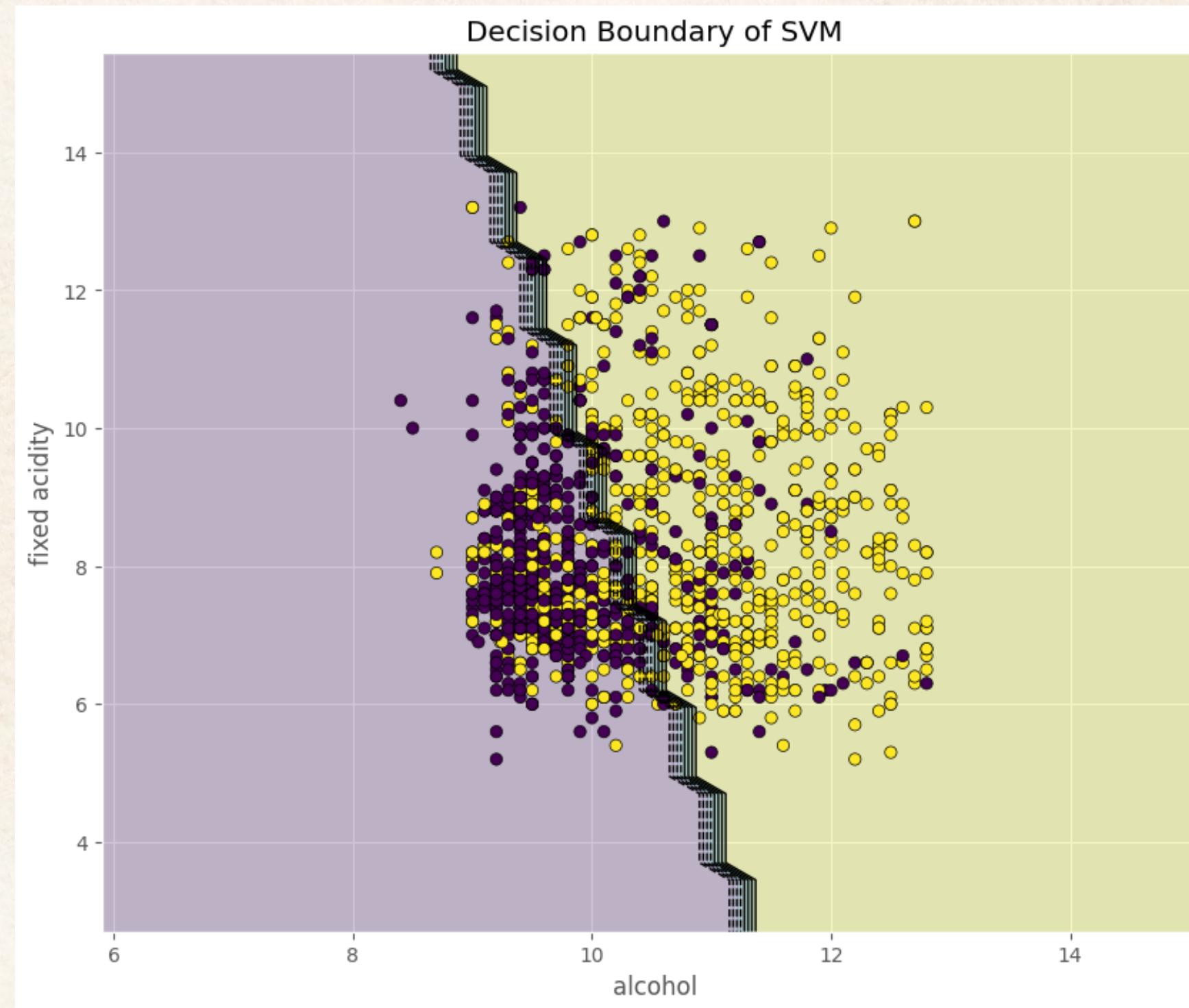
```
[All]Predicted quality of new red wine samples: ['BAD' 'GOOD' 'BAD' 'GOOD' 'GOOD']  
[All]Actual quality of new red wine samples: [-1 1 -1 1 1]
```

Selected Features

```
[Sec]Predicted quality of new red wine samples: ['BAD' 'GOOD' 'GOOD' 'GOOD' 'BAD']  
[Sec]Actual quality of new red wine samples: [-1 1 1 1 -1]
```



DECISION BOUNDARY OF SVM



Predict the labels for all points in the mesh grid

RED_WINE_SVM_1.00 (1)

