WEKA ile VERİ MADENCİLİĞİ

SINIFLANDIRMA (CLASSİFİCATİON) ÖRNEĞİ

Dr.Günay TEMÜR

Sınıflandırma: Örnek

Sağlıklı sonuçları alabilmek adına Sınıflandırma (CLASSİFİCATİON) işlemi için yaygın kullanılan İRİS veri seti ile çalışılacaktır.

İris Veri Seti

İris Veri Seti Özellikleri

İris adındaki bitkinin 3 farklı türüne ait (***İris Setosa,İris Virginica, Iris Versicolor***) her türden 50 örnek olmak koşuluyla 150 örneğe sahip bir veri setidir.

Her bir çiçeğe ait 4 özellik tanımlanmıştır;

sepal-length (alt-çanak yaprak uzunluğu)

sepal with (alt yaprak genişliği)

pedal-length (üst-taç yaprak uzunluğu)

pedal-with (üst yaprak genişliği)

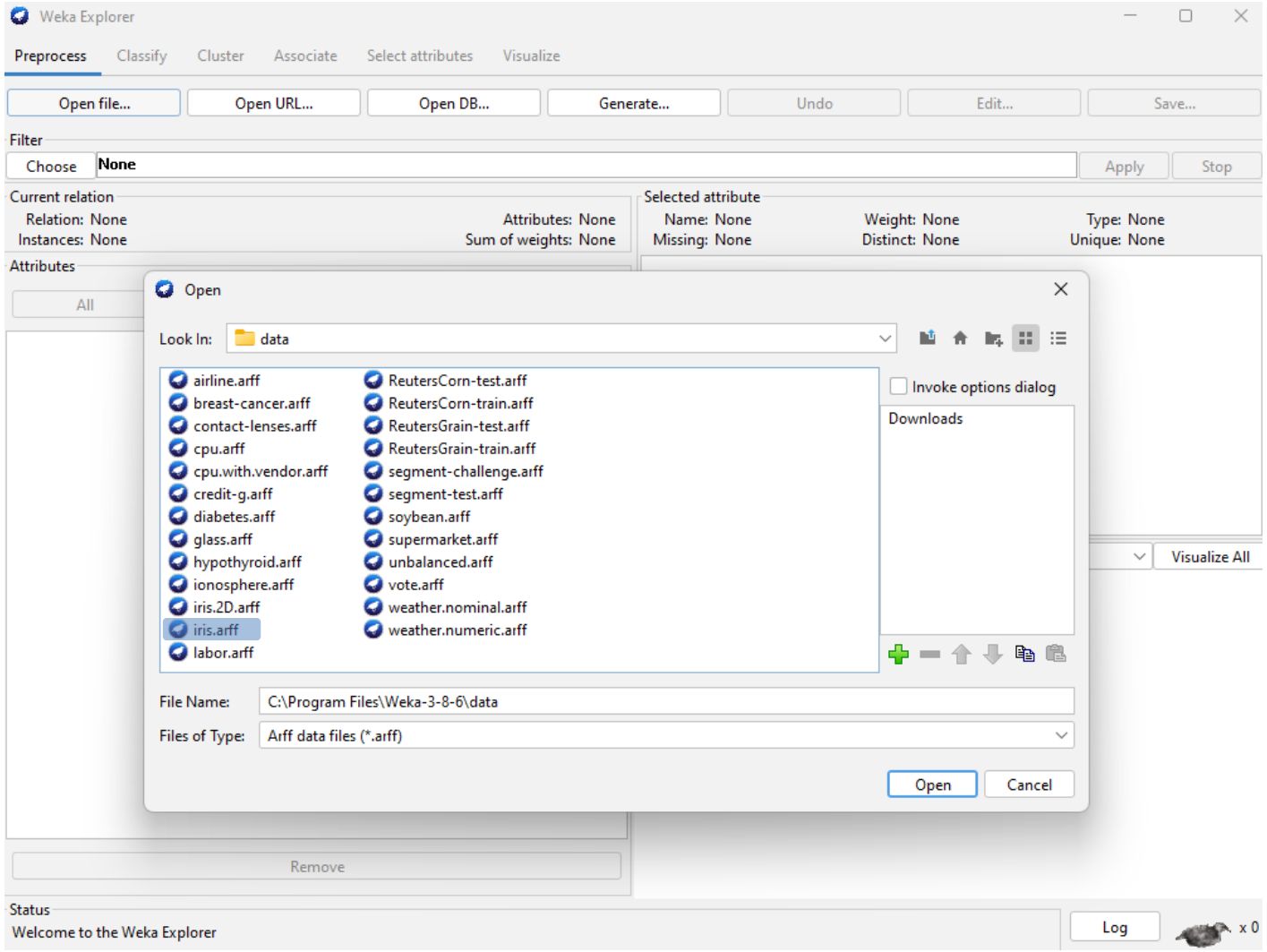
İris Veri Seti Özellikleri

Bu özelliklere bakılarak bitkinin 3 türden hangisine ait olduğu tahmin edilebilmektedir. Biz de bu 4 özelliğini kullanarak İris çiçeğinin hangi türe ait olduğunu makine öğrenmesi yoluyla tahmin etmeye çalışacağız.

