STAT636 Project

Xiaohan Wei, Qinye Jiang, Corina Ramont

2023-11-11

Step 1: import and clean the data

```r
# import data
data = read.csv("data.csv")

# First step: Code target response into categorical
# Define the levels and labels; we only focus on dropout and not dropout
levels = c("Dropout", "Graduate", "Enrolled")
labels = c(1,0,0)

# create a new column "outcome"
data$outcome = factor(data$Target, levels=levels, labels=labels)
data$outcome = as.numeric(data$outcome)
data$outcome = data$outcome-1

# drop original Target, create new dataset
dat = subset(data, select=-Target)

# Summarize dat
summary(dat)
```

```
##  Marital.status  Application.mode Application.order     Course
##  Min.   :1.000   Min.   : 1.00    Min.   :0.000     Min.   :  33
##  1st Qu.:1.000   1st Qu.: 1.00    1st Qu.:1.000     1st Qu.:9085
##  Median :1.000   Median :17.00    Median :1.000     Median :9238
##  Mean   :1.179   Mean   :18.67    Mean   :1.728     Mean   :8857
##  3rd Qu.:1.000   3rd Qu.:39.00    3rd Qu.:2.000     3rd Qu.:9556
##  Max.   :6.000   Max.   :57.00    Max.   :9.000     Max.   :9991
##  Daytime.evening.attendance. Previous.qualification
##  Min.   :0.0000              Min.   : 1.000
##  1st Qu.:1.0000              1st Qu.: 1.000
...
```

Step 2: splitting data into training and testing datasets

```r
index = createDataPartition(y = dat$outcome, p = 0.8, list = F)

train = dat[index, ]
test = dat[-index, ]
```