ເວຼອຣໜັນນີ້ ກໍາຮັດໃຫ້

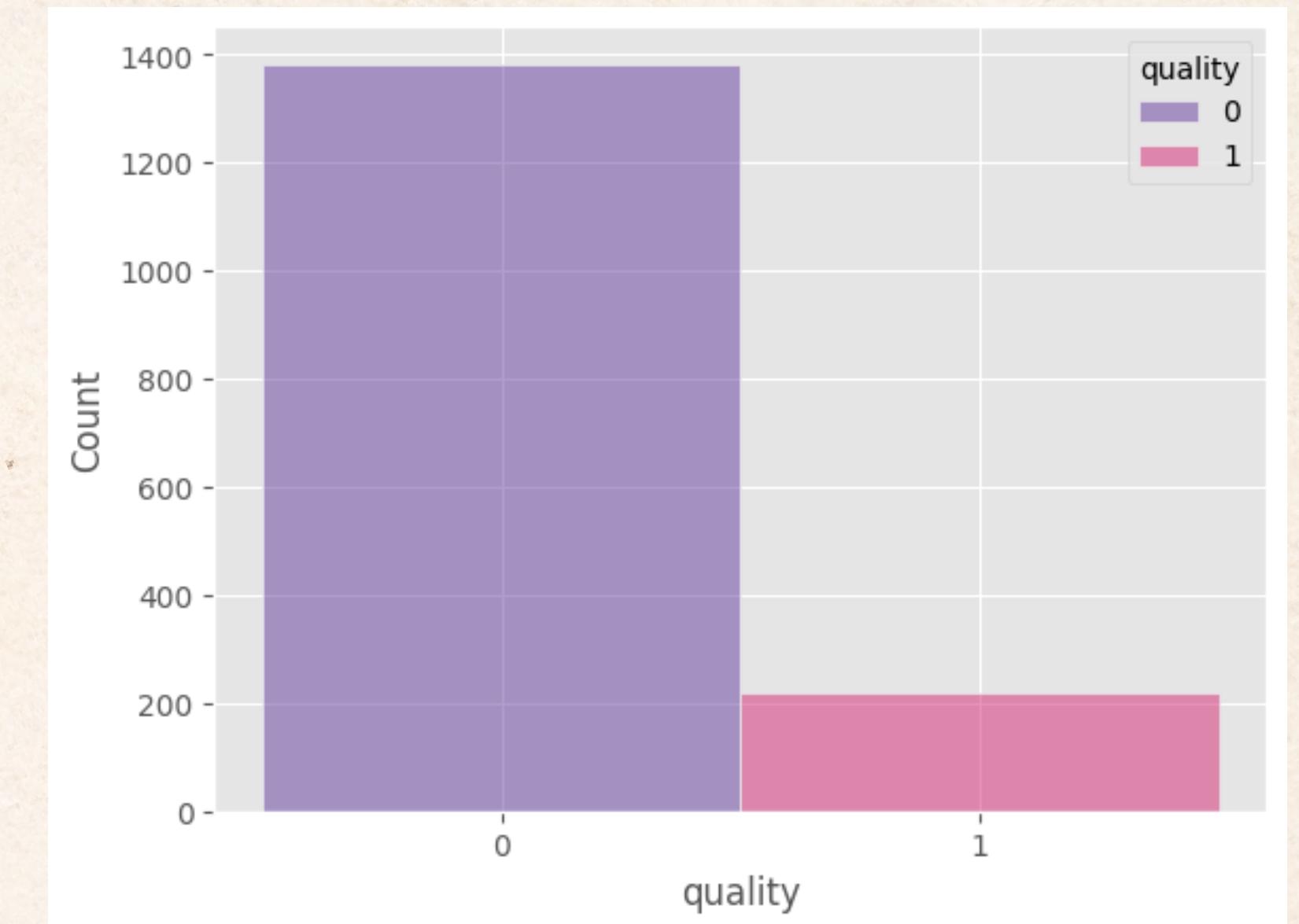
quality 2 ດຶງ 6 = "0" ມູນກາພທີມດີ

quality 6 ດຶງ 8 = "1" ມູນກາພທີດີ

ເນື່ອລອງນັບຈຳນວນພບວ່າ

quality = 0 ນັບໄດ້ 1382

quality = 1 ນັບໄດ້ 217



RED_WINE_SVM_1.00 (2)

