VERİ MADENCİLİĞİNDE CART VE LOJİSTİK REGRESYON ANALİZİNİN YERİ: İLAÇ PROVİZYON SİSTEMİ VERİLERİ ÜZERİNDE ÖRNEK BİR UYGULAMA

CART AND LOGISTIC REGRESSION ANALYSIS IN DATA MINING: AN APPLICATION ON PHARMACY PROVISION SYSTEM DATA

Zeynep Burcu GÜNER*

ÖZET

Bilimsel çalışmalarda kullanılan veri setleri zaman zaman karmaşık bir yapı teşkil etmektedir. Bu noktada veri madenciliği, büyük veri tabanlarından faydalı bilgileri ortaya çıkararak hizmet kalitesinin artırılması bakımından büyük katkılar sağlamaktadır. Genellikle araştırmalarda büyük veri kümelerini sınıflandırarak önemli veri sınıflarını ortaya koyan veya gelecek veri eğilimlerini tahmin etmede faydalanılan yöntemlerden, veri madenciliği teknikleri içerisinde en yaygın kullanıma sahip olanlarından bir tanesi sınıflama ve regresyon modelleridir. Bu çalışmada veri madenciliği metotları içerisinde, sınıflama ve regresyon modellerinden en çok kullanılan karar ağacı algoritmalarından biri olan sınıflama ve regresyon ağaçları (CART) algoritması ile lojistik regresyonun sınıflama özellikleri karşılaştırılarak gerçek bir veri seti üzerinde uygulama yapılmış ve söz konusu iki yöntemin başarısını göstermek amaçlanmıştır. Bu sayede mevcut veriler ile yapılan analiz sonuçlarına göre; aynı özellikte verilerle yapılacak ileriki çalışmalarda genel geçer kurallar tanımlanmasında, söz konusu analizleri kullanmanın uygun olacağı gösterilmek istenmiştir. Bu kapsamda, penisilin grubu antibiyotik kullanan hastaların profilini belirlemek amacıyla bir uygulama yapılmış ve çalışmaya alınan veri seti için CART analizinin lojistik regresyon analizine göre daha iyi bir doğru sınıflandırma oranına sahip olduğu görülmüştür

Anahtar Kelimeler: Veri madenciliği, CART, Lojistik Regresyon

* Sosyal Güvenlik Uzmanı, Sosyal Güvenlik Kurumu, Aktüerya ve Fon Yönetimi Daire Başkanlığı, e-posta: zkiran@sgk.gov.tr, tel: 0312 207 87 06.

ABSTRACT

The data sets used in scientific studies pose a very complex structure from time to time. At this point, data mining is making a big contribution in terms of improving the quality of services by revealing useful information from large databases. Generally on studies, to predict future data trends utilization of the methods, data mining techniques in one of the most widely used are classification and regression models. In this study, among data mining methods, classification and regression models most commonly used ones are decision tree algorithms. By comparing Classification and Regression Trees (CART) algorithm which belongs to decision trees and logistic regression shows classification characteristics on real data set and success rates of these two methods. In this context, taken by the Social Security Administration pharmacy provision system, from the respiratory disease which is one of 11 diagnoses for 6,772,313 entries in the prescribed antibiotics in the penicillin group was used to analyze that profiling the patients and the analysis found the CART analysis has better classification success than logistic regression analysis.

Key Words: Data mining, CART, Logistic regression

1. Giriş

Bilişim teknolojilerinde yaşanan hızlı gelişmeler ve bilgisayarların bilgi saklama kapasitelerinin artmasıyla birlikte depolanan veriler çok daha büyük boyutlara ulaşmakta, artan bilgi miktarı karşısında bilgi kaydı yapılan alanların sayısı da giderek artmaktadır. Dünyadaki bilgi miktarının her 20 ayda bir ikiye katlandığı tahmin edilmektedir [Köktürk, F., 2009, 20-25]. Veri tabanı sistemlerinin artan kullanımı ve sakladıkları veri miktarlarındaki böylesine büyük artış organizasyonları elde toplanan bu verilerden nasıl faydalanılabileceği problemi ile karşı karşıya bırakmıştır [Ahmad, I., 2000, 194–203]. Bilgisayar sistemleri ile üretilen bu veriler kendi başına değersizdir, çünkü tek başlarına herhangi bir anlam ifade etmemektedir. Veriler belirli bir amaca yönelik olarak işlenerek bilgiye dönüştürüldüğünde bir anlam ifade etmeye başlamaktadır. Bu nedenle çok büyük veri yığınlarını bilgiye dönüştürerek anlamlı hale dönüştüren teknikler son yıllarda büyük önem kazanmıştır.

1990'lı yılların başından itibaren kullanılmaya başlanan, büyük veri kümeleri içinde saklı durumda bulunan ve işlenmemiş bilgiyi anlaşılabilir ve yorumlanabilir hale getiren işlemlerden biri veri madenciliğidir. Maliyetli ve zahmetli bir süreç olan veri toplama yatırımlarından en yüksek faydayı sağlamak veri madenciliği ile mümkündür [Kecman, V., 2001, 1-4]. Veri madenciliğini bir veri kümesi içerisinde keşfedilmemiş örüntüleri bulmayı hedefleyen teknikler bütünü olarak ifade etmek mümkündür. Veri madenciliği, verilerden daha önceden bilinmeyen ve muhtemelen faydalı enformasyonun monoton olmayan bir süreçte çıkartılması işlemi olarak tanımlanabilir [Fayyad, U., 1996, 27-34].

Veri madenciliği aslında disiplinler arası bir çalışma alanı olup, sınıflama problemleri ve örüntü tanıma (pattern recognition) üzerinde yoğunlaşan yapay zekâ ve amacı yığın hakkında anlamlı bilgi elde etmek ve yorumlamak olan istatistik bilimindeki gelişmeler veri madenciliğinin temellerini oluşturmaktadır. Benzer şekilde veri madenciliği; disiplinler arası doğasından dolayı veri tabanları, makine öğrenmesi, bilgi toplama, görselleştirme, paralel ve dağıtık hesaplama ve optimizasyon gibi birçok disiplinden de etkilenmektedir [Zhou, Z., 2003, 139-146]. Veri madenciliği alanında başta makine öğrenme ve istatistik olmak üzere ayrı araştırma alanlarından türetilmiş birçok türde algoritma vardır. Bu algoritmalar veriden bilgiyi elde etmede kullanılmaktadır.

Operasyonel kararların ötesinde, stratejik ve politik karar verme süreçlerinde önemli bir yere sahip olan veri madenciliği günümüzde gerek kamuda gerekse özel sektörde karar verme sürecine ihtiyaç duyulan birçok alanda kullanılmaktadır. İstatistik ile olan yakın ilişkisi, veri madenciliğini özellikle tıp ve ekonomi gibi bilim dalları başta olmak üzere pek çok bilim dalı için de önemli kılmaktadır. Bilginin bu denli değerli olduğu çağımızda bilgiye ulaşmak için kat edilen yolda veri madenciliği oldukça önemli bir safhadır. Veri madenciliği astronomi, biyoloji, bankacılık, finans, pazarlama, sigorta, tıp ve birçok başka alanda başarılı bir şekilde kullanılmaktadır.

Bilimsel çalışmalarda kullanılan verilerin analizinde istatistiksel yöntemlerden diskriminant, kümeleme ve lojistik regresyon analizi gibi sınıflama ve regresyon modelleri sıklıkla kullanılmaktadır. Modellerde

kullanılan karmaşık verilerin sınıflandırılması, her ne kadar çok değişkenli istatistiksel analizlerin önemli bir bölümünü oluştursa da sağlık başta olmak üzere çeşitli bilim dallarında çok geniş bir kullanım alanına sahiptir. Genellikle araştırmalarda büyük veri kümelerini sınıflandırarak önemli veri sınıflarını ortaya koyan veya gelecek veri eğilimlerini tahmin etmede faydalanılan yöntemlerden veri madenciliği teknikleri içerisinde en yaygın kullanıma sahip olanlarından bir tanesi de sınıflama ve regresyon modelleridir. Bu modeller içerisinden ise sıklıkla tercih edilen yöntemler lojistik regresyon, karar ağaçları ve yapay sinir ağları gibi teknikleridir.

İstatistiksel uygulamalarda sınıflama ve regresyon yöntemleri, bağımlı ve bağımsız değişken arasındaki ilişkiyi tanımlamaya yönelik veri analizlerinin önemli bir parçası olmaya başlamıştır. Uygulamada genellikle modelleme örneklerinin en yaygın olanları bağımlı değişkeninin sürekli olduğu doğrusal regresyon modelleri olsa da, son yıllarda bağımlı değişkenin kategorik olması halinde normallik varsayımının bozulması ve tipik doğrusal modelin uygulanamadığı durumlarda lojistik regresyon modelinin kullanımı standart bir yöntem haline gelmiştir [Hosmer, D. W., Lemeshow, S., 1989, 5-50]. Lojistik regresyon ile en az değişkenin kullanılmasıyla en iyi uyuma sahip olacak biçimde bağımlı ve bağımsız değişkenler arasındaki ilişkiyi tanımlayabilen ve istatistiksel olarak kabul edilebilir bir model kurmak amaçlanmaktadır.

Bağımsız değişkenler için herhangi bir varsayım olmaksızın kategorik bağımlı değişkeni tahmin etmek için sadece lojistik regresyon değil aynı zamanda karar ağaçları da kullanılmaktadır [Kim, M., 2009, 6727–6734]. Çeşitli şekillerde elde edilmiş veriyi analiz ederek anlaşılır ve faydalı bir yapıya dönüştürmeyi hedefleyen veri madenciliği metotlarından biri olan karar ağaçları; kolay anlaşılır olması, görsel sunumunun ön planda olması gibi nedenlerle sıklıkla tercih edilmektedir.

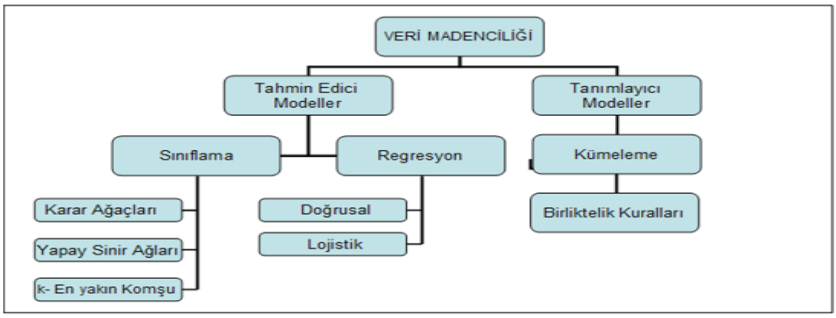
Bu çalışmada, giriş bölümünü takip eden ikinci bölümde sınıflama ve regresyon modellerinden karar ağaçları ve karar ağaçlarında en çok kullanılan analizlerin yapısı ile çalışmanın diğer konusu olan lojistik regresyon modeline değinilmiştir. Uygulamanın yapıldığı üçüncü

bölümde, lojistik regresyon ile karar ağacı algoritmalarından en çok kullanılan Classification and Regression Trees (CART) algoritmasının, ilaç provizyon sisteminden alınan solunum sistemi hastalıkları için yazılan antibiyotik veri seti üzerinde, penisilin grubu antibiyotiklerin analizi yapılmış ve çalışma yapılan analizlerin karşılaştırılmasının ve açıklanmasının yer verildiği dördüncü bölüm olan sonuç ve öneriler ile sona erdirilmiştir.

2. Veri Madenciliğinde Sınıflama ve Regresyon Modelleri

Veri Madenciliğinde kullanılan modeller, temel olarak Şekil 1'de görüldüğü üzere tahmin edici (predictive) ve tanımlayıcı (descriptive) olmak üzere iki ana başlık altında incelenmektedir [Bigus, J. P., 1996].

Şekil 1. Veri madenciliği modelleri



Çalışmada da kullanılan karar ağaçları ve lojistik regresyon gibi tahmin edici modellerin amacı, verilerden hareket ederek bir model geliştirmek ve kurulan bu model yardımıyla sonuçları bilinmeyen veri kümelerinin sonuç değerlerini tahmin etmektir. Tanımlayıcı modellerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır.

Sınıflama en çok bilinen veri madenciliği tekniklerinden birisidir; resim, örüntü tanıma, hastalık tanıları, dolandırıcılık tespiti, kalite kontrol çalışmaları ve pazarlama konuları sınıflama tekniklerinin sıklıkla kullanıldığı alanlardır. Sınıflama tahmin edici bir model olup, havanın bir

sonraki gün nasıl olacağı veya bir kutuda kaç tane mavi top olduğunun tahmin edilmesi bir sınıflama işlemidir [Silahtaroğlu, G., (2008 33, 45-47, 58].

Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler

- Karar ağaçları (Decision Trees),
- Lojistik regresyon (Logistic Regression),
- Yapay sinir ağları (Artificial Neural Networks),
- Genetik algoritmalar (Genetic Algorithms),
- K-en yakın komşu (K-Nearest Neighbor),
- Bellek temelli nedenleme (Memory Based Reasoning),
- Naïve-Bayes,
- Bulanık Küme Yaklaşımı (Fuzzy Set Approach) 'dır.