Step 3: Compare different classification models

Step 3.1: LDA

```r
# performing LDA on the training set of data

lda.fit = lda(outcome ~ ., data = train)

lda.fit # summary of the obtained LDA
```

```
## Call:
## lda(outcome ~ ., data = train)
##
## Prior probabilities of groups:
##         0         1
## 0.3214689 0.6785311
##
## Group means:
##   Marital.status Application.mode Application.order   Course
## 0       1.269772         24.01406          1.586116 8811.200
## 1       1.139467         16.35595          1.796420 8907.542
##   Daytime.evening.attendance. Previous.qualification
## 0                   0.8453427               5.466608
## 1                   0.9063281               4.279767
##   Previous.qualification..grade. Nacionality Mother.s.qualification
## 0                       131.3009    2.173111               21.07557
## 1                       133.3513    1.895504               18.96211
##   Father.s.qualification Mother.s.occupation Father.s.occupation
## 0               22.64236            9.884007            10.34359
## 1               21.88843           11.640300            11.69442
##   Admission.grade Displaced Educational.special.needs     Debtor
## 0        125.0763 0.4604569                0.01142355 0.22759227
## 1        128.0393 0.5870108                0.01082431 0.06369692
##   Tuition.fees.up.to.date    Gender Scholarship.holder Age.at.enrollment
## 0               0.6766257 0.4920914          0.1001757          26.18102
## 1               0.9796003 0.2772689          0.3180683          21.94671
##   International Curricular.units.1st.sem..credited.
## 0    0.02724077                           0.5808436
## 1    0.02706078                           0.7668609
##   Curricular.units.1st.sem..enrolled. Curricular.units.1st.sem..evaluations.
## 0                            5.840070                               7.685413
## 1                            6.470858                               8.535387
##   Curricular.units.1st.sem..approved. Curricular.units.1st.sem..grade.
## 0                            2.517575                         7.133116
## 1                            5.704413                        12.222220
##   Curricular.units.1st.sem..without.evaluations.
## 0                                      0.1985940
## 1                                      0.1161532
##   Curricular.units.2nd.sem..credited. Curricular.units.2nd.sem..enrolled.
## 0                           0.4279438                            5.789982
## 1                           0.5915903                            6.437136
##   Curricular.units.2nd.sem..evaluations. Curricular.units.2nd.sem..approved.
## 0                               7.175747                            1.872583
## 1                               8.453789                            5.601998
##   Curricular.units.2nd.sem..grade.
## 0                         5.778525
## 1                        12.248600
##   Curricular.units.2nd.sem..without.evaluations. Unemployment.rate
## 0                                      0.2469244          11.63515
## 1                                      0.1161532          11.48805
##   Inflation.rate         GDP
## 0       1.295079 -0.09526362
## 1       1.215196  0.09228560
##
## Coefficients of linear discriminants:
##                                                          LD1
## Marital.status                                  5.230772e-02
## Application.mode                               -1.910180e-03
## Application.order                              -3.419221e-02
## Course                                         -3.320106e-05
## Daytime.evening.attendance.                    -1.957273e-02
## Previous.qualification                          6.792109e-03
## Previous.qualification..grade.                 -3.833464e-03
## Nacionality                                    -1.335884e-02
## Mother.s.qualification                         -3.345256e-03
## Father.s.qualification                          1.826072e-03
## Mother.s.occupation                             7.598691e-03
## Father.s.occupation                            -4.037716e-03
## Admission.grade                                 1.958610e-03
## Displaced                                      -7.248311e-02
## Educational.special.needs                      -1.916159e-01
## Debtor                                         -2.470329e-01
## Tuition.fees.up.to.date                         1.213261e+00
## Gender                                         -1.634892e-01
## Scholarship.holder                              1.413596e-01
## Age.at.enrollment                              -1.903183e-02
## International                                   7.129866e-01
## Curricular.units.1st.sem..credited.            -4.988451e-02
## Curricular.units.1st.sem..enrolled.             7.555990e-04
## Curricular.units.1st.sem..evaluations.         -1.238160e-04
## Curricular.units.1st.sem..approved.             1.085085e-01
## Curricular.units.1st.sem..grade.               -2.437888e-02
## Curricular.units.1st.sem..without.evaluations.  4.466994e-02
## Curricular.units.2nd.sem..credited.            -1.217256e-01
## Curricular.units.2nd.sem..enrolled.            -3.265901e-01
## Curricular.units.2nd.sem..evaluations.          1.708304e-02
## Curricular.units.2nd.sem..approved.             4.299066e-01
## Curricular.units.2nd.sem..grade.                3.507169e-02
## Curricular.units.2nd.sem..without.evaluations.  6.065729e-02
## Unemployment.rate                               -2.596559e-02
## Inflation.rate                                  6.552265e-03
## GDP                                            -1.464136e-03
```

```r
# To determine the test error of the model obtained
lda.pred = predict(lda.fit, test) # Using the LDA model to predict through the test data
lda.class = lda.pred$class # predicted `outcome` based in the fitted LDA

table(lda.class, test$outcome) # confusion matrix
```

```
##
```

```
## lda.class   0   1
##         0 177  27
##         1 106 574
```

```
test_err.lda = mean(lda.class != test$outcome) # test error
test_err.lda
```

```
## [1] 0.1504525
```

Test error for LDA is 0.1504525.

