



와인 품질 분류 및 예측

데이터 마이닝 12조


김재영, 박성환, 이종기, 이찬우

2022.05.25.

CONTENTS

- 01 Overview(개요)
- 02 Data(데이터)
- 03 Method(방법론)
- 04 Conclusion(결론)



A glass of red wine is shown on the left side of the slide, partially filled with a deep red liquid. The background is a soft, out-of-focus bokeh of warm, golden-yellow and orange lights, suggesting a festive or intimate setting. A single, delicate pink petal is visible in the lower right quadrant, adding a touch of elegance. The overall color palette is warm and romantic, with a diagonal pink band running across the middle of the slide.

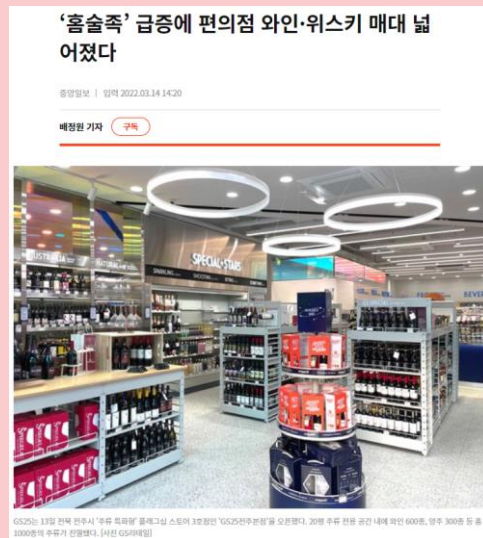
Data tell truth.

01 Overview(개요)

1. 주제 선정 이유
2. 문제 정의
3. 가설 및 목표

1. 주제 선정 이유

- 국내 와인 수입액 : 2018년 3118억 ➡ 2021년 7154억
- 지난 2년간 코로나로 인한 '홈술족'이 늘며 와인 판매량 증가
- 주류 스마트오더 시스템으로 인한 수요 증가
- 종류가 다양하여 입맛에 맞는 와인을 고르기 어려운데,
고르기 쉽게 도와주는 구독 서비스 생겨남



2. 문제 정의

기존에는 미각, 후각으로 측정하던 와인의 품질을
와인의 화학 측정 데이터로부터 추정한다.



3. 가설 및 목표

- 가설 수립

산성도, 알코올 도수 등 정량적으로 측정되는 화학데이터를 통해 특징 데이터를 구성하여 미각, 후각 측정 없이 와인의 품질을 추정할 수 있다.

- 목표

정량적으로 측정되는 화학데이터를 통해 와인의 품질이 "좋은지" "나쁜지" 예측할 수 있는 분류 모델을 SVM 알고리즘을 이용하여 모델링하고 정확도(Accuracy)로 성능을 측정한다.

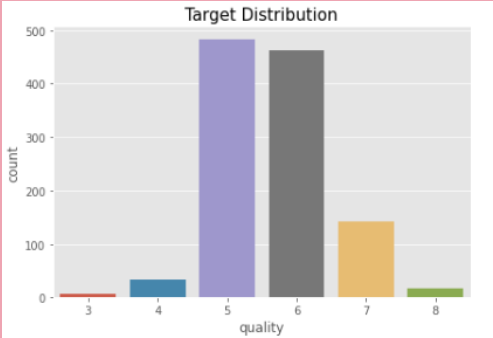




Data tell truth.

02 Data(데이터)

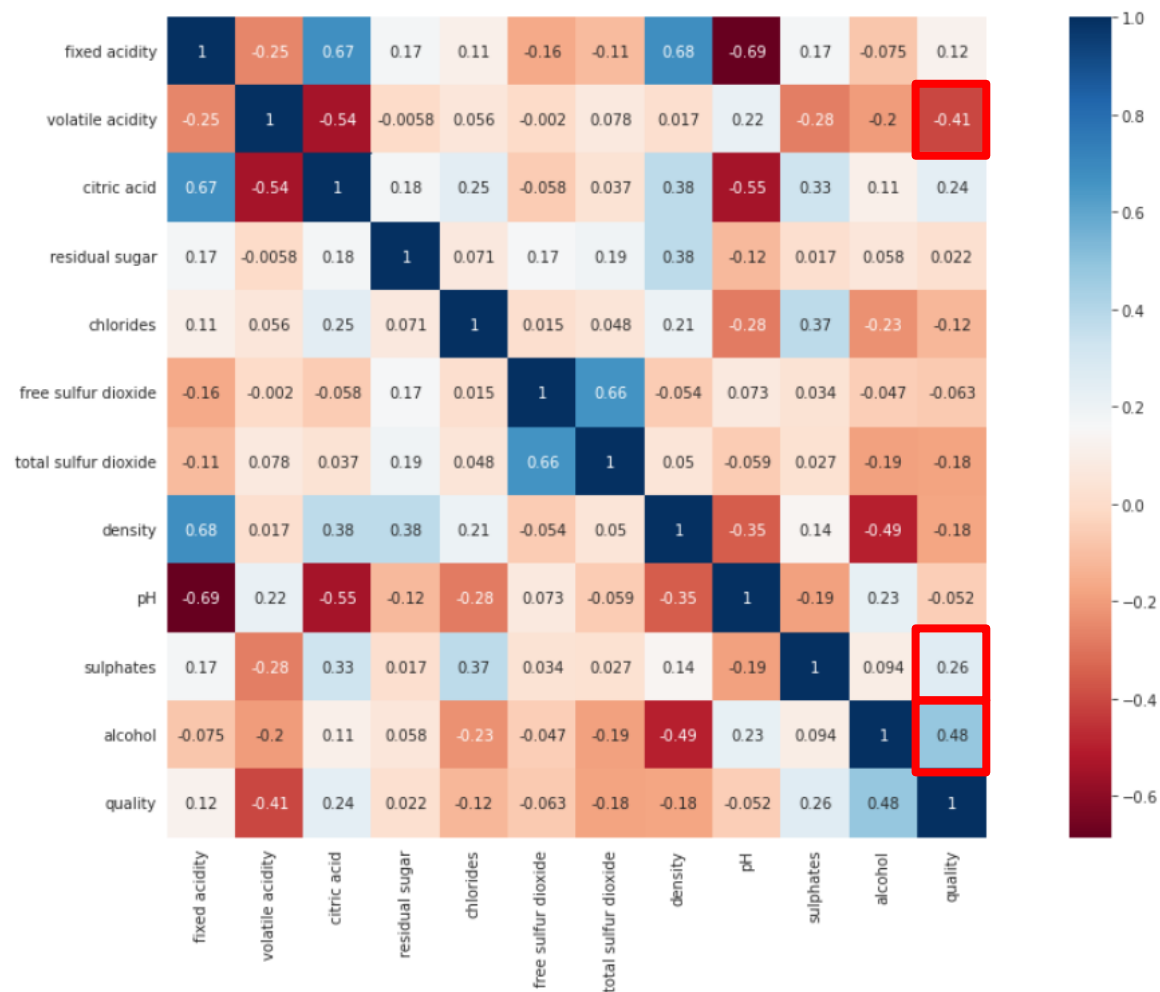
1. Info(정보)
2. EDA(탐색적 자료 분석)
3. PreProcessing(전처리)

