Twoing Algoritmasıile Sınıflandırma : Kalp Hastalığı Uygulaması

Kıhan Uysal [1] ı Mehmet Bilen [2] ı Sami Ulukuık [3]

[1] Mehmet Akif Ersoy Üniversitesi, Çavdır Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, Burdur

[2] Mehmet Akif Ersoy Üniversitesi, Çavdır Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, Burdur

[3] Mehmet Akif Ersoy Üniversitesi, Çavdır Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, Burdur

iuysal@mehmetakif.edu.tr, mbilen@mehmetakif.edu.tr, samiulukus@mehmetakif.edu.tr

Özet : Veri madenciliıinin kökleri yapay zekâ, makine öırenimi, matematik, istatistik ve bilgisayar uygulamaları alanlarına dayanmaktadır. Veriden bilgi çıkarımı insanoılunun yüzyıllar boyunca yaptığı bir ıktır. Ancak günümüzde bilgisayar, bilgisayar aıları ve depolama sistemleri teknolojilerindeki gelikmelerle birlikte hayatı büyük miktarlarda veri kuıatmaktadır ve bu verinin büyük bir kısmı ticari, bilimsel, stratejik ve benzeri amaçlarla büyük veritabanları veya veri ambarlarında saklanmaktadır.

Bu çalıkmada tıbbi veriler üzerinde veri madenciliıi uygulamasına bir örnek olarak kalp hastalığı verileri twoing algoritması ile sınıflandırılmıktır. Kalp hastalığı veri setinin sınıflandırma sonucunda karar kurallarıolukturulmuk ve kalp hastalığının belirtileri ortaya konmuktur. Çalıkma sonucunda yak, cinsiyet, maksimum kalp hızı, göıüs aırısı tipi, açlık kan kekeri, talasemi, büyük damarlar, anjine baılı depresyon gibi niteliklerin kalp hastalığı ile ne derece ilgili olduıu belirlenmiktir. Sınıflandırma sonucu olukturulan karar kurallarına göre göıüs aırısı, büyük damarlar ve st eıiminin kalp hastalığı belirtisi olduıu ortaya çıkmıktır.

Anahtar Kelimeler : *Veri Madenciliıi, Saılıkta Veri Madenciliıi, Twoing Algoritması, Karar Aıacı Algoritması, Kalp Hastalığı*

Abstract : The roots of data mining is based on fields of artifical intelligence, machine learning, mathematics, statistic and computer applications. Information extraction from a data is a job which is done for centuries by mankind. But today, because of development in technologies of computers, computer networks and storage systems, large amounts of data surrounds the life. A large portion of these data are stored in large databases or large data warehouses for commercial, scientific, strategic and similar purposes.

In this study, an example of data mining application working with medical datas , heart disease has been classified with twoing algorithm. Decision rules has been generated and Symptoms of heart disease has been displayed according to the results of heart disease data sets classification. As a result of working, age, sex, maximum heart rate, chest pain type, fasting blood glucose, thalassemia, large vessels, such as angina, depending on the nature of depression is associated with heart disease has been determined to what extent. According to the classification results generated decision rule chest pain, major vessels, and heart disease is a symptom of st inclination that has emerged.

Keywords : *DataMining, Health Data Mining, Twoing Algorithm, Decision Tree Algorithm, Heart Disease*

1. Giriş

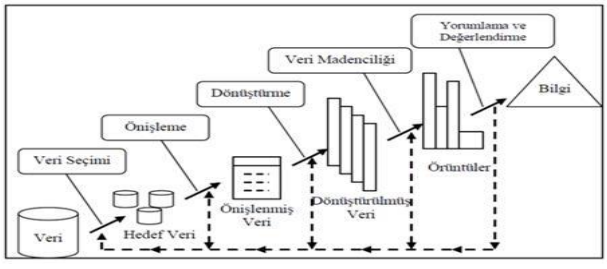
Veri madenciliği kavramını anlayabilmek için ikin en bakında kelimelerin yalın anlamlarından yola çıkılabilir. Madencilik yeryüzünün gizli ve kıymetli kaynaklarının açığa çıkarılması süreci olup, bu kelimenin veri kelimesi ile ilişkilendirilmesi ise veri yığınlarının içerisinde ilk bakışta fark edilemeyen kıymetli bilgilerin bulunması ve çıkartılması fikrini uyandırmaktadır [3].

Veri madenciliğinin kendine özgü karakteristik özellikleri olmalıdır. Bir sistemin veri madenciliği sistemi olabilmesi için büyük miktarlarda veri ile çalışabilmesi, birleşik sorgulara cevap verebilir bir yapıda veri ve bilgi geri alma işlemlerini gerçekleştirebilmesi gerekmektedir [4].

1.1 Veri Madenciliği Süreci :

Veri madenciliği sadece bir yöntem ve teknikler topluluğu olarak değil, probleme özgü tasarlanmış, ilgili yöntem, teknik ve uygulamaları da içine alan, sonuçları itibariyle probleme özgü olmak üzere ilişkileri, kuralları, örüntüleri, eğilimleri, vb. modelleyen ve gösteren bir süreç olarak algılanmalıdır [5].

