Classification with R Machine Learning

load data

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
## Warning: package 'tibble' was built under R version 3.6.3
```

```
## Warning: package 'tidyr' was built under R version 3.6.3
```

```
## Warning: package 'readr' was built under R version 3.6.3
```

```
## Warning: package 'purrr' was built under R version 3.6.3
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
## Warning: package 'stringr' was built under R version 3.6.3
```

```
## Warning: package 'forcats' was built under R version 3.6.3
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 3.6.3
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
houseData = read.csv('../data/housing.csv')
head(houseData)
```

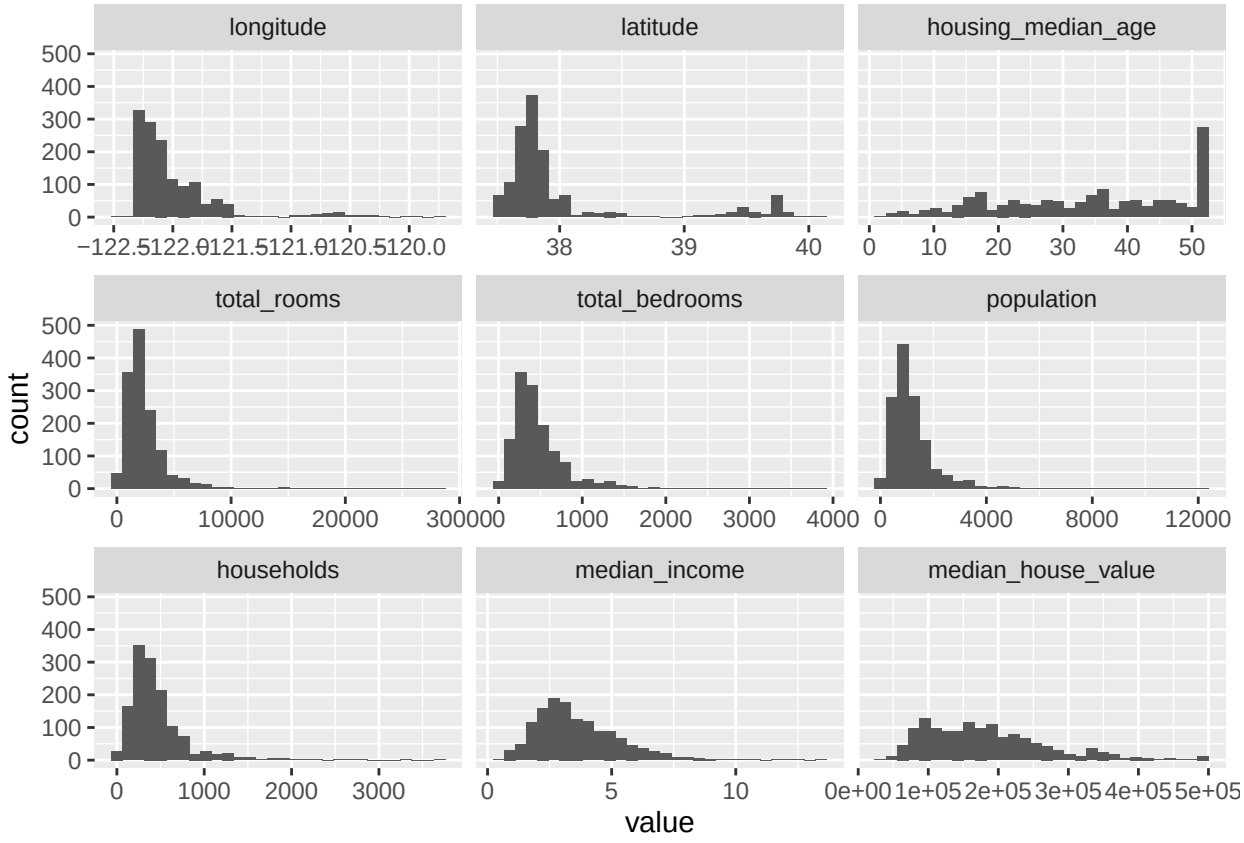
```
##   longitude latitude housing_median_age total_rooms total_bedrooms population
## 1   -122.23    37.88                41         880            129        322
## 2   -122.22    37.86                21        7099           1106       2401
## 3   -122.24    37.85                52        1467            190        496
## 4   -122.25    37.85                52        1274            235        558
## 5   -122.25    37.85                52        1627            280        565
## 6   -122.25    37.85                52         919            213        413
##   households median_income median_house_value ocean_proximity
```

```
## 1        126        8.3252            452600        NEAR BAY
## 2       1138        8.3014            358500        NEAR BAY
## 3        177        7.2574            352100        NEAR BAY
## 4        219        5.6431            341300        NEAR BAY
## 5        259        3.8462            342200        NEAR BAY
## 6        193        4.0368            269700        NEAR BAY
```