Burada her bir türe ait tanımlanan özellikler (alt-çanak yaprak uzunluğu, alt yaprak genişliği, üst-taç yaprak uzunluğu, üst yaprak genişliği) bağımsız değişkenler ,tür isimleri (İris Setosa, İris Virginica, Iris Versicolor) ise bağımlı değişkenlerimiz olacaktır.

Burada her bir türe ait tanımlanan özellikler (alt-çanak yaprak uzunluğu, alt yaprak genişliği, üst-taç yaprak uzunluğu, üst yaprak genişliği) bağımsız değişkenler ,tür isimleri (İris Setosa, İris Virginica, Iris Versicolor) ise bağımlı değişkenlerimiz olacaktır.

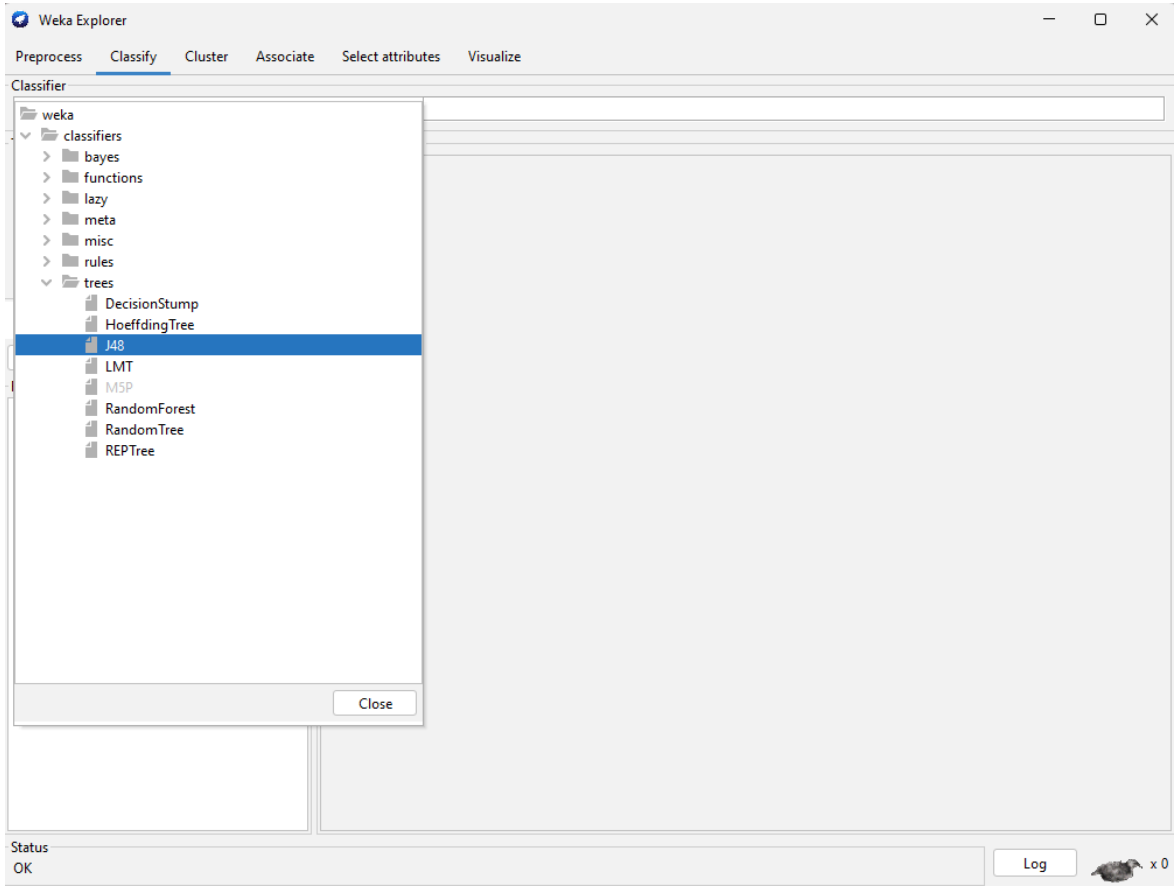
Öncelikle Preprocess- Open File kısmından Weka'nın bize hazır olarak sunduğu kütüphaneyi kullanarak weka-data dosyası içinden iris.arff dosyamızı açıyoruz.(dosya .arff formatında olduğu için dönüştürme işlemi yapmamıza gerek yok)



