Wine Quality Prediction

4조

독하게 데이터 사이언스

김태영 류채환 박준영 신해솔 홍성희

Introduction

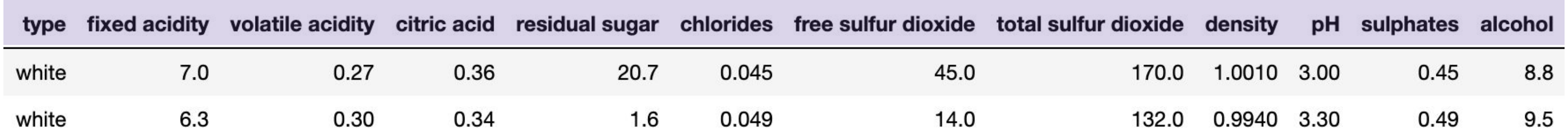
Data and Purpose

Problems of Prior Research

Our Improvement

분석 수행 목적: y = f(X)

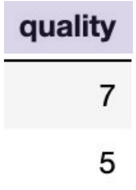
X : 와인의 화학적 성분 데이터로



type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
white	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8
white	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5

타입, 알코올, 밀도, pH, 산도, 구연산, 잔류 설탕, 염화물, 이산화황, …

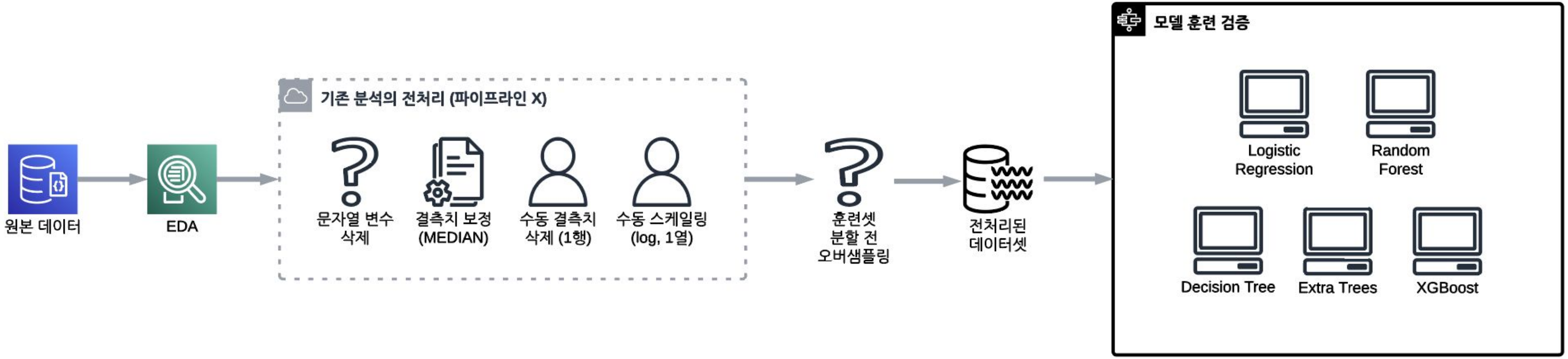
y : 와인의 품질 등급 맞추기



quality
7
5

품질(0~10점) → 실질적으로는 3~9의 값

선행 연구 Workflow

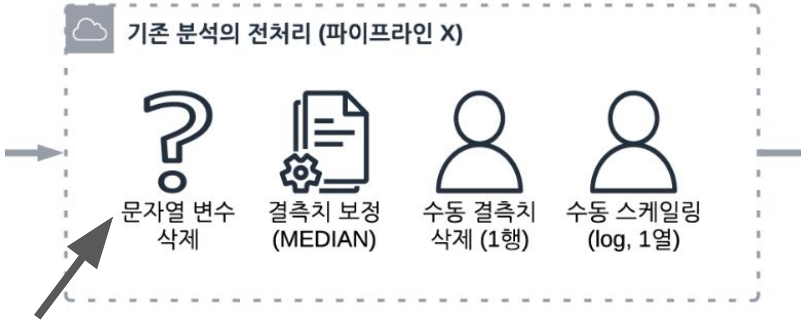


모델 평가 결과

최종모델선택: Extra Trees

Accuracy: 89.9

선행 연구의 문제점 1: 불완전한 전처리



기존 분석의 전처리 (파이프라인 X)

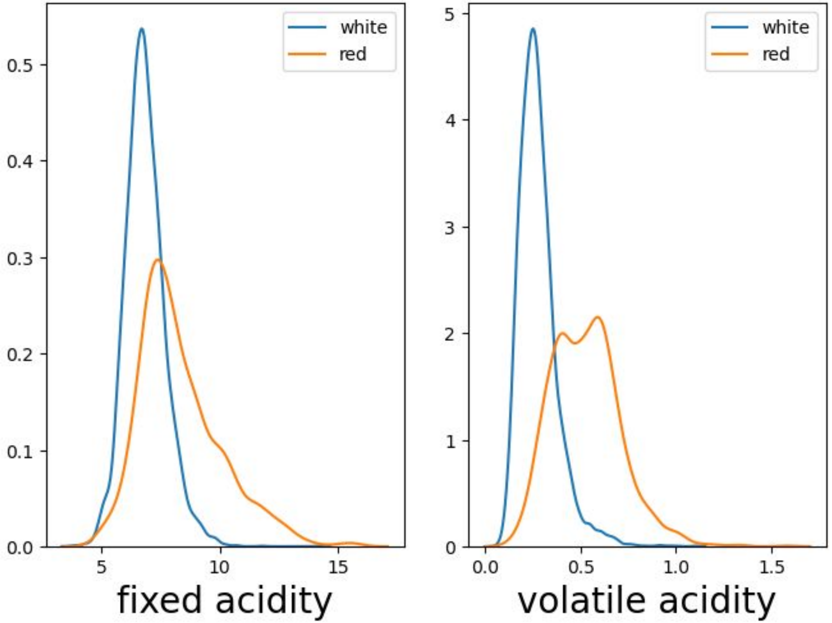
문자열 변수 삭제

결측치 보정 (MEDIAN)

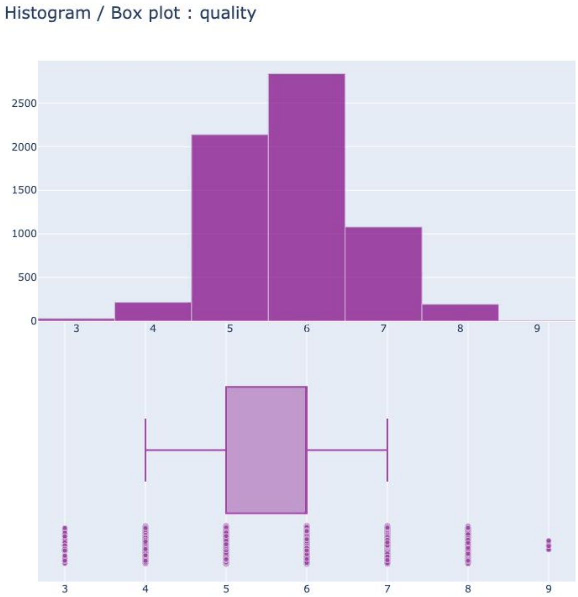
수동 결측치 삭제 (1행)

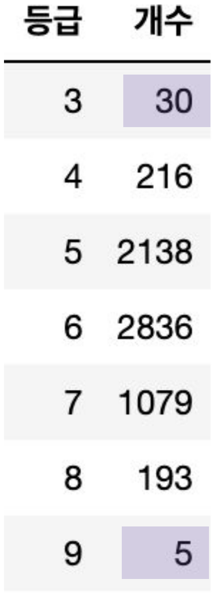
수동 스케일링 (log, 1열)

type (white, red)



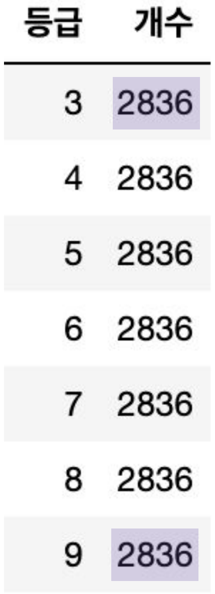
선행 연구의 문제점 2: 비합리적 오버샘플링





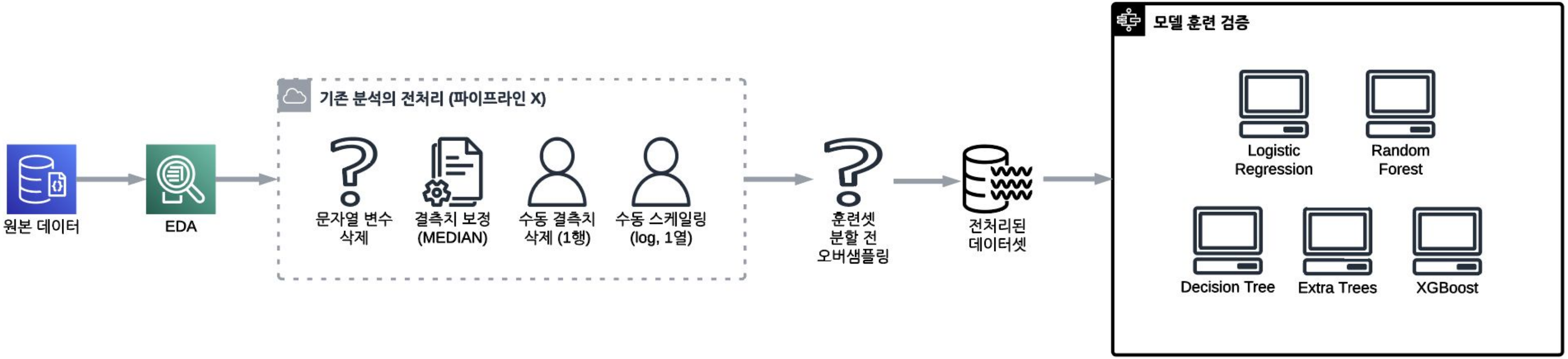
등급	개수
3	30
4	216
5	2138
6	2836
7	1079
8	193
9	5

훈련셋 분할 전
오버샘플링



등급	개수
3	2836
4	2836
5	2836
6	2836
7	2836
8	2836
9	2836

선행 연구 재평가



원본 데이터

EDA

기존 분석의 전처리 (파이프라인 X)

문자열 변수 삭제

결측치 보정 (MEDIAN)

수동 결측치 삭제 (1행)

수동 스케일링 (log, 1열)

훈련셋 분할 전 오버샘플링

전처리된 데이터셋

모델 훈련 검증

Logistic Regression

Random Forest

Decision Tree

Extra Trees

XGBoost

잘못된 선행 연구 평가 결과

최종모델선택: Extra Trees

Accuracy: 89.9

F1: 89.2

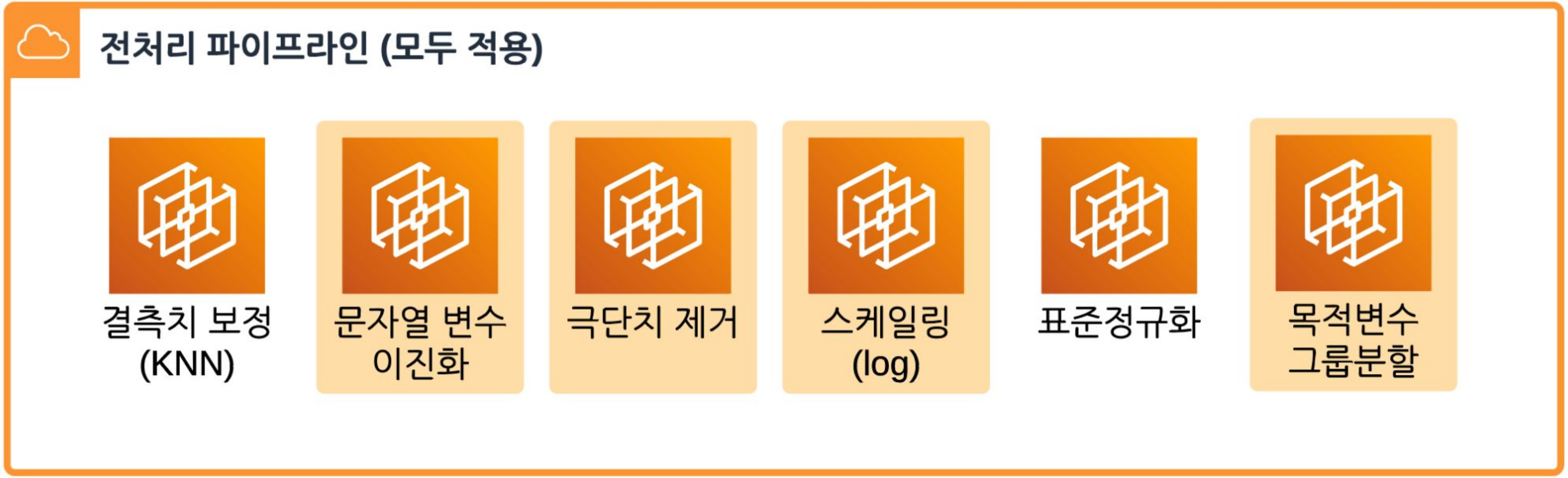
비합리적 가정 제거 후 선행 연구 평가 결과

최종모델선택: Extra Trees

Accuracy: 67.5

F1: 66.9

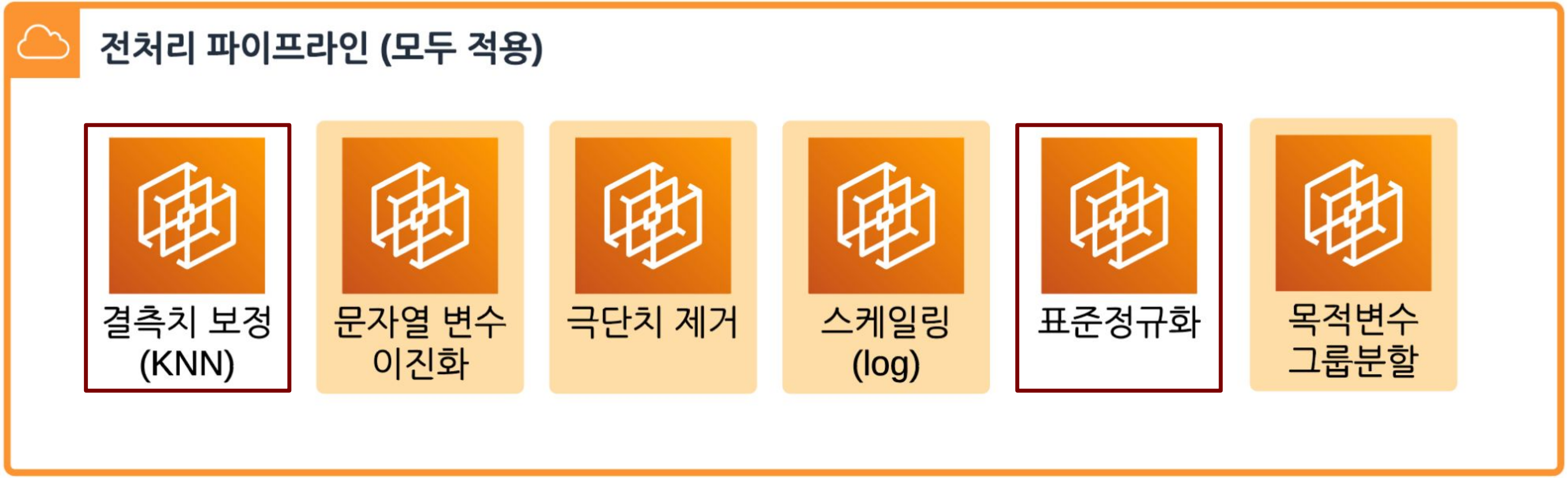
선행 연구 개선 1: 전처리



전처리 파이프라인 (모두 적용)

