P-SAT

3주차 패키지

- 분석 툴은 R/Python 둘 다 가능합니다. 1-3주차 패키지 문제의 조건 및 힌트는 R을 기준으로 하지만, Python을 사용해도 무방합니다.
- 제출형식은 피피티(템플릿은 P-SAT기본 템플릿), 마크다운, HTML, PDF 모두 됩니다. .R이나 .ipynb 등의 소스코드 파일은 안됩니다. 완료 후 psat2009@naver.com로 보내주세요.
- 패키지 과제 발표는 세미나 쉬는시간 이후에 하게 되며, 역시 랜덤으로 5시00분에 발표됩니다.
- 제출기한은 목요일 자정까지이며 지각시 벌금 5000원이 있습니다. 미제출 시 만원입니다. 패키지 2회 무단 미제출 시 퇴출이니 유의해 주세요.

Chapter 1. 모델링을 위한 전처리

이번주는 범주형 자료형이 대부분인 데이터의 'stroke'의 여부를 분류하는 모델링을 해보겠습니다. Chapter1에서는 모델링을 위한 전처리 및 간단한 시각화를 통해 데이터에 대한 이해를 해봅시다.

[조건: tidyverse, data.table, gridExtra 패키지 모두 사용(이외 패키지 금지), %>% 최대 활용]

문제 0 기본 세팅. tidyverse, data.table, gridExtra를 부른 후, setwd로 'data.csv' 및 'test.csv'가 있는 폴더로 경로를 설정하고, fread로 'data.csv'와 'test.csv'를 불러오세요.

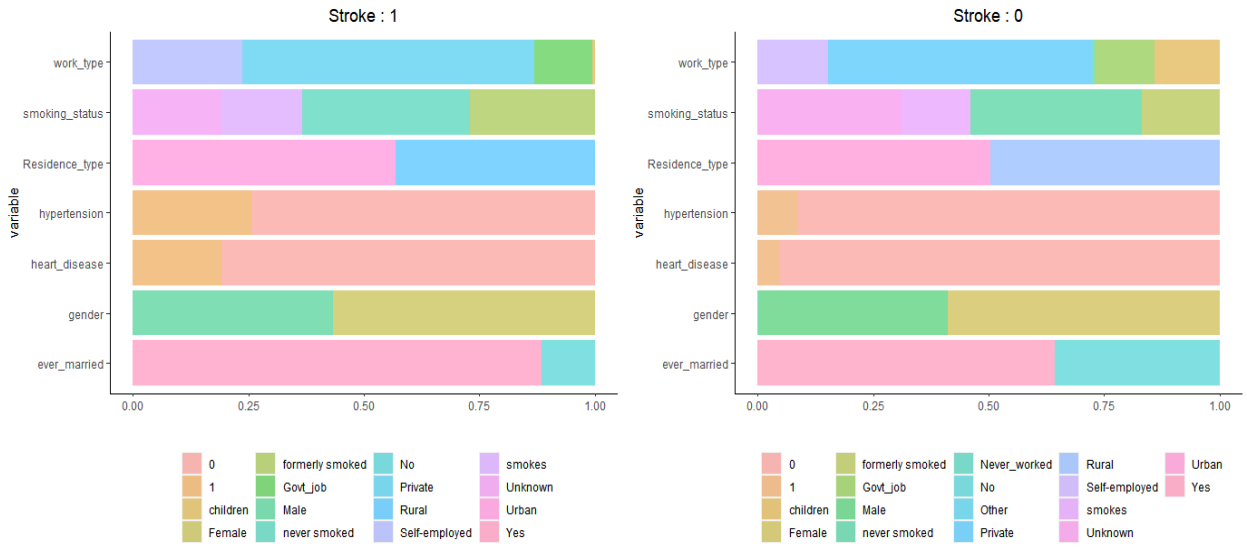
[Train data(data.csv) 전처리 및 EDA]

문제 1. 'bmi(bmi 지수)' 변수를 numeric 자료형으로 바꾸고, NA값을 mean imputation으로 채우세요.

