Datamining Project Report

환자 정보를 이용한 간암 병기 예측 모델 비교

2019년 12월 16일

서울대학교

통계학과 고우진

통계학과 권용찬

통계학과 김민국

INDEX

1. Introduction

1.1 연구배경 및 목적

암이란 자기 기능을 잃은 채 끊임없이 증식하는 세포를 의미한다. 무한증식을 통하여 주변으로 퍼져나가는 것이 암의 특징이다. 간세포가 고유 기능을 잃은 채 암세포로 변하여 끊임없는 자기 증식을 통하여 암으로 발전하게 된 것이 간암이다. 간암은 한국인에게 가장 흔하게 발생하고 있다. 이는 간암의 발병원인 중 하나인 B형 간염 바이러스 보유율이 한국이 5~8%정도로 높기 때문이라고 생각할 수 있다. 2017년 통계청의 사망원인통계 결과에 따르면 간암은 폐암에 이은 암으로 인한 사망률 2위를 하였다. 40대와 50대에서는 간암으로 인한 사망률이 가장 높았다. 간암의 이상적인 치료 방법은 수술적 절제이다. 하지만 간암의 조기진단이 힘들며 조기진단을 하더라도 간경변증이 심할 경우 수술을 할 수 없다. 즉, 간암의 치료를 위해서는 조기진단의 필요성이 중요하다. 현재 혈액검사, 방사선학적 영상검사, 조직검사 등이 간암의 진단에 이용되고 있다. 영상검사는 주된 진단 방법으로 혈액검사는 단독으로 진단이 어려우나 간편하여 선별검사로 이용되며 영상진단의 보조적인 검사로 시행되고 있다. 정확한 진단을 하기 위해서는 조직검사가 필요하다. 하지만 조직검사는 피부와 조직의 절개가 동반되며 통증과 출혈로 인해 많은 환자들이 불편함을 호소하였다. 본 연구는 혈액검사를 통해 얻을 수 있는 간의 기능적인 지표들, 신장, 나이, 암의 가족력 등 환자로부터 얻을 수 있는 신상정보를 바탕으로 간암의 여부 및 간암의 기수를 분류할 수 있는 효율적인 방법을 찾는데 목적을 두고 있다. 간암의 진단 방법 중 가장 간편한 혈액검사를 통해 간암의 기수를 제대로 분류할 수 있다면 영상진단에 도움이 될 수 있으며 사람들이 불편함을 표현한 조직검사를 최소화하는 기대해볼 수 있다.

1.2 생물학 이론 배경

1.2.1 알부민(Albumin)

알부민은 혈청 단백질의 50~60%를 차지하고 있다. 간에서 만드는 단백질의 25%를 차지 하고 있는 대표적인 단백질이다. 알부민은 필수 단백질로 인체에서 합성이 되어야 하지만 간 기능이 좋지 않으면 간에서 알부민을 만들지 못해 혈청 알부민 농도가 낮아지게 된다. 따라서 알부민 농도는 간 기능을 판단하는 데 지표로 이용되고 있다.

1.2.2. 빌리루빈(Bilirubin)

황달이 생기는 중요한 원인 중 하나이다. 간에서 생성되는 담즙은 적혈구가 죽어 생긴 빌리루빈을 몸 밖으로 내보내는 역할을 한다. 간 기능이 좋지 않으면 담즙의 생성이 저하되게 된다. 결과적으로 빌리루빈이 몸 밖으로 나가기 힘들기 때문에 혈액 속에 빌리루빈의 농도가 증가하게 된다.

1.2.3 크레아틴(Creatine)

크레아틴은 근육의 에너지원으로 쓰이는 물질이다. 간에서 크레아틴의 생성과 분해를 관장하고 있다. 간 기능의 저하로 단백질의 생성이나 분해가 되지 않게 되면 크레아틴의 대사산물인 크레아티닌 수치가 낮아지게 된다. 이처럼 부족해진 크레아틴은 진행성 간질환에 큰 영향을 미치게 된다

1.2.4 태아단백(Feto-Protein)

태아단백은 태아의 간에서 만들어져 분비되는 단백질이다. 태생기에 주로 만들어지며 생후 점점 감소하게 된다. 하지만 성인이 된 후 간 손상 및 간 질환의 경우 태아단백의 농도가 증가하게 된다.

1.2.5 혈소판(Platelet)

혈소판은 혈액을 응고시키는 역할을 하는 성분이다. 혈소판 수치가 정상범위 이하이면 간암 또는 간경변을 의심할 수 있다. 간암말기의 증상으로 혈소판감소증이 있다.

1.2.6 프로트롬빈 시간(Prothrombin Time)

프로트롬빈 시간은 혈액의 응고 시간을 의미한다. 프로트롬빈 시간은 Ⅰ, Ⅱ, Ⅴ, Ⅹ 인자들의 영향을 받는다. 혈액응고는 13개 이상의 응고 인자들이 관여하고 있으며 이러한 인자들 중 Ⅰ, Ⅱ, Ⅴ, Ⅶ, Ⅸ, Ⅹ, ⅩⅢ 인자들이 간에서 합성된다. 간의 기능이 떨어지면 응고 인자들의 합성이 적어지게 되며 혈액응고가 늦어지게 된다.

1.2.7 TNM 독립병기(TNM Staging System)

암의 진행 단계를 나타내는 기준이다. 정해진 병기를 바탕으로 암의 치료 방법 및 계획을 세우고 치료 예후와 결과를 예측한다. 종양병기(T), 림프절병기(N), 원격전이병기(M) 3가지로 구성이 된다.

1.2.7.1 종양병기(T)(Primary tumor)

TX : 주 종양이 측정될 수 없다.

T0 : 주 종양을 찾을 수 없다.

T1~T4 : 주 종양의 크기와 관련이 있다. 숫자가 커질수록 종양의 크기가 크거나 주변 조직 가까이에서 자랐음을 의미한다.

1.2.7.2. 림프절병기(N)(Regional lymph nodes)

NX : 가까운 림프절의 암이 측정될 수 없다.

N0 : 가까운 림프절에 알이 없다.

N1~N3 : 암을 포함한 림프절의 수와 위치와 관련이 있다. 숫자가 커질수록 암을 포함하는 림프절이 많음을 의미한다.

1.2.7.3 원격전이병기(M)(Distant metastasis)

MX : 전이가 측정될 수 없다.

M0 : 전이되지 않았다.

M1 : 전이되었다.

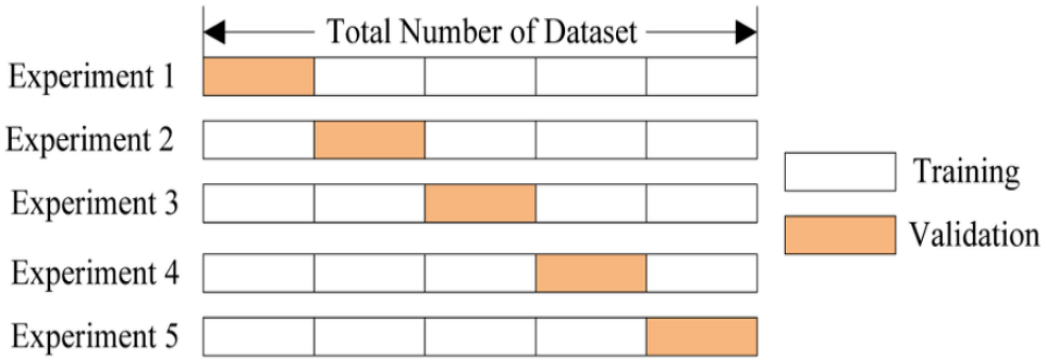
1.2.8 Modified UICC 병기분류

암의 개수, 크기, 혈관 침범여부, 림프절 전이 여부, 타 장기 전이 여부 등을 기준으로 1~4기로 나뉘게 된다. TNM 분류의 각 상태를 기준으로 암의 최종병기를 결정한다

1.3 관련 통계학 이론

1.3.1 K-fold cross validation

K-fold cross validation은 Training set을 K개의 fold로 나눠준 후 그 중 한 개는 validation set으로 이용을 하고 나머지 fold는 training set으로 이용한다. 각 validation set을 이용해 얻은 K개의 error의 평균을 통해 test error의 값을 추정한다. K-fold cross validation을 통해 개수가 적은 data에 대해 정확도를 향상시킬 수 있다.



