LOJİSTİK REGRESYON VE CART ANALİZİ TEKNİKLERİYLE SOSYAL GÜVENLİK KURUMU İLAÇ PROVİZYON SİSTEMİ VERİLERİ ÜZERİNDE BİR UYGULAMA

Zeynep Burcu KIRAN

YÜKSEK LİSANS TEZİ
İSTATİSTİK

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

MAYIS 2010
ANKARA

Zeynep Burcu KIRAN tarafından hazırlanan LOJİSTİK REGRESYON VE CART ANALİZİ TEKNİKLERİYLE SOSYAL GÜVENLİK KURUMU İLAÇ PROVİZYON SİSTEMİ VERİLERİ ÜZERİNDE BİR UYGULAMA adlı bu tezin Yüksek Lisans tezi olarak uygun olduğunu onaylarım.

Yrd. Doç. Dr. Necla GÜNDÜZ TEKİN ……..
Tez Danışmanı, İstatistik Anabilim Dalı

Bu çalışma, jürimiz tarafından oy birliği ile İstatistik Anabilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir.

Prof. Dr. Cevriye GENCER …….
Endüstri Mühendisliği Anabilim Dalı, G.Ü.

Yrd. Doç. Dr. Necla GÜNDÜZ TEKİN ……
İstatistik Anabilim Dalı, G.Ü.

Prof. Dr. Semra ORAL ERBAŞ …….
İstatistik Anabilim Dalı, G.Ü.

Tarih: 24./ 05 / 2010

Bu tez ile G.Ü. Fen Bilimleri Enstitüsü Yönetim Kurulu Yüksek Lisans derecesini onamıştır.

Prof. Dr. Bilal TOKLU …….
Fen Bilimleri Enstitüsü Müdürü

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

Zeynep Burcu KIRAN

LOJİSTİK REGRESYON VE CART ANALİZİ TEKNİKLERİYLE SOSYAL GÜVENLİK KURUMU İLAÇ PROVİZYON SİSTEMİ VERİLERİ ÜZERİNDE BİR UYGULAMA

(Yüksek Lisans Tezi)

Zeynep Burcu KIRAN

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
Mayıs 2010

ÖZET

Veri madenciliği teknikleri içerisinde en yaygın kullanıma sahip olan yöntemlerden 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 özelliklerinin karşılaştırılması amaçlanmaktadır. Bu kapsamda 2007–2009 yılları arası Sosyal Güvenlik Kurumu ilaç provizyon sisteminden alınan solunum sistemi hastalıklarında reçeteye yazılan antibiyotikler içerisinden, penisilin gurubu 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.

Bilim Kodu : 205.1.066
Anahtar Kelimeler : Veri madenciliği, CART, Lojistik Regresyon
Sayfa Adedi : 94
Tez Yöneticisi : Yrd. Doç. Dr. Necla GÜNDÜZ TEKİN

AN APPLICATION ON PHARMACY PROVISION SYSTEM DATA OF SOCIAL SECURITY INSTITUTION BY LOGISTIC REGRESSION AND CART ANALYSIS TECHNICS

(M.Sc.Thesis)

Zeynep Burcu KIRAN

GAZİ UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY
May 2010

ABSTRACT

One of the most widely used method of data-mining technics is classification and regression models. In this study, it was aimed at comparing the classification features of logistic regression, with Classification and Regression Trees (CART) algorithm which is one of the most widely used decision tree algorithm in data mining methods. In this context, an application are made with the aim of determining patient profile using penicillin group of antibiotics in antibiotics filling prescription for respiratory system diseases with the years of 2007–2009 data ensuring from pharmacy provision system of Social Security Institution. For the data set included, CART analysis was found to have a beter correct classification ratio than the logistic regression analysis.

Science Code : 205.1.066
Key Words : Data mining, CART, Logistic regression
Page Number : 94
Adviser : Assistant Prof. Dr Necla GÜNDÜZ TEKİN

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren Hocam Yrd. Doç. Dr. Necla GÜNDÜZ TEKİN'e, çalışmamda uygulama bölümü için verileri sağlayan ve analizleri yapmamda yardımlarını esirgemeyen Sosyal Güvenlik Kurumu çalışanlarından Dr. Rasim KÖSELERLİ'ye teşekkür ederim. Ayrıca aynı dönemde tez yazdığımız ve birbirimize destek olduğumuz arkadaşlarım Tuna GENÇ, Nadide YİĞİTELİ ve Veli AĞÖREN ile yazım sürecinde beni teşvik eden Altuğ GÜNER'e her koşulda yanımda oldukları için teşekkür ederim. Bugüne kadar manevi destekleriyle hep yanımda olan canım aileme de teşekkürü borç bilirim.

İÇİNDEKİLER

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

Sayfa

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

KISALTMALAR

Bu çalışmada kullanılmış bazı kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklama
CART	Sınıflama ve Regresyon Ağaçları (Classification and Regression Trees)
CHAID	Otomatik Ki-Kare Etkileşim Belirleme (Chi-Squared Automatic Interaction Detector)
EKK	En Küçük Kareler
MARS	Çok Değişkenli Uyumlu Regresyon Uzanımları (Multivariate Adaptive Regression Splines)
QUEST	Hızlı, Yansız, Etkin İstatistiksel Ağaç (Quick, Unbiased, Efficient Statistical Tree)
ÇGOHK	Çapraz Geçerlilik Ortalama Hata Karekökü

1. GİRİŞ

Bilimsel çalışmalarda kullanılan verilerin analizinde 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. Özellikle tıp ve biyoloji alanında yapılan çalışmalarda, veri setleri oldukça karmaşık bir yapı teşkil etmektedir. Bu noktada veri madenciliği sağlık ve tıp alanındaki büyük veri tabanlarından faydalı bilgileri ortaya çıkararak hem tıp hem de 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 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 [1]. 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 [2].

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, veri madenciliği metotları içerisinde sınıflama ve regresyon modellerinden en çok kullanılan karar ağaçları ile lojistik regresyonun sınıflama özellikleri karşılaştırılarak gerçek bir veri seti üzerinde uygulama yapılmış ve söz konusu iki yöntemin başarısını göstermek amaçlanmıştır.

Bu nedenle çalışmanın ikinci bölümünde, öncelikle veri madenciliği ve uygulama alanları hakkında genel bilgiler verilerek veri madenciliği modelleri tanıtılmıştır. Sınıflama ve regresyon modellerinden karar ağaçları ve karar ağaçlarında en çok kullanılan analizlerin yapısı ve algoritmaları genel olarak tanımlanmıştır.

Üçüncü bölümde ise, çalışmanın diğer konusu olan lojistik regresyon ile ilgili son yıllarda yapılan çalışmalar, doğrusal olasılık modeli ve lojistik fonksiyonla birlikte lojistik regresyon modeli ve varsayımlarının neler olduğuna değinilmiştir. Sınıflandırma analizlerinin sıklıkla tercih edilenlerinden biri olan ve birçok konuda uygulama alanı bulunan lojistik regresyonun parametrelerinin anlamlılık testleri ve modelin uyum iyiliği de detaylı bir şekilde incelenmiştir.

Uygulamanın yapıldığı dördüncü bölümde, lojistik regresyon ile karar ağacı algoritmalarından en çok kullanılan Classification and Regression Trees (CART) algoritmasının, Sosyal Güvenlik Kurumu 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 beşinci bölüm olan sonuç ve öneriler ile sona erdirilmiştir.

2. VERİ MADENCİLİĞİ

2.1. Veri Madenciliğine Genel Bir Bakış

Bilişim teknolojilerinde yaşanan hızlı gelişmeler ve bilgisayarların bilgi saklama kapasitelerinin artmasıyla birlikte depolanan veriler çok daha büyük boyutlara ulaşmaktadır. Yaşanan bu gelişmeler doğrultusunda bilgi miktarı hızla artarak karşımıza çıkmakta ve bilgi kaydı yapılan alanların sayısı da giderek artmaktadır. Dünyadaki bilgi miktarının her 20 ayda bir ikiye katlandığı tahmin edilmektedir [3]. Veri tabanı sistemlerinin artan kullanımı ve sakladıkları veri miktarlarındaki böylesine büyük artış organizasyonları elde toplanan bu verilerden nasıl faydalanılabileceği problemi ile karşı karşıya bırakmıştır [4]. 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 [5]. Veri madenciliği bir veri kümesi içerisinde keşfedilmemiş örüntüleri bulmayı hedefleyen teknikler bütününü ifade etmektedir. Veri madenciliğinin amacı, geçmiş faaliyetlerin analizini göz önünde bulundurarak gelecekteki davranışların tahminine yönelik karar verme modelleri yaratmaktır [6].