항목	내용			
데이터 출처	Kaggle wine-quality-dataset			
행 개수	1143			
열 개수	12			
컬럼명	fixed acidity	residual sugar	total sulfur dioxide	sulphates
	volatile acidity	chlorides	density	alcohol
	citric acid	free sulfur dioxide	pH	quality
타겟 컬럼	quality			
타겟 컬럼 분포도	<div>3: 6건 4: 33건 5: 483건 6: 462건 7: 143건 8: 16건</div> <div></div>			
훈련/평가 데이터셋	Train : 70%(800건) Test : 30%(343건)			

2. EDA(1/4)



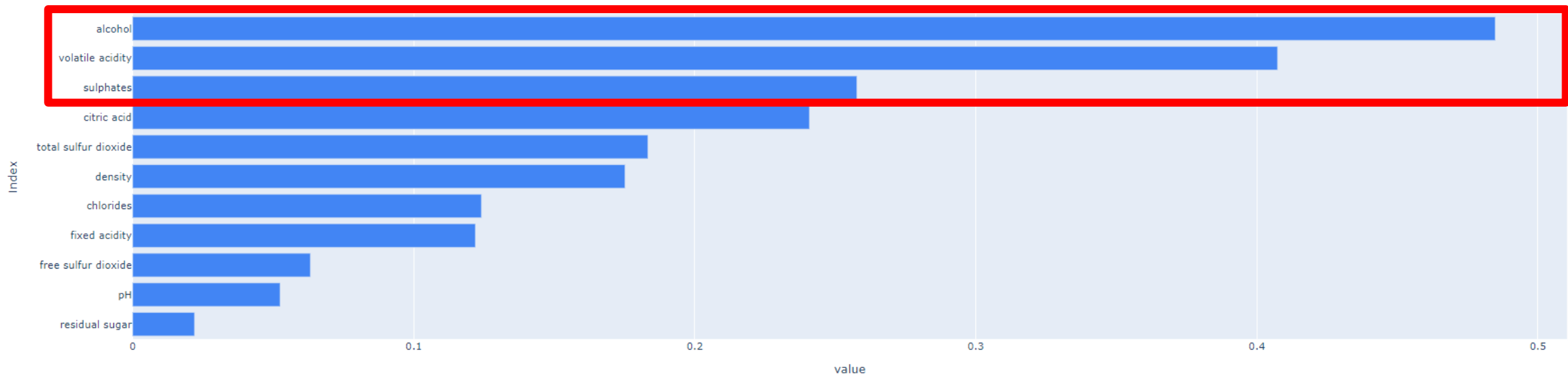
✓ Correlation Matrix(상관계수) HeatMap



2. EDA(2/4)



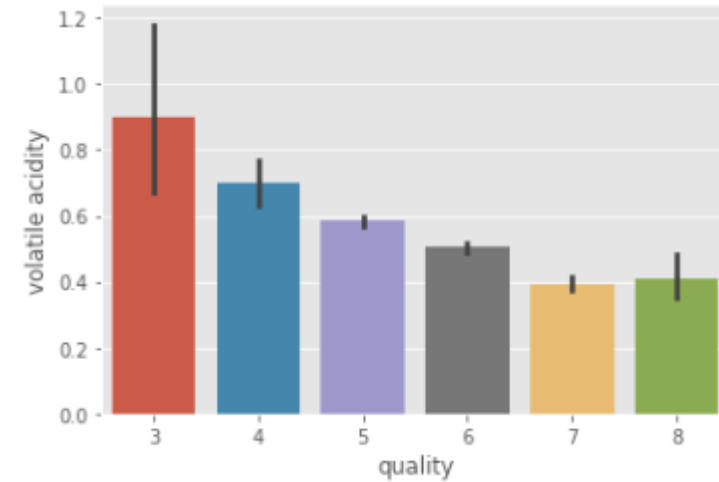
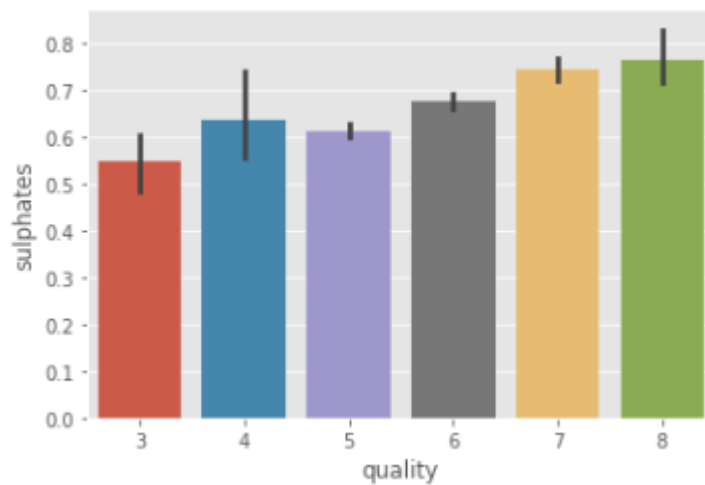
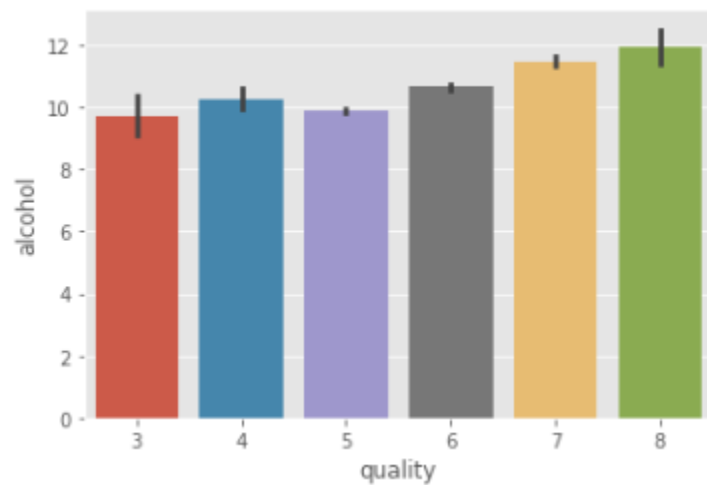
✓ Correlation Matrix(상관계수) Bar



