N 분석

데이터분석 스케치

화이트 와인에 대한 분석

임 정

임정(19910219)

jayjunglim@gmail.com

임정(19910219) / 허현(19930422) / 양혜리(19900729)

요 약

데이터 셋
Wine Quality Data Set (white wine)
탐색적 자료분석

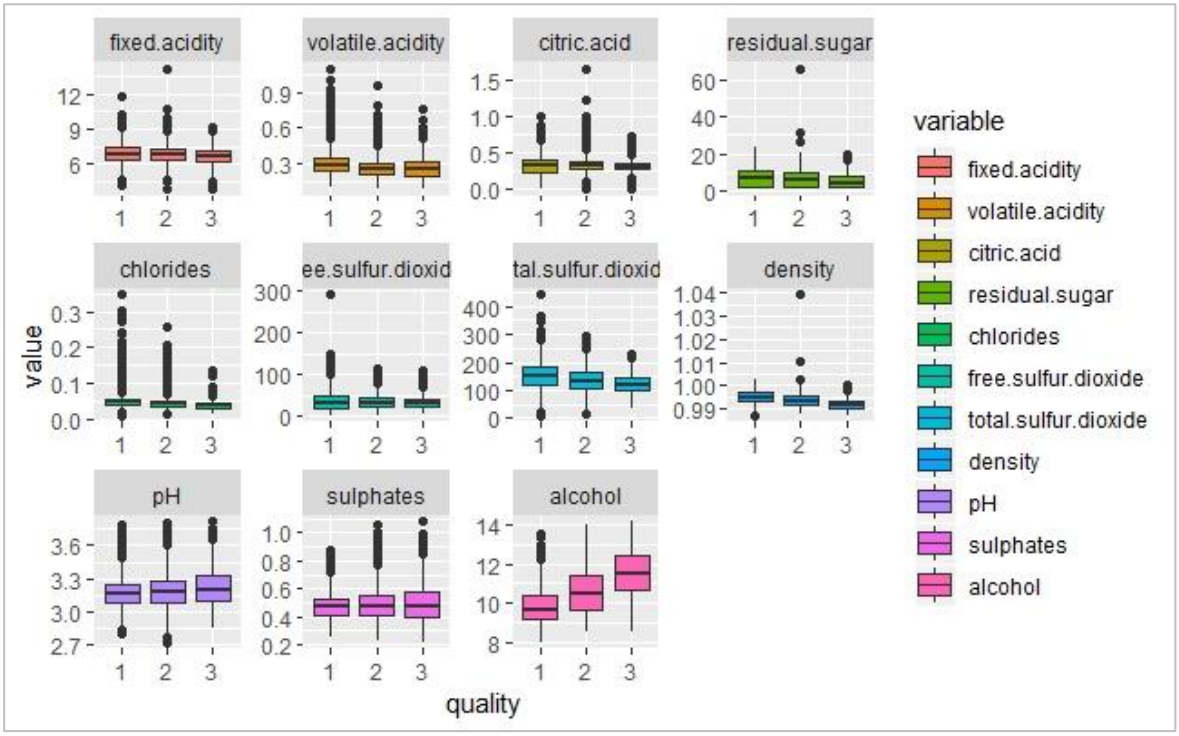

그림 1 quality(상:3, 중:2 하:1)에 대한 설명변수의 그림상자

문제 정의

화이트 와인의 품질은 11개의 설명 변수 중 다음 변수들에 상관관계를 가질 것이다.

1. 양의 상관관계: pH, alcohol
2. 음의 상관관계: volatile acidity, citric acid, total sulfur dioxide, density

분류

1. 다중 회귀분석

표 1 오차에 대한 가정 만족

alcohol	독립성	△	free.sulfur.dioxide	독립성	○
	정규성	○		정규성	X
	등분산성	△		등분산성	△
volatile.acidity	독립성	△	total.sulfur.dioxide	독립성	△
	정규성	△		정규성	X
	등분산성	○		등분산성	○
residual.sugar	독립성	○	pH	독립성	○
	정규성	△		정규성	○
	등분산성	△		등분산성	○
sulphates	독립성	○	citric.acid	독립성	△
	정규성	○		정규성	△
	등분산성	○		등분산성	△

범례: 회귀직선가정 만족, 절반 만족, 위험

표 2 최종모형 회귀계수들에 대한 평가

Regression Coefficients	Estimate	Std.Error	t-value	Pr(>|t|)	VIF
(Intercept)	-1.93432	0.25752	-7.51	<0.001	
alcohol	0.31609	0.00992	31.87	<0.001	1.2751
residual.sugar	0.02187	0.00241	9.08	<0.001	1.2995
volatile.acidity	-1.66983	0.10711	-15.59	<0.001	1.0222
sulphates	0.49067	0.09583	5.12	<0.001	1.0375
pH	0.18122	0.07365	2.46	<0.001	1.079

Residual standard error: 0.628 on 3422 degrees of freedom
Multiple R-squared: 0.267, Adjusted R-squared: 0.266
F-statistic: 249 on 5 and 3422 DF, p-value: <0.0000000000000002

다중 회귀분석에 의한 **adj-R2의 값은 0.266** 으로 상대적으로 낮음을 알 수 있었다.

2. 의사결정나무

- 랜덤 포레스트

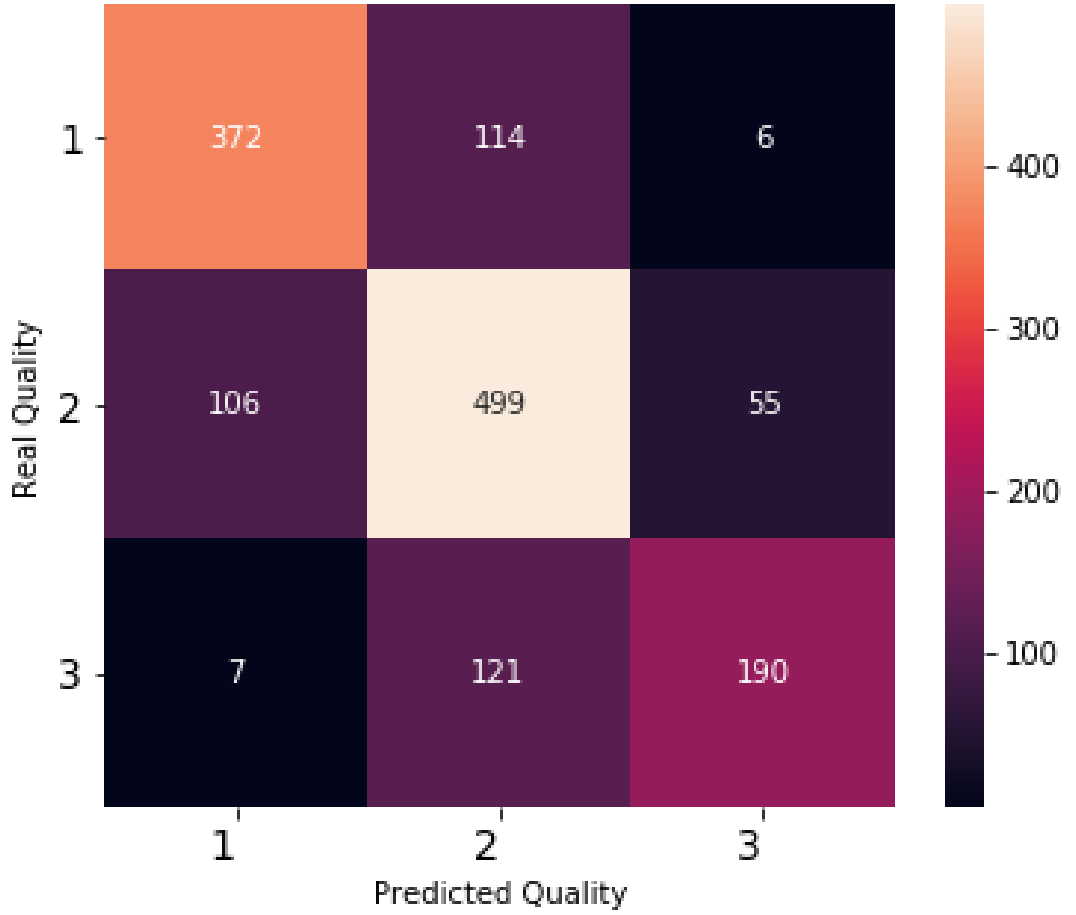

그림 2 최종 랜덤포레스트 분류 매트릭스

각 행은 실제 와인의 품질 정보를 나타내고 각 열은 예측한 와인의 품질 정보를 나타낸다. 랜덤포레스트 모델에서 품질 1(하)를 예측한 것 중 372개의 테스트 관측치가 맞았다는 것을 뜻한다. 품질 예측율은 각 품질1(하): 76.7%, 품질2(중) 67.9%, 품질3(상) 75.68%가 되었다. 총 예측율은 **f1-score: 0.72**으로 **72% 였다.**

3. 서포트 벡터 머신

- Gaussian Kernel

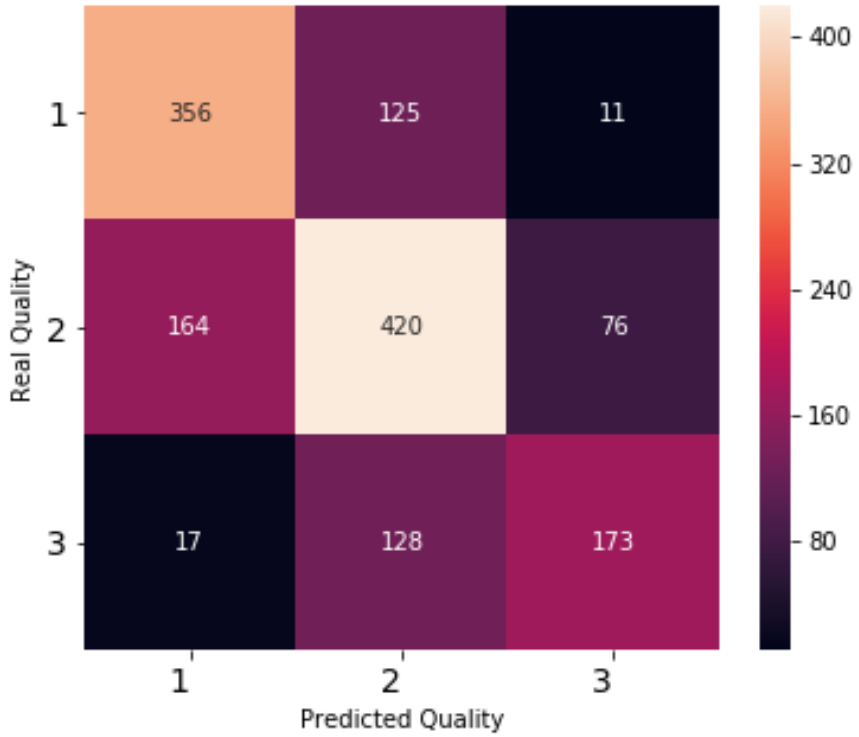

그림 3 최종 SVM 분류 매트릭스

SVM모형 결과 품질1(하): 66.29%, 품질2(중): 62.4%, 품질3(하):66.53% 가 도출되었으며 SVM 모형의 최종 예측율은 **f1-score: 0.64으로 64%** 였다.