Veri madenciliğine ilişkin yapılan tanımlardan bir tanesi, verilerden daha önceden bilinmeyen ve muhtemelen faydalı enformasyonun monoton olmayan bir süreçte çıkartılması işlemidir [7]. Diğer bir tanımlama ise büyük

miktarlardaki verinin içinden geleceğin tahmin edilmesinde yardımcı olacak anlamlı ve yararlı bağlantı ve kuralların bilgisayar programları aracılığıyla aranması ve analizidir. Ayrıca Veri Madenciliği, çok büyük miktardaki verilerin içindeki ilişkileri inceleyerek aralarındaki bağlantıyı bulmaya yardımcı olan veri analizi tekniğidir [8].

Sınıflama problemleri ve örüntü tanıma (pattern recognition) üzerinde yoğunlaşan yapay zeka 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 etkilenmektedir [9].

2.2. Veri Madenciliğinin Uygulama Alanları

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 tıp ve ekonomi gibi bilim dalları için de önemli kılmaktadır. Bilginin bu denli değerli olduğu çağımızda bilgiye ulaşmak için katedilen yolda veri madenciliği oldukça önemli bir safhadır.

Veri madenciliği astronomi, biyoloji, bankacılık, finans, pazarlama, sigorta, tıp ve birçok başka alanda başarılı bir şekilde kullanılmaktadır. Veri madenciliğinin yaygın olarak kullanıldığı alanlardan bazıları ana başlıklar halinde aşağıda özetlenmiştir.

2.2.1. Bankacılık ve sigortacılık

Kredi kartı ve internet üzerinden yapılan işlemlerdeki dolandırıcılıkların tespitinde, kredi taleplerinin değerlendirilmesinde, belli gruplara farklı çeşit hizmet sağlamak adına kredi kartı harcamalarına göre müşteri gruplarının belirlenmesinde, müşteri kazanma ve mevcut müşterileri elde tutma analizlerinde, risk yönetimi ve risk analizinde, iflas ve kaza sigortası gibi yüksek meblağlı taleplerdeki örüntülerin incelenmesiyle sahtekarlıkların azaltılması, poliçe fiyatlarının ve yeni poliçe talep edecek müşterilerin tahmin edilmesinde kullanılmaktadır.

2.2.2. Pazarlama

Müşterilerin satın alma örüntülerinin belirlenmesinde, müşteri ilişkileri yönetimi ve müşteri değerlendirmesinde, satış tahminlerinde, müşterilerin hangi malları birlikte satın alma eğiliminde olduğunu açıklayan pazar sepeti analizlerinde, mağaza yerleşim optimizasyonu ve çeşitli pazarlama kampanyalarında kullanılmaktadır.

2.2.3. Telekomünikasyon

Hatların yoğunluk tahmininde, müşteri gruplarına göre cazip fiyatlandırma programı geliştirme ile kalite ve hizmet geliştirme analizlerinde kullanılmaktadır.

2.2.4. Endüstri ve mühendislik

Lojistik, üretim süreçlerinin optimizasyonu, kalite kontrol analizleri ve ampirik veriler üzerinde modeller kurularak bilimsel ve teknik problemlerin çözümlenmesinde kullanılabilmektedir.

2.2.5. Sağlık ve ilaç

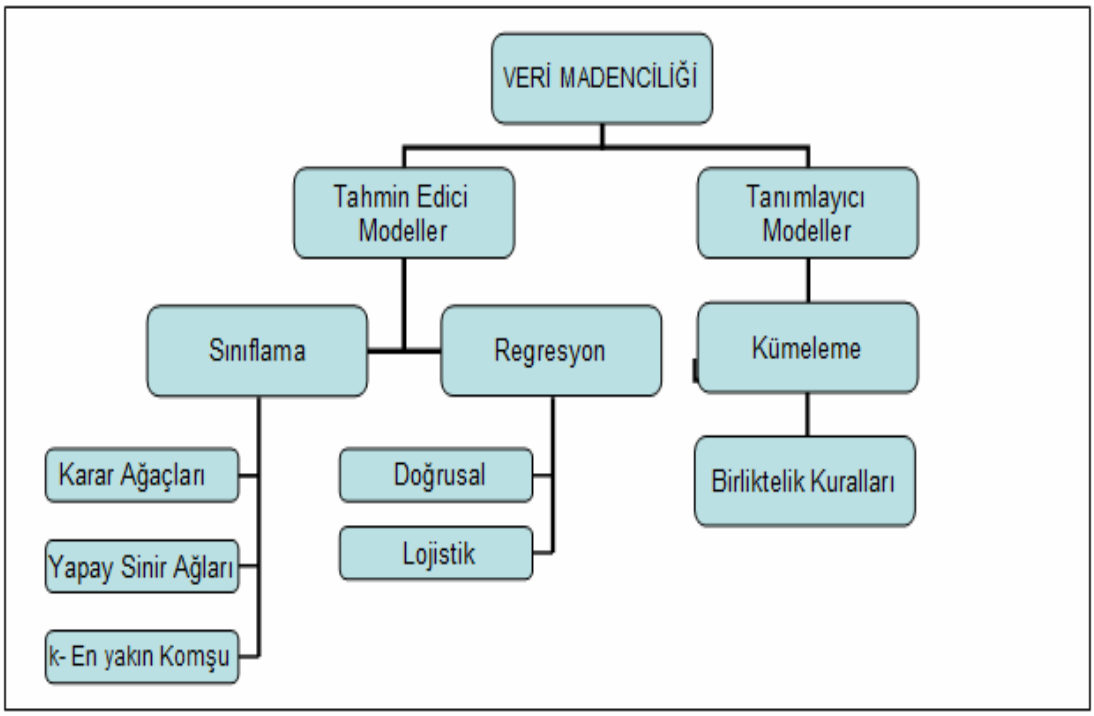
İnsanların ortalama ömürlerinin uzaması ve sağlık sektöründeki gelişmeler beraberinde bazı sorunları da getirmiştir. Örneğin pek çok insan kalp, astım ve diyabet gibi kronik hastalıklarla yaşamak zorundadır. Bu gibi hastalıkların hem tıbbi yönden hem de hastane maliyetleri açısından ele alınarak doğru yönetilmesi gerekmektedir. Bilgi sistemlerinin klasik yöntemlerinin yetersiz kaldığı bu analizlerde yardıma veri madenciliği teknikleri yetişmektedir.

Sağlık alanında yapılan birçok veri madenciliği araştırmasında elektronik tıbbi kayıtlar ve idari işlem belgelerine ait veriler kullanılmaktadır. Söz konusu bu veriler doğrultusunda yapılan veri madenciliği; tıbbi teşhis ve tanı koymada, belirli bir hastalığa sahip kişilerin ortak özelliklerinin tahmin edilmesinde, uygun tedavi sürecinin belirlenmesinde, hastane maliyetlerinin ve test sonuçlarının tahmin edilmesinde ve ürün geliştirmede kullanılmaktadır.

Tıbbın önemli araştırma konularından biri olan ilaçlar için de, ilaç etkileri analizi, ilaç üretimi ve geliştirilmesi gibi konularda da veri madenciliği tekniklerinden sıklıkla faydalanılmaktadır. İlaç alanında yapılan antipsikotik ilaçların kalp kası hastalıkları üzerine etkisi, ilaç yan etkilerinin tanımlanması gibi çeşitli çalışmalar veri madenciliği uygulamalarına örnek olarak verilebilir [10,11].