```
library(ggplot2)
ggplot(data = melt(houseData), mapping = aes(x = value)) + geom_histogram(bins = 30) + facet_wrap(~varia
```

```
## Using ocean_proximity as id variables
```

```
## Warning: Removed 16 rows containing non-finite values (stat_bin).
```



explor missing data

```
summary(houseData)
```

```
##    longitude          latitude     housing_median_age  total_rooms
##  Min.   :-122.5   Min.   :37.47   Min.   : 2.00      Min.   :   12
##  1st Qu.:-122.2   1st Qu.:37.70   1st Qu.:22.00      1st Qu.: 1326
##  Median :-122.1   Median :37.79   Median :35.00      Median : 1966
##  Mean   :-122.0   Mean   :38.02   Mean   :33.98      Mean   : 2456
##  3rd Qu.:-121.9   3rd Qu.:37.95   3rd Qu.:48.00      3rd Qu.: 2935
##  Max.   :-119.8   Max.   :40.06   Max.   :52.00      Max.   :28258
##                   NA's   :1       NA's   :1          NA's   :1
##  total_bedrooms     population      households   median_income
##  Min.   :   4.0   Min.   :   18   Min.   :   7   Min.   : 0.4999
```

```
##  1st Qu.: 270.0   1st Qu.:  662   1st Qu.: 254   1st Qu.: 2.4206
##  Median : 394.0   Median :  979   Median : 371   Median : 3.2552
##  Mean   : 493.2   Mean   : 1228   Mean   : 462   Mean   : 3.6189
##  3rd Qu.: 587.0   3rd Qu.: 1461   3rd Qu.: 543   3rd Qu.: 4.6000
##  Max.   :3864.0   Max.   :12203   Max.   :3701   Max.   :13.4990
##  NA's   :9        NA's   :1       NA's   :1      NA's   :1
##  median_house_value  ocean_proximity
##  Min.   : 39400              :  1
##  1st Qu.:112600     <1H OCEAN: 76
##  Median :170000     INLAND   :394
##  Mean   :184216     NEAR BAY :907
##  3rd Qu.:231200
##  Max.   :500001
##  NA's   :1
```

fill missing data with column mean value

```
houseData$latitude[is.na(houseData$latitude)] = mean(houseData$latitude, na.rm = TRUE)
houseData$housing_median_age[is.na(houseData$housing_median_age)] = median(houseData$housing_median_age
houseData$total_rooms[is.na(houseData$total_rooms)] = median(houseData$total_rooms, na.rm = TRUE)
houseData$total_bedrooms[is.na(houseData$total_bedrooms)] = median(houseData$total_bedrooms, na.rm = TRU
houseData$population[is.na(houseData$population)] = median(houseData$population, na.rm = TRUE)
houseData$households[is.na(houseData$households)] = median(houseData$households, na.rm = TRUE)
houseData$median_income[is.na(houseData$median_income)] = median(houseData$median_income, na.rm = TRUE)
houseData$median_house_value[is.na(houseData$median_house_value)] = median(houseData$median_house_value
summary(houseData)
```

```
##    longitude         latitude     housing_median_age  total_rooms
##  Min.   :-122.5   Min.   :37.47   Min.   : 2.00      Min.   :   12
##  1st Qu.:-122.2   1st Qu.:37.70   1st Qu.:22.00      1st Qu.: 1326
##  Median :-122.1   Median :37.79   Median :35.00      Median : 1966
##  Mean   :-122.0   Mean   :38.02   Mean   :33.98      Mean   : 2456
##  3rd Qu.:-121.9   3rd Qu.:37.95   3rd Qu.:48.00      3rd Qu.: 2934
##  Max.   :-119.8   Max.   :40.06   Max.   :52.00      Max.   :28258
##  total_bedrooms     population        households     median_income
##  Min.   :   4.0   Min.   :   18.0   Min.   :   7.0   Min.   : 0.4999
##  1st Qu.: 271.0   1st Qu.:  662.2   1st Qu.: 254.2   1st Qu.: 2.4213
##  Median : 394.0   Median :  979.0   Median : 371.0   Median : 3.2552
##  Mean   : 492.6   Mean   : 1227.6   Mean   : 462.0   Mean   : 3.6186
##  3rd Qu.: 582.8   3rd Qu.: 1459.8   3rd Qu.: 543.0   3rd Qu.: 4.5978
##  Max.   :3864.0   Max.   :12203.0   Max.   :3701.0   Max.   :13.4990
##  median_house_value  ocean_proximity
##  Min.   : 39400              :  1
##  1st Qu.:112650     <1H OCEAN: 76
##  Median :170000     INLAND   :394
##  Mean   :184205     NEAR BAY :907
##  3rd Qu.:231200
##  Max.   :500001
```

feature engineering

