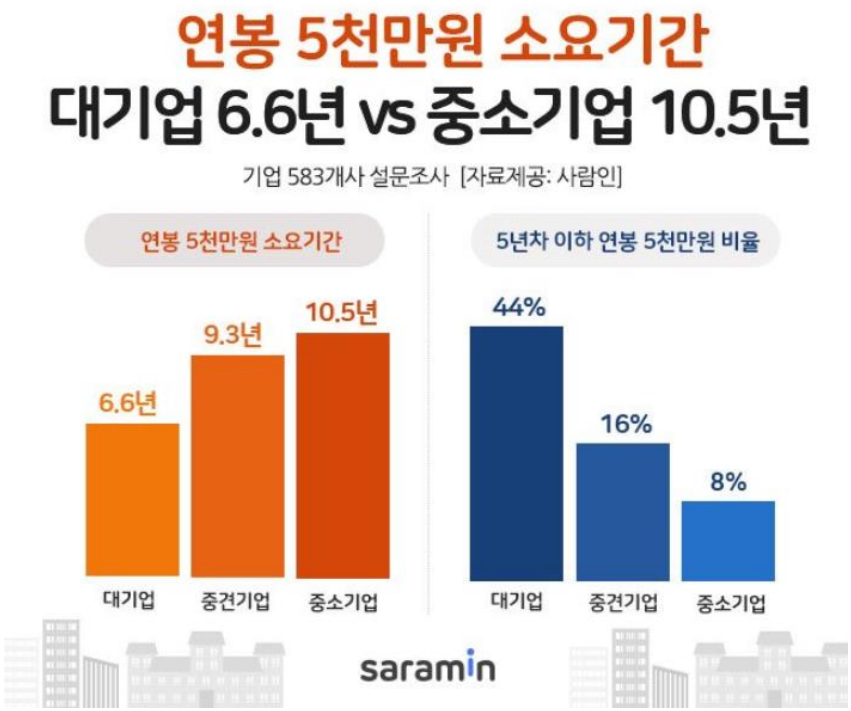
Data Mining Team Project

- 개인의 특성을 고려한 수입분류

데이터마이닝 프로세스

1. 목적설정
2. 데이터 수집
3. 데이터 전처리 (탐색 및 정제)
4. 데이터 분할
5. 데이터마이닝 문제설정
6. 데이터 분할
7. 데이터마이닝 방법선택
8. 최종모델 결정
9. 성능평가
10. 모델적용

#1 목적설정



우리나라에서 상징적인 "연봉 5,000만원"

인구통계학적 정보를 활용하여

어떤 특성을 지닌 사람들이 연봉 5,000만원을 넘는지 안 넘는지 분류

#2 데이터 전처리 - 구조 파악

Variable Name	Data Type	Variable Description
Y	범주형	연 소득 $50,000 이상 여부
Age	연속형	연령
Workclass	범주형	종사형태
Education	범주형	학력
Martial status	범주형	혼인유무
Occupation	범주형	직업
Relationship	범주형	가족관계
Race	범주형	인종
Sex	범주형	성별
Capital.gain	연속형	자본수입
Capital.loss	연속형	자본손실
Hours.per.week	연속형	주당 근무시간
Native.country	범주형	국적

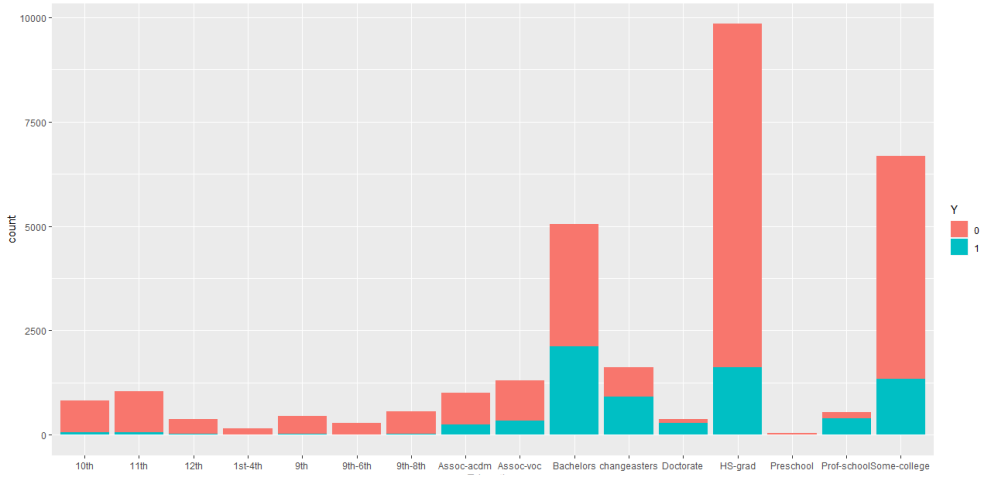
데이터 구조

30,161 (n)