1.3.2 Confusion matrix

Confusion matrix을 통해 분류 결과를 확인할 수 있다. 이를 통해 Accuracy, Sensitivity, Specificity와 같은 값을 구해 ROC curve를 얻을 수 있다.

Predicted True	Positive	Negative
Positive	TP	FN
Negative	FP	TN

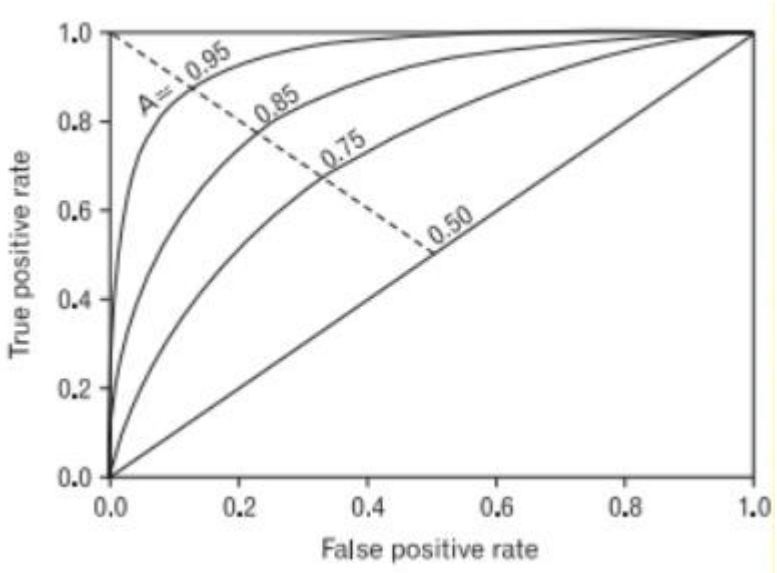
Accuracy = (TP + TN) / N

Specificity = TN / N

Sensitivity = TP / N

1.3.3 ROC curve, AUC

ROC curve는 sensitivity와 1-specificity를 축으로 높은 그래프이다. ROC curve의 커브의 x축과의 면적이 AUC가 된다. ROC curve가 왼쪽 위에 가까울수록 좋은 성능을 가진 분류이다. 이 때의 AUC는 1에 가까운 값을 가지게 된다.

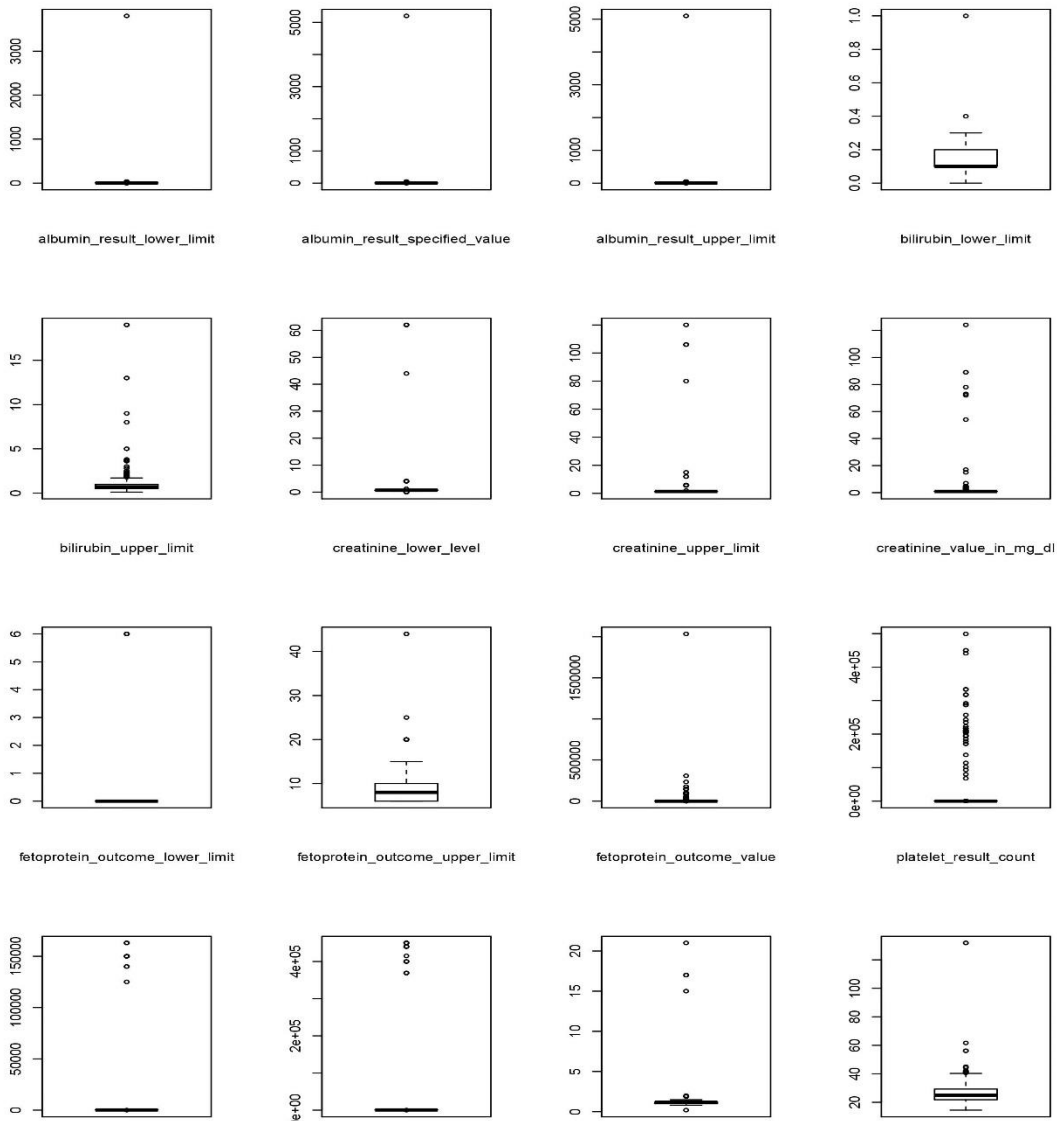


2. Data preprocessing

데이터는 CDC TCGA에서 제공하는 Liver Cancer phenotype dataset이며 469개의 샘플과 119개의 변수로 이루어져 있다. 변수중에 결측값(missing data)이 너무 많거나(90%이상), 클래스가 1개인 경우는 분산이 0이 되므로 분석에 도움이 되지 않는다. 따라서 이러한 변수 45개는 제외시켰다. 또한 우리의 목표인 혈액검사를 통한 간암병기 예측을 하려면 데이터는 오로지 혈액검사를 통해서 얻을 수 있는 변수만을 남겨두어야 한다. 따라서 조직검사나 영상검사로 얻을 수 있는 데이터를 제외하면 총 40개의 변수가 남게 된다.