2. EDA(3/4)



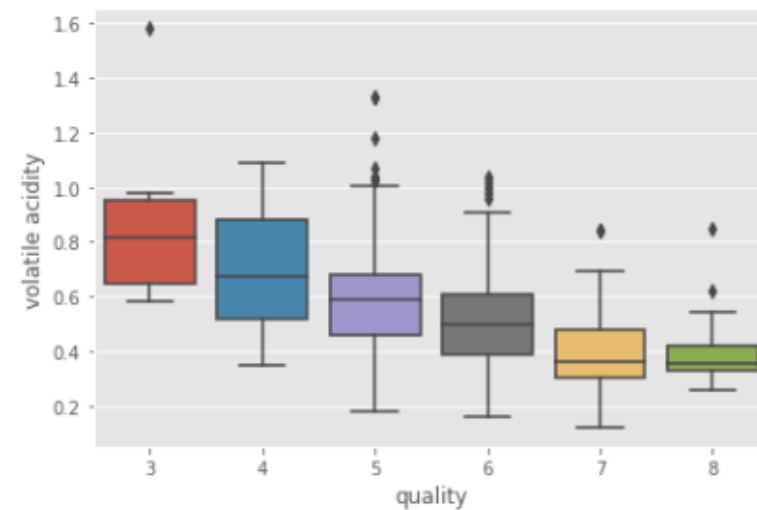
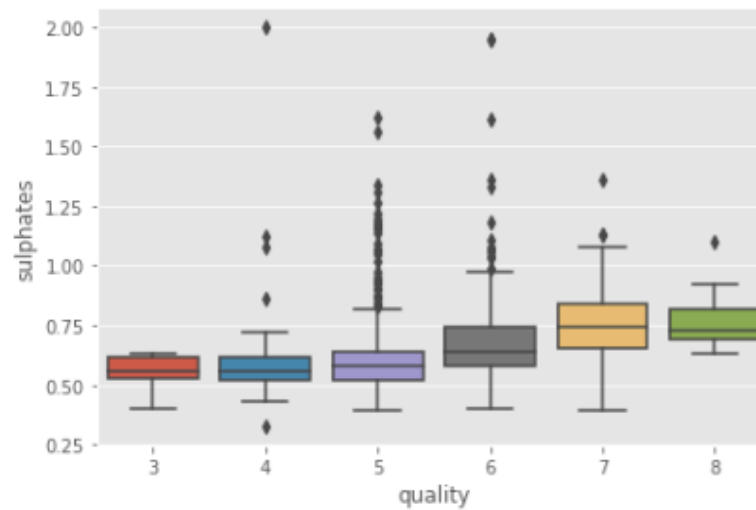
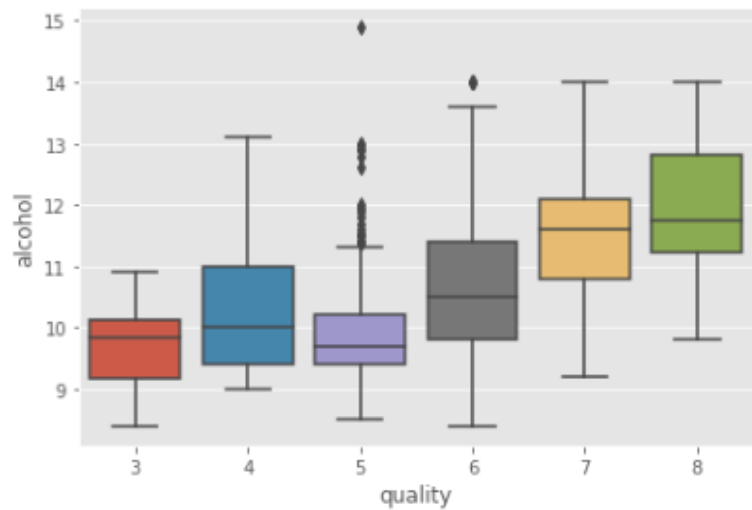
✓ 상관관계 상위 변수 별 수치 분포도



2. EDA(4/4)



✓ 상관관계 상위 변수 별 박스플롯



3. PreProcessing(1/2)



✓ Categorize

```
5    483
6    462
7    143
4     33
8     16
3      6
Name: quality, dtype: int64
```



6.5 기준으로 분리

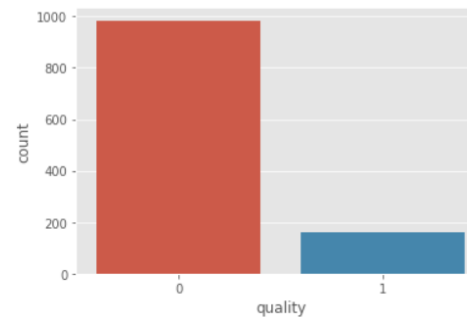
```
bad    984
good   159
Name: quality, dtype: int64
```