Step 3.2: QDA

```
qda.fit = qda(outcome ~ ., data = train)

qda.fit # summary of the obtained QDA
```

```
## Call:
## qda(outcome ~ ., data = train)
##
## Prior probabilities of groups:
##         0         1
## 0.3214689 0.6785311
##
## Group means:
##   Marital.status Application.mode Application.order   Course
## 0       1.269772         24.01406          1.586116 8811.200
## 1       1.139467         16.35595          1.796420 8907.542
##   Daytime.evening.attendance. Previous.qualification
## 0                   0.8453427               5.466608
## 1                   0.9063281               4.279767
##   Previous.qualification..grade. Nacionality Mother.s.qualification
## 0                       131.3009    2.173111               21.07557
## 1                       133.3513    1.895504               18.96211
##   Father.s.qualification Mother.s.occupation Father.s.occupation
## 0               22.64236            9.884007            10.34359
## 1               21.88843           11.640300            11.69442
##   Admission.grade Displaced Educational.special.needs     Debtor
## 0        125.0763 0.4604569                0.01142355 0.22759227
## 1        128.0393 0.5870108                0.01082431 0.06369692
##   Tuition.fees.up.to.date    Gender Scholarship.holder Age.at.enrollment
## 0               0.6766257 0.4920914          0.1001757          26.18102
## 1               0.9796003 0.2772689          0.3180683          21.94671
##   International Curricular.units.1st.sem..credited.
## 0    0.02724077                           0.5808436
## 1    0.02706078                           0.7668609
##   Curricular.units.1st.sem..enrolled. Curricular.units.1st.sem..evaluations.
## 0                            5.840070                               7.685413
## 1                            6.470858                               8.535387
##   Curricular.units.1st.sem..approved. Curricular.units.1st.sem..grade.
## 0                            2.517575                         7.133116
## 1                            5.704413                        12.222220
##   Curricular.units.1st.sem..without.evaluations.
## 0                                      0.1985940
## 1                                      0.1161532
##   Curricular.units.2nd.sem..credited. Curricular.units.2nd.sem..enrolled.
## 0                           0.4279438                            5.789982
## 1                           0.5915903                            6.437136
##   Curricular.units.2nd.sem..evaluations. Curricular.units.2nd.sem..approved.
## 0                               7.175747                            1.872583
## 1                               8.453789                            5.601998
##   Curricular.units.2nd.sem..grade.
## 0                         5.778525
## 1                        12.248600
##   Curricular.units.2nd.sem..without.evaluations. Unemployment.rate
## 0                                      0.2469244          11.63515
## 1                                      0.1161532          11.48805
##   Inflation.rate         GDP
## 0       1.295079 -0.09526362
## 1       1.215196  0.09228560
```

```
# To determine the test error of the model obtained

qda.pred = predict(qda.fit, test) # Using the QDA model to predict through the test data
qda.class = qda.pred$class # predicted `mpg01` based in the fitted QDA

table(qda.class, test$outcome) # confusion matrix
```

```
##
## qda.class   0   1
##         0 199  71
##         1  84 530
```

```
test_err.qda = mean(qda.class != test$outcome) # test error
test_err.qda
```

```
## [1] 0.1753394
```

Test error for QDA is 0.1753394.