ทำให้ค่าของ quality เกิดความไม่ balance กัน และ dataset นี้ก็มีค่า 5,6 ซึ่งเป็น 0 มากกว่า 7,8 อยู่แล้ว เมื่อเทรนนิ่งโมเดลตัวเดียวกัน ได้ผลดังต่อไปนี้

for All features prediction

Accuracy Score: 92.2 %

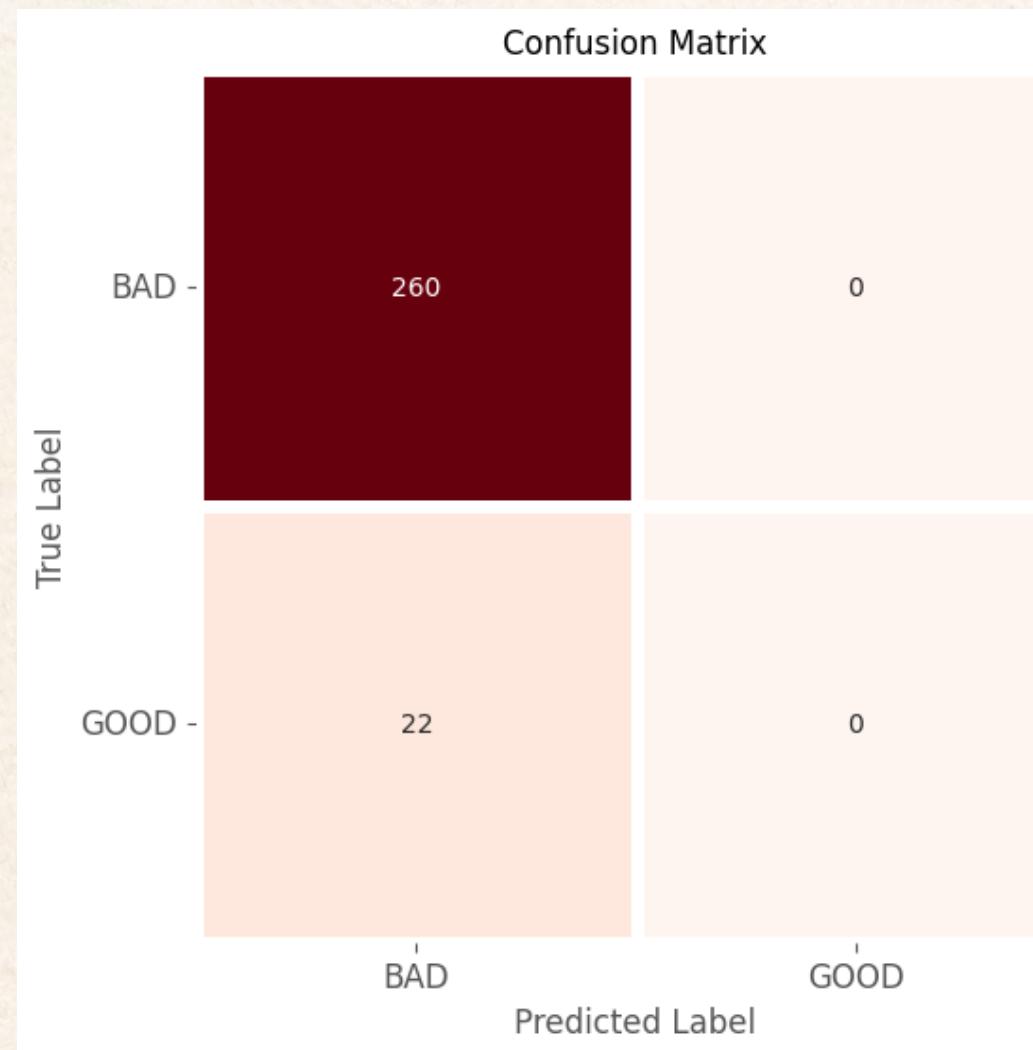
**for selected features :
alcohol & fixed acidity**