Çalışmanın kapsamında yukarıda sayılan söz konusu tekniklerden sadece karar ağaçları ve lojistik regresyon üzerinde durulacaktır.

2.1 Veri Madenciliğinde Karar Ağaçları

Sınıflama ve regresyon modellerinin bir yöntemi olan karar ağaçları, kurulmasının ucuz olması, yorumlanmalarının kolay olması, veri tabanı sistemleri ile kolayca entegre edilebilmeleri ve güvenilirliklerinin iyi olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahip olup ağaç yapısı ile kolay anlaşılabilen kurallar yaratabilen, bilgi teknolojileri işlemleri ile kolay entegre olabilen en popüler sınıflama tekniğidir [Ayık Y. Z., Özdemir A., Yavuz U., 2007, 441-454]. Karar ağaçları, basit karar verme adımları uygulanarak, çok sayıda kayıt içeren bir veri kümesini çok küçük kayıt gruplarına bölmek için kullanılan bir yapıdır [Berry, M. J., Linoff, G. S., 2004]. Her başarılı bölme işlemiyle, sonuç gruplarının üyeleri bir diğeriyle çok daha benzer hale gelmektedir.

Bu teknikte sınıflandırma için bir ağaç oluşturulur, daha sonra veri tabanındaki her bir kayıt bu ağaca uygulanır ve çıkan sonuca göre de bu kayıt sınıflandırılır. Karar ağaçları veri setinin çok karmaşık olduğu durumlarda bile, bağımlı değişkeni etkileyen değişkenleri ve bu

değişkenlerin modeldeki önemini basit bir ağaç yapısı ile görsel olarak sunabilmektedir.

Karar ağacı yöntemini kullanarak verinin sınıflanması temel olarak iki adımdan oluşmaktadır. Birinci adım; önceden bilinen bir eğitim verisinin model oluşturmak amacı ile sınıflama algoritması tarafından çözümlendiği öğrenme basamağıdır. Öğrenilen model, sınıflama kuralları veya karar ağacı olarak gösterilir. İkinci adım ise eğitim verisinin sınıflama kurallarının veya karar ağacının doğruluğunu belirlemek amacıyla test edilerek kullanıldığı sınıflamadır. Eğer doğruluk kabul edilebilir oranda ise, kurallar yeni verilerin sınıflanması amacıyla kullanılır.

Karar ağaçlarının kök, dallar ve yapraklardan oluşan ağaca benzeyen bir yapısı olup, örnekteki tüm gözlemleri kapsayan bir kök ile başlayıp aşağıya doğru inildikçe veriyi alt gruplara ayıran dallara ayrılırlar. Bu kökten dallara doğru büyüyen ağaç yapısında her boğum "düğüm" dür, oluşan ağaçlarda homojen olmayan düğümlere "çocuk düğümü (child node)", homojen düğümlere ise "terminal düğüm (parent node)" adı verilir [Pehlivan, G., 2006, 17]. Düğümler üzerinde niteliklerin test işlemi yapılmakta ve test işleminin sonucu ağacın veri kaybetmeden dallara ayrılmasına neden olmaktadır. Her düğümde test ve dallara ayrılma işlemleri ardışık olarak gerçekleşmekte ve sonuç olarak ağaç sınıflar ile son bulmaktadır.

Karar ağacında, tanımlanmış olan soruya ilişkin cevap gruplara ayrılmaktadır. Cevaplar soruya verilecek bir ölçüt belirlendikten sonra setler arasındaki riski maksimize edecek şekilde bölünmekte ve en iyi bölünmeyi bulmak için her soruda aynı işlem tekrar edilmektedir. Bir soru için grup oluşturulduktan ve gruplar arasındaki risk maksimize edildikten sonra oluşan iki grup için bu işlemler devam ettirilmektedir. Bu işlemlere istatistiksel olarak anlamlı bir fark bulunana kadar devam edilmekte, istatistiksel olarak anlamlı bir fark bulunmadığında ise son verilmektedir. Ayrıştırma işlemi tamamlandıktan sonra ise o grup içerisinde yer alan gözlemlerin oranına göre grup değerlendirilmektedir [Thomas, Lyn. C., 2000, 149–172].

Karar ağaçları oluşturulurken kullanılan algoritmanın ne olduğu önemli bir husustur. Kullanılan algoritmaya göre ağacın şekli değişebilir. Bu durumda değişik ağaç yapıları da farklı sınıflandırma sonuçları verecektir. Kök denilen ilk düğümün farklı olması, en uçtaki yaprağa ulaşırken izlenecek yolu ve dolayısıyla sınıflandırmayı da değiştirecektir [Silahtaroğlu, G., (2008 33, 45-47, 58].

Değişkenlerin seçiminde yinelemeli olan algoritmanın döngüden çıkması için o düğümdeki tüm öğelerin aynı sınıfa dahil olması şartı vardır. Eğer kalan değerler sadece bir sınıfa aitse veya sınıflandırılabilecek değer kalmadıysa döngüsel algoritma sonlanır ve karar ağacı oluşturulmuş olur. Sonuçta oluşan sınıflardaki her bir eleman aynı sınıfın diğer elemanları ile benzer özellikler gösterir. Yani ağaç yapısı heterojen yapıdaki veri kümesinin daha küçük ve homojen bir yapıya dönüşmesi için kurallar tanımlar.

Karar ağacı temelli analizlerin yaygın olarak kullanıldığı alanlar ve belli başlı uygulamalar;

- Belirli bir sınıfın olası üyesi olacak elemanların belirlenmesi (Segmentation),
- Çeşitli vakaların yüksek, orta, düşük risk grupları gibi bazı kategorilere ayrılması (Stratification),
- Sadece belirli alt gruplara özgü olan ilişkilerin tanımlanması,
- Gelecekteki olayların tahmin edilebilmesi için kurallar oluşturulması,
- Bireylerin kredi geçmişlerini kullanarak kredi kararlarının verilmesi (Credit Scoring),
- Üretim verilerinin incelenmesiyle ürün hatalarına yol açan değişkenlerin belirlenmesidir [Akpınar H., 2000, 1-22, Masseglia, F., Poncelet, P., Teisseire, M., 1999, 1-19].

Karar ağaçlarına dayalı olarak geliştirilen birçok algoritma vardır. Bu algoritmalar kök, düğüm ve dallanma kriteri seçimlerinde izledikleri yol açısından birbirlerinden ayrılırlar. Karar ağacı oluşturmak için geliştirilen bu algoritmalar arasında;

- CHAID (Chi-Squared Automatic Interaction Detector : Otomatik Ki-Kare Etkileşim Belirleme),
- CART (Classification and Regression Trees: Sınıflama ve Regresyon Ağaçları),
- MARS (Multivariate Adaptive Regression Splines: Çok Değişkenli Uyumlu Regresyon Uzanımları),
- QUEST (Quick, Unbiased, Efficient Statistical Tree: Hızlı, Yansız, Etkin İstatistiksel Ağaç),
- SLIQ (Supervised Learning in Quest),
- SPRINT (Scalable Parallelizable Induction of Decision Trees)
- ID3, C4.5 ve C5.0

yer almaktadır.

Bu çalışmadaki uygulamada CART algoritması kullanılmış olup bir sonraki bölümde anlatılmıştır.

2.1.1 CART algoritması

Bilimsel çalışmalardan elde edilen verilerin analizinde sınıflama ve regresyon ağaçları, kümeleme, diskriminant ve lojistik regresyon analizlerini içeren sınıflama teknikleri ve regresyon modelleri sıklıkla kullanılmaktadır [Teng, J., Lin, K., Ho, B., 2007, 741-748]. Ancak bu tür modellerin gerektirdiği varsayımlar pek çok alanda istatistiksel analiz imkanlarını kısıtlamaktadır. İncelenen veri seti üzerinde hiçbir varsayım gerektirmemesi nedeniyle, sınıflama ve regresyon ağaçları (CART) bu tür parametrik tekniklere karşı güçlü bir alternatif olarak ortaya çıkmaktadır [Temel, G. O., Çamdeviren, H., Akkuş, Z., 2005, 111-117].

CART, hem kategorik hem de sürekli değişkenleri kullanarak sınıflama ve regresyon problemlerinin çözümünde karar ağaçlarını kullanan parametrik olmayan istatistiksel bir metottur. Ele alınan bağımlı değişken kategorik ise yöntem sınıflama ağaçları (Classification Tree), sürekli ise regresyon ağaçları (Regression Tree) olarak adlandırılmaktadır [Deconinck, E., Hancock, T., 2005, 91–103]. Bu yönüyle CART, hem

çoklu regresyon analizini hem de bağımlı değişkenin kategorik olduğu durumlarda kullanılan lojistik regresyon analizini kapsamaktadır.

Yapılan çalışmalarda kullanılan CART algoritması, her aşamada ilgili kümeyi kendinden daha homojen olan iki alt kümeye ayırarak ikili karar ağaçları oluşturan bir yapıya sahiptir. Diğer bir ifadeyle CART, iki çocuk düğümü oluşturup bütün bağımsız değişkenleri kullanarak veriyi alt gruplara ayırmak üzerine kurulmuştur. En iyi bağımsız değişken safsızlık (impurity) ve değişim ölçülerindeki (gini, twoing, en küçük kareler sapması) değişkenliği kullanarak seçilir. Burada, amaç-hedef değişkene ilişkin mümkün olabilen en homojen veri alt gruplarını üretmektir [Kurt, I., Ture, M., Kurum, A. T., 2008, 366–374].

CART, sadece bağımlı değişken ile bağımsız değişken arasındaki ilişkinin yapısını araştırmakla kalmayıp, aynı zamanda bağımsız değişkenlerin birbirleri ile olan etkileşimlerini de ortaya koymaya çalışmaktadır. CART algoritmasının, bağımsız değişkenlerin bağımlı değişkenle ilişkisini değerlendirmede ve model içindeki etkileşim yapısını çözümlemede önemli avantajları mevcuttur.

CART'ın sahip olduğu algoritma, benzerlik gösteren değişkenlerin aynı ağaç düğümünde toplanmasına dayalı olup, bütün oluşturduğu alt dalları bağımlı değişken olan kök düğüme bağlamayla son bulmaktadır [Teng, J., Lin, K., Ho, B., 2007, 741-748].

Her bir düğümün her aşamada ikiye ayrıldığı CART algoritmasında, her bir bölünme noktasının belirlenmesinde Gini, Twoing gibi en iyi bölmeyi seçmek için geliştirilen söz konusu safsızlık ölçütlerinden Gini indeksi kullanılmaktadır.

Tablo 1'de verilen İşe başvuru sırası, eğitim durumu, yaş, cinsiyet ve işe kabul edilip edilmeme durumu isimli 5 nitelikten oluşan bir eğitim verisinde, veriyi daha küçük alt kümelere bölmek için en iyi bölünmenin seçilmesinde kullanılan Gini indeksi aşağıdaki gibi hesaplanmaktadır [Özkan, Y., 2008, 106-113].

Tablo 1. Eğitim verileri

İşe Başvuru Sırası	Eğitim Durumu	Yaş	Cinsiyet	İşe Kabul Durumu
1	Ortaokul	Yaşlı	Erkek	Evet
2	İlkokul	Genç	Erkek	Hayır
3	Yüksekokul	Orta	Kadın	Hayır
4	Ortaokul	Orta	Erkek	Evet
5	İlkokul	Orta	Erkek	Evet
6	Yüksekokul	Yaşlı	Kadın	Evet
7	İlkokul	Genç	Kadın	Hayır

1) Her nitelik değerleri ikili olacak biçimde gruplanmakta ve bu şekilde elde edilen sol ve sağ bölünmelere karşılık gelen sınıf değerleri gruplandırılmaktadır.

2) Her bir nitelikle ilgili olarak sol ve sağ taraftaki bölünmeler için

$Gini_{sol}$ ve $Gini_{sağ}$ değerleri;

k : Sınıfların sayısı, T : Bir düğümdeki örnekler,

T_{sol} : Sol düğümdeki örneklerin sayısı, $T_{sağ}$: Sağ düğümdeki örneklerin sayısı,

L_i : Sol düğümde i kategorisindeki örneklerin sayısı,

R_i : Sağ düğümde i kategorisindeki örneklerin sayısı olmak üzere;

$$Gini_{sol} = 1 - \sum_{i=1}^{k} \left(\frac{L_i}{T_{sol}} \right)^2 \quad (2.1)$$

$$Gini_{sağ} = 1 - \sum_{i=1}^{k} \left(\frac{R_i}{T_{sağ}} \right)^2 , \quad (2.2)$$

şeklinde hesaplanmakta ve her j niteliği için, eğitim verisindeki satır sayısı n olmak üzere genel $Gini$ indeks değeri ise;

$$Gini_j = \frac{1}{n} \left(T_{sol} \times Gini_{sol} + T_{sağ} \times Gini_{sağ} \right) \quad (2.3)$$

formülü ile hesaplanmaktadır.

