Veri Madenciliği

Güz 2023

Ders 8

WEKA

Sınıflandırma

- Tahmin edilen hedef kategorik/nominal olmalıdır
- Uygulanan yöntemler
  - Karar ağaçları (J48, vb.)
  - Kurallar (ZeroR, OneR, vb.)
  - Naif Bayes
- Değerlendirme yöntemleri
  - Test veri kümesi
  - Çapraz doğrulama

Sınıflandırma

- Algoritma
  - ZeroR: Tüm öznitelikleri yoksayar ve yalnızca hedef sınıfa dayanır. Her zaman çoğunluk değerini tahmin eder.
  - OneR: Her öznitelik için bir kural yapın (özniteliğin her değeri için sonuçların sıklığına bağlı olarak). En küçük hatayı veren kuralı/özniteliği seçin.
  - Naif Bayes: Bayes Teorem'ine dayanan olasılıksal bir sınıflandırıcı. Tüm özniteliklerin bağımsız olduğunu varsayar.

Değerlendirme Yöntemleri

- Test Veri Kümesi
  - Tüm veriler üzerinde eğitim; Tüm verileri sınama (önerilmez)
  - Verileri bölün (örneğin eğitim için% 66, test için% 34).
  - Biri eğitim örnekleriyle, diğeri test örnekleriyle ayrı dosyalar kullanın.
- Çapraz Doğrulama:
  - Veri kümesini gruplara bölme (ör. 10 örnek grubu)
  - Test için bir grup seçin, gerisini eğitim için kullanın
  - Her seferinde test etmek için farklı grupla birden çok kez tekrarlayın. (Örneğin, her seferinde test için 10 orijinal gruptan birini kullanarak 10 kez tekrarlayın ve geri kalanı eğitim için).
  - Tüm testlerin sonuçlarının ortalamasını al.

WEKA Veri Biçimleri

- Veriler bir dosyadan çeşitli biçimlerde içe aktarılabilir:
  - ARFF 'nin (Öznitelik İlişkisi Dosya Biçimi) iki bölümü vardır:
    - Üstbilgi öznitelik adını, türünü ve ilişkilerini tanımlar.
    - Veri bölümü veri kayıtlarını (örnekleri) listeler.
  - CSV: Virgülle Ayrılmış Değerler (metin dosyası)
  - C4.5: Karar indüksiyon algoritması tarafından kullanılan bir biçim, iki ayrı dosya gerektirir
    - Ad dosyası: özniteliklerin adlarını tanımlar
    - Veri dosyası: kayıtları listeler (örnekler)
  - ikili (binary)
- Veriler bir URL'den veya SQL veritabanından da okunabilir (JDBC kullanılarak; Java DataBase Bağlantısı, bir istemcinin veritabanına nasıl erişebileceğini tanımlayan bir Java API'sidir)

Öznitelik İlişkisi Dosya Biçimi (arff)

ARFF dosyaları iki ayrı bölümden oluşur:

- Üstbilgi bölümü öznitelik adını, türünü ve ilişkilerini tanımlar, bir anahtar sözcükle başlar.
  - @relation <veri adı>
  - @attribute <öznitelik adı> <tür> veya {aralığı}
- Veri bölümü veri kayıtlarını listeler,
  - @data
  - veri örnekleri listesi ile başlar
  - Açıklama: % ile başlayan herhangi bir satır

ARFF'de Meme Kanseri verileri

% Meme Kanseri verileri*: 286 örnek (tekrarlanmayan-olaylar: 201, tekrarlanan-olaylar: 85)

% Bölüm 1: Öznitelik adı, türleri ve ilişkileri tanımları

@relation meme-kanseri

@attribute yaş {'10-19','20-29','30-39','40-49','50-59','60-69','70-79','80-89','90-99'}
@attribute menopoz {'lt40','ge40', 'menopoz-öncesi'}
@attribute tümör-boyutu
{'0-4','5-9','10-14','15-19','20-24','25-29','30-34','35-39','40-44','45-49','50-54','55-59'}
@attribute inv-düğümleri
{'0-2','3-5','6-8','9-11','12-14','15-17','18-20','21-23','24-26','27-29','30-32','33-35','36-39'}
@attribute düğüm-başı {‘evet', 'hayır'}
@attribute kötülük-derecesi {'1','2','3'}
@attribute meme {‘sol','sağ'}
@attribute göğüs-dörtgeni {'sol-üst', 'sol-alt', 'sağ-üst', 'sağ-alt', 'merkez'}}
@attribute ışınlama {‘evet', 'hayır'}
@attribute **Sınıf** {'tekrarlanmayan-olaylar', 'tekrarlanan-olaylar'}