✓ Label Encoder

```
bad    984
good   159
Name: quality, dtype: int64
```



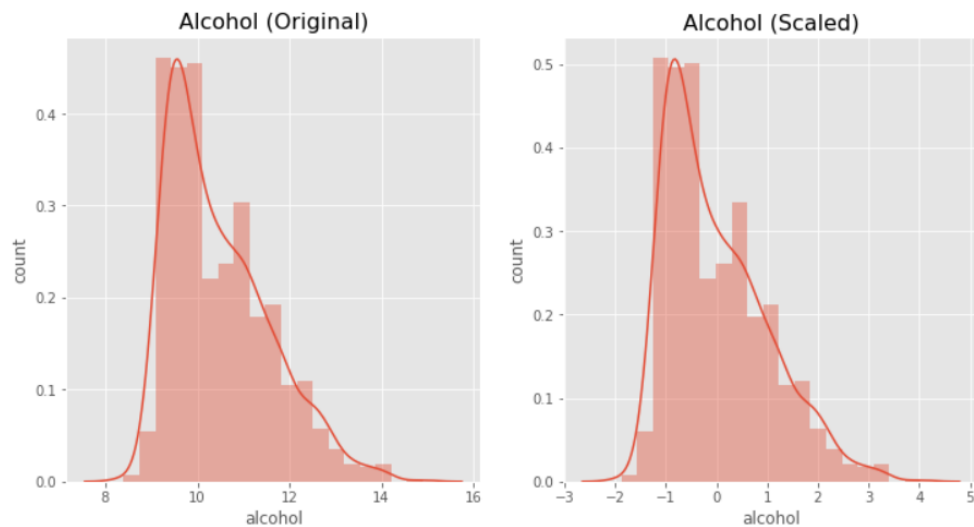
Label Encoder 실행하여
bad = 0 , good = 1 로 변경



3. PreProcessing(2/2)



✓ StandardScaler



✓ MinMaxScaler

- 최대값이 1, 최소값이 0이 되도록 스케일링



Data tell truth.

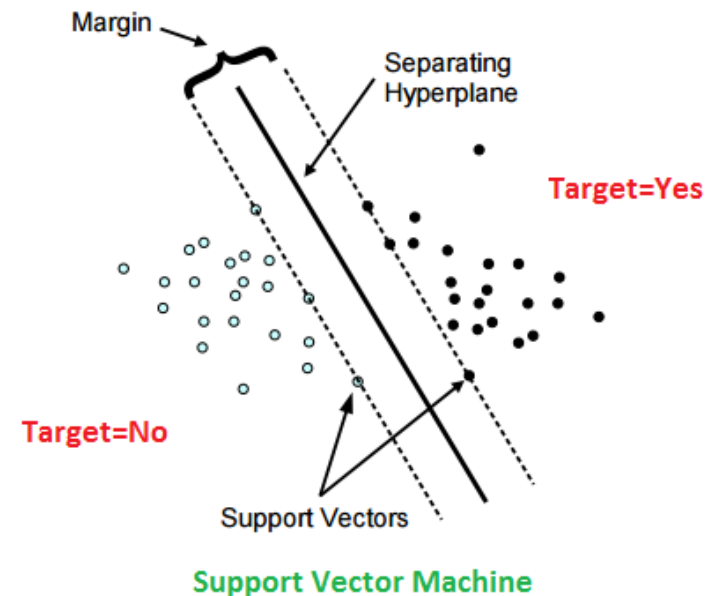
03 Method(방법론)

1. SVM이란?
2. 핵심 요소
3. 구현 내용
4. Library VS Coding

1. SVM이란?

- ✓ Support Vector Machine의 약자
- ✓ 결정 경계(Decision Boundary)를 정의하는 모델
 - 결정 경계 : 분류를 위한 기준 선
 - 마진이 가장 큰 결정 경계를 찾는 모델
 - 초평면(Hyperplane) : 속성이 3개 초과인 경우의 결정 경계
- ✓ Support Vectors는 결정 경계와 가까이 있는 데이터 포인트들을 의미
- ✓ 기본 수식

$$\sum_i \alpha_i k(x_i, x) = constant.$$



2. 핵심 요소



1) 마진(Margin)

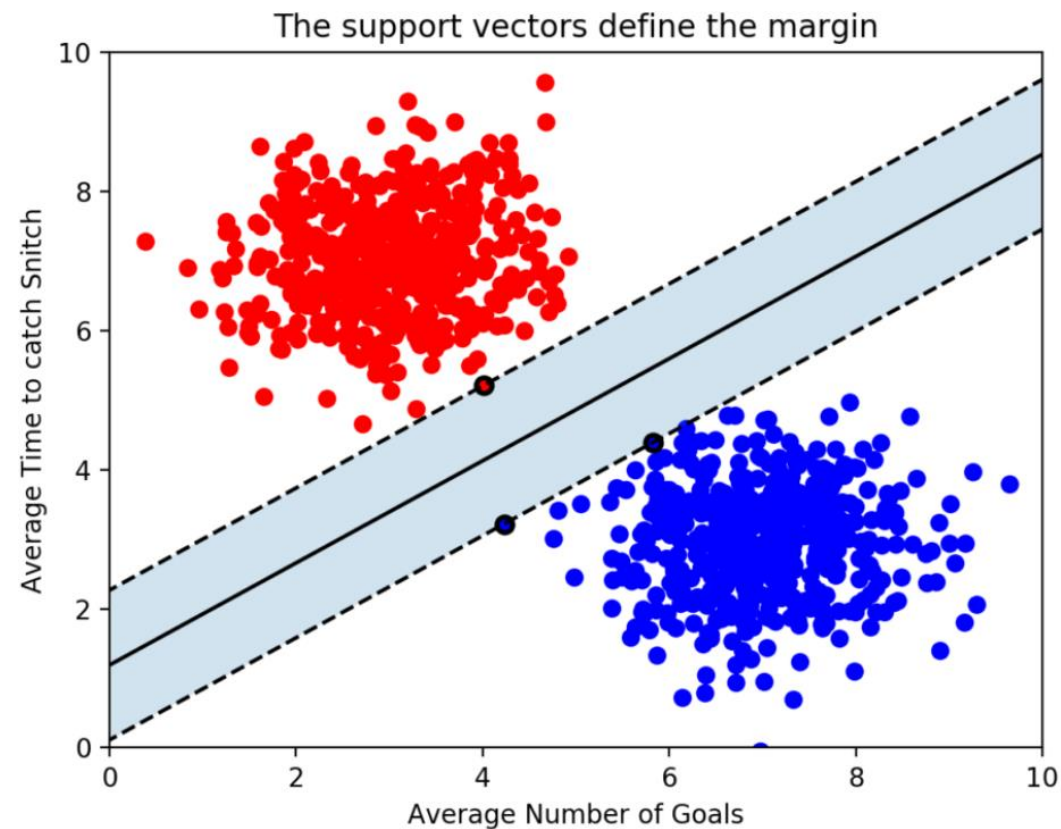
- 결정 경계와 서포트 벡터 사이의 거리
- n 개의 속성을 가진 데이터는 최소 $n+1$ 개의 서포트 벡터가 존재