Literatürde bu sürecin tanımlanması konusunda henüz bir ortak fikir oluşmadığı görülmektedir. Veri madenciliği bazı kaynaklarda veritabanlarında bilgi keşfi kavramı ile birbirinin yerine geçebilir şekilde kullanılmaktadır ve süreç veri madenciliği süreci veya veritabanlarında bilgi keşfi süreci olarak adlandırılmaktadır [10]. Bazı kaynaklarda ise süreç veri madenciliği süreci [1] ve bu işlevi gerçekleştiren bir sistem de veri madenciliği sistemi [7] olarak adlandırılmaktadır. Ancak birtakım kaynaklar veri madenciliğini veritabanlarında bilgi keşfi sürecinde bir aşama olarak konumlandırmaktadırlar [2] ve sürecin bütününü veritabanlarında bilgi keşfi kavramı ifade etmektedir. Birkaç kaynakta ise bu kavram karmaşası içerisinde iki kavram birleştirilerek süreç bilgi keşfi ve veri madenciliği süreci [8] olarak isimlendirilmektedir. Literatürde yaygın olarak kullanılan gösterimlerden biri olan veri madenciliği süreci şekil 1'de sunulmuştur.



Şekil 1 : Veri Madenciliği Sürecini Oluşturan Aşamalar[11]

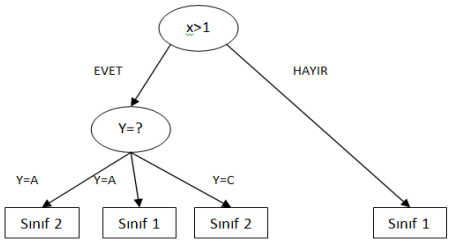
Veri madenciliği sürecinde öncelikle araştırılmak istenilen veri seti içerisinden hedeflenen verinin seçimi gerçekleştirilir. Bu veri seti veritabanı veya veri ambarı olabileceği gibi yalın metin dosyaları içerisinde tutulan veri de olabilir. Seçilen hedef veri üzerinde veri önişleme işlemleri gerçekleştirilerek hatalı veriler, eksik değerler gibi sorunlar halledilir. Veri madenciliği modellemeleri gerçekleştirilmeden önce eğer ihtiyaç varsa veri üzerinde dönüştürme işlemleri yapılır yeni değişkenler elde edilebilir veya mevcut değişkenlerin değerleri yeniden düzenlenebilir. Daha sonra veri madenciliği algoritmaları kullanılarak modeller elde edilir. Elde edilen modeller yorumlanarak bilgiye dönüştürme işlemi gerçekleştirilir.

1.2 Sınıflandırma

Verilerin ortak özellikleri kullanılarak verileri sınıflandırmak mümkündür. Sınıflandırma bir öğrenme algoritmasına dayanır. Tüm veriler kullanılarak eğitme işi yapılmaz. Bu veri topluluğuna ait bir örnek veri üzerinde gerçekleştirilir. Öğrenmenin amacı, bir sınıflandırma modelinin yaratılmasıdır. Diğer bir deyişle sınıflandırma, hangi sınıfa ait olduğu bilinmeyen bir kayıt için bir sınıf belirleme sürecidir [9].

1.2.1 Karar ağaçları ile sınıflandırma :

Verileri sınıflandırma yöntemlerinden karar ağaçları (*decisiontrees*) ile sınıflandırma adını taşımaktadır. Uygulamalı istatistikte makine öğrenmesi bağlığı altında birçok karar ağacı algoritması geliştirilmiştir. Örneklerden oluşan bir küme kullanılarak karar ağacının oluşturulmasını sağlayan çok sayıda öğrenme yöntemi vardır. Karar ağaçları akış şemalarına benzeyen yapılardır. Her bir nitelik bir düğüm tarafından temsil edilir. Dallar ve yapraklar ağaç yapısının elemanlarıdır. En son yapı yaprak, en üst yapı kök ve bunların arasında kalan yapılar ise dal olarak isimlendirilir. Tipik bir karar ağacı örneği şekil 2'de gösterilmektedir. [9]



Şekil 2: Tipik Bir Karar Ağacı[9]

1.2.1.1 Cart ile sınıflandırma

Sınıflandırma ve regresyon ağaçları(*CART*) veri madenciliğinin önemli konularından arasında yer alır. Bu yöntem 1984'te Breiman tarafından ortaya atılmıştır. CART karar ağacı, her bir karar düğümünden itibaren ağacın iki dala ayrılmasına esasına dayanır. Yani bu tür karar ağaçlarında ikili dallanmalar söz konusudur. O halde bir düğümde seçme işlemi yapıldığında, düğümlerden sadece iki dal ayrılabilir.