Tablo 2.1'e göre işe kabul durumu niteliğinde "Evet" sınıfına ilişkin olarak eğitim durumu niteliğinin "ilkokul" değerinden bir tane bulunmaktadır. Benzer şekilde "ortaokul" ve "yüksekokul" değerlerinden ise üç tane bulunmaktadır. Bu şekilde diğer değerler de hesaplanarak nitelik değerlerinin ikili gruplandırılması sonucunda Tablo 2 oluşmaktadır.

Tablo 2. Nitelik değerlerinin ikili gruplandırıldığı eğitim verisi

İşe Kabul Durumu	Eğitim Durumu		Yaş		Cinsiyet	
	İlkokul	Ortaokul Yüksekokul	Genç	Orta Yaşlı	Kadın	Erkek
Evet	1	3	0	4	1	3
Hayır	2	1	2	1	2	1

Nitelik değerlerinin ikili gruplandırılmasından sonra $Gini$ indeks değerleri ise aşağıdaki gibi hesaplanmaktadır:

Eğitim Durumu için:

$$Gini_{sol} = 1 - \left[\left(\frac{1}{3} \right)^2 + \left(\frac{2}{3} \right)^2 \right] = 0{,}444$$

$$Gini_{sağ} = 1 - \left[\left(\frac{3}{4} \right)^2 + \left(\frac{1}{4} \right)^2 \right] = 0{,}375$$

$$Gini_{egitim} = \frac{3 \times 0{,}444 + 4 \times 0{,}375}{7} = 0{,}405$$

Yaş için:

$$Gini_{sol} = 1 - \left[\left(\frac{0}{2} \right)^2 + \left(\frac{2}{2} \right)^2 \right] = 0$$

$$Gini_{sağ} = 1 - \left[\left(\frac{4}{5} \right)^2 + \left(\frac{1}{5} \right)^2 \right] = 0{,}320$$

$$Gini_{yaş} = \frac{2 \times 0 + 5 \times 0{,}320}{7} = 0{,}229$$

Cinsiyet için:

$$Gini_{sol} = 1 - \left[\left(\frac{1}{3} \right)^2 + \left(\frac{2}{3} \right)^2 \right] = 0{,}444$$

$$Gini_{sağ} = 1 - \left[\left(\frac{3}{4} \right)^2 + \left(\frac{1}{4} \right)^2 \right] = 0{,}375$$

$$Gini_{cinsiyet} = \frac{3 \times 0{,}444 + 4 \times 0{,}375}{7} = 0{,}405$$

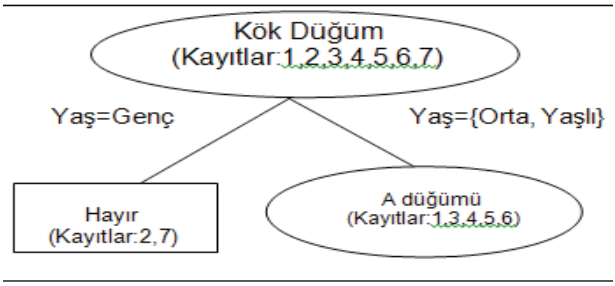
3) Son olarak her j niteliği için hesaplanan $Gini_j$ değerleri arasından en küçük olanı seçilmekte ve bölünme bu nitelik üzerinden gerçekleştirilmektedir.

Tablo 3. Her nitelik için hesaplanan $Gini$ indeks değerleri

İşe Kabul Durumu	Eğitim Durumu		Yaş		Cinsiyet	
	İlkokul	Ortaokul(ve) Yüksekokul	Genç	Orta Yaşlı	Kadın	Erkek
$Gini_{sol}$, $Gini_{sağ}$	0,444	0,375	0	0,32	0,444	0,375
$Gini_j$	0,405		**0,229**		0,405	

Yukarıdaki tabloda hesaplanan değerler göz önüne alındığında $Gini_{yaş} = 0{,}229$ değerinin $Gini_j$ değerleri içinde en küçüğü olduğu anlaşılmaktadır. Bu durumda kök düğümünden itibaren bölünme Yaş=Genç ve Yaş={Orta, Yaşlı} biçiminde olacaktır. Bölünmeyi elde etmek için Tablo 1 üzerinde yaşa ilişkin değerler aranarak, bölünme Yaş=Genç olan (2,7) satırları ve geri kalan (1,3,4,5,6) satırlarından oluşacak ve bölünme Şekil 2'de gösterildiği gibi olacaktır.

Şekil 2. Birinci bölünme sonucu oluşan karar ağacı



Birinci bölünme sonucu oluşan ağaç yapısından sonra yukarıda üç madde halinde sayılan adımlar tekrarlanmakta ve ikinci bölünmenin hangi niteliğe göre olacağı belirlenmektedir. Bunun için öncelikle eğitim verisinden (2,7) satırları çıkarılmakta ve hesaplamalar yeni oluşturulan Tablo 4'e ve eğitim verisinin gruplandırılmış hali olan Tablo 5'e göre tekrarlanmaktadır.

Tablo 4. (2,7) satırları çıkarıldıktan sonra oluşturulan yeni eğitim verisi

İşe Başvuru Sırası	Eğitim Durumu	Yaş	Cinsiyet	İşe Kabul Durumu
1	Ortaokul	Yaşlı	Erkek	Evet
3	Yüksekokul	Orta	Kadın	Hayır
4	Ortaokul	Orta	Erkek	Evet
5	İlkokul	Orta	Erkek	Evet
6	Yüksekokul	Yaşlı	Kadın	Evet

Tablo 5. Yeni eğitim verisinin gruplandırılmış hali

İşe Kabul Durumu	Eğitim Durumu		Yaş		Cinsiyet	
	İlkokul	Ortaokul Yüksekokul	Orta	Yaşlı	Kadın	Erkek
Evet	1	3	2	2	1	3
Hayır	0	1	1	0	1	0

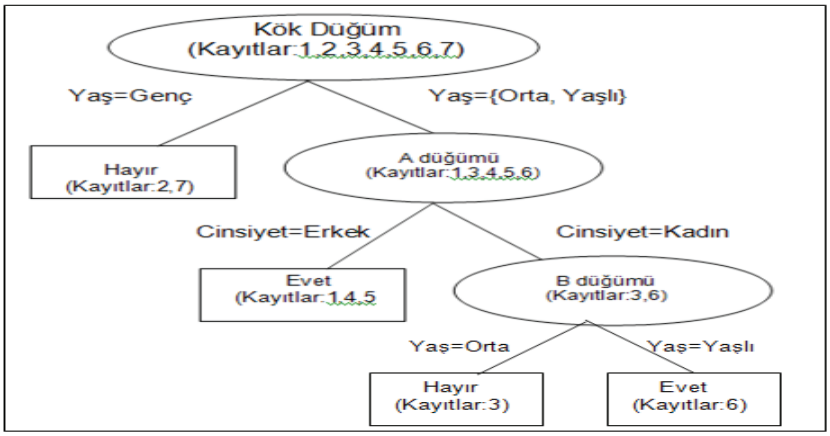
Nitelik değerlerinin ikili gruplandırılmasından sonra yeni bölünme için hesaplanan Gini indeks değerleri ise aşağıdaki gibidir:

Tablo 6. Yeni eğitim verisinde her nitelik için hesaplanan Gini indeks değerleri

İşe Kabul Durumu	Eğitim Durumu		Yaş		Cinsiyet	
	İlkokul	Ortaokul Yüksekokul	Orta	Yaşlı	Kadın	Erkek
$Gini_{sol}$, $Gini_{sağ}$	0,00	0,375	0,444	0,00	0,500	0,00
$Gini_j$	0,300		0,267		**0,200**	

Tablo 6'da gösterilmiş olan hesaplanan bu yeni $Gini_j$ değerleri arasından en küçük olanı seçilmekte ve yeni bölünme de en küçük değere sahip olan $Gini_{cinsiyet} = 0{,}200$ değerinin üzerinden yinelenmektedir. Bu durumda yeni bölünme cinsiyet niteliğinin "kadın" ve "erkek" değerlerine göre olacaktır. Benzer şekilde işlemlerin tekrarlanması sonucunda ise son bölünme iki ayrı sınıf olarak tanımlanan yaş niteliğine göre olup bu durum sonucunda oluşan nihai karar ağacı da Şekil 3'te gösterilmiştir [Özkan, Y., 2008, 106-113].

Şekil 3. Nihai karar ağacı



2.2. Lojistik Regresyon Analizi

Son yıllarda tıp, biyoloji, tarım ve ekonomi gibi alanlarda kolay kullanımı ve yorumlanması nedeniyle lojistik regresyon yaygın olarak kullanılan ve tercih edilen bir yöntem haline gelmiştir.

İstatistiksel uygulamalarda araştırmacılar tarafından genellikle bağımsız değişkenlerle bağımlı değişken arasında ilişki olup olmadığı analiz edilmek istenir. Yapılan istatistiksel analiz yöntemlerinde, verilerin yapısına göre en uygun yöntemin belirlenmesi büyük önem arz etmektedir. Bağımlı değişken sürekli olduğunda, genellikle doğrusal regresyon modeli kullanılmaktadır. Doğrusal modellerin önemli bir varsayımı hata terimlerinin normal dağılıma sahip olmasıdır. Tipik doğrusal regresyon modeli:

$$Y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2) \tag{2.4}$$

biçiminde tanımlanmaktadır. Belirtilen doğrusal regresyon modelinde bağımsız değişkenlerin kesikli veya sürekli olmaları modelin tahmininde kullanılacak yöntemi ve bu yöntemle elde edilen parametre tahminlerinin özelliklerini etkilemez. Bu nedenle modele girecek bağımsız değişkenler hem kesikli hem de sürekli değişkenler olabilirler. Buna karşın modeldeki bağımlı değişkenin kesikli bir yapıya sahip olmasının etkisi büyüktür. Bağımlı değişkenin kategorik olması durumunda normallik varsayımı bozulmakta ve tipik doğrusal model uygulanamamaktadır.

Bağımlı değişken iki değer aldığında model çeşitli dağılımlara dayalı olarak doğrusal regresyon modelinden farklı biçimde tanımlanmaktadır. Bağımlı değişkenin iki ya da çok sınıflı kesikli değişken olması durumunda kullanılabilecek modeller çok çeşitlidir. Bu modellerden doğrusal olasılık modeli, lojit ve probit modeller arasında en fazla tercih edilen yöntem lojistik regresyondur. Lojistik regresyon analizini, doğrusal regresyon analizinden ayıran en belirgin özellik de lojistik regresyon analizinde bağımlı değişkenin iki ya da çok sınıflı olmasıdır. Lojistik regresyon ve doğrusal regresyon analizi arasındaki bu farklılık hem parametrik model seçimine hem de varsayımlara yansımaktadır [Hosmer, D. W., Lemeshow, S., 1989, 5-50]. Lojistik regresyon, normallik varsayımının bozulması nedeniyle doğrusal regresyon analizine alternatif olmaktadır. Doğrusal regresyon analizinde bağımlı değişkenin değeri, lojistik regresyon analizinde ise bağımlı değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı tahmin edilmektedir. Temel olarak lojistik regresyonda bağımsız değişkenler ile iki ya da çok sınıflı kategorik bağımlı değişken arasındaki ilişkinin tanımlanması için matematiksel modelleme yapmak amaçlanmaktadır [Kleinbaum,G., D., 1994].

Lojistik regresyon analizinde modelleme kısmında kullanılacak olan lojistik modeli elde etmek için izlenen adımlara aşağıda kısaca değinilmiştir.

Herhangi bir i'inci gözlem için;

$$y_i = \sum_{k=0}^{p} \beta_k x_{ik} + \varepsilon_i \qquad (2.5)$$

şeklinde ifade edilen modelde bağımsız değişkenler üzerinde bir kısıt yoktur. Aynı zamanda y_i bağımlı değişken değeri de -∞ ile +∞ arasında tüm değerleri alabilmektedir. Bağımlı değişkenin 0 ve 1 gibi değerler aldığı durumda bu kural bozulmakta ve $P(y_i = 1)$, i'inci gözlemin 1 değerini alma olasılığı olmak üzere, beklenen değer:

$$E(y_i) = 1 \times P(y_i = 1) + 0 \times P(y_i = 0) = P(y_i = 1) \qquad (2.6)$$

olmaktadır.

Bu sonuç regresyon denklemi olarak yazılacak olursa:

$$E(y_i) = P(y_i = 1) = \sum_{k=0}^{p} \beta_k x_{ik}, \quad i = 1,\ldots,n \tag{2.7}$$

ifadesi elde edilmektedir. Sol tarafı 0-1 arasında olasılık değerleri alan bu denkleme "Doğrusal olasılık modeli" adı verilmektedir [Tatlıdil, H., 1996].