Tree Algoritması

Daha sonra sınıflandırma işlemi için CLASSİFY bölümüne geçip CHOOSE düğmesine tıkladıktan sonra J48 algoritması seçilir.

J48 algoritması %96.53 duyarlılık oranına sahiptir. Ayrıca doğruluk oranı (%86.36) en fazla olan algoritmalardan biridir. Bu sebeple bir çok uygulamada J48 algoritması tercih edilmektedir.



Test Ayar Seçenekleri

Algoritmayı çalıştırmadan önce test options kısmına biraz bakalım.

Use Training Set: Verilen örneklerinin ne kadar iyi sınıflandırıldığını kontrol eder.

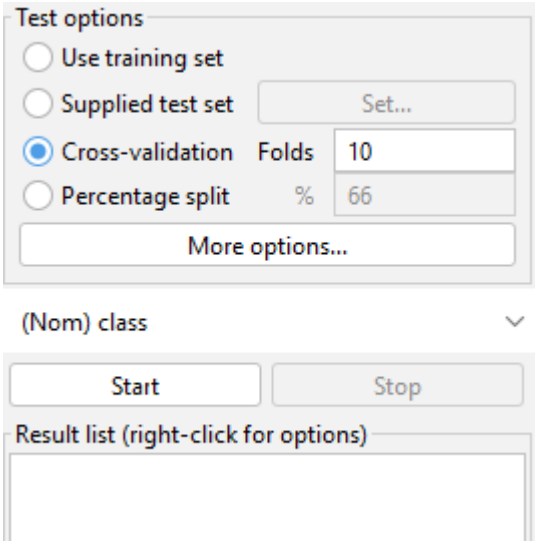
Supplied Test Set : Set kısmından bizim yükleyeceğimiz bir dosya için ne kadar iyi sınıflanrılma yapıldığıını kontrol eder.

Cross-validation : Folds kısmına girilen değere göre Örneğin 10 girilmişse veri setinin %10 ile test, %90 ile de eğitim yapar.

Percentage split: Vereceğimiz yüzdeye göre sınıflandırının ne kadar iyi olduğunu test eder.

Ben Cross-validation seçeneğini işaretleyip Folds değerine 10 vererek devam. edeceğim. Weka veri setimizin %10 luk kısmını test %90 lık kısmını da eğitim için kullansın.

Start diyerek algoritmamızı çalıştıralım.



Değerlendirme Sonuçları

Görüldüğü gibi algoritma başarılı bir şekilde sınıflandırma işlemini gerçekleştirdi.

