Denetimli İstatistiksel Öğrenme Final

Buse BALTACIOĞLU ~ 2019900540

07 02 2021

Doğrusal regresyon (LM)

Değişken seçim yöntemlerinden adımsal regresyon uyguladık ve yeni modelimizi kurduk.

```
##
## Call:
## lm(formula = train_a$Bwt ~ train_a$UI + train_a$Ltw + train_a$Race +
##     train_a$Smoke, data = train_a)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1768.95  -405.89    51.59   478.70  1658.55
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2762.561    279.649   9.879  < 2e-16 ***
## train_a$UI1    -573.872    159.174  -3.605 0.000448 ***
## train_a$Ltw       4.625      1.975   2.342 0.020740 *
## train_a$Race2  -544.495    184.754  -2.947 0.003823 **
## train_a$Race3  -349.218    131.983  -2.646 0.009185 **
## train_a$Smoke1 -357.621    125.754  -2.844 0.005203 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 646.4 on 126 degrees of freedom
## Multiple R-squared:  0.2252, Adjusted R-squared:  0.1944
## F-statistic: 7.323 on 5 and 126 DF,  p-value: 4.695e-06
```

UI1,Ltw,Race2,Race3,Smoke1 %5 önem düzeyinde anlamlı çıkmıştır. p<alfa olduğu içinde model anlamlı çıkmıştır. 5 bağımsız değişken doğum ağırlığını %23 açıklamaktadır.

H_0: Bwt değişkeni normal dağılır. H_1: Bwt değişkeni normal dağılmaz.

```
##
##  Shapiro-Wilk normality test
##
## data:  train_a$Bwt
## W = 0.99331, p-value = 0.7909
```

0.7909>0.05 olduğu için H_0 reddedilemez. Bwt değişkeni normal dağılır.

H_0: Artıklar normal dağılır. H_1: Artıklar normal dağılmaz.

```
##
##  Shapiro-Wilk normality test
##
## data:  Lm_new$residuals
## W = 0.99422, p-value = 0.8727
```

0.8727>0.05 olduğu için H_0 reddedilemez. Hatalar normal dağılır.

H_0:Artıkların varyansı homojendir H_1:Artıkların varyansı heterojendir

```
##
##  studentized Breusch-Pagan test
##
## data:  Lm_new
## BP = 5.7093, df = 5, p-value = 0.3355
```

0.3355>0.05 olduğu için H_0 reddedilemez. Artıkların varyansı homojendir.

```
##                   GVIF Df GVIF^(1/(2*Df))
## train_a$UI    1.029034  1        1.014413
## train_a$Ltw   1.162759  1        1.078313
## train_a$Race  1.298782  2        1.067540
## train_a$Smoke 1.145507  1        1.070284
```

Tüm değişkenlerin vif değerleri 5'ten küçük olduğu için aralarında çoklu doğrusal bağlantı sorunu yoktur.

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##   rstudent unadjusted p-value Bonferroni p
## 2 -2.91993          0.0041544      0.54838
```

2. gözlem uç değer çıkmıştır.

Tüm doğrusal regresyon varsayımlarını modelimiz sağlamıştır.

Doğrusal regresyon modeli için temel performansları

lm	rmse	mae	mape
train	631.5292	519.1709	0.1988
test	890.4845	689.9774	0.3512

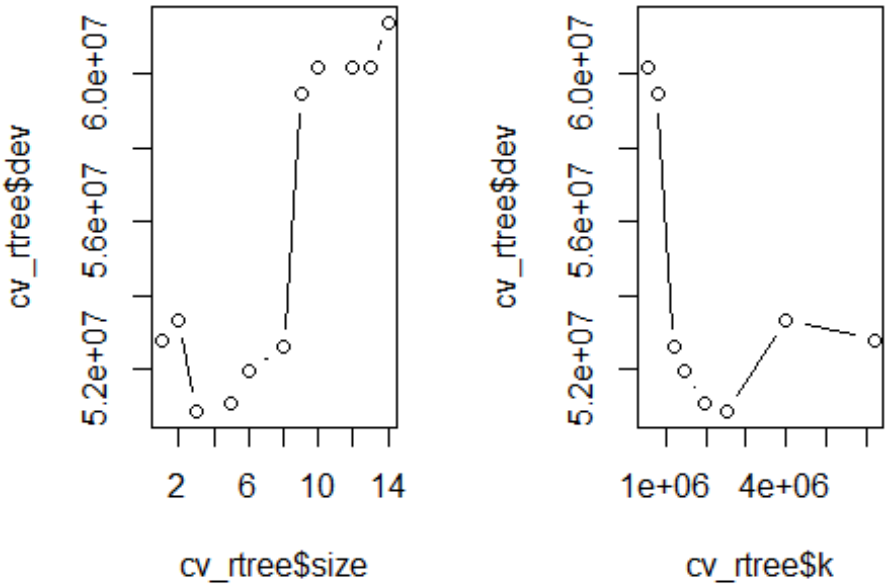
Regresyon Ağacı (RT)