Doğrusal olasılık modelinde bağımlı değişken değeri olarak ifade edilen olasılık değerinin çeşitli dönüşümlerle -∞, +∞ arasında tanımlı hale getirilmesi amacıyla yapılacak dönüşümlerden birisi lojit dönüşüm olup, lojit dönüşümde ilk olarak;

$$E(y_i) = P(y_i = 1) = \sum_{k=0}^{p} \beta_k x_{ik} \qquad i = 1,\ldots,n \quad \infty \tag{2.8}$$

modelinde olasılık değerleri üzerinde $P/1-P$ dönüşümü yapılarak bağımlı değişkenin sınırları 0, +∞ yapılmakta, daha sonra ise bu oran değerinin logaritması alınarak bağımlı değişkenin sınırları -∞, +∞ yapılmaktadır. Bu dönüşümlerden sonra elde edilen yeni fonksiyon:

$$E(y_i) = P(y_i = 1) = L_i = ln(\frac{P_i}{1-P_i}) = \sum_{k=0}^{p} \beta_k x_{ik} \tag{2.9}$$

olarak yazılabilir. Bu modele de "Lojistik model" ya da kısaca "Lojit" denmektedir. Ayrıca kullanılan $\ln(\frac{P}{1-P})$ dönüşümü de "lojit dönüşüm" adını almaktadır [Hosmer, D. W., Lemeshow, S., 1989, 5-50].

Lojistik fonksiyonun elde edildiği modelde kullanılan P_i olasılık değeri ise:

$$P_i = \frac{\exp\left(\sum_{k=0}^{p} \beta_k x_{ik}\right)}{1+\exp\left(\sum_{k=0}^{p} \beta_k x_{ik}\right)} \tag{2.10}$$

biçiminde tanımlanmaktadır [Collet, D., 2003].

P_i olasılık değerine sahip lojistik analizde en önemli noktalardan biri kurulan modelin katsayılarının yorumlanmasıdır. Bağımsız bir x_k değişkeninin katsayısı β_k, x_k'da meydana gelen bir birim değişikliğin y bağımlı değişkeni üzerinde yarattığı değişimin miktarını ve yönünü vermektedir. Bunun için öncelikle bağımlı ve bağımsız değişkenler arasındaki fonksiyonel ilişkinin bulunması gereklidir.

Bir modeldeki bağımsız değişkenler ile bağımlı değişken arasındaki lineer ilişkiyi veren fonksiyona "link fonksiyonu" adı verilmektedir. Bağımlı değişkenin tanımı gereği parametrelerinde doğrusal olan doğrusal regresyon modelinde link fonksiyonu birim fonksiyon (matris) iken; lojistik regresyonda söz konusu fonksiyon logit dönüşümdür ve Eş. 2.9 tanımından yararlanarak hatırlanacağı üzere lojistik regresyon modelindeki lojit değişim de,

$$g(x) = \ln\{P(x)[1-P(x)]\} = \beta_0 + \beta_1 x$$

şeklinde idi. Buna göre lojistik regresyon modelinde β_1 katsayısı, x bağımsız değişkeninin bir birim değişiminin lojitte sağlayacağı değişim olup, $\beta_1 = g(x+1) - g(x)$ olarak ifade edilmektedir. Yani lojistik regresyon modelinde katsayının yorumu, iki lojit arasındaki farka anlam kazandırılması esasına dayanmaktadır.

Bağımsız değişken x'in iki sınıflı olduğu, yani 0 ve 1 değerlerini aldığı durumda $P(x)$ ve $1-P(x)$'in iki ayrı değeri söz konusudur ve çizelge 3.1'de bağımsız değişkenin iki sınıflı olması durumunda lojistik regresyon modelinin alacağı bu değerler gösterilmiştir [Hosmer, D. W., Lemeshow, S., 1989, 5-50].

Tablo 7. Bağımsız değişken iki sınıflı olduğunda lojistik modele ilişkin değerler

Bağımlı Değişken (y)	Bağımsız Değişken (x)	
	$x=1$	$x=0$
$y=1$	$P(1)=\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$P(0)=\frac{e^{\beta_0}}{1+e^{\beta_0}}$
$y=0$	$1-P(1)=\frac{1}{1+e^{\beta_0+\beta_1}}$	$1-P(0)=\frac{1}{1+e^{\beta_0}}$
TOPLAM	**1**	**1**

x = 1 iken sonucun olma olasılığı, $P(1)/[1-P(1)]$, x = 0 iken sonucun olma olasılığı da $P(0)/[1-P(0)]$ şeklinde tanımlanmaktadır. İki sınıflı bağımlı değişkenin iki kategorisinin görülme olasılıklarının birbirine oranlanmasına "Odds" adı verilir. Lojistik regresyon konusundaki önemli kavramlardan birisi "Odds oranı"dır ve katsayıların yorumlanması için "Odds"lar ve "Odds oranı"ndan yararlanılmaktadır. "Odds oranı" herhangi bir olayda tercih etmenin tercih etmemeye oranı olarak tanımlanabilmektedir. Örneğin, ilgilenilen türden bir olayın olma olasılığı (p) ise, diğer olayın olma olasılığı (1-p) olacaktır. Odds w ile gösterilecek olursa Odds oranı bu iki olasılığın oranlanması ile bulunmaktadır.

$$w=\frac{\pi}{1-\pi} \qquad (2.11)$$

Tablo 2.7'de olduğu gibi bağımlı değişken 0 ve 1 değerleri verilerek kodlanırsa, $P(x)$ olasılığı ile bağımlı değişken verilen x değeri için 1'e eşit olur. Öte yandan $1-P(x)$ olasılığı ile verilen x değeri için 0 olur. Bu durumda, eğer bağımsız değişken de iki kategorili bir değişken ise ve 0, 1 olarak kodlanmışsa;

x =1 için: $w_1 = \frac{P(1)}{1-P(1)}$, x =0 için: $w_0 = \frac{P(0)}{1-P(0)}$ (2.12)

olarak tanımlanabilir. Bu durumda Odds oranı ise:

$$\Psi = \frac{w_1}{w_0} = \frac{\frac{P(1)}{1-P(1)}}{\frac{P(0)}{1-P(0)}} \qquad (2.13)$$

olacaktır. Odds oranı bağımlı değişkenin y =1 görülme olasılığı bakımından, x =1 olanları,

x =0 olanlarla karşılaştırmada kullanılır. Tablo 2.7 için odds oranına bakılacak olursa:

$$\psi = \frac{(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}})(1+e^{\beta_0+\beta_1})}{(\frac{e^{\beta_0}}{1+e^{\beta_0}})(1+e^{\beta_0})} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1} \qquad (2.14)$$

olacaktır. Bu, iki sonuçlu bağımsız değişkenin lojistik regresyonu için odds oranıdır. Bu odds oranı için lojit farkı ise,

$$\ln\psi = \ln e^{\beta_1} = \beta_1 \qquad (2.15)$$

katsayısına eşittir.

Odds oranı, nispi risk[1] ile yakından ilgili olup bir bağlantı ölçümüdür. Şöyle ki, eğer ilgilenilen durumun olma olasılığı düşük ise odds oranı nispi riske yakın sonuçlar verir. Bu oran x = 1 için sonucun olma olasılığının, x = 0'ın olma olasılığından ne kadar çok ya da az olduğunun tahminini yapar. Örneğin y bağımlı değişkeni akciğer kanseri olup olmamasını gösteriyor ve x bağımsız değişkeni de bir kişinin sigara kullanıp kullanmadığını gösteriyorsa $\psi = 2$'nin yorumu; bir yığında sigara kullananlar arasında akciğer kanseri olma olasılığının, sigara kullanmayanlara göre 2 kat daha fazla olacağı şeklindedir.

Odds oranının anlamının daha iyi anlaşılması için başka bir örnek vermek gerekirse; bir hastalığa karşı 25 kişilik kontrol gurubu seçilmiş olsun. Tedavi grubu yeni bir ilaç alarak bu hastalığa yakalanma riskini azaltmaya çalışsın. Deneyin sonunda tedavi gurubundan iki ve kontrol gurubundan üç kişi bu hastalığa yakalanmış olsun.

Tedavi gurubu için risk, r_t = 2/25=0,08

Kontrol gurubu için risk, r_k = 3/25=0,12 olacaktır.

Buradan nispi risk hesaplanmak istendiğinde;

$$\frac{r_t}{r_k} = \frac{0{,}08}{0{,}12} = 0{,}667$$

olacaktır. Yani tedavi gurubundaki risk, kontrol gurubundan 0,667 kat daha fazla ya da 1/0,667=1,5 olduğundan kontrol gurubundaki risk, tedavi gurubundan 1,5 kat daha fazladır.

Bu örneğe ilişkin odds oranı ise:

$$\psi = \frac{r_t/(1-r_t)}{r_k/(1-r_k)} = \frac{0{,}08/(1-0{,}08)}{0{,}12/(1-0{,}012)} = 0{,}64$$

olacaktır.

1 Nispi risk: Araştırılan etkene maruz kalan grupta elde edilen sonucun, etkene maruz kalmayan kontrol grubundakilere oranı olarak tanımlanmaktadır.

r_t ve r_k değerleri küçük ise odds oranı nispi riske yakın sonuçlar verecektir. En yaygın kullanım alanları iki kategorili değişken arasındaki ilişkinin ölçüldüğü alanlar olan odds oranları hassastır ve lojistik regresyon analizinde önemli bir ölçüttür [Allison, D. P., 2000]. Özet olarak, lojistik regresyon katsayısı ve olasılık oranı arasındaki ilişki, lojistik regresyon sonuçlarını açıklamamız için bir temel teşkil etmektedir.

3. Uygulama

Çalışmanın bu bölümünde, çok geniş bir çalışma alanı olan veri madenciliğinin sınıflama ve regresyon modellerine ilişkin teknikleri ile sınırlandırılan, Sosyal Güvenlik Kurumu (SGK) İlaç Provizyon Sistemi veri tabanından elde edilen veri kümesi üzerinde bir uygulama yapılmıştır. Bu kapsamda, veri madenciliğinin söz konusu teknikleri arasında yaygınlıkları dikkate alınarak karar ağaçları algoritmalarından CART algoritması ve lojistik regresyon analizi kullanılmıştır.

Uygulamaya konu olan veri kümesi, İlaç Provizyon Sisteminden solunum sistemi hastalıkları için antibiyotik kullanan yaklaşık 50 milyon hasta içerisinden örnekleme yoluyla seçilen 18.931.000 hastanın 12 farklı değişkene ilişkin değerlerini içermektedir. Uygulamada söz konusu veri kümesi üzerinde Clementine 12.0 yazılımı kullanılarak veri madenciliğinin sınıflama ve regresyon problemine ilişkin CART ile lojistik regresyon teknikleri için örnek bir uygulama ortaya konmuştur. Çalışmanın amacı bu veri kümesi[2] için, veri madenciliği uygulaması ile penisilin grubu antibiyotik kullanan hastaların profilini belirleyen önemli faktörlerin araştırılarak ortaya çıkarılmasıdır.

Verilerin analizinden önce, uygulanacak teknikler için nihai veri kümesinin oluşturulması amacıyla verileri temizleme işlemi gerçekleştirilmiştir. Veri kümesi üzerinde yapılan ilk incelemede bazı veri kalitesi sorunları tespit edilmiştir. Kayıtlarda yer alan hastaların bazı demografik değişkenlere ilişkin değerlerinin bilinmediği görülmüştür. Ayrıca bazı çelişkili veri değerleri olduğu ve bazı veri değerlerinin yanlış kodlandığı gözlemlenmiştir.

2 Uygulamada kullanılan veri kümesi, Sosyal Güvenlik Kurumu'na ait olması ve hastaların kişisel bilgilerini içermesi sebebiyle gizlilik içermektedir. Bu açıdan ilgili hastalara ilişkin bilgi düzeyleri paylaşılmamıştır.

Değişken bazında yapılan incelemede hasta profili açısından bilgi içermeyen verilerin belirlenmesi ile bu kayıtların veri kümesinden çıkarılması sağlanmıştır. Bazı değişkenlerde değer olduğu halde bu değerleri üretecek diğer değişkenlerin bulunmaması oranların bozulmasına neden olacağından bu tür veriler de değerlendirme dışı tutulmuştur. Örneğin doktorun branş kodu bulunmamasına rağmen tanı kodu veri setinde yer almakta ise verileri eşleştirmek mümkün olmadığından bu kayıt inceleme dışında bırakılmıştır.

Aynı zamanda veri kayıtlarında tutarsızlıkların tespit edilmesi sonucu tutarsızlıklarının düzeltilmesi mümkün görülmeyen kayıtlar da benzer şekilde veri kümesinden çıkarılmıştır. Özellikle hastalara konulan tanıların kodlanmasında yanlışlık olduğu tespit edildiğinden verilerin içerdiği toplam 30 tanıdan sadece 11 tanının analize alınması uygun bulunmuştur. Bu işlemlerle verilerin temizlenmesi sağlanmış ve veri kümesi modelleme için hazır hale getirilmiştir.