2.3. Veri Madenciliği Modelleri

Veri Madenciliğinde kullanılan modeller, temel olarak şekil 2.1'de görüldüğü üzere tahmin edici (predictive) ve tanımlayıcı (descriptive) olmak üzere iki ana başlık altında incelenmektedir [12].



Şekil 2.1. Veri madenciliği modelleri

Tahmin edici modellerin amacı, verilerden hareket ederek bir model geliştirmek ve kurulan bu model yardımıyla sonuçları bilinmeyen veri kümelerinin sonuç değerlerini tahmin etmektir. Eğer tahmin edilecek değişken sürekli bir değişkense tahmin problemi regresyon, kategorik bir değişkense sınıflama problemi olarak nitelendirilmektedir [3]. Tanımlayıcı modellerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır.

Veri Madenciliği modelleri fonksiyonlarına göre ise;

- Sınıflama (Classification) ve Regresyon,
- Kümeleme (Clustering),
- Birliktelik kuralları (Association Rules)

şeklinde sınıflandırılmaktadır. Şekil 2.1'de gösterilen veri madenciliği modellerinden sınıflama ve regresyon modelleri tahmin edici, kümeleme, birliktelik kuralları ve ardışık zamanlı örüntüler ise tanımlayıcı modellerdir [8].

Veri madenciliği teknikleri içerisinde en yaygın kullanıma sahip olan, büyük veri kümelerini sınıflandırarak önemli veri sınıflarını ortaya koyan veya gelecek veri eğilimlerini tahmin etmede faydalanılan yöntemlerden bir tanesi sınıflama ve regresyon modelleridir. Tezin konusuyla bağlantısından dolayı söz konusu bu modeller bölüm 2.3.1'de detaylı olarak anlatılacaktır.

Veri madenciliğinin en önemli alanlarından biri olan kümeleme, nesneleri birbirlerine olan benzerliklerine göre gruplara ayırmaktadır. Yani kümeleme modellerinde amaç, küme üyelerinin birbirlerine çok benzediği ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesidir. Böylece nesneler, örneklenen kitle özelliklerini iyi yansıtan etkili bir temsil gücüne sahip olacaktır.

Veri madenciliği araştırmalarında özel bir uygulama alanı olan ve ticaret, mühendislik, fen ve sağlık sektörlerinin içinde bulunduğu birçok alanda da uygulanan birliktelik kuralları; aynı işlem içinde çoğunlukla beraber görülen nesneleri içeren kurallar olup büyük veri kümeleri arasında birliktelik ilişkileri bulurlar. Bu anlamda birliktelik kuralı belirli türlerdeki veri ilişkilerini tanımlayan bir modeldir. Herhangi bir veri tabanında birliktelik kurallarının tanımlanması veri tabanı bilgi keşfi sürecinin ilk adımıdır [13]. Çok sayıda verinin depolandığı bir veri tabanı içinde nesneler arasında ilk anda fark edilmeyen ilişkilerin ortaya çıkartılması stratejik kararların alınmasına da yardımcı olabilir.

Genellikle satın alma eğilimlerinin tanımlanmasında kullanılan birliktelik kuralları veri madenciliğinde yaygın olarak pazar sepeti çözümlemesinde kullanılmaktadır. Pazar sepeti çözümlemesinde müşteri ile ilgili veri

hareketlerinden gelecekte müşterinin nasıl bir tercih yapacağına dair sonuçlar tahmin edilmektedir. Ayrıca tıp, finans ve farklı olayların birbirleriyle ilişkili olduğunun belirlenmesi sonucunda değerli bilgi kazanımının söz konusu olduğu alanlarda da bu teknikler önemli bir yere sahiptir [14].

2.3.1. Sınıflama ve regresyon modelleri

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 [13]. Matematiksel olarak sınıflama;

$D = \{t_1, t_2, ..., t_n\}$ bir veri tabanı ve her bir t_i bir kayıt (gözlem),

$C = \{C_1, C_2, ..., C_m\}$ ise m adet sınıftan oluşan sınıflar kümesini temsil etmek üzere,

$f : D \to C$ ve her bir t_i bir sınıfa dahildir.

Ayrıca her bir C_j ayrı bir sınıftır ve her bir sınıf kendisine ait kayıtları içerir. Yani,

$C_j = \{t_i / f(t_i) = C_j, 1 \le i \le n, ve \ \ t_i \in D\}$ olarak tanımlanmaktadır.

Veri madenciliği yöntemleri içerisinde en yaygın kullanıma sahip olan sınıflama ve regresyon modelleri arasındaki temel farklılık bağımlı değişkenin kategorik veya süreklilik gösteren bir değere sahip olmasıdır. Daha önce de bahsedildiği gibi, eğer bağımlı değişken sürekli ise problem regresyon

problemi, değil ise problem sınıflama problemi olarak adlandırılır. Ancak bölüm 3'de ayrıntılı anlatılacak olan lojistik regresyon gibi kategorik değerlerin de tahmin edilmesine imkan veren tekniklerle, her iki model giderek birbirine yaklaşmakta ve bunun bir sonucu olarak aynı tekniklerden yararlanılması mümkün olmaktadır.

Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler;

- Yapay sinir ağları (Artificial Neural Networks),
- Karar ağaçları (Decision Trees),
- Lojistik regresyon (Logistic Regression),
- 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.

2.4. Veri Madenciliğinde Karar Ağaçları

Sınıflama ve regresyon modellerinin bir yöntemi olan karar ağaçları, kurulmasının ucuz olması, yorumlanmalarının kolay olması, veritabanı 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 [15]. Karar ağaçları, basit karar verme adımları uygulanarak, çok sayıda kayıt içeren bir veri kümesini çok küçük kayıt gruplarına bölmek için kullanılan bir yapıdır

[16]. 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 [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.

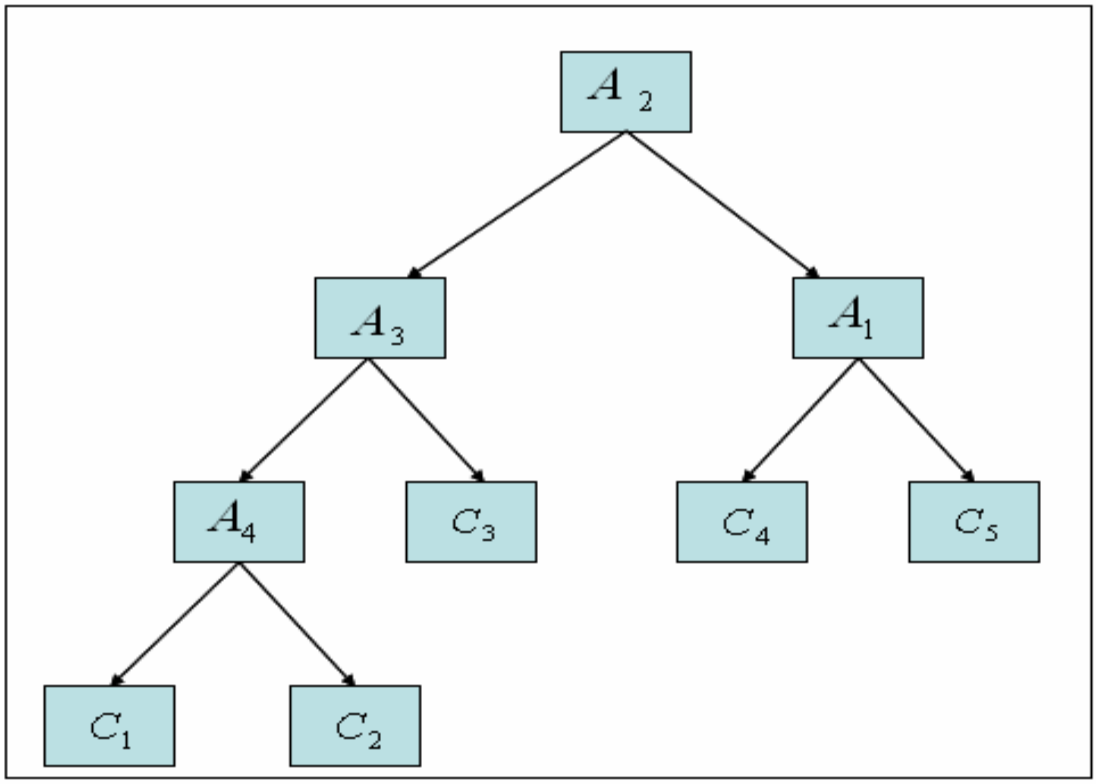
Karar ağacında, tanımlanmış olan soruya ilişkin cevap gruplara ayrılmaktadır. Cevaplar soruya verilecek bir ölçüt belirlendikten sonra setler arasındaki riski maksimize edecek şekilde bölünmekte ve en iyi bölünmeyi bulmak için her

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 [18].