Step 3.3: Logistic regression

```
logistic.fit = glm(outcome ~ .,
                   data = train,
                   family = binomial)

summary(logistic.fit) # summary of the logistic regression model
```

```
##
## Call:
## glm(formula = outcome ~ ., family = binomial, data = train)
##
## Coefficients:
##                                                   Estimate Std. Error z value
## (Intercept)                                      1.021e+00  8.494e-01   1.202
## Marital.status                                  1.074e-01  1.078e-01   0.996
## Application.mode                               -3.155e-03  4.148e-03  -0.761
## Application.order                              -1.028e-01  4.841e-02  -2.123
## Course                                         -1.009e-04  4.118e-05  -2.451
## Daytime.evening.attendance.                    -5.316e-02  2.010e-01  -0.264
## Previous.qualification                          1.471e-02  6.079e-03   2.419
## Previous.qualification..grade.                 -5.767e-03  5.126e-03  -1.125
## Nacionality                                    -3.561e-02  1.134e-02  -3.140
## Mother.s.qualification                         -9.768e-03  4.591e-03  -2.128
## Father.s.qualification                          4.334e-03  4.426e-03   0.979
## Mother.s.occupation                             1.783e-02  5.606e-03   3.181
## Father.s.occupation                            -8.007e-03  5.770e-03  -1.388
## Admission.grade                                 4.099e-03  4.847e-03   0.846
## Displaced                                      -2.902e-01  1.317e-01  -2.203
## Educational.special.needs                      -4.158e-01  4.806e-01  -0.865
## Debtor                                         -4.228e-01  1.906e-01  -2.218
## Tuition.fees.up.to.date                         2.565e+00  2.177e-01  11.784
## Gender                                         -3.522e-01  1.200e-01  -2.935
## Scholarship.holder                              4.100e-01  1.545e-01   2.654
## Age.at.enrollment                              -5.066e-02  1.061e-02  -4.773
## International                                   1.876e+00  6.044e-01   3.104
## Curricular.units.1st.sem..credited.            -6.854e-02  8.909e-02  -0.769
## Curricular.units.1st.sem..enrolled.             6.769e-03  1.132e-01   0.060
## Curricular.units.1st.sem..evaluations.          4.912e-03  2.770e-02   0.177
## Curricular.units.1st.sem..approved.             2.141e-01  6.008e-02   3.563
## Curricular.units.1st.sem..grade.               -3.913e-02  2.714e-02  -1.442
## Curricular.units.1st.sem..without.evaluations.  1.237e-01  1.030e-01   1.201
## Curricular.units.2nd.sem..credited.            -2.526e-01  9.709e-02  -2.602
## Curricular.units.2nd.sem..enrolled.            -4.862e-01  1.105e-01  -4.401
## Curricular.units.2nd.sem..evaluations.          3.391e-02  2.597e-02   1.306
## Curricular.units.2nd.sem..approved.             6.779e-01  5.492e-02  12.344
## Curricular.units.2nd.sem..grade.                6.499e-02  2.526e-02   2.573
## Curricular.units.2nd.sem..without.evaluations.  1.599e-01  8.524e-02   1.876
## Unemployment.rate                              -8.514e-02  2.395e-02  -3.555
## Inflation.rate                                 -3.184e-02  4.112e-02  -0.774
## GDP                                            -1.345e-02  2.885e-02  -0.466
##                                                Pr(>|z|)
## (Intercept)                                    0.229280
## Marital.status                                 0.319145
## Application.mode                               0.446875
## Application.order                              0.033784 *
## Course                                         0.014264 *
## Daytime.evening.attendance.                    0.791449
## Previous.qualification                         0.015546 *
## Previous.qualification..grade.                 0.260653
## Nacionality                                    0.001690 **
## Mother.s.qualification                         0.033356 *
## Father.s.qualification                         0.327468
## Mother.s.occupation                            0.001468 **
## Father.s.occupation                            0.165224
## Admission.grade                                0.397788
## Displaced                                      0.027598 *
## Educational.special.needs                      0.386944
## Debtor                                         0.026533 *
## Tuition.fees.up.to.date                          < 2e-16 ***
## Gender                                          0.003336 **
## Scholarship.holder                              0.007963 **
## Age.at.enrollment                               1.82e-06 ***
## International                                   0.001910 **
## Curricular.units.1st.sem..credited.             0.441667
## Curricular.units.1st.sem..enrolled.             0.952297
## Curricular.units.1st.sem..evaluations.          0.859258
## Curricular.units.1st.sem..approved.             0.000366 ***
## Curricular.units.1st.sem..grade.                0.149357
## Curricular.units.1st.sem..without.evaluations. 0.229896
## Curricular.units.2nd.sem..credited.             0.009281 **
## Curricular.units.2nd.sem..enrolled.             1.08e-05 ***
## Curricular.units.2nd.sem..evaluations.          0.191524
## Curricular.units.2nd.sem..approved.              < 2e-16 ***
## Curricular.units.2nd.sem..grade.                0.010074 *
## Curricular.units.2nd.sem..without.evaluations. 0.060598 .
## Unemployment.rate                               0.000378 ***
## Inflation.rate                                  0.438662
## GDP                                             0.641223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4446.0  on 3539  degrees of freedom
## Residual deviance: 2144.7  on 3503  degrees of freedom
## AIC: 2218.7
##
## Number of Fisher Scoring iterations: 6
```

```r
# To determine the test error of the model obtained
# storing the predicted values of `outcome` from the fitted logistic regression
logistic.pred = ifelse(predict(logistic.fit, type = "response", test) > 0.32, 1, 0)
table(logistic.pred, test$outcome) # confusion matrix
```

```
##
## logistic.pred   0   1
##             0 168  18
##             1 115 583
```

```r
test_err.logistic = mean(logistic.pred != test$outcome) # test error
test_err.logistic
```

```
## [1] 0.1504525
```