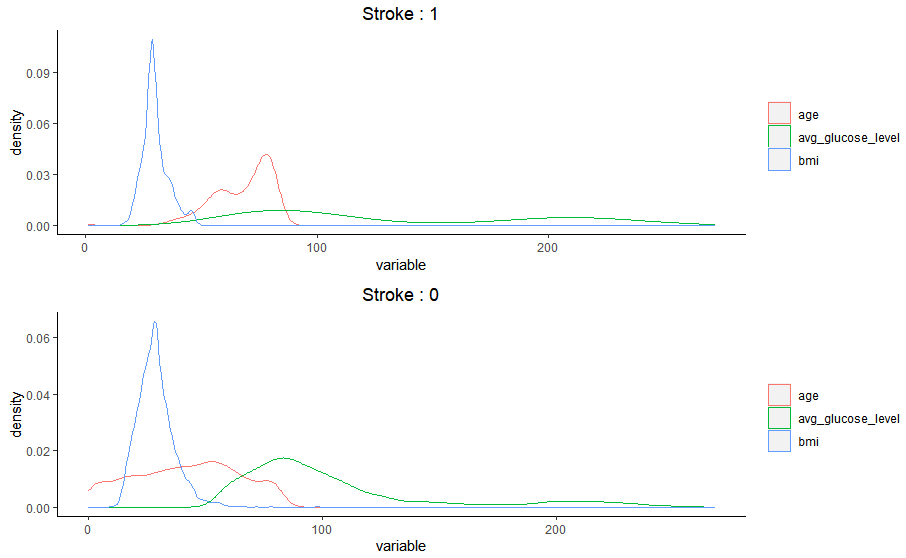
문제 2. 문자형(character) 변수를 명목형 변수(factor)로 바꾸세요.

문제 3. 'id' 변수를 제거하세요.

문제 4. 타겟(stoke)값 별로 범주형 변수의 분포를 다음과 같이 시각화 하고, 간단히 해석해보세요. (gather, gridExtra 패키지 내 함수 이용).



문제 5. 타겟(stoke)값 별로 수치형 변수의 분포를 다음과 같이 시각화 하고, 간단히 해석해보세요. (gather, gridExtra 패키지 내 함수 이용).



문제 6. 타겟 변수와 범주형 변수에 대한 카이스퀘어 독립성 검정을 진행하고 다음과 같이 출력하세요. (데이터프레임을 만든 후, for 문으로 독립성 검정을 진행하여 chi 변수에 검정 결과를 넣으세요.)

```
        cate_var    chi
1         gender accept
2   hypertension denied
3  heart_disease denied
4   ever_married denied
5      work_type denied
6 Residence_type accept
7 smoking_status denied
```

문제 7. 카이스퀘어 독립성 검정에서 가설을 기각하지 못한 범주형 변수를 제거하세요.

[Test data(test.csv) 전처리]

문제 8. train data에서 했던 전처리 방법들을 사용하여 전처리 하세요.

-'bmi' 변수의 자료형을 바꾸고 mean imputation하기

-문자형 변수 자료형 바꾸기

-'id' 및 train 데이터에서 제거된 범주형 변수 제거하기

Chapter 2. Catboost

부스팅 모델 중 Catboost 모델은 범주형 변수가 많을 때 선호되는 모델입니다. 이번 챕터에서는 **Grid Search 5-fold CV**를 통해서 파라미터를 튜닝하고, Test set의 Logloss를 평가해 봅시다.

[조건: catboost, caret, MLmetrics 패키지 모두 사용]

**catboost 라이브러리는 먼저 devtools 패키지를 설치한 후 아래 코드를 실행하여 설치해야 합니다.

```
devtools::install_url('https://github.com/catboost/catboost/releases/download/v0.21/catboost-R-Windows-0.21.tgz',
INSTALL_opts = c("--no-multiarch"))
```

문제 0. Catboost 모델의 특성 및 대표적인 파라미터에 대해 간단히 설명하세요.

문제 1. expand.grid를 사용하여 다음과 같은 데이터 프레임을 만드세요. (데이터 프레임명: logloss_cb)

```
> logloss_cb
  depth iterations logloss
1     4        100      NA
2     6        100      NA
3     8        100      NA
4     4        200      NA
5     6        200      NA
6     8        200      NA
```

문제 2. Catboost에 대해 depth와 iteration 파라미터 튜닝을 위한 grid search 5-fold CV를 진행하세요.

<조건>

-튜닝 기준: LogLoss

-시드 넘버: 1234

-Sys.time을 이용하여 걸린 시간을 측정할 것

<힌트>

-catboost.load_pool을 이용하여 모델에 맞는 데이터 형식으로 바꾸고 학습시킬 것

문제 3. logloss_cb에서 가장 낮은 logloss 값의 행을 출력하세요.

문제 4. 가장 낮은 logloss 값의 파라미터로 전체 데이터를 학습시켜 test set에 대한 logloss값을 구하세요.