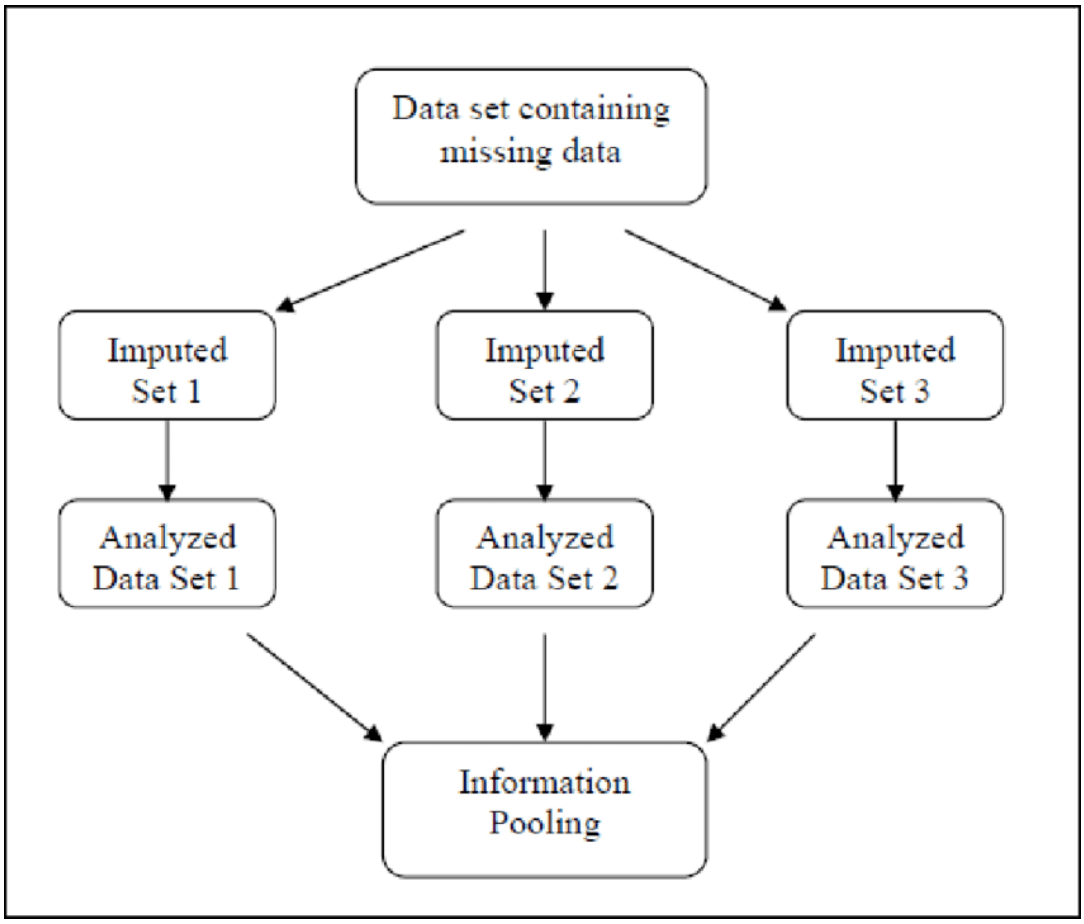
40개의 변수중에서 16개의 연속형 변수에 대해서 box plot을 그려보았다.(Figure[1]) 대부분의 변수에서 out-lier가 다수 식별되었다. 이는 조사자가 데이터를 입력하는 과정에서 발생하는 것으로 인위적으로 처리해주지 않으면 bias의 원인이 된다. 따라서 변수별로 적당한 Threshold를 지정하여 out-lier를 결측값으로 처리해주었다. 이 과정에서 일부 변수는 Dataset에서 제외되었다. 예를들어 feto_protein_lower_limit은 태아단백질의 하한을 의미하는 변수다. 이 변수는 1개의 out-lier를 제외하고 나머지(468개)는 0의 값을 가지므로 분산이 0이 되었다. 태아단백은 정상적인 간에서는 없는게 당연하므로 하한은 0이 되는

것이 자명하다. 이처럼 적절하게 out-lier를 처리하는 과정을 4번 반복하였으며 그 결과 변수는 34개로 정리되었고 out-lier는 적절하게 처리되었다.



Figure[1] 16개 연속형 변수의 Boxplot

이외에도 결측값의 처리는 model fitting 이전에 가장 중요한 이슈다. 단순히 평균이나 빈도가 높은 값으로 일괄적으로 대체하는 경우 표본분산이 낮아져 분석 결과가 왜곡될 수 있다. 그렇다고 결측치를 임의로 제거하면 bias가 심해질 수 있다. Dataset은 out-lier 처리 전에는 2795개의 NA값이 있었으며 전처리 이후에는 3058개의 NA값이 발생했다. 이러한 오류를 최소화하기 위해 multiple imputation을 이용하여 결측치를 보충해주었다. Multiple-Imputation(다중대체법)은 널리 사용되는 결측치 처리 방법으로 3단계로 이루어진다.(Figure[2]) 첫째로 불완전한 original dataset으로부터 Imputed dataset을 생성한다. 이때 생성된 dataset은 변수별로 observed data를 통해 추정된 분포를 이용하여 결측치를



Figure[2] Multiple Imputation(다중대체법)의 Diagram

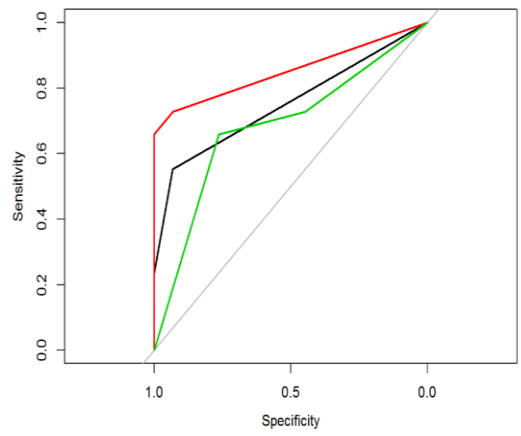
random하게 채워준다. 이는 true value를 알 수 없는 결측치의 불확실성을 보장해준다. 두번째로 생성된 Imputed dataset을 이용하여 model을 fitting 해준다. 각각의 dataset의 model fitting만으로는 각각 data의 차이가 있기 때문에 각기 다른 결과를 갖지만 이를 pooling해주면 의미가 있는 결과값을 얻을 수 있는 것이다. 많은 통계 소프트웨어가 MI를 지원하고 있으며 분석간에는 R의 "mice" package를 이용하여 Dataset을 완성하였다. 이를 이용해서 KNN, Naïve Bayes, Support vector machine, Tree decision method, Multiple logistics regression 방법을 이용하여 Liver hepatocellular carcinoma의 stage를 예측하는 모델을 만들어 보겠다.

3. Data analysis

3.1 KNN

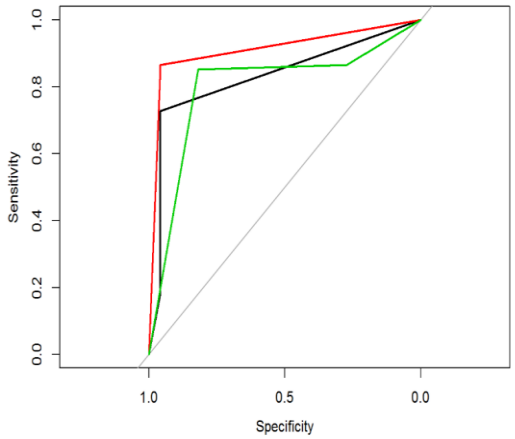
k-nearest neighbor(knn) 분류 알고리즘은 k개의 가장 가까운 training data를 이용하여 test data의 값을 예측한다. 적절한 k개의 값은 data를 통해서 직접 정해주저야 한다. 본 프로젝트에서는 training data에서 5-fold cross validation 방법을 이용하여 k의 값을 정해주었다. 또한 knn 알고리즘은 거리를 이용하는 방법이기 때문에 qualitative 변수들에 대해서는 one-hot encoding을 통해 거리를 구할 수 있도록 데이터 처리를 해주었다. Tumor stage를 'stage i', 'stage ii', 'stage iii이상'로 분류(A모형)하거나 'not reported', 'stage i', 'stage ii이상'으로 분류하는데(B모형)따라서 2가지 모형으로 분석한 결과는 다음과 같다. (Figure[3])

5-fold cross validation 방법을 통해 얻은 k 값은 A모형에서는 10이며, B모형에서는 14이다.



Predicted / True	Stage i	Stage ii	Stage iii
Stage i	68	5	0
Stage ii	17	12	9
Stage iii	12	3	29

AUC : 0.762 / Accuracy : 0.70 / k : 10



Predicted / True	Stage i	Stage ii	Not reported
Stage i	67	3	0
Stage ii	10	63	1
Not reported	3	2	6

AUC : 0.847 / Accuracy : 0.78 / k : 14

Figure[3] 좌측 plot은 A모형의 multiple ROC curve, AUC, confusion matrix이고 우측 plot은 B모형의 ROC, AUC, confusion matrix이다.