% Bölüm 2: Veri Bölümü

@data

'40-49', 'menopoz-öncesi','15-19','0-2', ‘evet','3', 'sağ', ‘sol-üst', ‘hayır', 'tekrarlanan-olaylar'
'50-59','ge40','15-19','0-2', 'hayır','1', 'sağ', 'merkez', 'hayır', 'tekrarlanmayan-olaylar'
'50-59','ge40','35-39','0-2', 'hayır','2', 'sol', 'sol_alt', 'hayır', 'tekrarlanan-olaylar'
.....

- % kaynak: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer

Çıktıyı Yorumlama: Karışıklık Matrisi

Karışıklık matrisi, her sınıflandırma kategorisinde her sınıf değerinden kaç tane sınıflandırıldığını gösterir.

Karışıklık Matrisi:

a	b	<-- olarak sınıflandırılır
56	8	a = tekrarlanmayan-olaylar
23	10	b = tekrarlanan-olaylar

- 56 tekrarlanmayan-olay (a) doğru olarak sınıflandırıldı (a)
- 8 tekrarlanmayan-olay (a) yanlış olarak (b) olarak sınıflandırıldı
- 23 tekrarlanan-olaylar (b) yanlış olarak (a) olarak sınıflandırıldı
- 10 tekrarlanan-olay (b) doğru olarak (b) olarak sınıflandırıldı
- Ana köşegendeki öğeler doğru sınıflandırmalardır

Çıktıyı Yorumlama

Ağacın metin gösterimi:

```
J48 budanmış ağaç
------------------

düğüm-başı = evet
|   kötülük-derecesi = 1: tekrarlanan-olaylar (1.01/0.4)
|   kötülük-derecesi = 2: tekrarlanmayan-olaylar (26.2/8.0)
|   kötülük-derecesi = 3: tekrarlanan-olaylar (30.4/7.4)
Düğüm-başı = hayır: tekrarlanmayan-olaylar (228.39/53.4)
```