Verilerin temizlenmesi işleminden sonra çalışma kapsamında ilaç provizyon sistemi üzerinde yer alan tanılardan 11 tanıya ilişkin 6.772.313 kayıtlık reçete verisi[3] kullanılmıştır. Reçete bilgilerinde yer alan penisilin kullanımı için, penisilin kullananlar "1" kullanmayanlar "2" olarak kodlanarak veriler ikiye ayrılmış ve analizde iki seviyeli kategorik bağımlı değişken olarak kullanılmıştır. Bu şekilde yapılan kodlama sonucunda penisilin kullanan toplam 2.484.352 kişinin olduğu tespit edilmiştir. Söz konusu bağımlı değişkeni önemli derecede etkileyen tanı grubu, hastane grubu, fiyat aralığı, cinsiyet ve yaş bağımsız değişkenler olarak ele alınmıştır. Analizde kullanılan bu bağımsız değişkenler aşağıda açıklanmıştır:

Tanı grubu:

İlk bağımsız değişken olan tanı grubu kategorik bir değişken olup, provizyon sisteminde kodlandığı haliyle aynen alınmış ve aşağıda sunulmuştur:

3 Veriler her hastaya bir reçete karşılık gelecek şekilde oluşturulduğundan hasta sayısı reçete sayısına eşittir.

711: Akut tonsillit, 712: Akut larenjit ve trakeit, 713: Akut obstrüktif larenjit ve epiglottit, 714: Akut üst solunum yolu enfeksiyonları birden fazla olan, 715: Belirlenmiş influenza virüsüne bağlı İnfluenza, 718: Streptococcus pneumoniae'ye bağlı Pnömoni, 720: Bakteriyel pnömoniler, 722: Başka yerde sınıflanmamış hastalıklarda bulunan Pnömoni, 724: Akut bronşit, 728: Kronik rinit, nazofarenjit ve farenjit, 744: Bronşiektazi

Hastane grubu:

İkinci bağımsız değişken olarak modele alınan ve hastanenin bağlı bulunduğu kurum ile işlevsel özelliğini gösteren bu değişkenin tanım ve atama değerleri aşağıdaki gibi 4 kategori ile yapılmıştır:

Özel Hastaneler: ..... "1",

2. Basamak Sağlık Bakanlığı Hastaneleri: ..... "2",

3. Basamak Sağlık Bakanlığı Hastaneleri: ..... "3",

Üniversite Hastaneleri: ..... "4"

şeklinde yapılmıştır.

Fiyat: Üçüncü bağımsız değişken olan ilaç fiyatı ise;

ilacın fiyatı;

<= 5 TL olanlar için: ..... 1",

5 TL ile 25 TL olanlar için: ..... "2"

> 25 TL olanlar için: ..... "3"

şeklinde kodlanarak analize dahil edilmiştir.

Cinsiyet:

Kesikli bir değişken olan cinsiyet bağımsız değişkeni de; kadın için "1" erkek için "2", şeklinde kodlanmıştır.

Yaş:

Yaş bağımsız değişkeni sürekli bir değişken olup yıl cinsinden ölçülmüştür ve yaşı;

0-15 olanlar için: ..... "1"
15-29 olanlar için: ..... "2"
30-44 olanlar için: ..... "3"
45-64 olanlar için: ..... "4"
65 ve üzeri yaşta olanlar için: ..... "5"

olarak kodlanmış ve CART analizi bu kodlamaya göre yapılmıştır. Ancak uygulamanın ikinci kısmı olan lojistik regresyon analizinde yaş değişkeni kodlama yapılmadan sürekli değişken olarak analize dahil edilmiştir. Uygulamaya dahil edilen 6.772.313 hasta "penisilin kullananlar" ve "penisilin kullanmayanlar" olarak sınıflandırılmış ve Tablo 8'de gösterilmiştir.

Tablo 8. Hastaların penisilin kullanma durumlarına göre dağılımı

Bağımlı Değişken	Sayı	Yüzde (%)
Penisilin Kullananlar	2.484.352	36.7
Penisilin Kullanmayanlar	4.287.961	63.3

Çalışmada daha önce de belirtildiği gibi veri madenciliğinin sınıflandırma fonksiyonuna ilişkin CART ve lojistik regresyon yöntemleri ile sınıflandırma modellerinin geliştirilmesi amaçlanmıştır. Sınıflandırma modelleri için iki temel başarı kriteri söz konusu olduğundan, öncelikle

geliştirilen modelin eğitim verisi üzerinde sınıflandırma başarısı birinci başarı kriteri; veri madenciliğinin amacı doğrultusunda geliştirilen modelin öngörü amaçlı kullanılabilmesi için modelin eğitim kümesinden tamamen farklı bir test kümesi üzerinde sınanması ise ikinci başarı kriteri olarak belirlenmiştir.

Bu doğrultuda analizlerde tüm verinin % 70'i model oluşturmak amacı ile eğitim verisi, geri kalan % 30'u ise sınıflama kurallarının doğruluğunu test etmek amacıyla test verisi olarak kullanılmıştır. Böylece sınıflandırma modeli öğrenme kümesi üzerinde geliştirilmiş olup test verisinden oluşan sınama kümesi üzerinde de öngörü başarılarının sınanması sağlanmıştır.

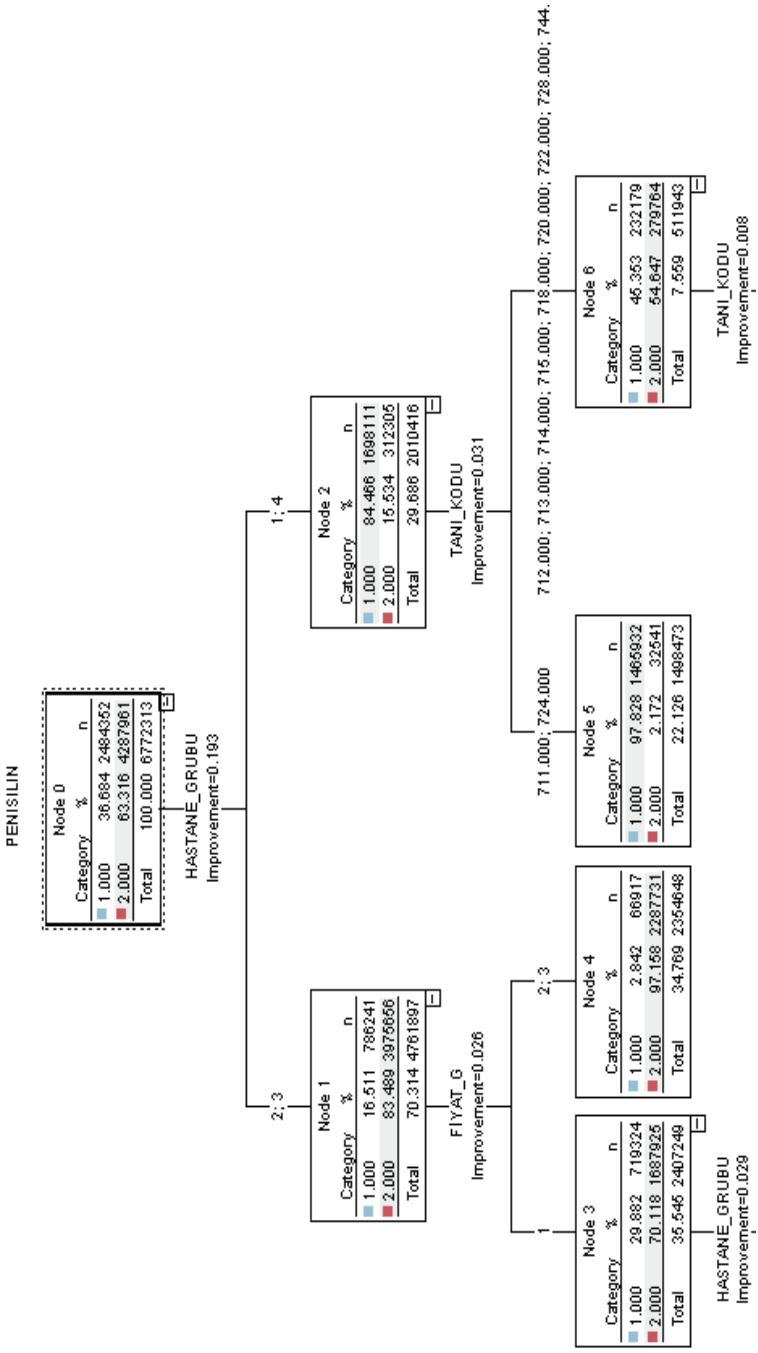
3.1. CART Analizi Uygulaması

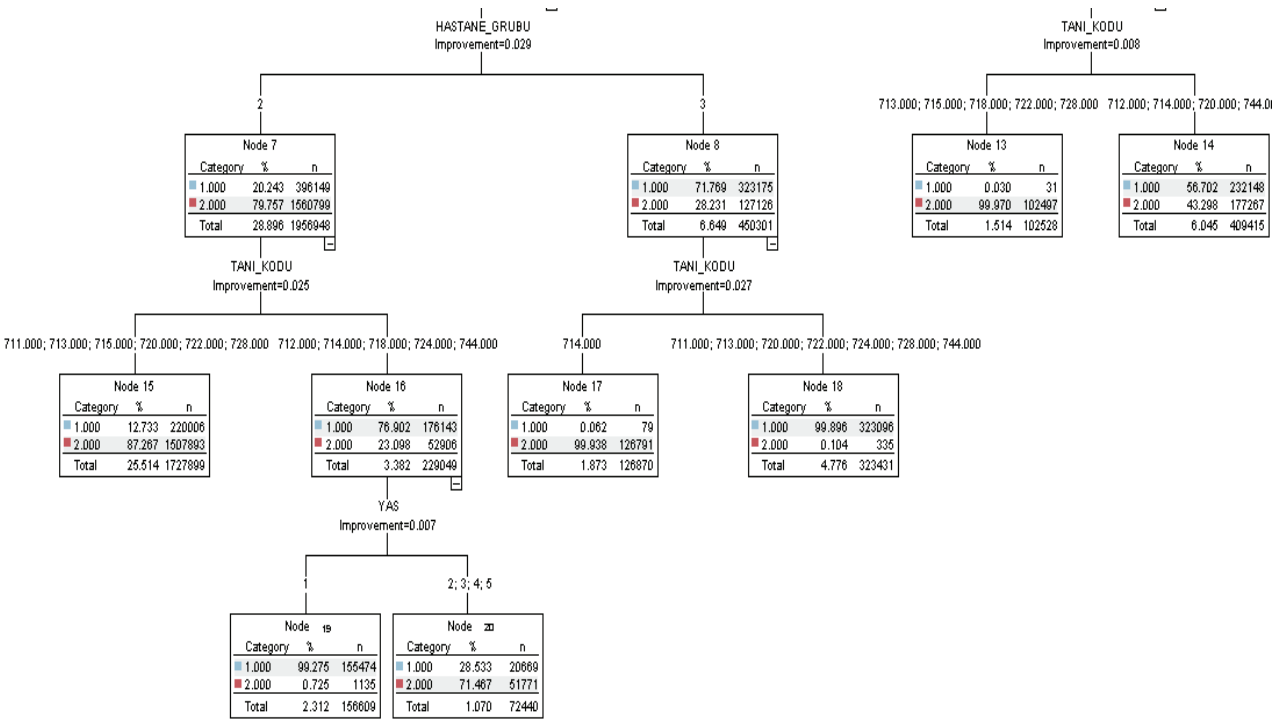
Bu tekniğin uygulanmasında öncelikle her değişken için hazırlanmış olan veri kümeleri clementine 12.0 yazılımına yüklenmiş, değişkenlerin seçimi "select" düğümü ile yapılmış ve modele girmesi istenmeyen değişkenler de "filter" düğümü ile elenmiştir. Ayrıca yaş ve fiyat gibi değişkenler "derive" düğümü ile formülüze edilerek gruplandırılmış, hedef (bağımlı) değişken de "type" düğümü ile tanımlanmıştır. Son olarak "partition" düğümü ile verilerin eğitim (training) ve test (testing) verisi olarak ayrılması sağlanarak model kurma aşamasına gelinmiştir.

Modelin tahmin edilmesinde "simple" modu seçeneği tercih edilerek maksimum ağacın derinliği yani ağacın büyüyebileceği katman sayısı 5 olarak belirlenmiştir. Ayrıca ağacın büyümesine rehberlik eden ve ana dallarda tek bir çıktı kategorisine yoğunlaşan tepkilerin düzeyini yakalayan safsızlık kriteri olarak da "Gini Safsızlık Ölçütü" tercih edilmiştir.

CART ağacı büyütürken daha önce de bahsedildiği gibi bir dalın dağınıklığını yani safsızlığını en çok azaltan tahminleyicide bölmekte ve ana daldan yavru dala doğru dağınıklıktaki bu değişime de gelişme (improvement) denmektedir. Analizde bu gelişme değeri için "default" değer olan 0,0001 seçilerek model oluşturulmuş ve oluşturulan modele ilişkin ağaç yapısı da Şekil 4'te gösterilmiştir.