3.2 Naïve bayes

Naïve bayes classifier은 베이즈 정리를 이용한 분류기이다. 베이즈 정리를 이용한 모델이기 때문에 조건부 확률 모델이라고 생각을 할 수 있다. 베이즈 정리에 따르면 사

후 확률을 다음과 같이 구할 수 있다.

$$\Pr(\mathrm{Y}|\mathrm{X}=\mathrm{x}) = \frac{\Pr(Y)\times\Pr(X=x|Y)}{\Pr(X=x)}$$

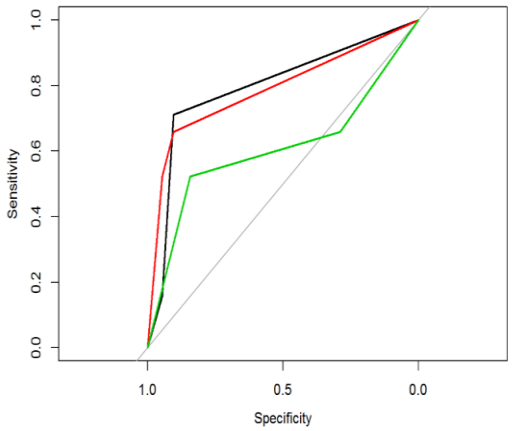
Naïve bayes classifier는 모든 특성들의 독립을 가정한 간단한 모델이다. 특성들의 독립을 가정하고 있기 때문에 모델이 매우 효율적이며 training data의 양이 적더라도 모델을 만들 수 있다. 특성들의 독립을 가정 한 단순한 모형이더라도 성능이 매우 좋다는 것이 알려져 있다. Naïve bayes classifier는 다음과 같은 식을 갖게 된다.

$$\hat{y} = argmax_{y\in\{1,2,\cdots,k\}}\, Pr(Y)\prod_{i=1}^{n}\Pr(x_i|Y)$$

즉, Naïve bayes classifier를 통해 예측되는 분류값은 사후확률을 가장 높게 만드는 값이다.

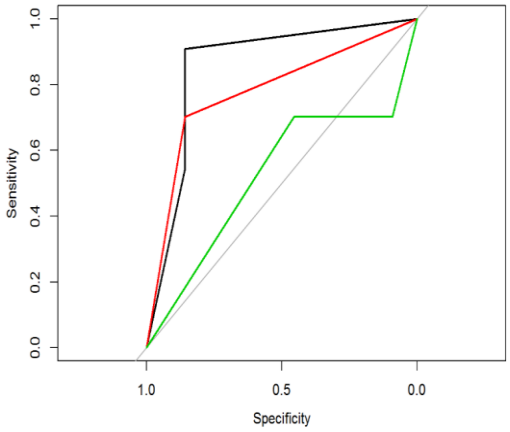
Naïve bayes classifier를 이용하기에 앞서 각 모형마다 모형에 이용할 변수를 선택해주는 과정을 선행하였다. 5-fold cross validation을 이용하여 accuracy를 최대로 만들어주는 변수의 개수를 선택을 한 후 전체 training set과 test set을 이용하여 해당하는 개수의 변수를 선택해주었다.

Tumor stage를 'stage i', 'stage ii', 'stage iii이상'로 분류(A모형)하거나 'not reported', 'stage i', 'stage ii이상'으로 분류하는데(B모형)따라서 2가지 모형으로 분석한 결과는 다음과 같다. (Figure[4])



Predicted / True	Stage i	Stage ii	Stage iii
Stage i	66	3	4
Stage ii	11	21	6
Stage iii	15	6	23

AUC : 0.731 / Accuracy : 0.701



Predicted / True	Stage i	Stage ii	Not reported
Stage i	60	10	0
Stage ii	22	52	0
Not reported	1	6	4

AUC : 0.721 / Accuracy : 0.748

Figure[4] 좌측 plot은 A모형의 multiple ROC curve, AUC, confusion matrix이고 우측 plot은 B모

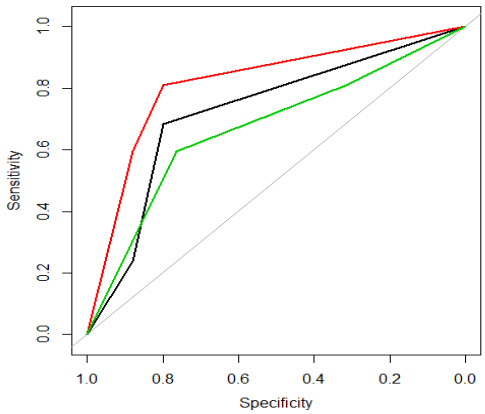
형의 ROC, AUC, confusion matrix이다.

A모형에서는 9개의 변수를 이용했을 때 가장 높은 accuracy 값을 얻을 수 있었다. 이 때 사용되는 9개의 변수는 "vascular tumor cell type", "fibrosis ishak score", "age at diagnosis", "radiation therapy", "ethnicity demographic", "inter norm ratio lower limit", "post op ablation embolization tx", "race demographic", "age at initial pathologic diagnosis" 였다. B모형에서는 6개의 변수를 이용했을 때 가장 높은 accuracy 값을 얻을 수 있었다. 이 때 사용되는 6개의 변수는 "vascular tumor cell type", "fibrosis ishak score", "creatinine lower level", "platelet result lower limit", "postoperative rx tx", "ethnicity demographic" 이었다.

3.3 Support vector machine

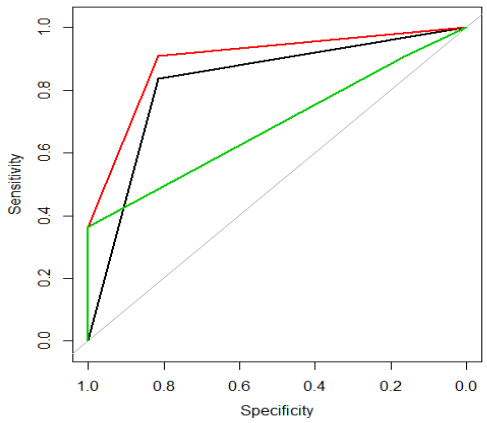
서포트 벡터 머신은 예측이 좋은 대표적인 모델중 하나이다. 일반적으로 서포트 벡터 머신은 maximal margin classifier, support vector classifier, support vector machine 3가지를 총칭해서 이른다. 이때 maximal margin classifier와 support vector classifier는 선형의 함수관계를 다루거나, 변수의 차원이 낮은 경우에는 잘 적합하나 우리가 다루는 dataset은 변수의 수가 34개로 위의 두가지 방법으로는 fitting하기에 적절하지 않다. 따라서 우리는 support vector machine을 사용하였다. Support vector machine(SVM)은 함수가 비선형이고 data의 차원이 높을때도 예측력이 높은 것으로 알려져 있다. SVM은 kernel function으로 비선형 함수를 표현하는데 주로 linear와 radial을 사용한다. 우리의 dataset으로는 linear가 예측력이 좋지 않았고 kernel function으로 radial을 사용하였다.

Radial function은 Gamma와 cost를 hyperparameter로 가진다. 최적의 값을 구하기 위해서 "e1071" package의 tune.out function을 이용했다. Tumor stage를 'stage i', 'stage ii', 'stage iii이상'로 분류(A모형)하거나 'not reported', 'stage i', 'stage ii이상'으로 분류하는데(B모형)따라서 2가지 모형으로 분석한 결과는 다음과 같다. (Figure[5])



Predicted True	Stage i	Stage ii	Stage iii
Stage i	60	6	9
Stage ii	12	17	9
Stage iii	8	9	25

AUC : 0.732 / Accuracy : 0.65



Predicted True	Stage i	Stage ii	Not reported
Stage i	57	13	0
Stage ii	12	62	0
Not reported	1	6	4

AUC : 0.8032 / Accuracy : 0.79

Figure[5] 좌측 plot은 A모형의 multiple ROC curve, AUC, confusion matrix이고 우측 plot은 B모형의 ROC, AUC, confusion matrix이다.

3.4 Tree Decision method

3.4.1 모델설명

의사결정나무(decision trees)는 주어진 입력값에 대하여 출력값을 예측하는 모형으로서 분류나무(classification trees)와 회귀나무(regression trees) 모형이 있다. 의사결정나무라는 이름은 그 결과를 나무 형태의 그래프로 표현할 수 있다는 사실에 기인한다.