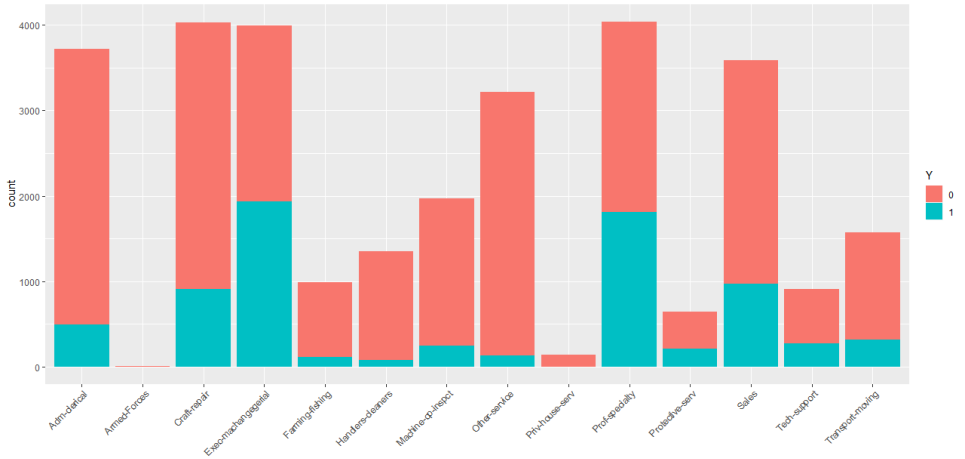
x

13 (p)