- Yaprak Sayısı : 4
- Ağacın boyutu : 6

WEKA Gezgini

- Weka GUI'de Explorer'a tıklayın

- Explorer penceresinde , "Dosyayı Aç"ı tıklatın
  - Veri dosyasını açmak için,

  örn. Meme Kanseri verileri:`meme kanseri tr.arff`

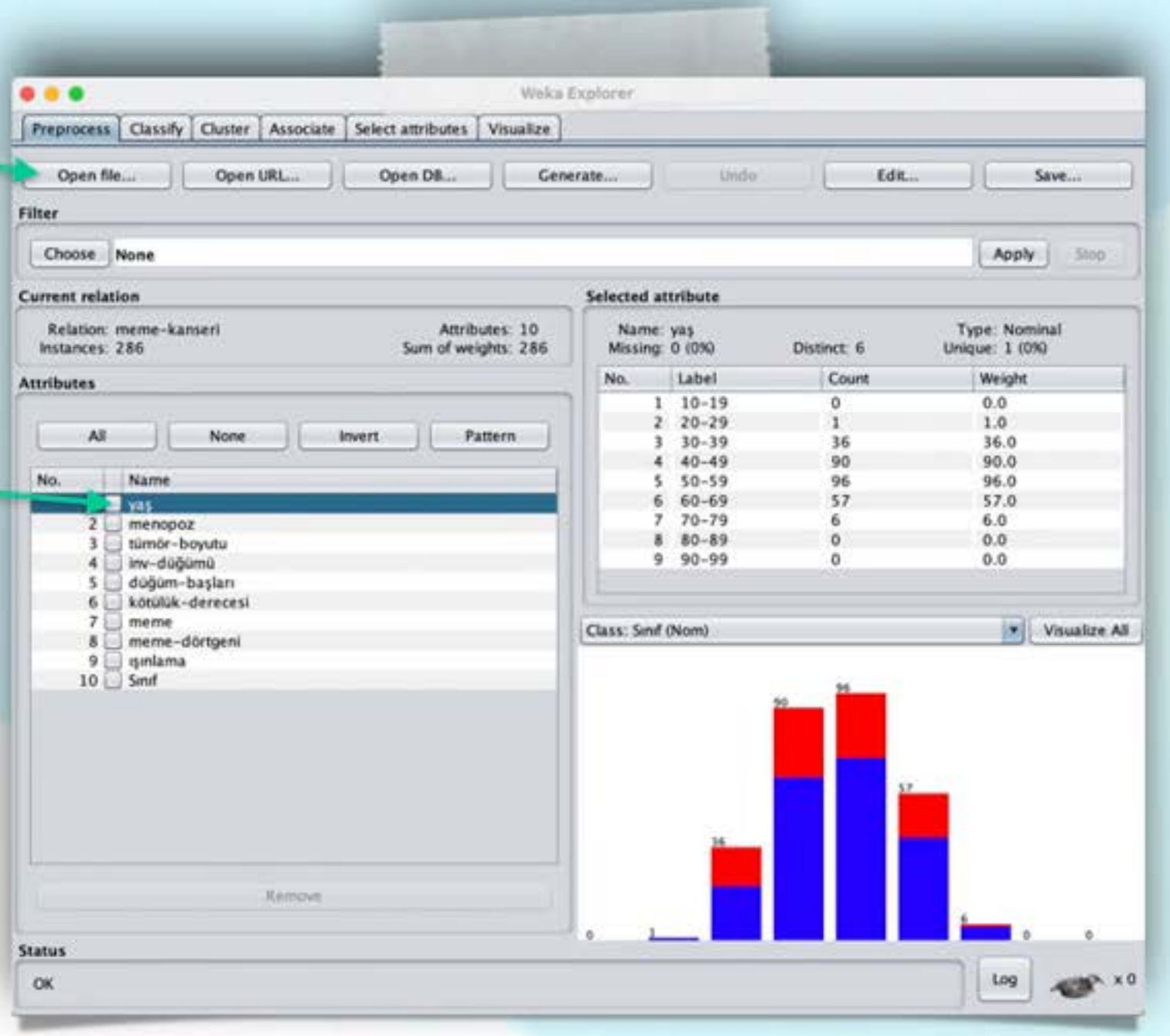
- Veya (bu veri kümeniz yoksa), WEKA paketi tarafından sağlanan veri klasörü

  Örn. `iris_tr.arff` veya `havaDurumu_nominal_tr.arff`

WEKA Explorer: Veri Dosyasını Açma

Meme-Kanseri verilerini açın.

Yaş gibi bir özniteliği tıklatın, dağılımı histogramda görüntülenir.



WEKA Explorer: Sınıflandırıcılar

Bir veri dosyası yükledikten sonra, Sekmeyi Sınıflandır'a tıklayın

- Sınıflandırıcı seçin, Sınıflandırıcı altında
  - Seç Düğmesi'ni tıklatın
  - Açılan menüden Ağaçlar Klasörü'nü tıklatın
  - J48'i seçin - bir karar ağacı algoritması

- Test seçeneği belirleme
  - Yüzde Bölme Düğmesini Seç
    - Eğitim için varsayılan oranı %66, test için %34 kullanın