2) 커널(Kernel)

- 선형으로 분류할 수 없는 경우, 고차원으로 변환하여 초평면으로 분류하기 위해 사용
 - 동차다항식(Homogeneous polynomial)
 - 다항식 커널(Polynomial kernel)
 - 가우시안 방사 기저 함수(Gaussian radial basis function)
 - 쌍곡탄젠트(Hyperbolic tangent)

2. 핵심 요소 - 마진(Margin) (1/4)

- ✓ 마진(Margin)
 - 하드 마진
 - 소프트 마진

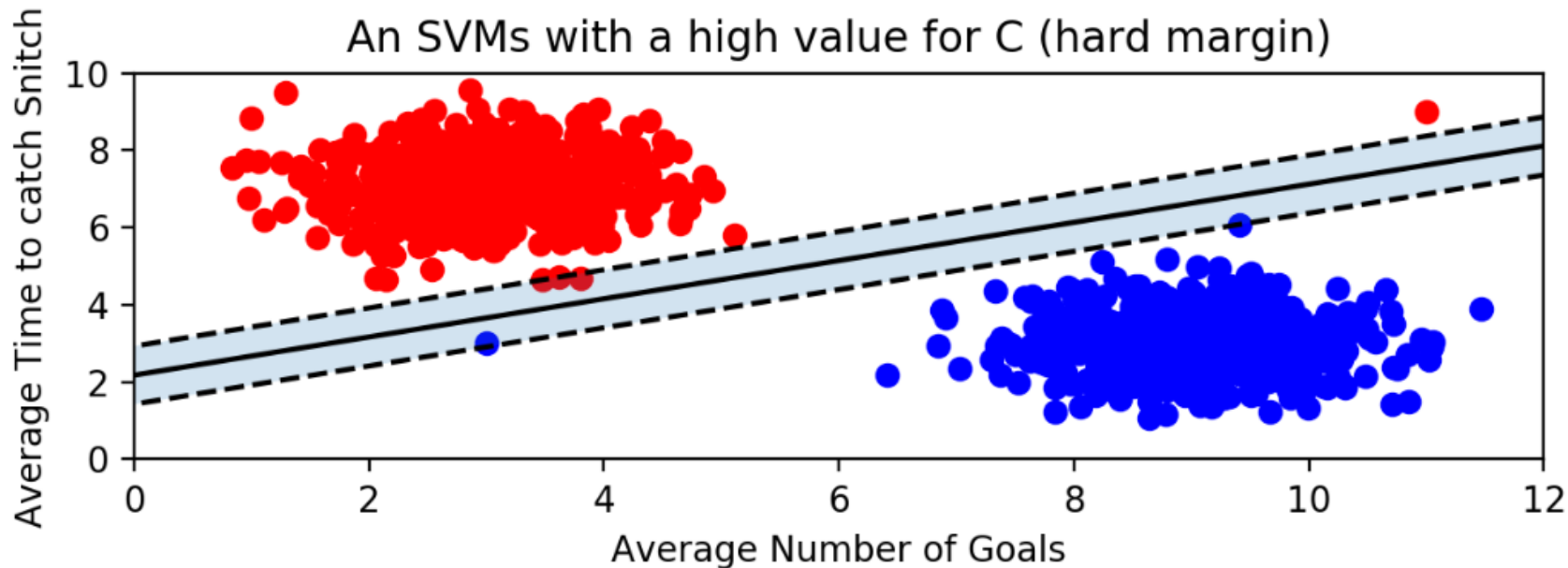


2. 핵심 요소 - 마진(Margin) (2/4)



✓ 하드 마진

- 이상치(Outlier)를 허용하지 않는 기준
- 서포트 벡터와 결정 경계 사이의 거리가 가까움 (마진이 작음)
- 과적합(Overfitting)이 발생할 수 있음

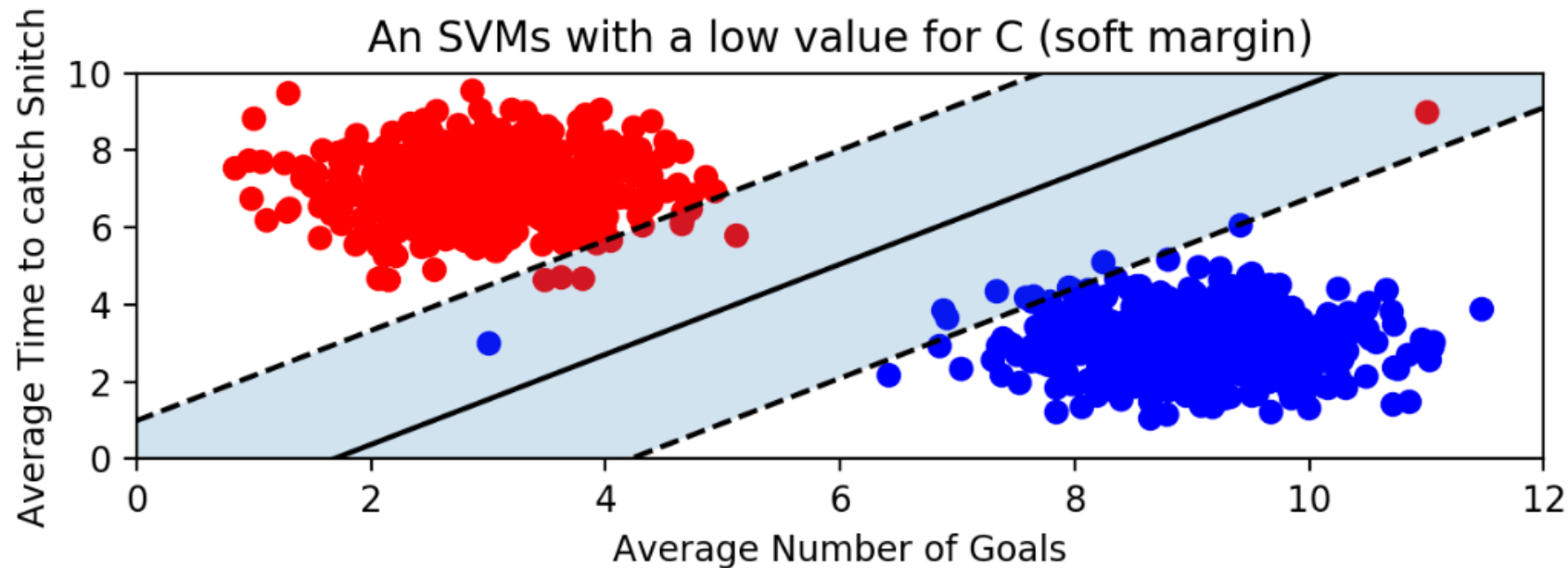


2. 핵심 요소 - 마진(Margin) (3/4)