- 결측치 보정 (KNN)
- 문자열 변수 이진화
- 극단치 제거
- 스케일링 (log)
- 표준정규화
- 목적변수 그룹분할

선행 연구 개선 1: 전처리

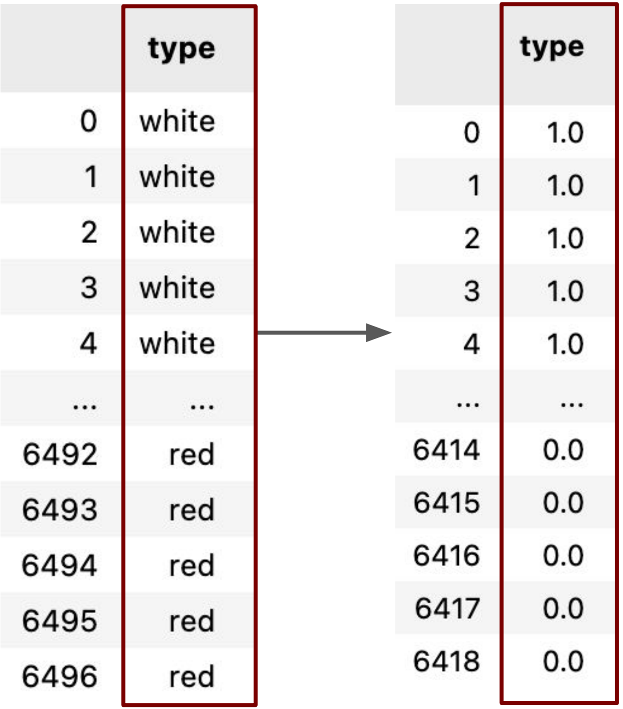


전처리 파이프라인 (모두 적용)

- 결측치 보정 (KNN)
- 문자열 변수 이진화
- 극단치 제거
- 스케일링 (log)
- 표준정규화
- 목적변수 그룹분할

선행 연구 개선 1: 전처리

문자열 변수 이진화

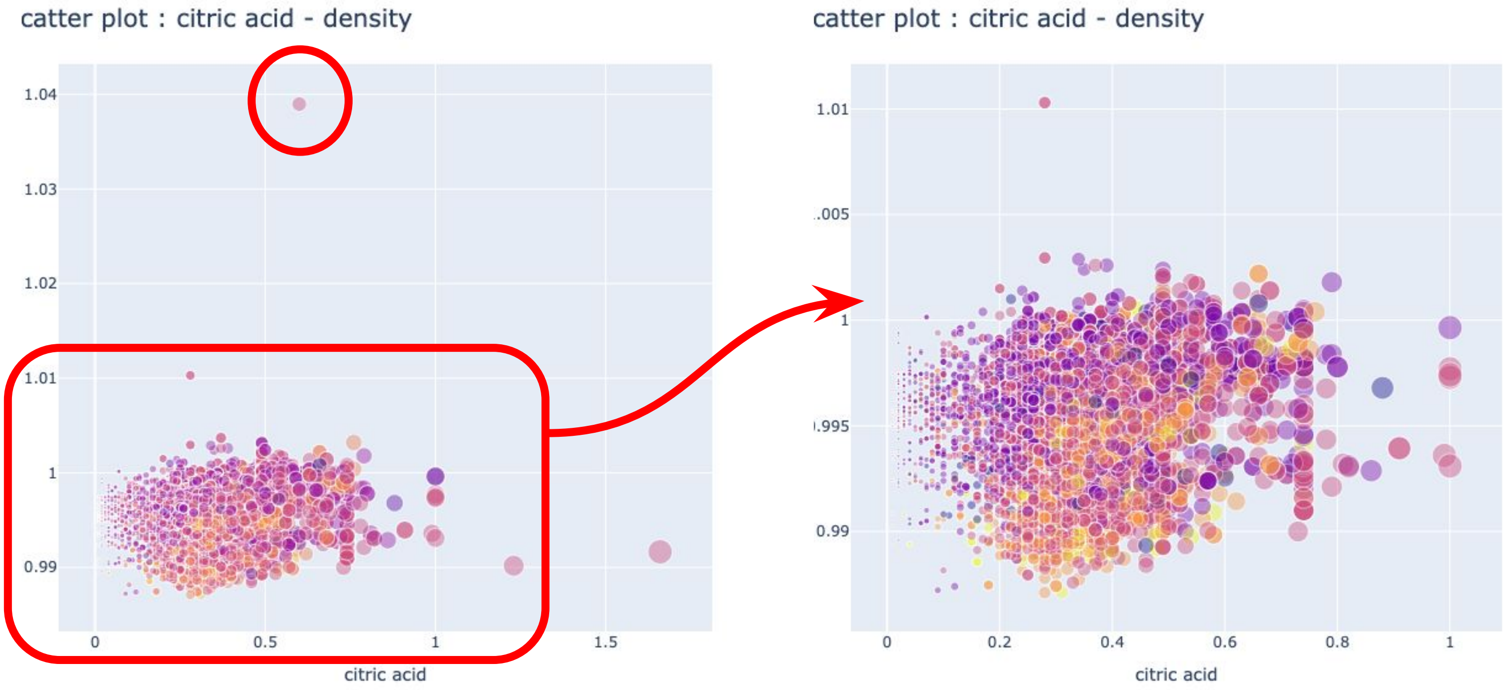


	type
0	white
1	white
2	white
3	white
4	white
...	...
6492	red
6493	red
6494	red
6495	red
6496	red

	type
0	1.0
1	1.0
2	1.0
3	1.0
4	1.0
...	...
6414	0.0
6415	0.0
6416	0.0
6417	0.0
6418	0.0

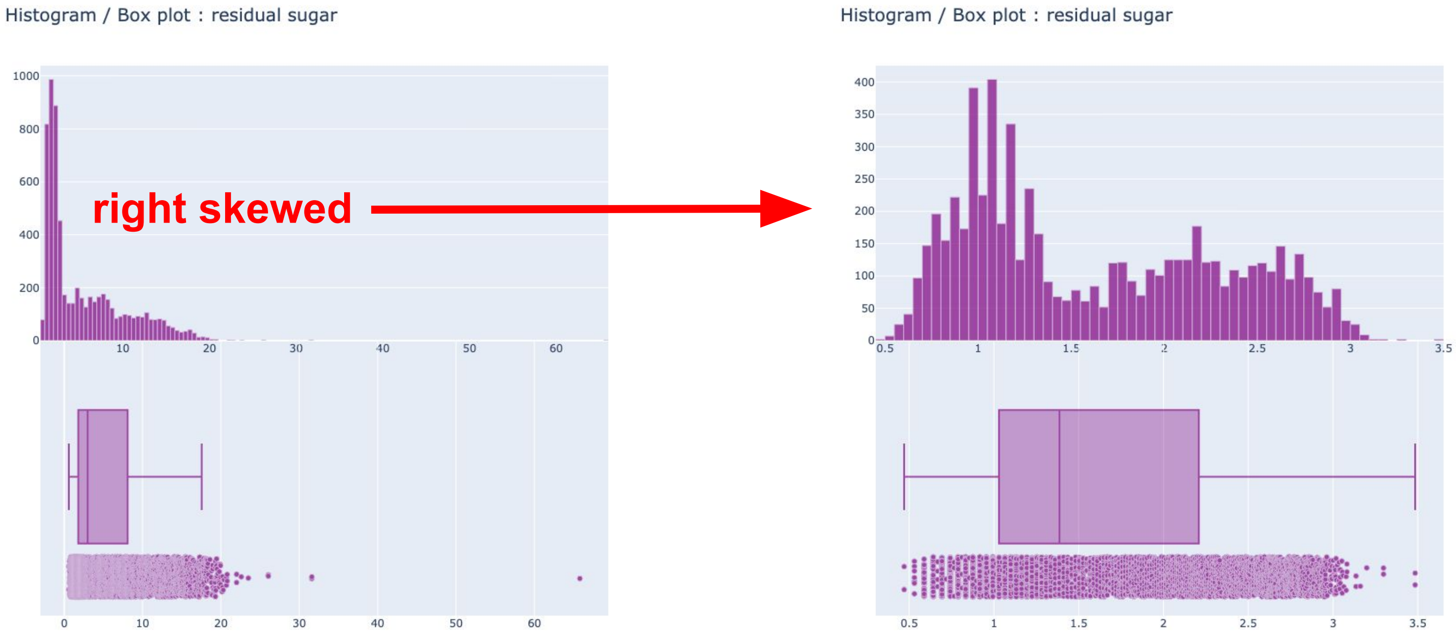
선행 연구 개선 1: 전처리

극단치 제거



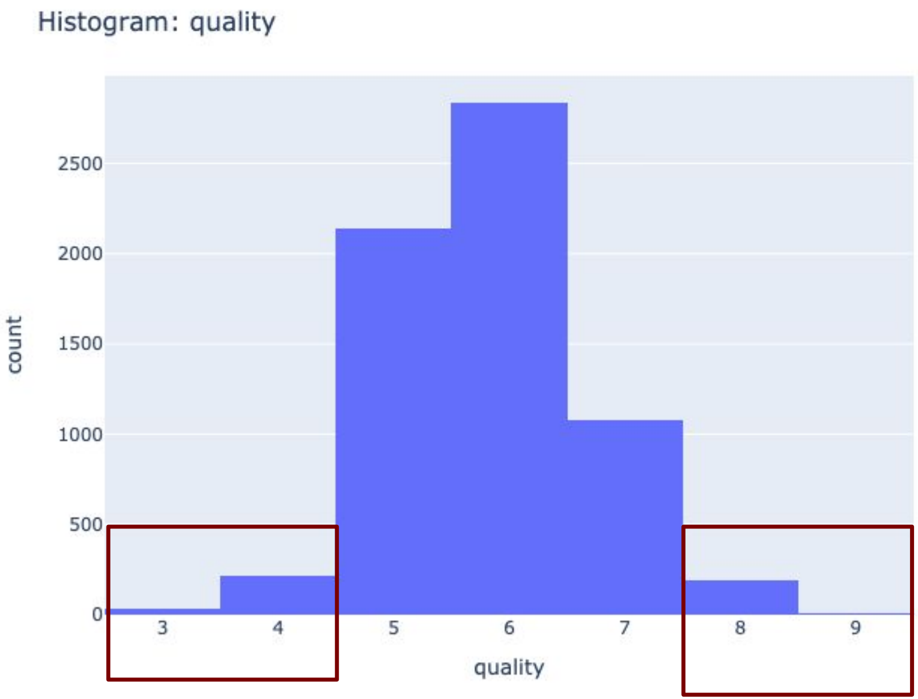
선행 연구 개선 1: 전처리

스케일링
(log)



선행 연구 개선 1: 전처리

목적변수
그룹분할

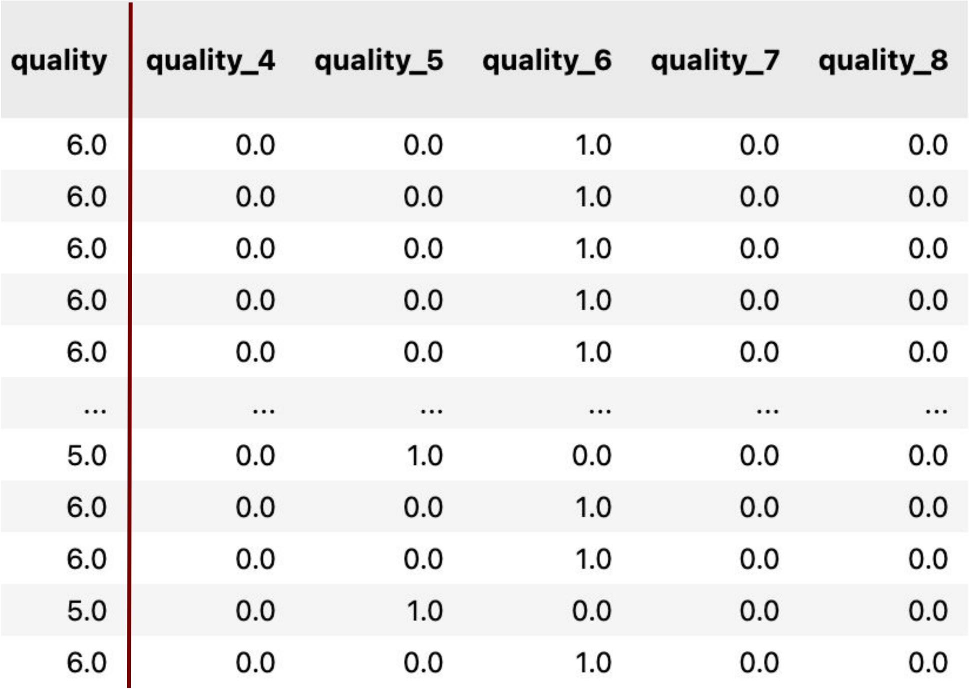


quality 3, 4
→ **quality 4**

quality 8, 9
→ **quality 8**

선행 연구 개선 1: 전처리

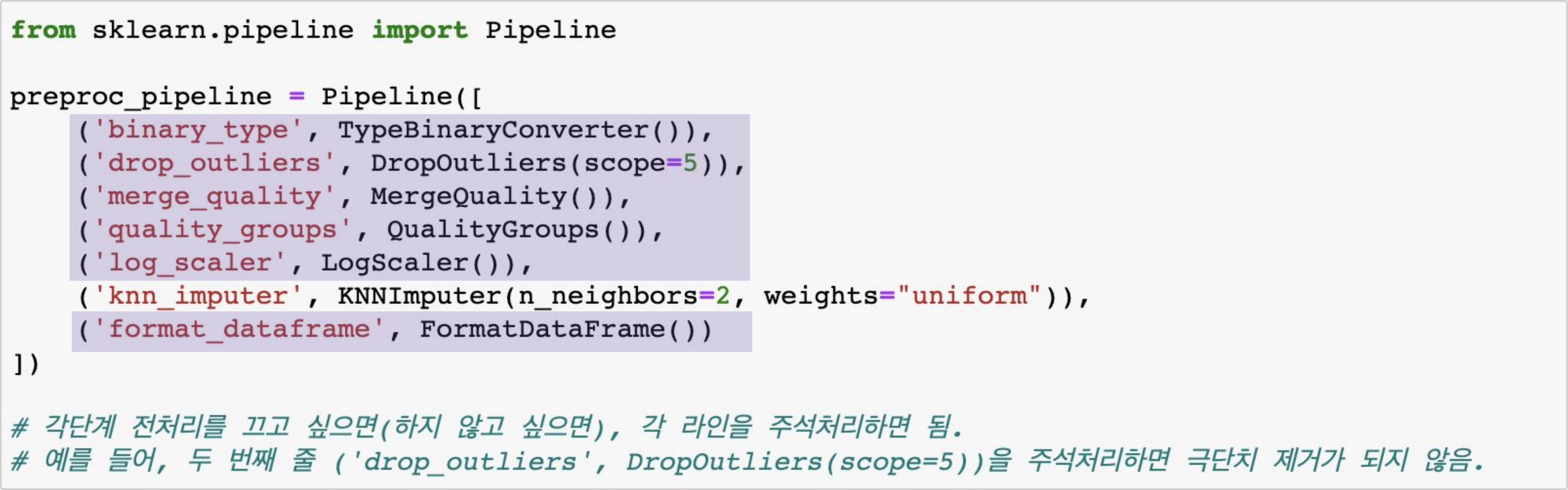
목적변수
그룹분할



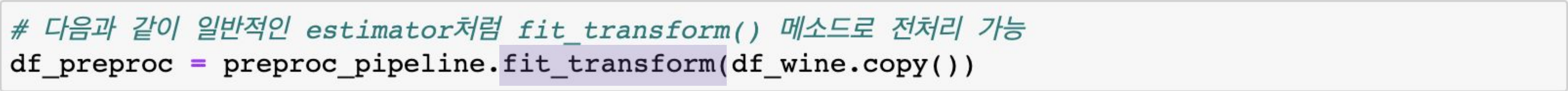
quality	quality_4	quality_5	quality_6	quality_7	quality_8
6.0	0.0	0.0	1.0	0.0	0.0
6.0	0.0	0.0	1.0	0.0	0.0
6.0	0.0	0.0	1.0	0.0	0.0
6.0	0.0	0.0	1.0	0.0	0.0
6.0	0.0	0.0	1.0	0.0	0.0
...	...	...	...	...	...
5.0	0.0	1.0	0.0	0.0	0.0
6.0	0.0	0.0	1.0	0.0	0.0
6.0	0.0	0.0	1.0	0.0	0.0
5.0	0.0	1.0	0.0	0.0	0.0
6.0	0.0	0.0	1.0	0.0	0.0

