VERİ MADENCİLİĞİNE GİRİŞ
DERS PROJESİ

ÖĞRENCİ İSİM&SOYİSİM : ÇETİN TEKİN
ÖĞRENCİ NO: 17011603
KONU: HABERMAN’S SURVIVAL

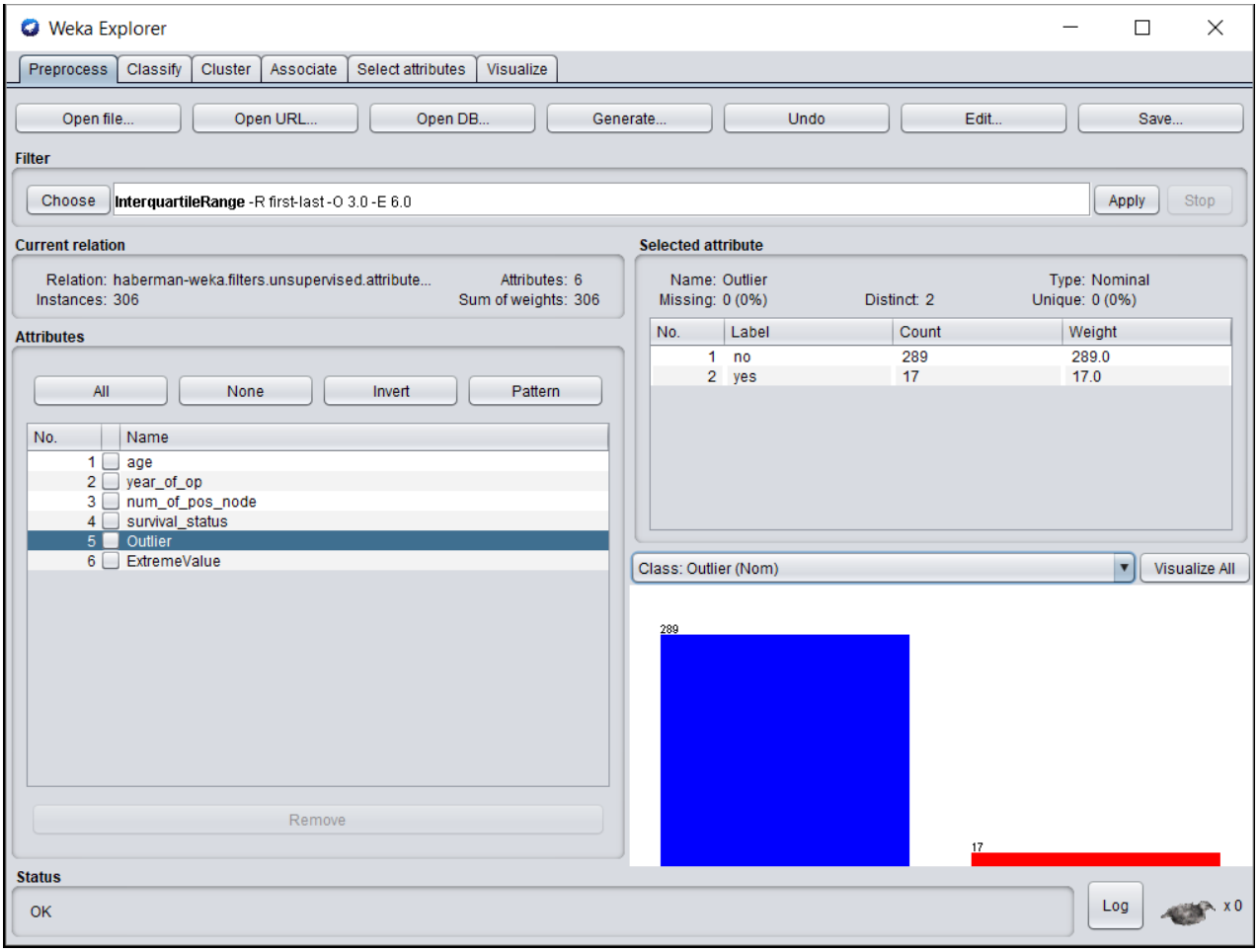
VERİ SETİ TANITIMI

Projede kullanılan Haberman’s Survival veri seti, 1958 ile 1970 arasında Chicago Üniversitesi Hastanesi’nde yapılan araştırmadan elde edilmiştir. Araştırmada meme kanseri tedavisi için ameliyat olmuş kişilerin hayatta kalıp kalmadıklarına odaklanılmıştır.