✓ 소프트 마진

- 이상치(Outlier)가 마진 안에 어느정도 포함되도록 기준
- 서포트 벡터와 결정 경계 사이의 거리가 멀어짐 (마진이 큼)
- 과소적합(Underfitting)이 발생할 수 있음

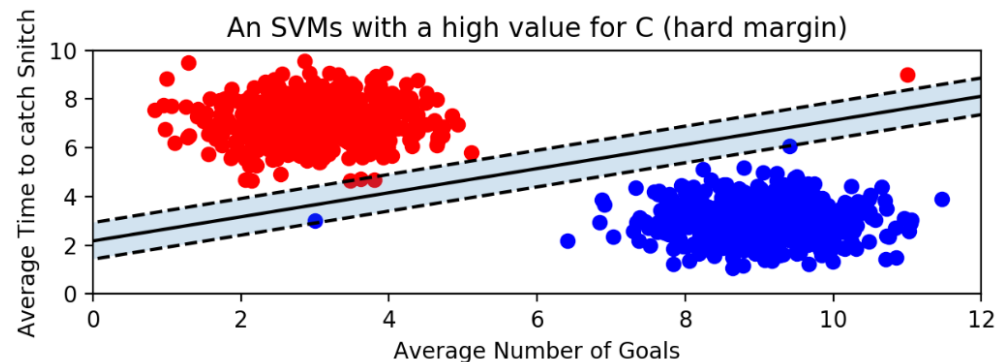


2. 핵심 요소 - 마진(Margin) (4/4)

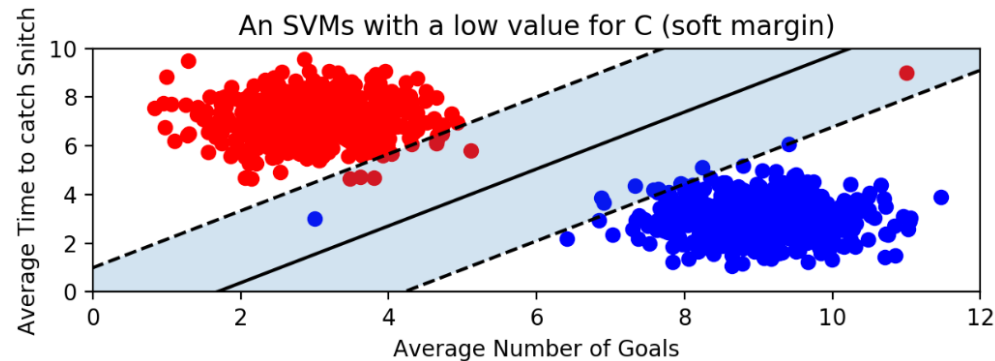


✓ 적용 파라미터 : C

- SVM 모델 오류 허용 범위 설정
- 클수록 하드마진(오류 허용 안 함)

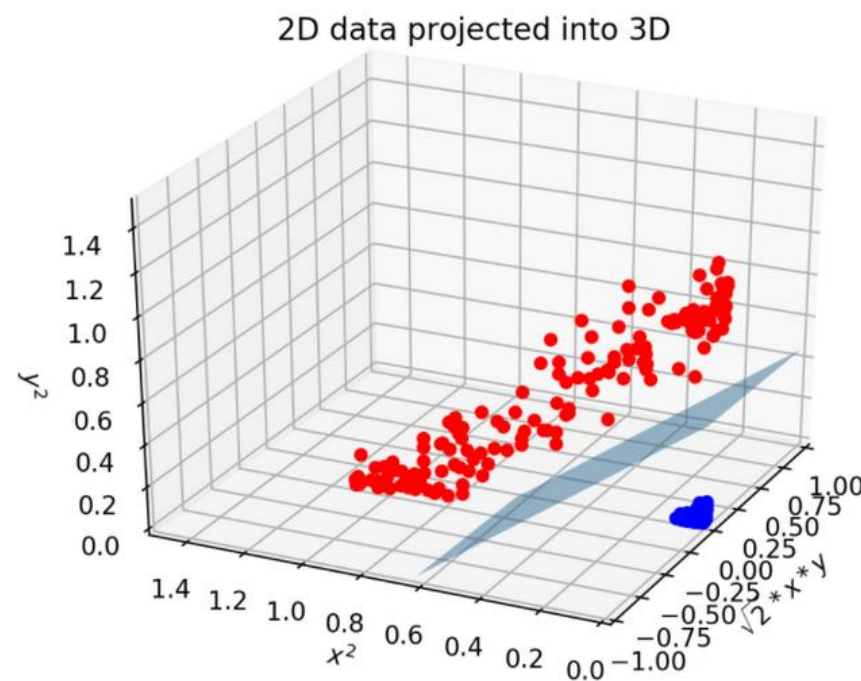
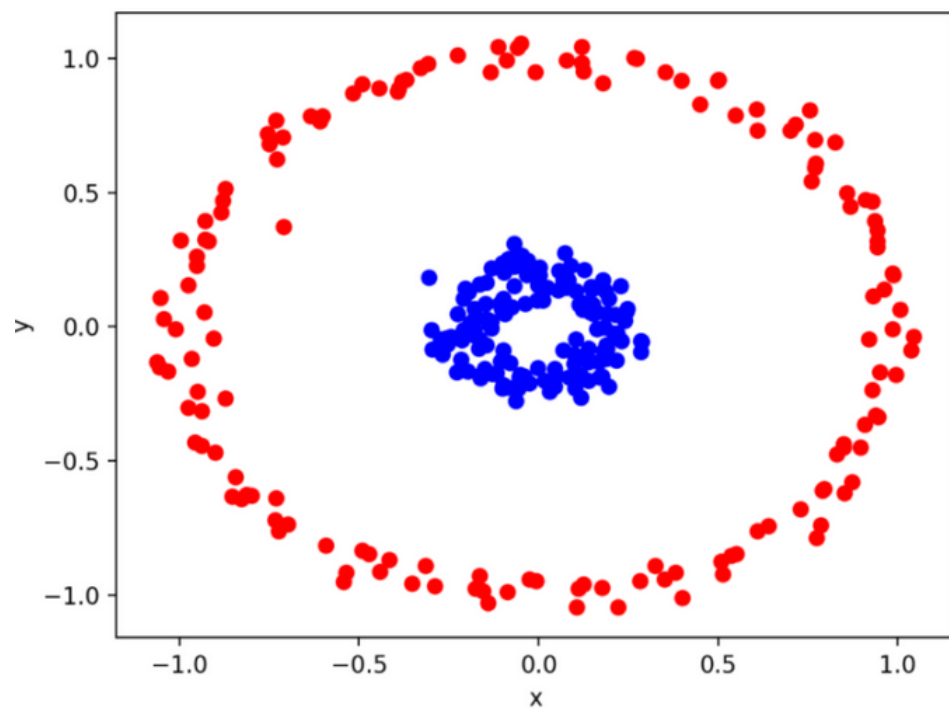