Test error for LR is 0.1504525

Step 3.4: naive Bayes

```r
# performing naive Bayes on the training set of data
nb.fit = naiveBayes(outcome ~ .,
                    data = train)
```

```
nb.fit # summary of the obtained naive Bayes model
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##         0         1
## 0.3214689 0.6785311
##
## Conditional probabilities:
##    Marital.status
## Y       [,1]      [,2]
##   0 1.269772 0.7373561
##   1 1.139467 0.5415969
##
##    Application.mode
## Y       [,1]     [,2]
##   0 24.01406 17.08115
##   1 16.35595 17.25892
##
##    Application.order
## Y       [,1]     [,2]
##   0 1.586116 1.213404
##   1 1.796420 1.344997
##
##    Course
## Y       [,1]     [,2]
##   0 8811.200 2193.612
##   1 8907.542 1955.836
##
##    Daytime.evening.attendance.
## Y        [,1]      [,2]
##   0 0.8453427 0.3617366
##   1 0.9063281 0.2914324
##
##    Previous.qualification
## Y       [,1]     [,2]
##   0 5.466608 10.45547
##   1 4.279767 10.23969
##
##    Previous.qualification..grade.
## Y       [,1]     [,2]
##   0 131.3009 13.11512
##   1 133.3513 13.36594
##
##    Nacionality
## Y       [,1]     [,2]
##   0 2.173111 8.715623
##   1 1.895504 6.678952
```

```
##
##    Mother.s.qualification
## Y       [,1]     [,2]
##   0 21.07557 15.48209
##   1 18.96211 15.60115
##
##    Father.s.qualification
## Y       [,1]     [,2]
##   0 22.64236 15.38487
##   1 21.88843 15.35201
##
##    Mother.s.occupation
## Y        [,1]     [,2]
##   0  9.884007 20.04688
##   1 11.640300 29.21780
##
##    Father.s.occupation
## Y       [,1]     [,2]
##   0 10.34359 20.03670
##   1 11.69442 27.53599
##
##    Admission.grade
## Y       [,1]     [,2]
##   0 125.0763 15.36911
##   1 128.0393 14.07472
##
##    Displaced
## Y        [,1]      [,2]
##   0 0.4604569 0.4986530
##   1 0.5870108 0.4924734
##
##    Educational.special.needs
## Y         [,1]      [,2]
##   0 0.01142355 0.1063155
##   1 0.01082431 0.1034969
##
##    Debtor
## Y         [,1]      [,2]
##   0 0.22759227 0.4194623
##   1 0.06369692 0.2442631
##
##    Tuition.fees.up.to.date
## Y        [,1]      [,2]
##   0 0.6766257 0.4679699
##   1 0.9796003 0.1413925
##
##    Gender
## Y        [,1]      [,2]
##   0 0.4920914 0.5001573
##   1 0.2772689 0.4477436
##
##    Scholarship.holder
## Y        [,1]      [,2]
##   0 0.1001757 0.3003662
##   1 0.3180683 0.4658231
##
##    Age.at.enrollment
## Y       [,1]     [,2]
##   0 26.18102 8.748148
##   1 21.94671 6.611148
##
##    International
## Y         [,1]      [,2]
##   0 0.02724077 0.1628558
##   1 0.02706078 0.1622944
##
##    Curricular.units.1st.sem..credited.
## Y        [,1]     [,2]
##   0 0.5808436 2.042009
##   1 0.7668609 2.494095
##
##    Curricular.units.1st.sem..enrolled.
## Y       [,1]     [,2]
##   0 5.840070 2.239584
##   1 6.470858 2.525873
##
##    Curricular.units.1st.sem..evaluations.
## Y       [,1]     [,2]
##   0 7.685413 4.794949
##   1 8.535387 3.719524
##
##    Curricular.units.1st.sem..approved.
## Y       [,1]     [,2]
##   0 2.517575 2.838073
##   1 5.704413 2.648358
##
##    Curricular.units.1st.sem..grade.
## Y        [,1]     [,2]
##   0  7.133116 6.039747
##   1 12.222220 3.105260
##
##    Curricular.units.1st.sem..without.evaluations.
## Y        [,1]      [,2]
##   0 0.1985940 0.8383496
##   1 0.1161532 0.6576982
##
##    Curricular.units.2nd.sem..credited.
## Y        [,1]     [,2]
##   0 0.4279438 1.614738
##   1 0.5915903 2.028981
##
##    Curricular.units.2nd.sem..enrolled.
## Y       [,1]     [,2]
##   0 5.789982 2.015508
##   1 6.437136 2.221493
##
##    Curricular.units.2nd.sem..evaluations.
## Y       [,1]     [,2]
##   0 7.175747 4.756595
##   1 8.453789 3.384272
##
##    Curricular.units.2nd.sem..approved.
## Y        [,1]     [,2]
##   0 1.872583 2.520719
##   1 5.601998 2.440420
##
##    Curricular.units.2nd.sem..grade.
## Y         [,1]     [,2]
##   0  5.778525 6.104216
##   1 12.248600 3.076044
##
##    Curricular.units.2nd.sem..without.evaluations.
## Y         [,1]      [,2]
##   0 0.2469244 1.0424644
##   1 0.1161532 0.6390699
##
##    Unemployment.rate
## Y        [,1]     [,2]
##   0 11.63515 2.774506
##   1 11.48805 2.609910
##
##    Inflation.rate
## Y        [,1]     [,2]
##   0 1.295079 1.388802
##   1 1.215196 1.372760
##
##    GDP
## Y           [,1]     [,2]
##   0 -0.09526362 2.249004
##   1  0.09228560 2.253303
```

```
# To determine the test error of the model obtained
nb.class = predict(nb.fit, test) # Using the naive Bayes model to predict through the test data

test_err.nb = mean(nb.class != test$outcome) # test error
test_err.nb
```

```
## [1] 0.1900452
```