```
##
## Regression tree:
## tree(formula = train_a$Bwt ~ ., data = train_a, subset = train_dataa)
## Variables actually used in tree construction:
## [1] "UI"    "Ltw"   "Race"  "Age"   "Smoke" "Ftv"   "Ptl"
## Number of terminal nodes:  14
## Residual mean deviance:  304900 = 24090000 / 79
```

```
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1091.0  -417.0    39.0     0.0   344.4  1037.0
```

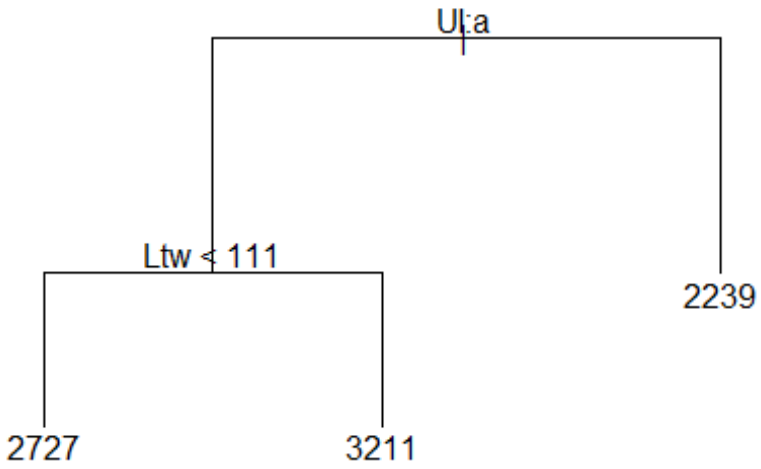
Terminal node sayısı 14, artıkların ortalamadan sapması 304900

Oluşan ağaca baktığımızda gereksiz dallanmalar olduğu görülmekte bu sebepten dolayı budamamız gerekir.



Dirsek noktalarını dikkate alırsak 3 terminal node incelenmelidir.