CART algoritmasında, bir düğümde belirli bir kriter uygulanarak bölümleme işlemi gerçekleştirilir. Bunun için önce tüm niteliklerin var olduğu değerler göz önüne alınır ve tüm eşleşmelerden sonra iki bölünme elde edilir. Bu bölünmelere üzerinde seçme işlemi uygulanır [9].

1.2.1.1.1 Twoing Algoritması

CART yöntemleri arasında yer alan Twoing algoritmasına 2 Adımdan oluşmaktadır.

Adım 1 :

a) Niteliklerin içerdiği değerler göz önüne alınarak eğitim kümesi iki ayrı dala ayrılır. Bunlara *aday bölünme* adı verilir. Bir t düğümünde nsağ ve nsolo olmak üzere iki ayrı dal bulunur. Bu bölümlenen kümeler t_{Sol} ve t_{Sag} biçimindedir.

b) Aday bölünmelerin her biri için P_{Sol}ve$P(j\backslash t_{Sol})$ olasılıklarını hesaplanır. Söz konusu olasılıklar eşitlik 1'de verilmektedir[22]. Burada $P(j\backslash t_{Sol})$ ifadesi bir j sınıf değerinin sol taraftaki bölümlemede olma olasılığını verir. Söz konusu j değerleri sınıf değerlerinin yer aldığı nitelik olarak göz önüne alınır.

$$P_{sol}=\frac{t_{sol}\text{ daki her bir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}}{\text{Eğitim kümesindeki kayıtların sayısı}}$$

$$p(j/_{t_{sol}})=\frac{t_{sol}\text{ daki kayıtların j sınıfları sayısı}}{t_{sol}\text{ daki her bir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}} \quad (1)$$

c) Aday bölünmelerin her biri için $P_{Sağ}$ ve $P(j\backslash t_{Sağ})$ olasılıklarını hesaplanır. Burada $P(j\backslash t_{Sağ})$ ifadesi bir j sınıf değerinin sağ taraftaki bölünme olma olasılığını verir. İşlade eşitlik 2'de görülmektedir[9].

$$P_{sağ}=\frac{t_{sağ}\text{ daki her bir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}}{\text{Eğitim kümesindeki kayıtların sayısı}}$$

$$p(j/_{t_{sağ}})=\frac{t_{sağ}\text{ daki kayıtların j sınıfları sayısı}}{t_{sağ}\text{ daki her bir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}} \quad (2)$$

d)$\bar{u}$ $(s\backslash t)$, t düğümündeki s aday bölümlemelerinin *uygunluk* (goodness) ölçüsü olsun. Söz konusu uygunluk ölçüsünün nasıl hesaplandığı eşitlik 3'de görülmektedir [9].

$$\Phi(s/t)=2P_{Sol}P_{Sağ}\sum_{j=1}^{n}|P(j/t_{Sol})-P(j/t_{Sağ})| \quad (3)$$

e)$\bar{u}$ $(s\backslash t)$ değerleri hesaplandıktan sonra içlerinde *en büyük olanı* seçilir. Bu değerin ilgili olduğu aday bölünme satırı dallanmanın yapılacağı satırı bildirir.

f) Dallanma bu şekilde yapıldıktan sonra, bu adıma ilişkin olarak karar ağacı çizilir.

Adım 2

Algoritmanın en başındaki adıma dönülerek ağacın alt kümesine aynı işlemler uygulanır.

1.2.2 Karar ağaçlarında dallanma kriterleri

Karar ağaçlarında en önemli sorunlardan birisi, herhangi bir kökten itibaren bölümlemenin veya bir başka deyişle dallanmanın hangi kıstasa göre yapılacağıdır. Aslında her farklı kriter için bir karar ağacı algoritması karşımıza gelmektedir.

1.2.3 Karar ağaçlarının budanması

Karar ağaçları çoğu kez karmaşık bir görünüme sahip olabilir. Bir karar ağacında, bir alt ağacı atarak yerine bir yaprak yerleştirmek söz konusu olabilir. Bu şekilde yapılan işleme karar ağacının budanması adı verilmektedir. Alt ağacın yerine yaprak yerleştirmekle, algoritma öngörülü hata oranını azaltmayı ve sınıflandırma modelinin kalitesini artırmayı amaçlar. Ancak hata oranının hesaplanması kolay değildir. Sadece çalışma verileri kümesine dayanan hata oranı uygun sonuçlar üretmez. Öngörülü hata oranını belirlemek için şöyle bir yol izlenebilir: İşlave test örneklerinden oluşan yeni bir küme kullanmak. Bu teknik önceden var olan örnekleri eşit boydaki bloklara böler ve her bir blok için bu bloğu oluşturan tüm örneklerden bir

aşaç oluşturulur, ardından bu aşaç verilmiş örnekler bloşu ile test edilir. Eldeki çalışma ve test örnekleri ile birlikte görünmeyen test örneklerinin, sınıflandırma doşruluşunda katkısı olmayan alt aşaçları çıkarmak budama işleminin temelini oluşturur. Böylece daha az karmaşık bir aşaç üretilmiş olur.