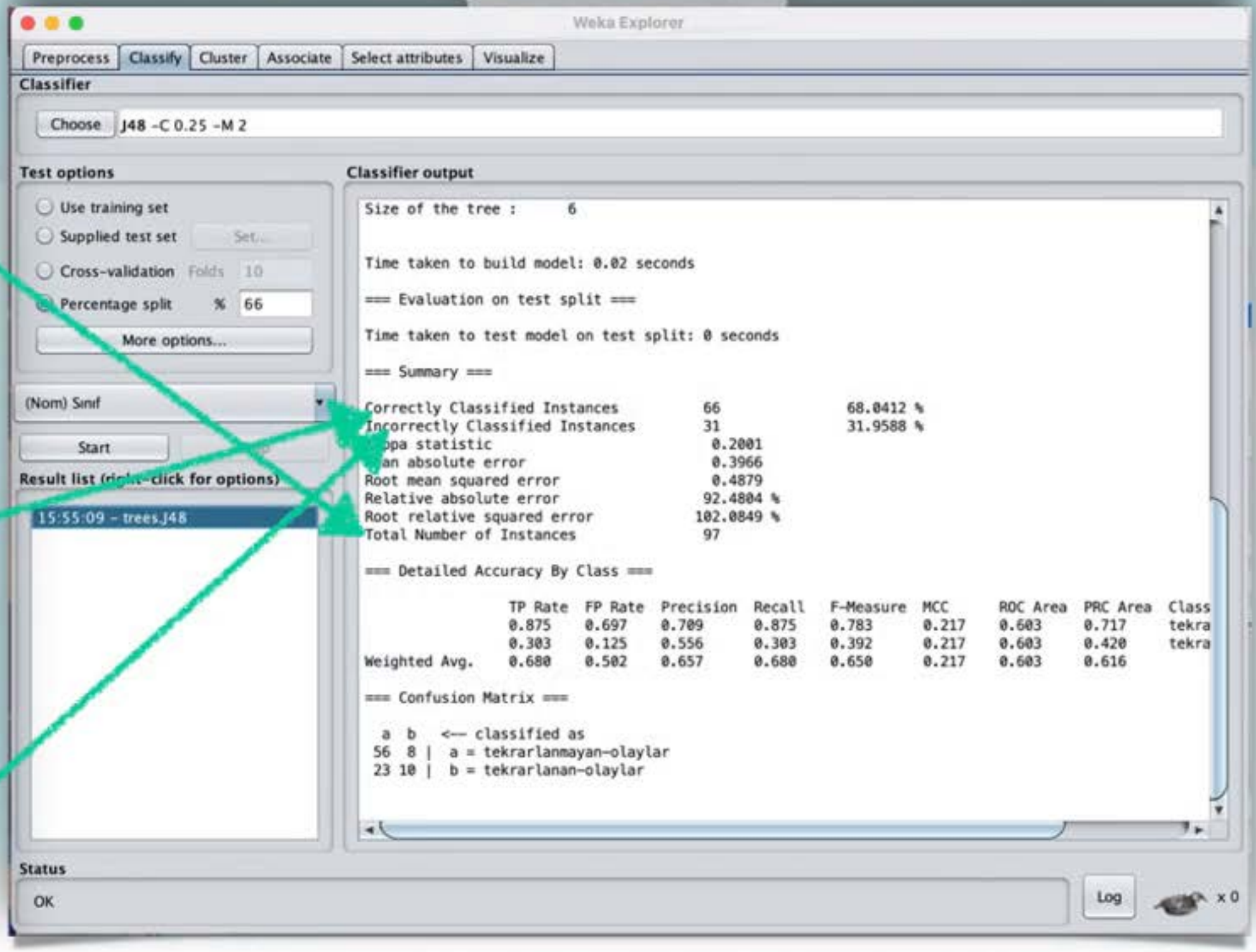
- Sınıflandırıcıyı eğitmek ve sınamak için Başlat Düğmesi'ni tıklatın.
  - Eğitim ve test bilgileri sınıflandırıcı çıktı penceresinde görüntülenecektir.

WEKA Explorer: Sonuçlar

Testte kullanılan 97 vaka.