```
##
## Regression tree:
## snip.tree(tree = Rt, nodes = c(4L, 3L, 5L))
## Variables actually used in tree construction:
## [1] "UI"  "Ltw"
## Number of terminal nodes:  3
## Residual mean deviance:  425100 = 38260000 / 90
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1482.00  -390.30    13.75     0.00   439.70  1382.00
```



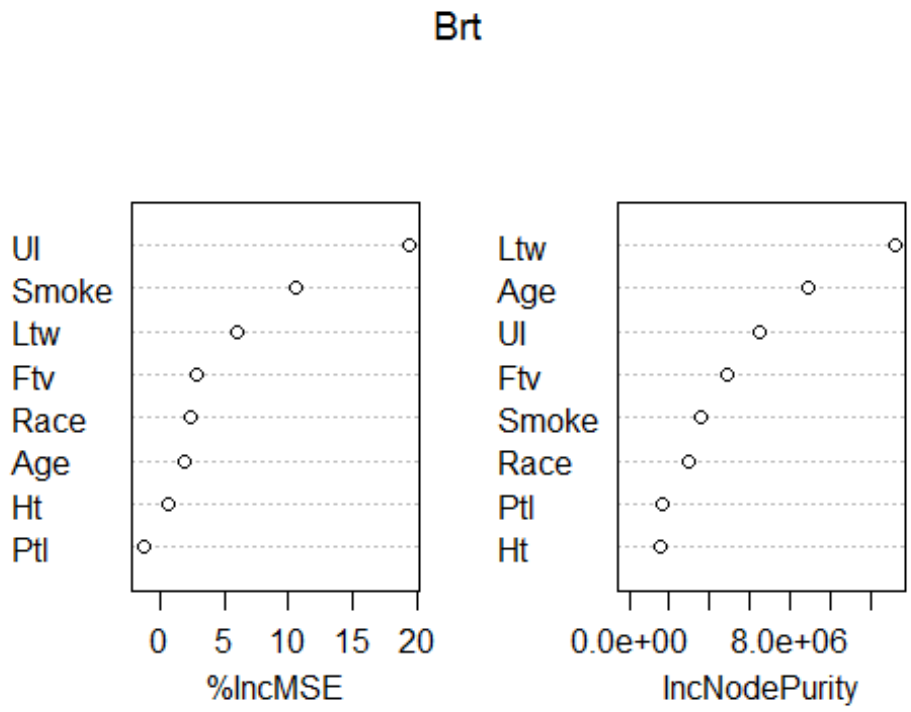
Regresyon ağacı modeli için temel performansları

Rt	rmse	mae	mape
train	661.9345	527.1941	0.1993
test	713.4994	563.1298	0.2585

Bagging ile regresyon ağacı (BRT)

```
##
## Call:
##  randomForest(formula = train_a$Bwt ~ ., data = train_a, mtry = 8,
importance = TRUE, subset = train_dataa, na.action = na.omit)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 8
##
##           Mean of squared residuals: 487849.2
##                     % Var explained: 6.35
```

MSR=497620 ve varyans açıklama oranı 6.35



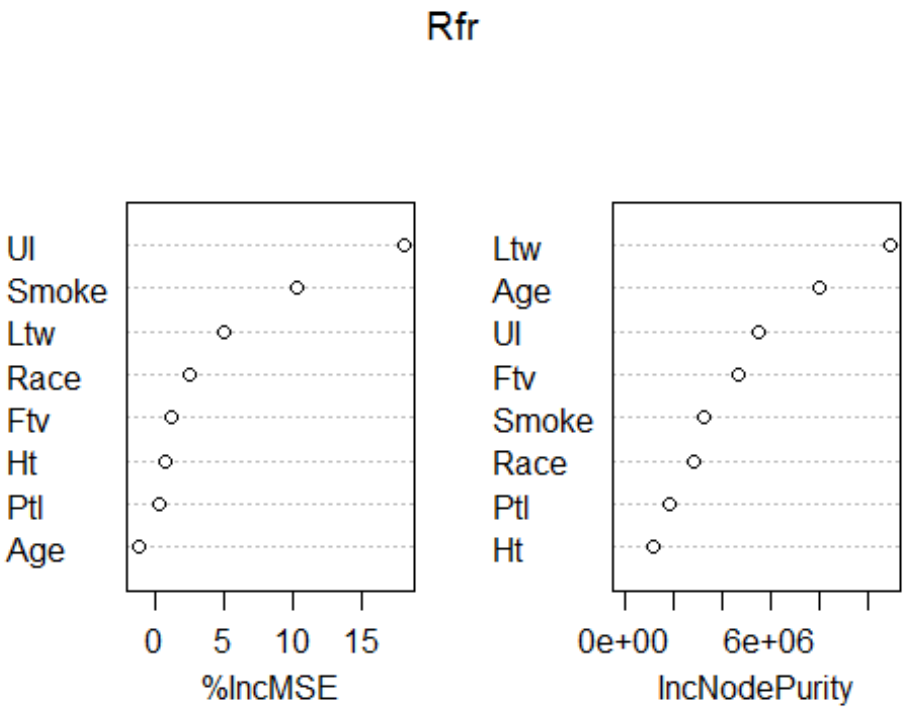
Bagging ile regresyon ağacı modeli için temel performansları

Brt	rmse	mae	mape
train	447.9892	335.6576	0.1315
test	684.1947	557.0798	0.2501

Rassal Ormanlar Regresyonu (RFR)