Şekil 4. CART analizi sonucu karar ağacı




HASTANE_GRUBU
Improvement=0.029
2
3
Node 7
Category % n
1.000 20.243 396149
2.000 79.757 1560799
Total 28.896 1956948
Node 8
Category % n
1.000 71.769 323175
2.000 28.231 127126
Total 6.649 450301
TANI_KODU
Improvement=0.008
713.000; 715.000; 718.000; 722.000; 728.000
712.000; 714.000; 720.000; 744.0
Node 13
Category % n
1.000 0.030 31
2.000 99.970 102497
Total 1.514 102528
Node 14
Category % n
1.000 56.702 232148
2.000 43.298 177267
Total 6.046 409415
TANI_KODU
Improvement=0.025
TANI_KODU
Improvement=0.027
711.000; 713.000; 715.000; 720.000; 722.000; 728.000
712.000; 714.000; 718.000; 724.000; 744.000
714.000
711.000; 713.000; 720.000; 722.000; 724.000; 728.000; 744.000
Node 15
Category % n
1.000 12.733 220006
2.000 87.267 1507893
Total 25.514 1727899
Node 16
Category % n
1.000 76.902 176143
2.000 23.098 52906
Total 3.382 229049
Node 17
Category % n
1.000 0.062 79
2.000 99.938 126791
Total 1.873 126870
Node 18
Category % n
1.000 99.896 323096
2.000 0.104 335
Total 4.776 323431
YAS
Improvement=0.007
1
2; 3; 4; 5
Node 19
Category % n
1.000 99.275 155474
2.000 0.725 1135
Total 2.312 156609
Node 20
Category % n
1.000 28.533 20669
2.000 71.467 51771
Total 1.070 72440

Şekil 4 incelendiğinde solunum sistemi hastalıklarında yazılan toplam 6.772.313 antibiyotiğin 2.484.352'sinin penisilin grubu antibiyotik olduğu görülmektedir. Yazılan penisilin grubu antibiyotiklerin en önemli belirleyicileri arasında sırasıyla hastane grubu, fiyat grubu, tanı kodu ve yaş değişkenleri yer almaktadır. Bu belirleyiciler penisilin grubu antibiyotik kullanımını 9 ayrı profile ayırmaktadır. Bu profiller Şekil 4 ve tablo 9'dan yararlanarak aşağıda açıklanmakta ve seçilmiş bazı profillere ilişkin yorumlar da aşağıda yer almaktadır.

Tablo 9. Reçeteye yazılan penisilin grubu antibiyotiklerin en önemli belirleyicileri ve profilleri

Profiller	Düğümler	Hastane Grubu	Fiyat Grubu	Tanı Kodu	Yaş
Profil 1	Node 19	2,3	1	712,714,718,724,744	1
Profil 2	Node 20	2,3	1	712,714,718,724,744	2
Profil 3	Node 15	2,3	1	711,713,715,720,722,728	
Profil 4	Node 17	2,3	1	714	
Profil 5	Node 18	2,3	1	711,713,720,722,724,728,744	
Profil 6	Node 4	2,3	2,3		
Profil 7	Node 13	1,4		712,713,714,715,718,720,722, 728,744	
Profil 8	Node 14	1,4		712,713,714,715,718,720,722, 728,745	
Profil 9	Node 5	1,4		711,724	

Penisilin kullanımı için ilk sınıflamanın hastane grubuna göre olduğu görülmektedir. Penisilin grubu antibiyotiklerin % 16'sının Sağlık Bakanlığı 2. ve 3. basamak hastanelerinde (Node 1), % 84'ünün ise özel hastaneler ve üniversite hastanelerinde (Node 2) reçetelendirildiği

ve bu hastanelerde yazılan antibiyotiklerin de alt kırılımların fiyat grubu, tanı kodu ve yaşa göre önemli bulunarak sınıflandırıldığı gözlemlenmektedir.

Birinci temel profilde (Node 1) yer alan solunum sistemi hastalıklarında yazılan 4.761.897 antibiyotik ilk altı alt profildeki (Node 19, Node 20, Node 15, Node 17, Node 18, Node 4) antibiyotiklerin sınıflandırılmasını içermektedir. İlk iki profilde (Node 19 ve Node 20) yer alan penisilin grubu antibiyotiklerin Sağlık Bakanlığı 2. basamak hastanelerinde, fiyatı 5 TL'nin altında ve 712 tanı kodlu akut larenjit ve trakeit, 714 tanı kodlu akut üst solunum yolu enfeksiyonları birden fazla olan, 718 tanı kodlu streptococcus pneumoniae'ye bağlı Pnömoni, 724 tanı kodlu akut bronşit ve 744 tanı kodlu bronşiektazi hastalarına yazıldığı görülmektedir.

Birinci temel profil altında aynı hastane grubu, fiyat aralığı ve tanı koduna göre sınıflandırılan, ayrıca yaşı 0 ile 15 yaş arasında ve penisilin türü antibiyotik kullananların oranı % 99 olan toplam 156.609 hasta profil 1'i (Node 19) oluştururken, 15 yaş üzeri ve penisilin kullananların oranı yaklaşık % 29 olan toplam 72.440 hasta profil 2'yi (Node 20) oluşturmaktadır. Bu profillerden solunum sistemi hastalıkları için antibiyotik yazılan hastalar içerisinde penisilin grubu antibiyotik kullanım oranı en yüksek olan hastalar profil 1'de yer almaktadır. Bu durum 0–15 yaş arası okul çağında olan nüfusun beta mikrobu gibi çeşitli mikroplar nedeniyle sıklıkla hastalanması ve hastalığın tedavisinde penisilin grubu antibiyotiklerin kullanımının yaygın olmasıyla açıklanabilmektedir.

Birinci temel profilde Sağlık Bakanlığı 2. ve 3. basamak hastaneleri için yapılan temel sınıflandırma içerisinden fiyatı 5 TL'nin altında, 3. Basamak hastanelerde ve tanı kodu 711 olan akut tonsillit, 713 olan akut obstrüktif larenjit ve epiglottit, 720 olan bakteriyel pnömoni, 722 olan başka yerde sınıflanmamış hastalıklarda bulunan Pnömoni, 724 olan akut bronşit, 728 olan kronik rinit ile tanı kodu 744 olan bronşiektazi hastası için yazılan toplam 323.431 antibiyotik de profil 5'i (Node 18) oluşturmakta ve bu profilde yer alan hastalar % 99 gibi yüksek bir oranda penisilin grubu antibiyotik kullanmaktadır.

Özel hastane ve üniversite hastanelerinde solunum sistemi hastalıkları için yazılan toplam 2.010.416 antibiyotik içerisinden penisilin grubu antibiyotik kullanımının % 84'lük oranla oldukça yüksek olduğu İkinci temel profil (Node 2) üç ayrı alt profile (Node 13, Node 14, Node 5) ayrılmaktadır. Bu profillerde yazılan penisilin grubu antibiyotikler için en önemli belirleyicinin tanı kodu değişkeni olduğu görülmektedir.

İkinci temel profil altında sınıflandırılan ilk iki profilde (Node13, Node 14) yer alan profil 8' de (Node 14) toplam 409.415; 712 tanı kodlu akut larenjit, 714 tanı kodlu akut üst solunum yolu enfeksiyonları birden fazla olan, 720 tanı kodlu bakteriyel pnömoni ve 744 tanı kodlu bronşiektazi hastası için yazılan antibiyotiklerden penisilin grubu antibiyotikleri kullananların oranının % 57 gibi yüksek bir oranda olduğu gözlenmektedir.

Ayrıca ikinci temel profil için tanı koduna göre yapılan sınıflamada, tanı kodu 711 olan akut tonsillit ve tanı kodu 724 olan akut bronşit hastaları için yazılan antibiyotikler de profil 9'u (Node 5) oluşturmakta ve bu profilde yer alan toplam 1.498.473 antibiyotik kullanan hastanın % 98'ini penisilin grubu antibiyotik kullanan hastalar oluşturmaktadır.

Yukarıda bazı profillerin açıklandığı penisilin hedef (bağımlı) değişkeni üzerinde gerçekleştirilen CART algoritması sonucunda kurulan modelin sınıflandırma başarısı % 92,34 olarak hesaplanmıştır. Ayrıca modelin öngörü başarısını sınamak amacıyla hazırlanan test verisi üzerindeki sınıflandırma başarısı da % 92,31 olarak hesaplanmış ve Tablo 10 ve 11'de söz konusu oranların sunulduğu sınıflandırma tabloları ortaya çıkmıştır.

Tablo 10. CART analizi sonucu elde edilen doğru sınıflandırma oranı tablosu

			Tahmin Değerleri			
			Penisilin			
			Kullananlar	Kullanmayanlar	Toplam	Doğru Sınıflandırma Oranı
Gerçek Değerler	Penisilin	Kullananlar	2.176.650	307.702	2.484.352	87,61%
		Kullanmayanlar	211.278	4.076.683	4.287.961	95,07%
Toplam			2.387.928	4.384.385	6.772.313	**92,34%**

Tablo 3.3'de görüldüğü gibi uygulama kapsamındaki 6.772.313 hastadan gerçekte 2.484.352'si penisilin grubu antibiyotik kullanırken geriye kalan 4.287.961 kişi penisilin grubu antibiyotik kullanmamaktadır. Penisilin kullanan hastaların 2.176.650'si doğru, 307.702'si ise hatalı olmak üzere % 87,61'lik doğruluk yüzdesiyle sınıflandırılmıştır. Penisilin grubu antibiyotik kullanmayan 4.287.961 hastanın ise 4.076.683 tanesi CART algoritması ile yapılan sınıflandırma işleminde doğru, 211.278 hasta ise hatalı olmak üzere % 95,07'lik doğruluk yüzdesiyle sınıflandırılmıştır. CART algoritması ile yapılan sınıflandırma işleminde genel doğruluk değeri ise 6.772.313 hastanın 6.253.333 tanesi doğru sınıflandırılarak % 92,34 olarak hesaplanmıştır.

Tablo 11. Test verisi üzerinden elde edilen doğru sınıflandırma oranı tablosu

			Tahmin Değerleri			
			Penisilin			
			Kullananlar	Kullanmayanlar	Toplam	Doğru Sınıflandırma Oranı
Gerçek Değerler	Penisilin	Kullananlar	932.928	932.928	1.064.850	87,61%
		Kullanmayanlar	91.266	1.747.923	1.839.189	95,04%
Toplam			1.024.194	1.879.845	2.904.039	**92,31%**

CART ile oluşturulan modelin öngörü başarısını sınamak amacıyla hazırlanan test verisi kapsamına alınan 2.904.039 hastadan gerçekte 1.064.850'si penisilin grubu antibiyotik kullanırken geriye kalan 1.839.189'u farklı grup antibiyotik kullanmaktadır. Penisilin kullanan hastaların 932.928'i doğru, 131.922'si ise hatalı olmak üzere % 87,61'lik doğruluk yüzdesiyle sınıflandırılmıştır. Penisilin grubu antibiyotik kullanmayan 1.839.189 hastanın ise 1.747.923 tanesi yapılan sınıflandırma işleminde doğru, 91.266 hasta ise hatalı olmak üzere % 95,04'lük doğruluk yüzdesiyle sınıflandırılmıştır. Test verisi üzerinde yapılan sınıflandırma işleminde genel doğruluk değeri ise 2.904.039 hastanın 2.680.851 tanesi doğru sınıflandırılarak % 92,31 olarak hesaplanmıştır.

3.2. Lojistik Regresyon Analizi Uygulaması

Bu tekniğin uygulanmasında da öncelikle her değişken için hazırlanmış olan veri kümelerinin Clementine 12.0 yazılımına yüklenmesi gerçekleştirilmiştir. Değişkenlerin seçimi "select" düğümü ile yapılmış ve ilacın fiyatı gibi gruplandırılmış değişkenler "derive" düğümü ile formülüze edilmiştir. Yaş değişkeni CART analizi uygulamasından farklı olarak bu bölümde sürekli değişken olarak analize dahil edilmiş ve

hedef (bağımlı) değişken olan penisilinin tanımlanması yine "type" düğümü ile yapılmıştır. Son olarak "partition" düğümü ile verilerin eğitim (training) ve test (testing) verisi olarak ayrılması sağlanarak parametrelerin tahmin edildiği "logistic" düğümü ile model kurma aşamasına gelinmiştir.

Nihai modelin oluşturulmasında lojistik regresyon denkleminde hiçbir değişken yokken başlayan ve sonra her adımda bir değişkenin eklendiği ya da çıkarıldığı adım adım seçim (enter) yöntemi seçilmiştir. Ayrıca model değişkenler arası etkileşimlerin modelde yer almayacağı, sadece temel etkileri içerecek şekilde oluşturulmuştur. Modele alınacak değişkenlerin seçiminde ise anlamlılık düzeyi 0,05 $(\alpha = 0{,}05)$ olarak belirlenmiştir. Diğer model parametreleri için yazılımın varsayılan yani "default" değerleri kullanılmıştır.

Solunum sistemi hastalıkları için penisilin grubu antibiyotik kullanımı olarak belirlenen y bağımlı değişkeni; kullanma durumunda 1, kullanmama durumunda ise 2 olarak kodlanmış ve lojistik regresyon analizi ile test edilecek hipotez aşağıdaki gibi kurulmuştur.

H_0 : Tanı grubu, hastane grubu, ilacın fiyatı, cinsiyet ve yaş bağımsız değişkenleri ile oluşturulan model anlamsızdır.