Doğru: 66 (%68)

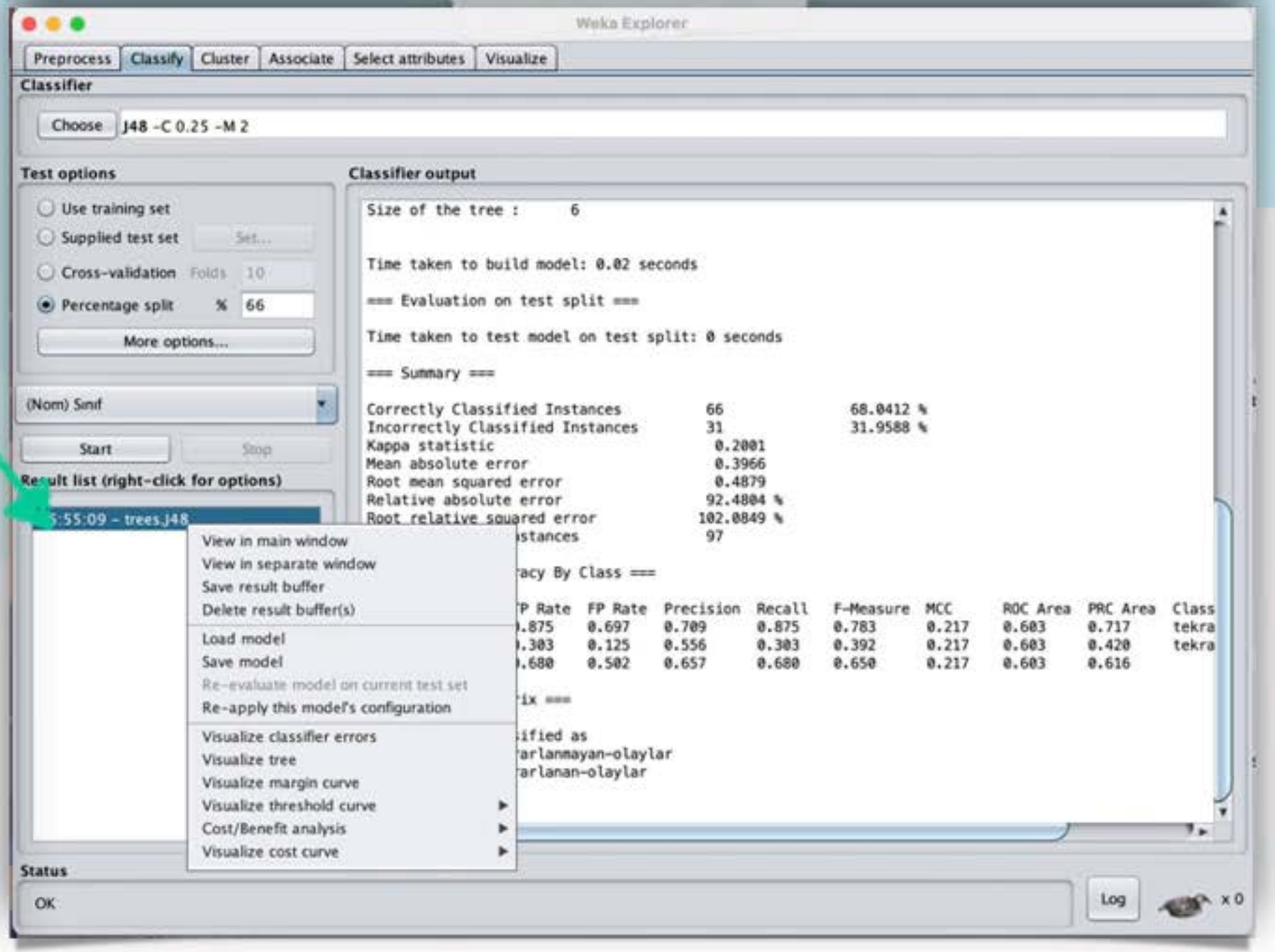
Yanlış: 31 (32%)