결론

표 3 최종결론: 설명변수 5가지

MR	SVM (Linear kernel)	DT (random forest)
Alcohol	Alcohol	Alcohol
Volatile acidity	Volatile acidity	Volatile acidity
Residual sugar	Chloride	Residual sugar
pH	pH	pH
sulphates	sulphates	Free sulfur dioxide
		Fixed acidity

결과적으로 3가지 모델에서 2가지이상 선택된 **Alcohol, Volatile acidity, Residual sugar, pH, Free sulfur dioxide** 상관관계를 입증할 수 있었다. 문제 정의에서 가정했던 설명변수들과 대부분 일치함을 알 수 있었다.

목 차

표 목차

그림 목차

서 론

목적

본 보고서는 Data Science Competition 2018 (이하 DSC2018)의 온라인 테스트인 데이터 분석 스케치에 대한 과제 제출이다. 본 보고서는 분석할 데이터 셋으로 UCI Machine Learning Repository의 Wine Quality Data Set[1] 중 화이트 와인 데이터를 사용하였다. 데이터 셋은 4898개의 관측치와 12개의 변수로 이루어진 데이터 셋으로 데이터 분석 프로세스를 익힐 때 필수적으로 쓰이는 데이터 셋이다 DSC2018의 본격적인 참여에 앞서 테스트의 일환으로 진행되었으며 본 보고서를 통해 데이터 분석의 프로세스를 익히고 역량을 증진하는데 목적이 있다.

와인 데이터를 통해 데이터에 제시된 11개의 설명변수 중에 품질을 결정하는데 가장 영향력이 있는 설명변수를 추론해 나가는 것이 본 분석의 최종 목적이다. 해당 데이터 셋의 70%을 훈련 데이터셋으로, 30%를 시험셋으로 선정하여 어떤 모형이 와인데이터셋을 잘 설명하는지 살펴보고 최종적으로 설명변수를 도출하고자 한다.

[1] https://archive.ics.uci.edu/ml/datasets/Wine+Quality

본 론

1. 데이터 분석 프로세스

와인 데이터분석에 앞서 전체적인 데이터 분석 프로세스를 파악하고 진행하고자 한다.

그림 4 데이터 분석 프로세스

① **문제 정의:** 문제 정의는 포괄적인 기준과 함께 질의한다. 예를 들면 문제는 '고혈압 환자의 재방문 횟수는 증상정도에 따라 어떤 경향을 보이는가?' 또는 '해당 제품의 고객 선호도의 차이를 어떻게 추적할 것인가?' 와 같이 분석하고자 하는 대상이 될 수 있다. 데이터 분석 프로젝트에 있어 성공의 핵심은 목표와 요구사항을 정확히 이해하고 이해한 내용을 바탕으로 명확한 문제를 정의하는 것이다. 우리는 본 와인 데이터 분석을 통해 어떤 요인이 품질에 영향을 주는지 보고자 하였고 이 분석과정을 통해 향후 DSC2018에서 마주하게 될 분석 프로세스의 기초를 학습하고자 하였다.

② **데이터 준비:** 데이터 준비는 데이터의 획득, 정리, 정규화, 변형 등을 통해 최적의 데이터 세트를 어떻게 얻을 것인가에 대한 것이다. 최종 분석에 있어서 기초 토대를 마련하는 과정이므로 데이터가 유실되어 있지 않은 지, 사용하려는 분류 모델에 있어 적합한지에 대한 타당성이 이루어져야 하는 단계이다. 우리는 여기서 와인 데이터를 받고 결측치 유무를 확인하는 과정을 통해 데이터 준비의 첫 단계를 시작하였다.

③ **탐색적 자료분석:** 탐색적 자료분석(Exploratory data analysis, EDA)는 데이터로부터 패턴, 연관성, 관계를 발견하기 위해 데이터를 그래픽이나 통계의 형태로 관찰하는 행동이다. 시각화 하는 것은 결국 데이터로부터 의미 있는 패턴을 찾기 위한 수단이다. 우리는 탐구의 기술적인 도구로써 각 팀원들의 역량에 맞는 SAS, R, Python을 동시에 이용하여 교차적 탐구

에 활용하였다.

④ **예측 모델링:** 예측 모델링은 결과의 개연성을 가장 잘 예측할 수 있는 통계 모델을 세우거나 기존의 모델을 선택하는 데이터 분석의 한 과정이다. 우리가 본 분석에서 다룰 알고리즘들은 다음과 같다.

- 다중 회귀분석 (Multiple Regression)
- 의사결정나무 (Decision Tree)

  랜덤 포레스트 (Random Forest)
- 서포트 벡터 머신 (Support Vector Machines)

첫 번째로는 연속형 자료형을 모델링할 때 다중 회귀분석은 가장 기본적인 접근 방식이기 때문에 예측 모델링으로 사용하였다. 가장 기본적인 모형이기에 다른 모형을 접근하는데 좋은 기초가 될 것으로 예상하였다.

두 번째로는 의사결정나무 모델을 도입하였다. 다중 회귀분석과 비슷하게 적용하기 비교적 쉽고 결과에 대한 해석이 편리하다는 장점이 있기에 분석의 도입 부분에 사용하기 적절하기 때문이다. 또한 의사결정규칙(decision rule)을 나무구조로 도표화하여 분류와 예측을 수행하기 때문에 분석자가 그 과정을 쉽게 이해하고 설명할 수 있는 장점이 있기에 도입하게 되었다. 그 자체가 분류와 예측 모형으로 사용될 수 있기 때문이다. 그 중 랜덤 포레스트 모형을 사용했다. 랜덤 포레스트는 의사결정나무를 이용해 만들어진 알고리즘으로 여러 개의 의사결정나무를 만들고 투표를 시켜 다수결로 결과를 결정하는 방법이다.

세 번째로는 서포트 벡터 머신을 사용하였다. 보다 더 높은 유연성과 비선형 학습능력(Non-linear Capacity) 때문에 서포트 벡터 머신은 데이터 마이닝 분야에서 주목 받았고 또한 높은 예측 성능을 나타내기 때문에 도입하게 되었다.

변수 선택은 불필요한 설명변수를 제거함에 있어서 유용하며 더 쉬운 모델들이 더 좋은 성능을 내고 해석하기 용이하게 해줄 수 있다. 복잡한 모델은 데이터를 과 적합 시킬 수 있고 일반화에 대한 효용성을 잃을 수도 있다. 반면 지나치게 간단한 모델은 제한된 학습 능력(limited

learning capacity)를 내놓을 수도 있기에 세 가지 모델 모두를 사용함으로써 그 차이를 확인하고자 한다.

예측 모델링과 관련하여 무엇보다 중요한 것은, 어떤 특정 문제에 있어 그 문제를 가장 잘 설명할 수 있는 최선의 모델을 선택했는 가이다. 이를 위해서 무엇보다 문제를 잘 이해해야 하고 또한 예측 모델링 알고리즘의 특성을 잘 이해해야만 이러한 평가가 가능 할 것이다.

⑤ **결과 시각화:** 결과 시각화는 데이터 분석 프로세스 중 가장 마지막 단계이며 테이블 형식, 2D 플롯, 3D 플롯, 인포그래픽 등에서 무엇을 이용할 것 인가 어떻게 결과를 표현할 것인가에 대한 대답을 하기 위한 과정이라고 할 수 있겠다.

2. 자료 개요

표 4 와인데이터구조

> str(Winequltiy-White) 'data.frame': 4898 obs. of 12 variables:		
변수명	**자료형**	**데이터**
$ fixed.acidity	num	7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 …
$ volatile.acidity		0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 …
$ citric.acid		0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 …
$ residual.sugar		20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 …
$ chlorides		0.045 0.049 0.05 0.058 0.058 0.05 0.045 …
$ free.sulfur.dioxide		45 14 30 47 47 30 30 45 14 28 …
$ total.sulfur.dioxide		170 132 97 186 186 97 136 170 132 129 …
$ density		1.001 0.994 0.995 0.996 0.996 …
$ pH		3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 …
$ sulphates		0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 …
$ alcohol		8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 …
$ quality	int	6 6 6 6 6 6 6 6 6 6 …

R의 str 함수를 이용한 분석결과 화이트 와인 데이터에는 총 12개의 변수 4898개의 관측치가 있으며 12개의 변수 중에 11개는 연속형 1개는 범주형 변수로 나누어지는 것을 확인할 수 있었다.