$D = \{t_1,...,t_n\}$ bir veri tabanı olmak üzere, her t_i, $t_i = \{t_{i1}...t_{in}\}$ 'den ve bu veri tabanı $\{A_1, A_2,..., A_n\}$ alanlarından oluşmaktadır.

Bunun dışında $C = \{C_1,..., C_n\}$ kadar da sınıf verilmiş olduğunda,

- Her bir düğümü A_i alanıyla tanımlanmış,
- Her düğümden ayrılan kollar bu alanla ilgili bir soruya yanıt veren ve
- Her yaprağın bir sınıf olduğu karar ağacı şekil 2.2'de gösterilmiştir [13].



Şekil 2.2. Örnek karar ağacı

Şekil 2.2'de görülen karar ağacındaki $A_1, A_2, ..., A_n$'den her biri bir düğümü oluşturmakta ve her düğüm kendinden sonra iki dala ayrılmaktadır. Bu ayrılma işlemi sürecinde, A_i düğümü hakkında cevabı veri tabanında bulunacak bir soru sorulmakta ve verilen yanıta göre de bir dal izlenmektedir. Ağaçtaki $C_1, C_2, ..., C_n$'lerin her biri birer yaprağı aynı zamanda sınıfı temsil etmektedirler.

Karar ağaçları oluşturulurken kullanılan algoritmanın ne olduğu önemli bir hususdur. Kullanılan algoritmaya göre ağacın şekli değişebilir. Bu durumda değişik ağaç yapıları da farklı sınıflandırma sonuçları verecektir. Kök denilen ilk düğümü oluşturan A_i'nin farklı olması, en uçtaki yaprağa ulaşırken izlenecek yolu ve dolayısıyla sınıflandırmayı da değiştirecektir [13].

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. Ağaç yapısı heterojen yapıdaki veri kümesinin daha küçük ve homojen bir yapıya dönüşmesi için kurallar tanımlar. Ağaç inşası sonunda elde edilen ağaç maksimum ağaç olarak adlandırılır ve öğrenme kümesindeki deney ünitelerine en uygun ağaçtır. Ancak maksimum ağaç pratikte iki dezavantaja sahiptir [19].

- Maksimum ağaç başlangıç veri setini (öğrenme kümesini) kusursuz biçimde tanımlar çünkü eklenen her bağımsız değişken hatalı sınıflama oranını düşürür. Bu durumda, maksimum ağaç veri için olması gerekenden daha iyi bir tahmin modeli sunar. Ancak, başlangıç veri setine aşırı uyumlu maksimum ağaçlar farklı bir veri seti söz konusu olduğunda iyi bir tahmin sağlayamazlar.

- Bir sınıflama ağacının karmaşıklık ölçüsü o ağacın terminal düğüm sayısına eşittir. Terminal düğüm sayıları ve dolayısıyla karmaşıklığı yüksek olan maksimum ağacın anlaşılması ve yorumlanması güçtür.

Maksimum ağacın pratikte ortaya çıkardığı bu sorunların çözümü için maksimum ağacın budanması gereklidir. Maksimum ağacın budanması daha küçük ağaçlar dizisi oluşturur ve oluşturulan bu dizi içerisinden optimum ağaç seçilir. Optimum ağaç maksimum ağaçtan daha az karmaşıklığa sahiptir ancak öğrenme kümesine maksimum ağaçtan daha az uyumludur ve hatalı sınıflama oranı da daha yüksektir [19].

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,
- Kategorilerin birleştirilmesi ve sürekli değişkenlerin kesikli değişkenlere dönüştürülmesi,
- Parametrik modellerin kurulmasında kullanılmak üzere çok sayıdaki değişkenden en önemlilerinin seçilmesi,
- Gelecekteki olayların tahmin edilebilmesi için kurallar oluşturulmasıdır [8].
- Bireylerin kredi geçmişlerini kullanarak kredi kararlarının verilmesi (Credit Scoring),
- Tıbbi gözlem verilerinden yararlanarak en etkin kararların verilmesi,
- Geçmişte işletmeye en faydalı olan bireylerin özelliklerini kullanarak işe alma süreçlerinin belirlenmesi,
- Satışları etkileyen değişkenlerin belirlenmesi,
- Üretim verilerinin incelenmesiyle ürün hatalarına yol açan değişkenlerin belirlenmesidir [20].

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 tezdeki uygulamada CART algoritması kullanılmış olup bir sonraki bölümde ayrıntılı olarak anlatılmıştır.

2.4.1. Karar ağaçlarında kullanılan bazı algoritmalar

C4.5 algoritması

C4.5 algoritması ilk olarak 1993 yılında Quinlan tarafından ortaya atılmıştır [21]. Verilerin yinelemeli olarak alt kümelere ayrılmasıyla bir sınıflama karar ağacı oluşturulur. Karar ağacı oluşturulurken kayıp veriler hesaba katılmaz

yani sadece verileri eksik olmayan kayıtlar kullanılır. Algoritma budama işlemleri yapabilmekte ve karar üretilmesi gerçekleştirebilmektedir.

SLIQ algoritması

SLIQ algoritması 1996 yılında Uluslararası İş Makinaları Şirketi (International Business Machines Corporation) Almaden Araştırma merkezinde Mehta M. ve diğ. tarafından önerilmiştir [22]. Algoritma hem sayısal hem de kategorik verilerin sınıflandırılmasında kullanılabilmektedir. Sayısal verilerin değerlendirilmesinde maliyeti azaltmak için ağacın oluşturulması sırasında önceden sıralama tekniği kullanılır. SLIQ algoritmasında en iyi dallara ayırma kriterinin bulunmasında teknik; verilerin sıralama işleminin her düğümde yapılması yerine öğrenme verilerinin sadece bir kere, o da ağacın büyüme aşamasının başlangıcında yapılarak gerçekleştirilmesidir [13].

SPRINT algoritması

SPRINT algoritması sınıf ve sıra (kayıt) numarasını tutan farklı bir nitelik listesi yapısı kullanarak her bir değişken için ayrı bir değişken listesi hazırlar. Dolayısıyla her tabloda kullanılacak olan değişken, sınıf ve sıra numarası bulunacağından veri tabanındaki değişken sayısı kadar tablo oluşturmaktadır. Algoritmada bir liste bölümlendiğinde listedeki kayıtların sıraları da yeniden düzenlenmekte ve sürekli değerleri taşıyan tablolar sürekli değişkene göre sıraya dizilirken, kategorik değer taşıyan diğer tablolar sıra numarasına göre sıralı olarak kalmaktadır.

MARS algoritması

MARS algoritması temeli regresyon analizine dayanan, büyük ve karmaşık veri tabanlarına uygulanabilen bir yöntemdir. MARS çok boyutlu verilerin içinde gizli kalmış karmaşık veri yapısını, en uygun veri dönüşümlerini ve verilerin karşılıklı etkileşimlerini belirleyebilme avantajı bakımından regresyon

modellemesinde yeni bir yaklaşımdır. MARS algoritmasıyla geniş veri tabanları ve çok karmaşık veri yapıları için regresyon modelleri kolayca geliştirilebilmektedir [23].

CART algoritması ve CART ile yapılan çalışmalar

Bilimsel çalışmalardan elde edilen verilerin analizinde sınıflama ve regresyon ağaçları, kümeleme, diskriminant ve lojistik regresyon analizlerini içeren sınıflama teknikleri ve regresyon modelleri sıklıkla kullanılmaktadır [24]. 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 [19].