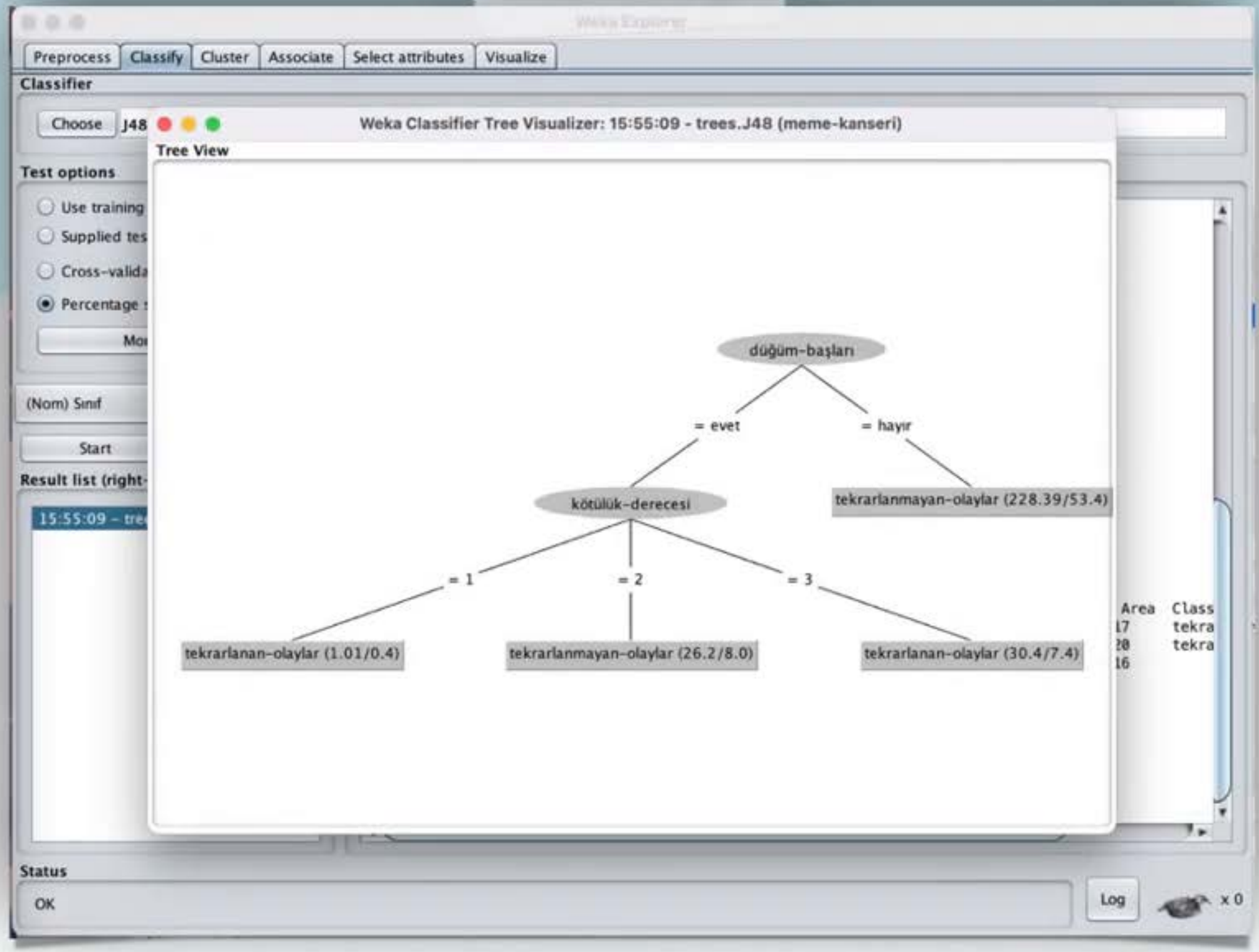
Sonuç ve Model Seçenekleri

Sonuç listesi penceresinin üzerine gelin ve fareyi sağ tıklatın.

Menü, modelle ilgili mevcut seçenekleri görüntüler.