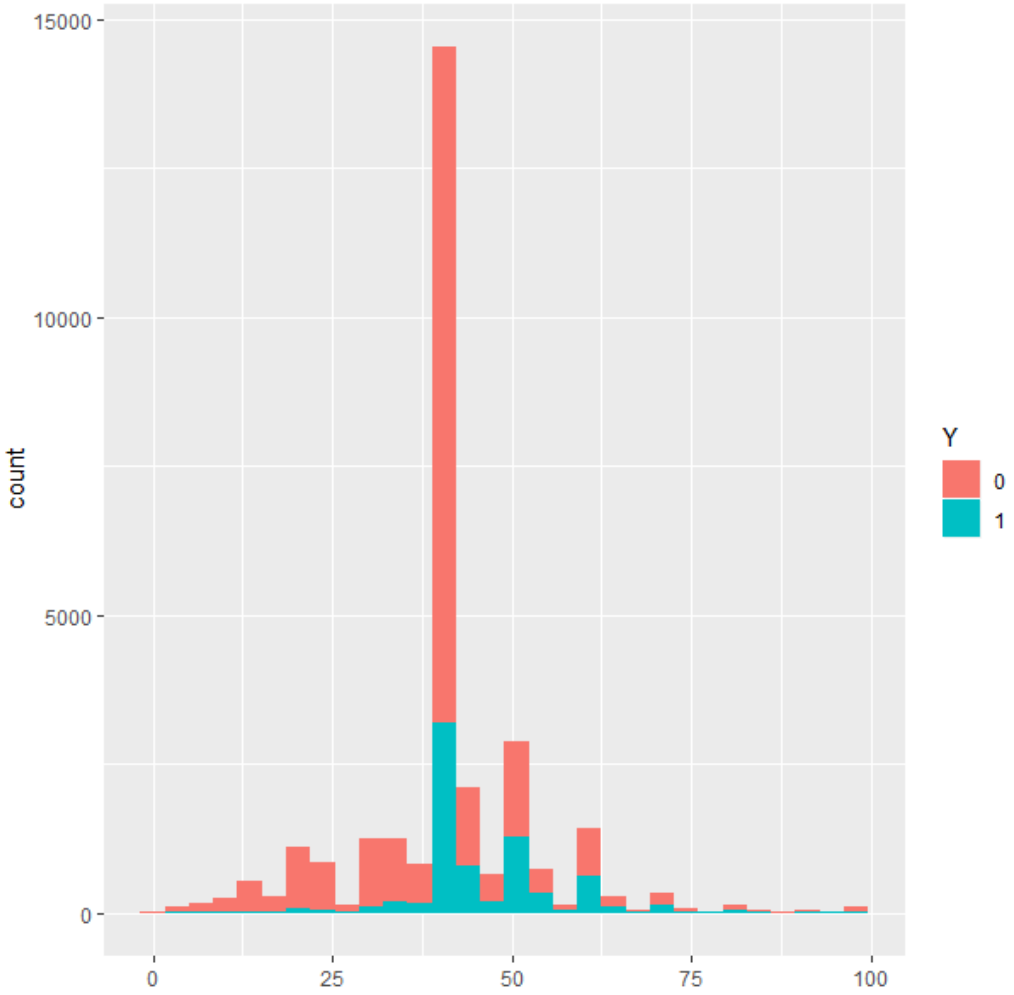
#2 데이터 전처리 - 탐색



Education

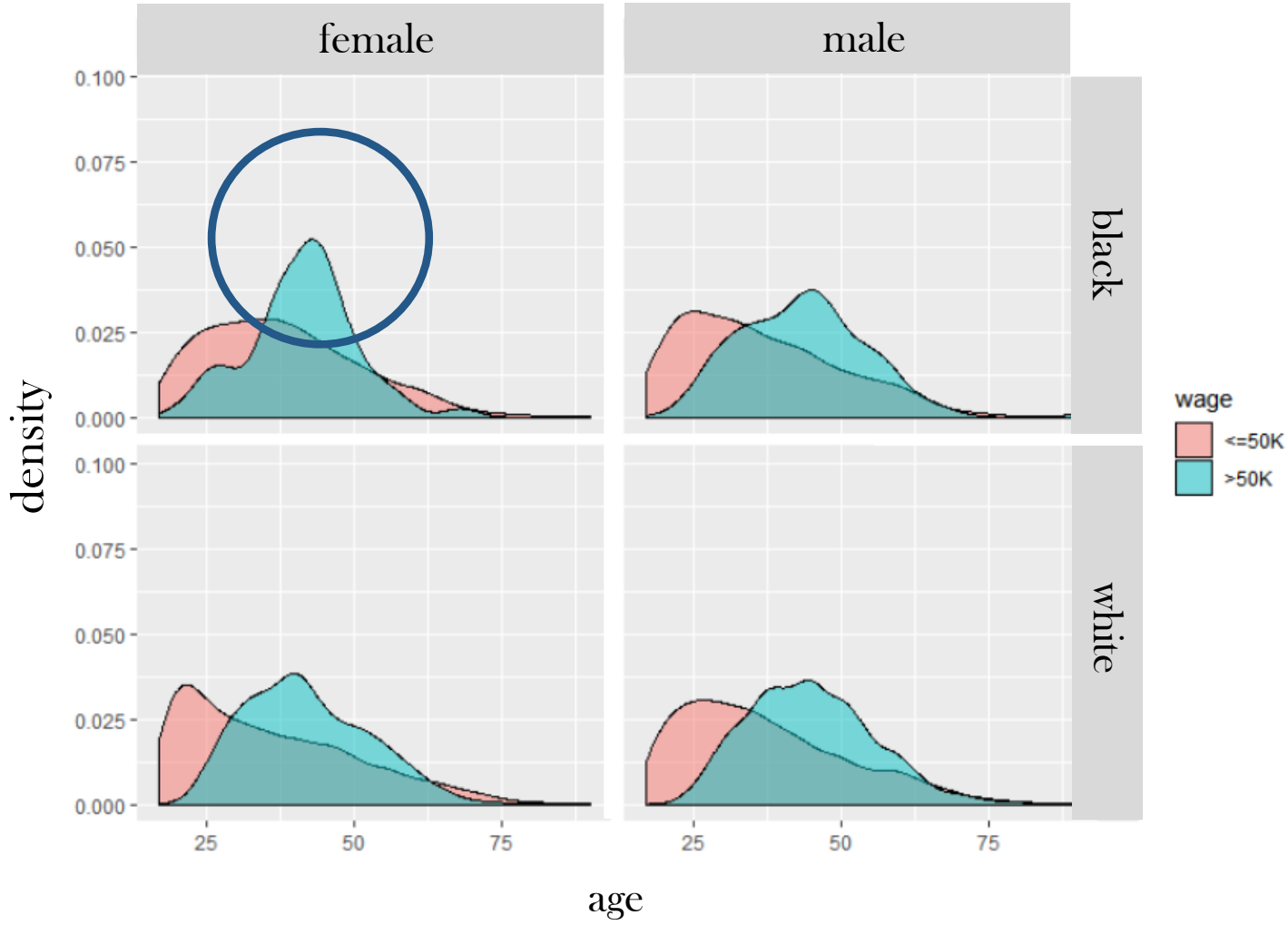


Occupation

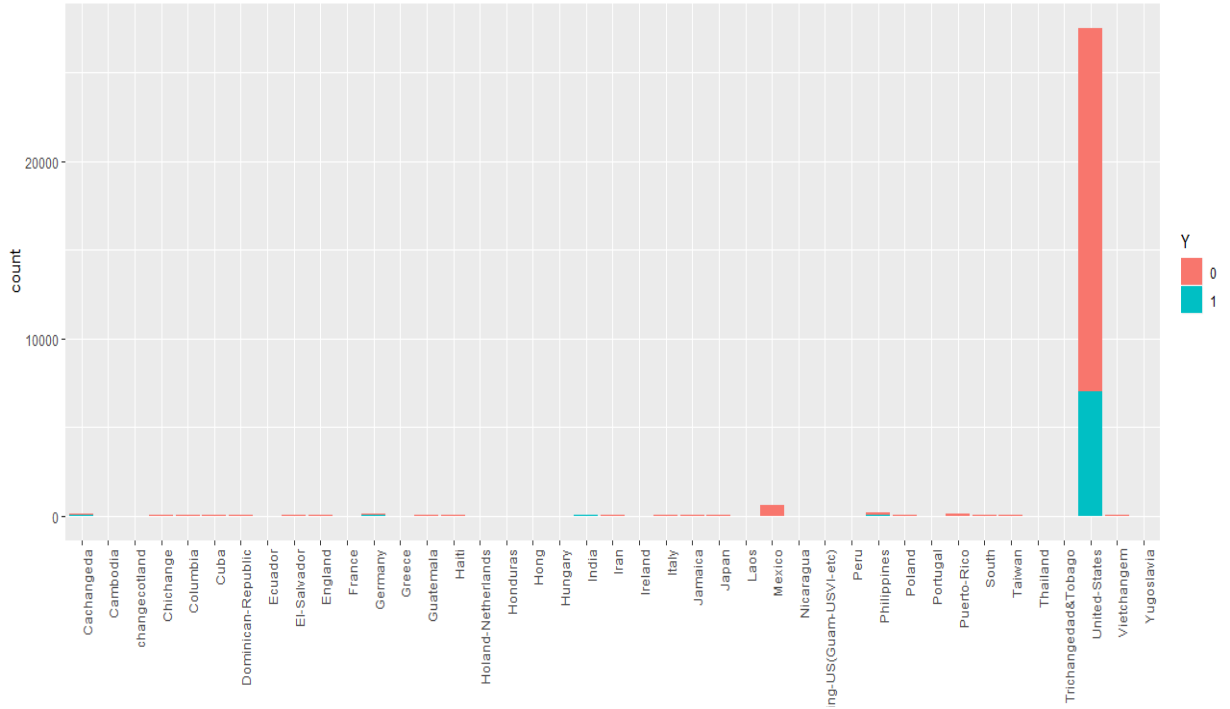