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 [25]. Bu yönüyle CART, hem çoklu regresyon analizini hem de bağımlı değişkenin kategorik olduğu durumlarda kullanılan lojistik regresyon analizini kapsamaktadır.

CART algoritması 1984 yılında Breiman tarafından geliştirilmiştir [26]. Deconinck ve diğerleri (2005), ilaçlar üzerinde yaptıkları çalışmada CART algoritmasını kullanarak mide ve bağırsaklarda emilim özelliklerine göre ilaçları sınıflandırmıştır [25]. Benzer şekilde Temel G. ve diğerleri (2005), CART algoritması yardımıyla mevcut verilerini sınıflandırarak Restless Legs Syndrome (RLS) hastalarına tanı koymayı kolaylaştırmışlardır. Çalışmada Mersin Üniversitesi Tıp Fakültesi Nöroloji bölümünde 206 denek hasta üzerinde yapılan anket çalışmasının sonuçları kullanılmış ve deneklerin RLS

hastası olup olmama durumunu belirleyen değişkenler sınıflama ağaçları analizi ile tespit edilmiştir [19].

Haughton ve Oulabi (2006), çalışmalarında doğrudan pazarlama modelini CHAID ve CART algoritmaları ile gerçekleştirmiş, CART ve CHAID analizlerinin sonuçlarını karşılaştırmıştır [27]. Lemon ve diğerleri (2003), halk sağlığı üzerinde yaptıkları çalışmada, riskli grupta olan ve benzer özellikleri gösteren hastaları sınıflandırmak için CART ile lojistik regresyon analizini kullanmış ve iki analizin sonuçlarını karşılaştırmıştır [28].

Çamdeviren ve diğerleri (2007), farklı doğum sonrası dönemlerinde 1447 kadının depresyon durumunu etkileyen sosyo-demografik risk faktörlerinin tespit edilmesinde CART ve lojistik regresyon analizlerini kullanmıştır [29]. Türe ve diğerleri (2009), 500 meme kanserli hasta üzerinde yinelemesiz sağkalım süresini etkileyen risk faktörlerinin belirlenmesinde karar ağacı yöntemlerinden CART, CHAID, QUEST, C4.5 ve ID3 ile Kaplan-Meier analizini birlikte kullanmıştır [30].

Kayri ve Boysan (2008), CART algoritması yardımıyla yaptıkları çalışmalarında, Yüzüncü Yıl Üniversitesi'nden 437 öğrenciye uygulanan sınırlılık şemaları envanteri, genel öz yeterlilik ölçeği ve beck depresyon envanteri kullanarak, sınırlılık algısının depresyon için bir bilişsel yatkınlık faktörü olduğunu tespit etmiştir [31]. Albayrak ve Akbulut (2008), İstanbul Menkul Kıymetler Borsası (İMKB) sanayi ve hizmet sektörlerinde faaliyet gösteren firmaların sermaye yapılarını etkileyen en önemli faktörleri CART analizi ile araştırmıştır. Çalışmada seçilen 38 farklı finansal göstergeden, işletmelerin sermaye yapılarının en önemli belirleyicilerinin likidite, varlık kullanım etkinliği ve işletme riski olduğunu göstermiştir [32].

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 [33].

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 [24]. CART analizi genellikle 3 adımdan oluşmaktadır. Birinci adım veri setini tanımlayan maksimum ağacın oluşturulmasıdır. İkinci adım; oluşturulan ağaçlar içerisinden bağımlı değişkenle önemli ilişkisi olan ağaçları seçmek için yapılan budama işlemi ve son adım ise en uygun ağaç yapısının seçimidir [26].

Maksimum ağacın oluşturulması

Maksimum ağaç, ağacın kökünde başlayan bir ikili bölme işlemi kullanan yapıdır. Ağacın kökü, veri seti içerisindeki her nesneyi içermekte ve her bir seviyede kendine özgü iki alt düğüm halinde bölünen bir ana düğüm olarak düşünülmektedir. Sonraki adımda, her alt grup bir ana grup olmaktadır. Her bölünme bir alt gruptaki tüm nesnelerin benzer bağımlı değişken değerlerine sahip olacak şekilde seçilen bir açıklayıcının değeri ile tanımlanmaktadır.

Sürekli değişkenlerin bölünmesi x_i'nin seçilmiş açıklayıcı değişken ve a_j'nin onun bölünme değeri olan "$x_i < a_j$" ile ifade edilmektedir.

Bir bölünme ve onun bölünme değeri için en uygun tanımlayıcıyı seçmek için CART, içinde tüm tanımlayıcıların ve tüm bölünme değerlerinin düşünüldüğü bir algoritma kullanmakta ve test koşulunun ne kadar iyi uygulandığını belirlemek için ana düğümün safsızlık derecesini alt düğümlerin safsızlık derecesiyle karşılaştırmaktadır. Ana ve alt düğümlerin safsızlıkları arasındaki fark ne kadar büyükse test koşulu o kadar daha iyi olduğundan, ana düğüm (t_p) ve alt düğümler (t_L ve t_R) arasındaki safsızlık ölçüsünü en iyi azaltan bölünme seçilmektedir. Matematiksel olarak bu durum aşağıdaki gibi ifade edilmektedir:

$$\Delta i(s,t_p) = i_p(t_p) - PLi(t_L) - PRi(t_R) \tag{2.1}$$

Burada i safsızlığı, s aday bölünme değerini ve PL ile PR sırasıyla sağ ve soldaki alt düğümlerdeki nesnelerin bölünmelerini ifade etmektedir. Bu eşitlikte $\Delta i(s,t_p)$ değerini maksimize edecek s değerinin seçilmesi amaçlanmakta ve t_p düğümünde bütün kayıtların katılımıyla hesaplanan bu değer, CART ağacında gelişme (improvement) kavramı ile ifade edilmektedir. CART algoritması ağacı geliştirirken $\Delta i(s,t_p)$'yi maksimize eden bir test koşulu seçtiğinden ve $i_p(t_p)$ bütün test koşulları için aynı olduğundan, $\Delta i(s,t_p)$'yi maksimize etmek alt düğümlerin safsızlık ölçülerinin ağırlıklı ortalamalarını minimize etmekle eşdeğer olmaktadır [25].

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.

Çizelge 2.1'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 [34].

Çizelge 2.1. 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;

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

$$Gini_{sağ} = 1 - \sum_{i=1}^{k} \left(\frac{R_i}{T_{sağ}} \right)^2 , \quad (2.3)$$

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

$$Gini_j = \frac{1}{n} \left(T_{sol} \times Gini_{sol} + T_{sağ} \times Gini_{sağ} \right) \quad (2.4)$$

formülü ile hesaplanmaktadır.

Çizelge 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 çizelge 2.2 oluşmaktadır.

Çizelge 2.2. 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:

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

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

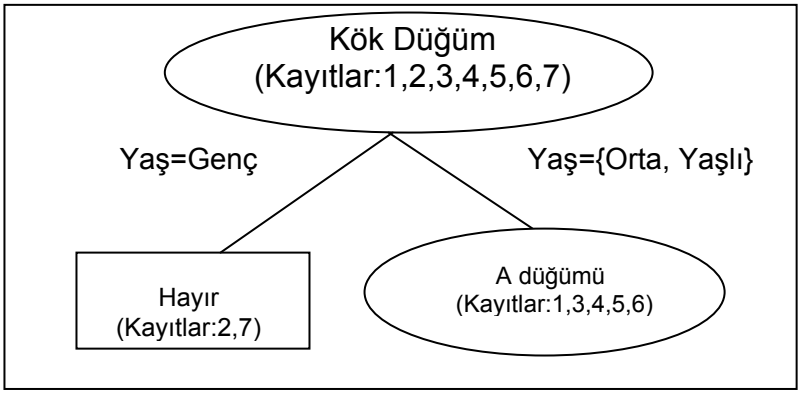
$$Gini_{cinsiyet} = \frac{3 \times 0,444 + 4 \times 0,375}{7} = 0,405$$

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

Çizelge 2.3. Her nitelik için hesaplanan $Gini$ indeks değerleri

İşe Kabul Durumu	**Eğitim Durumu**		**Yaş**		**Cinsiyet**	
	İlkokul	Ortaokul 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 çizelgede 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 çizelge 2.1 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 2.3'de gösterildiği gibi olacaktır.