의사결정나무는 지도학습 기법으로 각 변수의 영역을 반복적으로 분할함으로써 전체 영역에서의 규칙을 생성한다. 의사결정나무의 예측력은 다른 지도학습 기법들에 비해 대체로 떨어지나 해석력이 좋다.

의사결정나무의 형성과정은 크게 growing(성장), pruning(가지치기), 타당성 평가, 해석 및 예측으로 이루어진다. 성장단계는 각 마디에서 적절한 최적의 분리규칙을 찾아서 나무를 성장시키는 과정으로서 적절한 정지규칙을 만족하면 중단한다. 가지치기 단계는 오차를 크게 할 위험이 높거나 부적절한 추론규칙을 가지고 있는 가지 또는 불필요한 가지를 제거한다. 타당성 평가 단계에서는 이익도표 혹은 시험자료를 이용하여 의사결정나무를 평가하게 된다. 해석 및 예측단계에서는 구축된 나무모형을 해석하고 예측모형을 설정한 후 예측에 적용한다. 우리는 출력변수가 범주형인 분류나무(classification tree)를 사용하여 모델을 적합해 보았다.

3.4.2 Tree, raprt, party packages

R에는 의사결정나무 분석을 할 수 있는 패키지가 여러 개 존재한다. 그 중 대표적인 3개의 패키지를 꼽자면 tree, rpart, party 가 있다. 각각의 패키지에는 의사결정나무를 만들 때 가지치기를 하는 방법에 차이가 존재한다.

tree 패키지는 binary recursive partitioning을, rpart 패키지는 CART(classification and regression trees) 방법론을 사용한다. 이 패키지들은 엔트로피, 지니계수를 기준으로 가지치기를 할 변수를 결정하기 때문에 상대적으로 연산 속도는 빠르지만 over-fitting의 위험성이 존재한다. 그래서 두 패키지를 사용할 경우에는 Pruning 과정을 거쳐서 의사결정나무를 최적화하는 과정이 필요하다.

party 패키지는 Unbiased recursive partitioning based on permutation tests 방법론을 사용한다. p-test를 거친 Significance를 기준으로 가지치기를 할 변수를 결정하기 때문에 biased 될 위험이 없어 별도로 Pruning할 필요가 없다는 장점이 있다.

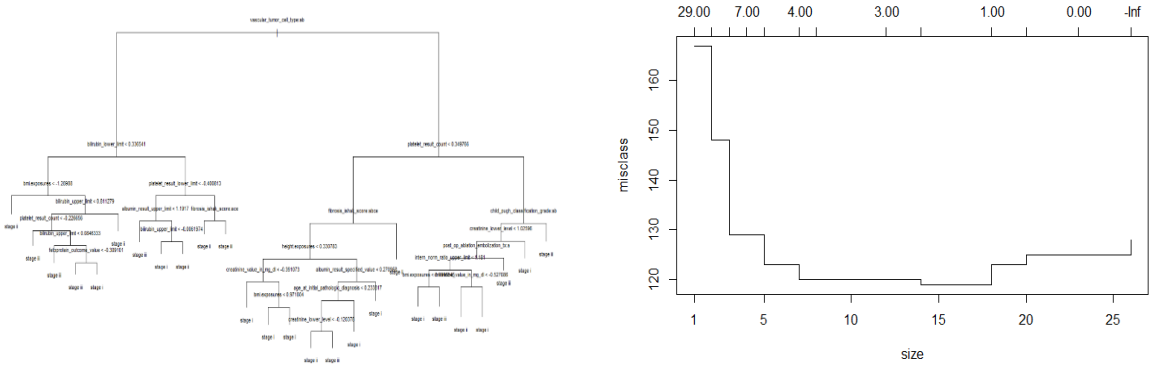
3.4.3 분석결과

3.4.3.1 tree package

① A 모델(stage i, stage ii, stage iii)

tree package를 full_train 데이터셋에 대해 적용하여 의사결정나무를 만들어 보았다.

이후에 over-fitting의 문제를 해결하기 위해서 가지치기(pruning) 단계를 진행하였다. 10-fold cross-validation 방법을 사용하여 full_train 데이터셋을 10개로 쪼개서 테스트한 후에 분산이 가장 낮은 가지의 수를 선택하여 주었다. 아래의 cv 그래프를 보면 가지의 수가 14, 15, 16, 17, 18일 때의 의사결정나무가 가장 분산이 낮은 걸 확인 해 볼 수 있다.

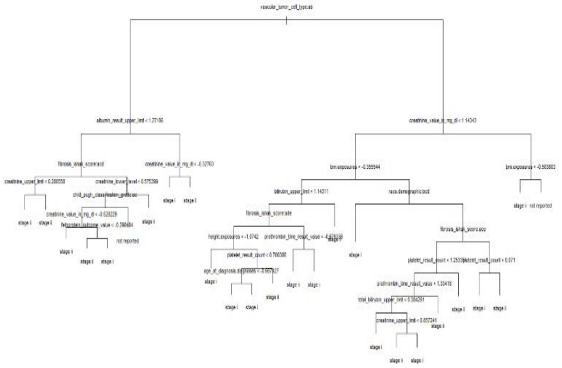


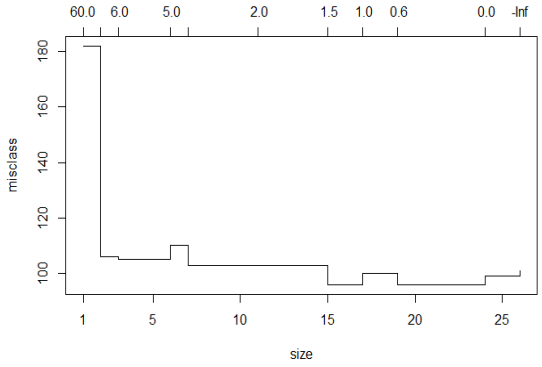
각 경우(14, 15, 16, 17, 18개의 가지수)에 맞게끔 의사결정나무를 pruning 할 때마다 predict 함수를 사용해서 test 데이터셋의 tumor_stage.diagnoses를 예측한 후, confusionMatrix함수를 사용해서 모델의 정확성을 평가해 보았다. 가지의 수를 14개로 했을 때가 Accuracy가 0.5962로 가장 높았다. 또한 해당 경우의 AUC는 0.7223이 나왔다.

② B 모델(nor reported, stage i, stage ii)

마찬가지로 tree package를 full_train 데이터셋에 대해 적용하여 의사결정나무를 만들어 보았다.

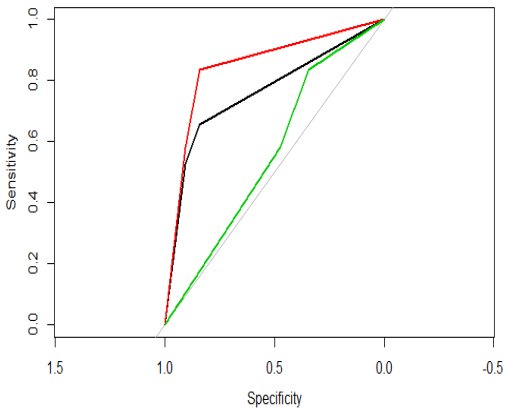
이후에 over-fitting의 문제를 해결하기 위해서 가지치기(pruning) 단계를 진행하였다. 10-fold cross-validation 방법을 사용하여 full_train 데이터셋을 10개로 쪼개서 테스트한 후에 분산이 가장 낮은 가지의 수를 선택하여 주었다. 아래의 cv 그래프를 보면 가지의 수가 15, 16, 17, 19, 20, 21, 22, 23, 24일 때의 의사결정나무가 가장 분산이 낮은 걸 확인 해 볼 수 있다.