Chapter 3. K-means Clustering

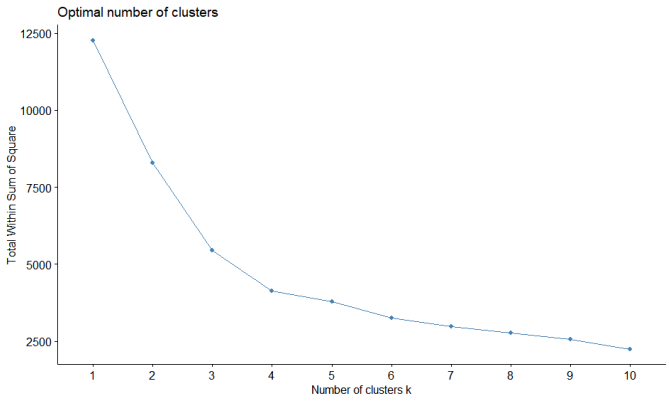
지금까지는 지도학습에 대해 알아보았습니다. 이번 챕터에서는 비지도학습의 대표적인 모델인 K-means Clustering을 해보겠습니다. 거리를 기반으로 비슷한 데이터를 묶어 주는 군집화 방법으로, 적절한 설명변수가 없을 때 변수들을 해석하기 위해 자주 사용합니다. 자세한 내용은 이번주 데이터마이닝 클린업에서 다룰 예정입니다.

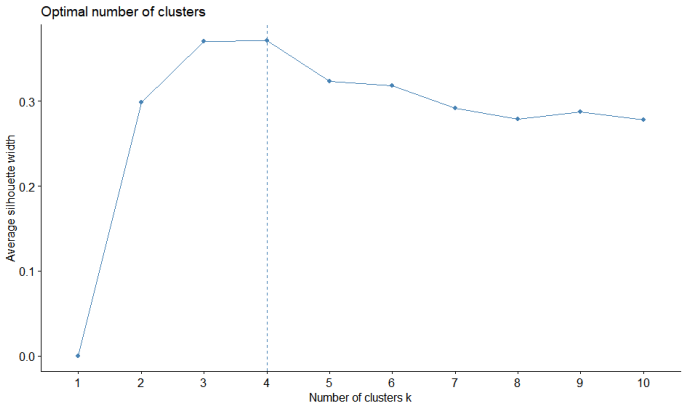
데이터는 앞서 전처리된 'data.csv'의 수치형 변수를 사용하도록 하겠습니다.

[조건: factoextra, cluster 패키지 모두 사용]

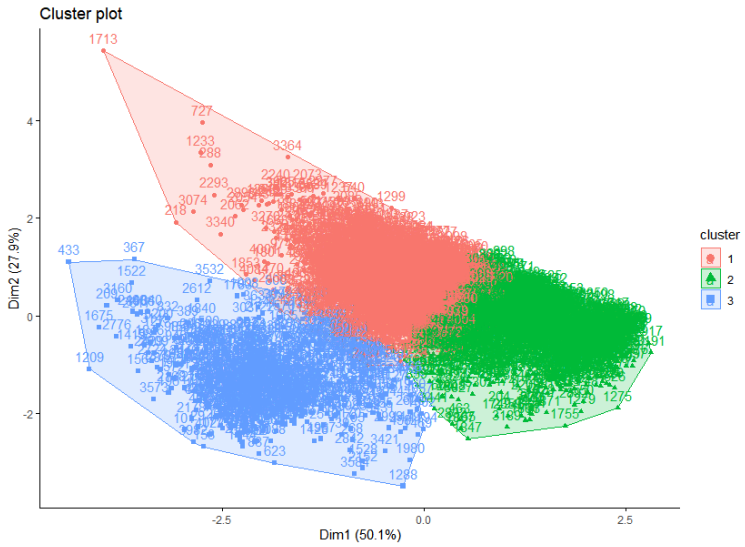
문제 1. 수치형 변수(age, avg_glucose_level, bmi)에 대해 scale 함수로 정규화 스케일링을 하세요.

문제 2. fviz_nbclust 함수로 다음과 같이 시각화 한 뒤, 적절한 K값이 무엇인지 설명하세요.





문제 3. K-means 클러스터링을 한 후, 다음과 같이 시각화하세요. (nstart = 1, iter.max =30, seed: 1234)



문제 4. 사용한 변수인 age, avg_glucose_level(평균 혈당), bmi(bmi 수치)에 대해 다음과 같이 box_plot 시각화를 하고, 클러스터 별로 해석해보세요. (사용 색 : #845ec2, #ffc75f, #ff5e78)