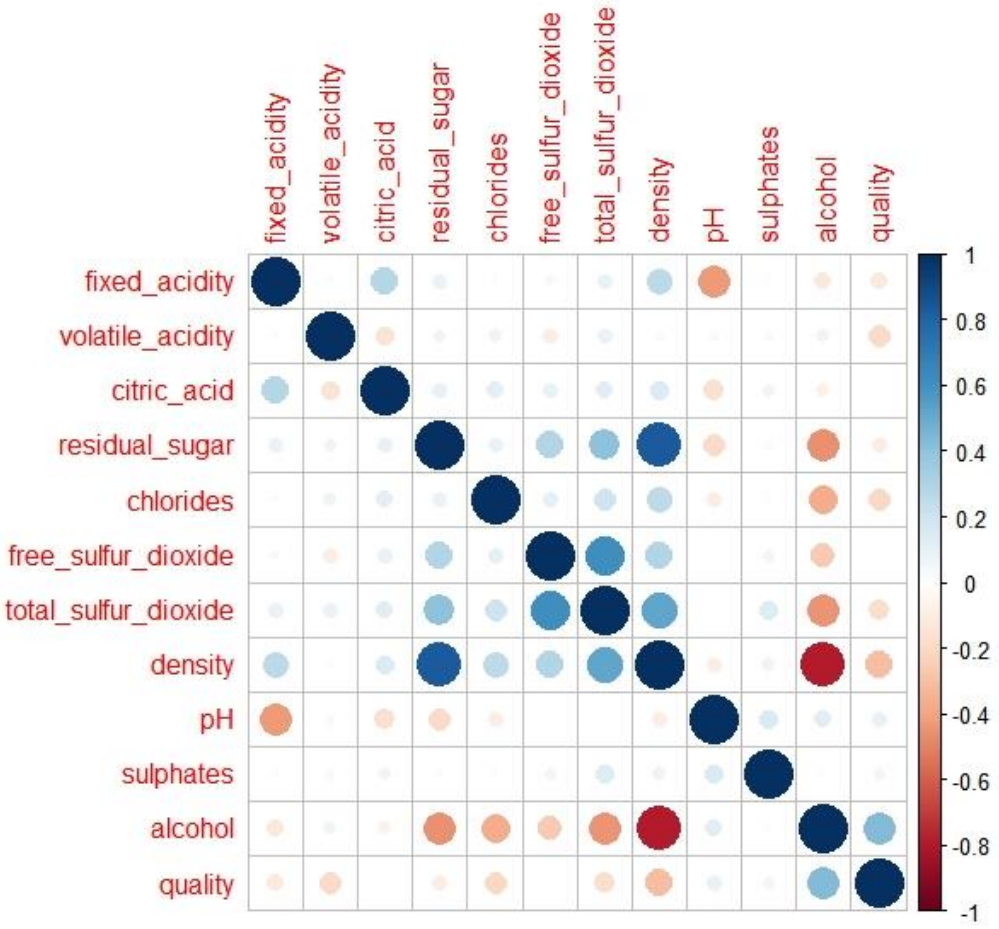

그림 5 각 변수 별 상관 정도

문제 정의

위의 그림3 은 설명변수 11개와 범주형 변수1개에 대하여 선형 상관정도를 정리하였다. 여기서 중점적으로 확인할 수 있는 것은 품질에 대하여 alcohol이 강한 양의 상관관계를 나타내며 그 다음으로 density, chlorides, volatile_acidity 순으로 음의 상관관계가 크다는 사실을 확인 할 수 있었다. 위 자료를 바탕으로 품질에 영향을 주는 변수 4가지를 확인할 수 있었고 후에 기술할 모델링을 통해 정말 품질에 영향을 주는 것이 다양한 모형을 통해 확인해보고자 한

다. 추가적으로 품질에 영향이 있을 것으로 추측되는 residual sugar, total sulfur dioxide의 연관성도 살펴보고자 한다. 알려진 정보에 따르면 residual sugar는 발효 후 잔류하는 설탕의 양이고, total sulfur dioxide는 산화 방지를 위해 넣는 첨가물의 총량에 대한 정보로 free sulfur dioxide가 50ppm 이상이면 맛과 향이 느껴진다고 한다.

3. 분석

가. 탐색적 자료분석

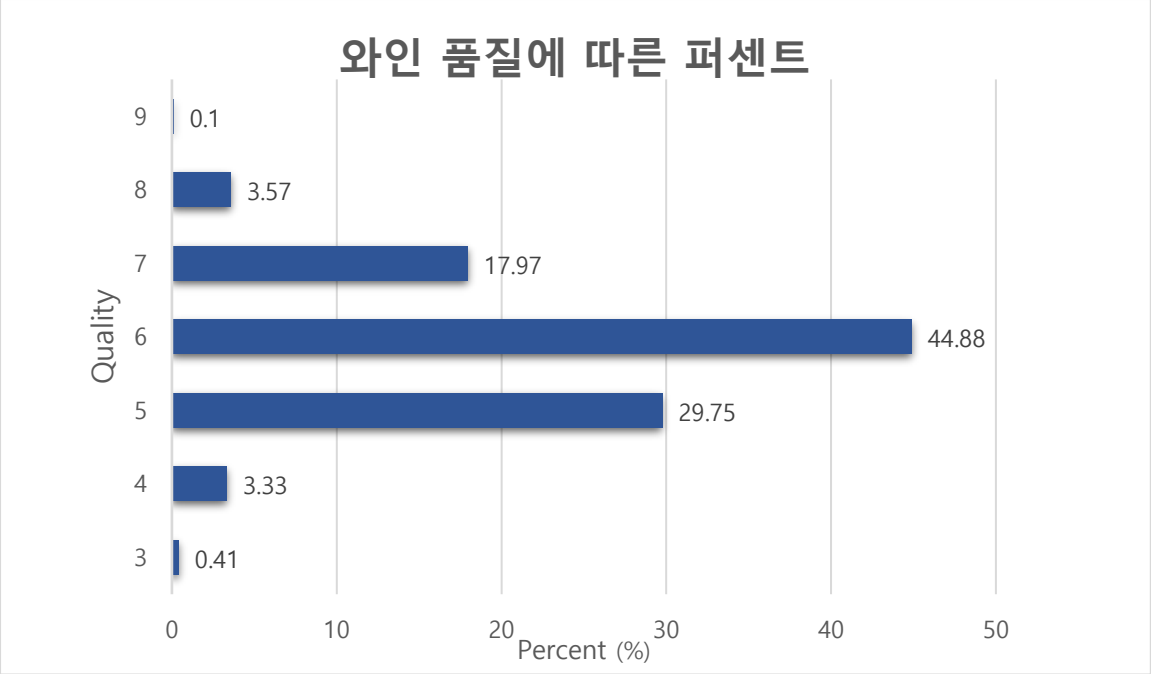

그림 6 품질에 따른 와인 퍼센트

위 그림은 기존 데이터셋에서 품질에 따른 와인의 퍼센트(%)를 도식화하였다. 품질 6을 기준으로 양쪽 품질에 대하여 빈도수가 감소하는 그래프를 확인할 수 있었다. 품질 별로 분석을 생각했었다. 하지만 와인 데이터의 문제점은 품질 별로 빈도수의 편차가 난다는 것이다. 데이터의 양 편차에 대해서 Up-sampling을 고려해 보았으나 원본 데이터를 증폭방법은 데이터의 왜곡이 우려되었다. 따라서 우리는 품질 별로 상, 중, 하로 순서형 범주를 통해 분석하고자 한다. 그에 따라 가장 데이터가 많은 품질 6을 품질 '중' 으로 설정하고 각각 위 아래를 상, 하로 분류하였다.

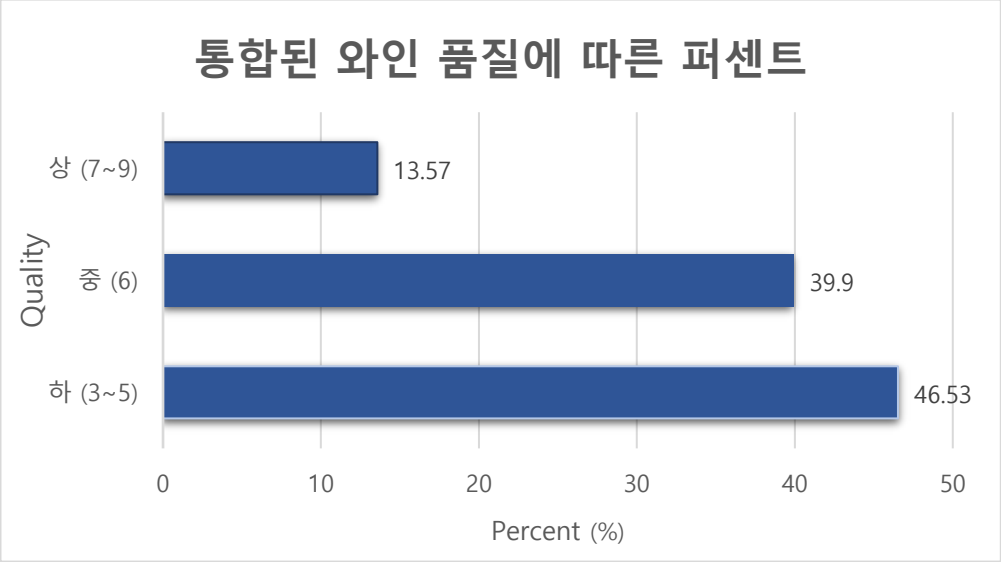

그림 7 통합된 와인 품질에 따른 퍼센트

품질 데이터를 통합함에 따라 단순 통계량을 다시 정리하였다.

표 5 설명변수들의 단순 통계량

단순 통계량						
변수	N	평균	표준편차	합	최솟값	최댓값
Quality (상:3, 중:2, 하 1)	4898	1.88158	0.73303	9216	1	3
fixed_acidity	4898	6.85479	0.84387	33575	3.8	14.2
volatile_acidity	4898	0.27824	0.10079	1363	0.08	1.1
citric_acid	4898	0.33419	0.12102	1637	0	1.66
residual_sugar	4898	6.39141	5.07206	31305	0.6	65.8
chlorides	4898	0.04577	0.02185	224.193	0.009	0.346
free_sulfur_dioxide	4898	35.30808	17.00714	172939	2	289
total_sulfur_dioxide	4898	138.36066	42.49806	677691	9	440
density	4898	0.99403	0.00299	4869	0.98711	1.03898
pH	4898	3.18827	0.151	15616	2.72	3.82
sulphates	4898	0.48985	0.11413	2399	0.22	1.08
alcohol	4898	10.51427	1.23062	51499	8	14.2