Veri setinde biri sınıf etiketi olmak üzere 4 adet sayısal özellik kullanılmıştır. Bunlar ameliyat olunan zamanda hastanın yaşı, hastanın ameliyat yılı, tespit edilen pozitif meme kanseri düğümleri ve kurtarılma durumudur. Kurtarılma durumu 1 ise hastanın 5 yıl veya daha fazla yaşadığı 2 ise 5 yıldan az yaşadığı anlamına gelmektedir.

VERİ ANALİZİ

Veri analizi aşamasında öncelikle veri setinde aykırı değer olup olmadığı araştırılmıştır. WEKA üzerinde InterQuartileRange filtresi kullanılarak aykırı değerler tespit edilmiştir. Elde edilen aykırı değer sayısı Şekil 1.1'de verilmiştir.



Şekil 1.1

Görüldüğü üzere veri setinde 17 adet aykırı değer tespit edilmiştir. Bu değerler veri setinden temizlenerek sınıflandırma ve kümeleme adımlarına geçilmiştir.

SINIFLANDIRMA

Projede WEKA ile Naive Bayes,Decision Tree ve K-NN algoritmaları kullanılmıştır. Naive Bayes yöntemi C ile kodlanmıştır.

Naive Bayes yöntemi sınıflandırma sonuçları

```
Correctly Classified Instances         219               76.8421 %
Incorrectly Classified Instances        66               23.1579 %
Kappa statistic                          0.2234
Mean absolute error                      0.3195
Root mean squared error                  0.4249
Relative absolute error                 86.8193 %
Root relative squared error             99.1928 %
Total Number of Instances              285

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,935    0,754    0,795      0,935    0,860      0,250    0,644     0,811     1
                 0,246    0,065    0,548      0,246    0,340      0,250    0,644     0,396     2
Weighted Avg.    0,768    0,587    0,736      0,768    0,734      0,250    0,644     0,710

=== Confusion Matrix ===

   a   b   <-- classified as
 202  14 |   a = 1
  52  17 |   b = 2
```

Naive Bayes yöntemi k=10 değeri kullanılarak yapılan K-fold cross validation ile WEKA üzerinde %76 başarı elde edilmiştir.
Naive Bayes yöntemi C ile kodlanarak k=10 değeri K-fold cross validation yapıldığında elde edilen sonuç aşağıdaki gibidir:

```
Please enter the k value for k fold cross validation: 10
Accuracy: 76.140351
Total number of instances: 285

Confusion Matrix:

200 16
52 17
tekin@tekin-Lenovo-Z50-70:~/Desktop/gitRepos/dataMiningProje$
```

Görüldüğü üzere karmaşıklık matrisleri üzerindeki ufak farklılıklar dışında genel olarak aynı sonuçlar elde edilmiştir.

K-NN yöntemi sınıflandırma sonuçları

```
Correctly Classified Instances         195               68.4211 %
Incorrectly Classified Instances        90               31.5789 %
Kappa statistic                          0.0949
Mean absolute error                      0.3293
Root mean squared error                  0.5653
Relative absolute error                 89.4701 %
Root relative squared error            131.9491 %
Total Number of Instances              285

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,815    0,725    0,779      0,815    0,796      0,095    0,559     0,794     1
                 0,275    0,185    0,322      0,275    0,297      0,095    0,559     0,274     2
Weighted Avg.    0,684    0,594    0,668      0,684    0,675      0,095    0,559     0,668

=== Confusion Matrix ===

   a   b   <-- classified as
 176  40 |   a = 1
  50  19 |   b = 2
```

K-NN yöntemi ile WEKA üzerinde %68 başarı elde edilmiştir.