H_1 : Tanı grubu, hastane grubu, ilacın fiyatı, cinsiyet ve yaş bağımsız değişkenleri ile oluşturulan model anlamlıdır.

Analizde bağımlı değişken olarak alınan Penisilin grubu antibiyotik kullanımı değişkenini etkileyen bağımsız değişkenlerin belirlenmesi aşamasında Wald testi kullanılmıştır. β parametreleri ile bu parametrelere ilişkin Wald istatistikleri, serbestlik dereceleri, anlamlılık düzeyleri, odds oranı değerleri ve bu değerlere ilişkin % 95'lik güven aralıkları tablo 3.5'de gösterilmiştir.

Tablo11'de yer alan modelde, değişkenlere ait katsayıların anlamlılığını test eden Wald istatistiğinin ($H_0: \beta_i = 0$ ve $H_1: \beta_i \neq 0$) anlamlılık düzeyi olan p değerlerine bakılıp her bir değişkenin anlamlılık testinin yapılması sonucunda, cinsiyet dışında kalan diğer bütün değişkenlerin bağımlı değişkenle istatistiksel olarak anlamlı bir ilişki içinde olduğu görülmektedir ($p < 0{,}05$, cinsiyet için $p = 0{,}103 > 0{,}05$).

Tablo 11. Lojistik regresyon analiz sonuçları

	β	S. Hata	Wald	Sd	P değeri	Odds Oranı	Odds Oranı İçin Güven Aralığı	
							Alt Sınır	Üst Sınır
Sabit	-0,953	0,017	3.088	1	0,000			
Yaş	-0,029	0,000	167.416	1	0,000	0,971	0,971	0,971
Tanı Grubu								
711=Akut tonsillit	1,205	0,014	7.207	1	0,000	3,337	3,246	3,432
712=Akut larenjit ve trakeit	-4,198	0,018	51.504	1	0,000	0,015	0,014	0,016
713=Akut obstrüktif larenjit ve epiglottit	1,233	0,034	1.311	1	0,000	0,291	0,273	0,312
714=Akut üst solunum yolu enfeksiyonları birden fazla olan	-0,795	0,014	3.063	1	0,000	0,452	0,439	0,464
715=Belirlenmiş influenza virüsüne bağlı İnfluenza	-4,846	0,041	13.697	1	0,000	0,008	0,007	0,009
718=Streptococcus pneumoniae'ye bağlı Pnömoni	-1,998	0,049	1.675	1	0,000	0,136	0,123	0,149
720=Bakteriyel pnömoni	0,069	0,022	10	1	0,001	1,072	1,027	1,119
722=Başka yerde sınıflanmamış hastalıklarda bulunan Pnömoni	-0,726	0,03	574	1	0,000	0,484	0,456	0,513
724=Akut bronşit	2,194	0,015	22.562	1	0,000	8,972	8,719	9,233
728=Kronik rinit, nazofarenjit ve farenjit	-8,55	0,035	61.268	1	0,000	0,000	0,000	0,000
744=Bronşiektazi (Referans)	.	.	.	.	.	.	.	.
Hastane Grubu								
Özel Hast.	3,498	0,009	141.375	1	0,000	33,035	32,438	33,642
SB 2. Bas. Hast.	-4,337	0,008	262.658	1	0,000	0,013	0,013	0,013
SB 3. Bas. Hast.	-1,575	0,008	35.806	1	0,000	0,207	0,204	0,21
Üniversite Hast.(Referans)	.	.	.	.	.	.	.	.
Cinsiyet								
Kadın	0,004	0,003	3	1	0,103	1,004	0,999	1,01
Erkek (Referans)	.	.	.	0	.	.	.	.
Fiyat Grubu								
Fiyat<=5 TL	3,384	0,007	232.994	1	0,000	29,501	29,099	29,909
5 TL<Fiyat<=25 TL	-0,445	0,006	6.098	1	0,000	0,641	0,634	0,648
Fiyat>25 TL (Referans)	.	.	.	.	.	.	.	.

Tablo 3.5'de elde edilen odds oranı değerlerine bakıldığında; 711 tanı kodlu akut tonsilit hastalarının 744 tanı kodlu bronşiektazi hastalarına göre 3,337 kat, 724 tanı kodlu akut bronşit hastalarının ise yine 744 tanı kodlu bronşiektazi hastalarına göre 8,972 kat daha fazla penisilin grubu antibiyotik kullandıkları görülmektedir. 714 tanı kodlu akut üst solunum yolu birden fazla olan hastalar ve 718 tanı kodlu streptococcus pneumoniae'ye bağlı Pnömoni hastaları ile penisilin kullanma durumu arasında anlamlı ancak negatif bir ilişki bulunduğundan, bu tanı gruplarına giren hastalarda penisilin kullanma oranının azaldığı gözlenmektedir. 744 tanı kodlu bronşiektazi hastaları referans olarak alındığında penisilin grubu antibiyotik kullanma oranı, 714 tanı kodlu akut üst solunum yolu birden fazla olan hastalarda 0,548 kat daha az (1-odds=0,452), 718 tanı kodlu streptococcus pneumoniae'ye bağlı Pnömoni hastalarında ise 0,864 (1-odds=0,136) kat daha az olmaktadır.

Hastane grubuna ilişkin odds oranları dikkate alındığında; üniversite hastanelerine göre özel hastanelerde solunum sistemi hastalıkları için penisilin grubu antibiyotik yazılma oranı 33,035 kat daha fazla, Sağlık Bakanlığı 2. basamak hastanelerinde ise 0,987 kat daha az olarak gerçekleşmektedir. Benzer şekilde ilacın fiyatına ilişkin odds oranlarına bakıldığında ise yüksek fiyatlı penisilin grubu antibiyotiklere göre, fiyatı 5 TL'den az olan düşük fiyatlı penisilin grubu antibiyotiklerin 29,501 kat daha fazla yazıldığı görülmektedir.

Diğer taraftan yaş değişkeninin katsayısının negatif işaretli olması da odds oranıyla arasında negatif bir ilişki olduğunu gösterdiğinden, solunum sistemi rahatsızlığı olan bireylerin yaşı azaldıkça penisilin grubu antibiyotik kullanma olasılıklarının her bir birim için 0,971 kat arttığı sonucunu ortaya çıkarmaktadır.

Katsayıların anlamlılığının test edilmesinden sonra, çizelge 4.10'da modele ilişkin genel anlamlılığının test edildiği model uygunluk tablosu sonucunda bulunan Ki-Kare değeri de $\alpha = 0{,}05$ düzeyinde anlamlı bulunduğundan ve modelin anlamsız olduğuna dair kurulan H_0 hipotezi reddedildiğinden oluşturulan lojistik regresyon modelinin verilere uygun olduğunu söylemek mümkündür.

Tablo 12. Modelin anlamlılığına ilişkin test sonucu

	Model Uygunluk Kriteri	Olabilirlik Oran Testi		
	-2 Log Olabilirlik	Ki-Kare	Serbestlik Derecesi	P Değeri
Sadece sabit terimin olduğu model	7054860,506			
Son model	1891272,644	5163587,862	17	0,000

Tablo 13'de sunulan doğru sınıflandırma oranı modelin uyum iyiliğini test etmeye yönelik diğer bir ölçüt olarak kullanılmaktadır. Sınıflandırma tablosunda bağımlı değişkenin gerçek değerleri ile tahmin değerleri çaprazlanmakta ve hesaplanan tahmin değerleri için 0,5 eşik değerinden küçük olanlara "1" değeri, büyük olanlara "2" değeri atanmaktadır. Bu şekilde yapılan atama değerleri sonucunda penisilin hedef (bağımlı) değişkeni üzerinde gerçekleştirilen lojistik regresyon analizi sonucunda kurulan modelin sınıflandırma başarısı % 91,37 olarak hesaplanmıştır. Ayrıca modelin öngörü başarısını sınamak amacıyla hazırlanan test verisi üzerindeki sınıflandırma başarısı da % 91,36 olarak hesaplanmış ve Tablo 14'te gösterilmiştir.

Tablo 13. Lojistik regresyon analizi sonucu elde edilen doğru sınıflandırma oranı tablosu

			Tahmin Değerleri			
			Penisilin			
			Kullananlar (1)	Kullanmayanlar (2)	Toplam	Doğru Sınıflandırma Oranı
Gerçek Değerler	Penisilin	Kullananlar (1)	2.069.709	414.643	2.484.352	83,31%
		Kullanmayanlar (2)	169.600	4.118.361	4.287.961	96,04%
Toplam			2.239.309	4.533.004	6.772.313	**91,37%**

Tablo 3.7'de görüldüğü gibi, uygulama kapsamındaki 6.772.313 hastadan gerçekte 2.484.352'si penisilin grubu antibiyotik kullanırken geriye kalan 4.287.961 kişi penisilin grubu antibiyotik kullanmamaktadır. Penisilin kullanan hastaların 2.069.709'u doğru, 414.643'ü ise hatalı olmak üzere % 83,31'lik doğruluk yüzdesiyle sınıflandırılmıştır. Penisilin grubu antibiyotik kullanmayan 4.287.961 hastanın ise 4.118.361 tanesi lojistik regresyon ile yapılan sınıflandırma işleminde doğru, 169.600 hasta ise hatalı olmak üzere % 96,04'lük doğruluk yüzdesiyle sınıflandırılmıştır. Lojistik regresyon ile yapılan sınıflandırma işleminde genel doğruluk değeri ise 6.772.313 hastanın 6.188.070 tanesi doğru sınıflandırılarak % 91,37 olarak hesaplanmıştır.

Tablo 14. Test verisi üzerinden elde edilen doğru sınıflandırma oranı tablosu

			Tahmin Değerleri			
			Penisilin			
			Kullananlar (1)	Kullanmayanlar (2)	Toplam	Doğru Sınıflandırma Oranı
Gerçek Değerler	Penisilin	Kullananlar (1)	886.892	177.958	1.064.850	83,29%
		Kullanmayanlar (2)	73.068	1.766.121	1.839.189	96,03%
Toplam			959.960	1.944.079	1.944.079	**91,36%**

Lojistik regresyon ile oluşturulan modelin öngörü başarısını sınamak amacıyla hazırlanan test verisi kapsamına alınan 2.904.039 hastadan gerçekte 1.064.850'si penisilin grubu antibiyotik kullanırken geriye kalan 1.839.189'u farklı grup antibiyotik kullanmaktadır. Penisilin kullanan hastaların 886.892'si doğru, 177.958'i ise hatalı olmak üzere % 83,29'luk doğruluk yüzdesiyle sınıflandırılmıştır. Penisilin grubu antibiyotik kullanmayan 1.839.189 hastanın ise 1.766.121 tanesi yapılan sınıflandırma işleminde doğru, 73.068 hasta ise hatalı olmak

üzere % 96,03'lük doğruluk yüzdesiyle sınıflandırılmıştır. Test verisi üzerinde yapılan sınıflandırma işleminde genel doğruluk değeri ise 2.904.039 hastanın 2.653.013 tanesi doğru sınıflandırılarak % 91,36 olarak hesaplanmıştır.

Bu doğrultuda oluşturulan uygun modelde P_i; solunum sistemi hastalıkları için reçeteye yazılan ilacın penisilin grubu antibiyotik olma olasılığı, $1-P_i$; penisilin grubu antibiyotik olmama olasılığı olmak üzere; $P_i = \frac{1}{1+e^{-z}}$ $(z = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k)$ lojistik dağılım fonksiyonundan ve tablo 3.5'de yer alan lojit katsayıları β 'lardan yararlanarak değişkenlerin çeşitli düzeylerine yönelik olasılıklar hesaplanabilmektedir. Örneğin, 20 yaşında 711 tanı kodlu akut tonsilit hastası için özel hastanede tedavi sonucunda reçeteye yazılan, fiyatı 5 TL'den az olan bir antibiyotiğin penisilin grubu antibiyotik olma olasılığı Tablo 15'te gösterildiği gibi hesaplanabilmektedir.

Tablo 15. Lojistik regresyon modeli ile olasılık tahmini

Olguların özellikleri	**Katsayı değerleri (β)**	
Sabit	-0,953	$P_i = \frac{1}{1+e^{-6,554}} = 0,998$
Yaş (20)	(-0,029)*20= -0,58	
Tanı Grubu (711)	1,205	
Hastane Grubu (Özel Hast.)	3,498	
Fiyat Grubu (Fiyat<=5 TL)	3,384	
z=	6,554	

Oluşturulan lojistik regresyon modeli yardımıyla tablo 3.9'da görüldüğü üzere, söz konusu özellikleri taşıyan bir hasta için reçeteye yazılan bir antibiyotiğin penisilin grubu antibiyotik olma olasılığı % 99,8 olarak hesaplanmıştır.

3.3. CART ve Lojistik Regresyon Analizlerinin Karşılaştırılması

Aynı donanım ve yazılım olanaklarının kullanıldığı uygulama bölümünde, CART ve lojistik regresyon analizi sonucunda elde edilen modellerin karşılaştırma kriteri her bir analizin işlem süresi, sınıflandırma ve öngörü başarısı olarak dikkate alınmıştır. Tablo 16'da her bir modelin üretilmesi için gerekli işlem süresi, modele ilişkin sınıflandırma ve öngörü başarısı yer almaktadır.