```
##
## Call:
##  randomForest(formula = train_a$Bwt ~ ., data = train_a, mtry = 3,
importance = TRUE, subset = train_dataa, na.action = na.omit)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 474943.6
##                     % Var explained: 8.82
```

MSR=474943.6 ve varyans açıklama oranı 8.82



Rassal ormanlar regresyon modeli için temel performansları

Rfr	rmse	mae	mape
train	469.0678	365.8806	0.1440
test	683.0983	542.8808	0.2505

Test verisi üzerinde performanslarını karşılaştırırsak

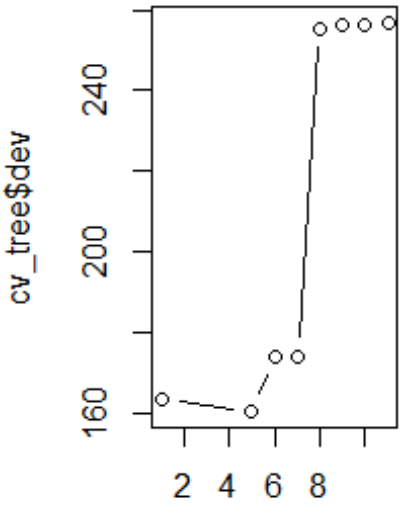
```
| rmse          | mae           | mape
```

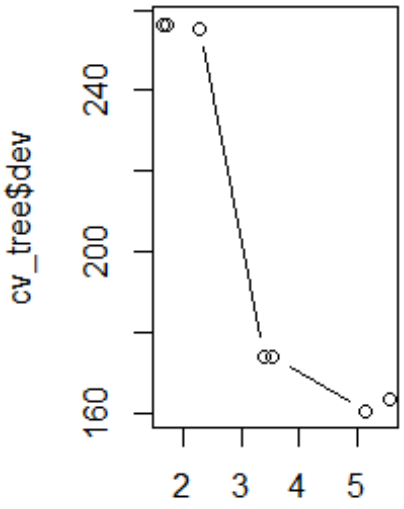
— | ——— | ——— | ———- lm | 890.4845 | 689.9774 | 0.3512 rt | 713.4994 | 563.1298 | 0.2585 brt | 684.1947 | 557.0798 | **0.2501** rfr | **683.0983** | **542.8808** | 0.2505

Modellerimizin test verisi üzerindeki performansları yukarudaki tabloda verilmiştir. En uygun modelin rassal ormanlar regresyonu olduğuna karar verilmiştir.

Sınıflandırma Ağacı (CT)

Terminal node sayısı 11, artıkların ortalamadan sapması 0.9398 ve accuracy=76 den oluşan ağaca baktığımızda gereksiz dallanmalar olduğunu görmekteyiz bu sebepten dolayı budamamız gerekir.

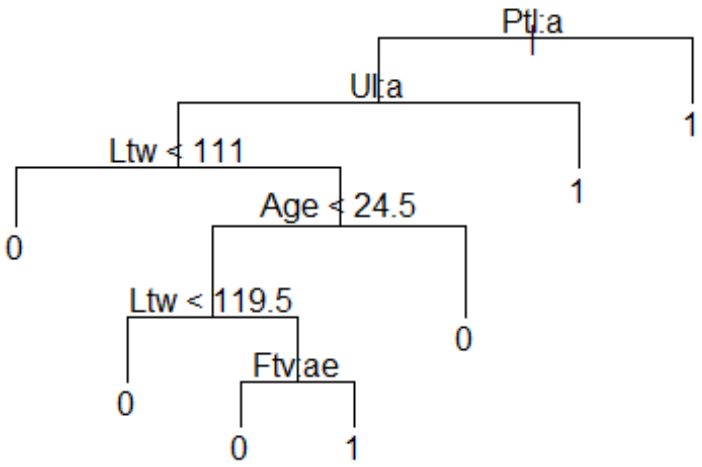




Grafiğe baktığımızda 5 denenmeli diyebiliriz fakat denediğimizde gereksiz dallanma olduğunu ve node değerini 5 in altında alamayacağımızı görüyoruz. Yapılan denemeler üzerine 7 terminal node sayısı olarak alınmıştır.