Classifier Output (Sınıflandırma işlemi sonucunda oluşan çıktılar) kısmını biraz daha yakından inceleyecek olursak;

1 Run Information kısmında veri seti için hangi algoritmayı kullandığımız (J48), veri setimizin isminin ne olduğu (iris), kaç adet veri içerdiği (150), hangi özellikleri içerdiği (sepalllength, sepalwidth, petallength-petalwidth) ve hangi test modelinin (10- fold cross-validation) seçildiğine dair bilgiler yer almaktadır.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    10-fold cross-validation
```

1

Değerlendirme Sonuçları

***2* Classifier Model** kısmında sınıflandırma işlemi için dallanmanın ne şekilde olacağı, ilk dallanmanın nereden başlayacağı hangi aralıklara göre dallanmanın gerçekleşeceği gibi bilgiler yer almaktadır

- *Dallanma **petalwidth (üst yaprak genişliği)** özelliğinden başlamış üst yaprak genişliğinin 0.6 değerine eşit veya ondan küçük olduğu durumda tür **Iris-setosa** kabul edilmiş.*

- ***Üst yaprak genişliği (petalwidth)** 0.6 dan büyük olduğunda 1.7 den büyük ve küçük eşit olma durumuna göre ayrılıp, 1.7 den küçük ve eşit olma durumu için tekrar bir dallanma olmuş, ve bu sefer **üst yaprak uzunluğu(petallength)** özelliğine bakılıp, üst yaprak uzunluğu 4.9 değerinden küçük ve eşitse tür **Iris-veriscolor** kabul edilmiş.*

- ***Üst yaprak uzunluğu (petallength)** 4.9 dan büyük ise **üst yaprak genişliğine(petalwidth)** göre tekrar bir dallanma olmuş üst yaprak genişliği 1.5 değerinden küçük ve ona eşitse tür Iris-virginica,1.5 den büyük olduğu durumlarda ise tür Iris-verticolor kabul edilmiş.*

- ***Pedalwidth (üst yaprak genişliğinin)** 1.7 den büyük olduğu durumlarda ise tür Iris-virginica olarak kabul edilmiştir.*

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5
```

2

Değerlendirme Sonuçları

3 Stratified cross-validation numaralı kısımda seçilen test modelinde ağacın tahmin performans sonuçları görülebilir. Test modeli olarak **Cross-validation** seçildiğinden bu başlığa göre sonuçlandırılmıştır.

Sonuçlara göre 150 adetlik veri seti için **144 tanesi doğru, 6 tanesinin yanlış** şekilde sonuçlandığını ve **%96** başarı elde edildiği görülmektedir. Mutlak hata ortalaması: 0.035 , karesel hata ortalaması: 0.1586, Göreceli mutlak hata: 7.8705' tir.

***4* Detailed Accuracy By Class (sınıflara göre ayrıntılı doğruluk oranları)** kısımda her türe **(Iris-setosa,Iris-versicolor,Iris-virginica)** ait F-measure ,TP Rate (Gerçek Pozitif Değeri), FP Rate (Gerçek Negatif Değeri) ROC, Recall gibi değerleri ayrı ayrı görebiliriz.

```
Size of the tree :      9


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         144               96      %
Kappa statistic                          0.94
Mean absolute error                      0.035
Root mean squared error                  0.1586
Relative absolute error                  7.8705 %
Root relative squared error             33.6353 %
Total Number of Instances              150
```

