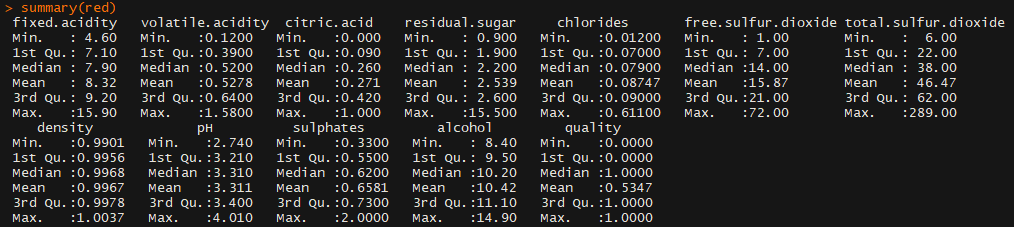
Red Wine 분석

1. 변수 설명
2. Fixed Acid : 불휘발성 산
3. Volatile Acid : 휘발성 산, 슬러지 소화에 있어서 복잡한 유기물이 혐기성 분해될 때의 중간생성물.
4. Citric Acid(구연산) : 감귤류에 속하는 과일에서 추출된 유기산(organic acid)이며 pH 조절을 위해 쓰인다. 산-알칼리 균형을 조절시키기 위해 일차적으로 사용되며 산화 방지제와 방부제의 성질을 약간 띤다.
5. Residual Sugar(잔당) : Residual Sugar는 사전적 의미로 “잔당, 발효 후 남은 당분, 첨가된 당분이 아닌 자연발효로 생성된 당분”이란 뜻이며, 와인용어로 “발효 후 와인 속에 남아있는 당분”을 말한다. 발효하는 동안 모든 당분이 이스트에 의해 알코올로 변하는 것이 자연스러우나, 알코올 제조자들은 와인의 달콤한 맛을 위해 인위적인 방법으로 발효를 중지시킨다.
6. Chlorides(염화물) : 염소와 염소보다 양성인 원소와의 화합물의 총칭
7. Free Sulfur Dioxide(이산화유황) : 이산화유황(二酸化硫黃). 이산화황. 무색의 불연(不燃)가스. SO2. 강하고 숨막히는 냄새가 있으며, 종종 조제 제품의 산화 방지제로 사용된다.
8. Total Sulfur Dioxide
9. Density : 밀도
10. pH : 산도
11. Sulphate(황산염) : 황산 분자에 들어 있는 수소 이온의 일부 또는 전부가 금속 이온 따위의 [양이온](https://terms.naver.com/entry.nhn?docId=2040819&ref=y)으로 [치환](https://terms.naver.com/entry.nhn?docId=2042293&ref=y)된 [화합물](https://terms.naver.com/entry.nhn?docId=2043180&ref=y)을 통틀어 말하며, 칼슘염, 바륨염, 납염을 제외하면 물에 잘 녹음.
12. Alcohol : 알코올 농도
13. Quality(Target) : 본디 1~10단계로 나누어져 있었으나, 해석의 용이함을 위하여 1~5단계는 0, 6~10단계는 1로 단순화.

Quality 제외한 모든 Input 변수들은 연속형임.

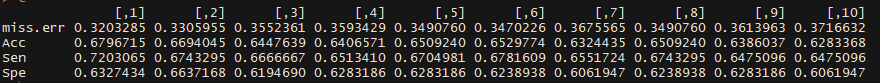


1. 기본 분석 방법

KNN, LDA/QDA, Logit Regression을 각각 사용하여 분석하였으며, 한 번은 Train-Test를 7:3으로 나누어 분석하였고, 한 번은 K-fold를 이용하여 모델을 평가하였다. Prediction Assess에는 AUC와 CV 에러를 사용하였다.

1. KNN
2. No Validation

K =1~10까지 넣어 Misclassification, Accuracy, Sensitivity, Specificity 계산



K=1 일 때 가장 좋은 정확도를 보임

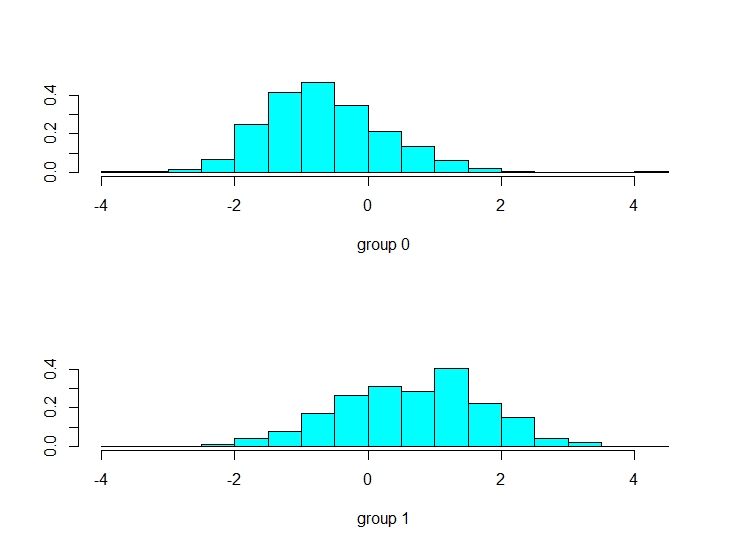
1. V-fold : V=10



K=1 일 때 역시 가장 좋은 정확도를 보임

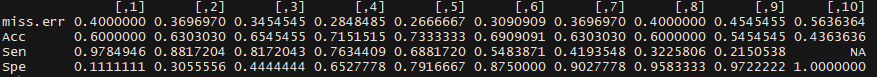
아마 KNN의 경우 변수의 개수가 많아질수록 그 예측력에 있어서 문제점을 보이는데, 이 데이터셋의 경우, p=11 이기 때문에 그리 좋지 않은 예측력을 보여주는 것으로 보임.