```
##
## Classification tree:
## snip.tree(tree = Ct, nodes = c(5L, 3L, 74L))
## Variables actually used in tree construction:
## [1] "Ptl" "UI"  "Ltw" "Age" "Ftv"
## Number of terminal nodes:  7
## Residual mean deviance:  1.001 = 86.13 / 86
## Misclassification error rate: 0.2581 = 24 / 93
```

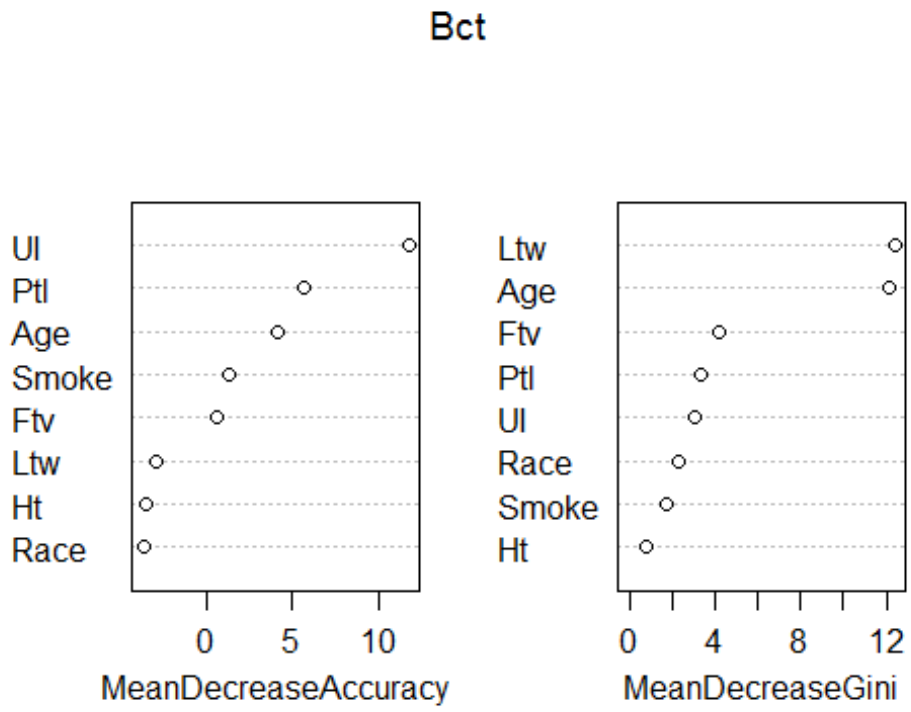
Terminal node sayısı 7, artıkların ortalamadan sapması 1.001 ve accuracy=74



accuracy	train	test
ct	0.72	0.67

Bagging ile sınıflandırma ağacı (BCT)

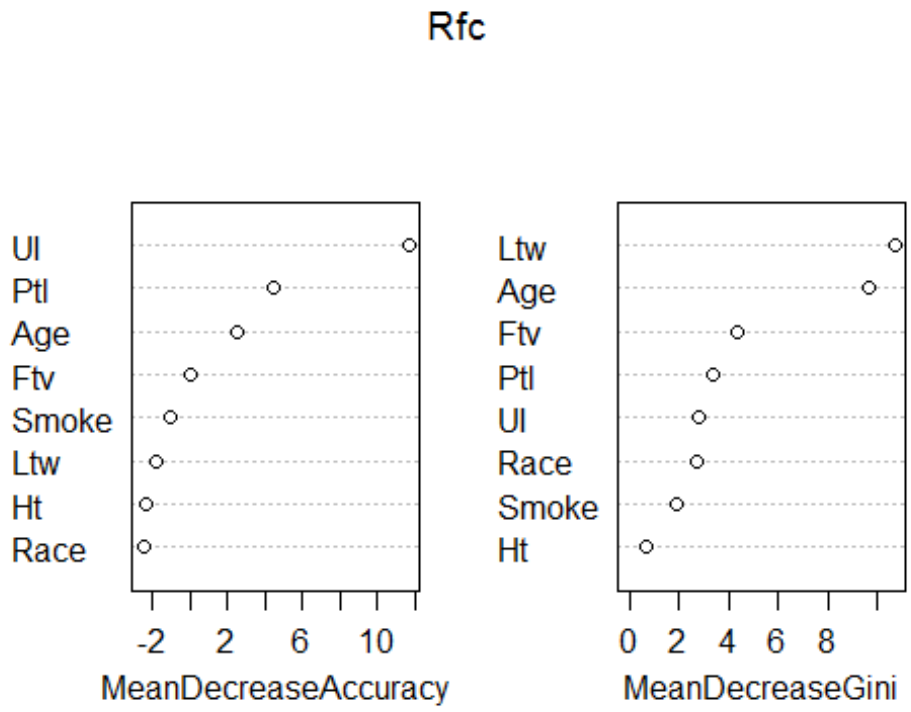
```
##
## Call:
##  randomForest(formula = train_b$Low ~ ., data = train_b, mtry = 8,
importance = TRUE, subset = train_datab, na.action = na.omit)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 8
##
##         OOB estimate of  error rate: 34.41%
## Confusion matrix:
##    0  1 class.error
## 0 53 10   0.1587302
## 1 22  8   0.7333333
```



accuracy	train	test
bct	0.89	0.72

Rassal Ormanlar ile Sınıflandırma Ağacı (RFC)