Tablo 16. Analizleri karşılaştırma kriterleri

	İşlem Süresi	Sınıflandırma Başarısı	Öngörü Başarısı
CART	8 dak. 27 sn.	% 92,34	% 92,31
Lojistik Regresyon	14 dak. 58 sn.	% 91,37	% 91,36

Veri madenciliği çok fazla miktardaki verilerin analizine dayandığından, kullanılan tekniklerin veri madenciliği açısından değerlendirilmesinde, tekniklerin işlem sürelerinin kısa olması ve sınıflandırma ile öngörü başarısının yüksek olması modellerin güvenilirliği konusunda oldukça önemli bir yere sahiptir. Ancak tekniklerin işlem süreleri ne kadar kısa olursa olsun eğer sınıflandırma başarısı düşük kalıyorsa bu modellere güvenilmesi söz konusu olmayacağından, modellerin karşılaştırılmasında sınıflandırma başarısını dikkate almak birinci öncelik olarak sayılabilir.

Tablo 3.10'da görüldüğü üzere işlem süresi açısından en hızlı analiz CART olmuştur. Bu nedenle hızlı karar alma ihtiyacına gerek duyulan süreçlerde CART analizinin kullanılmasının daha uygun olacağı söylenebilir.

Yine aynı tablo 3.10 incelendiğinde, oluşturulan modellerin sınıflandırma başarısı açısından farklılık göstermeyip çok yakın sonuçlar verdiği gözlenmiştir. CART analizi sonucunda kurulan modelin sınıflandırma başarısı % 92,34, lojistik regresyon analizi sonucunda kurulan modelin sınıflandırma başarısı da % 91,37 olarak hesaplanmıştır.

Eğitim verisi üzerindeki yüksek sınıflandırma başarısının öngörü başarısının sınandığı test verisi üzerinde tekrar etmemesi durumunda model güvenilir olmaktan uzak olacağından, modelleri karşılaştırmada diğer önemli kriter öngörü başarısı olarak dikkate alınmıştır. CART analizi sonucunda kurulan modelin öngörü başarısı % 92,31 olarak bulunmuşken, lojistik regresyon analizi sonucunda kurulan modelin öngörü başarısı % 91,36 olarak bulunmuştur.

4. Sonuç ve Öneriler

Son yıllarda yaygın olarak kullanılmaya başlanan ve büyük veri kümeleri içinde saklı durumda bulunan işlenmemiş veriyi anlaşılabilir ve yorumlanabilir hale getiren işlemlerden biri veri madenciliğidir. Veri madenciliğinin temel amacı, bilgisayar sistemleri ile üretilen kendi başına bir anlam ifade etmeyen verilerin, uygun programlar çerçevesinde derlenerek, bu verilerden bilgi çıkarılması ve geçmiş faaliyetlerin analizini göz önünde bulundurarak gelecekteki davranışların tahminine yönelik karar verme modelleri yaratmaktır. Bu çalışmada söz konusu amaca uygun olarak ilaç provizyon sistemi üzerinden alınan solunum sistemi hastalıkları için kullanılan antibiyotik verileri içerisinden, penisilin gurubu antibiyotikleri sınıflandırmak için profesyonel bir program kullanılarak analizler yapılmıştır.

Genellikle verilerin sınıflandırılmasında bugüne kadar daha çok kümeleme, diskriminant analizi ve lojistik regresyon analizi gibi çok değişkenli istatistik tekniklerden yararlanıldığı görülmektedir. Bu tekniklere göre daha yeni olan karar ağacı algoritmaları da ülkemizde yaygın olarak kullanılmaya başlanmıştır. Karar ağacı algoritmalarının en önemli avantajı, parametrik olmayan yöntemler arasında olması nedeniyle diğer çok değişkenli tekniklerde sağlanması gereken

istatistiksel varsayımların olmamasıdır. Bu nedenle daha yeni bir yöntem olan karar ağacı algoritmalarıyla daha klasikleşmiş bir metot olan lojistik regresyonun sınıflama özelliklerini karşılaştırmak için teknik alt yapısı oldukça zengin olan Clementine 12.0 programının, modelleme modülünde yer alan CART ve lojistik regresyon veriler üzerinde denenmiş ve iki yöntemin sınıflama özellikleri dikkate alınmıştır.

Uygulamaya konu olan veri kümesi başlangıçta 18.931.000 hastanın 12 farklı değişkene ilişkin değerlerini içerirken, veri hazırlama aşamasında yürütülen işlemlerle 6.772.313 hastanın 6 değişkene ilişkin değerlerini içerecek şekilde biçimlendirilmiştir.

CART analizi uygulamasında penisilin kullanımı için ilk sınıflama hastane grubuna göre oluşmuş ve penisilin grubu antibiyotiklerin % 16'sının Sağlık Bakanlığı 2. ve 3. basamak hastanelerinde % 84'ünün ise özel hastaneler ve üniversite hastanelerinde reçetelendirildiği sonucu bulunmuştur.

Ayrıca bu temel sınıflamanın altında 2. basamak Sağlık Bakanlığı hastanelerinde, fiyatı 5 TL'nin altında akut bronşit, streptococcus pneumoniae'ye bağlı Pnömoni ve bronşiektazi gibi solunum yolu hastalıkları için yazılan antibiyotikler içinden, penisilin grubu antibiyotik kullanan hastaların çoğunun 15 yaş ve altı hastalar olduğu sonucuna ulaşılmıştır. Bu durum ise 0–15 yaş arası okul çağında olan nüfusun beta mikrobu gibi çeşitli mikroplar nedeniyle sıklıkla hastalanması ve hastalığın tedavisinde penisilin grubu antibiyotiklerin kullanımının yaygın olmasıyla açıklanabilmektedir.

Modeldeki risk faktörleri için tahmin edilen odds oranları yardımıyla yorumlamanın yapıldığı lojistik regresyon analizi uygulamasında; hastane grubuna ilişkin odds oranları dikkate alındığında, üniversite hastanelerine göre özel hastanelerde solunum sistemi hastalıkları için penisilin grubu antibiyotik yazılma oranının 33 kat daha fazla olduğu sonucu bulunmuştur. Benzer şekilde ilacın fiyatına ilişkin odds oranları değerlendirildiğinde ise yüksek fiyatlı penisilin grubu antibiyotiklere göre, fiyatı 5 TL'den az olan düşük fiyatlı penisilin grubu antibiyotiklerin 29 kat daha fazla yazıldığı sonucu ortaya çıkmıştır.

Söz konusu yöntemlerle yapılan analizlerde tüm verinin % 70'i model oluşturmak amacı ile eğitim verisi, geri kalan % 30'u ise sınıflama kurallarının doğruluğunu test etmek amacıyla test verisi olarak kullanılmıştır. Bu doğrultuda sınıflandırma modelinin öğrenme kümesi üzerinde geliştirilmesi ve test verisinden oluşan sınama kümesi üzerinde öngörü başarılarının sınanması sağlanmıştır.

Oluşturulan modellerin, geliştirildikleri veri kümesi üzerinde sınıflandırma başarısının bir ölçüsü olan doğru sınıflandırma oranları açısından farklılık göstermeyip çok yakın sonuçlar verdiği gözlenmiştir. Penisilin hedef (bağımlı) değişkeni üzerinde gerçekleştirilen CART analizi sonucunda kurulan modelin sınıflandırma başarısı % 92,34, modelin öngörü başarısını sınamak amacıyla hazırlanan test verisi üzerindeki sınıflandırma başarısı da % 92,31 olarak hesaplanmıştır. Benzer şekilde aynı hedef değişken üzerinde gerçekleştirilen lojistik regresyon analizi sonucunda kurulan modelin sınıflandırma başarısı % 91,37, test verisi üzerindeki öngörü başarısı da % 91,36 olarak hesaplanmıştır.

Her iki modelin % 90'ın üzerinde sınıflandırma başarısı gösterdiği dikkat çekerken, CART analizinin daha yüksek sınıflandırma başarısına sahip olduğu tespit edilmiştir. Bu noktadan hareketle CART ve lojistik regresyon analizi ile yapılan çalışmalarda hata riskini en aza indirmek amacıyla CART analizi tekniğinin kullanılması daha uygun bulunmuştur. Bununla birlikte çalışma kapsamında oldukça fazla sayıda veri ile gerçekleştirilen uygulamada, oluşturulan modelin sınıflandırma başarısı ile bu modele ilişkin test verisi üzerinde gerçekleştirilen öngörü başarısının birbirine paralel bir şekilde yüksek çıkması, yapılan analizlerde veri kalitesinin de önemli bir rol oynadığı gerçeğini ortaya koymuştur.

Bu çalışma mevcut veriler ile yapılan analizlere bakılarak aynı özellikte verilerle yapılacak diğer çalışmalarda da genel geçer kurallar tanımlanmasında kullanılabilir. Böylece analizde kullanılan veriler ışığında, aynı türde yeni veriler ortaya çıktığında bu verilerin hangi sınıfta yer alması gerektiğine ilişkin ileriye yönelik tahminler kolaylıkla yapılacaktır.

KAYNAKÇA

Ahmad, I., "Data Warehousing in Construction Organizations", Construction Congress VI, Florida, 194–203 (2000).

Akpınar H., "Veri Tabanlarında bilgi keşfi ve Veri Madenciliği", İ.Ü. İşletme Fakültesi Dergisi, 29 (1), 1-22 (2000).

Allison, D. P., "Logistic Regression Using The SAS System 2nd ed.", SAS Institute, (2000).

Ayık Y. Z., Özdemir A., Yavuz U., "Lise Türü Ve Lise Mezuniyet Başarısının, Kazanılan Fakülte İle İlişkisinin Veri Madenciliği Tekniği İle Analizi", Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 10(2): 441-454 (2007).

Berry, M. J., Linoff, G. S., "Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management 2nd ed.", Wiley, USA, (2004).

Bigus, J. P., "Data Mining With Neural Networks: Solving Business Problems from Application Development to Decision Support", McGraw Hill, (1996).

Collet, D., "Modelling Binary Data", Chapman & Hall, Florida, (2003).

Deconinck, E., Hancock, T., Coomans, D., Massart, D.L., Heyden, Y.V., "Classification of drugs in absorption classes using the classification and regression trees (CART) methodology", Journal of Pharmaceutical and Biomedical Analysis, 39 : 91–103 (2005).

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., "The KDD Process for Extracting Useful Knowledge From Volumes of Data", Communications of the ACM, 39 (11): 27-34 (1996).

Hosmer, D. W., Lemeshow, S., "Applied Logistic Regression", John Wiley & Sons, New York, 5-50 (1989).

Kecman, V., "Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models", The MIT Pres, Cambridge, MA, 1-4 (2001).

Kim, M., "Two-stage Logistic Regression Model", Expert Systems with Applications, 36: 6727–6734 (2009).

Kleinbaum,G., D., "A Self-learning Text Logistic Regression", Springer, Atlanta, (1994).

Köktürk, F., Ankaralı, H., Sümbüloğlu, V., "Veri Madenciliği Yöntemlerine Genel Bakış", Türkiye Klinikleri Journal of Biostatistics, 1 (1): 20-25 (2009).

Kurt, I., Ture, M., Kurum, A. T., "Comparing Performances of Logistic Regression, Classification and Regression Tree, and Neural Networks for Predicting Coronary Artery Disease", Expert Systems with Applications, 34 : 366–374 (2008).

Masseglia, F., Poncelet, P., Teisseire, M., "Using Data Mining Techniques on Web Access Logs to Dynamically Improve Hypertext Structure", ACM Sigweb Newsletter, 8 (3): 1-19 (1999).

Özkan, Y., "Veri Madenciliği Yöntemleri", Papatya Yayıncılık Eğitim, İstanbul, 106-113 (2008).

Pehlivan, G., "Chaid Analizi ve Bir Uygulama", Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, 17 (2006).

Silahtaroğlu, G., "Kavram ve Algoritmalarıyla Temel Veri Madenciliği", Papatya Yayıncılık Eğitim, İstanbul, 33, 45-47, 58 (2008).

Tatlıdil, H., "Uygulamalı Çok Değişkenli İstatistiksel Analiz", Cem Web Ofset, Ankara, (1996).

Temel, G. O., Çamdeviren, H., Akkuş, Z., "Sınıflama Ağaçları Yardımıyla Restless Legs Syndrome (RLS) Hastalarına Tanı Koyma", İnönü Üniversitesi Tıp Fakültesi Dergisi, 12 (2): 111-117 (2005).

Teng, J., Lin, K., Ho, B., "Application of Classification Tree and Logistic Regression for The Management and Health İntervention Plans in A Community-Based Study", Journal of Evaluation in Clinical Practice, 13 : 741-748 (2007)

Thomas, Lyn. C., "A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Consumer", International Journal of Forecasting, 16 (2): 149–172 (2000).

Zhou, Z., "Three Perspectives of Data Mining", Artificial Intelligence, 143 (1): 139-146 (2003).