Şekil 2.3. 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 çizelge 2.4'e ve eğitim verisinin gruplandırılmış hali olan çizelge 2.5'e göre tekrarlanmaktadır.

Çizelge 2.4. (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

Çizelge 2.5. 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 $Gini$ indeks değerleri ise aşağıdaki gibi hesaplanmaktadır:

<u>Eğitim Durumu için:</u>

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

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

$$Gini_{egitim} = \frac{1 \times 0 + 4 \times 0,375}{5} = 0,300$$

Yaş için:

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

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

$$Gini_{yaş} = \frac{3 \times 0,444 + 2 \times 0}{5} = 0,267$$

Cinsiyet için:

$$Gini_{sol} = 1 - \left[\left(\frac{1}{2} \right)^2 + \left(\frac{1}{2} \right)^2 \right] = 0,500$$

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

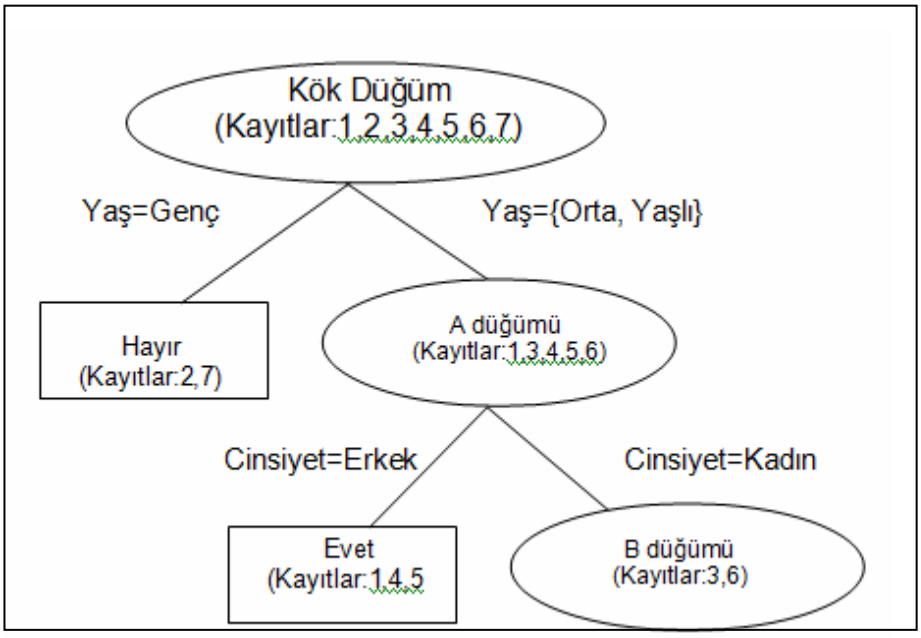
$$Gini_{cinsiyet} = \frac{2 \times 0,500 + 3 \times 0}{5} = 0,200$$

Çizelge 2.6'da gösterilmiş olan hesaplanan bu yeni $Gini_j$ değerleri arasından en küçük olanı seçilmekte ve yeni bölünme de en küçük değere sahip olan nitelik üzerinden yinelenmektedir.

Çizelge 2.6. 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**	

Yukarıdaki çizelgede hesaplanan yeni değerler göz önüne alındığında $Gini_{cinsiyet} = 0{,}200$ değerinin $Gini_j$ değerleri içinde en küçüğü olduğu anlaşılmaktadır. Bu durumda yeni bölünme cinsiyet niteliğinin "kadın" ve "erkek" değerlerine göre olacaktır. Bölünmeyi elde etmek için çizelge 2.1 üzerinde cinsiyete ilişkin değerler arandığında, bölünmenin (3,6) ve (1,4,5) satırları biçiminde gerçekleşeceği anlaşılmakta olup söz konusu bölünme şekil 2.4'de gösterildiği gibi olacaktır.



Şekil 2.4. İkinci bölünme sonucunda karar ağacının görünümü

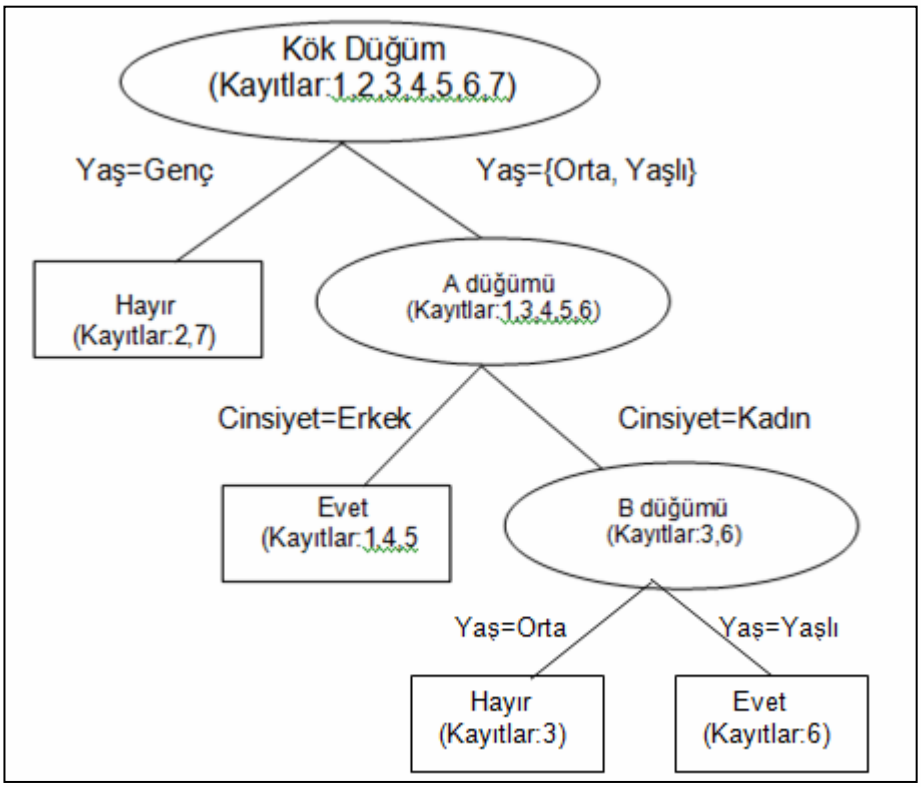
Şekil 2.4 üzerinde görüldüğü gibi (1,4,5) satırları "Evet" ile sonlanmıştır. Bu durumda bu satırların da eğitim verisinden çıkarılması sağlanarak çizelge

2.7'de sunulan üçüncü bölünme için oluşturulan eğitim verisi elde edilmektedir.

Çizelge 2.7. Üçüncü bölünme için oluşturulan eğitim verisi

İşe Başvuru Sırası	Eğitim Durumu	Yaş	Cinsiyet	İşe Kabul Durumu
3	Yüksekokul	Orta	Kadın	Hayır
6	Yüksekokul	Yaşlı	Kadın	Evet

Elde edilen son tablo iki satırdan oluşmakta ve iki ayrı sınıfı tanımlamaktadır. Bu durum sonucunda ise şekil 2.5'de gösterilen karar ağacı elde edilmektedir [34].