**quality 이진
처리**

전처리기: scikit-learn 방식 구현



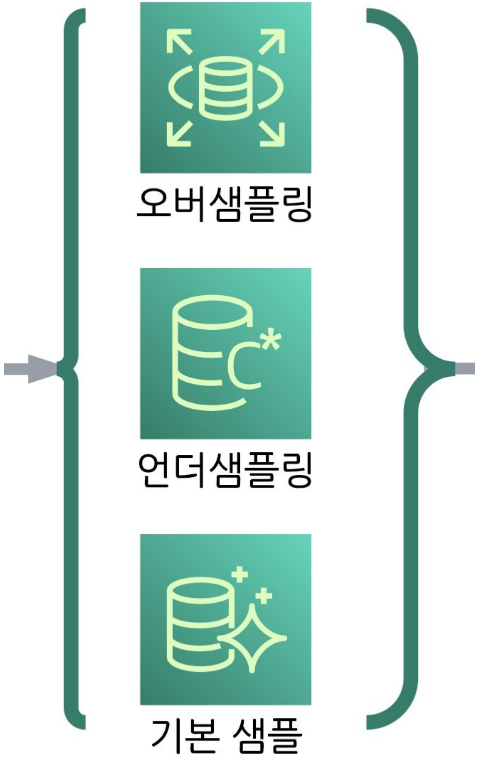
```python
from sklearn.pipeline import Pipeline

preproc_pipeline = Pipeline([
    ('binary_type', TypeBinaryConverter()),
    ('drop_outliers', DropOutliers(scope=5)),
    ('merge_quality', MergeQuality()),
    ('quality_groups', QualityGroups()),
    ('log_scaler', LogScaler()),
    ('knn_imputer', KNNImputer(n_neighbors=2, weights="uniform")),
    ('format_dataframe', FormatDataFrame())
])

# 각단계 전처리를 끄고 싶으면(하지 않고 싶으면), 각 라인을 주석처리하면 됨.
# 예를 들어, 두 번째 줄 ('drop_outliers', DropOutliers(scope=5))을 주석처리하면 극단치 제거가 되지 않음.
```



```python
# 다음과 같이 일반적인 estimator처럼 fit_transform() 메소드로 전처리 가능
df_preproc = preproc_pipeline.fit_transform(df_wine.copy())
```

ref) 오렐리앙 제롱, 핸즈온머신러닝 2판 “2.5.3 나만의 변환기” p106

선행 연구 개선 2: 오버샘플링



오버샘플링

언더샘플링

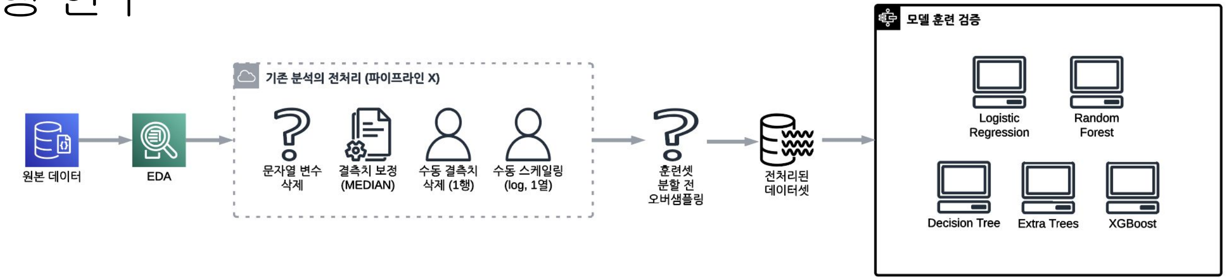
기본 샘플

1. Train set 분할 후
   Train set에 대해서만 오버샘플링

2. 오버샘플링/언더샘플링/기본샘플
   모델 분석 x3 수행
   (샘플링 방식에 따른 모델 학습 편향 방지)

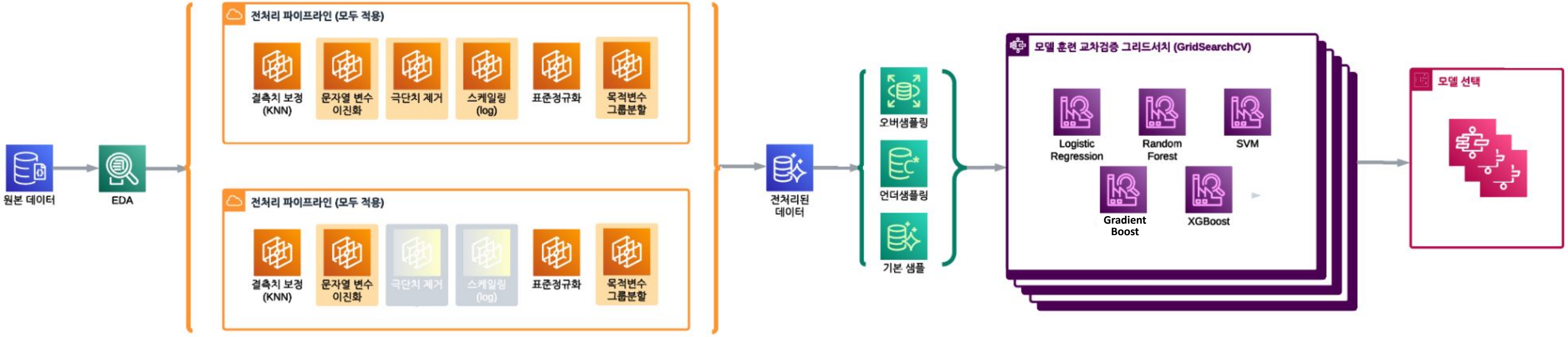
선행 연구와의 비교: Workflow Diagram

선행 연구



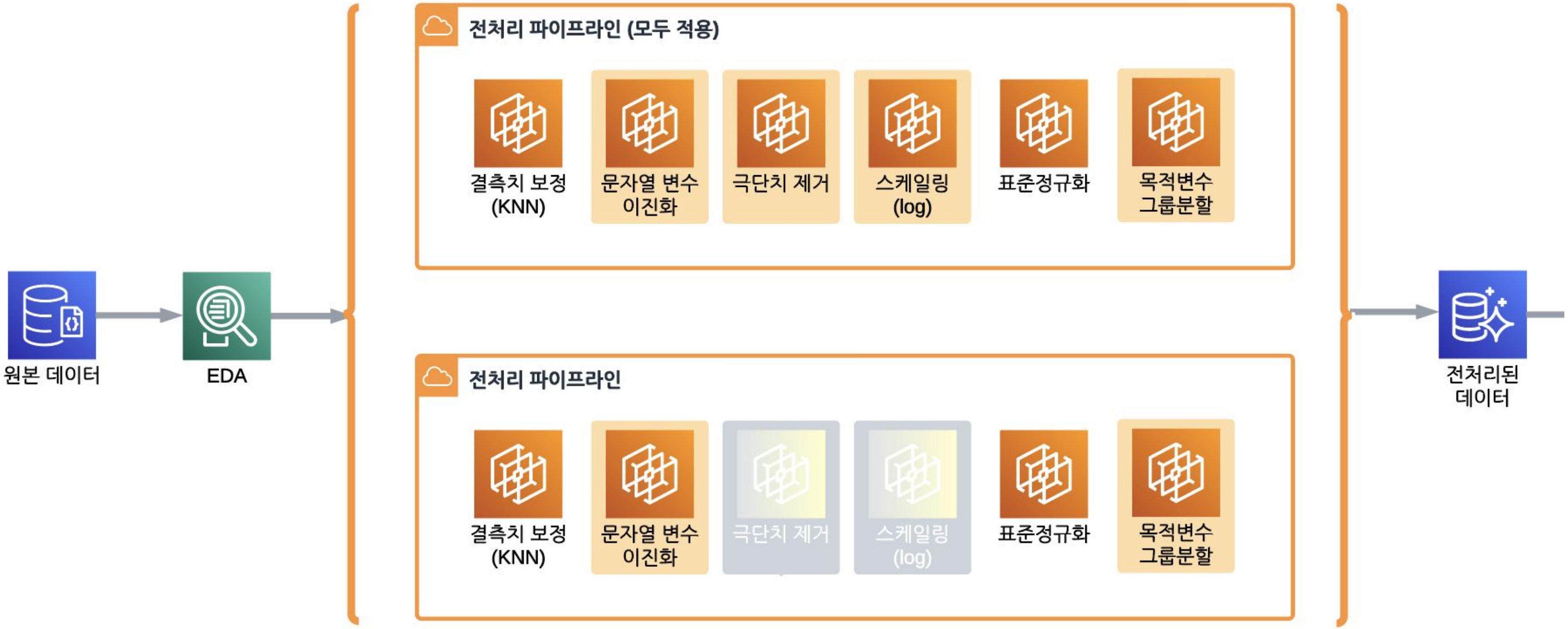
원본 데이터
EDA
기존 분석의 전처리 (파이프라인 X)
문자열 변수 삭제
결측치 보정 (MEDIAN)
수동 결측치 삭제 (1행)
수동 스케일링 (log, 1열)
훈련셋 분할 전 오버샘플링
전처리된 데이터셋
모델 훈련 검증
Logistic Regression
Random Forest
Decision Tree
Extra Trees
XGBoost

수행한 연구

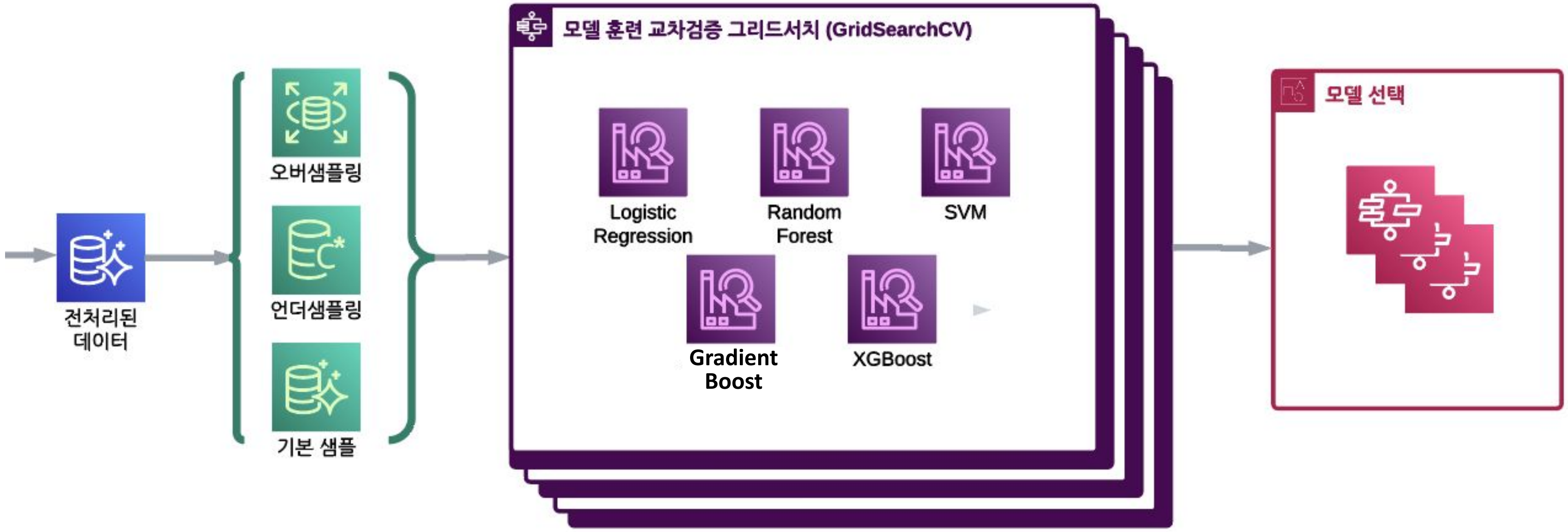


원본 데이터
EDA
전처리 파이프라인 (모두 적용)
결측치 보정 (KNN)
문자열 변수 이진화
극단치 제거
스케일링 (log)
표준정규화
목적변수 그룹분할
전처리 파이프라인 (모두 적용)
결측치 보정 (KNN)
문자열 변수 이진화
극단치 제거
스케일링 (log)
표준정규화
목적변수 그룹분할
전처리된 데이터
오버샘플링
언더샘플링
기본 샘플
모델 훈련 교차검증 그리드서치 (GridSearchCV)
Logistic Regression
Random Forest
SVM
Gradient Boost
XGBoost
모델 선택

수행한 분석 Workflow (1/2)



수행한 분석 Workflow (2/2)



Models
(Supervised Machine Learning)