Hours.per.week

#2 데이터 전처리 - 탐색

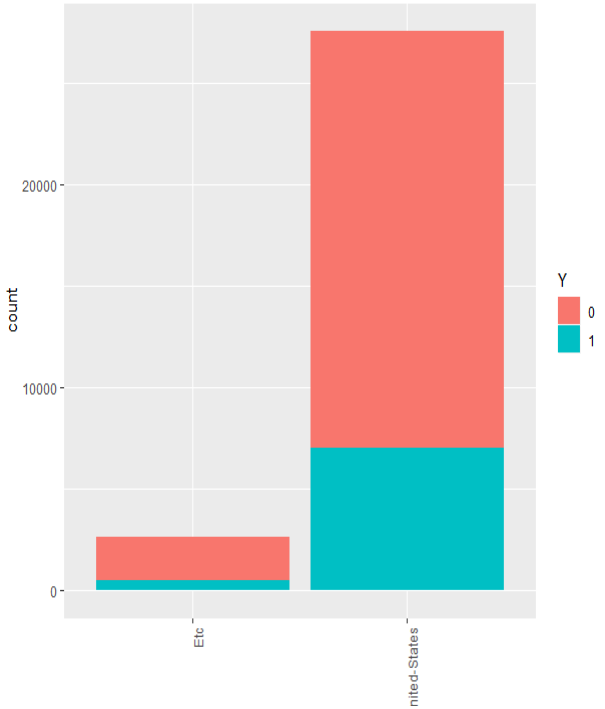

female
male
black
white
density
0.100
0.075
0.050
0.025
0.000
25
50
75
age
wage
<=50K
>50K