```
##
## Call:
##  randomForest(formula = train_b$Low ~ ., data = train_b, mtry = 3,
importance = TRUE, subset = train_datab, na.action = na.omit)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 33.33%
## Confusion matrix:
##    0  1 class.error
## 0 53 10   0.1587302
## 1 21  9   0.7000000
```



accuracy	train	test
rfc	0.89	0.70

Lojistik Regresyon (LR)

Full model için %95 güvenle 0.0002<0.05 Ho reddedilir ve model geçerlidir.Ptl1 değişkenleri %95 güvenle anlamlı çıkmıştır. Fakat biz değişken seçimi için AIC yöntemi kullanılmıştır. % değişkenli yeni modelimizi kurduk.

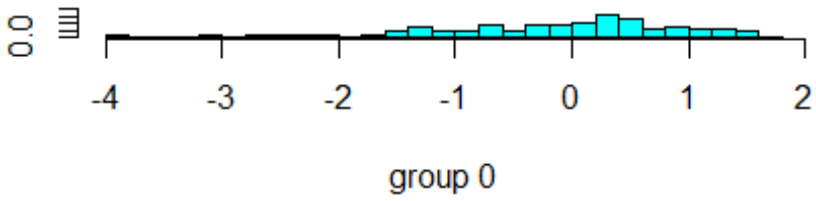
```
##
## Call:
## glm(formula = train_b$Low ~ train_b$Ltw + train_b$Race + train_b$Smoke +
##     train_b$Ptl + train_b$Ht + train_b$UI, family = binomial,
##     data = train_b)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0102  -0.8219  -0.5366   0.9251   2.1795
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.343887   1.148244   0.299  0.76457
## train_b$Ltw    -0.018996   0.008485  -2.239  0.02518 *
## train_b$Race2   1.465833   0.670113   2.187  0.02871 *
## train_b$Race3   0.824589   0.518135   1.591  0.11151
```

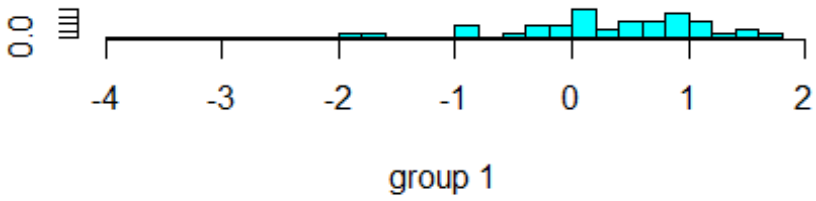
```
## train_b$Smoke1  0.820135   0.488278   1.680  0.09303 .
## train_b$Ptl1    1.694124   0.642109   2.638  0.00833 **
## train_b$Ptl2    0.471768   1.090607   0.433  0.66532
## train_b$Ht1     1.314175   0.871294   1.508  0.13148
## train_b$UI1     1.048274   0.554565   1.890  0.05872 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.62  on 131  degrees of freedom
## Residual deviance: 140.04  on 123  degrees of freedom
## AIC: 158.04
##
## Number of Fisher Scoring iterations: 4
```

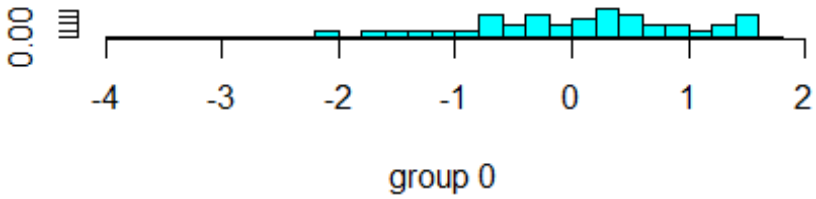
accuracy	train	test
lr	0.70	0.63

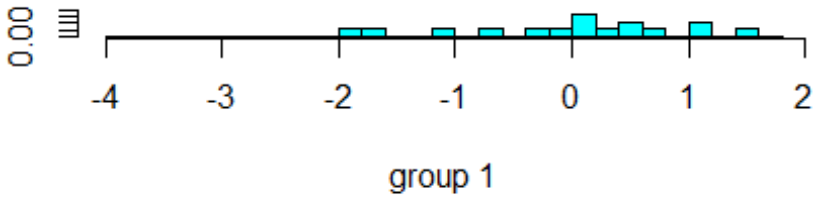
Doğrusal Ayırma Analizi (LDA)