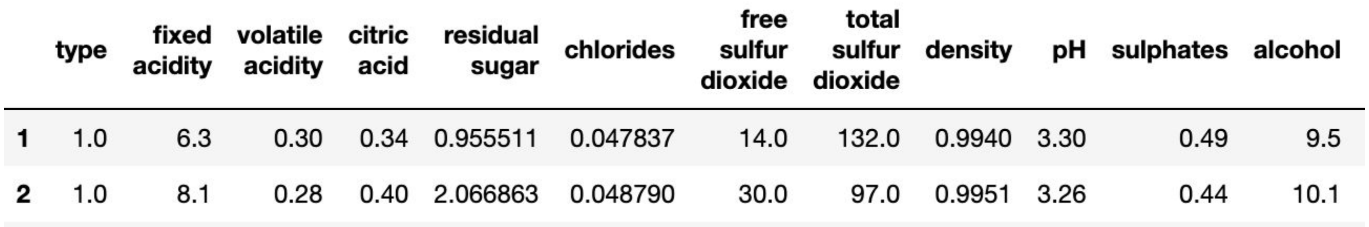
Logistic Regression | Random Forest

SVM | Gradient Boost | XGBoost

Trial 1.
Multi Classification

Multi Classification

X: Wine Features



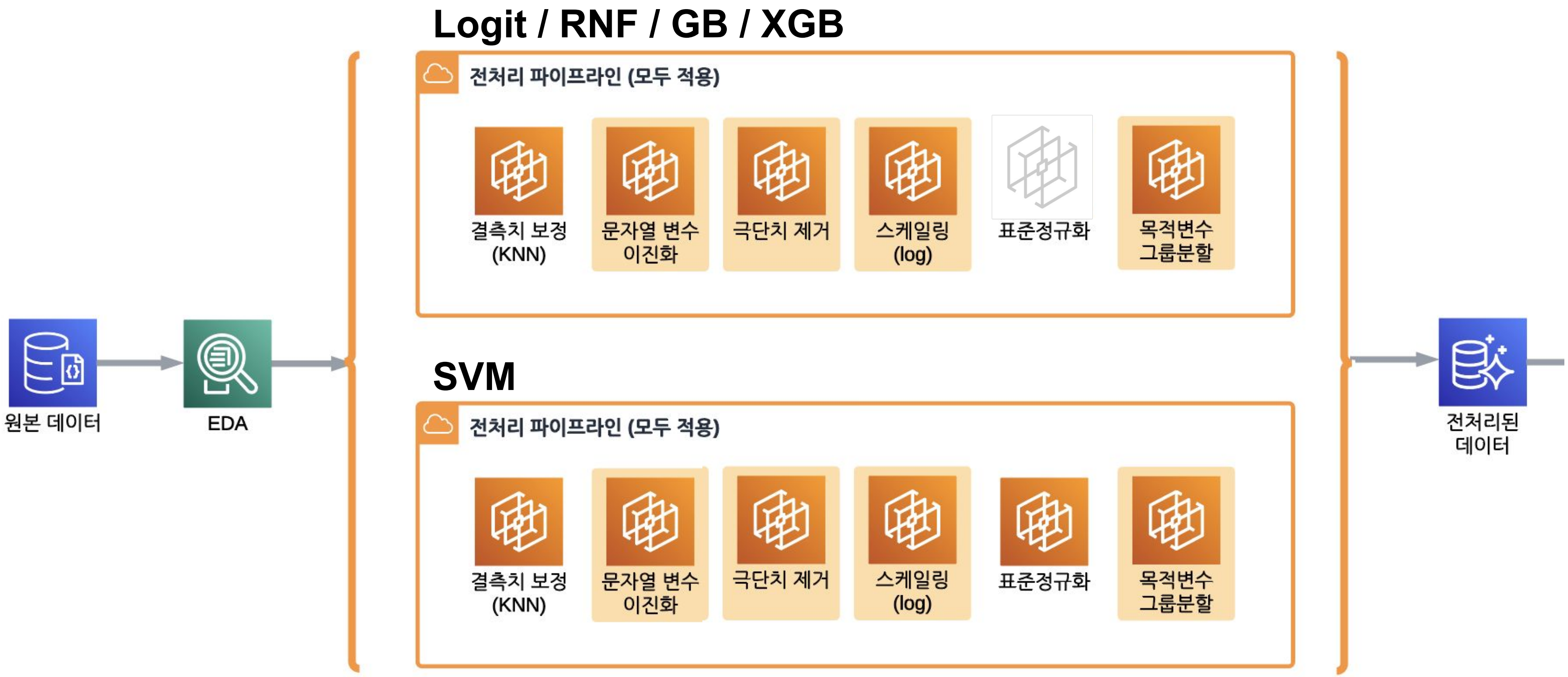
	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
1	1.0	6.3	0.30	0.34	0.955511	0.047837	14.0	132.0	0.9940	3.30	0.49	9.5
2	1.0	8.1	0.28	0.40	2.066863	0.048790	30.0	97.0	0.9951	3.26	0.44	10.1

Model

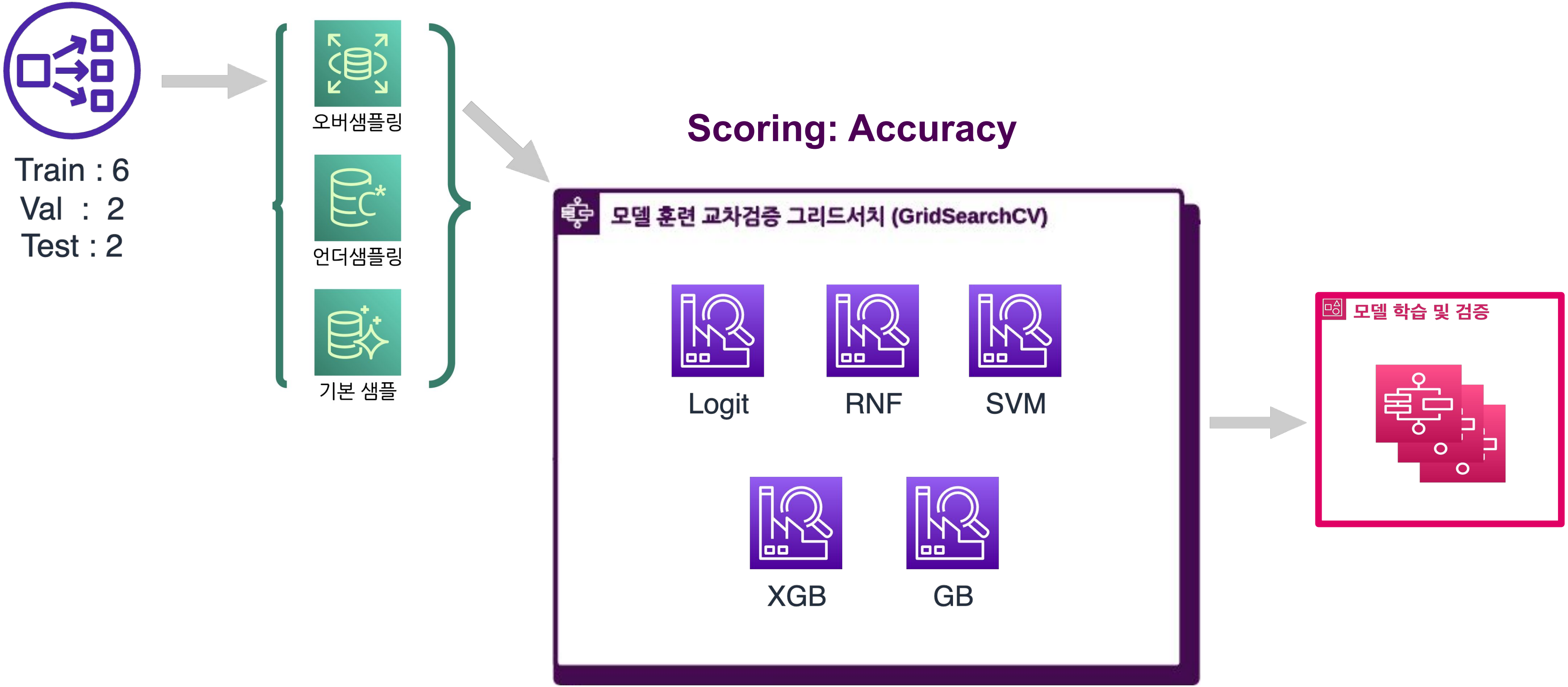
Y: Wine Quality

4 | 5 | 6 | 7 | 8

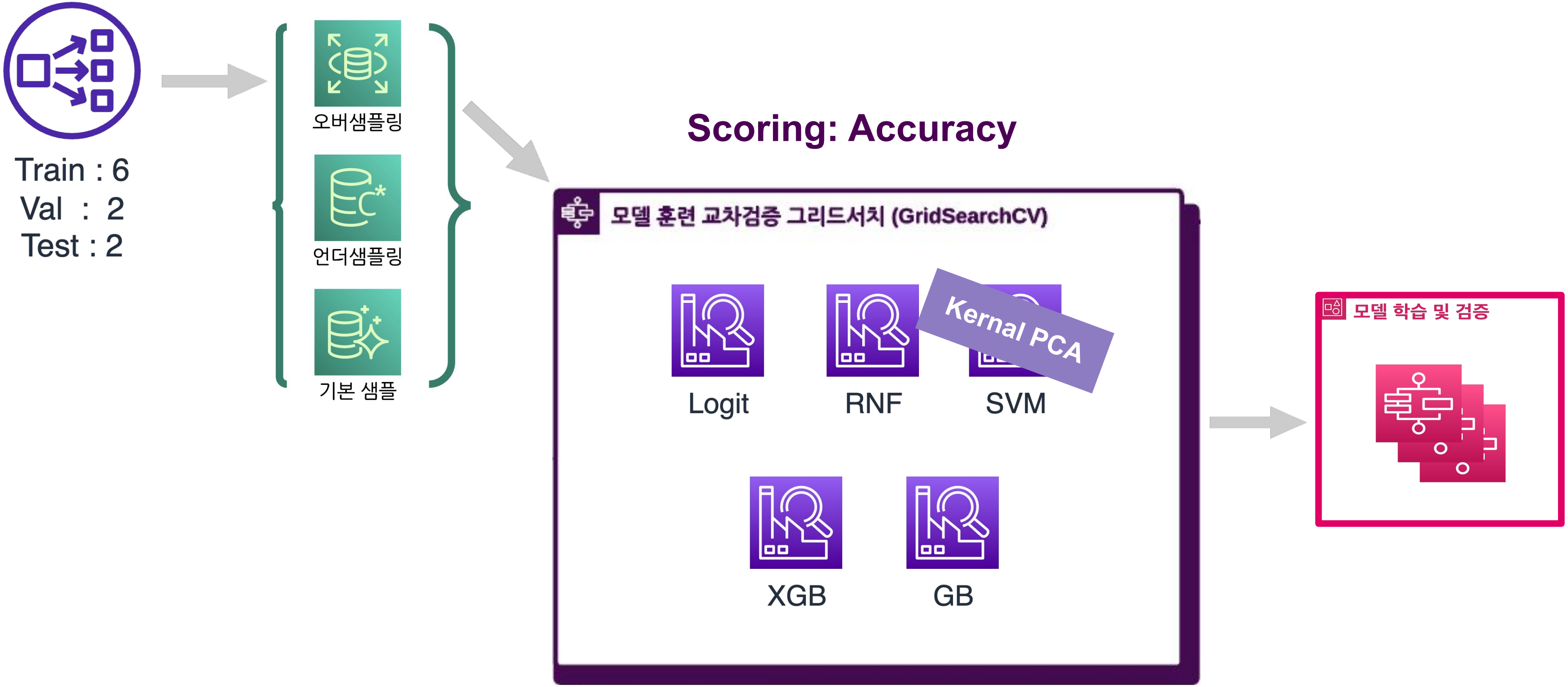
Multi Classification Workflow



Multi Classification Workflow



Multi Classification Workflow



Train : 6
Val : 2
Test : 2

오버샘플링

언더샘플링

기본 샘플

Scoring: Accuracy

모델 훈련 교차검증 그리드서치 (GridSearchCV)

Logit

RNF

SVM

Kernal PCA

XGB

GB

모델 학습 및 검증

Logistic Regression (oversampling)

Confusion Matrix:

```
                 pred
     [  26   15    0    3    5 ]
     [  94  174   38   30   63 ]
real [  63  141   69  157  134]
     [  14   23   16   96   79 ]
     [   3    4    2   19   16  ]
```

Total Accuracy:
0.2967

Class 5
Accuracy: 0.436
Precision: 0.4873
Recall: 0.436
f1: 0.4603

Class 7
Accuracy: 0.421
Precision: 0.3147
Recall: 0.4210
f1: 0.3602

Class 4
Accuracy: 0.5306
Precision: 0.13
Recall: 0.5306
f1: 0.2088

Class 6
Accuracy: 0.1223
Precision: 0.552
Recall: 0.1223
f1: 0.2002

Class 8
Accuracy: 0.3636
Precision: 0.0538
Recall: 0.3636
f1: 0.0938

Random Forest (oversampling)

Confusion Matrix:

```
                 pred
     [  8    20    17     3     0  ]
     [  6   296   104     9     0  ]
real [  3   108   383    85     5  ]
     [  0    10    82   119     3  ]
     [  0     0     9    17    12  ]
```

Total Accuracy:
0.6297

Class 5
Accuracy: 0.7132
Precision: 0.6833
Recall: 0.7518
f1: 0.7159

Class 7
Accuracy: 0.556
Precision: 0.6414
Recall: 0.5570
f1: 0.5962

Class 4
Accuracy: 0.1666
Precision: 0.4444
Recall: 0.1632
f1: 0.2388

Class 6
Accuracy: 0.6558
Precision: 0.6557
Recall: 0.7092
f1: 0.6814

Class 8
Accuracy: 0.3157
Precision: 0.7894
Recall: 0.3409
f1: 0.4761

Gradient Boost (oversampling)

Confusion Matrix:

```
               pred
     [  8   20   10    0    0 ]
     [ 10  292  128    3    1 ]
real [  4  109  437   25    1 ]
     [  0    4   98  107    0 ]
     [  0    0   14   13    9 ]
```

Total Accuracy:
0.6597

Class 4
Accuracy: 0.2105
Precision: 0.3636
Recall: 0.2105
f1: 0.2666

Class 5
Accuracy: 0.6728
Precision: 0.6870
Recall: 0.6728
f1: 0.6798

Class 6
Accuracy: 0.7586
Precision: 0.6360
Recall: 0.7586
f1: 0.6920

Class 7
Accuracy: 0.5119
Precision: 0.7229
Recall: 0.5119
f1: 0.5994

Class 8
Accuracy: 0.25
Precision: 0.8181
Recall: 0.25
f1: 0.3829

XGBoost (orign)

Confusion Matrix:

```
                pred
     [  1    31    17     0     0  ]
     [  2   274   121     2     0  ]
real [  0    96   439    29     0  ]
     [  0     9   117   102     0  ]
     [  0     2    20     8    14  ]
```

Total Accuracy:
0.6464

Class 4
Accuracy: 0.0204
Precision: 0.3333
Recall: 0.0204
f1: 0.0384

Class 5
Accuracy: 0.6867
Precision: 0.6650
Recall: 0.6867
f1: 0.6757

Class 6
Accuracy: 0.7783
Precision: 0.6148
Recall: 0.7783
f1: 0.6870

Class 7
Accuracy: 0.4473
Precision: 0.7234
Recall: 0.4473
f1: 0.5528

Class 8
Accuracy: 0.3181
Precision: 1.0
Recall: 0.3181
f1: 0.4827

KPCA_SVC (orign)

Confusion Matrix:

```
                pred
     [ 3    1    44    0    0 ]
     [ 0  129   286    0    0 ]
real [ 0   16   567    1    0 ]
     [ 0    1   149   64    0 ]
     [ 0    0    25    1   12 ]
```