1.2.4 Karar kuralları oluşturmak

Eşitim kümesine başlı olarak elde edilen karar aşacından yararlanarak karar kuralları oluşturulabilir. Karar kurallarının aynen programlama dillerindeki *IF-THEN-ELSE* yapılarına benzer.

1.3 Sağlık Alanında Veri Madenciliği

Sağlık ve tıp, çağımızın en önemli bilimsel araştırma alanlarından olduğu için bu alandaki bilgi sistemleri de araştırmalar için en büyük veri kaynaklarıdır. Son otuz yılda dünyada sağlık bilgi sistemlerinde büyük gelişmeler yaşanmıştır. Sağlık Bilişiminin yeni bir alan olmasına raşmen özellikle bilgi modelleme ve tanı araçlarında hızlı yenilikler yapılmıştır. Sağlık bilgi sistemlerindeki veri madenciliği tekniklerinin ilk kullanımı 1970'lerde ve daha sonraki yıllarda geliştirilen uzman sistemlerle olmuştur. Uzman sistemlerin tıp alanında güçlü araçlar sunmasına raşmen, bu alandaki verilerin hızlı değişmesi ve uzmanlar arasındaki görüş farklılıkları nedeniyle çok yaygınlaşmamıştır. Daha sonraki yıllarda özellikle 1990'lı yıllarda hastaların gelecekteki sağlık durumları ve Sağlık alanında yapılan birçok veri madenciliği araştırmalarında hastaların elektronik tıbbi kayıtları ve idari işleri belgeleyen veriler kullanılmaktadır. Bu verilerden yararlanılarak farklı tahminler yapılabilir. Örneğin bunlardan bazıları şunlardır:

- Belirli bir hastalığa sahip kişilerin ortak özelliklerinin tahmin edilmesi
- Tıbbi tedaviden sonra hastaların durumlarının tahmin edilmesi
- Hastane maliyetlerinin tahmin edilmesi
- Ölüm oranları ve salgın hastalıkların tahmin edilmesi [6].

maliyet tahminleri gibi konuları araştırmak için sinir aşları kullanılmaya başlanmıştır [6].

Günümüzde tıbbın gelişmesi ve insanların ortalama yaşam sürelerinin uzaması beraberinde bazı sorunları da getirmiştir. Örneğin birçok insan, kalp hastalıkları, diyabet ve astım hastalıkları gibi kronik hastalıklarla yaşamak zorundadır. Bu hastalıkların hem tıbbi açıdan hem de hastane kaynak ve maliyetleri açısından ele alınarak doğru yönetilmesi gerekmektedir. Bu noktada bilgi sistemleri üzerinde çalışılabilecek klasik sorgulama yöntemleri yeterli gelmemektedir. Veri madenciliği yöntemleri kullanılarak bu sistemlerdeki gizli ve önemli bilgiler keşfedilmelidir. Keşfedilen bu bilgiler hem tıbbi araştırmalar hem de yönetim planları için değerlendirilmelidir.

1.3.1 Veri Setinin Seçilmesi

Çeşitli veritabanları incelendikten sonra kalp hastalığı verileri için en uygun veri seti Makine Öğrenmesi ve Akıllı Sistemler Merkezi adresinde bulunmuştur [12]. En uygun veri seti aranırken nitelik sayısı, örnek sayısı, daha önce kullanılıp kullanılmadığı gibi bilgiler göz önüne alınmıştır. Seçilen kalp hastalığı veri seti 13 adet nitelik ve 270 satır örnek içermektedir ve güvenilir bir karar kuralı oluşturmak için idealdir. Nitelikler aşağıdaki özellikleri içermektedir.

Özellik Bilgi:

1. Yaş
2. Cinsiyet
3. Göğüs aşrısı tipi (4 değerleri)
4. Istirahat kan basıncı
5. Mg / dl serum kolestrol
6. Açlık kan şekeri> 120 mg / dl
7. Dinlenme elektrokardiyografik sonuçları (değerleri 0,1,2)
8. Maksimum kalp hızı
9. Anjine başlı egzersiz
10. Egzersiz ile oluşturulan st depresyonu
11. Mak. Egzersiz st parçasının eğimini
12. Büyük damarlar (0-3) sayısı
13. Talasemi: 3 = normal; 6 = sabit defekt; 7 = tersinir kusur

Özellikleri türleri :Gerçek: 1,4,5,8,10,12 Sıralı: 11, Ikili: 2,6,9 Nominal: 7,3,13

Tahmin edilecek değişken : 1-Kalp hastalığı yok 2-Kalp hastalığı var

1.4 Tıbbi Veriler Üzerinde Twoing Algoritmasıyla Sınıflandırma

Bir veri seti CART ile sınıflandırılırken iki yöntem kullanılabilir. Birincisi veri setini ikiye bölmek, ikincisi ise sınıf niteliğini ikiye bölmek. Bu uygulamada kalp hastalığını sınıflandırmak için veri setinin ilk 135 satırı eğitim kümesi için ayrılmıştır.