3

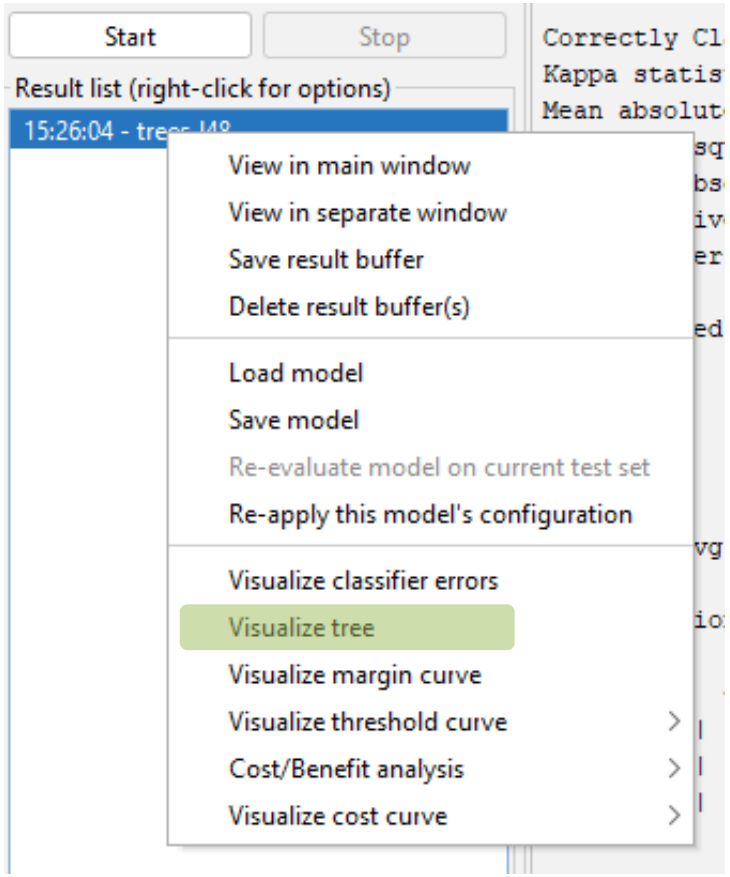
```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0,980     0,000     1,000      0,980    0,990      0,985    0,990     0,987     Iris-setosa
                 0,940     0,030     0,940      0,940    0,940      0,910    0,952     0,880     Iris-versicolor
                 0,960     0,030     0,941      0,960    0,950      0,925    0,961     0,905     Iris-virginica
Weighted Avg.    0,960     0,020     0,960      0,960    0,960      0,940    0,968     0,924
```