- 작을수록 소프트마진(오류 허용)



2. 핵심 요소 - 커널(Kernel) (1/3)

✓ 커널 (Kernel)



2. 핵심 요소 - 커널(Kernel) (2/3)



✓ 커널 함수 종류

- 동차다항식(Homogeneous polynomial)
 - 수식 : $k(x_i, x_j) = (x_i \cdot x_j)^d$
- 다항식 커널(Polynomial kernel)
 - 수식 : $k(x_i, x_j) = (x_i \cdot x_j + 1)^d$
- 가우시안 방사 기저 함수(Gaussian radial basis function)
 - 수식 : $k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$ for $\gamma > 0$
- 쌍곡탄젠트(Hyperbolic tangent)
 - 수식 : $k(x_i, x_j) = \tanh(\kappa x_i \cdot x_j + c)$ for some $\kappa > 0$ and $c < 0$

2. 핵심 요소 - 커널(Kernel) (3/3)



✓ 적용 파라미터 : kernel, Gamma

➤ kernel

- 커널 종류를 선택
 - 동차다항식 : linear
 - 다항식커널 : poly
 - 가우시안 방사 기저 함수 : rbf
 - 쌍곡탄젠트 : sigmoid
- linear 이외의 커널은 Gamma값이 필요

➤ Gamma

- 학습 데이터 민감 반응 정도 설정
- 값이 높을 경우
 - 학습 데이터에 많이 의존
 - 결정 경계가 곡선의 형태
 - 과적합(Overfitting) 발생 가능
- 값이 낮은 경우
 - 학습 데이터에 많이 의존하지 않음
 - 결정 경계가 직선의 형태
 - 과소적합(Underfitting) 발생 가능

2. 핵심 요소 - C와 Gamma 이해 (1/1)