1.4.1 Adım 1

Adım 1.a: Aday Bölünmeler

Twoing algoritmasını uygulamak için, niteliklerin her bir değeri için iki ayrı dizi oluşturulur. Burada **s** aday bölünmenin her bir satırını ifade etmektedir. Örneğin *Yaş<=53* olarak alınırsa bu sol taraf dizisinin elemanı olacaktır. Geriye kalan *Yaş>53* sağ taraf dizi elemanını oluşturur. İki diziden sol tarafta bulunanı t_{Sol}, sağ tarafta yer alanı ise $t_{Sağ}$ dizisi olarak değerlendirilir.

Tablo 1 – Aday Bölünmeler

1	**Yaş<=55**	**Yaş>55**	**14**	**Anjin=0**	**Anjin=1**
2	Cinsiyet=1	Cinsiyet=0	**15**	ST Depresyonu<=2,1	ST Depresyonu>2,1
3	GöğüsAğrısıTipi=1	GöğüsAğrısıTipi ∈{2,3,4}	**16**	ST Eğimi=1	ST Eğimi ∈{2,3}
4	GöğüsAğrısıTipi=2	GöğüsAğrısıTipi {1,3,4}	**17**	ST Eğimi=2	ST Eğimi ∈{1,3}
5	GöğüsAğrısıTipi=3	GöğüsAğrısıTipi ∈{1,2,4}	**18**	ST Eğimi=3	ST Eğimi ∈{1,2}
6	GöğüsAğrısıTipi=4	GöğüsAğrısıTipi {1,2,3}	**19**	Büyük Damarlar=0	Büyük Damarlar {1,2,3}
7	Kan Basıncı<=147	Kan Basıncı>147	**20**	Büyük Damarlar=1	Büyük Damarlar ∈{0,2,3}
8	Kolestrol<=345	Kolestrol>345	**21**	Büyük Damarlar=2	Büyük Damarlar {0,1,3}
9	Açlık Kan Şekeri=0	Açlık Kan Şekeri=1	**22**	Büyük Damarlar=3	Büyük Damarlar ∈{0,1,2}
10	Elektrokardiyografi=0	Elektrokardiyografi {1,2}	**23**	Talasemi=3	Talasemi ∈{6,7}
11	Elektrokardiyografi=1	Elektrokardiyografi ∈{0,2}	**24**	Talasemi=6	Talasemi ∈{3,7}
12	Elektrokardiyografi=2	Elektrokardiyografi {0,1}	**25**	Talasemi=7	Talasemi ∈{3,6}
13	Maksimum Kalp Hızı<=128	Maksimum Kalp Hızı>128			

Adım 1.b: Sol Taraf Olasılıklarının Hesaplanması

Eğitim verisi için toplam 270 örnekten ilk yarısı olan 135 örnek kullanılmıştır.

Yaş<=55 için Hesaplamalar ;

Ek-A'daki her bir nitelik değerinin Tablo 1'de *Yaş* niteliği içindeki tekrar sayılarını belirlemek gerekir. Örneğin *Yaş<=55* değerini Ek-A'da *Yaş* sütununda 64 kez tekrar edildiği görülmektedir. Eğitim kümesinde 135 satır yer almaktadır. Bu durumda *Yaş<=55* elde etme olasılığı olan P_{Sol} değeri şu şekilde hesaplanır:

P_{Sol}= 64/135 = 0,474

şimdi $P(j/t_{Sol})$ değerini hesaplanır. Burada j sınıfları gösterir. Sınıf isimli sınıf niteliğinin 1 ve 2 biçiminde iki değeri vardır. O halde $P(1/t_{Sol})$ ve $P(2/t_{Sol})$ değerlerinin hesaplanması gerekmektedir.

$P(1/t_{Sol})$= 42/64 = 0,656 $P(2/t_{Sol})$= 22/64 = 0,344

elde edilir. Bu hesaplamalar Tablo 1'deki tüm aday bölünmeler için tekrarlanırsa aşağıdaki tablo elde edilir:

Tablo 2 – Sol Taraf Olasılıkları (İlk 5 Satır)

Aday Bölünme	t_{Sol} kayıt sayısı	P_{Sol}	1 sayısı	2 sayısı	$P(1\|t_{Sol})$	$P(2\|t_{Sol})$
1	64	0,474	42	22	0,656	0,344
2	89	0,659	38	51	0,427	0,573
3	9	0,067	7	2	0,778	0,222
4	23	0,17	19	4	0,826	0,174
5	34	0,252	27	7	0,794	0,206

Adım 1.c: Sağ Taraf Olasılıklarının Hesaplanması

Yaş>55 için hesaplamalar ;

Yaş>55 değerlerinin eğitim kümesi içindeki tekrar sayılarını belirlememiz gerekiyor. Ek-A'da Yaş sütununda 71 kez tekrar edildiği görülmektedir. Eğitim kümesinde 135 satır yer almaktadır. Bu durumda Yaş>55 elde etme olasılığı olan $P_{Sağ}$ değeri şu şekilde hesaplanır:

$P_{Sağ}$ = 71/135 = 0,526

Eğitim kümesinde *Yaş>53* değerlerinin yer aldığı satırları göz önüne alınır. Bu satırlardan kaç

tanesinde **1** kaç tanesinde **2** sınıf değerlerinin var olduğunu belirlemek gerekir. Yani $P(1,t_{Sağ})$ ve $P(2,t_{Sağ})$ koşullu olasılık değeri hesaplanır.