```
## Call:
## lda(train_b$Low ~ train_b$Age + train_b$Ltw, data = train_b)
##
## Prior probabilities of groups:
##         0         1
## 0.6742424 0.3257576
##
## Group means:
##   train_b$Age train_b$Ltw
## 0    23.39326    132.9775
## 1    22.18605    119.7442
##
## Coefficients of linear discriminants:
##                       LD1
## train_b$Age -0.07486431
## train_b$Ltw -0.02895282
```









accuracy	train	test
lda	0.68	0.63

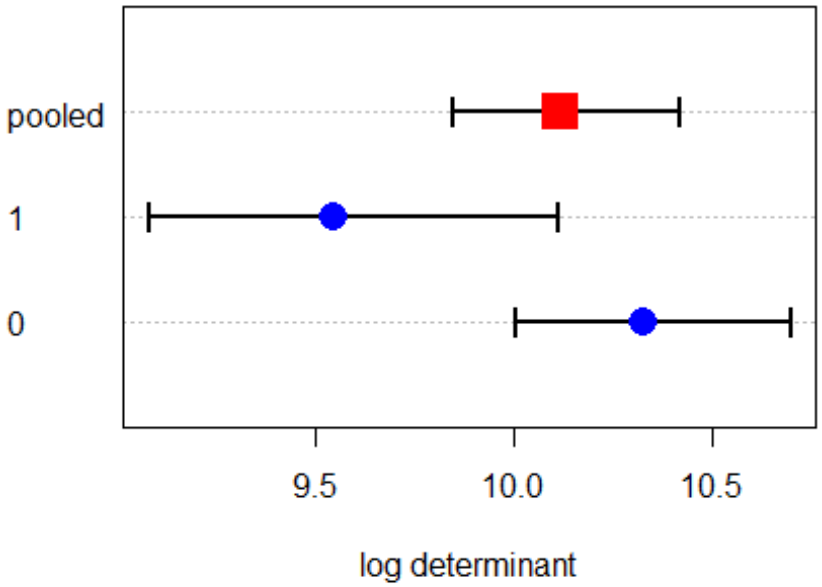
```
##             Henze-Zirkler test for Multivariate Normality
##
##   data : norm_0[, -1]
##
##   HZ              : 3.53203
##   p-value         : 3.563469e-08
##
##   Result  : Data are not multivariate normal (sig.level = 0.05)
```