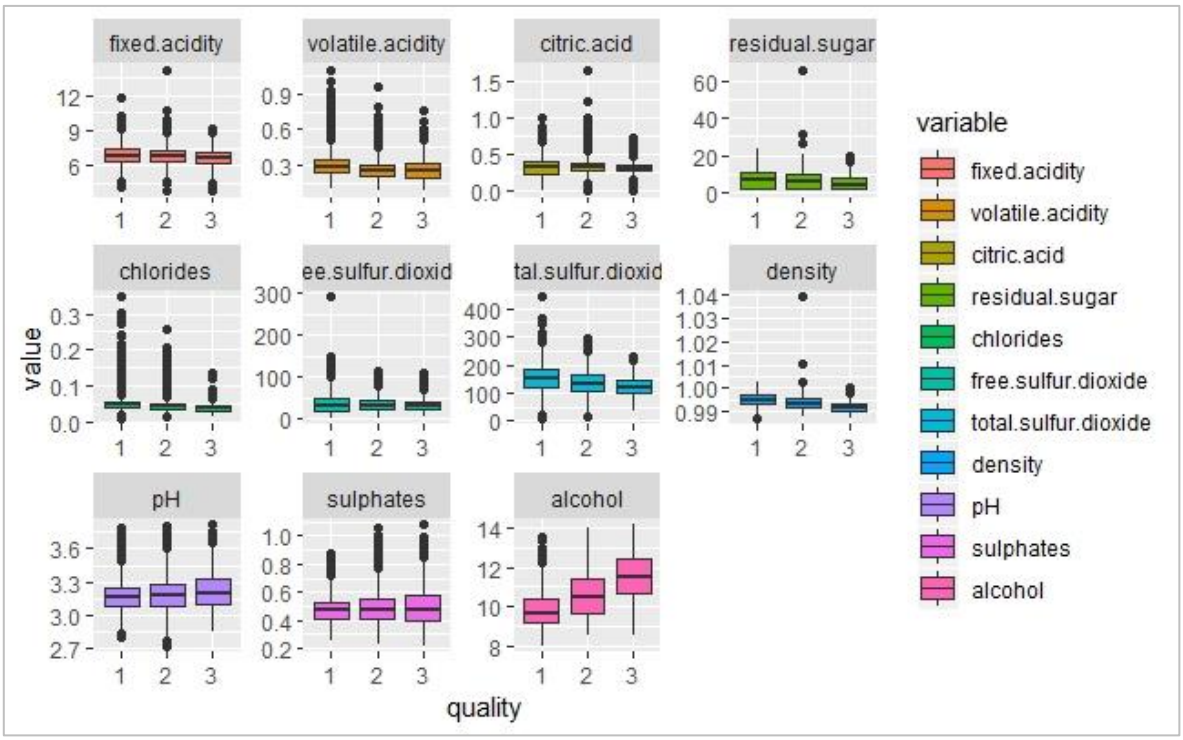

그림 8 quality(상:3, 중:2 하:1)에 대한 설명변수의 그림상자

위 요약 통계량과 그림상자를 보면 품질이 증가할 수록(quality가 1에서 3으로 증가할 수록) pH와 alcohol이 양의 상관관계를 같는 것처럼 보인다. 또한 volatile acidity와 citric acid, total sulfur dioxide, density가 음의 상관관계를 갖는 것처럼 보인다. 위의 탐색적 자료분석을 바탕으로 예측 모델링을 통해 위에 기술한 설명변수들이 정말 품질에 상관이 있는지 살펴보자.

나. 예측 모델링

① Multiple Regression (MR)

q개의 예측변수 X_1, X_2, X_q 와 하나의 반응변수 Y에 관한 다음 형식의 다중선형 회귀모형인 $y_i = \beta_0 + \sum_{j=1}^{q} (\beta_j x_{ij}) + \varepsilon_i$ 식에 데이터를 적용시켰다. 전체 11개의 예측변수의 가능한 예측변수를 모두 포함하는 모형을 고려하는 방법으로, 모든 예측변수가 포함된 완전모형에서, VIF 및 계수가 0이 아닌 변수선택방법을 이용하여, 예측변수들의 일부분만 포함하는 부분모형을 데이터에 적용시켜 예측변수를 찾아보려 한다.₩

Model Ⅰ : All predictors in the model

: 주어진 데이터에 대해 귀무가설을 '각각의 변수들이 영향을 미치지 않는다'로 가정하고, 대립가설을 '적어도 하나는 0이 아니다' 라고 가정하고 다중회귀분석을 해보았다. 또한, 모든 예측변수들(11개)를 최소제곱추정값을 이용하여 가능한 모든 회귀방정식들의 평가를 하였다.

아래 표는 모든 변수들을 넣고 모델 및 계수들의 유의성에 대해 검정을 해본 결과이다.

표 6 가능한 모든 회귀계수들의 평가

Regression Coefficients	Estimate	Std.Error	t-value	pr(>|t|)	VIF
(Intercept)	96.990536	17.917659	5.41	<0.001	
fixed.acidity	0.062933	0.020866	3.02	0.0026	2.5981
volatile.acidity	-1.480035	0.112539	-13.15	<0.001	1.1481
citric.acid	-0.102578	0.094276	-1.09	0.2766	1.164
residual.sugar	0.058221	0.007303	7.97	<0.001	12.1609
chlorides	-0.319585	0.55248	-0.58	0.563	1.2438
free.sulfur.dioxide	0.003481	0.000831	4.19	<0.001	1.7785
total.sulfur.dioxide	-0.000771	0.000373	-2.06	0.0391	2.2247
density	-99.933556	18.190708	-5.49	<0.001	26.5658
pH	0.483218	0.104254	4.64	<0.001	2.1998
sulphates	0.673364	0.099984	6.73	<0.001	1.1491
alcohol	0.190142	0.023301	8.16	<0.001	7.1589

Residual standard error: 0.623 on 3416 degrees of freedom
Multiple R-squared: 0.281, Adjusted R-squared: 0.279
F-statistic: 121 on 11 and 3416 DF, p-value: <0.0000000000000002

모형을 검정해 본 결과, 모형에 대한 유의성은 좋게 나왔으나, 다중공선성의 측도인 VIF를 확인해본 결과, Density의 변수가 너무 높게 나온 것을 확인 할 수 있다. 교호작용의 영향이 있으므로, 모형이 적합하지 않다. 이 모형의 VIF를 개선하고자 density를 제거하고 새로운 모델로 분석해보았다.

Model Ⅱ : After removal of density VIFs improved

: 변수 density를 제거한 후, Model Ⅰ에서 적용한같은 방법으로 계수들의 유의성검정을 해본 결과로 아래와 같이 나왔다.

표 7 변수 density를 제외한 부분 회귀계수의 평가

Regression Coefficients	Estimate	Std.Error	t-value	pr(>|t|)	VIF
(Intercept)	-1.424895	0.344925	-4.13	<0.001	
fixed.acidity	-0.016942	0.01503	-1.13	0.2597	1.3366
volatile.acidity	-1.56153	0.112032	-13.94	<0.001	1.1282
citric.acid	-0.14173	0.094407	-1.5	0.1334	1.1574
residual.sugar	0.020517	0.002508	8.18	<0.001	1.422
chlorides	-0.821435	0.547197	-1.5	0.1334	1.2098
free.sulfur.dioxide	0.004137	0.000826	5.01	<0.001	1.7419
total.sulfur.dioxide	-0.001123	0.000369	-3.04	0.0024	2.159
pH	0.121577	0.081187	1.5	0.1344	1.3228
sulphates	0.523759	0.096614	5.42	<0.001	1.0639
alcohol	0.302139	0.011332	26.66	<0.001	1.6789

Residual standard error: 0.625 on 3417 degrees of freedom
Multiple R-squared: 0.274, Adjusted R-squared: 0.272
F-statistic: 129 on 10 and 3417 DF, p-value: <0.0000000000000002

변수 density를 제거하고나서, 다중공선성(VIF)가 개선된 것을 보아, 유의하지 않은 변수들을 제거하고, 모델을 만들어보는 것이 좋을 것이라고 생각되어, Akaike정보기준인 AIC를 이용하여, 변수선택방법인 전진선택법(forward selection), 후진제거법(backward elimination), 단계별 방법(stepwise method)를 사용해 유의한 변수선택을 해보았다. 그 결과 중, 대표적으로 전진선택법(forward method)을 이용하여 분석한 내용에 대한 최종결과내용을 아래와 같이 표로 나타냈다.

표 8 전진선택법(forward selection)에 의한 최종 변수

```
Step:  AIC=-3208.1
TARGET ~ alcohol + volatile.acidity + residual.sugar + sulphates +
    free.sulfur.dioxide + total.sulfur.dioxide + pH + citric.acid
```

표 9 최종회귀식에서 제거된 예측변수

	Df	Sum of Sq	RSS	AIC
<none>			1338	-3208
+ chlorides	1	0.780	1337	-3208
+ fixed.acidity	1	0.396	1337	-3207

Model Ⅲ: Working Model

Model Ⅱ에서 전진선택방법(forward selection)에 의해 선택된 모형으로, 선택된 변수들의 유의성 검정을 해보면, 다음과 같다.

표 10 전진선택방법으로 선택된 회귀계수들의 평가

Regression Coefficients	Estimate	Std.Error	t-value	pr(>|t|)	VIF
(Intercept)	-1.755047	0.268393	-6.54	<0.001	
alcohol	0.308448	0.010637	29	<0.001	1.4787
volatile.acidity	-1.576463	0.111243	-14.17	<0.001	1.1119
residual.sugar	0.021065	0.002487	8.47	<0.001	1.398
sulphates	0.520169	0.096611	5.38	<0.001	1.0634
free.sulfur.dioxide	0.004289	0.000816	5.26	<0.001	1.6975
total.sulfur.dioxide	-0.001197	0.000366	-3.27	0.0011	2.1169
pH	0.162648	0.07471	2.18	0.0295	1.1197
citric.acid	-0.1823	0.091016	-2	0.0453	1.0753

```
Residual standard error: 0.625 on 3419 degrees of freedom
Multiple R-squared:  0.274,     Adjusted R-squared:  0.272
F-statistic:  161 on 8 and 3419 DF,  p-value: <0.0000000000000002
```

Train데이터의 Residual Standard Error는 자유도 3419에서 0.625였고, R^2의 값은 0.274 adj-R^2의 값은 0.272로 나타났다. 먼저, 연속형 종속변수로 Quality 범주가 1~9인 데이터에 대해 똑같은 분석을 해봤을 때, Train데이터의 R^2의 값은 0.256, adj-R^2의 값은 0.255였다. 이 값을 통해, quality에 대해 세분류의 집단으로 나눠 분석한 값이 더 나은 결과를 얻을 수 있었다. 허나, R^2에 대해 낮은 값이 나와서 회귀분석의 진단을 통해 좀 더 확인해보기 위해, Student residual, Cook's D, DFFITS를 통해, 잔차 결과에 대해서 확인해보았고, 시각적으로 그려본 그림은 다음 아래와 같다

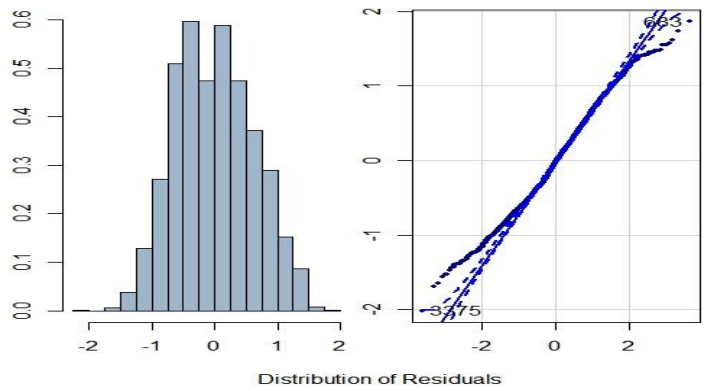

그림 9 잔차의 히스토그램 및 Q-Q Plot

먼저, 잔차에 대한 히스토그램과, Q-QPlot 을 그려서 정규성을 따르는지 확인해보았다. 잔차는 아래와 같이 대칭분포를 가지고 있고, 양끝에 약간의 이상치가 있는 것으로 보인다. 허나, [-2,2]사이에서 잔차가 있음을 확인할 수 있었고, 종속변수가 실제로는 정수값을 취하지만, 연속적이라고 가정되었기 때문에 이러한 패턴을 보이는 것으로, 어느 정도 정규성에 따른다고 할 수 있다. 더 자세히 살펴보기 위해, 각각의 변수에 대해 부분 잔차를 확인해보아, 얼마나 잔차 가정의 만족하는지 확인해보았다. 회귀계수에 대한 부분 잔차 그림을 그려보면, 다음과 같다.

