통계적기계학습

분류 모델 간 성능 비교

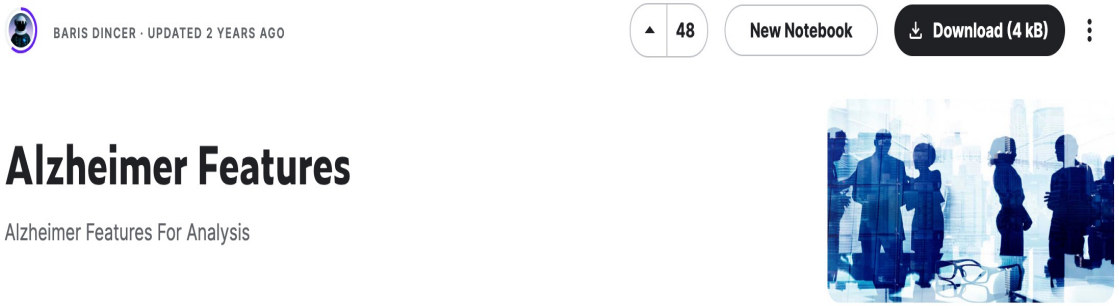
컴퓨터전자시스템공학부

202002203 유승리

목차

Kaggle – Alzheimer Features

```
> df <- read.csv("/Users/yooseungli/Downloads/alzheimer.csv", sep = ",", head = T)
> summary(df)
    Group               M.F                 Age             EDUC            SES
 Length:373         Length:373         Min.   :60.00   Min.   : 6.0   Min.   :1.00
 Class :character   Class :character   1st Qu.:71.00   1st Qu.:12.0   1st Qu.:2.00
 Mode  :character   Mode  :character   Median :77.00   Median :15.0   Median :2.00
                                       Mean   :77.01   Mean   :14.6   Mean   :2.46
                                       3rd Qu.:82.00   3rd Qu.:16.0   3rd Qu.:3.00
                                       Max.   :98.00   Max.   :23.0   Max.   :5.00
                                                                      NA's   :19
      MMSE            CDR              eTIV           nWBV            ASF
 Min.   : 4.00   Min.   :0.0000   Min.   :1106   Min.   :0.6440   Min.   :0.876
 1st Qu.:27.00   1st Qu.:0.0000   1st Qu.:1357   1st Qu.:0.7000   1st Qu.:1.099
 Median :29.00   Median :0.0000   Median :1470   Median :0.7290   Median :1.194
 Mean   :27.34   Mean   :0.2909   Mean   :1488   Mean   :0.7296   Mean   :1.195
 3rd Qu.:30.00   3rd Qu.:0.5000   3rd Qu.:1597   3rd Qu.:0.7560   3rd Qu.:1.293
 Max.   :30.00   Max.   :2.0000   Max.   :2004   Max.   :0.8370   Max.   :1.587
 NA's   :2
```



- Alzheimer's Disease(AD)에 대한 정보를 포함하고 있는 데이터셋
- 373명의 실험자
- 9가지 특징
  ⇒ Converted / Demented / Nondemented 분류

- M.F : 성별
- Age : 연령
- EDUC : 교육 수준
- SES : 사회경제적 지위 (1 ~ 5)
- MMSE : 미니멘탈 상태 검사 (Mini Mental State Examination)
  * 인지 기능 평가
- CDR : 임상 치매 등급 (Clinical Dementia Rating)
  * 치매의 정도 평가
- eTIV : 추정 총 두뇌 내부 부피 (Estimated total intracranial volume)
- nWBV : 정규화된 전체 뇌 부피 (Normalized Whole Brain Volume)
- ASF : 아틀라스 스케일링 팩터 (Atlas Scaling Factor)
  * 뇌 부피 조정