Decision Tree yöntemi sınıflandırma sonuçları

```
Correctly Classified Instances         195               68.4211 %
Incorrectly Classified Instances        90               31.5789 %
Kappa statistic                          0.0949
Mean absolute error                      0.3293
Root mean squared error                  0.5653
Relative absolute error                 89.4701 %
Root relative squared error            131.9491 %
Total Number of Instances              285

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,815    0,725    0,779      0,815    0,796      0,095    0,559     0,794     1
                 0,275    0,185    0,322      0,275    0,297      0,095    0,559     0,274     2
Weighted Avg.    0,684    0,594    0,668      0,684    0,675      0,095    0,559     0,668

=== Confusion Matrix ===

   a   b   <-- classified as
 176  40 |   a = 1
  50  19 |   b = 2
```

Decision Tree yöntemi ile WEKA üzerinde %73 başarı elde edilmiştir.

KÜMELEME

Projede WEKA ile K-means clustering, Hierarchical Clustering ve Cobweb yöntemleri kullanılmıştır. K-means clustering yöntemi C ile kodlanmıştır.

K-means clustering yöntemi kümeleme sonuçları

```
=== Model and evaluation on training set ===

Clustered Instances

0       131 ( 46%)
1       154 ( 54%)


Class attribute: survival_status
Classes to Clusters:

   0   1  <-- assigned to cluster
 101 115 | 1
  30  39 | 2

Cluster 0 <-- 2
Cluster 1 <-- 1

Incorrectly clustered instances :       140.0    49.1228 %
```

K-means clustering yöntemi ile oldukça başarısız bir sonuç elde edilmiştir. Bu başarısızlığın arkasında yapılan kontroller ile ne olduğu anlaşılmıştır. Clusterlardaki sınıf etiketlerine göre majority voting yapılırken her iki cluıster için de aynı sınıf etiketi baskın çıkmaktadır. Durum böyle olunca aslında veri setinin kendi sınıf etiketlerinin (2 tane) veriyi ikiye ayırmada yetersiz olduğu görülmüştür.

K-means clustering yöntemi C ile kodlanmıştır. Elde edilen sonuçlar aşağıdaki gibidir:

```
Please enter number of clusters:
2


K means clustering total number of iterations:
4

Final cluster centroids:

Centroid 1: (44.443039 ,62.575951 ,2.721519)
Centroid 2: (62.496063 ,63.149605 ,2.157480)
Number of elements in clusters:

Cluster 1: 158
Cluster 2: 127
Cluster majors are same!!!
Incorrectly clustered samples percentage: 47.017544
tekin@tekin-Lenovo-Z50-70:~/Desktop/gitRepos/dataMiningProje/clustering$
```

Hierarchical clustering yöntemi kümeleme sonuçları

```
=== Model and evaluation on training set ===

Clustered Instances

0      281 ( 99%)
1        4 (  1%)


Class attribute: survival_status
Classes to Clusters:

   0   1  <-- assigned to cluster
 214   2 | 1
  67   2 | 2

Cluster 0 <-- 1
Cluster 1 <-- 2

Incorrectly clustered instances :       69.0     24.2105 %
```

Hierarchical clustering yöntemi ile veri setinin mevcut sınıf etiketleri arasında %76 gibi oldukça yüksek bir eşleşme olmuştur.

Cobweb clustering yöntemi kümeleme sonuçları

```
Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

1      217 ( 76%)
2       68 ( 24%)


Class attribute: survival_status
Classes to Clusters:

   1   2  <-- assigned to cluster
 178  38 | 1
  39  30 | 2

Cluster 1 <-- 1
Cluster 2 <-- 2

Incorrectly clustered instances :       77.0     27.0175 %
```

Cobweb kümeleme yöntemi ile 0.1 threshold kullanılarak yapılan kümelemede oldukça yüksek bir başarı elde edilmiştir.