4

Karar Ağacı

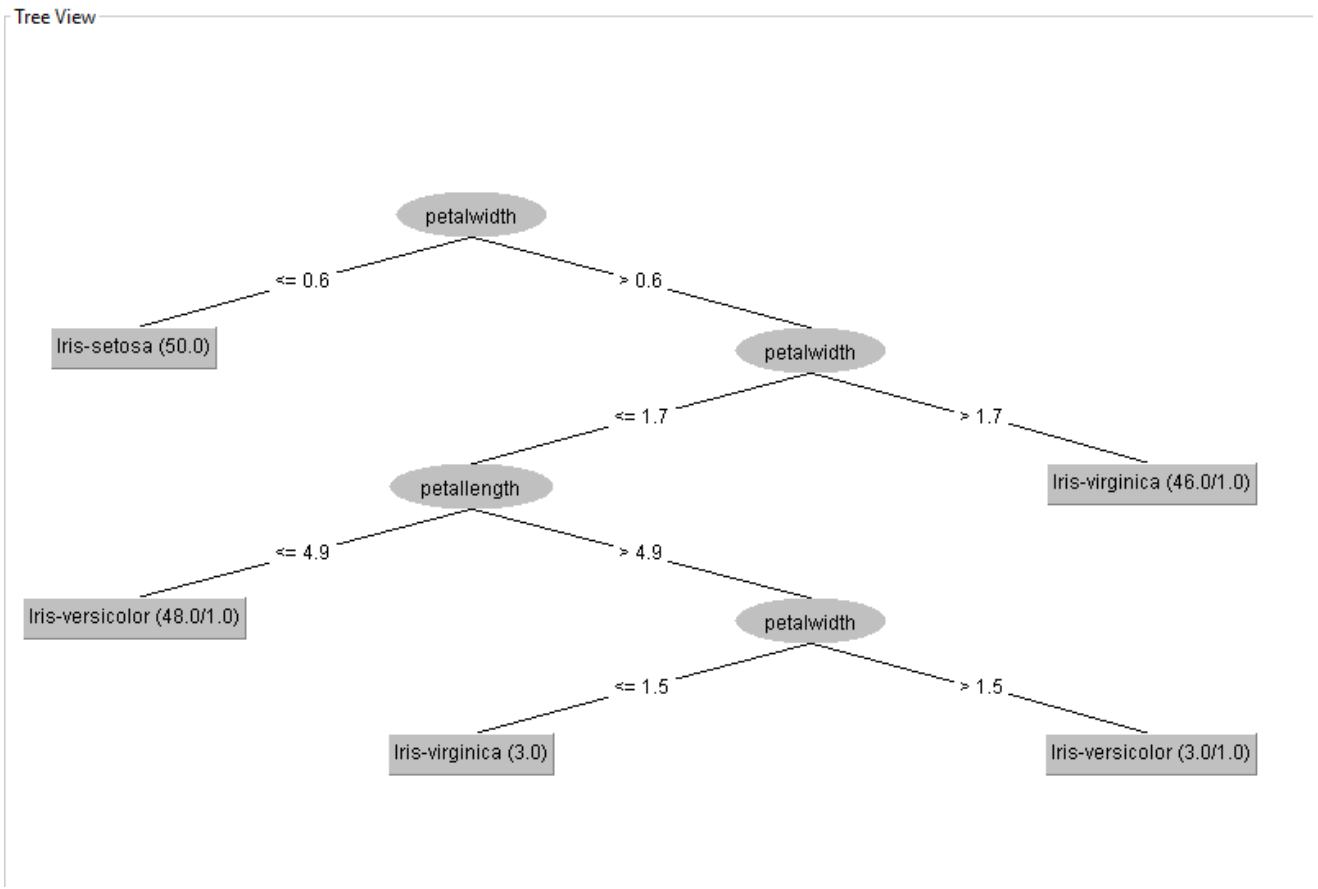
Sınıflandırma işleminden sonra karar ağacımızın sonuçlarını görselleştirmek isteyelim.. Bunun İçin **Result list** kısmında algoritmamıza sağ tıklayarak ***Visualize tree*** seçeneğini seçiyoruz.



Karar Ağacı

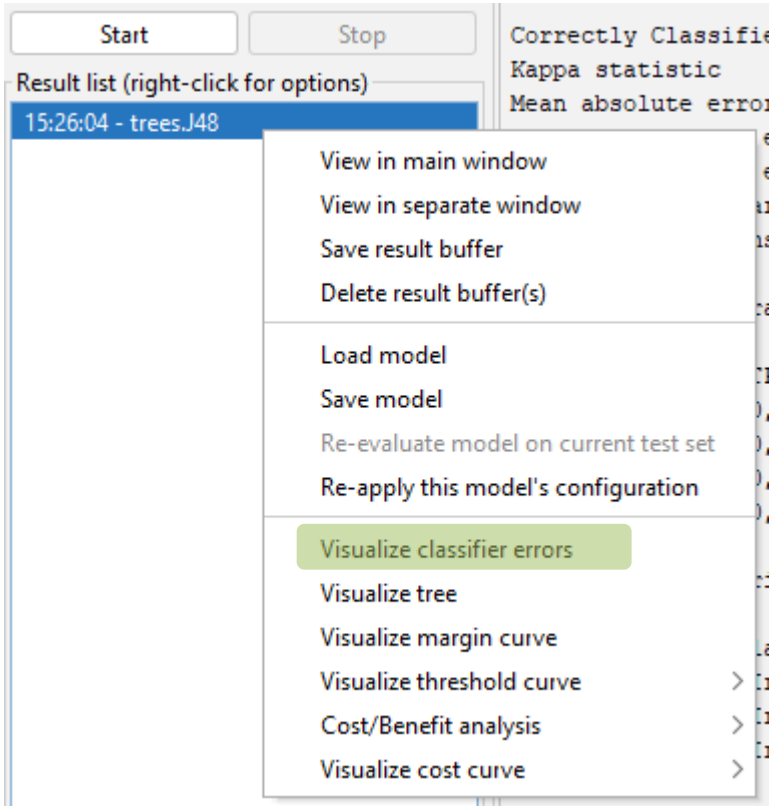
Görüldüğü gibi karar ağacı görselleştirilmiş olarak elde edilmiştir.