결측치 처리

```
       Group      M.F          Age             EDUC          SES
 Converted  : 37   F:213   Min.   :60.00   Min.   : 6.0   Min.   :1.00
 Demented   :146   M:160   1st Qu.:71.00   1st Qu.:12.0   1st Qu.:2.00
 Nondemented:190           Median :77.00   Median :15.0   Median :2.00
                           Mean   :77.01   Mean   :14.6   Mean   :2.46
                           3rd Qu.:82.00   3rd Qu.:16.0   3rd Qu.:3.00
                           Max.   :98.00   Max.   :23.0   Max.   :5.00
                                                          NA's   :19
      MMSE            CDR              eTIV           nWBV            ASF
 Min.   : 4.00   Min.   :0.0000   Min.   :1106   Min.   :0.6440   Min.   :0.876
 1st Qu.:27.00   1st Qu.:0.0000   1st Qu.:1357   1st Qu.:0.7000   1st Qu.:1.099
 Median :29.00   Median :0.0000   Median :1470   Median :0.7290   Median :1.194
 Mean   :27.34   Mean   :0.2909   Mean   :1488   Mean   :0.7296   Mean   :1.195
 3rd Qu.:30.00   3rd Qu.:0.5000   3rd Qu.:1597   3rd Qu.:0.7560   3rd Qu.:1.293
 Max.   :30.00   Max.   :2.0000   Max.   :2004   Max.   :0.8370   Max.   :1.587
 NA's   :2
```

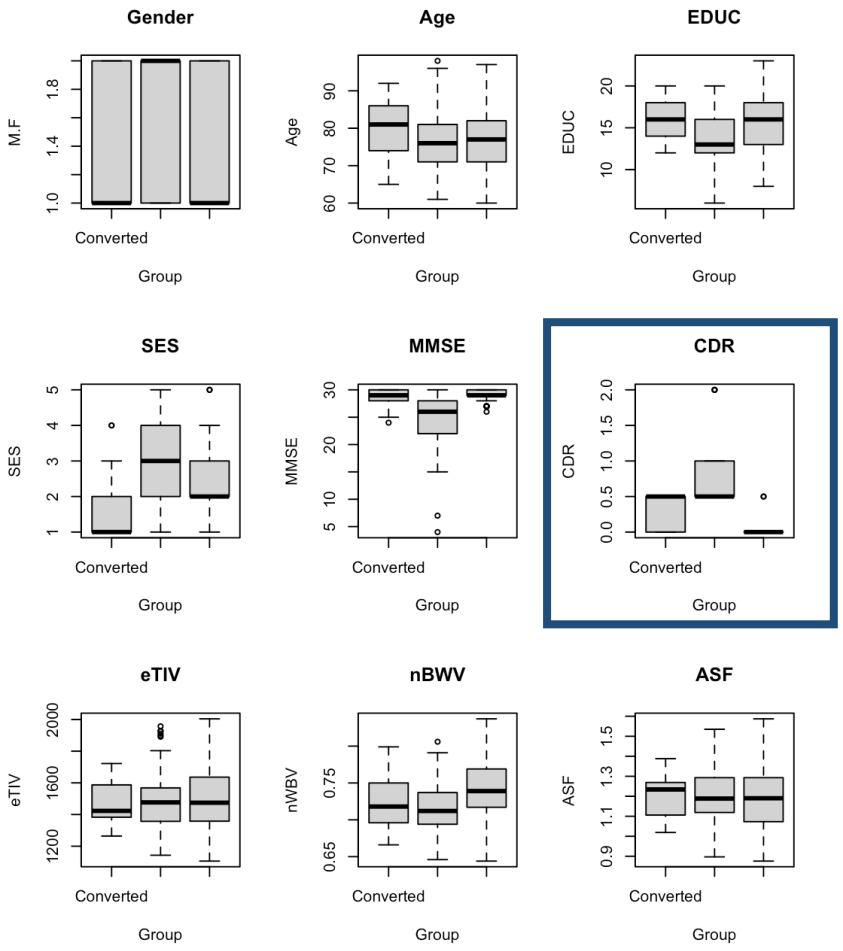
```
       Group      M.F          Age             EDUC          SES             MMSE            CDR
 Converted  : 37   F:213   Min.   :60.00   Min.   : 6.0   Min.   :1.000   Min.   : 4.00   Min.   :0.0000
 Demented   :146   M:160   1st Qu.:71.00   1st Qu.:12.0   1st Qu.:1.833   1st Qu.:27.00   1st Qu.:0.0000
 Nondemented:190           Median :77.00   Median :15.0   Median :2.000   Median :29.00   Median :0.0000
                           Mean   :77.01   Mean   :14.6   Mean   :2.464   Mean   :27.33   Mean   :0.2909
                           3rd Qu.:82.00   3rd Qu.:16.0   3rd Qu.:3.000   3rd Qu.:30.00   3rd Qu.:0.5000
                           Max.   :98.00   Max.   :23.0   Max.   :5.000   Max.   :30.00   Max.   :2.0000
     eTIV           nWBV            ASF
 Min.   :1106   Min.   :0.6440   Min.   :0.876
 1st Qu.:1357   1st Qu.:0.7000   1st Qu.:1.099
 Median :1470   Median :0.7290   Median :1.194
 Mean   :1488   Mean   :0.7296   Mean   :1.195
 3rd Qu.:1597   3rd Qu.:0.7560   3rd Qu.:1.293
 Max.   :2004   Max.   :0.8370   Max.   :1.587
```