Total Accuracy:
0.5966

Class 4
Accuracy: 0.0625
Precision: 1.0
Recall: 0.0625
f1: 0.1176

Class 5
Accuracy: 0.3108
Precision: 0.8775
Recall: 0.3108
f1: 0.4590

Class 6
Accuracy: 0.9708
Precision: 0.5294
Recall: 0.9708
f1: 0.6851

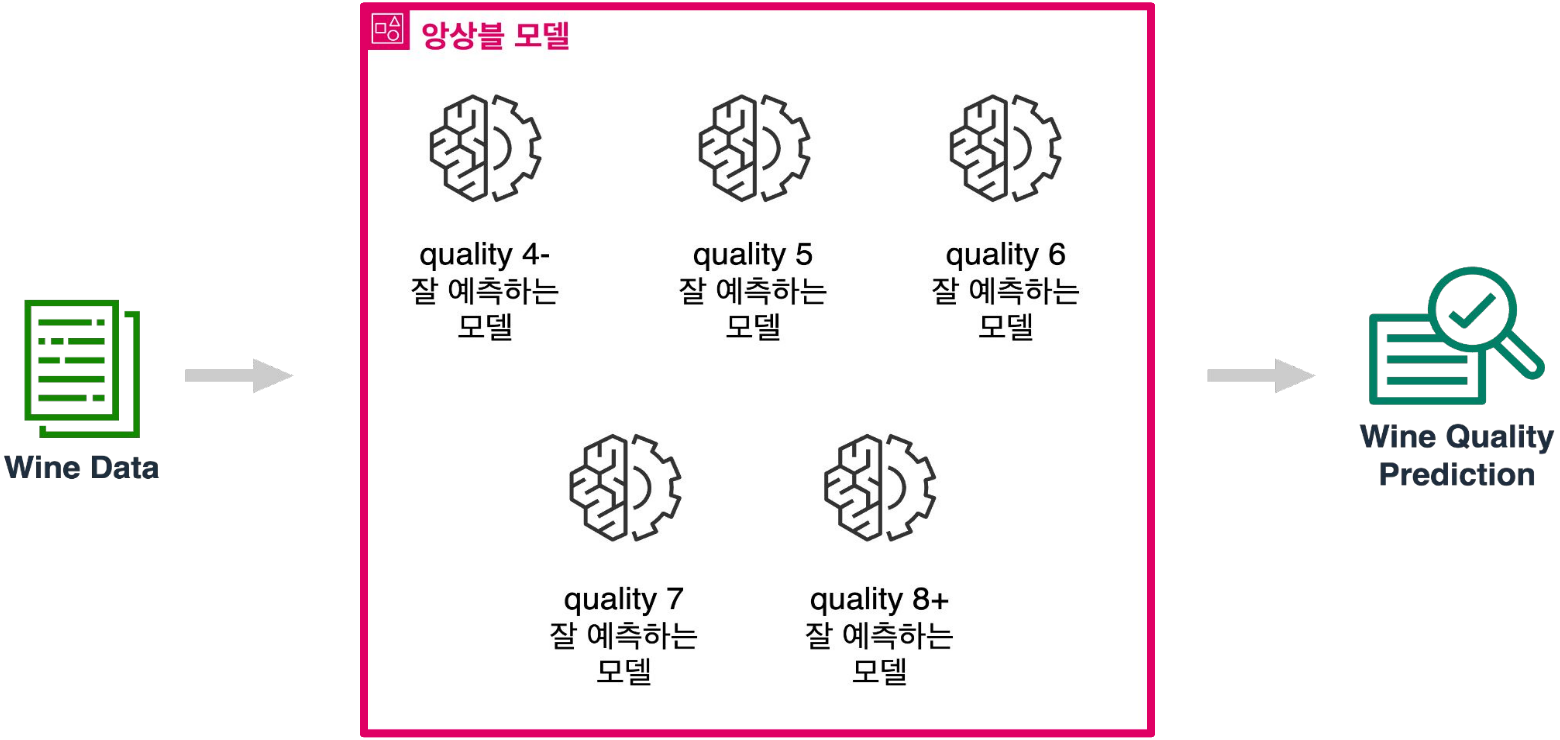
Class 7
Accuracy: 0.2990
Precision: 0.9696
Recall: 0.2990
f1: 0.4571

Class 8
Accuracy: 0.3157
Precision: 1.0
Recall: 0.3157
f1: 0.4799

모델 성능 비교 (Accuracy)

	Logistic Regression (over)	Random Forest (over)	Gradient Boost (over)	XGBoost (orign)	KPCA-SVM (orign)
Total	0.2967	0.6297	0.6597	0.6464	0.5966
quality 4 -	0.5306	0.1666	0.2105	0.0204	0.0625
quality 5	0.436	0.7132	0.6728	0.6867	0.3108
quality 6	0.1223	0.6558	0.7586	0.7783	0.9708
quality 7	0.421	0.556	0.5119	0.4473	0.2990
quality 8 +	0.3636	0.3157	0.25	0.3181	0.3157

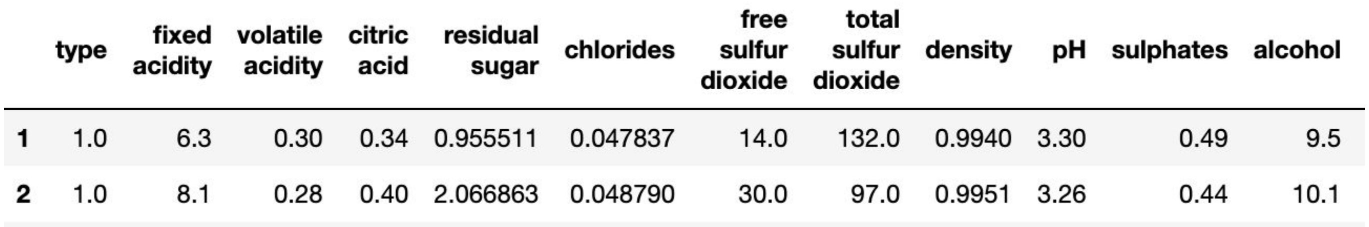
Model Essemble



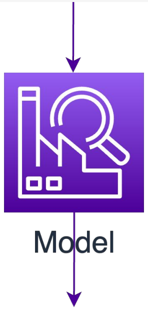
Trial 2.
Binary classification

Binary Classification

X: Wine Features



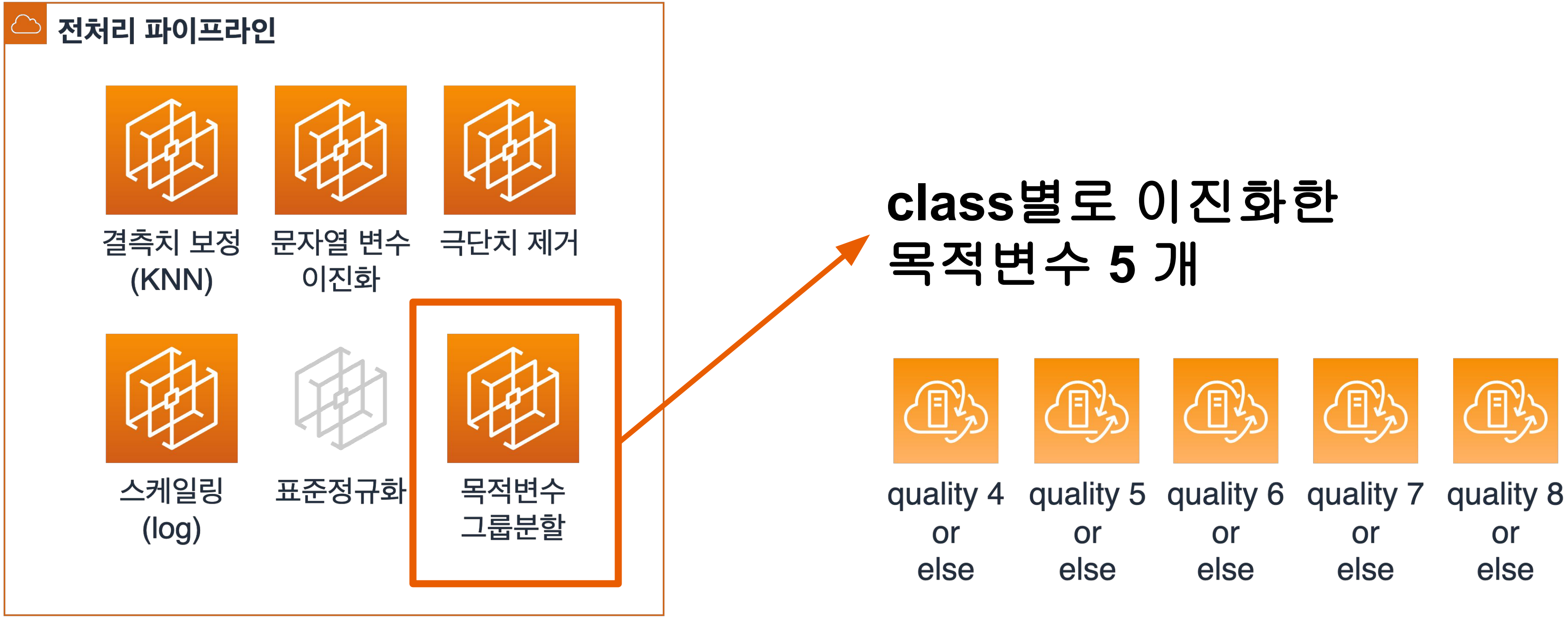
	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
1	1.0	6.3	0.30	0.34	0.955511	0.047837	14.0	132.0	0.9940	3.30	0.49	9.5
2	1.0	8.1	0.28	0.40	2.066863	0.048790	30.0	97.0	0.9951	3.26	0.44	10.1



Y_4: Wine Quality

4 or else

Binary Classification Workflow



전처리 파이프라인

결측치 보정 (KNN)

문자열 변수 이진화

극단치 제거

스케일링 (log)

표준정규화

목적변수 그룹분할

class별로 이진화한 목적변수 5 개

quality 4 or else

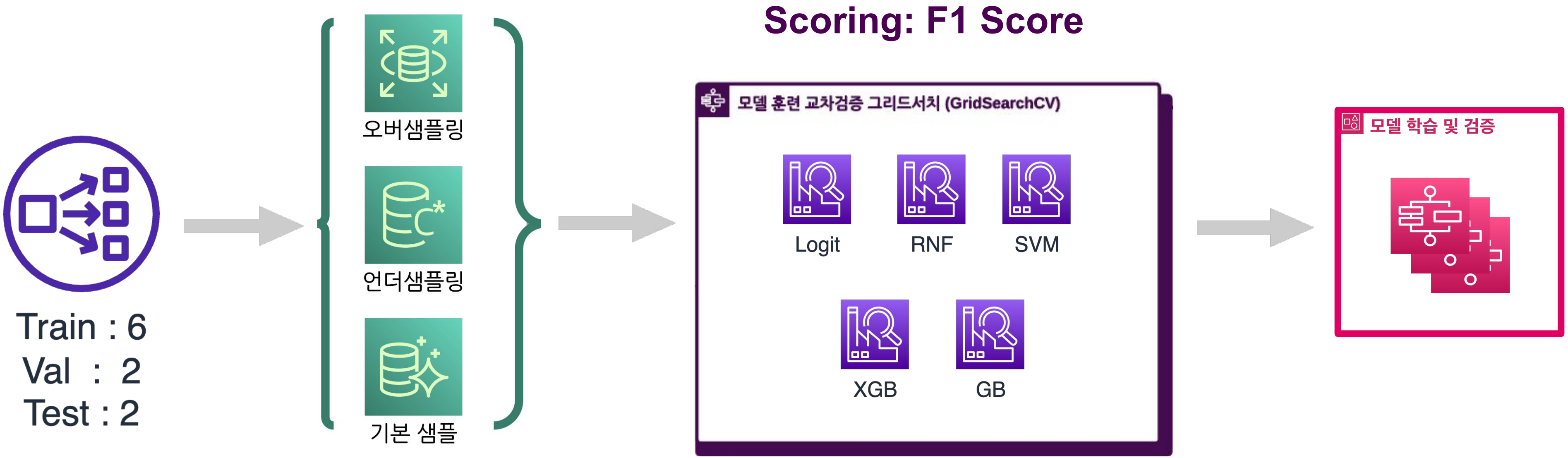
quality 5 or else

quality 6 or else

quality 7 or else

quality 8 or else

Binary Classification Workflow



Train : 6
Val : 2
Test : 2

오버샘플링

언더샘플링

기본 샘플

Scoring: F1 Score

모델 훈련 교차검증 그리드서치 (GridSearchCV)

Logit

RNF

SVM

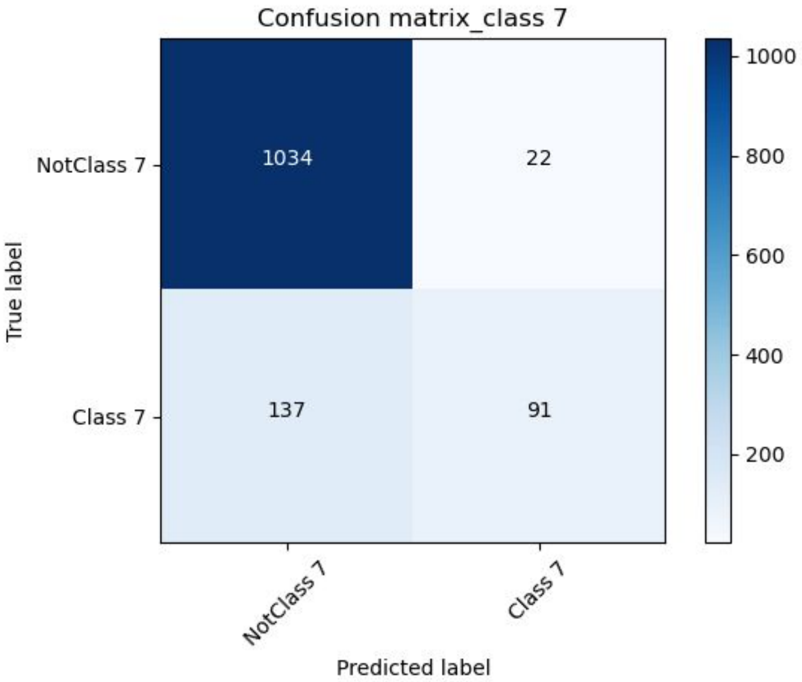
XGB

GB

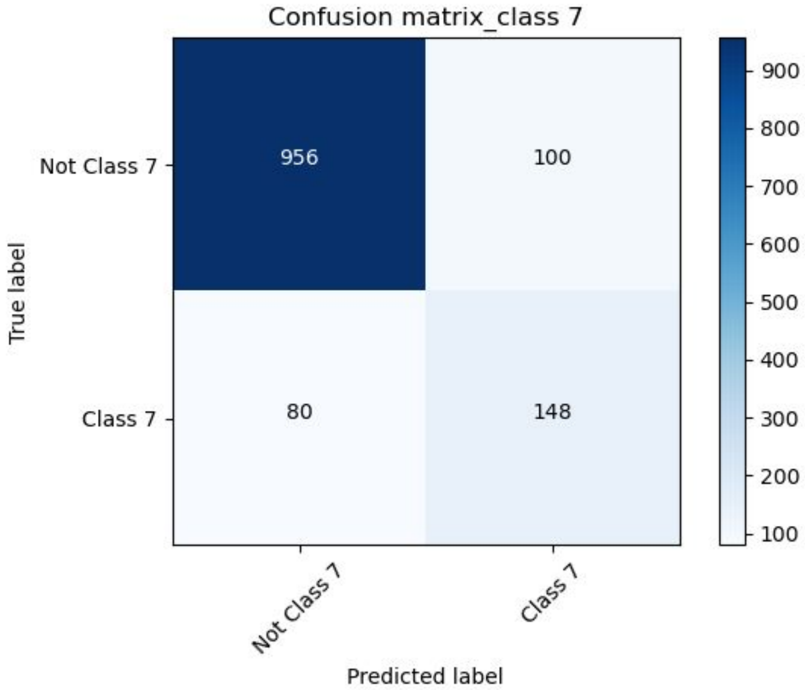
모델 학습 및 검증

Random Forest (class 7)