#2 데이터 전처리 – 정제



Native.country

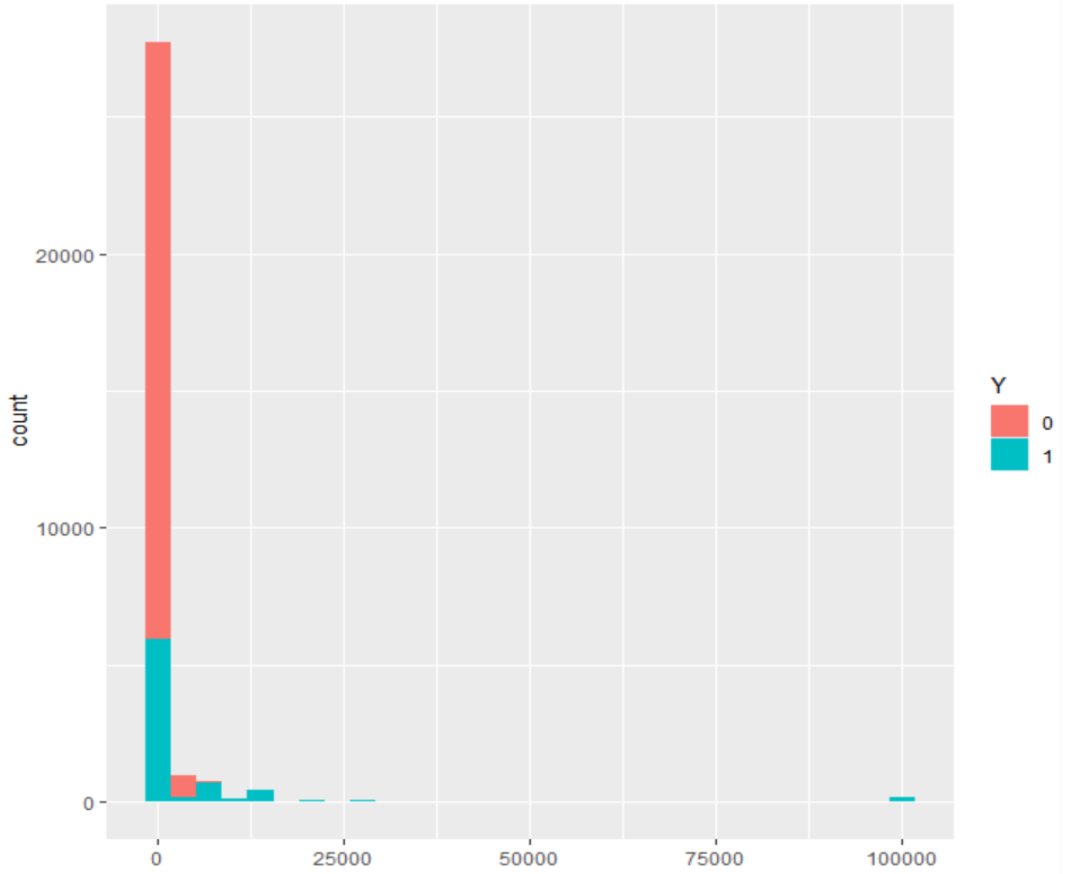
변환



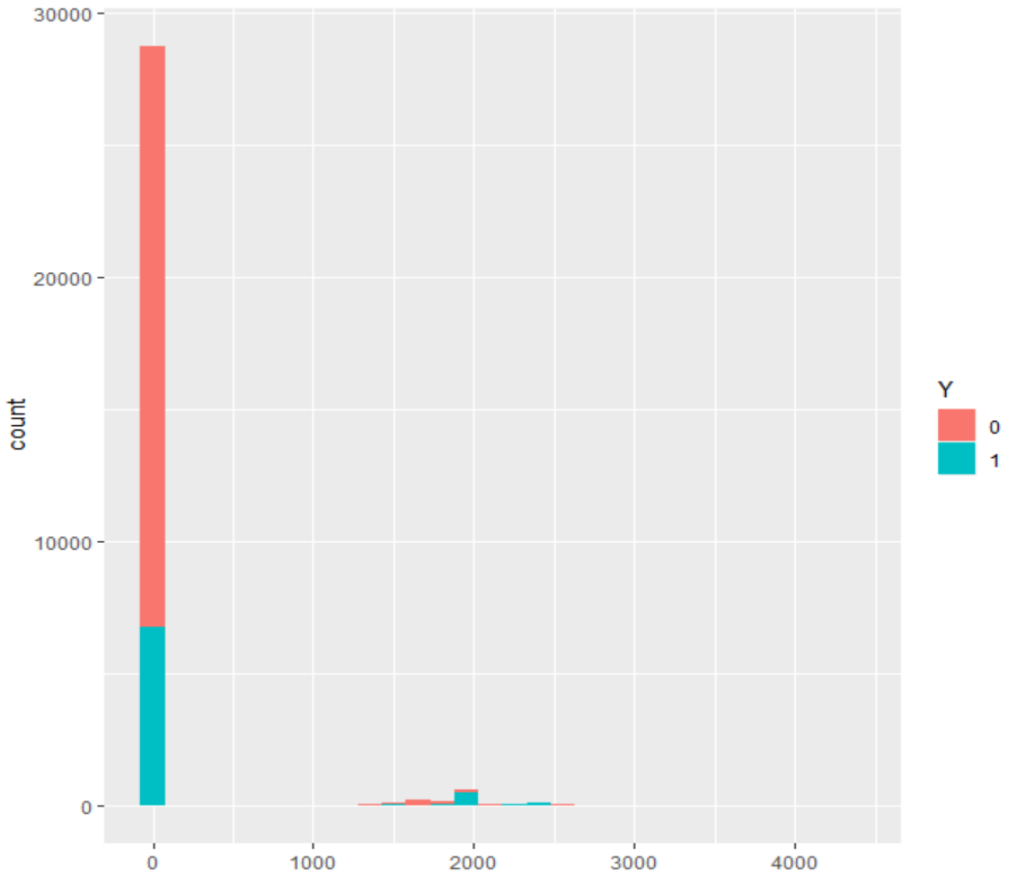
Native.country

미국을 제외한 나머지 국가들의 케이스가 상대적으로 작기 때문에,
"Etc"라는 하나의 범주로 변환

#2 데이터 전처리 – 정제



Capital.gain



Capital.loss

대부분의 케이스(98% 이상)가 자본 획득과 자본 손실이 0으로 나타나,
필요 없는 변수로 판단 되어 변수 삭제

#3 데이터마이닝 방법선택

$50,000을 넘는지 안 넘는지에 대한 "분류문제"