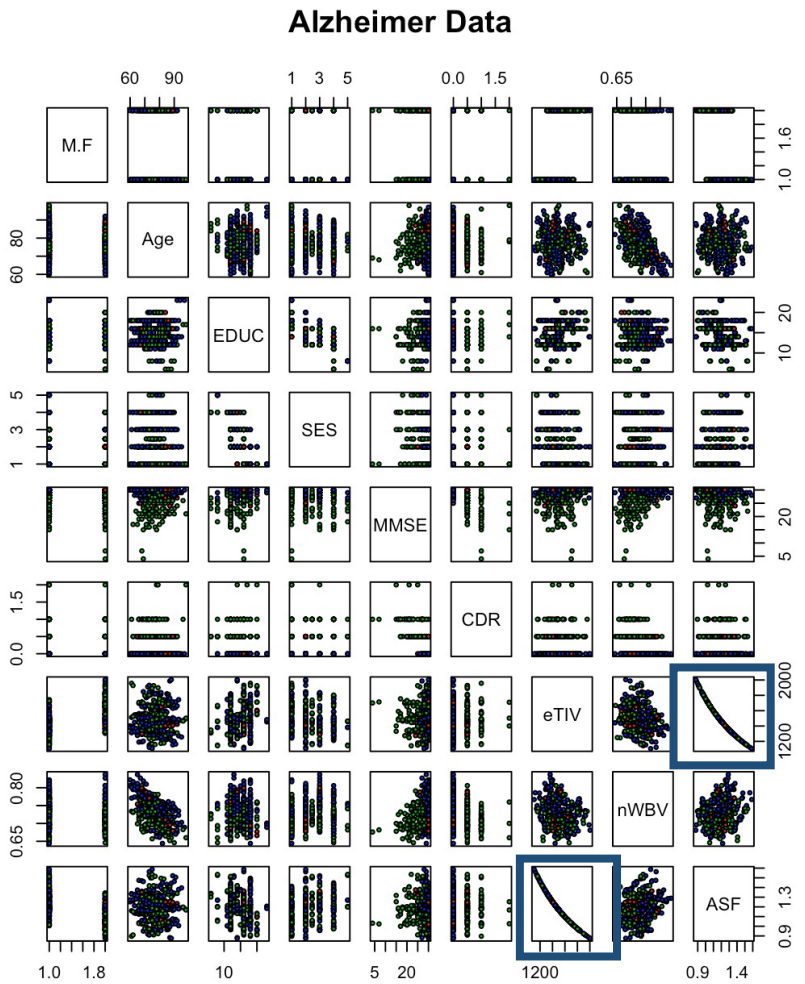
- 데이터의 수가 많지 않기 때문에
  결측치가 있는 행을 아예 제거하는 것보다
  다른 값을 통해 결측치를 대체하는 방법을 선택함