Aynı zamanda weka üzerinden seçilen algoritmaya göre oluşan sınıflandırma hatalarını da görebilmek mümkündür.



Sınıflandırma Hataları

Bunun için de aynı şekilde **Result list** kısmında algoritmaya ait listlede sağ tıkladıktan sonra ***visualize classifier errors*** seçeneğiniz seçilir.



Sınıflandırma Detayı

Jiter kısmındaki düğmeyi sağa doğru kaydırarak daha ayrıntılı hale getirilir.

Her sınıfa belirli bir renk atanmış;

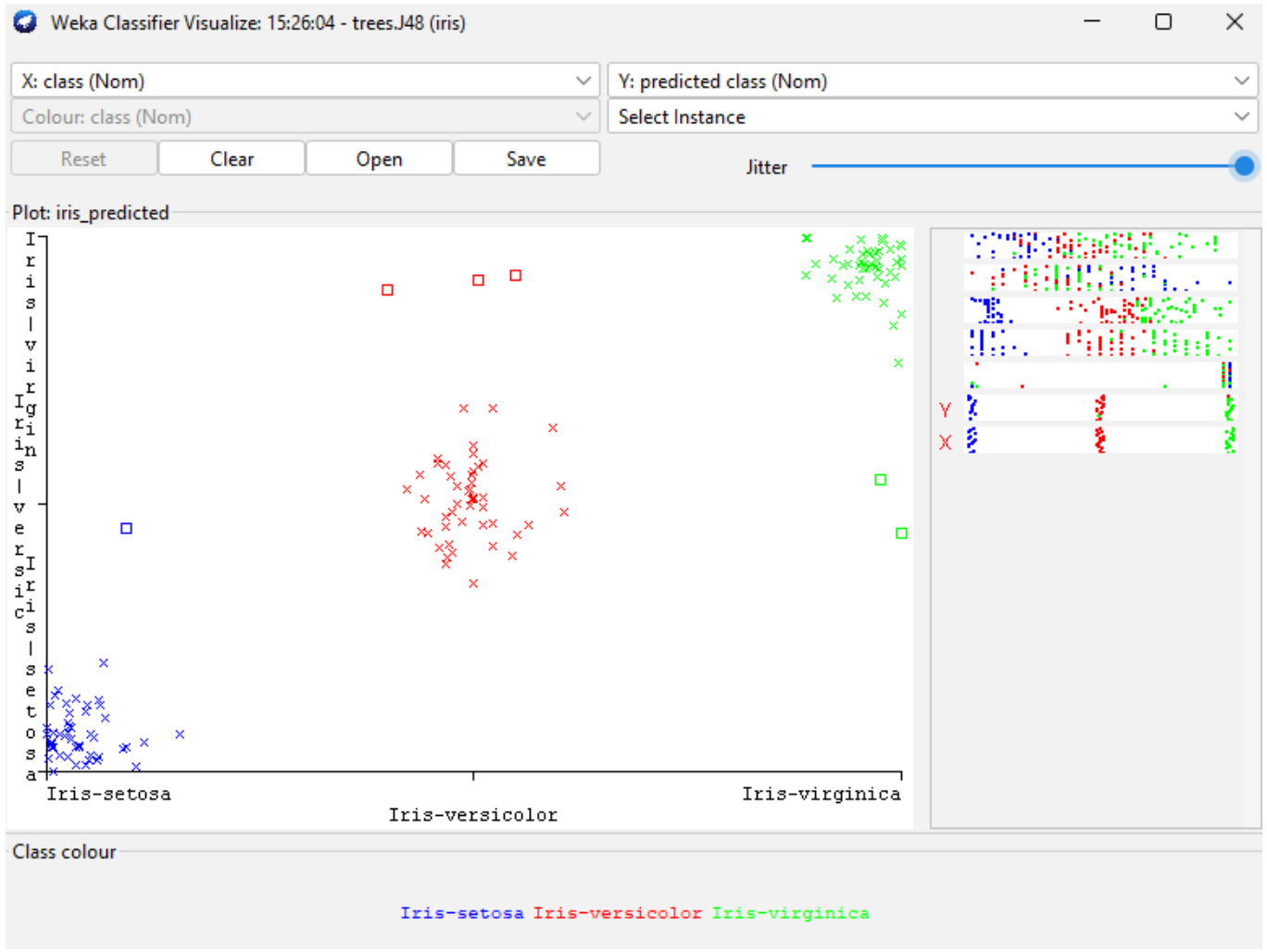
Iris-Setosa : Mavi

Iris-versicolor: Kırmızı

Iris-virginica: Yeşil

Grafikte ise **x** ile gösterilenler doğru sınıflandırılmış verileri □ şeklinde gösterilenler ise yanlış sınıflandırılmış verileri ifade ediyor.

Örneğin; ***Iris-virginica*** kısmında bulunan **yeşil** bir □ simge bu değerin Iris-virginica sınıfında olduğunu ama yanlış şekilde sınıflandırılarak başka bit sınıfa dahil edildiğini gösteriyor.