* 시도한 알고리즘

K-최근접이웃 알고리즘
분류나무
나이브 베이즈
신경망

* 모델선택기준

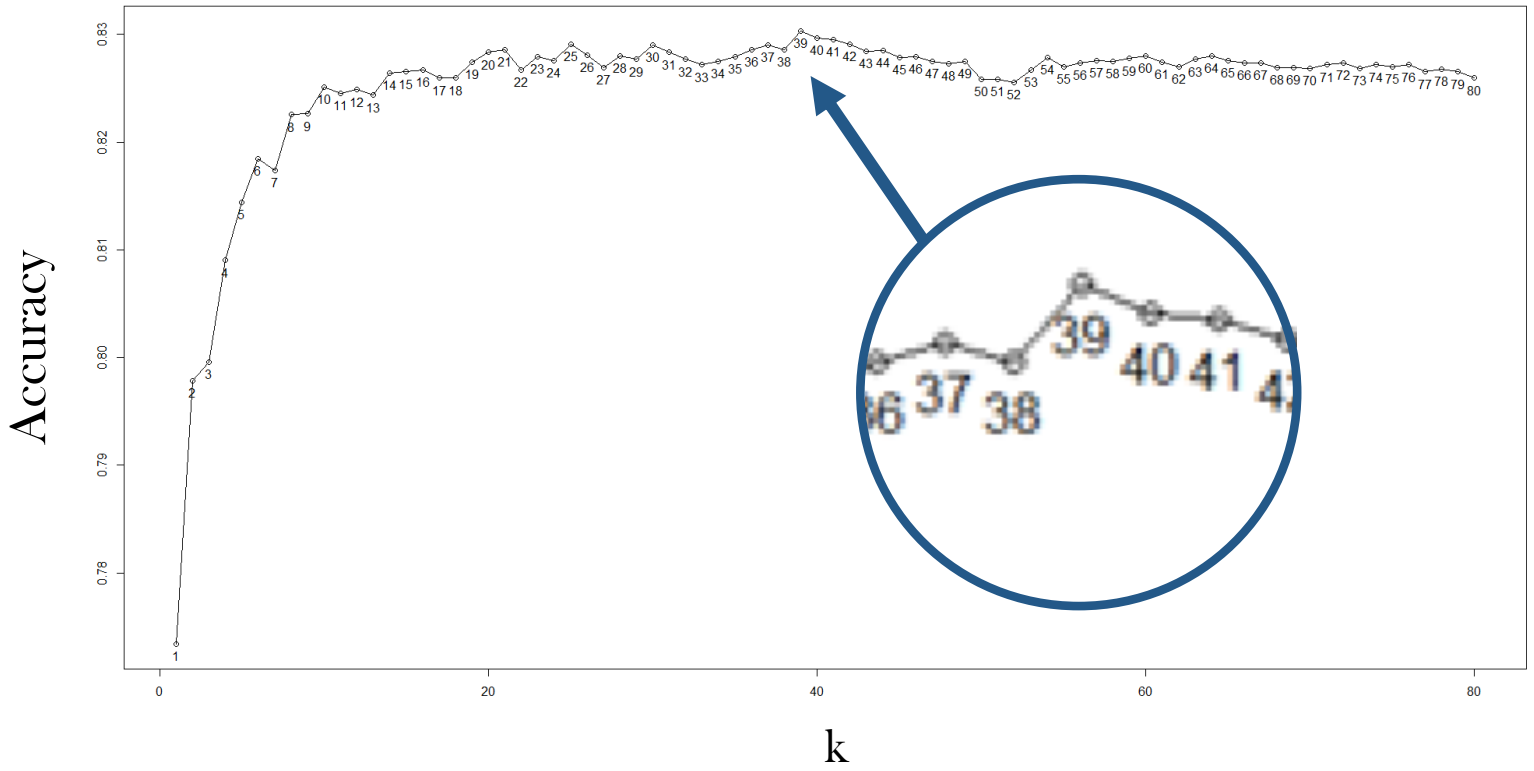
1. 정확도 (Accuracy) =(A+D)/(A+B+C+D)
2. 특이도 (Specificity) = A/(A+C)

참고)

prediction	Reference	
	0	1
0	A	B
1	C	D

#3 데이터마이닝 방법선택 kNN

검증데이터에 대한 정확도가 가장 높은 홀수인 k = 39



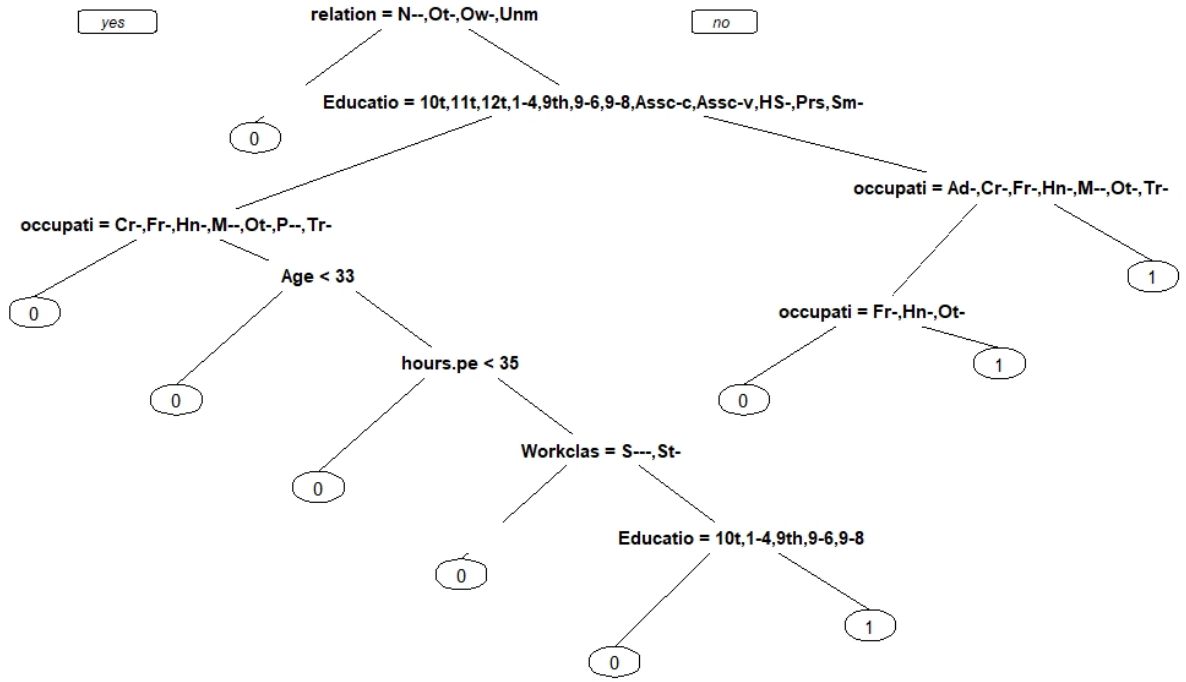
prediction \ Reference	0	1
0	8191	1193
1	854	1827