Scoring: Accuracy

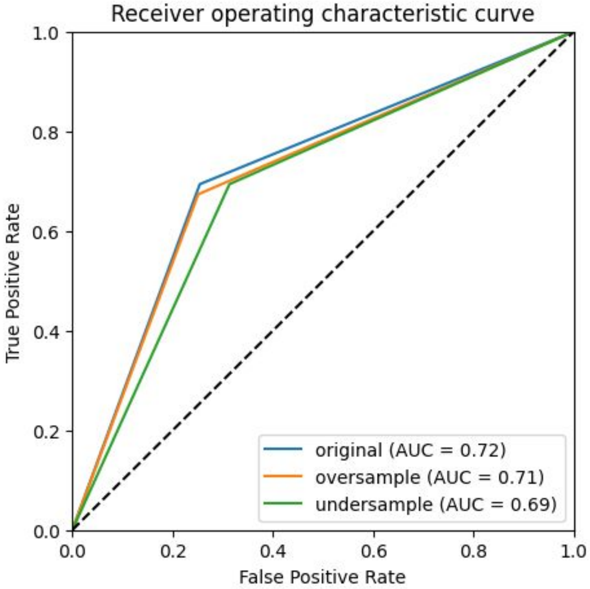


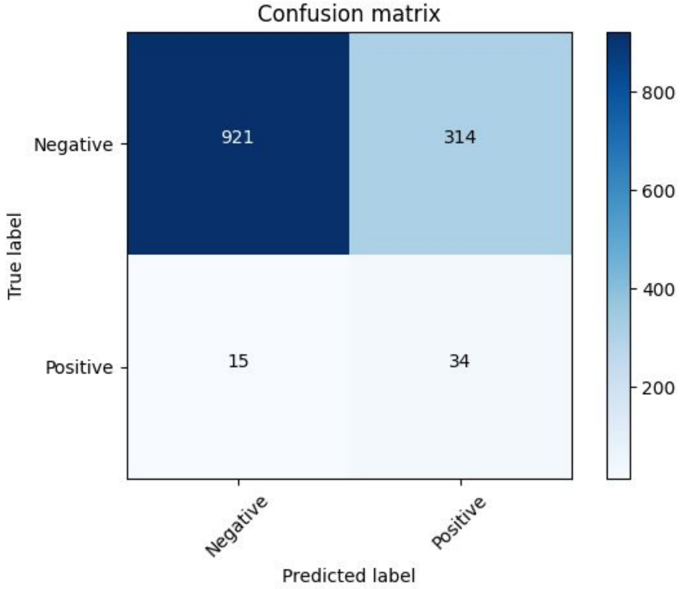
Scoring: F1



Logistic Regression (class 4)

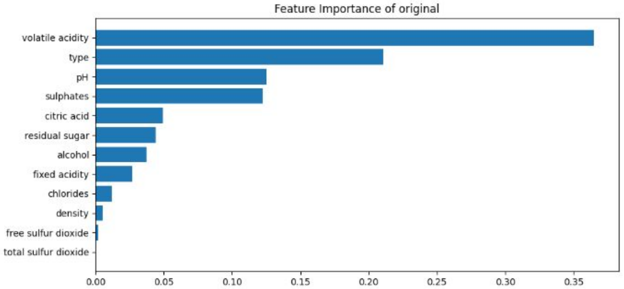
	Accuracy	Precision	Recall	F1-score	AUC
original	0.7438	0.0977	0.6939	0.1713	0.7198
oversample	0.7461	0.0962	0.6735	0.1684	0.7112
undersample	0.6869	0.0808	0.6939	0.1447	0.6903

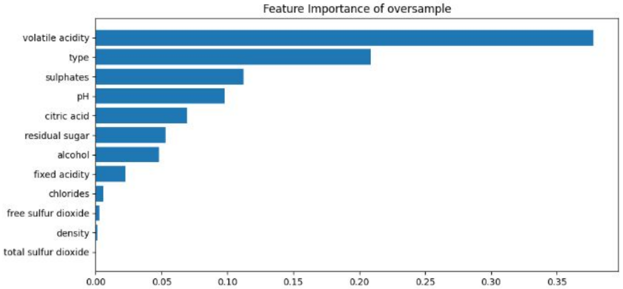


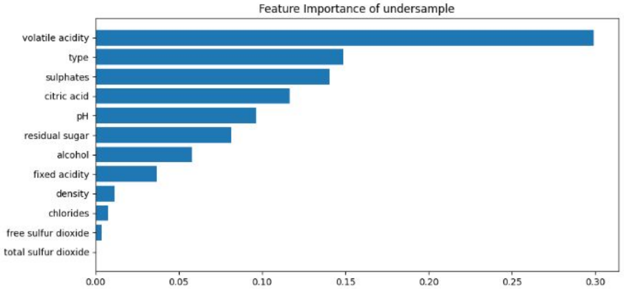


Logistic Regression (class 4)

	Accuracy	Precision	Recall	F1-score	AUC
original	0.7438	0.0977	0.6939	0.1713	0.7198
oversample	0.7461	0.0962	0.6735	0.1684	0.7112
undersample	0.6869	0.0808	0.6939	0.1447	0.6903

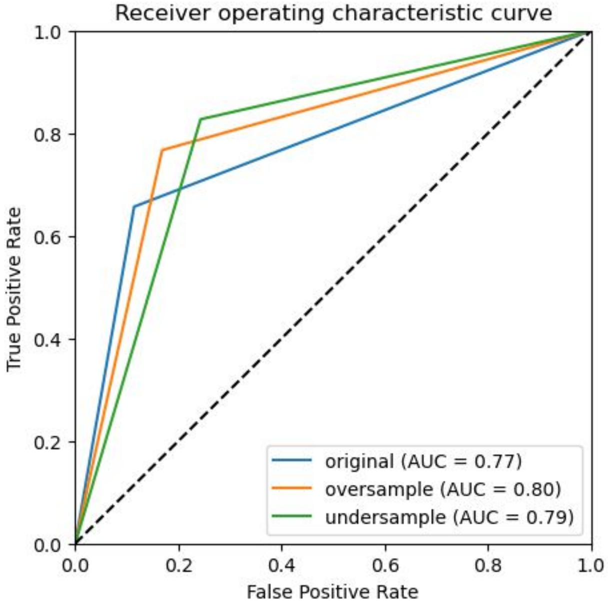

Feature Importance of original

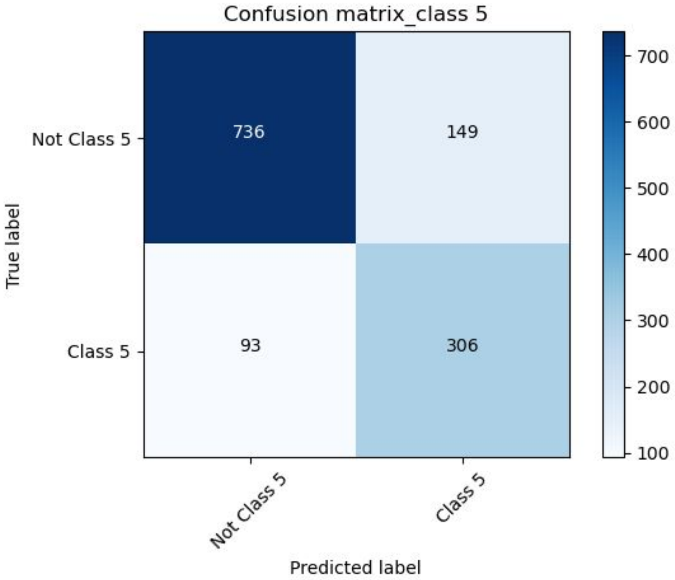

Feature Importance of oversample


Feature Importance of undersample

Random Forest (class 5)

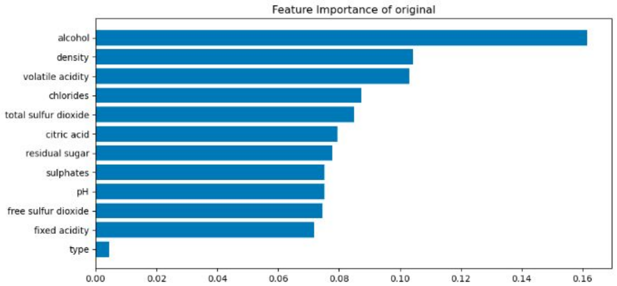
	Accuracy	Precision	Recall	F1-score	AUC
original	0.8146	0.7218	0.6566	0.6877	0.7713
oversample	0.8115	0.6725	0.7669	0.7166	0.7993
undersample	0.7788	0.6055	0.8271	0.6992	0.7921

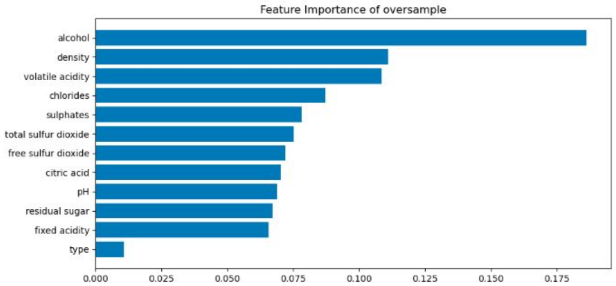


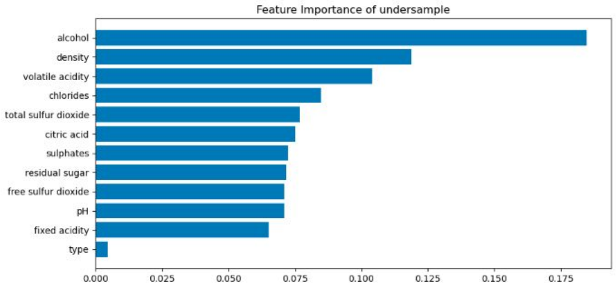


Random Forest (class 5)

	Accuracy	Precision	Recall	F1-score	AUC
original	0.8146	0.7218	0.6566	0.6877	0.7713
oversample	0.8115	0.6725	0.7669	0.7166	0.7993
undersample	0.7788	0.6055	0.8271	0.6992	0.7921

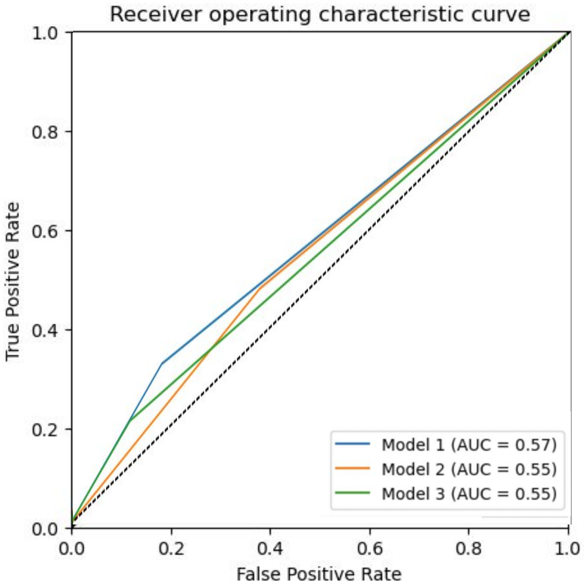

Feature Importance of original

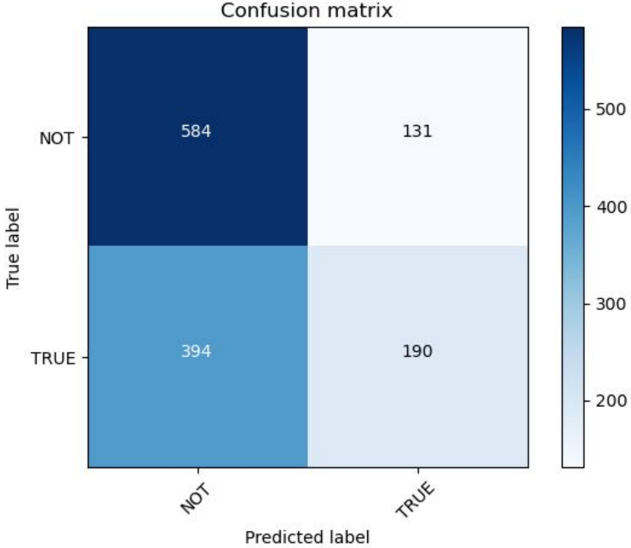

Feature Importance of oversample


Feature Importance of undersample

KPCA-SVC (class 6)

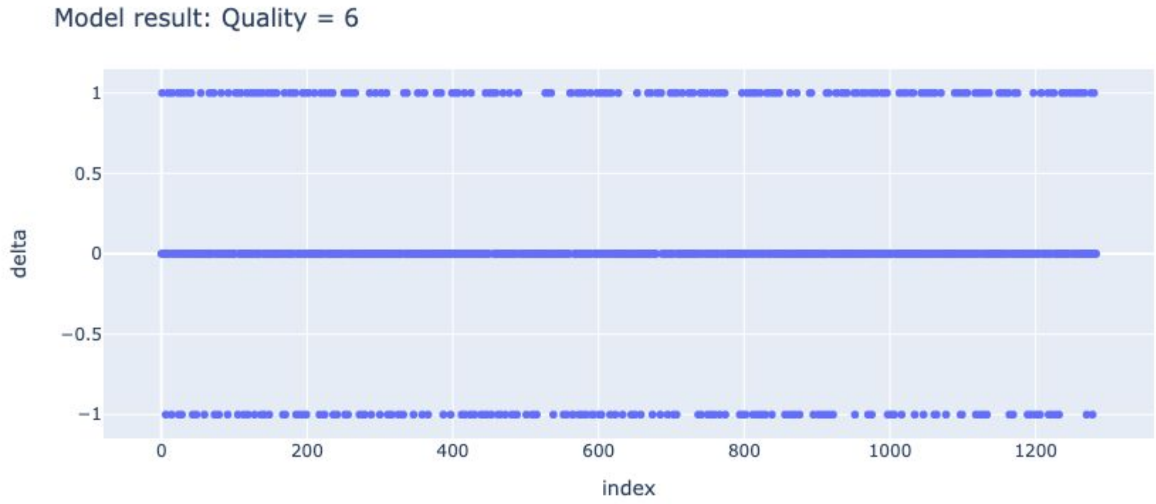
	Accuracy	Precision	Recall	F1-score	AUC
original	0.5958	0.5919	0.3253	0.4199	0.5711
oversample	0.5566	0.5073	0.4777	0.4921	0.5494
undersample	0.5789	0.5885	0.2106	0.3102	0.5452





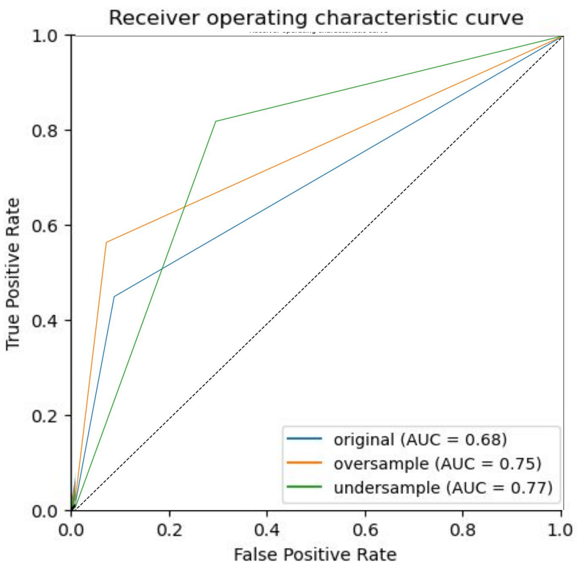
KPCA-SVC (class 6)