Şekil 2.5. Üçüncü bölünme sonucunda elde edilen karar ağacı

Ağaç budama

Maksimum ağaç genellikle aşırı öğrenme (overfitting)[1] eğilimi göstermektedir. Diğer modelleme tekniklerinde olduğu gibi ağacın karmaşıklığı ve tahmin gücü arasında bir orta yol bulmak için budama işlemi gerekmektedir.
Budama işlemi esnasında maksimum ağaçtan türetilen bir seri daha küçük alt ağaçlar arka arkaya gelen uç dallardan elde edilmekte böylece farklı alt ağaçlar en uygun olanla karşılaştırılmaktadır. Bu karşılaştırma, hem ağaç doğruluğu hem de karmaşıklığın düşünüldüğü bir maliyet-karmaşıklık ölçüsü üzerine kurulmaktadır. Maliyet-karmaşıklık parametresi $R_\alpha(T)$ olmak üzere her bir alt ağaç T için aşağıdaki gibi tanımlanmaktadır:

$$R_\alpha(T) = R(T) + \alpha\left|\tilde{T}\right| \tag{2.5}$$

$R(T)$ ortalama düğüm (node) içi kareler toplamı, $\left|\tilde{T}\right|$ alt ağacın toplam düğüm sayısı olarak tanımlanan ağaç karmaşıklığı ve α her bir ilave terminal düğüm için bir ceza olan karmaşıklık parametresidir. Budama işlemi sırasında α, dereceli olarak 0 dan 1'e doğru artmakta ve α'nın her bir değeri için $R_\alpha(T)$'yi minimize eden ağaç seçilmektedir. α'nın 0'a eşit olduğu değer için $R_\alpha(T)$ maksimum ağaç tarafından minimize edilmekte ve böylece dereceli olarak artan bir α ile karmaşıklığı azalmış bir seri ağaç elde edilmiş olmaktadır [25].

Optimum ağacın seçimi

Elde edilmiş alt ağaçlar serisinden optimal olanı seçilmek zorundadır ve seçim işlemi de tahmin hatasının değerlendirilmesi üzerine kuruludur. Tahmin hatası ise sıklıkla çapraz geçerlilik testi kullanarak değerlendirilmektedir. Çapraz geçerlilik testinde pek çok sayıda nesne, veri setinden rastgele olarak çıkarılmakta ve geri kalan veri ile uyumlu şekilde ağacın tahmin gücünü

[1] Ağaç çok büyük ve eğitim örneklerine ait hata oranı düşük olduğu halde, test verisi için sınıflandırma hatası büyük ise bu duruma aşırı öğrenme (overfitting) denir.

değerlendirmek için bir test grubu olarak kullanılmaktadır. Genellikle tercih edilen test ise 10 katlı çapraz geçerlilik testidir.

10 katlı çapraz geçerlilik testi sırasında veri seti her biri bağımlı değişkenin benzer dağılımını içeren 10 alt gruba bölünmektedir. Böylece bu alt gruplardan biri, diğer dokuz alt grupla uyumlu ağacın tahmin hatasını değerlendirmek için kullanılmaktadır. Bu işlem her defasında bir diğer alt grubu test grubu olarak kullanarak on defa tekrarlanmaktadır. En doğru ağaç, aşağıda Eş. (2.6) ile verilen çapraz geçerlilik ortalama hata karekökü (ÇGOHK) olarak tanımlanan, ortalama en küçük çapraz geçerlilik hatasına sahip olandır.

$$ÇGOHK = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{2.6}$$

Eş. 2.6'da yer alan y_i, i nesnesinin bağımlı değişken değerini, $\hat{y}_i$, i nesnesi için tahmin edilen bağımlı değişken değerini ve n ise toplam nesne sayısını göstermektedir.

Bu durumda en doğru ağacın bir standart hata sınırları içinde ÇGOHK ile birlikte en az karmaşık olan optimum ağacı tespit ederek en doğru olan ağacın yani, tahmin hatasıyla karşılaştırılabilir daha az karmaşık olan ağacın seçilmesi amaçlanmaktadır [25].

CHAID algoritması

CHAID, CART algoritmasına benzemektedir ancak CART ikili ağaçlar türetirken CHAID çoklu ağaçlar türetmektedir. CHAID algoritmasında kategorik değişkenlere ilişkin veri kümesi, bağımlı değişkeni en iyi açıklayacak şekilde detaylı homojen alt gruplara bölünmektedir. Verinin bölümlere ayrılmasında ise en uygun bölümleri seçmek için kullanılan

entropy ve gini gibi ölçütler yerine Ki-Kare testine dayanan bir ölçme tekniği kullanılmaktadır. CHAID hem sürekli hem de kategorik değişkenler üzerinde çalışabilmesi, ağaçta her düğümü ikiden fazla alt gruba ayırabilmesi gibi nedenler dolayısıyla günümüzde de tercih edilen bir algoritmadır.

QUEST algoritması

Hızlı, yansız, etkin istatistiksel ağaç olarak bilinen QUEST algoritmasının CART algoritmasında olduğu gibi ikili karar ağaçları oluşturmak üzerine kurulu bir yapısı vardır. Algoritmanın ikili ağaç türetme nedeni; bölme, CART' ta olduğu gibi maksimum ağacı budama ve durdurma kuralları gibi tekniklere izin veren ikili ağaçların kullanılmasıdır.

3. LOJİSTİK REGRESYON ANALİZİ

3.1. Lojistik Regresyon ile Yapılmış Çalışmalar

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.

Lojistik regresyon modelinin kullanımına ilişkin ilk çalışmalar Berkson (1944) tarafından yapılmış ve model Finney (1972) tarafından biyolojik deneylerde probit analize bir alternatif olarak önerilmiştir [35]. Cox (1970) lojistik modele ilişkin çeşitli uygulamalar yapmıştır, onu izleyen yıllarda Anderson (1979, 1983) tarafından gelişmeler verilmiştir [36–37]. Daha sonra Lesaffre (1986), Lesaffre ve Albert (1989)'in incelemiş olduğu çoklu grup lojistik modellerde etkin ve aykırı gözlemlerle belirleme ölçütleri lojistik modeller üzerine yapılan diğer çalışmalar arasındadır [38].

Lojistik regresyon seyrek gerçekleşen olayların ortaya çıkmasını ya da bu olayların ortaya çıkma sıklıklarının olasılıklarını tahmin etmek için de kullanılmaktadır [39]. Bu doğrultuda George Washington Üniversitesi'nden Langche Zeng ve Harward Üniversitesi'nden Gary King (2001) çok seyrek gerçekleşen olaylar karşısında lojistik regresyonun ve veri seçiminin nasıl uygulanması gerektiğini açıklamaya çalışmışlardır [40].

Daha önce lojistik regresyon analiziyle (Bergh, Baigi, Mansson, Mattsson, & Marklund,2007; Chen, Rosenheck, Greenberg, & Seibyl, 2007) veya probit model ile (Huberman, Iyengar, & Jiang, 2007) emeklilik sistemi üzerine de pek çok çalışma yapılmıştır [41]. Benzer şekilde Amerika'da Robin Fisher ve Jennifer Campbell (2002) tarafından sağlık sigortasıyla ilgili olarak sigortası olmayanların yaş, cinsiyet ve ekonomik durumlarına göre sınıflandırılmasında lojistik regresyon modelinden yararlanılmıştır. Avustralya'da Jeromey Temple (2002) tarafından Avustralyalı hane halkının yıllar itibariyle özel sağlık

sigortasına ekonomik, sağlık ve sosyal olarak geçiş nedenleri çoklu lojistik regresyon analiziyle açıklanmıştır [42].

Harward Üniversitesi'nden Amitabh Chandra ve Dartmouth Koleji'nden Andrew A. Samwick (2005) Amerika'daki malullük riskini ve malullük risk sigortalarını incelemiş, bu çalışmada lojistik regresyon kullanarak malulen emekli olanları çalışabilme sınırlarına, çalışamama durumlarına ve sağlık yardımı almak zorunda olmalarına göre sınıflandırmıştır [43].

3.2. Doğrusal Model ve Lojistik Regresyon Arasındaki İlişki

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:

$$Y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2) \tag{3.1}$$

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 kategorik değişken olması durumunda yaygın olarak kullanılan modellerden birisi lojistik regresyon modelidir. 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 [1].

Doğrusal regresyon analizinde olduğu gibi lojistik regresyon analizinde de bazı değişken değerlerine dayanarak tahmin yapılmaya çalışılır. Fakat söz konusu bu iki yöntem arasında temel olarak üç önemli fark mevcuttur.

1) Doğrusal regresyon analizinde tahmin edilecek bağımlı değişken sürekli iken lojistik regresyon analizinde bağımlı değişken kesikli (kategorik) bir değer almaktadır.

2) Doğrusal regresyon analizinde bağımlı değişkenin değeri, lojistik regresyon analizinde ise bağımlı değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı tahmin edilmektedir.