$P(1,t_{Sağ})$= 31/71 = 0,437 $P(2,t_{Sağ})$= 40/71 = 0,563

Benzer biçimde diğer satırlarda hesaplamalar yapılırsa tablo 3'ün elde edildiği görülür.

Tablo 3 – Sağ Taraf Olasılıkları (İlk 5 Satır)

Aday Bölünme	$t_{Sağ}$ kayıt sayısı	$P_{Sağ}$	1 sayısı	2 sayısı	$P(1\|t_{Sağ})$	$P(2\|t_{Sağ})$
1	71	0,526	31	40	0,437	0,563
2	46	0,341	35	11	0,761	0,239
3	126	0,933	66	60	0,524	0,476
4	112	0,83	54	58	0,482	0,518
5	101	0,748	46	55	0,455	0,545

Adım 1.d: Uygunluk Ölçütünün Hesaplanması

Uygunluk ölçütü eşitlik 4'daki gibi hesaplanır [22];

$$\Phi(s/t) = 2P_{sağ}P_{sağ}\sum_{j} |P(j/t) - P(j/t_{sağ})| \quad (4)$$

Tablo 2 ve Tablo 3'te elde edilen değerleri burada yerine yazarak her satır için uygunluk ölçütü hesaplanır. Bu t düğümünde Yaş<=53 ve Yaş>53 biçimindeki ilk aday bölümleme için söz konusu hesaplamayı sadece birinci satır için yapılır. Burada s=1 olarak kabul edilir. $\Phi(1\backslash t)=2(0.407)(0.593)[|0.673-0.450|+|0.327-0.550|]=0,108$

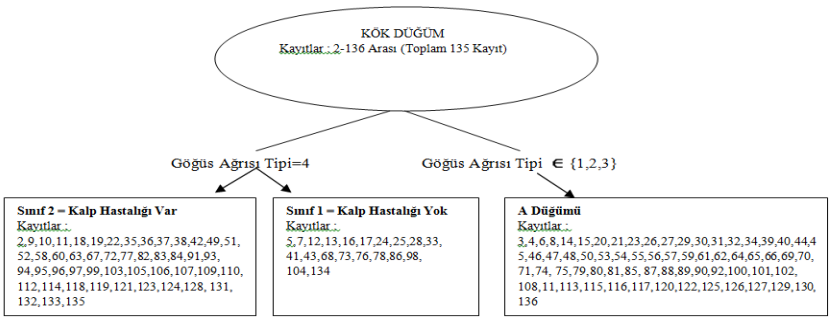
Burada belirtilen hesaplamalar diğer tüm satırlar için benzer biçimde yapılır. Sonuç olarak tablo 4 elde edilir.

Tablo 4 : Her bir aday bölüm için $\Phi(s/t)$ değerleri (İlk 5 Satır)

Aday Bölünme	P_{Sol}	$P_{Sağ}$	$2P_{Sol}P_{Sağ}$	$\Phi(s\backslash t)$
1	0,474	0,526	0,499	0,219
2	0,659	0,341	0,449	0,3
3	0,067	0,933	0,125	0,064
4	0,17	0,83	0,282	0,194
5	0,252	0,748	0,377	0,256

Tablo 4 üzerinde Φ(s/t) sütununda en büyük değer 6. satır üzerinde yer alan 0,513 değeridir. Bu değer en büyük uygunluk ölçütü olarak seçilir. Tablo 1 de 6. satırda birinci bölünmede ***Göğüs Ağrısı=4*** yer aldığına göre, Ek-A'da Göğüs Ağrısı niteliği içinde 4 değerleri aranılır. Bu değerler aşağıda karar ağacında ayrılmıştır. Bu durumda, eğitim serisinde kök düğümden itibaren nasıl bir ayrım yapılacağı belli olmuştur. Bu ayrım şekil 3'te karar ağacında görülmektedir.