Accuracy Score: 90.4 %

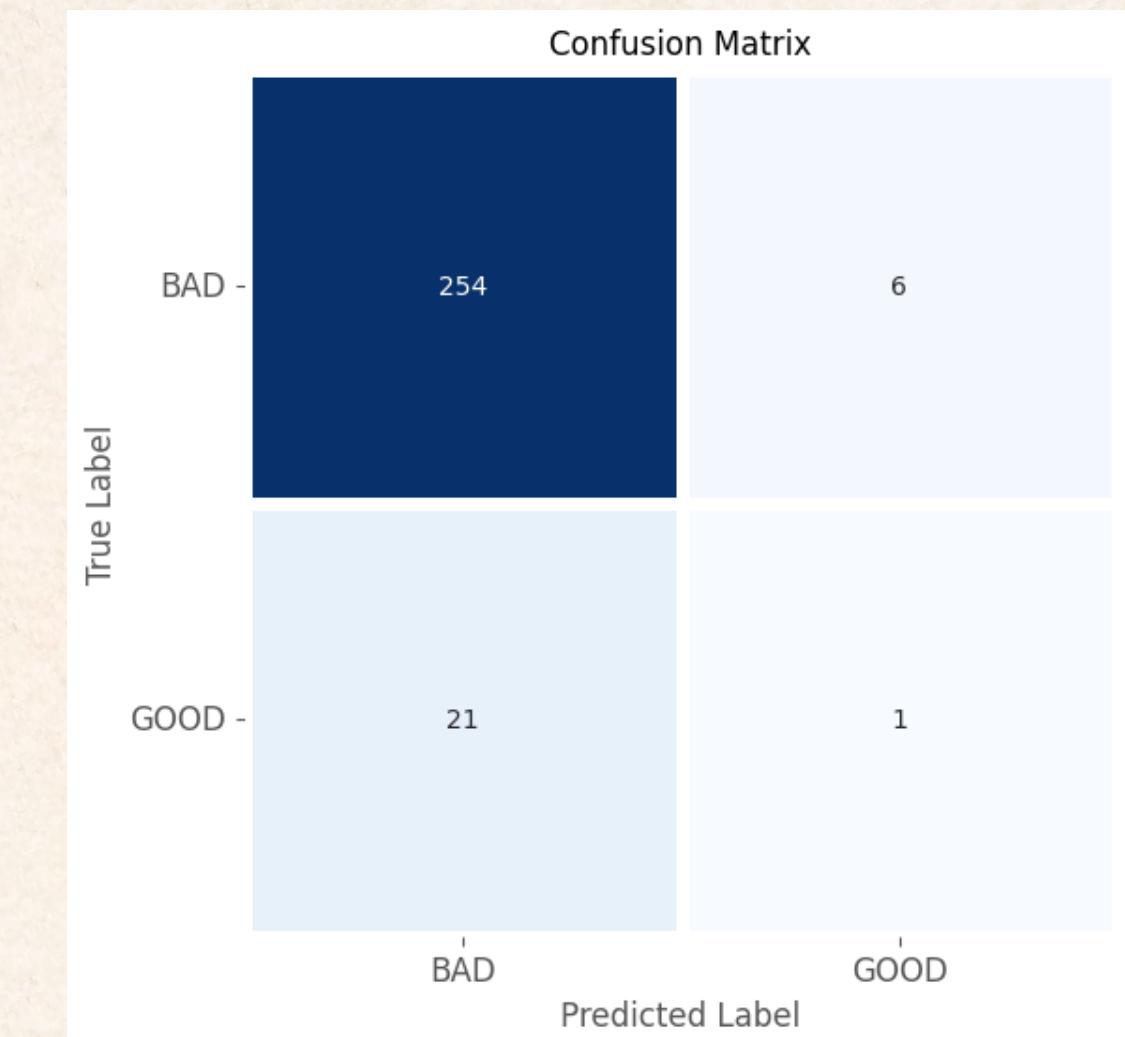


RED_WINE_SVM_1.00 (3)

All Features



Selected Features



เมื่อดูจาก confusion matrix ปรากฏว่าค่า accuracy ถีมาก
แต่ ไม่เดลกหมายได้แค่ -1 หรือ 'BAD' เท่านั้น



CALCULATE THE HINGE LOSS

Hinge Loss เป็นฟังก์ชันการสูญเสีย (Loss Function)

Hinge Loss มีประสาทภาพในการจัดการกับปัญหา binary classification

สูตรของ Hinge Loss:

$$\text{Hinge Loss}(y, h(x)) = \max(0, 1 - y * h(x))$$

- y : (True Label) เป็นค่า 1 หรือ -1
- $h(x)$: (Predicted Score)

Hinge Loss จะ penalize โมเดล เมื่อ: **Hinge Loss จะไม่ penalize โมเดล เมื่อ:**

- โมเดลทำนายผิด ($y * f(x) < 0$)
- โมเดลทำนายถูก และคะแนนไม่มากพอ ($y * f(x) < 1$)
- โมเดลทำนายถูก และคะแนนมากพอ ($y * f(x) \geq 1$)

SUMMARY

เปรียบเทียบค่าของ hinge loss function ระหว่าง V1.0.0 และ V2.0.0

	V1.0.0	V2.0.0
Accuracy		
All Features	92.20%	71.28%
Selected Features	90.40%	72.70%

SUMMARY

เปรียบเทียบค่าของ hinge loss function ระหว่าง V1.0.0 และ V2.0.0

	V1.0.0	V2.0.0	V1.0.0	V2.0.0
	Loss (Test)		Loss(Future)	
All Features	0.156	0.574	0.400	0.960
Selected Features	0.191	0.546	0.400	0.960

Hinge Loss พยายามผลักดันโมเดลให้กำหนดค่าแบบที่ "ห่าง" จาก 0 มากที่สุด



Thank you for
your attention!