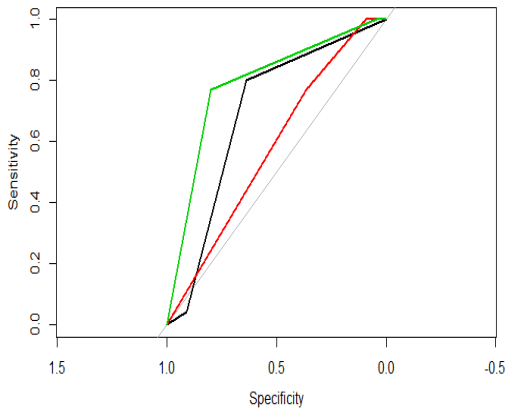
각 경우(15, 16, 17, 19, 20, 21, 22, 23, 24개의 가지 수)에 맞게끔 의사결정나무를 pruning 할 때마다 predict 함수를 사용해서 test 데이터셋의 tumor_stage.diagnoses를 예측한 후, confusionMatrix함수를 사용해서 모델의 정확성을 평가해 보았다. 가지의 수를 19개로 했을 때가 Accuracy가 0.7161로 가장 높았다. 또한 해당 경우의 AUC는 0.6857이 나왔다.

A모델과 B모델의 accuracy, ROC curve, AUC를 비교하면 아래와 같다.



Predicted / True	Stage i	Stage ii	Stage iii
Stage i	63	5	7
Stage ii	13	5	20
Stage iii	7	11	25

AUC : 0.7223 / Accuracy : 0.5962



Predicted / True	Stage i	Stage ii	Not reported
Stage i	53	14	3
Stage ii	17	57	0
Not reported	3	7	1

AUC : 0.6857 / Accuracy : 0.7161

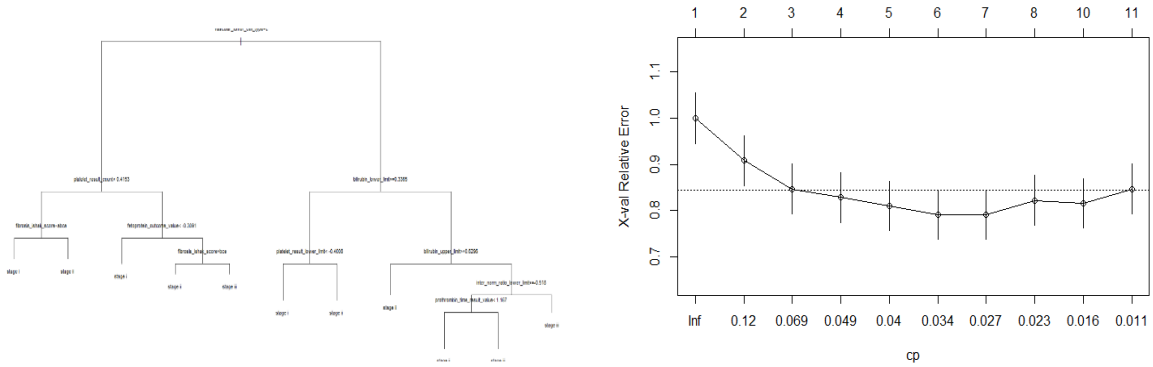
Figure[6] 좌측 plot은 A모형의 multiple ROC curve, AUC, confusion matrix이고 우측 plot은 B모형의 ROC, AUC, confusion matrix이다.

3.4.3.2 rpart package

① A 모델(stage i, stage ii, stage iii)

rpart package를 full_train 데이터셋에 대해 적용하여 의사결정나무를 만들어 보았다.

이후에 over-fitting의 문제를 해결하기 위해서 가지치기(pruning) 단계를 진행하였다. 10-fold cross-validation 방법을 사용하여 full_train 데이터셋을 10개로 쪼개서 테스트한 후에 xerror가 가장 낮은 split 개수를 선택해 주었다. 아래의 그래프를 보면 split이 6개일 때 가장 낮은 error를 보이고 있다.



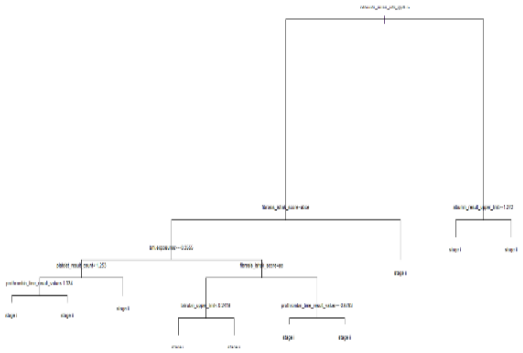
해당 경우에 맞게끔 의사결정나무를 pruning 하고 predict 함수를 사용해서 test 데이터셋의 tumor_stage.diagnoses를 예측한 후, confusionMatrix함수를 사용해서 모델의 정확성을 평가해 보았다. Accuracy는 0.5576923가 나왔고 또한 해당 경우의 AUC는 0.6664가 나왔다.

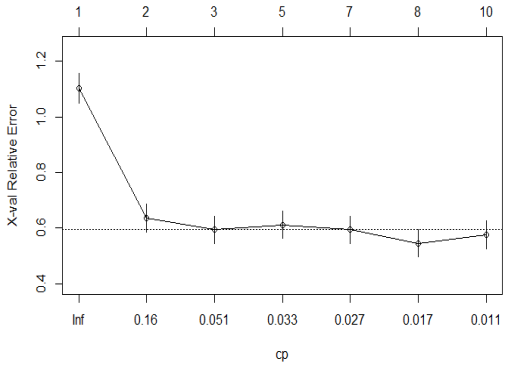
② B 모델(nor reported, stage i, stage ii)

rpart package를 full_train 데이터셋에 대해 적용하여 의사결정나무를 만들어 보았다.

이후에 over-fitting의 문제를 해결하기 위해서 가지치기(pruning) 단계를 진행하였다. 10-fold cross-validation 방법을 사용하여 full_train 데이터셋을 10개로 쪼개서 테스트한 후에 xerror가 가장 낮은 split 개수를 선택해 주었다. 아래의 그래프

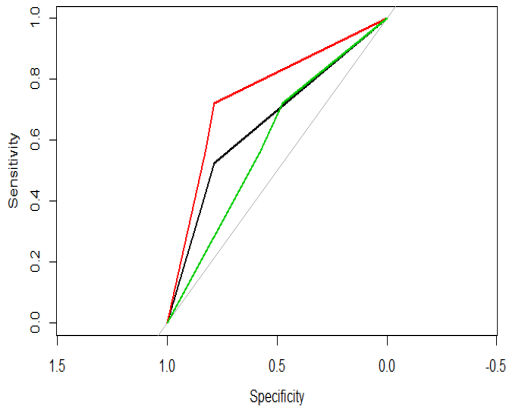
를 보면 split이 8개일 때 가장 낮은 error를 보이고 있다.





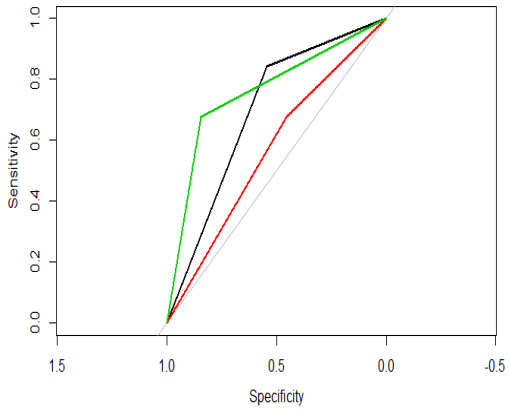
해당 경우에 맞게끔 의사결정나무를 pruning 하고 predict 함수를 사용해서 test 데이터셋의 tumor_stage.diagnoses를 예측한 후, confusionMatrix함수를 사용해서 모델의 정확성을 평가해 보았다. Accuracy는 0.7032가 나왔고 또한 해당 경우의 AUC는 0.6728가 나왔다.

A모델과 B모델의 accuracy, ROC curve, AUC를 비교하면 아래와 같다.