```
##             Henze-Zirkler test for Multivariate Normality
##
##   data : norm_1[, -1]
##
##   HZ              : 0.8375591
##   p-value         : 0.0734444
##
##   Result  : Data are multivariate normal (sig.level = 0.05)
```

H_0 = Varyans kovaryans matrisi eşittir. H_1 = Varyans kovaryans matrisi eşit değildir.

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  data_b[, c(2, 3)]
## Chi-Sq (approx.) = 5.7226, df = 3, p-value = 0.1259
```

p-value = 0.1259 > 0.05 olduğu için %95 güvenle Ho reddedilemez. Değişkenlerin varyans kovaryans matrisi eşittir.



Eğrisel Ayırma Analizi (QDA)

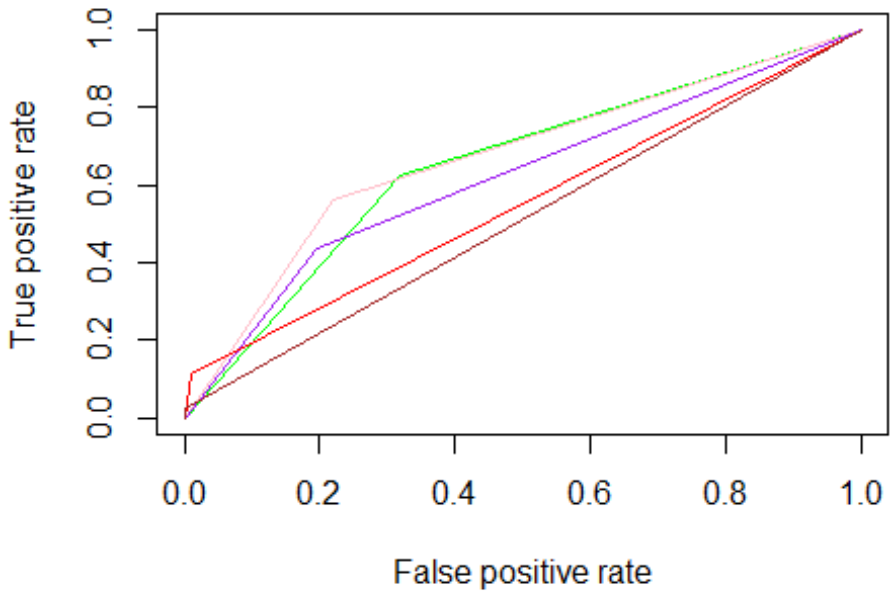
```
## Call:
## qda(train_b$Low ~ train_b$Age + train_b$Ltw, data = train_b)
##
## Prior probabilities of groups:
##         0         1
## 0.6742424 0.3257576
##
## Group means:
##   train_b$Age train_b$Ltw
## 0    23.39326    132.9775
## 1    22.18605    119.7442
```

accuracy	train	test
qda	0.67	0.64

Test verisi üzerinde performanslarını karşılaştırırsak

accuracy	ct	bct	rfc	lr	lda	qda
test	0.67	**0.72**	0.70	0.63	0.63	0.64

Tüm modellere ait ROC eğrisini tek bir grafik üstünde göstererek, eğri altında kalan alan (AUC) hesaplaması



auc	ct	bct	rfc	lr	lda	qda
test	0.65	**0.67**	0.62	0.54	0.51	0.5

Bu veri seti için genel olarak en uygun modelleme seçimi

Model performanslarını karşılaştırdığımızda en iyi modelin bagging ile sınıflandırma ağacı olarak seçilmiştir.