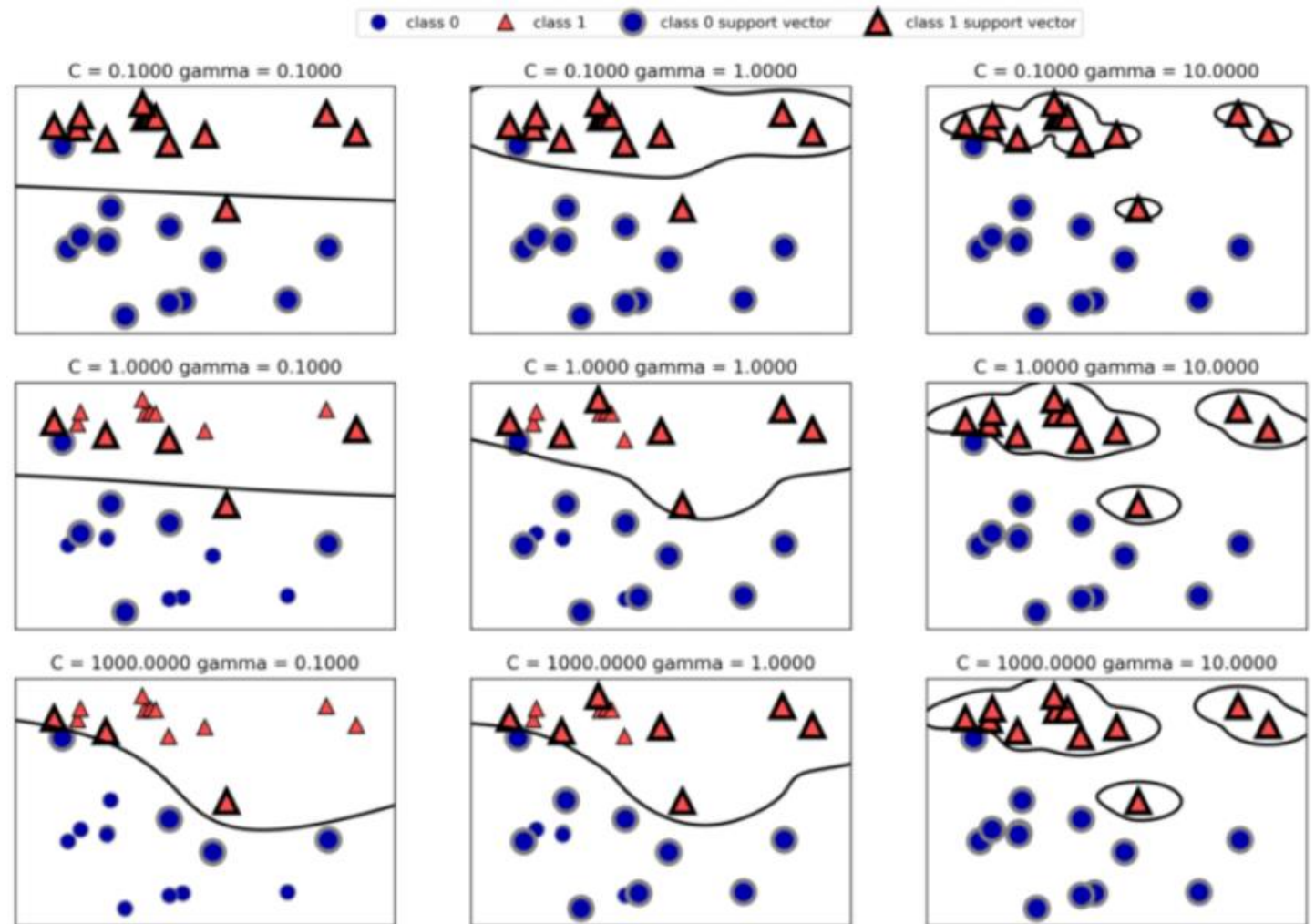
✓ C & Gamma

➤ C

- 클수록 하드마진
- 작을수록 소프트마진

➤ Gamma

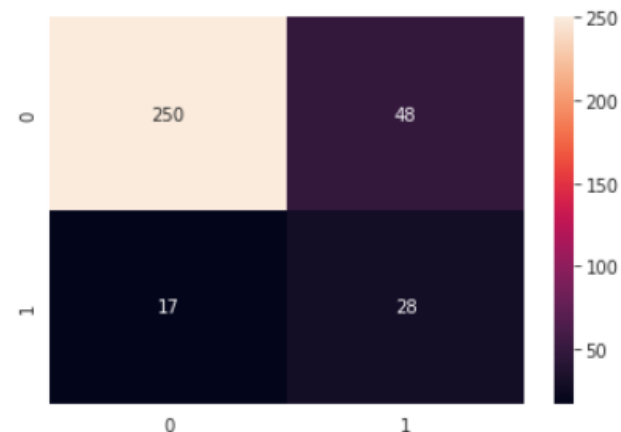
- 클수록 데이터 의존
- 작을수록 데이터 의존 적음



3. 구현 내용

✓ Case1) LinearSVC

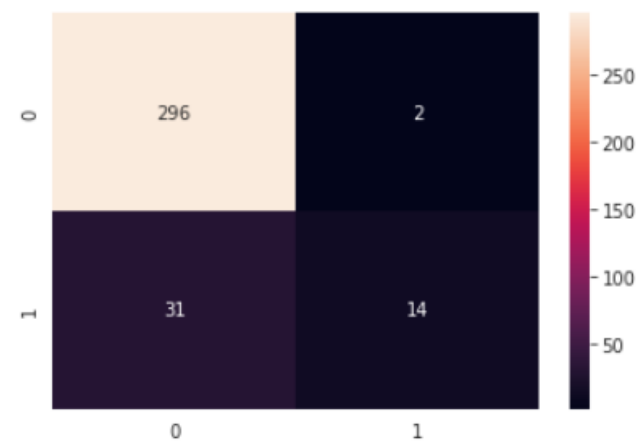
- confusion matrix : 정확도 - 278개
- 정확도 결과값 : 0.8104956268221575



case1

✓ Case2) SVC

- GridSearch 결과
 - 최적 : $C = 1$, kernel = 'rbf', gamma = 0.5
- confusion matrix : 정확도 - 310개
- 정확도 결과값 : 0.9037900874635568



case2



Data tell truth.

04 Conclusion(결론)

1. 타 모델 비교
2. 향후 과제

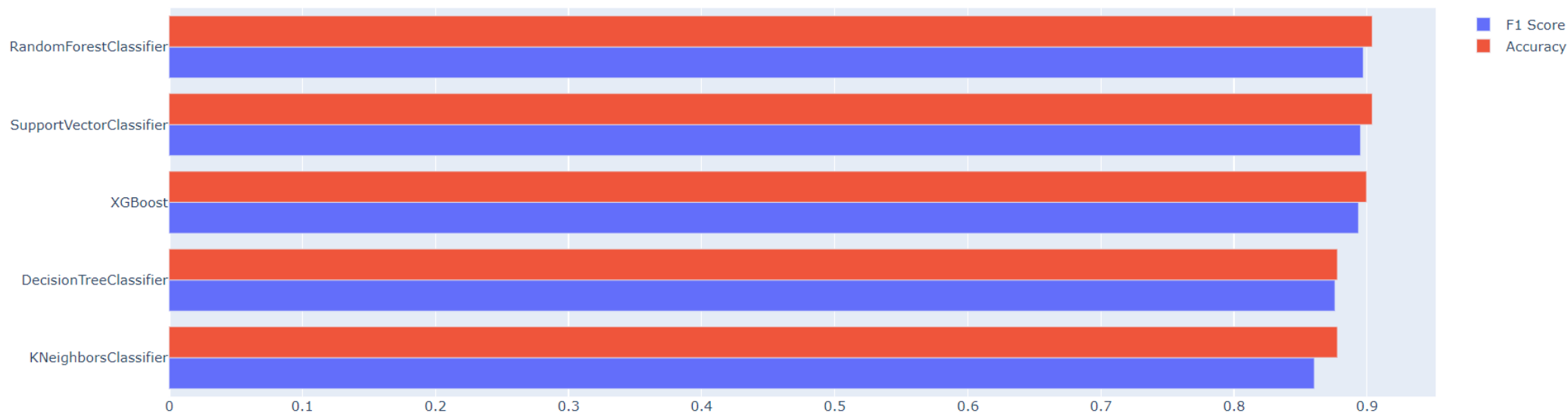
1. 타 모델 비교

✓ 적용 모델

- SVM(Support Vector Machine)
- Decision Tree
- KNN(K-Near Neighbors)
- Random Forest
- XGBoost

✓ 결과

	Accuracy	F1 Score
KNeighborsClassifier	0.877729	0.860344
DecisionTreeClassifier	0.877729	0.875850
XGBoost	0.899563	0.893534
SupportVectorClassifier	0.903930	0.895094
RandomForestClassifier	0.903930	0.897186



2. 향후 과제

- ✓ PCA(주성분분석)
 - 주성분분석을 통해 주성분을 추출하여 분류 및 예측 진행
- ✓ 타 모델 학습
 - Random Forest, XGBoost 등의 방법론에 대한 연구



A romantic dinner table setting featuring a wine glass filled with red wine, a bouquet of pink roses, and warm bokeh lights in the background. The scene is overlaid with a large, diagonal, semi-transparent pink shape that serves as a background for the text.

THANK
YOU