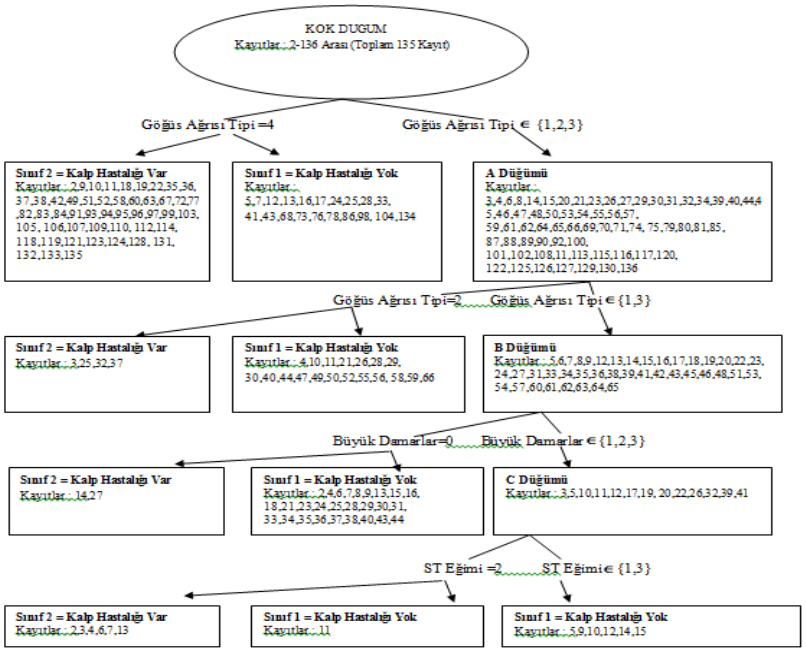
Adım 1.f : Karar Ağacı



Şekil 3**:** Başlangıç bölünmesi ardından elde edilen karar ağacı

4 Adım Sonunda Elde Edilen Karar Ağacı

1.adımdaki karar ağacından sonra sınıfları belli olan kayıtlar tablodan çıkarılarak 2. Adıma geçilmiş ve 1. Adımdaki işlemler tekrar edilerek karar ağacı oluşturulmuş, sınıfları belli olmayan kayıtlar tablodan çıkarılarak 3. Adıma geçilmiştir. Aynı işlemler tekrar edilerek 3. Adım ve 4. Adım uygulanmıştır. 4 adım sonunda sınıflandırma işlemi tamamlanmıştır. Şekil 4'de elde edilen karar ağacına göre sınıfı belli olmayan kayıt kalmadığı için sınıflandırma işlemi bitmiştir. Elde edilen karar ağacına göre karar kuralları oluşturulmuştur.



Şekil 4 - Bakılan üç bölünmesi ardından elde edilen karar ağacı

1.4.5 Kural Tablosu

Elde edilen karar ağacına uygun olarak aşağıdaki kural tablosu düzenlenebilir:

Kural 1 :

Eğer Göğüs Ağrısı Tipi=4 ise Sınıf=2 (Kalp Hastalığı Var) veya Sınıf =1 (Kalp Hastalığı Yok)

Kural 2 :

Eğer Göğüs Ağrısı Tipi =1 veya 2 veya 3 ise ve Eğer Göğüs Ağrısı Tipi=2 ise Sınıf=2 (Kalp Hastalığı Var) veya Sınıf=1 (Kalp Hastalığı Yok)

Kural 3 : Eğer Göğüs Ağrısı Tipi=1 veya 3 ise ve Eğer Büyük Damarlar=0 ise Sınıf=2 (Kalp Hastalığı Var) veya Sınıf=1 (Kalp Hastalığı Yok)

Kural 4 :

Eğer Büyük Damarlar=1 veya 2 veya 3 ise ve Eğer ST Eğimi=2 ise Sınıf=2 (Kalp Hastalığı Var) veya Sınıf=1 (Kalp Hastalığı Yok)

Kural 5 :

Eğer Büyük Damarlar=1 veya 2 veya 3 ise ve Eğer ST Eğimi=1 veya 3 ise Sınıf=1 (Kalp Hastalığı Yok)

2. Sonuç ve Öneriler

Günümüzde en çok bilgi ihtiyacı olan araştırma alanları sağlık ve tıptır. Veri Madenciliği, sağlık ve tıp alanındaki büyük veritabanlarından değerli bilgileri ortaya çıkartarak, hem tıp açısından hem de hizmet kalitesinin artırılması açısından büyük katkılar sağlar. Günümüzde uluslararası ortak projeler kapsamında geliştirilen ve veri setlerinin saklandığı veritabanları, bu veritabanlarına erişim ve veri madenciliği sistemleri de klinik araştırmaların önemli bir parçası haline gelmişlerdir.