Test error for NB is 0.1900452

Step 3.5: KNN

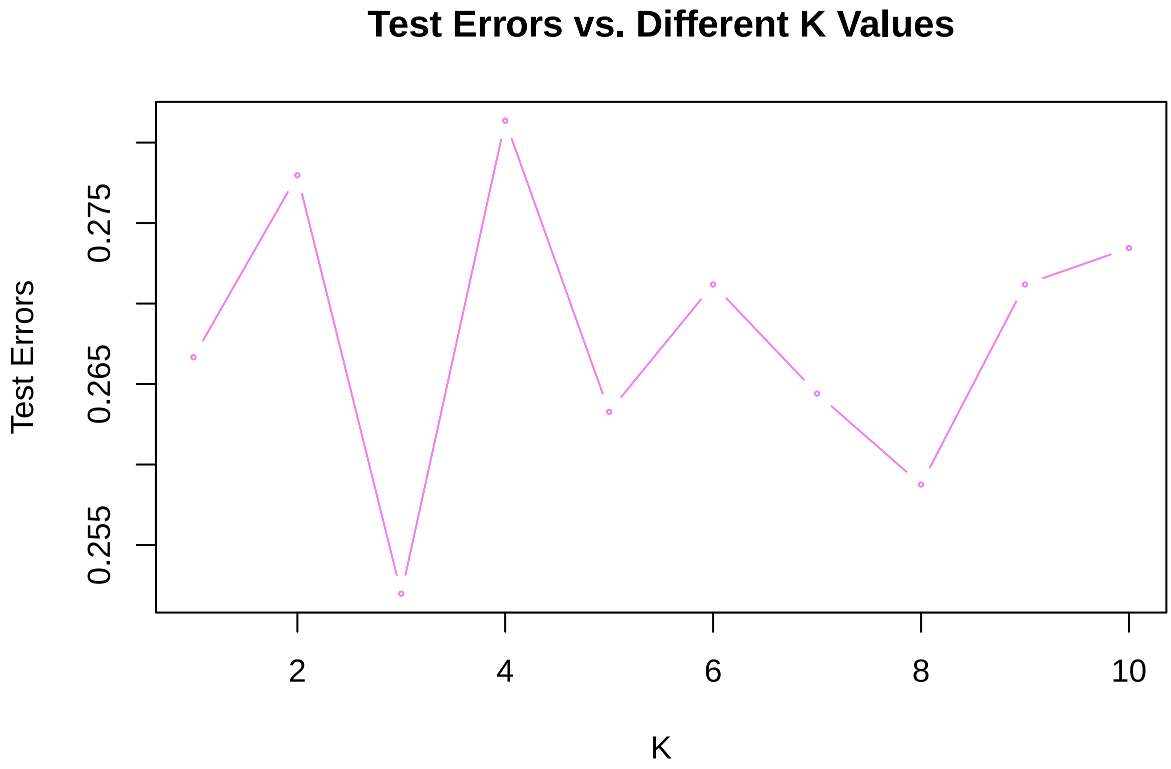
```
# separate original training data into a training and tuning set for KNN
# overall percentages: 60% training, 20% tuning, 20% testing
index2 = createDataPartition(y = train$outcome, p = 0.25, list = F)
knn_train = train[-index2,]
knn_tune = train[index2,]
knn_test = test

c1 = as.factor(knn_train$outcome)
```

```r
# performing KNN for different values of K on the training set
n = 10 # total number of choices for K
test_err.knn = array(0) # to store the test errors for different values of K

for(j in 1:n){
  knn.fit = knn(knn_train, knn_tune, c1, k = j) # fitting the KNN model with K = j
  test_err.knn[j] = mean(knn.fit != knn_tune$outcome) # test error for the jth value of K
}

# plotting the test errors for different values of K for which the KNN has been fitted
plot(1:n, test_err.knn, type = "b", cex = 0.3, col = "violet",
     xlab = "K", ylab = "Test Errors", main = "Test Errors vs. Different K Values")
```