Predicted True	Stage i	Stage ii	Stage iii
Stage i	59	3	13
Stage ii	18	4	16
Stage iii	12	4	24

AUC : 0.6664 / Accuracy : 0.5577



Predicted True	Stage i	Stage ii	Not reported
Stage i	59	11	0
Stage ii	24	50	0
Not reported	5	6	0

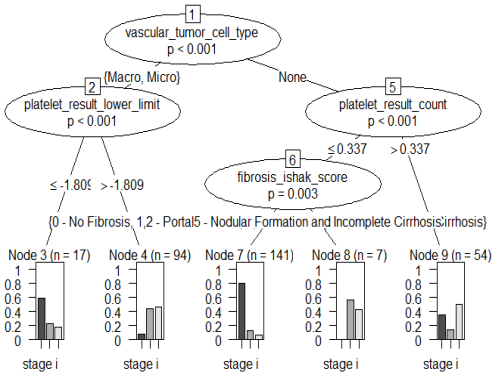
AUC : 0.6728 / Accuracy : 0.7032

Figure[7] 좌측 plot은 A모형의 multiple ROC curve, AUC, confusion matrix이고 우측 plot은 B모형의 ROC, AUC, confusion matrix이다.

3.4.3.3 party package

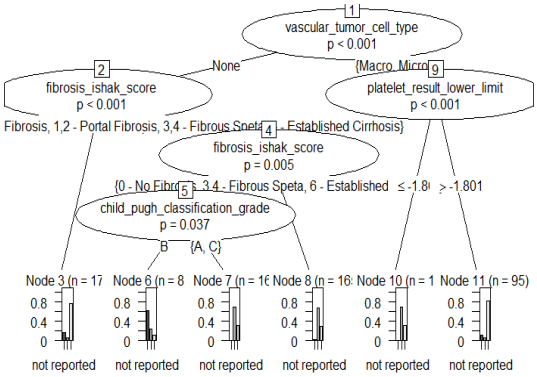
① A 모델(stage i, stage ii, stage iii)

party package는 significance를 사용하기 때문에 별도의 pruning 과정이 필요 없다. 따라서 test 데이터셋을 사용해서 바로 결과를 확인해 볼 수 있었다. Accuracy는 0.5833333이 나왔고 AUC는 0.6951이 나왔다.

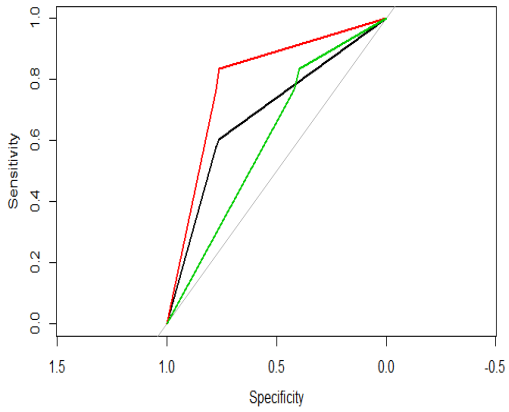


② B 모델(nor reported, stage i, stage ii)

마찬가지로 party package는 significance를 사용하기 때문에 별도의 pruning 과정이 필요 없다. 따라서 test 데이터셋을 사용해서 바로 결과를 확인해 볼 수 있었다. Accuracy는 0.7161이 나왔고 AUC는 0.6608이 나왔다.

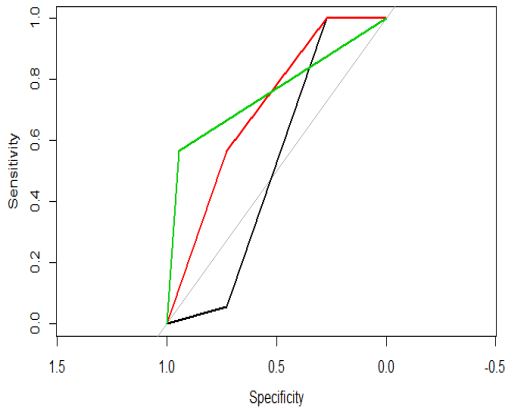


A모델과 B모델의 accuracy, ROC curve, AUC를 비교하면 아래와 같다.



Predicted / True	Stage i	Stage ii	Stage iii
Stage i	57	15	7
Stage ii	1	1	3
Stage iii	17	22	33

AUC : 0.6951 / Accuracy : 0.5833



Predicted / True	Stage i	Stage ii	Not reported
Stage i	66	4	0
Stage ii	32	42	0
Not reported	5	3	3

AUC : 0.6608 / Accuracy : 0.7161

Figure[8] 좌측 plot은 A모형의 multiple ROC curve, AUC, confusion matrix이고 우측 plot은 B모형의 ROC, AUC, confusion matrix이다.

3.5 Multiple logistics regression

3.5.1 모델설명

종속변수 Y의 수준수가 k개 일 때, 일반적으로 범주표시는 1부터 k까지 숫자가 부여된다. $x=(1, x_1, x_2, \cdots, x_p)'$ 는 상수항을 포함한 독립변수들이다. 독립변수들의 수준이 x일 때 종속변수의 결과가 $Y=m$이 될 확률을 $P(Y=j|x)=\pi_m(x)$ 로 표시하자. 일반화 로짓모형(generalized logit model)에 근거하여 범주 1을 참고범주로 둔다면, 공변량 $x_i=(1, x_{i1}, x_{i2}, \cdots, x_{ip})'$ 에서 종속변수인 Y가 j번째 범주일 확률 $\pi_j(x_i)$ 는 다음과 같이 정의된다.

$$\pi_j(x_i)=\frac{1}{1+\sum_{l=2}^{k}\exp(x_i'\beta_l)} \quad \text{for } j=1$$

$$\pi_j(x_i)=\frac{\exp(x_i'\beta_j)}{1+\sum_{l=2}^{k}\exp(x_i'\beta_l)} \quad \text{for } j=2, \cdots, k$$

(모수벡터: $\beta_j = (\beta_{j0}, \beta_{j1}, \cdots, \beta_{jp})' \quad \text{for } j = 2, \cdots, k$)

Multiclass logistic model에서 자료들이 독립이라면 가능도함수는 다음과 같이 설정된다.

$$L(\beta_1, \cdots, \beta_k | x) = \prod_{j=1}^{k} \prod_{Y_i = j} \frac{\exp(x_i' \beta_j)}{\sum_{m=1}^{k} \exp(x_i' \beta m)}$$

일반적으로 Newton-Raphson 유형의 알고리즘을 사용하여 β_j에 대한 최대가능도 추정량 $\hat{\beta}_j$를 얻는다. 이후 개별 범주에 속할 확률은 아래와 같은 식을 통해서 추정될 수 있고 해당 확률이 가장 큰 범주로 분류를 하게 된다.

$$\hat{\pi}_j(x_i) = \frac{\exp(x_i' \hat{\beta}_j)}{1 + \sum_{l=2}^{k} \exp(x_i' \hat{\beta}_l)} \quad \text{for } j = 2, \cdots, k$$

$$\hat{\pi}_j(x_i) = \frac{1}{1 + \sum_{l=2}^{k} \exp(x_i' \hat{\beta}_l)} \quad \text{for } j = 1$$

3.5.2 Multinom() function in nnet package

nnet 패키지에 들어있는 multinom() 이란 함수는 neural networks를 통해서 multinomial log-linear 모델을 적합해 준다. 해당 분석에서는 3개의 class를 갖는 종속변수 Y: tumor_stage.diagnoses를 분류하고자 multinom() 함수를 통해 multiclass logistic regression을 적합해보았다.

3.5.3 분석결과

① A 모델(stage i, stage ii, stage iii)

"full_train" 데이터셋에는 1개의 종속변수(tumor_stage.diagnoses)와 33개의 반응변수가 들어있다. full_train 데이터셋에 대해 nnet package에 들어있는 multinom() 함수를 사용하여 5-fold cross-validation 과정을 수행하였다. 우선 313개의 observation을 가지는 full_train 데이터를 각각 62개, 62개, 63개, 63개, 63개의 observation을 갖는 train1, train2, train3, train4, train5인 5개의 fold로 구분하였다.

k=1,2,3,4,5에 대해서 k번째 fold를 제외시키고 남은 4개의 fold를 이용해서 반응변수를 1개 갖는 총 33종류의 multiclass logistic regression model을 적합하였다. 이후 k번째 fold를 대입하여 각각의 accuracy를 구했고, 이들을 평균 내어 가장 큰 accuracy를 보이는 모델에 들어있는 변수를 선택하였다.