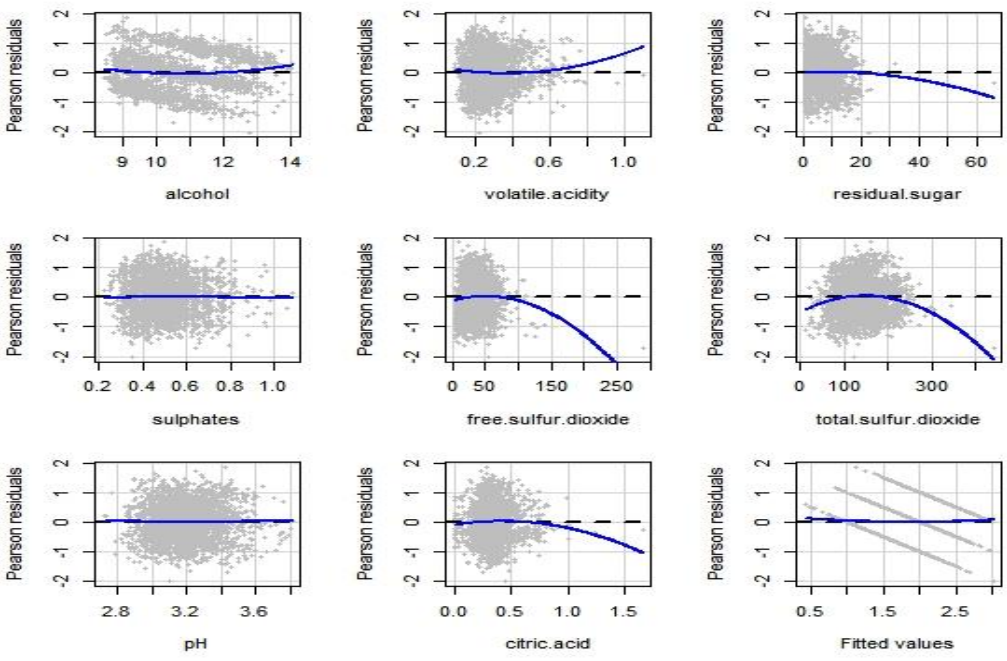

그림 10 MODELⅢ에 대한 변수들의 부분잔차

위 그림를 확인해보면, 오차에 대한 가정인 서로 독립이고 동일한 정규확률변수로서 평균 0과 공통분산을 가지는 가정을 만족하는 변수를 위 그림을 보고, 오차의 가정에 어느정도 부합한 Alcohol, volatile acidity, sulphates, pH로 선택하였다. 또한, Alcohol의 그림을 보면, 범주가 Quality의 범주를 세 그룹으로 나눈 영역에 꽤 비슷한 PLOT을 그림으로 확인할 수 있다. 이 부분에 대해 더 나아가 확실한 연관성에 대해 확인하기 위해서, 범주로 나눠, 순서형 로지스틱회귀모형을 통해 분석하면 좋을 것이라고 생각한다.

표 11 오차에 대한 가정 만족

alcohol	독립성	△	free.sulfur.dioxide	독립성	○
	정규성	○		정규성	X
	등분산성	△		등분산성	△
volatile.acidity	독립성	△	total.sulfur.dioxide	독립성	△
	정규성	△		정규성	X
	등분산성	○		등분산성	○
residual.sugar	독립성	○	pH	독립성	○
	정규성	△		정규성	○
	등분산성	△		등분산성	○
sulphates	독립성	○	citric.acid	독립성	△
	정규성	○		정규성	△
	등분산성	○		등분산성	△

범례: 만족, 절반 만족, 위험

오차에 대한 가정을 만족하는 것에 대해서, 좀 더 재밌는 접근으로 변수들의 부분잔차의 그림을 보고, 모형가정의 위반되는 정도에 대해 O, △, X 로 평가하여, 의미있는 변수를 선택하였다. 먼저, 오차항의 가정인 독립성은 일정한 패턴을 보이는가에 대한 측도로, 그림에 패턴이 있는가를 확인하였다. alcohol이 약간의 집단형태를 보였지만, 연속형 종속변수가 아닌 범주형 종속변수이기 때문에, 한계점에 대해서는 점수를 좀 더 줘서 평가하였다. 두번째로, 정규성은 파란색 부분이 Q-Qplot을 대신해서, residual값이 0으로 일직선일때, 정규성을 만족하는 부분으로, free sulfur dioxide, total sulfur dioxide는 정규성 부분에서 최저점으로 평가하였다. 그리고 마지막으로, 등분산성에 대해서는 데이터가 잘 퍼져있는지 확인하여 평가하였다.

ModelⅣ: Final Model

ModelⅢ로 선택된 변수들에 대해서 오차 및 유의성 검정을 통해, 최종 기준에 따라 모형을 아래와 같이 결정하였다.

표 12 최종모형 회귀계수들에 대한 평가

Regression Coefficients	Estimate	Std.Error	t-value	Pr(>|t|)	VIF
(Intercept)	-1.93432	0.25752	-7.51	<0.001	
alcohol	0.31609	0.00992	31.87	<0.001	1.2751
residual.sugar	0.02187	0.00241	9.08	<0.001	1.2995
volatile.acidity	-1.66983	0.10711	-15.59	<0.001	1.0222
sulphates	0.49067	0.09583	5.12	<0.001	1.0375
pH	0.18122	0.07365	2.46	<0.001	1.079

Residual standard error: 0.628 on 3422 degrees of freedom
Multiple R-squared: 0.267, Adjusted R-squared: 0.266
F-statistic: 249 on 5 and 3422 DF, p-value: <0.0000000000000002

최종 모형에 대한, Residual Standard Error는 자유도 3422에서 0.628였고, R2의 값은 0.267 adj-R2의 값은 0.266으로 나타났다. R2의 값은 낮게 나왔으나, 이 점은 종속변수가 순서형 범주이기 때문에 예측에 대한 정확도측도가 많이 낮다는 한계점이 있었기 때문에, 모형 및 예측변수에 대한 유의성과 공분산 및 오차항에 대한 여러 검토를 거쳐 alcohol, residual sugar, volatile acidity, sulphates, pH가 와인품질집단에 대해 영향이 있었다고 할 수 있다.

또한, 아래의 그래프는 Quality를 세분류로 나누기 전과 후의 오차의 그래프를 비교해보았다.

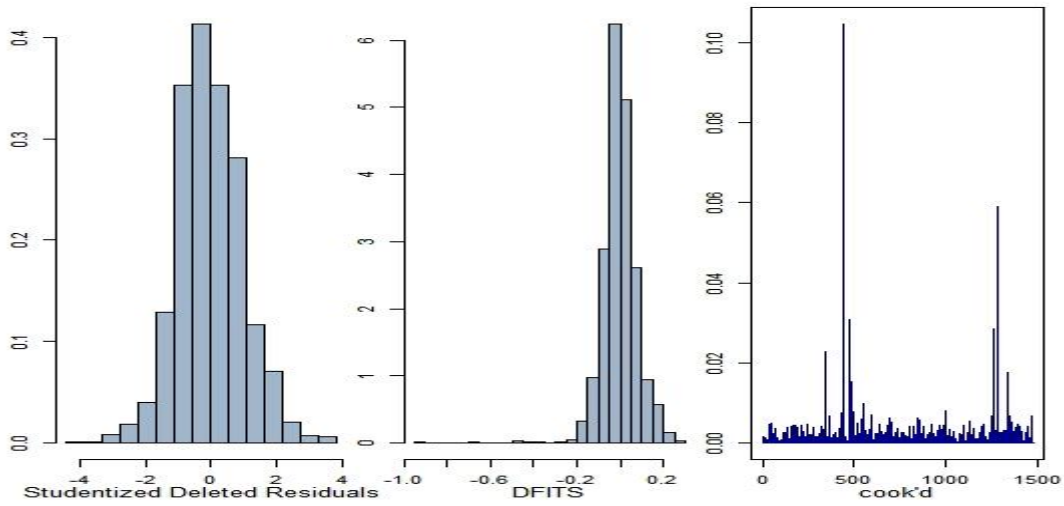

그림 11 Quality의 범주가 3~9에 해당하는 오차

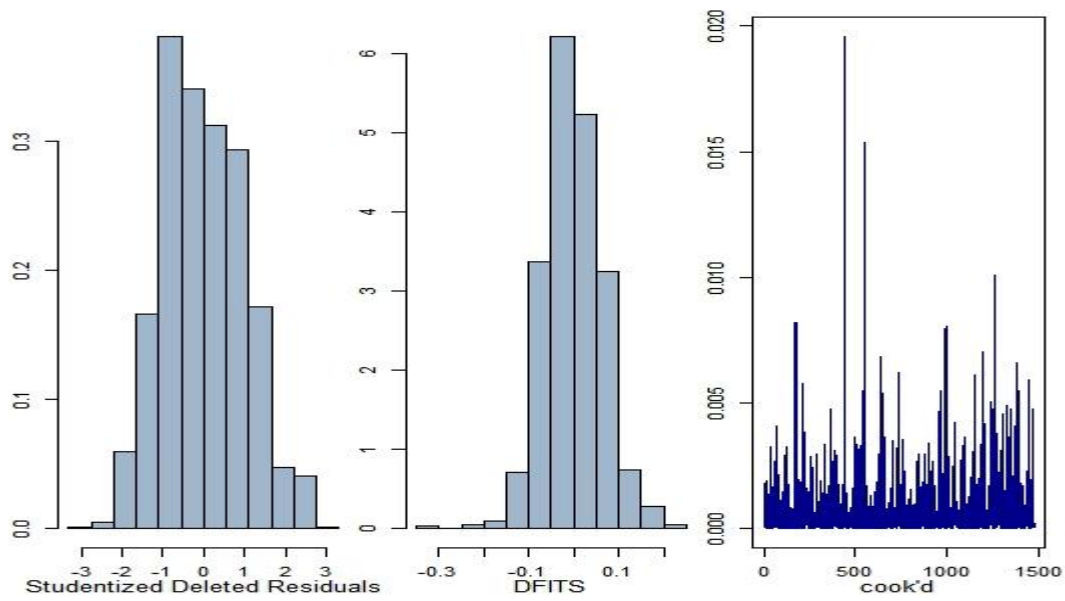

그림 12 Quality의 범주가 상중하에 해당하는 오차

위 결과를 볼 때, 잔차의 범주가 Quality가 1~9로 이뤄졌을 때 보다, Quality의 범주를 상중하로 나눠 모형을 설정하는 것이, 정규성에 더 가까운 모형을 만드는 것이 더 적합하다는 결과를 얻게 되었다.