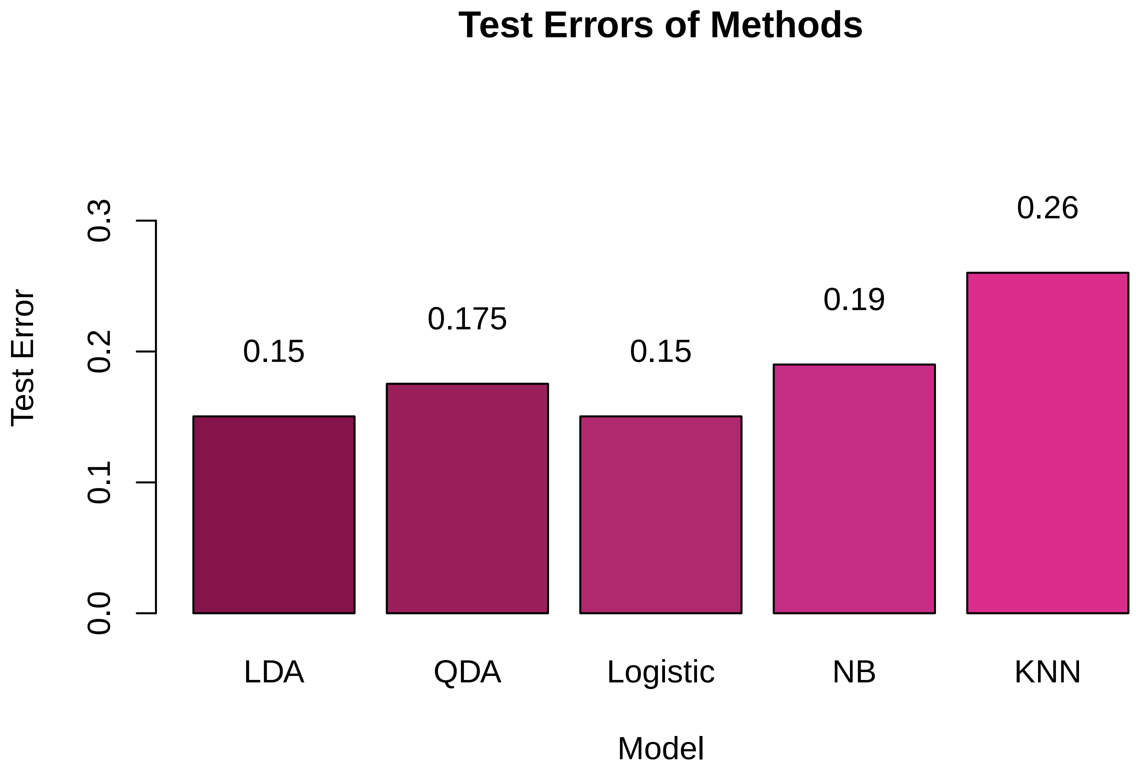
```r
# fitting the KNN model with K = 5
knn.fit.final = knn(knn_train, knn_test, c1, k = 5)
# test error for K = 5 on testing set
err.knn = mean(knn.fit.final != knn_test$outcome)
err.knn
```

```
## [1] 0.260181
```

Test error for KNN (K=5) is 0.2567873.

Step 4: visualize all test errors

```
### Test error
test_error = c(test_err.lda, test_err.qda, test_err.logistic,
                test_err.nb, err.knn)
model_names = c("LDA", "QDA", "Logistic", "NB", "KNN")
my_colors1 = c("#85144b", "#9B1F5A", "#B0296F", "#C62D85", "#DB2E8C",
                         "#FF69B4", "#FFADD8", "#FFC1E0", "#FFD1E7", "#FFE5F2")
barplot(test_error, col=my_colors1, names.arg=model_names,
        xlab="Model", ylab="Test Error", ylim=c(0, max(test_error)*1.5),
        main = "Test Errors of Methods")
text(x = seq(from=0.7, to=12, by=1.2), y = test_error + 0.05,
     labels = round(test_error, 3),col = "black")
```



The smallest test error is 0.15 from LDA and logistic.