Bu çalışmada kalp hastalığı veri seti kullanılarak kalp hastalığı twoing algoritması ile sınıflandırılmış ve karar kuralları oluşturulmuştur. Yaş, cinsiyet, göğüs ağrıları, kalp atış hızı, kan şekeri, talasemi gibi niteliklerin kalp hastalığının belirtisi olup olmadığı araştırılmıştır. 135 satırdan oluşan örnekte oluşturulan karar kurallarına göre belirtilen niteliklerin kalp hastalığı belirtisi olup olmadığı sınıflandırılmıştır. Sınıflandırmaya göre göğüs ağrısı, büyük damarlar ve st eğimi gibi niteliklerin tiplerine göre kalp hastalığı belirtisi olduğu sonucu ortaya çıkmıştır. Büyük damarlar özelliği en büyük bilgi kazancına sahip olduğu için ağacın kökünde yer alacak özellik olarak seçilir. Bu özellik en ayırt edici özellik olarak bulunmuştur. Bundan sonraki yapılacak çalışmalarda, bu çalışma sonucu elde edilen bilgilere göre bir adım ileriye gidilebilir ve daha net belirtiler elde edilebilir. Kalp hastalığı verileri twoing algoritması ile sınıflandırılmıştır. Bundan sonraki yapılacak çalışmalarda aynı veri seti kullanılarak gini algoritması ile sınıflandırılıp karşılaştırma yapılabilir.

3. Kaynaklar

[1] Delavari, N.,Beikzadeh, M. R. ve Phon-Amnuaisuk, S., ñApplication of Enhanced Analysis Model for Data MiningProcesses in HigherEducationalSystemsò, IEEE ITHET 6th Annual International Conference, JuanDolio, DominicanRepublic, F4B/1-6, 2005.

[2] Dunham, M. H., Data Mining: Introductoryand Advanced Topics, Prentice-Hall, UpperSaddleRiver, NJ, USA, 2003.

[3] Gëudëcë, P.,Applied Data Mining: Statistical Methodsfor Business andIndustry, John Wiley&Sons, West Sussex, England, 2003.

[4] Han, J., Kamber, M., ñData Mining:ConceptsandTechniquesò, Morgan Kaufmann, 2006.

[5] Kantardzic, M., ñData Mining: Concepts, MethodsandAlgorithmsò, Wiley, 2003.

[6] Kudyba, S., ñManaging Data Miningò, CyberTech Publishing, 2004, 146-163

[7] Li, Q. ve Khosla, R., ñPerformanceOptimization of Data Mining Applications Using a Multi-layered Multi-agent Data Mining Architectureò, CIMSA 2005 ï IEEE International Conference on ComputationalIntelligenceforMeasurementSystemsand Applications, GiardiniNaxos, Italy, 227-231, July, 2005.

[8] Microsoft(2006), Microsoft Research, Data Mining: Efficient Data Exploration andModeling, http://research.microsoft.com/dmx/DataMining, (12.05.2006).

[9] Özkan, Y., "Veri Madenciliği Yöntemleri", Papatya Yayëncëlëk Eğitim, 2013.

[10] Tsai, C. Y. ve Tsaë, M. H., ñA Dynamic Web Service based Data MiningProcessSystemò, Proceedings of TheFifth International Conference on Computerand Information Technology (CITô05), Washington, DC, USA, IEEE ComputerSociety, 1033-1039, 2005.

[11] U. Fayyad, G. Piatetsky-Shapiro, P. Symth, P. ñFrom Data Miningto Knowledge Discovery in Databasesò, AI Magazine, 17(3), 37-54, 1996.

[12] http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29- Makine Öğrenmesi ve Akëllë Sistemler Merkezi

4. Ekler

Ek-A : Uygulamada kullanëlacak eğitim tablosu-Adëm 1 için (Çok 5 satër)

Sëra	Yaş	Cinsiyet	Göğüs ağrësëtipi	Hareketsiz kan basëncë	Serum kolestrol	Açlëk kan şekeri	Elektrokardiyografi	Maksimum kalp hëzë	Anjine başlëegzersiz	ST depresyonu	ST eğimi	Büyük damarlar	Talasemi	Sënëf
1	70	1	4	130	322	0	2	109	0	2,4	2	3	3	2
2	67	0	3	115	564	0	2	160	0	1,6	2	0	7	1
3	57	1	2	124	261	0	0	141	0	0,3	1	0	7	2
4	64	1	4	128	263	0	0	105	1	0,2	2	1	7	1
5	74	0	2	120	269	0	2	121	1	0,2	1	1	3	1

Kaynak : http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29

Ek-B : 3. Adëmdan Sonra Elde Edilen Yeni Eğitim Kümesi ï Adëm 4 için (Çok 5 Satër)

Sëra	Yaş	Cinsiyet	Göğüs Ağrësë Tipi	Kan Basëncë	Kolestrol	Açlëk Kan şekeri	Elektrokardiyografi	Maksimum kalp hëzë	Anjin	ST Depresyonu	ST Eğimi	Büyük Damarlar	Talasemi	Sënëf
1	56	1	3	130	256	1	2	142	1	0,6	2	1	6	2
2	61	1	1	134	234	0	0	145	0	2,6	2	2	3	2
3	58	1	3	112	230	0	2	165	0	2,5	2	1	7	2
4	71	0	3	110	265	1	2	130	0	0	1	1	3	1
5	57	1	3	128	229	0	2	150	0	0,4	2	1	7	2