0 : 연 소득 $50,000 이하
1 : 연 소득 $50,000 초과

1. 정확도 : 0.8303

2. 특이도 : 0.9055832

#3 데이터마이닝 방법선택 분류나무



CP (복잡도 파라미터) : 0.00258253
잎노드의 개수 : 10

Reference

prediction	0	1
0	8207	1265
1	803	1790

0 : 연 소득 $50,000 이하
1 : 연 소득 $50,000 초과

1. 정확도 : 0.8286

2. 특이도 : 0.9108768

#3 데이터마이닝 방법선택 나이브 베이즈

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1
0.7526525 0.2473475

Conditional probabilities:
   workclass
Y          change   Federal-gov     Local-gov Never-worked      Private Self-emp-inc Self-emp-not-inc     State-gov  Without-pay
  0 0.0000000000 0.0261380323 0.0649045521 0.0000000000 0.7649045521 0.0220998532     0.0791483113 0.0420704846 0.0007342144
  1 0.0000000000 0.0460232350 0.0810991957 0.0000000000 0.6548257373 0.0779714030     0.0918230563 0.0482573727 0.0000000000

   education
Y           10th         11th         12th      1st-4th          9th      9th-6th      9th-8th    Assoc-acdm    Assoc-voc    Bachelors
  0 0.0324522761 0.0435389134 0.0154185022 0.0065345081 0.0195301028 0.0131424376 0.0234948605 0.0328193833 0.0403083700 0.1300293686
  1 0.0078194817 0.0091599643 0.0031277927 0.0006702413 0.0029043789 0.0011170688 0.0055853441 0.0332886506 0.0462466488 0.2768096515
   education
Y       Doctorate      HS-grad      Masters    Preschool  Prof-school Some-college
  0 0.0038913363 0.3650513950 0.0324522761 0.0017621145 0.0061674009 0.2334067548
  1 0.0357462020 0.2216264522 0.1262287757 0.0000000000 0.0549597855 0.1747095621

   marital_status
Y       Divorced Married-AF-spouse Married-civ-spouse Married-spouse-absent Never-married    Separated      Widowed
  0 0.1618942731      0.0003671072       0.3388399413          0.0149779736  0.4100587372 0.0408223201 0.0330396476
  1 0.0589812332      0.0015638963       0.8507596068          0.0051385165  0.0638963360 0.0091599643 0.0105004468