  i. 선형 보간법
  → 로지스틱 회귀 : 정확도가 약간 증가
  SVM과 Random Forest : 정확도 감소

  ii. 평균으로 결측치 채우기 ☑

데이터 분포 확인





- SES, MMSE, CDR을 제외한 변수들은 세 그룹의 분포 형태가 비슷함

- **CDR** (임상 치매 등급)
  - Demented : 0.5 ~ 1.0
  - Nondemented : 0.0
  - Converted : 0.0 ~ 0.5

⇒ CDR이 클래스 분류에 큰 역할을 할 것으로 추측

- eTIV와 ASF의 상관계수 : -0.98887652
  → 뇌 전체 부피가 클수록 아틀라스 스케일링 팩터는 감소한다

분류 모델 학습

1. 로지스틱 회귀

Confusion Matrix and Statistics

Prediction	Converted	Demented	Nondemented
Converted	3	0	0
Demented	1	48	0
Nondemented	6	0	54

Accuracy : 0.9375
95% CI : (0.8755, 0.9745)
No Information Rate : 0.4821
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.8867

Mcnemar's Test P-Value : NA

Statistics by Class:

	Class: Converted	Class: Demented	Class: Nondemented
Sensitivity	0.30000	1.0000	1.0000
Specificity	1.00000	0.9844	0.8966
Pos Pred Value	1.00000	0.9796	0.9000
Neg Pred Value	0.93578	1.0000	1.0000
Prevalence	0.08929	0.4286	0.4821
Detection Rate	0.02679	0.4286	0.4821
Detection Prevalence	0.02679	0.4375	0.5357
Balanced Accuracy	0.65000	0.9922	0.9483

2. SVM

Confusion Matrix and Statistics

Prediction	Converted	Demented	Nondemented
Converted	2	2	0
Demented	2	45	0
Nondemented	6	1	54

Accuracy : 0.9018
95% CI : (0.8311, 0.9499)
No Information Rate : 0.4821
P-Value [Acc > NIR] : <2e-16

Kappa : 0.8228

Mcnemar's Test P-Value : 0.0719

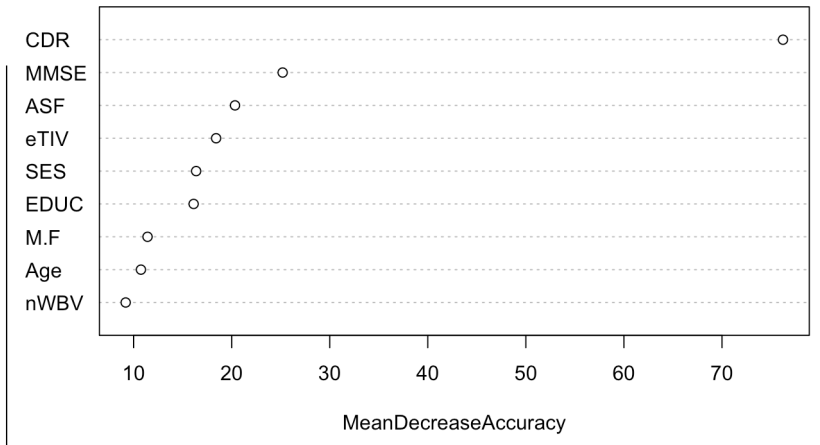
Statistics by Class:

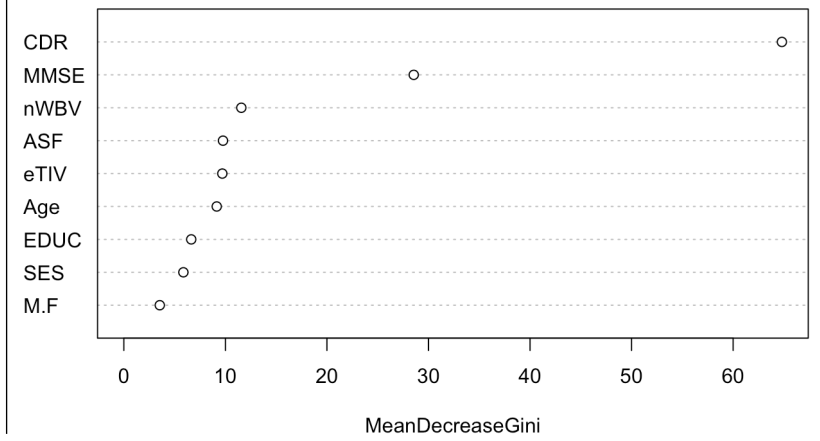
	Class: Converted	Class: Demented	Class: Nondemented
Sensitivity	0.20000	0.9375	1.0000
Specificity	0.98039	0.9688	0.8793
Pos Pred Value	0.50000	0.9574	0.8852
Neg Pred Value	0.92593	0.9538	1.0000
Prevalence	0.08929	0.4286	0.4821
Detection Rate	0.01786	0.4018	0.4821
Detection Prevalence	0.03571	0.4196	0.5446
Balanced Accuracy	0.59020	0.9531	0.9397

분류 모델 학습

3. **Random Forest**

```
Confusion Matrix and Statistics

                Reference
Prediction      Converted Demented Nondemented
  Converted             3        1           0
  Demented              2       47           0
  Nondemented           5        0          54

               Accuracy : 0.9286
                 95% CI : (0.8641, 0.9687)
    No Information Rate : 0.4821
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8714

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Converted Class: Demented Class: Nondemented
Sensitivity                   0.30000          0.9792              1.0000
Specificity                   0.99020          0.9688              0.9138
Pos Pred Value                0.75000          0.9592              0.9153
Neg Pred Value                0.93519          0.9841              1.0000
Prevalence                    0.08929          0.4286              0.4821
Detection Rate                0.02679          0.4196              0.4821
Detection Prevalence          0.03571          0.4375              0.5268
Balanced Accuracy             0.64510          0.9740              0.9569
```





- Variable Importance

정확도, 지니계수 측면에서
정도에 차이가 있지만
CDR, MMSE가 클래스 분류에
가장 기여도가 높은 변수로 나타남

결과 해석

1. 로지스틱 회귀
정확도 : 0.9375
Demented / Nondemented 분류 시에는 모두 정확히 분류함

2. SVM
정확도 : 0.9018

3. Random Forest
정확도 : 0.9286
Demented / Nondemented 분류 시에는 SVM 3건, RF 1건 제외하고는 모두 정확히 분류함

⇒ 정확도와 Demented / Nondemented 분류에서는 로지스틱 회귀의 성능이 가장 좋지만
Converted는 로지스틱 회귀, SVM, RF 모두 제대로 분류하지 못함

전체 데이터 중 Converted는 10% 밖에 되지 않기 때문에
편향된 데이터로 인해 제대로 학습이 이루어지지 않은 것으로 예상됨

추가 진행 사항

Converted	Demented	Nondemented
27	98	136

Converted	Demented	Nondemented
0	0	0

```
install.packages("reticulate")
reticulate::py_install("imbalanced-learn")
library(reticulate)
imblearn <- import("imblearn")
```

- 언더샘플링을 통해 데이터 불균형을 해소하고 다시 모델 학습을 진행한 후 결과를 비교해보고자 했음

⇒ Converted에 맞춰서 언더샘플링한 결과, 모든 데이터 수가 0이 되어 학습 진행 불가

감사합니다