이후 앞서 뽑힌 변수를 고정시키고 나머지 32개의 변수를 하나씩 추가해 만든 두 개의 변수를 갖는 총 32개에 모델에 대해서 동일한 방법으로 5-fold cross-validation을 진행해 다음 변수를 선택하였다. 해당 과정을 5-fold에서 구한 최대평균 accuracy의 값이 더이상 증가하지 않을 때까지 반복해 주었다.

그 결과 5-fold cross-validation을 통해 최적의 변수 개수가 5개라는 정보를 얻었고 test 데이터셋을 이용하여 최종 accuracy를 기준으로 모델은 선정하고자 하였다. 원래대로라면 full_train 데이터셋을 이용해 5개의 반응변수를 갖는 총 237336($_{33}C_5$)개의 모델을 적합한 후에 test 데이터셋을 넣어주어 가장 높은 accuracy를 갖는 모델을 선정해야 한다. 하지만 computing 시간이 너무 오래 걸리기 때문에 변수의 개수를 5개라 정해 놓고 forward selection 방법을 사용해 최종 모델을 선정해 주었다.

그 결과 최종 모델은 "vascular tumor cell type", "child pugh classification grade", "fibrosis ishak score", "albumin result lowe limit", "age at initial pathologic diagnosis" 순으로 변수를 선정하여 해당하는 5개의 변수를 가지게 되었고 test accuracy는 0.7115, AUC는 0.7597을 가졌다.

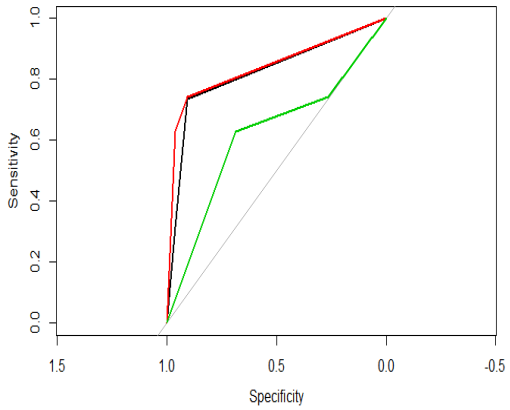
② B 모델(nor reported, stage i, stage ii)

A 모델의 경우 마찬가지 방법으로 분석을 진행하였다. B 모델의 경우에는 314개의 observation을 가지는 full_train 데이터를 각각 62개, 63개, 63개, 63개, 63개의 observation을 갖는 train1, train2, train3, train4, train5인 5개의 fold로 구분하였다.

5-fold cross-validation을 통해 최적의 변수 개수가 4개라는 정보를 얻었고 test 데이터셋을 이용하여 최종 accuracy를 기준으로 모델은 선정하고자 하였다. 원래대로라면 full_train 데이터셋을 이용해 4개의 반응변수를 갖는 총 40920개의 모델을 적합한 후에 test 데이터셋을 넣어주어 가장 높은 accuracy를 갖는 모델을 선정해야 한다. 하지만 computing 시간이 너무 오래 걸리기 때문에 변수의 개수를 4개라 정해 놓고 forward selection 방법을 사용해 최종 모델을 선정해 주었다.

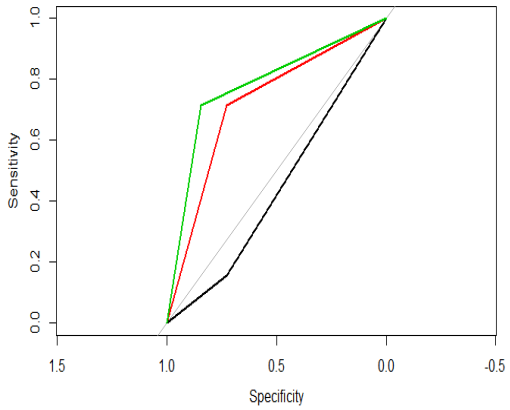
그 결과 최종 모델은 "vascular tumor cell type" "neoplasm histologic grade" "platelet result lower limit" "platelet result count" 순으로 변수를 선정하여 해당하는 4개의 변수를 가지게 되었고 test accuracy는 0.7226, AUC는 0.6478을 가졌다.

A모델과 B모델의 accuracy, ROC curve, AUC를 비교하면 아래와 같다.



Predicted / True	Stage i	Stage ii	Stage iii
Stage i	68	4	3
Stage ii	10	16	12
Stage iii	11	5	27

AUC : 0.7597 / Accuracy : 0.7115



Predicted / True	Stage i	Stage ii	Not reported
Stage i	59	11	0
Stage ii	21	53	0
Not reported	8	3	0

AUC : 0.6478 / Accuracy : 0.7226

Figure[9] 좌측 plot은 A모형의 multiple ROC curve, AUC, confusion matrix이고 우측 plot은 B모형의 ROC, AUC, confusion matrix이다

.

4. Results

분류방식 모델명	stage1 stage2 stage3 & stage4	stage1 stage2 & stage3 & stage4 Not reported
KNN	AUC : 0.762 / Accuracy : 0.70	AUC : 0.847 / Accuracy : 0.78
Naïve bayes	AUC : 0.731 / Accuracy :0.701	AUC : 0.721 / Accuracy : 0.748
SVM	AUC : 0.732 / Accuracy : 0.65	AUC : 0.8032 / Accuracy : 0.79
Decision tree(tree)	AUC : 0.722 / Accuracy : 0.60	AUC : 0.686 / Accuracy : 0.716
Decision tree(rpart)	AUC : 0.666 / Accuracy : 0.56	AUC : 0.673 / Accuracy : 0.703
Decision tree(party)	AUC : 0.695 / Accuracy : 0.58	AUC : 0.661 / Accuracy : 0.716
Multiple logistics	AUC : 0.760 / Accuracy : 0.71	AUC : 0.648 / Accuracy : 0.723

Stage1 / stage2/ stage3 & stage4 로 분류한 A모형일 때는 AUC를 기준으로 한다면 KNN과 Multiple logistics가 가장 좋은 결과를 보였다. Accuracy를 추가적으로 고려한다면 Multiple logistics 모델이 가장 좋은 효율을 보였다.

Stage1 / stage2 & stage3 & stage4 / Not reported로 분류한 B모형일 때는 AUC를 기준으로 한다면 KNN이 가장 좋은 결과를 보였다. 하지만 SVM도 매우 높은 AUC를 보였으며 Accuracy에서는 SVM이 가장 높은 값을 보였다. AUC와 Accuracy를 모두 고려한다면 KNN이 가장 좋은 결과를 보였다. SVM도 유사한 성능을 갖고 있다고 볼 수 있다.

5. Conclusion & Discussion

두 모형 모두 AUC와 Accuracy가 0.7 이상의 값을 갖는 모델들이 있다. 이러한 모델들을 이용한다면 조직검사를 하지 않더라도 혈액정보를 포함한 환자들의 정보만을 이용해서 높은 예측력으로 환자들의 간암 stage를 예측할 수 있다.

5-fold cross validation을 이용하여 Naïve bayes 모델과 Multiple logistic 모델에서는 variable selection 과정을 진행하였다. Variable selection 방법에는 여러 가지가 있지만 본 프로젝트에서는 forward selection 방법만을 이용하였다. 또한 forward selection을 할 때 기준이 되는 값이 test error가 아닌 accuracy를 이용했다는 한계점이 있다.

AUC와 Accuracy를 기준으로 성능이 좋다고 판단된 KNN과 SVM 모델의 경우 해석이 제한적인 부분이 있어서 특정 변수가 병기 결정에 어떠한 영향을 하는지 파악하기가 어렵다는 한계점이 있다. 또한 해석을 하기 쉬운 Decision tree 같은 경우 다른 모델들에 비해 성능이 좋지 못했다는 결과를 얻을 수 있었다.

6. Reference

[1] Jonathan AC Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, James R Carpenter et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls

[2] Gareth James, Daniela Witten, Trevor Hastie, and Rob Tibshirani (2013). An introduction to Statistical Learning with application in R

[3] Umberto Cillo, Alessandro Vitale, Francesco Grigoletto, Fabio Farinati, Alberto Brolese, Giacomo Zanus et al. (2006). Prospective validation of the Barcelona Clinic Liver Cancer staging system

[4] 이병석, 박남환 (2010). 간암의 진단적 분류법들에 대한 고찰