```
target = c('ocean_proximity') # set ocean_proximity as the target
houseDataX = houseData[, 1: 9]
houseDataX = scale(houseDataX)
houseDataY = houseData[target]
```

```r
# numeric the target column
houseDataY = as.numeric(as.factor(houseDataY$ocean_proximity))
houseDataY = matrix(houseDataY, ncol = 1)
```

split into training and testing data set

```r
data = cbind(houseDataX, houseDataY)
colnames(data) = c(colnames(data)[1: 9], 'target')

# split
set.seed(1)
rows = nrow(data)
# set ratio of training:testing = 3: 1
sample = sample.int(n = rows, size = round(0.75 * rows), replace = F)
trainingData = data[sample, ]
testingData = data[-sample, ]
paste('after sampling, the ratio of traing : testing is: ', nrow(trainingData) %/% nrow(testingData))
```

```
## [1] "after sampling, the ratio of traing : testing is:  3"
```

build model1 DecisionTree

```r
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 3.6.3
```

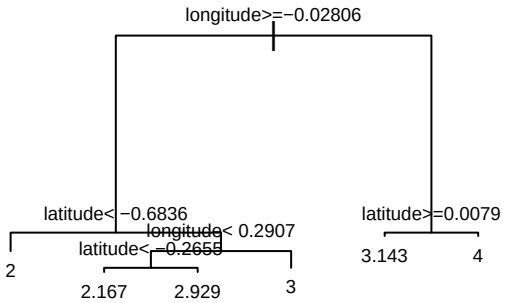
```r
trainingData = as.data.frame(trainingData)
testingData = as.data.frame(testingData)
testingTarget = testingData$target
testingData = testingData[, 1: ncol(testingData) - 1]
tree = rpart(target~., data = trainingData)
summary(tree)
```

```
## Call:
## rpart(formula = target ~ ., data = trainingData)
##   n= 1034
##
##           CP nsplit  rel error     xerror       xstd
## 1 0.85168864      0 1.00000000 1.00110267 0.05188164
## 2 0.07978515      1 0.14831136 0.14951608 0.01831059
## 3 0.01668912      2 0.06852621 0.07450272 0.01654165
## 4 0.01454193      4 0.03514797 0.07209559 0.01674609
## 5 0.01000000      5 0.02060604 0.05033281 0.01488997
##
## Variable importance
##          longitude           latitude housing_median_age median_house_value
##                 43                 32                 16                  5
##        total_rooms         population      median_income
##                  3                  1                  1
##
## Node number 1: 1034 observations,    complexity param=0.8516886
##   mean=3.617988, MSE=0.3385932
##   left son=2 (337 obs) right son=3 (697 obs)
##   Primary splits:
##       longitude          < -0.02806409 to the right, improve=0.85168860, (0 missing)
##       latitude           < -0.1850749  to the right, improve=0.34020800, (0 missing)
```