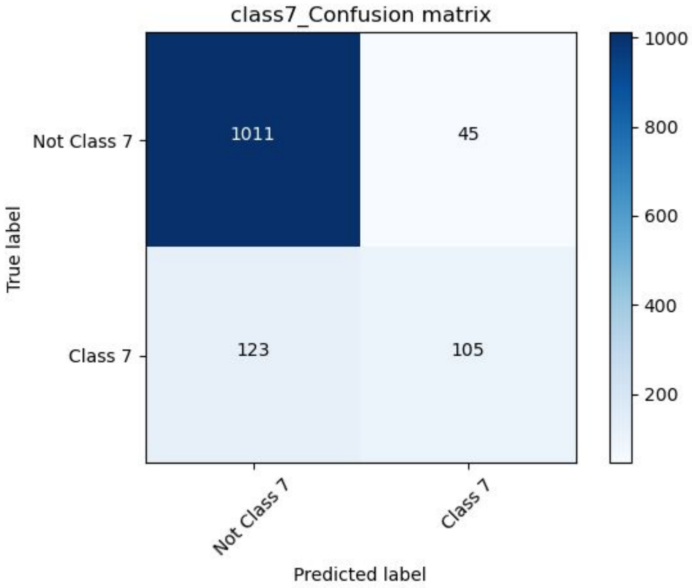
	Accuracy	Precision	Recall	F1-score	AUC
original	0.5958	0.5919	0.3253	0.4199	0.5711
oversample	0.5566	0.5073	0.4777	0.4921	0.5494
undersample	0.5789	0.5885	0.2106	0.3102	0.5452



Gradient Boost (class 7)

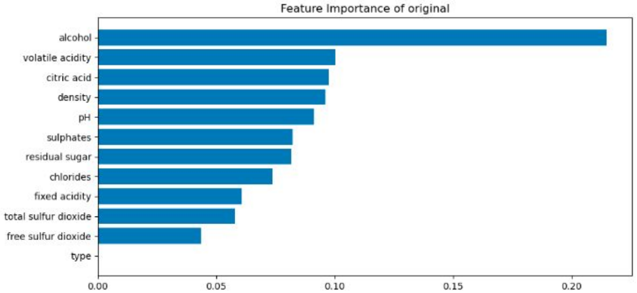
	Accuracy	Precision	Recall	F1-score	AUC
original	0.8692	0.7000	0.4605	0.5556	0.7090
oversample	0.8575	0.6154	0.5263	0.5674	0.7276
undersample	0.7360	0.3856	0.8202	0.5245	0.7690

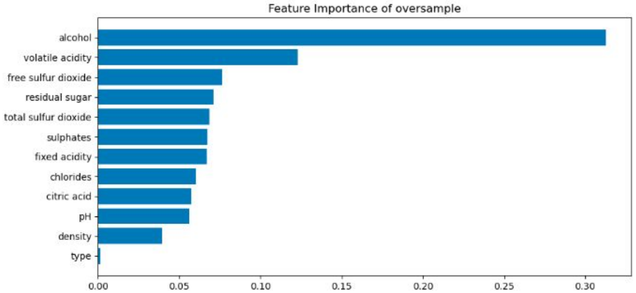


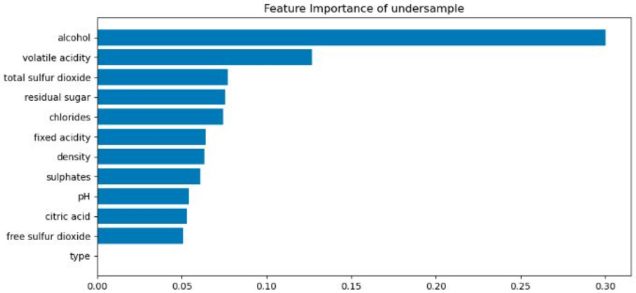


Gradient Boost (class 7)

	Accuracy	Precision	Recall	F1-score	AUC
original	0.8692	0.7000	0.4605	0.5556	0.7090
oversample	0.8575	0.6154	0.5263	0.5674	0.7276
undersample	0.7360	0.3856	0.8202	0.5245	0.7690

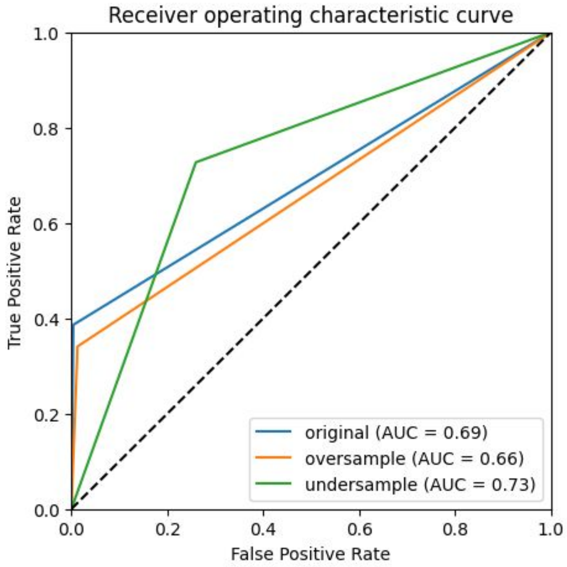

Feature Importance of original

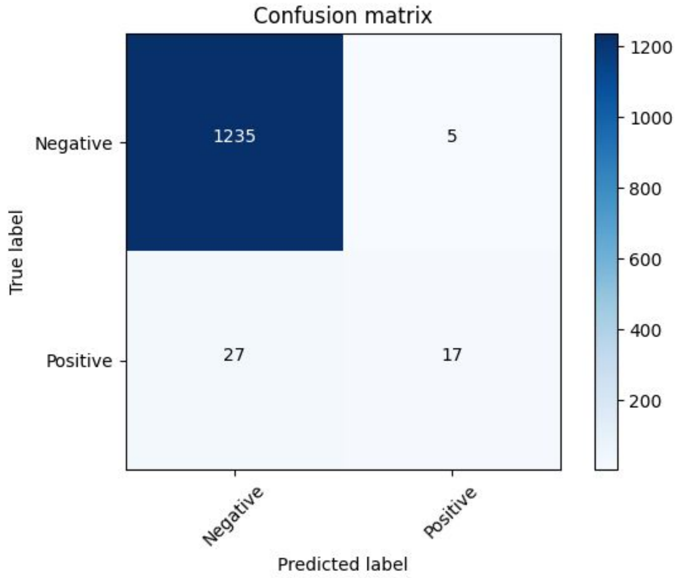

Feature Importance of oversample


Feature Importance of undersample

XGBoost (class 8)

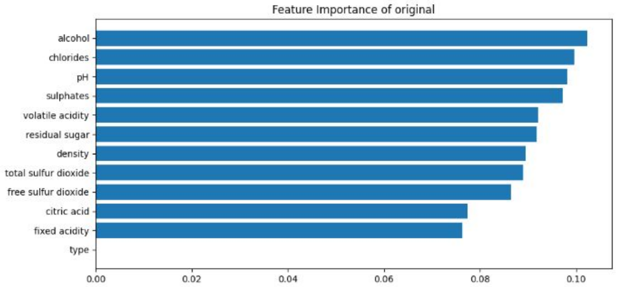
	Accuracy	Precision	Recall	F1-score	AUC
original	0.9751	0.7727	0.3864	0.5152	0.6912
oversample	0.965	0.4839	0.3409	0.4	0.664
undersample	0.7399	0.0904	0.7273	0.1608	0.7338

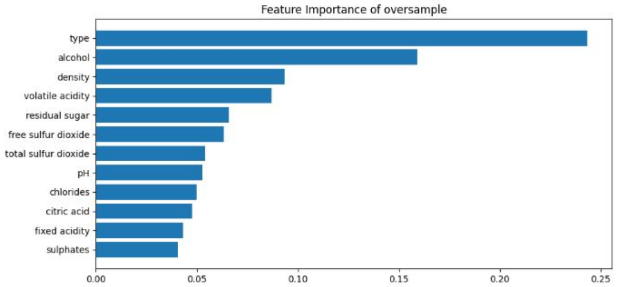

Receiver operating characteristic curve

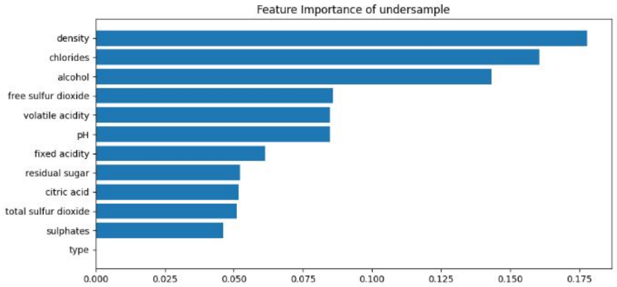

Confusion matrix

XGBoost (class 8)

	Accuracy	Precision	Recall	F1-score	AUC
original	0.9751	0.7727	0.3864	0.5152	0.6912
oversample	0.965	0.4839	0.3409	0.4	0.664
undersample	0.7399	0.0904	0.7273	0.1608	0.7338