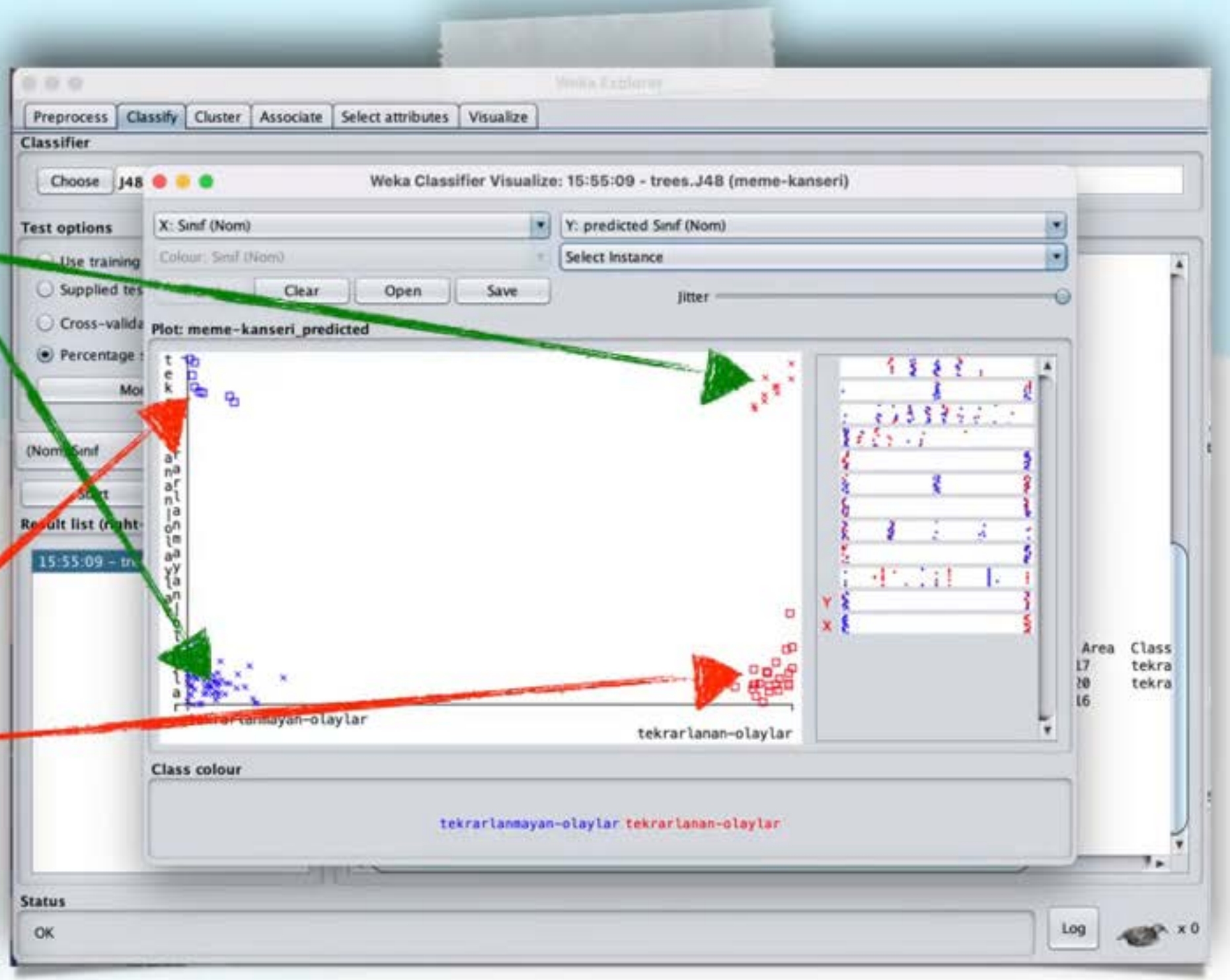
Ağacı Görselleştir'i seçin



Sınıflandırıcı Hatalarını Görüntüle

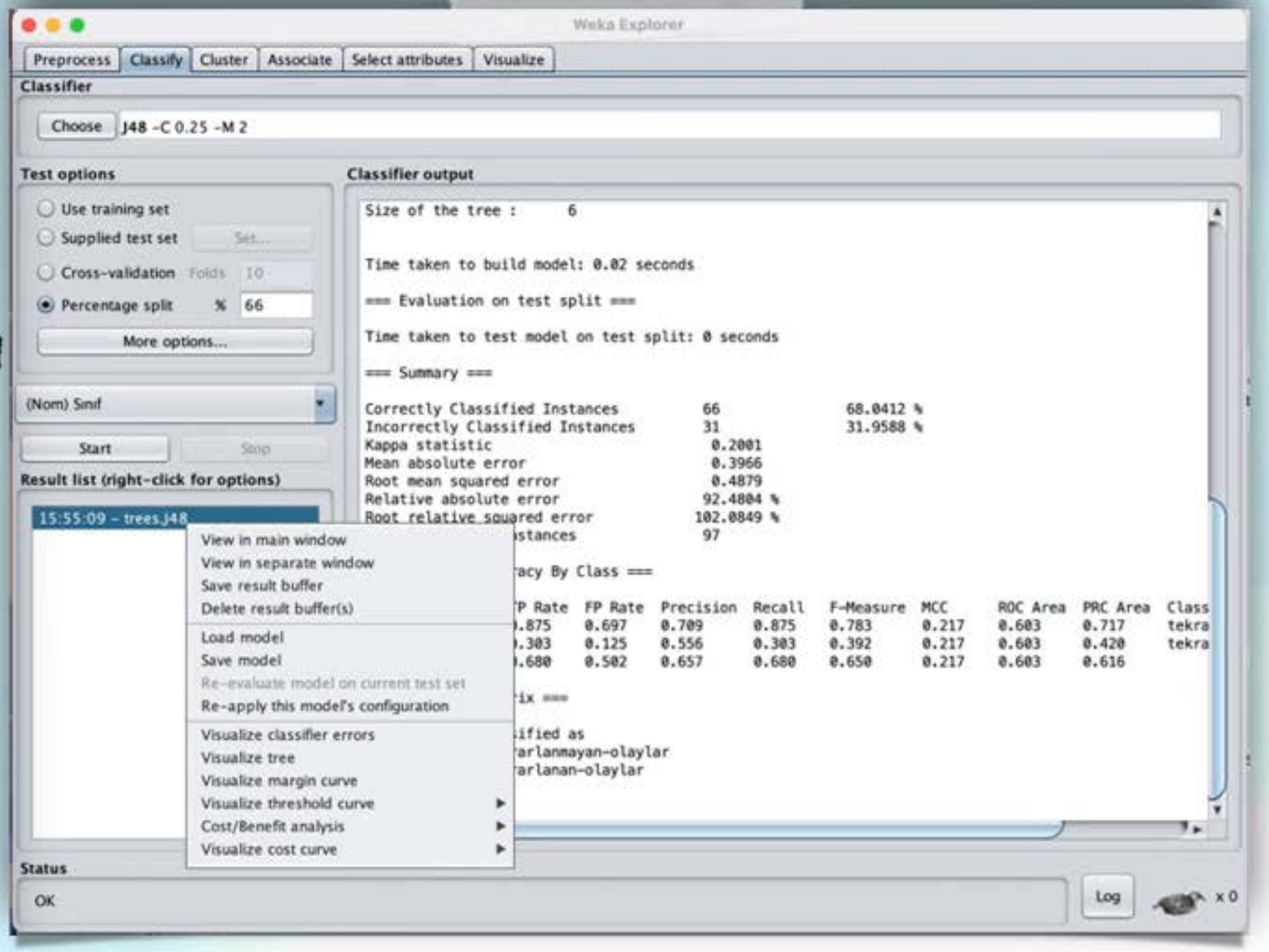
Doğru tahmin edilen vakalar

Yanlış vakalar



Modeli ve Sonuçları Kaydetme

Sağ/seçenek düğmesi sonucu tıklatın. Sınıflandırıcıyı ve sonuçları kaydetmek için **Modeli kaydet** ve **Sonuç arabelleği kaydet**'i seçin,



Özet

Weka, açık kaynak kodlu veri madenciliği yazılımıdır.

- GUI arayüzü:
- **Gezgin, Deneyci, Bilgi Akışı**
- Fonksiyon ve Araçlar
  - Sınıflandırma yöntemleri: karar ağaçları, kural öğrenenler, naif Bayes, vb.
  - Regresyon/tahmin yöntemleri: lineer regresyon, model ağacı üreteçleri, vb.
  - Kümeleme yöntemleri
    - Özellik seçimi yöntemleri
    - Ve Daha Fazlası...