Untitled

# 기계학습(50점)  
# 1.1 데이터 탐색  
# 1.1.1 탐색적 데이터 분석 수행하시오(시각화 포함)  
# 1.1.2 이상치 처리하시오  
# 1.1.3 앞선 두 단계에서 얻은 향후 분석시 고려사항 작성  
  
# 2.1 클래스 불균형을 처리하시오  
# 2.1.1 업 샘플링 과정 설명하고 결과 작성  
# 2.2.2 언더 샘플링 과정 설명하고 결과 작성  
# 2.2.3 둘 중 선택하고 이유 설명  
  
# 3.1 모델링 하시오  
# 3.1.1 최소 3개 이상 알고리즘 제시하고   
# 정확도 측면의 모델 1개와 속도 측면의 모델 1개를 꼭 구현(총 2개 이상)  
# 3.1.2 모델 비교하고 결과 설명  
# 3.1.3 속도 개선을 위한 차원 축소 설명하고 수행, 예측 성능과 속도 비교하고 결과 작성  
  
# 2. 통계분석(50점)  
  
# 1.1 금속 성분 함유량 변수 1개. (1열 데이터)   
# 제품에 금속 재질 함유량의 분산이 1.3을 넘으면 불량이라고 보고있는데,   
# 제조사별로 차이가 난다고 제보를 받음 분산에 대해 검정을 수행하시오.  
# 1.1.1 연구가설과 귀무가설 작성  
# 1.1.2 양측 검정 어쩌고  
# 1.1.3 검정통계량, 가설 채택  
  
  
# 2.1 Lot별 불량 제품 수량 데이터. lot 번호와 불량제품수 두 개의 열. 각 lot별 200개에 대한 불량제품 수.  
# 2.1.1 p관리도에 따라 관리중심선(center line), 관리 상한선, 하한선 구하시오  
# 2.1.2 관리도 시각화 하시오  
  
# 3.1 데이터 없음. 표에 제품 1,2를 만드는데 사용되는 재료 a b c 컬럼 있고   
# 재료에 따라 최종 만들어지는 제품 두 개에 대한 수량 있음.   
# 최하단 행에는 수익이 있음.   
# 제품 수량을 최대로 뽑으면서 수익이 최적이 되도록 하라고 함.(10점)  
  
# 4.1 데이터 없음. 상품 a와 b가 있을 때 다음과 같은 구매 패턴이 있다고 함.   
# aa bb bbbb aa aaa bb bbb aa bb a b 정확히 기억 안나지만 대충 비슷함.   
# 4.1.1 구매하는 패턴으로 봐서 두 상품이 연관이 있는지 가설 세우고 검정하시오  
# 4.1.2 연구가설 귀무가설 세우시오  
# 4.1.3 가설 채택하시오  
  
  
############################ 문제 풀이 ##########################  
# 1.1 데이터 탐색  
# 1.1.1 탐색적 데이터 분석 수행하시오(시각화 포함)  
# 1.1.2 이상치 처리하시오  
# 1.1.3 앞선 두 단계에서 얻은 향후 분석시 고려사항 작성  
  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.4 v dplyr 1.0.7  
## v tidyr 1.1.4 v stringr 1.4.0  
## v readr 2.0.2 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(magrittr)

##   
## 다음의 패키지를 부착합니다: 'magrittr'

## The following object is masked from 'package:purrr':  
##   
## set\_names

## The following object is masked from 'package:tidyr':  
##   
## extract

library(caret)

## 필요한 패키지를 로딩중입니다: lattice

##   
## 다음의 패키지를 부착합니다: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(recipes)

##   
## 다음의 패키지를 부착합니다: 'recipes'

## The following object is masked from 'package:stringr':  
##   
## fixed

## The following object is masked from 'package:stats':  
##   
## step

library(funModeling)

## 필요한 패키지를 로딩중입니다: Hmisc

## 필요한 패키지를 로딩중입니다: survival

##   
## 다음의 패키지를 부착합니다: 'survival'

## The following object is masked from 'package:caret':  
##   
## cluster

## 필요한 패키지를 로딩중입니다: Formula

##   
## 다음의 패키지를 부착합니다: 'Hmisc'

## The following objects are masked from 'package:dplyr':  
##   
## src, summarize

## The following objects are masked from 'package:base':  
##   
## format.pval, units

## funModeling v.1.9.4 :)  
## Examples and tutorials at livebook.datascienceheroes.com  
## / Now in Spanish: librovivodecienciadedatos.ai

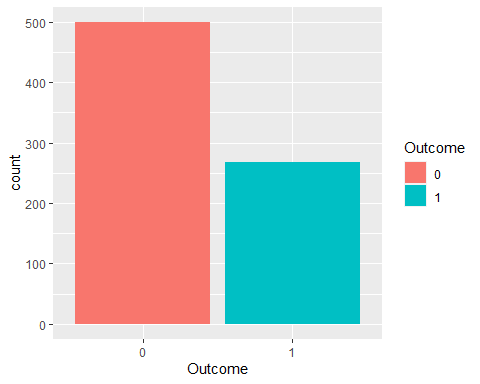
library(knitr)  
  
rm(list=ls())  
setwd("D:/#000.자격증공부/#003. ADP실기/각종주제별코드/R과외/study/ADP기출/")  
  
read.csv("data.csv") -> data   
data %>% names()

## [1] "Pregnancies" "Glucose"   
## [3] "BloodPressure" "SkinThickness"   
## [5] "Insulin" "BMI"   
## [7] "DiabetesPedigreeFunction" "Age"   
## [9] "Outcome"

# [1] "Pregnancies" "Glucose" "BloodPressure"   
# [4] "SkinThickness" "Insulin" "BMI"   
# [7] "DiabetesPedigreeFunction" "Age" "Outcome"  
  
data %>% glimpse()

## Rows: 768  
## Columns: 9  
## $ Pregnancies <int> 6, 1, 8, 1, 0, 5, 3, 10, 2, 8, 4, 10, 10, 1, ~  
## $ Glucose <int> 148, 85, 183, 89, 137, 116, 78, 115, 197, 125~  
## $ BloodPressure <int> 72, 66, 64, 66, 40, 74, 50, 0, 70, 96, 92, 74~  
## $ SkinThickness <int> 35, 29, 0, 23, 35, 0, 32, 0, 45, 0, 0, 0, 0, ~  
## $ Insulin <int> 0, 0, 0, 94, 168, 0, 88, 0, 543, 0, 0, 0, 0, ~  
## $ BMI <dbl> 33.6, 26.6, 23.3, 28.1, 43.1, 25.6, 31.0, 35.~  
## $ DiabetesPedigreeFunction <dbl> 0.627, 0.351, 0.672, 0.167, 2.288, 0.201, 0.2~  
## $ Age <int> 50, 31, 32, 21, 33, 30, 26, 29, 53, 54, 30, 3~  
## $ Outcome <int> 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, ~

data$Outcome %<>% as.factor()  
  
  
# 목적변수 탐색   
data %>% ggplot() + geom\_bar() +   
 aes(x=Outcome, fill = Outcome)



data$Outcome %>% table() %>% prop.table() %>% round(3)

## .  
## 0 1   
## 0.651 0.349

# 탐색 결과   
# 분류 모델이다.   
# 0의 비율이 1에 비해 2배 많다. (Imbalanced data 이다.)