② Decision Tree Model (DTM)

의사결정나무 모델은 분류하는 규칙의 기준을 트리구조로 만들어 복합적으로 분류를 해주면서도 설명력이 좋은 모델이다. 이 모델을 최대 깊이를 5로 해서 적용해보았고 이 때 분류 규칙은 뒤의 부록**(그림 19 의사결정나무 1/4)**에 시각화 했다. 이 때 사용된 노드들을 토대로 중요도를 분석해 본 결과 다음의 변수들이 분류에 중요했다.

표 13 의사결정나무 모델을 이용한 각 변수의 중요도

alcohol	0.452
volatile acidity	0.270
free sulfur dioxide	0.0887
fixed acidity	0.041
residual sugar	0.038
pH	0.030
chlorides	0.025
citric acid	0.017
sulphates	0.016
density	0.011
total sulfur dioxide	0.010

이 모델을 활용한 예측의 성능은 다음과 같았다.

표 14 의사결정나무 모델 분류 성능표

	precision	recall	f1-score	support
1	0.64	0.67	0.65	492
2	0.56	0.69	0.62	660
3	0.70	0.30	0.42	318
avg / total	0.62	0.60	0.59	1470

단순한 트리형태로 분류한 것임에도 높은 성능을 보였기에 앙상블 방법으로 트리 모델을 보완한 랜덤포레스트를 활용해보았다. 트리 모델을 1000개를 만들어 훈련시켰으며, 위의 Decision Tree에서 한 것과 같은 중요도를 분석해 본 결과 비교적 모든 설명변수가 분류에 많이 활용되었음을 알게 되었다.

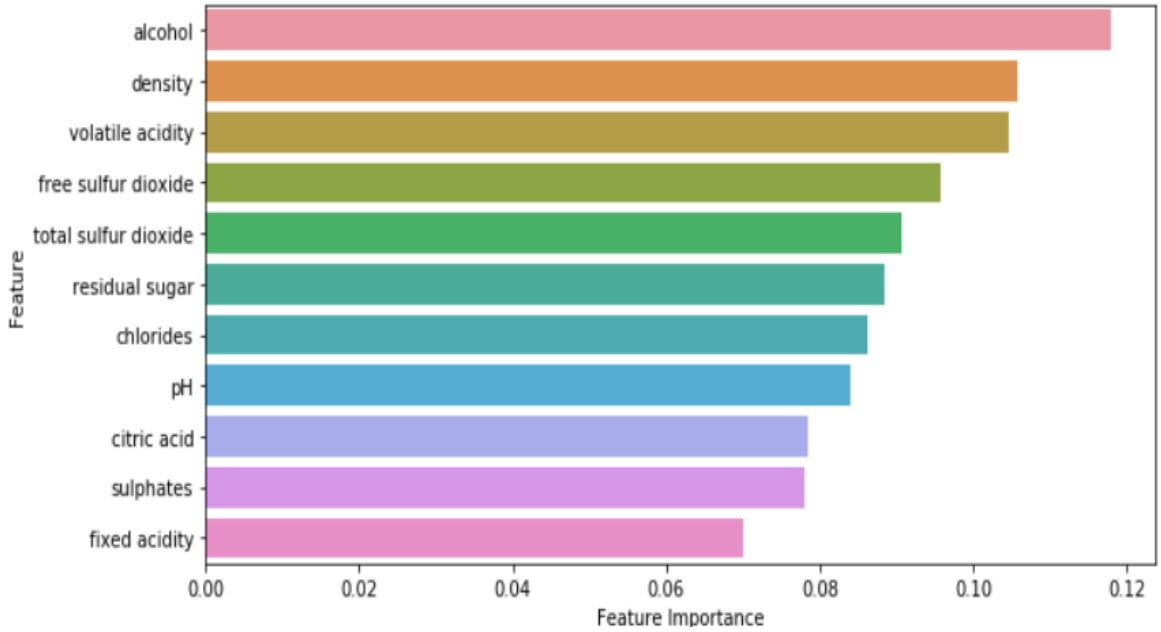

그림 13 랜덤포레스트 모델을 이용한 각 변수의 중요도

이 모델의 예측 결과 다음과 같이 더 좋은 성능을 보였다.

표 15 랜덤포레스트 모델 분류 성능표

	precision	recall	f1-score	support
1	0.77	0.76	0.76	492
2	0.68	0.76	0.72	660
3	0.76	0.60	0.67	318
avg / total	0.73	0.72	0.72	1470

이 모델이 어떻게 예측했고 그 실제 값이 어떤 값이었는지 알기 위해 히트맵 매트릭스를 그려보았다. 이 시각화를 통해 "상" 품질에 해당하는 와인에 대해 "중" 품질로 예측한 경우가 조금 많았음을 확인할 수 있었지만, 대체적으로 잘 분류했다는 것을 알 수 있었다.

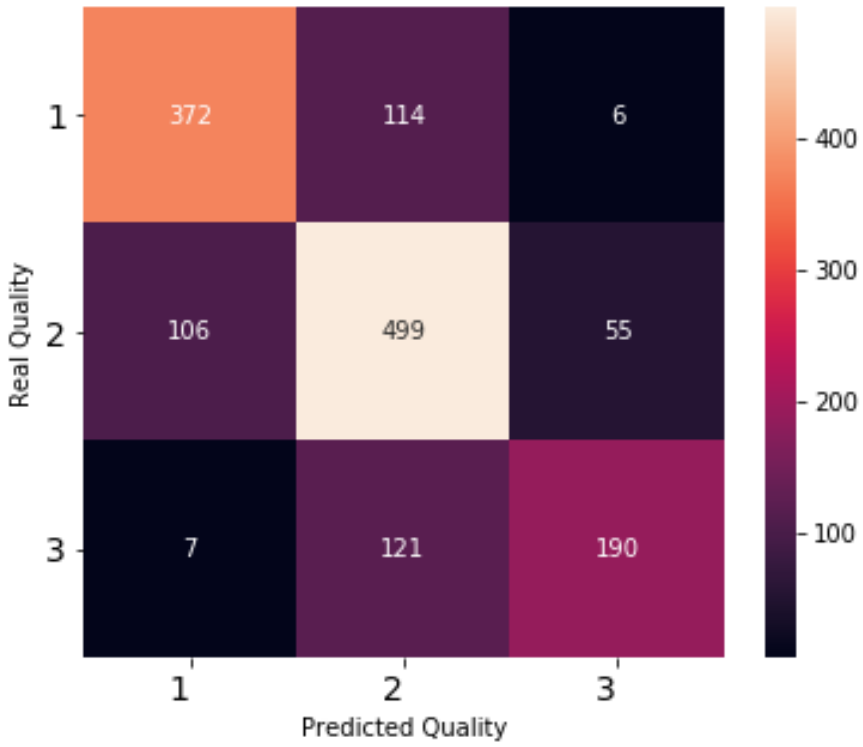

그림 14 랜덤포레스트 분류 매트릭스

③ Support Vector Machine (SVM)

Model I : Linear Kernel

선형적으로 데이터를 분류하는 Linear Kernel을 이용한 SVM의 분류 예측 결과는 아래와 같다.

표 16 Linear Kernel SVM 분류 성능표

	precision	recall	f1-score	support
1	0.64	0.58	0.61	492
2	0.50	0.78	0.61	660
3	0.00	0.00	0.00	318
avg / total	0.44	0.54	0.48	1470

위 결과를 통해 가장 좋은 '상' 종류 와인에 대한 예측이 매우 좋지 않음을 알 수 있었다. 이런 결과가 Linear Kernel의 선형성에 있을 것이라고 보고 더 개선하기 위해 Gaussian Kernel을 적용했다. 비록 Linear Kernel을 활용한 SVM의 성능이 좋지는 않았지만 각 변수에 대한 계수를 구할 수 있어 그 계수들의 절대값을 활용해 더 중요하게 사용된 변수를 알 수 있었다.

표 17 Linear Kernel SVM 모델을 이용한 각 변수의 중요도

volatile acidity	9.604
sulphates	3.314
chlorides	2.110
alcohol	1.861
pH	0.863
citric acid	0.430
density	0.242
fixed acidity	0.241
residual sugar	0.144
free sulfur dioxide	0.028
total sulfur dioxide	0.010

Model Ⅱ: Gaussian Kernel

Gaussian Kernel을 활용한 SVM은 비선형적으로 분류한다는 강점을 가지고 있다. 이렇게 분류한 결과의 성능은 다음과 같았다.

표 18 Gaussian Kernel SVM 분류 성능표

	precision	recall	f1-score	support
1	0.63	0.53	0.58	492
2	0.55	0.73	0.63	660
3	0.65	0.39	0.48	318
avg / total	0.60	0.59	0.58	1470

이전 모델에 비해 3번 집단을 잘 분류했고 그에 따라 전체적인 예측성능도 증가한 것을 알 수 있다. 또한 실제 값과 예측된 값에 대한 매트릭스를 만들어 히트맵 형태로 시각화 해 보았고 그 결과 이 모델에서 와인의 '중' 으로 예측하는 경향성이 있다는 것을 알게 되었다.