1. LDA / QDA
2. LDA – No validation



데이터는 어느 정도 잘 나뉘어 있음.

컷오프 0.1~1.0 까지 비교



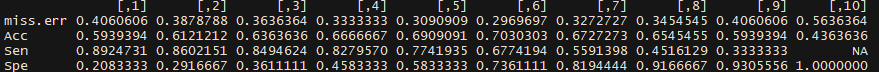
컷 오프 0.5에서 가장 좋은 정확도를 보여줌



Train AUC는 0.827, Test AUC는 0.770 임

1. QDA – No validation

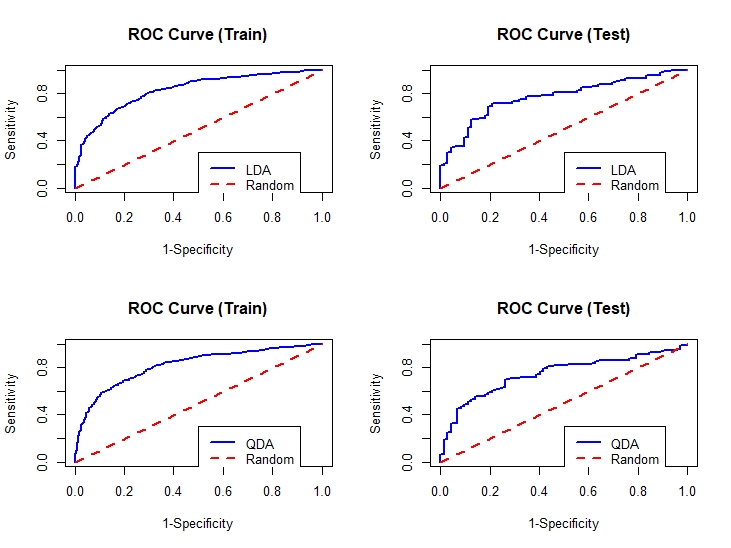
컷오프 0.1~1.0 까지 비교



컷 오프 0.6에서 가장 좋은 정확도를 보여줌



Train AUC는 0.816, Test AUC는 0.744 임.



1. V-fold – V=10

* LDA



Cut off 0.1~1.0으로 계산한 것인데, 0.5일 때 가장 작은 에러를 보인다.

* QDA

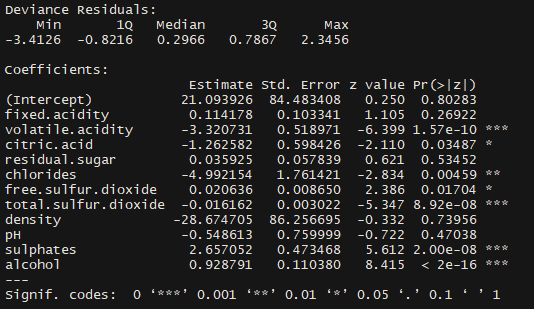


Cut off 0.1~1.0으로 계산한 것인데, 0.5일 때 가장 작은 에러를 보인다.

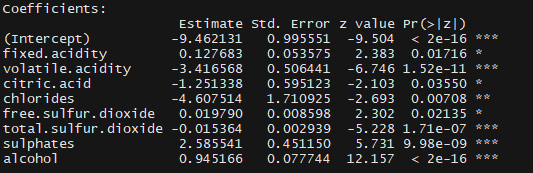
KNN보다는 적은 에러값을 보여준다.

1. Regression
2. 모델 설정

Logit Regression 모델을 사용한다.

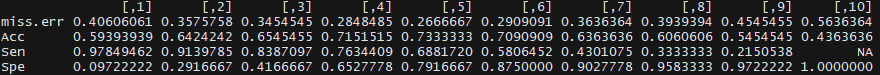


P-value가 유의하지 않은 변수들이 있기 때문에, 이것들을 stepwise로 제외한다.

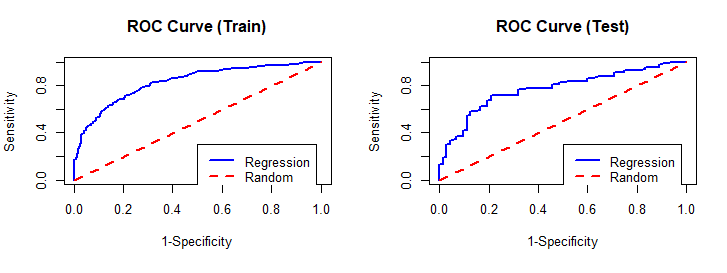


Residual sugar, density, pH 세 가지의 변수를 없앤 모델을 fitting한다.

1. No Validation



Cutoff=0.5 일 때, 가장 높은 정확도를 보여준다.





Train AUC는 0.827, Test AUC는 0.771 이다.

1. V-fold, V=10



Cutoff=0.5 일 때, 가장 작은 에러를 보여준다.

1. 종합

가장 좋은 성능을 보여준 것은 Logit Regression 모델이었고, 순서대로 LDA, QDA, KNN 이었다. 데이터가 어느 정도 잘 나뉘어져 있긴 하지만, 관측된 개체의 수가 크기 때문에 Regression이 가장 높은 성능을 보였고, Domain이 없는 상태에서, Red Wine의 Quality에 영향을 끼치는 변수를 골라내는 데에는 그나마 Logit Regression이 낫다고 할 수 있겠다.