```
##       housing_median_age < -0.3147501  to the left,  improve=0.29312810, (0 missing)
##       total_rooms        < 0.3809576   to the right, improve=0.07500078, (0 missing)
##       population         < 0.356865    to the right, improve=0.05523714, (0 missing)
##   Surrogate splits:
##       latitude           < -0.1850749  to the right, agree=0.894, adj=0.674, (0 split)
##       housing_median_age < -0.4553647  to the left,  agree=0.801, adj=0.389, (0 split)
##       median_house_value < -0.647096   to the left,  agree=0.705, adj=0.095, (0 split)
##       total_rooms        < 0.8367953   to the right, agree=0.694, adj=0.062, (0 split)
##       population         < 0.9569996   to the right, agree=0.682, adj=0.024, (0 split)
##
## Node number 2: 337 observations,    complexity param=0.07978515
##   mean=2.845697, MSE=0.1364281
##   left son=4 (35 obs) right son=5 (302 obs)
##   Primary splits:
##       latitude           < -0.6835947  to the left,  improve=0.60755730, (0 missing)
##       longitude          < 0.2375425   to the left,  improve=0.53975100, (0 missing)
##       median_house_value < 0.05627763  to the right, improve=0.39961260, (0 missing)
##       median_income      < 0.2976095   to the right, improve=0.20109750, (0 missing)
##       population         < 0.7423092   to the right, improve=0.09938862, (0 missing)
##   Surrogate splits:
##       longitude          < 0.1047392   to the left,  agree=0.941, adj=0.429, (0 split)
##       median_house_value < 2.819571    to the right, agree=0.899, adj=0.029, (0 split)
##
## Node number 3: 697 observations,    complexity param=0.01454193
##   mean=3.991392, MSE=0.008534218
##   left son=6 (7 obs) right son=7 (690 obs)
##   Primary splits:
##       latitude           < 0.0079005   to the right, improve=0.85590240, (0 missing)
##       median_house_value < -1.207675   to the left,  improve=0.36957310, (0 missing)
##       longitude          < -0.05462475 to the right, improve=0.06247779, (0 missing)
##       population         < -1.070077   to the left,  improve=0.02142428, (0 missing)
##       households         < -1.119368   to the left,  improve=0.02142428, (0 missing)
##   Surrogate splits:
##       median_house_value < -1.273495   to the left,  agree=0.993, adj=0.286, (0 split)
##       longitude          < -0.8514445  to the left,  agree=0.991, adj=0.143, (0 split)
##
## Node number 4: 35 observations
##   mean=2, MSE=0
##
## Node number 5: 302 observations,    complexity param=0.01668912
##   mean=2.943709, MSE=0.05974519
##   left son=10 (46 obs) right son=11 (256 obs)
##   Primary splits:
##       longitude          < 0.2906638   to the left,  improve=0.29516390, (0 missing)
##       median_house_value < 0.2381173   to the right, improve=0.19030160, (0 missing)
##       latitude           < -0.474538   to the left,  improve=0.17878590, (0 missing)
##       median_income      < 1.301823    to the right, improve=0.10505240, (0 missing)
##       population         < 1.033389    to the right, improve=0.03600976, (0 missing)
##   Surrogate splits:
##       latitude           < -0.6031883  to the left,  agree=0.854, adj=0.043, (0 split)
##       housing_median_age < -2.072432   to the left,  agree=0.851, adj=0.022, (0 split)
##       median_income      < 1.301823    to the right, agree=0.851, adj=0.022, (0 split)
##
## Node number 6: 7 observations
```

```
##   mean=3.142857, MSE=0.122449
##
## Node number 7: 690 observations
##   mean=4, MSE=0
##
## Node number 10: 46 observations,    complexity param=0.01668912
##   mean=2.630435, MSE=0.276465
##   left son=20 (18 obs) right son=21 (28 obs)
##   Primary splits:
##       latitude           < -0.2654813  to the left,  improve=0.50012210, (0 missing)
##       median_house_value < -0.1807833  to the right, improve=0.46769560, (0 missing)
##       median_income      < 0.3519818   to the right, improve=0.20683760, (0 missing)
##       total_rooms        < -0.3842071  to the right, improve=0.08827173, (0 missing)
##       housing_median_age < -0.9475158  to the right, improve=0.06254144, (0 missing)
##   Surrogate splits:
##       median_house_value < 0.1081521   to the right, agree=0.978, adj=0.944, (0 split)
##       median_income      < 0.3519818   to the right, agree=0.848, adj=0.611, (0 split)
##       total_rooms        < 0.5070863   to the right, agree=0.696, adj=0.222, (0 split)
##       total_bedrooms     < 0.5359917   to the right, agree=0.696, adj=0.222, (0 split)
##       population         < 0.9355306   to the right, agree=0.696, adj=0.222, (0 split)
##
## Node number 11: 256 observations
##   mean=3, MSE=0
##
## Node number 20: 18 observations
##   mean=2.166667, MSE=0.1388889
##
## Node number 21: 28 observations
##   mean=2.928571, MSE=0.1377551
```

```
plot(tree, margin=0.5)
text(tree, cex=0.6)
```



build model2 RandomForest

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 3.6.3
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
rm = randomForest(target~., data = trainingData)
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values. Are you sure you want to do regression?
```

```
importance(rm)
```

```
##                   IncNodePurity
## longitude            215.503016
## latitude              63.640637
```

```
## housing_median_age     33.133608
## total_rooms             8.572998
## total_bedrooms          2.697816
## population              4.887195
## households              2.540248
## median_income           5.175905
## median_house_value     11.570122
```

predition and evaluation

```
library(ROCR)
```

```
## Warning: package 'ROCR' was built under R version 3.6.3
```

```
## Loading required package: gplots
```

```
## Warning: package 'gplots' was built under R version 3.6.3
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
pre = predict(tree, testingData)
print('Accuracy using Decision Tree:')
```

```
## [1] "Accuracy using Decision Tree:"
```

```
print(sum(pre == testingTarget) / length(testingTarget))
```

```
## [1] 0.9244186
```

```
pre = predict(rm, testingData)
print('Accuracy using Random Forest:')
```

```
## [1] "Accuracy using Random Forest:"
```

```
print(sum(pre == testingTarget) / length(testingTarget))
```

```
## [1] 0.122093
```