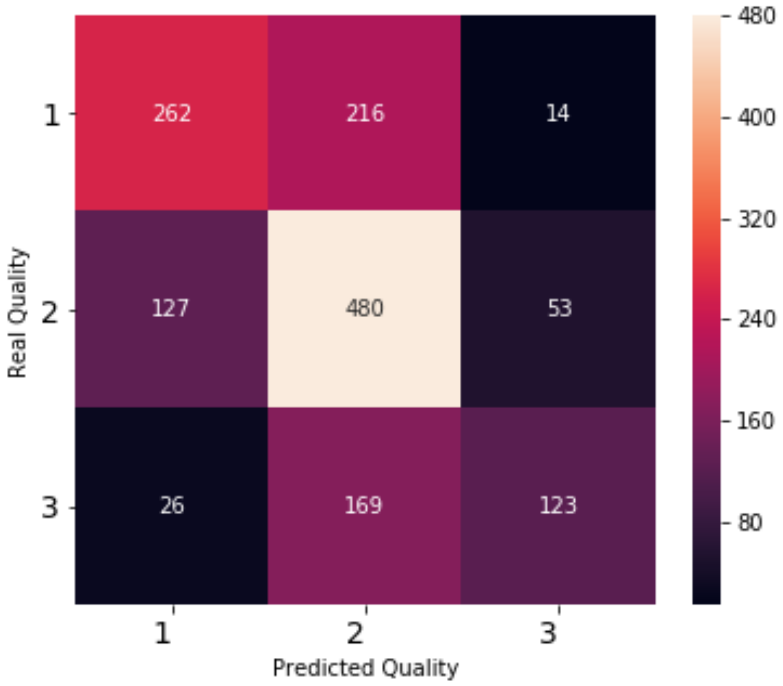

그림 15 Gaussian Kernel SVM 분류 매트릭스

이런 문제를 보완하고자 두가지 시도를 했다. 첫째는 SVM 모델이 scale 차이에 민감하다는 것을 고려해 Min-Max 정규화를 했다. 아래의 그림은 정규화한 데이터의 일부이다.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
0	0.173077	0.107843	0.156627	0.012270	0.062315	0.034843	0.155452	0.065356	0.481818	0.174419	0.467742
1	0.288462	0.215686	0.210843	0.033742	0.086053	0.027875	0.359629	0.078851	0.336364	0.360465	0.645161
2	0.298077	0.313725	0.180723	0.153374	0.071217	0.076655	0.180974	0.106805	0.390909	0.267442	0.774194
3	0.355769	0.333333	0.084337	0.154908	0.109792	0.055749	0.199536	0.169462	0.454545	0.127907	0.435484
4	0.298077	0.294118	0.174699	0.200153	0.115727	0.174216	0.417633	0.206863	0.254545	0.441860	0.241935

그림 16 정규화한 White wine 데이터

둘째는 상대적으로 적은 "하"와 "상" 품질의 와인의 영향력을 늘리기 위해 Gaussian Kernel SVM에서 경계에 영향을 주는 C 값과 gamma 값을 조정하여 각 데이터 포인트의 영향 범위를 넓혔다. 아래 그림은 C 값과 gamma 값의 조정의 영향력에 대해 표현한 그림이다.

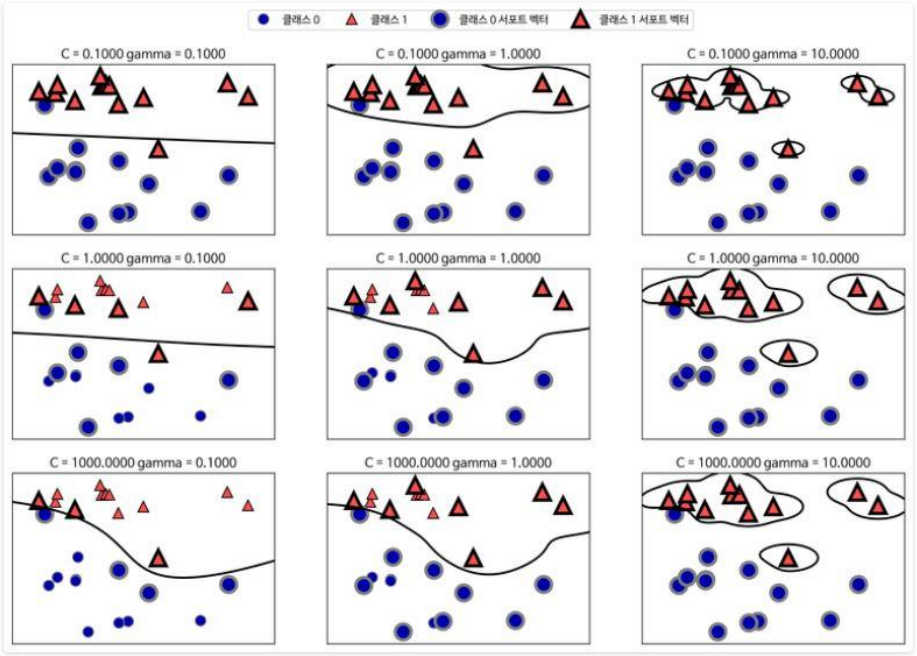

그림 17 C와 gamma 파라미터 조정에 따른 차이

이렇게 개선한 모델은 더 좋은 성능을 보였다.

표 19 개선한 Gaussian Kernel SVM 분류 성능표

	precision	recall	f1-score	support
1	0.66	0.72	0.69	492
2	0.62	0.64	0.63	660
3	0.67	0.54	0.60	318
avg / total	0.65	0.65	0.64	1470

이렇게 개선한 모델을 다시 분류 매트릭스로 만들었을 때 이전 모델보다 "중"으로 예측하는 경향이 덜해졌다. 이렇게 제약을 수정하는 방법을 통해 우리가 원하는 방향으로 모델을 만들기는 했지만, 대신에 실제로 "중"일 때 "중"이 아니라고 예측한 경우가 늘어났다. 다만 이 경우는 trade-off라는 성질을 고려했을 때, 전체적 성능이 개선되었기 때문에 괜찮다고 볼 수 있다.

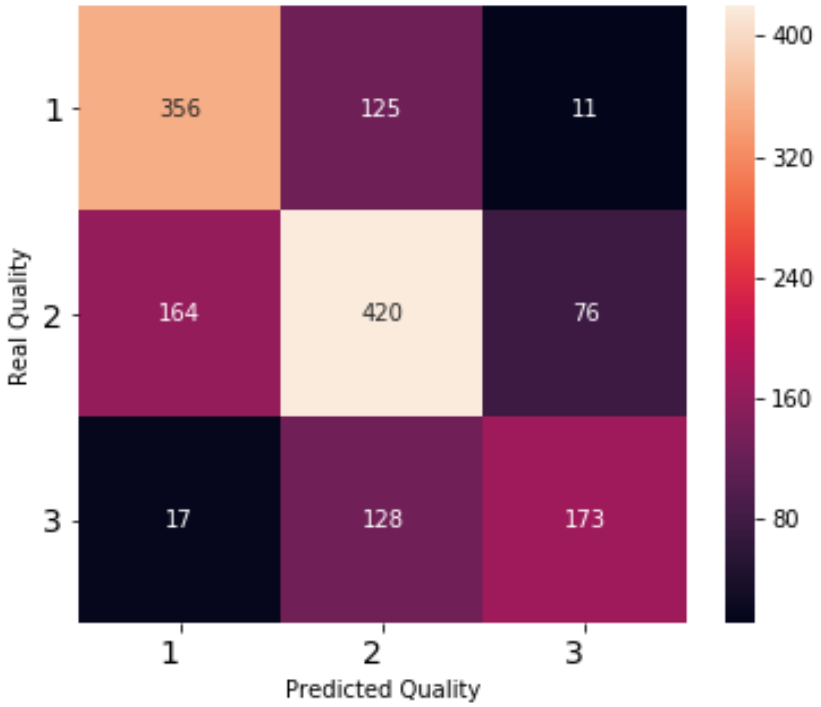

그림 18 개선한 Gaussian Kernel SVM 분류 매트릭스

결 론

1. 분석 모형들의 한계

다중 회귀분석의 한계

다중 회귀분석에서는 공선성 문제를 VIF가 높은 값을 지닌 변수를 제거하여 모형을 개선하고, AIC를 이용하여 변수를 선택하였으며, 더 나아가 잔차에 대한 진단을 통해 최종적인 모형을 만들었다. 하지만 모형의 영향력의 측도인 R^2의 값이 현저하게 낮았다는 한계가 있었다. 이 한계점에 대해 살펴볼 때, 종속변수가 와인의 품질이 연속형 변수로 나타내어 있지만, 실제적으로는 범주형 변수이기 때문에, 와인품질에 대해 설명력이 있는 모형이 되기 위해서는 로지스틱회귀분석을 통해 분석해야 하는 점이 명확한 한계로 나타났다.

의사결정나무의 한계

의사결정나무에서는 연속형 변수를 비연속적인 값으로 취급하기 때문에 분리의 경계점 근방에서는 예측 오류가 클 가능성이 있으며 선형성 또는 주 효과의 결여가 일어날 수 있다. 또한, 분석용 자료에만 의존하는 의사결정나무는 새로운 자료 예측에서 불안전할 가능성이 높다. 화이트와인 데이터를 바탕으로 와인의 품질 데이터를 예측하고자 한다면 이점을 분명히 구분해야 할 것이다. 마지막으로 의사결정나무는 깊이가 깊어질수록 오버피팅이 되는 문제가 있다. 따라서 랜덤포레스트를 통해 여러 개의 트리를 만들어 오버피팅의 문제에 대비하였다.

서포트 벡터 머신의 한계

서포트 벡터 머신은 변수 선택 및 스케일 차이에 민감하며, 위에서 다룬 C, gamma 등의 값을 직접 조절해야 한다는 한계가 있다. 스케일 차이는 Min-Max 정규화를 통해 해결하였다. 나머지 한계에 대해서는 더 좋은 방법이 경우가 있을 수 있지만 직접 모든 경우에 대해 적용해 본다는 것이 불가능하기 때문에 최적의 솔루션은 아닐 수 있다는 것을 인지해야 한다. 또한 Kernel 기법도 다양하게 존재하지만 제한된 시간에 모두 해볼 수 없었다. Gaussian Kernel에서는 변수 중요도를 파악하기가 어렵다는 점 때문에 Linear Kernel에서 나온 중요 변수로 대체했다는 점에서 명확한 한계가 있다.

2. 와인 선택에 영향을 미치는 요소

와인의 품질을 평가하는 데에는 다양한 방법론이 존재할 수 있다. 우리의 목표는 좋은 와인을 고르고자 할 때 무엇을 기준으로 정할지 안내하는 것이었고, 이 기준이 되는 설명변수의 선택을 위해 여러 방법으로 분석하고 각각 영향력이 높다고 평가된 변수들 중 다수의 모델에서 중복되게 나온 변수를 고르기로 했다. 본 분석은 다중회귀분석, 의사결정나무, SVM을 분류 기준으로 선정하였다. 그리고 결과적으로 세가지 모델 중 두가지 모델 이상에서 영향력을 가진 설명변수들과 그에 대한 특성은 다음과 같았다. 최종적으로 Alcohol, volatile acidity, Residual Sugar, pH, Free sulfur dioxide 5가지 변수들이 품질에 중요한 요소라고 결론 지으며 본 분석을 마무리하고자 한다.

