sevvalTASYONAN-hw3

#LIBRARIES

```
# libraries
##install.packages("tidyverse")
library(caret)
```

```
Warning: package 'caret' was built under R version 4.2.3
```

```
Zorunlu paket yükleniyor: ggplot2
```

```
Warning: package 'ggplot2' was built under R version 4.2.3
```

```
Zorunlu paket yükleniyor: lattice
```

```
library(rpart)
library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.2.3
```

```
Warning: package 'tibble' was built under R version 4.2.3
```

```
Warning: package 'tidyr' was built under R version 4.2.3
```

```
Warning: package 'readr' was built under R version 4.2.3
```

```
Warning: package 'purrr' was built under R version 4.2.3
```

```
Warning: package 'dplyr' was built under R version 4.2.3

Warning: package 'stringr' was built under R version 4.2.3

Warning: package 'forcats' was built under R version 4.2.3

Warning: package 'lubridate' was built under R version 4.2.3

-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v lubridate 1.9.2     v tibble    3.2.1
v purrr     1.0.1     v tidyr     1.3.0

-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
x purrr::lift()   masks caret::lift()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
library(ranger)
```

```
Warning: package 'ranger' was built under R version 4.2.3
```

1-) Detail your task with the problem, features, and target.

Here we are analyzing the hotel booking dataset.The dataset contains 32 columns with booking information of two hotels, like how many customers took reservation,arrived,canceled,room types, meals and so on

read data and delete missing values

```
hotel_bookings <- read.csv("hotel_bookings.csv")
hotel_bookings <- na.exclude(hotel_bookings)
```

#split train and test

```r
set.seed(123)
index <- sample(1:nrow(hotel_bookings), round(nrow(hotel_bookings) * 0.8))
train <- hotel_bookings[index, ]
test  <- hotel_bookings[-index, ]
```

logistic regression model with 5-fold cross validation

```r
set.seed(123)
lr_model <- train(
  as.factor(is_canceled) ~ . -country -agent -company -assigned_room_type -reservation_sta
  data = train,
  method = "glm",
  family = "binomial",
  trControl = trainControl(method = "cv", number = 5)
)
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
prediction from a rank-deficient fit may be misleading

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
lr_model
```

```
Generalized Linear Model

95509 samples
   31 predictor
```

```
    2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 76407, 76408, 76406, 76407, 76408
Resampling results:

  Accuracy   Kappa
  0.8094316  0.5665569
```

decision tree model with 5-fold cross validation

```
set.seed(123)
dt_model <- train(
  as.factor(is_canceled) ~ . -country -agent -company -assigned_room_type -reservation_sta
  data = train,
  method = "rpart",
  trControl = trainControl(method = "cv", number = 5)
)
dt_model
```

```
CART

95509 samples
   31 predictor
    2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 76407, 76408, 76406, 76407, 76408
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.01283389  0.8021757  0.55273751
  0.03508871  0.7701686  0.44703431
  0.32536034  0.6526176  0.07469488

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.01283389.
```

Logistic regression model accuracy is 0.8094316 and Decision tree model accuracy is 0.8021757.Because accuracy of logistic model is higher we use this model