Feature Importance of original


Feature Importance of oversample

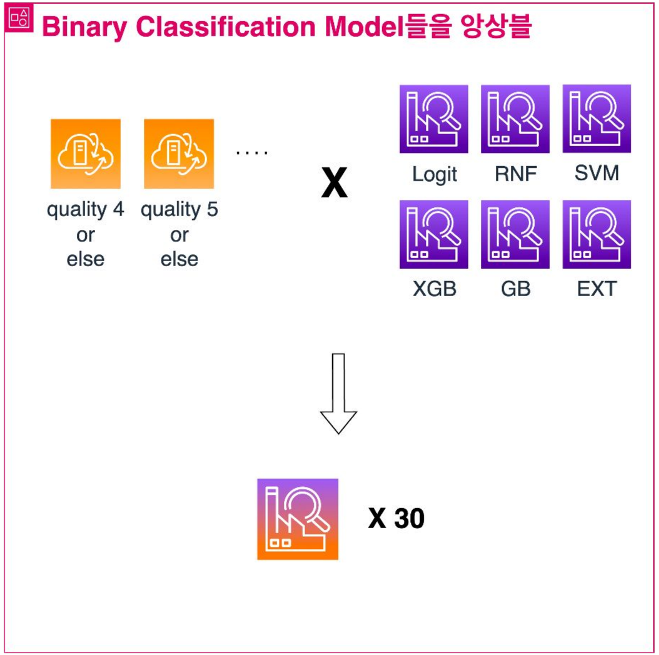

Feature Importance of undersample

Ensemble Model

VotingClassifier | StackingClassifier

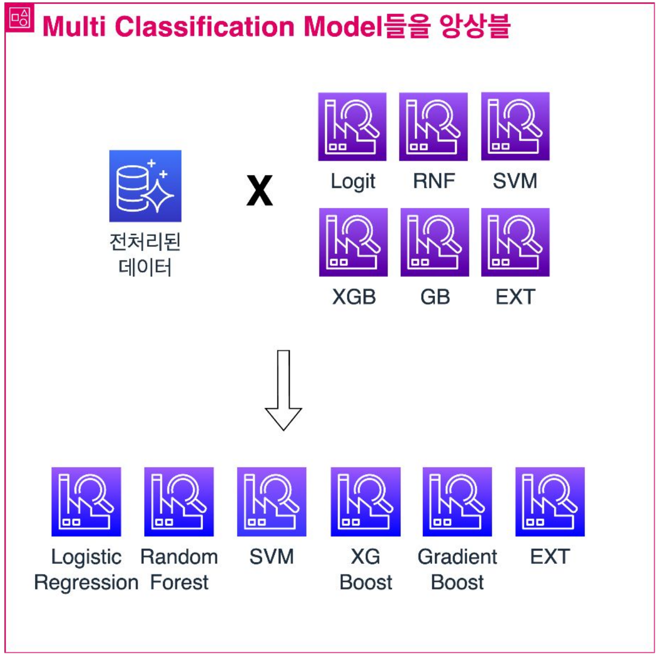
Model Ensemble

Trial 2



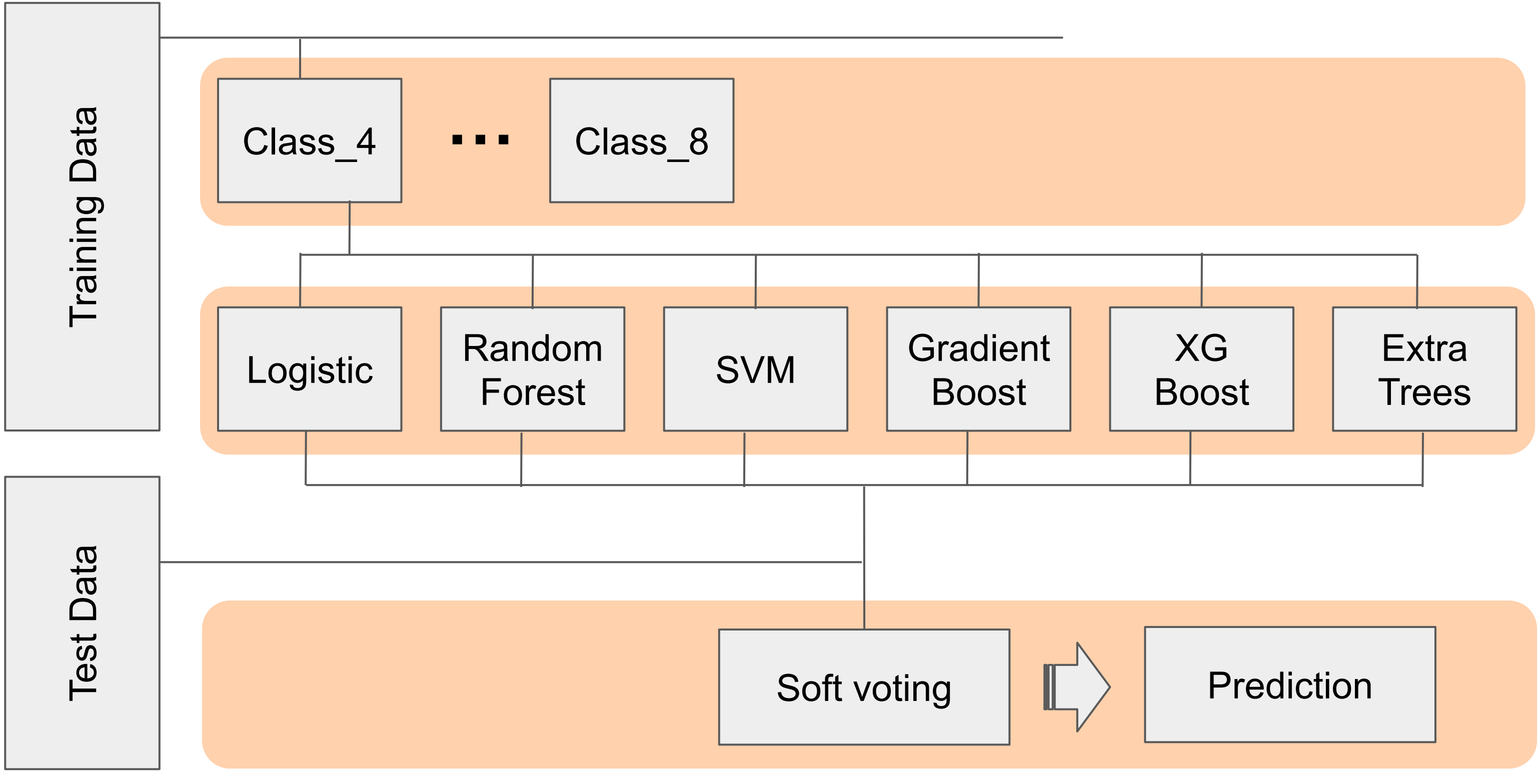
VotingClassifier

Trial 1



StackingClassifier

VotingClassifier



[scale]
ordinal
to
binary

[fitting]
original
over
under

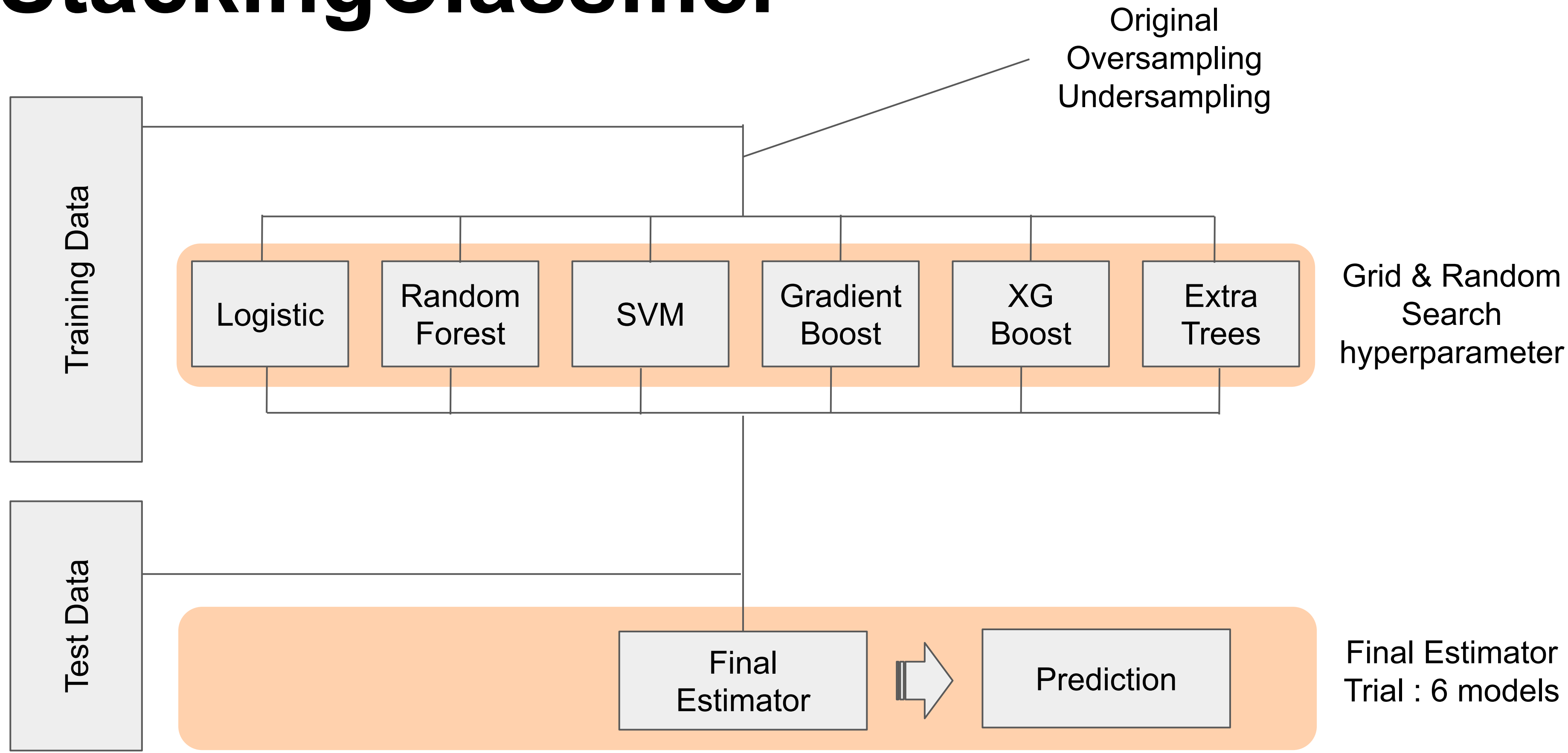
[weight]
uniformly
f1-score
precision
best

VotingClassifier

index weight	Accuracy	Precision	Recall	F1-score
uniformly	0.5522	0.5412	0.5522	0.5426
f1-score	0.5428	0.5473	0.5428	0.5361
precision	0.5405	0.5442	0.5405	0.5297
best	0.5506	0.5763	0.5506	0.5478

Poor prediction

StackingClassifier



StackingClassifier

		Accuracy	Precision	Recall	F1-score
log	orign	0.6838	0.6936	0.6838	0.6725
	over	0.6745	0.6761	0.6745	0.6678
	under	0.4455	0.5281	0.4455	0.4538
rnf	orign	0.6776	0.688	0.6776	0.6675
	over	0.6628	0.6728	0.6628	0.6487
	under	0.4587	0.549	0.4587	0.4755
svm	orign	0.6854	0.7239	0.6854	0.6678
	over	0.6783	0.6783	0.6783	0.6742
	under	0.4587	0.5283	0.4587	0.4635

StackingClassifier

		Accuracy	Precision	Recall	F1-score
xgb	orign	0.6745	0.6897	0.6745	0.6657
	over	0.655	0.6899	0.655	0.6362
	under	0.3894	0.5179	0.3894	0.4098
gdb	orign	0.6838	0.6994	0.6838	0.6766
	over	0.641	0.676	0.641	0.6259
	under	0.4073	0.5126	0.4073	0.4244
ext	orign	0.6822	0.6969	0.6822	0.6723
	over	0.6846	0.6933	0.6846	0.6752
	under	0.4478	0.5438	0.4478	0.4584

결론

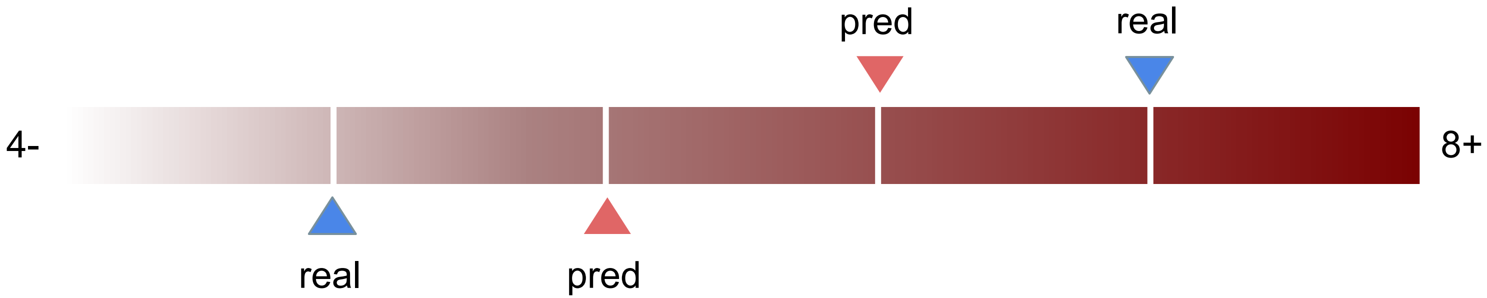
Quality grade | Ordinal scale

Unsupervised learning

Quality

PROBLEM β
예측등급 < 실제등급

- 적절한 가격산정 실패로 인한 판매손실 가능성



PROBLEM α
예측등급 > 실제등급

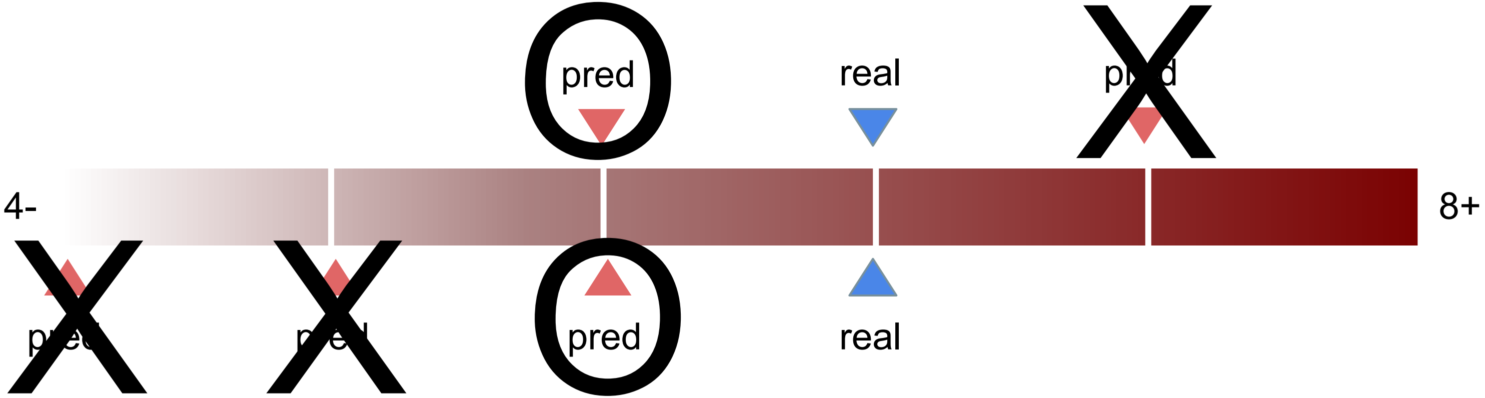
- 와인 산지 평판 하락 가능성
- 소비자 피해 가능성

Quality

PROBLEM α

예측등급 > 실제등급

예측등급이 실제등급보다 낮은 범위



PROBLEM β

예측등급 < 실제등급

1등급 차까지만

Quality

Accurate

+Tolerance

Prediction

Observed

	4-	5	6	7	8+
4-	5	34	15	0	0
5	3	286	108	2	0
6	2	84	462	29	1
7	0	13	103	109	3
8+	0	0	14	12	15

(ExtraTreesClassifier, oversampling)

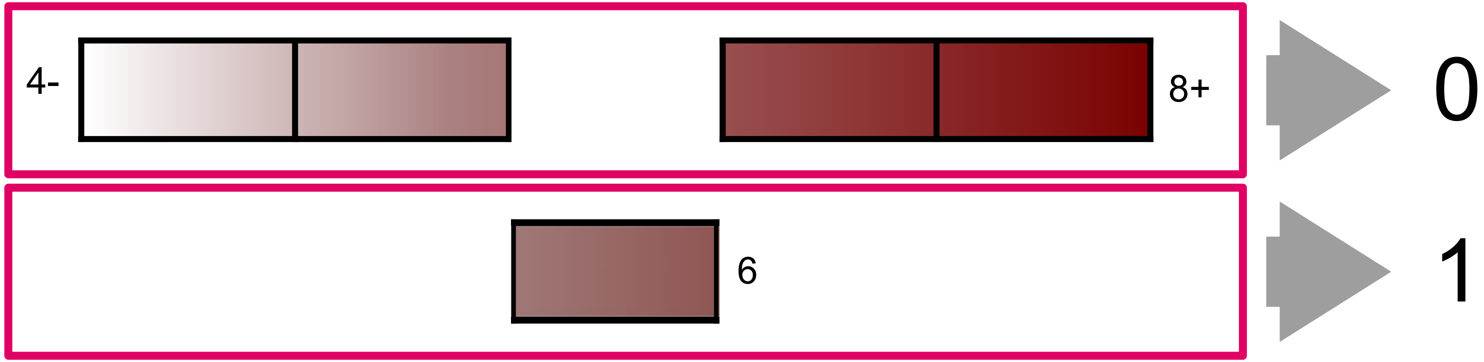
	4-	5	6	7	8+
4-	7	32	8	0	0
5	3	363	71	0	0
6	3	153	381	1	2
7	0	19	121	75	1
8+	0	1	25	3	15

(StackingClassifier, oversampling)

Quality

	Tutorial		StackingClassifier	
	(ExtraTrees, oversampling)		(SVM, original)	(XGB, oversampling)
	Accurate	Tolerance	Accurate	Tolerance
Accuracy	0.6746	0.7232	0.6854	0.8012
Precision	0.6746	0.8747	0.7239	0.9036
Recall	0.6746	0.8300	0.6854	0.8731
F1	0.6590	0.8431	0.6678	0.8822

Ordinal



Loss

Expected Gain

이진화하면
순서정보 상실

>

이진화하면 모델별로
잘 예측하는 변수의 결과만
취합 가능

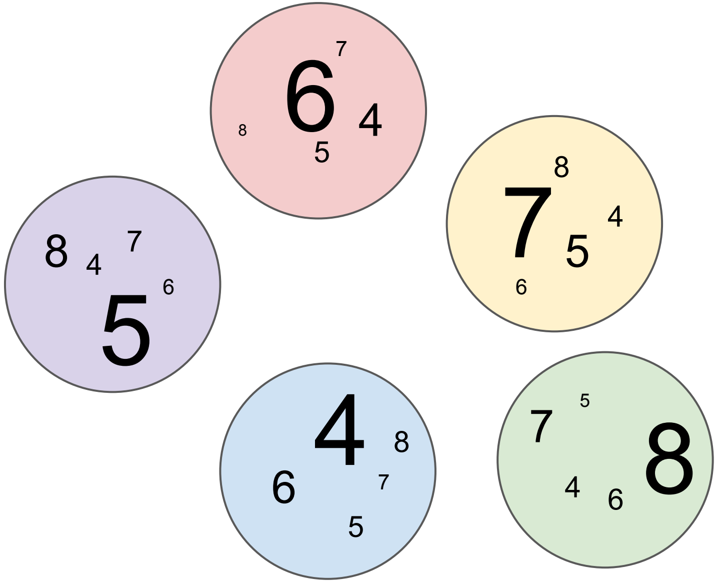
+

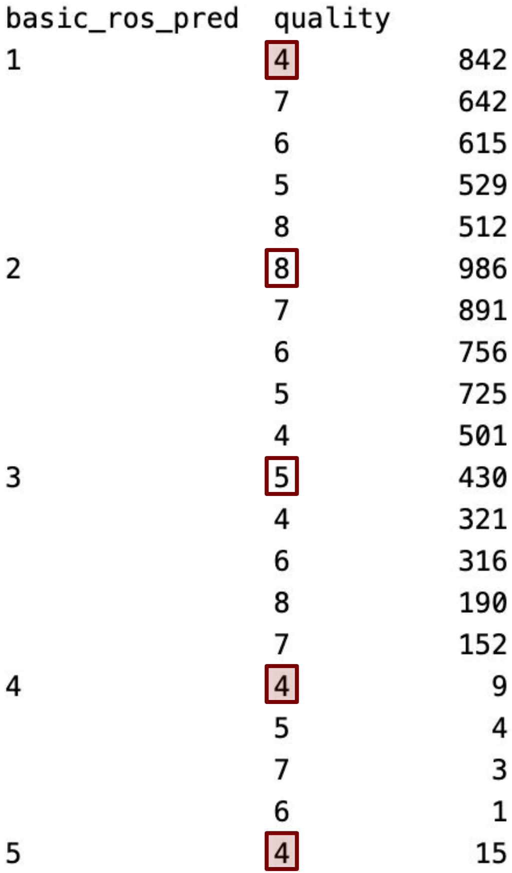
클래스별
over/undersampling 적용
가능

Unsupervised

K-Means Clustering

1. 5 clusters vs 15 clusters and re-clustering(by fcluster())
2. No Dimension Reduction vs PCA vs Kernel PCA
3. Origin / Oversampling / Undersampling Data





```
basic_ros_pred  quality
1               4          842
                7          642
                6          615
                5          529
                8          512
2               8          986
                7          891
                6          756
                5          725
                4          501
3               5          430
                4          321
                6          316
                8          190
                7          152
4               4            9
                5            4
                7            3
                6            1
5               4           15
```

감사합니다.