표 20 최종결론: 설명변수 5가지

MR	SVM (Linear kernel)	DT (random forest)
Alcohol	Alcohol	Alcohol
Volatile acidity	Volatile acidity	Volatile acidity
Residual sugar	Chloride	Residual sugar
pH	pH	pH
Free sulfur dioxide	Sulphates	Free sulfur dioxide
		Fixed acidity

부 록

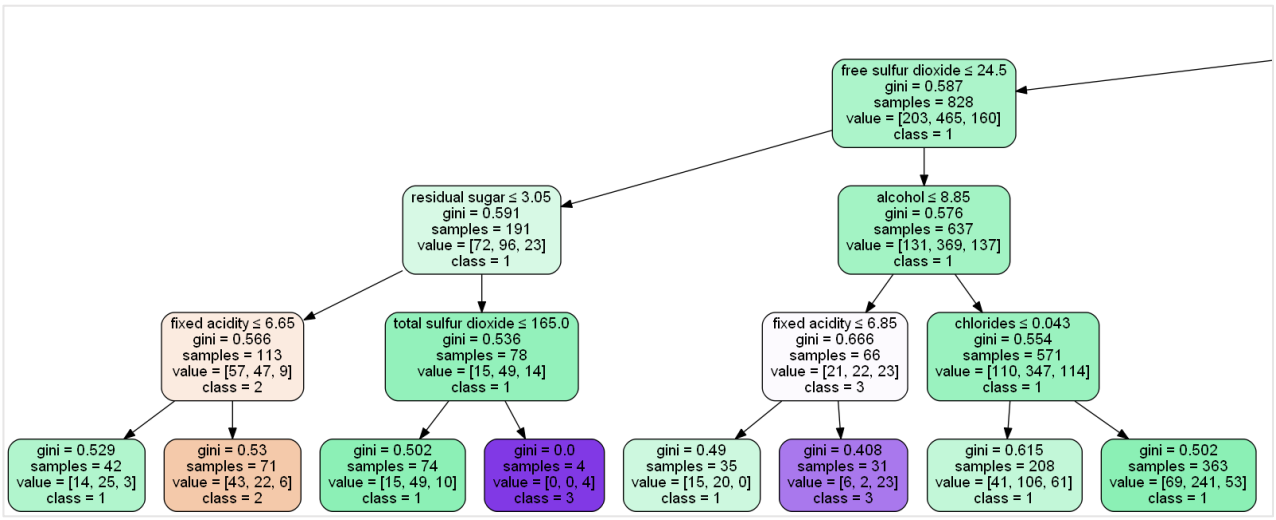

그림 19 의사결정나무 1/4

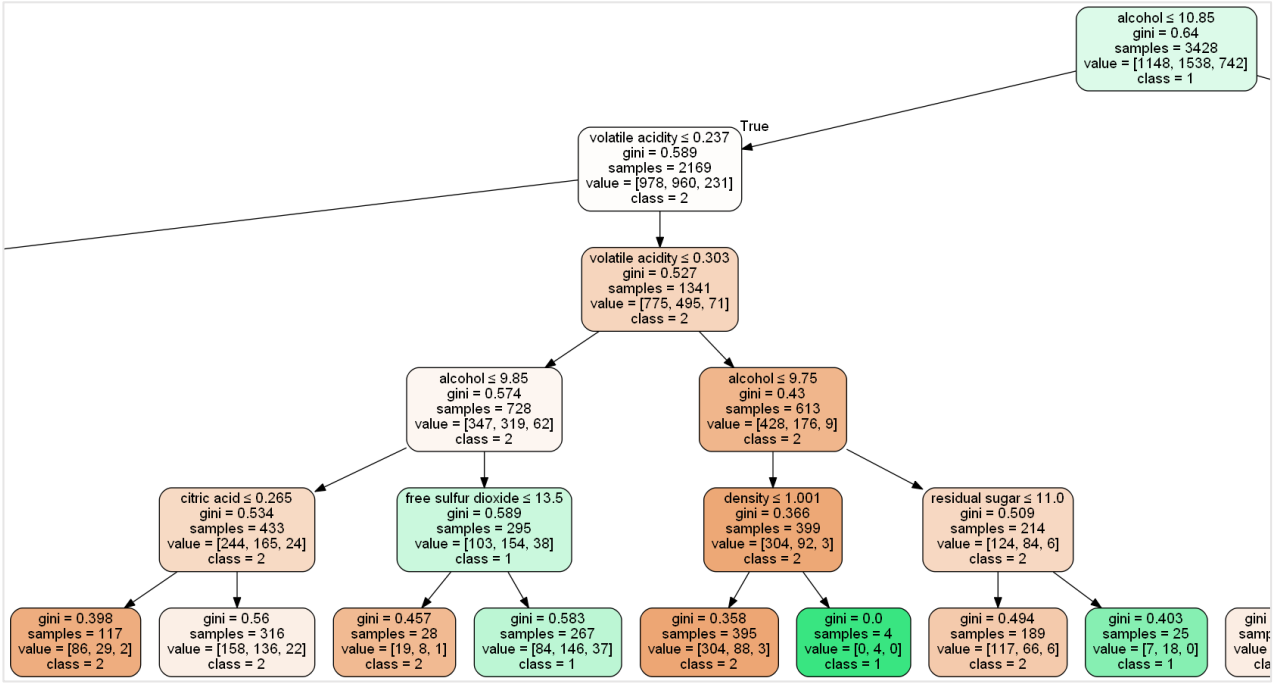

그림 20 의사결정나무 2/4

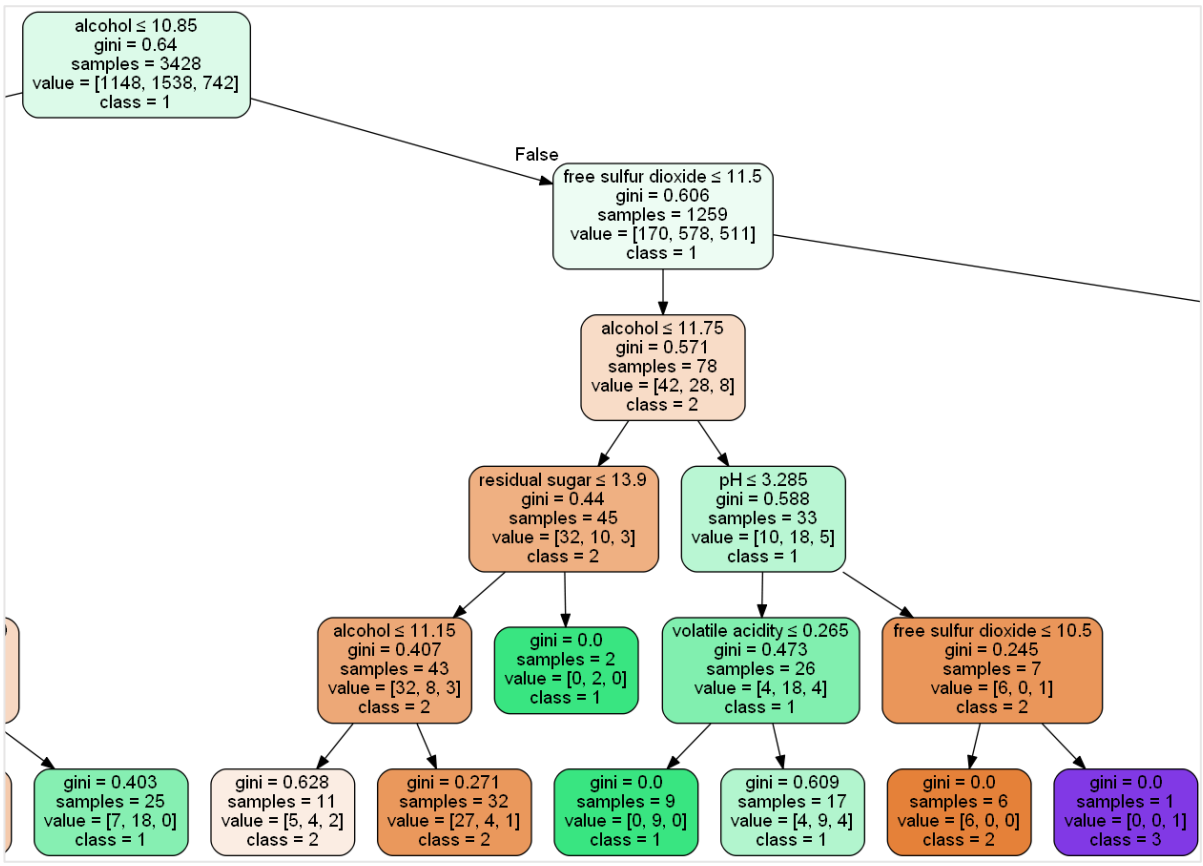

그림 21 의사결정나무 3/4

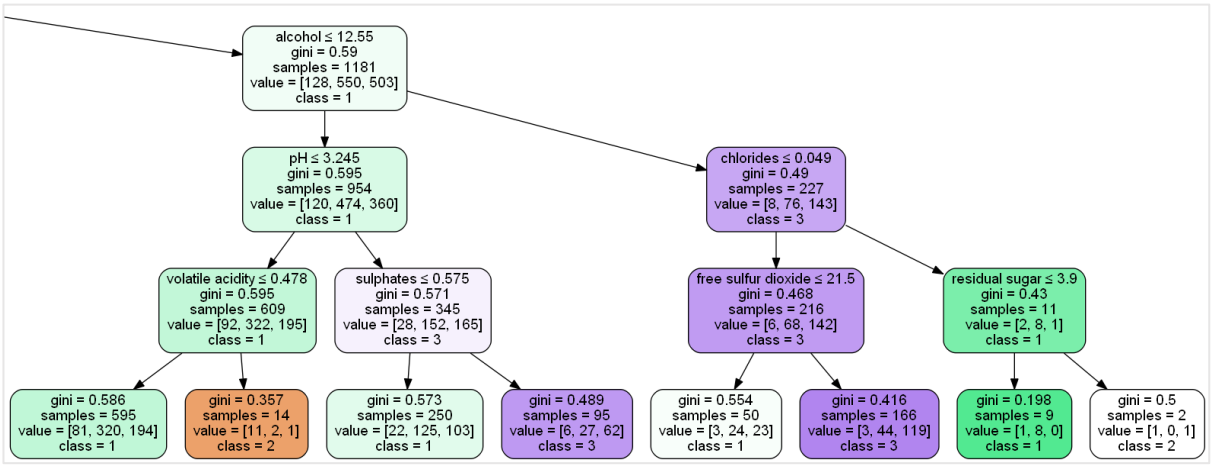

그림 22 의사결정나무 4/4

b