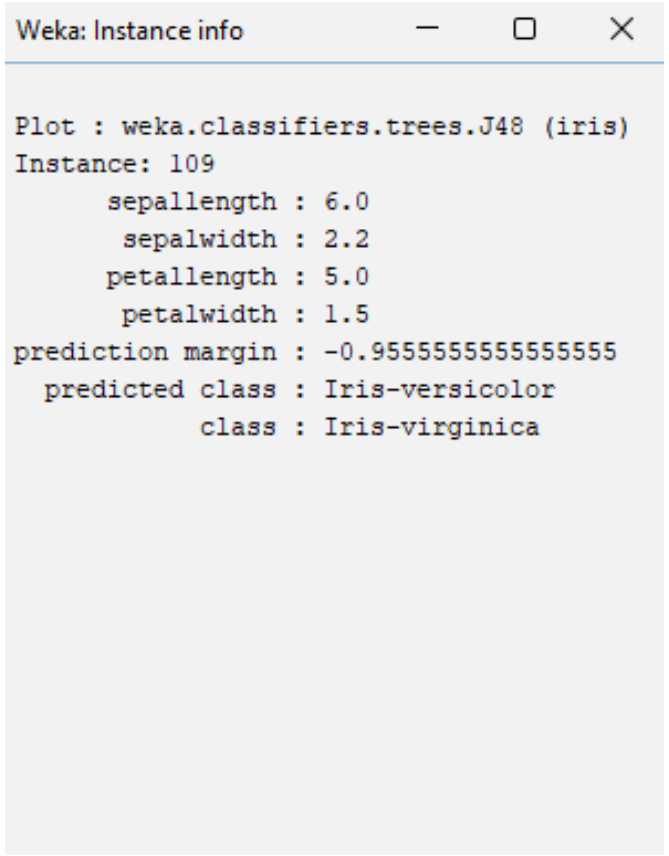


Hatalı Sınıf Detayı

Karelerden bir tanesinin üzerine çift tıklayarak ayrıntıları hakkında bilgi alınabilir. Örneğin Iris-virginica kısmında bulunan yeşil karelerden birine çift tıklatıldığında.

sınıfın (prediction class) Iris-versicolor olduğunu ancak aslında bu değerin (class) Iris-virginica sınıfına ait olduğu görülebilir.

Aynı zamanda değere ait özellik (üst yaprak uzunluğu-genişliği, alt yaprak uzunluğu-genişliği) bilgilerine de ulaşabilir.



BİTTİ ☺

Kaynaklar:
https://bilgisayarkavramlari.com
https://www.veribilimiokulu.com
https://zeynepozturkk.wordpress.com
https://kubracosar.blogspot.com
https://tr.myservername.com

Dr.Günay TEMÜR