   occupation
Y   Adm-clerical Armed-Forces       change Craft-repair Exec-machangegerial Farming-fishing Handlers-cleaners Machine-op-inspct Other-service
  0 0.1410425844 0.0002202643 0.0000000000 0.1393538913        0.0875917768    0.0395741557      0.0544052863     0.0748164464  0.1387665198
  1 0.0656836461 0.0002234138 0.0000000000 0.1255585344        0.2587131367    0.0138516533      0.0111706881     0.0308310992  0.0167560322
   occupation
Y   Priv-house-serv Prof-specialty Protective-serv        Sales Tech-support Transport-moving
  0    0.0066813510    0.1011747430     0.0195301028 0.1137298091 0.0260646109     0.0570484581
  1    0.0000000000    0.2432975871     0.0285969616 0.1251117069 0.0352993744     0.0449061662
```

Ex) $P(Y=1|Education = MASTER, Occutation = Prof-Specialty) =$

$0.2473475 \times 0.1262287757 \times 0.2432975871$

$P(Y=0|Education = MASTER, Occutation = Prof-Specialty) =$

$0.0324522761 \times 0.1011747430 \times 0.7526525$

나이브 베이즈로 계산한 확률 : 0.754536

Reference

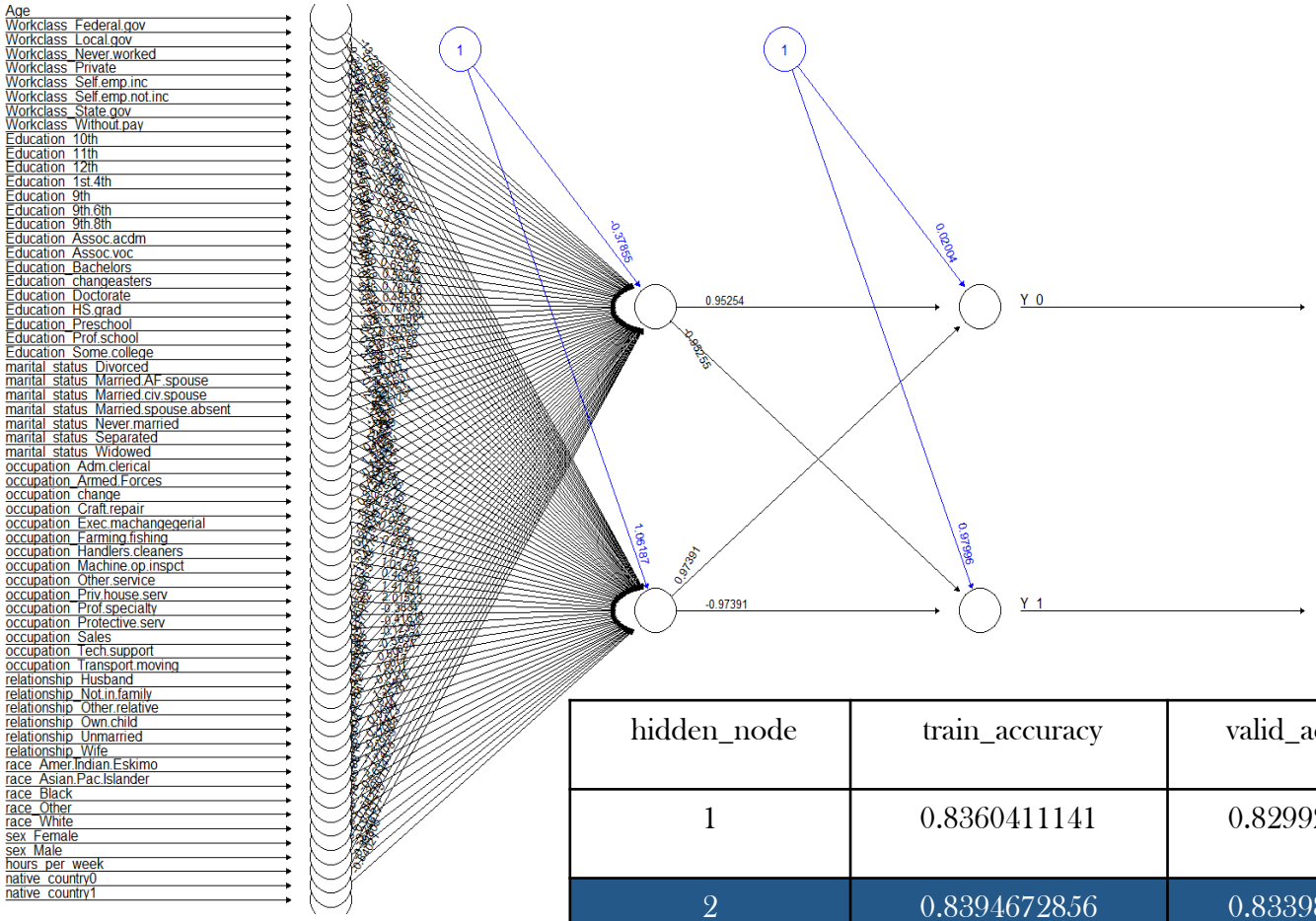
prediction	0	1
0	7319	796
1	1714	2236

0 : 연 소득 $50,000 이하
1 : 연 소득 $50,000 초과

1. 정확도 : 0.792

2. 특이도 : 0.8102513

#3 데이터마이닝 방법선택 신경망



hidden_node	train_accuracy	valid_accuracy
1	0.8360411141	0.8299212598
2	0.8394672856	0.8339825943
3	0.8454354553	0.8276833817
4	0.8432250221	0.8317447161

Reference

prediction	0	1
0	8271	1229
1	774	1791

0 : 연 소득 $50,000 이하
1 : 연 소득 $50,000 초과

1. 정확도 : 0.8339826

2. 특이도 : 0.9144279

#4 최종모델 결정

kNN

	0	1
0	8191	1193
1	854	1827

1. 정확도 : 0.8303

2. 특이도 : 0.9055832

나이브 베이즈

	0	1
0	7319	796
1	1714	2236

1. 정확도 : 0.792

2. 특이도 : 0.9055832

분류나무

	0	1
0	8207	1265
1	803	1790

1. 정확도 : 0.8286

2. 특이도 : 0.8102513

신경망

	0	1
0	8271	1229
1	774	1791

1. 정확도 : 0.8339826

2. 특이도 : 0.9144279

모델 선택 기준 2가지를 바탕으로, 가장 적절한 신경망 기법으로 결정

#5 새로운 사례 적용

새로운 관측치

나이	35
종사형태	비자영업
학력	박사
혼인유무	기혼
직업	교수
가족관계	아내
인종	기타
성별	여성
주당근무시간	52
국적	기타

실제로 모델에 적용

```
> predict2_lee$net.result
             [,1]         [,2]
[1,] 0.1831098334 0.8168915826
> class2_lee
[1] 1
```

모델 적용 결과,
새로운 관측치는 약 0.8의 경향값으로,
이는 컷오프값 0.5 기준으로
클래스 1 (연 소득 $50,000 초과)에
속한다고 할 수 있다.

Thank you.