3) Doğrusal regresyon analizinde bağımsız değişkenlerin çoklu normal dağılım göstermesi şartı aranırken lojistik regresyon analizinde böyle bir şart bulunmamaktadır [35].

3.3. Doğrusal Olasılık Modeli ve Lojistik Fonksiyon

Regresyon analizinde bağımlı değişkenin kategorik olması basit doğrusal regresyon analizinin bazı varsayımlarının yerine getirilememesine yol açmaktadır. Doğrusal olasılık modeli yardımıyla bu durum aşağıdaki gibi açıklanabilir:

Çeşitli gösterim biçimleri olan genel doğrusal regresyon modelini; n gözlem sayısı, p bağımsız değişken sayısı olmak üzere;

$$E(y_i/x_{i1},...,x_{ip}) = \sum_{k=0}^{p} \beta_k x_{ik} \qquad i = 1,...,n \quad ; \tag{3.2}$$

biçiminde koşullu beklenen değer olarak da yazmak mümkündür.

Eş. 3.2 ile ifade edilen modelde bağımsız değişkenler üzerinde kısıt yok iken *y* bağımlı değişkeninin sürekli olması koşulu vardır. Herhangi bir i'inci gözlem için;

$$y_i = \sum_{k=0}^{p} \beta_k x_{ik} + \varepsilon_i \tag{3.3}$$

şeklinde ifade edilen modelde bağımsız değişkenler üzerinde bir kısıt olmadığından, y_i bağımlı değişken değeri -∞ 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:

$$E(y_i) = 1 \times P(y_i = 1) + 0 \times P(y_i = 0) = P(y_i = 1) \tag{3.4}$$

olmaktadır. Bu sonuç regresyon denklemi olarak yazılacak olursa:

$$E(y_i) = P(y_i = 1) = \sum_{k=0}^{p} \beta_k x_{ik}, \qquad i = 1,...,n \tag{3.5}$$

ifadesi elde edilmektedir. Sol tarafı 0-1 arasında olasılık değerleri alan bu denkleme "Doğrusal olasılık modeli" adı verilmektedir [44].

Eş. 3.3'de y_i'nin iki ayrı değeri için ε_i hata terimi $i = 1,...,n$ ve y_i=0 ile y_i=1 için,

$$\sum_{k=0}^{p} \beta_k x_{ik} + \varepsilon_i = 0 \Rightarrow \varepsilon_i = -\sum_{k=0}^{p} \beta_k x_{ik} \tag{3.6}$$

$$\sum_{k=0}^{p} \beta_k x_{ik} + \varepsilon_i = 1 \Rightarrow \varepsilon_i = 1 - \sum_{k=0}^{p} \beta_k x_{ik} \tag{3.7}$$

şeklinde elde edilebilir.

Hata terimleri üzerindeki varsayımların beklenen değeri,

$$E(\varepsilon_i) = P(y_i = 0)\left(\sum_{k=0}^{p} - \beta_k x_{ik}\right) + P(y_i = 1)\left(1 - \sum_{k=0}^{P} \beta_k x_{ik}\right) = 0 \tag{3.8}$$

ve varyansı,

$$V(\varepsilon_i) = E(\varepsilon_i^{\,2}) = P(y_i = 0)\left(\sum_{k=0}^{p} - \beta_k x_{ik}\right)^2 + P(y_i = 1)\left(1 - \sum_{k=0}^{P} \beta_k x_{ik}\right)^2$$

$$= \left(\sum_{k=0}^{p} \beta_k x_{ik}\right)\left(1 - \sum_{k=0}^{p} \beta_k x_{ik}\right) \tag{3.9}$$

olarak elde edilmektedir [44]. ε_i'lerin varyansı bağımsız değişkenlere bağlı olarak değişmektedir. Bu durumda β_k katsayılarının En Küçük Kareler (EKK) tahmini yansız, ancak en iyi olmamaktadır. Değişen varyanslılık durumu söz konusu olduğu için doğrusal olasılık modeli ağırlıklandırılarak sabit bir duruma getirilmelidir. Bunun için önce y'lerin x'ler üzerindeki regresyonundan $\hat{\beta}$ tahminleri bulunur ve bu tahmin değerlerine göre ağırlık değerleri:

$$W_i = 1 \Big/ \left(\sum_{k=0}^{p} \hat{\beta}_k x_{ik} \right) \left(1 - \sum_{k=0}^{p} \hat{\beta}_k x_{ik} \right) \tag{3.10}$$

şeklinde elde edilir. Eş. 3.3 ile verilen model belirtilen bu ağırlık ile ağırlıklandırıldığında yeni hata terimi $W_i * \varepsilon_i$ sabit varyanslı olacağı için minimum varyanslı yansız kestirim değerleri elde edilebilecektir [44].

Eş. 3.5 ile verilen modelde $P(y_i = 1)$ olasılığı x'e bağlı olarak değişecektir. Her bir bağımlı değişken değeri farklı Bernoulli dağılımı, n bağımsız tekrarlı bağımlı değişken ise Binom dağılımı gösterecektir.

Verilen Eş. 3.5 doğrusal olasılık modeli olduğundan, bu eşitliğin sol tarafı 0-1 arasında sınırlı olasılık değerleri aldığından ve bu değerlerin sonsuz değerler alabilen bağımsız değişkenlerle ilişkilendirildiğinden söz edilmişti. Bağımsız değişkenlerin sınırsız değerler alması nedeniyle söz konusu eşitlik her zaman sağlanamamaktadır. Böylesi bir durumla karşılaşılmaması için en iyi çözüm bağımlı değişken değeri olarak ifade edilen olasılık değerinin çeşitli dönüşümlerle -∞, +∞ arasında tanımlı hale getirilmesidir. Bu amaçla geliştirilen dönüşümlerden en yaygın olarak kullanılanı lojit dönüşümdür.

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \tag{3.11}$$

şeklinde ifade edilen doğrusal olasılık modelinde, $E(\varepsilon_i) = 0$ varsayımı altında,

$$E(y_i \mid x_i) = \beta_1 + \beta_2 x_i \tag{3.12}$$

dir. y_i,1 ve 0 değerleri alabildiğinden,

$P_i = 1$ ve $1 - P_i = 0$ olmaktadır.

Bu durumda y_i değişkeninin olasılık dağılımı aşağıdaki şekilde olacaktır.

y_i	Olasılık
0	$1-P_i$
1	P_i

Beklenen değer tanımından,

$$E(y_i) = 0\times(1-P_i)+1\times(P_i)$$
$$= P_i$$

olacaktır. Bu durumda,

$$E(y_i \mid x_i) = \beta_1 + \beta_2 x_i = P_i \qquad (3.13)$$

olur. Yani modelin koşullu beklenen değeri aslında y_i' nin koşullu olasılığıdır ve P_i olasılığı [0,1] aralığında olduğuna göre,

0≤ $E(y_i \mid x_i)$≤1 sınırlaması vardır [45].

Doğrusal olasılık modeli x veri iken y'nin gerçekleşmesinin koşullu olasılığını ölçtüğüne göre, $E(y_i \mid x_i)$ zorunlu olarak [0,1] aralığında olmalıdır. Ancak $E(y_i \mid x_i)$'nin tahmin edilecek değeri $\hat{y}_i$'lerin bu sınırlamaya uyacağının bir güvencesi yoktur.

0≤ $E(y_i \mid x_i)$≤1 sınırlamasının yerine gelmeyişiyle doğrusal olasılık modelinin geçerliliği şüphe altındadır. Bu nedenle lojistik model bu tip bir yaklaşımda yani bağımlı değişkenin kategorik olduğu durumda kullanılabilmektedir. Bu

durumda bağımlı değişken bağımsız değişkenlere doğrusal bir yapı ile bağlı olacaktır.

Lojit fonksiyonları P=0,5 etrafında simetriktir ve lojit dönüşüm aşağıdaki gibi uygulanmaktadır:

$$z_i = \beta_1 + \beta_2 x_i \tag{3.14}$$

ise,

$$P_i = \frac{1}{1+e^{z_i}} \tag{3.15}$$

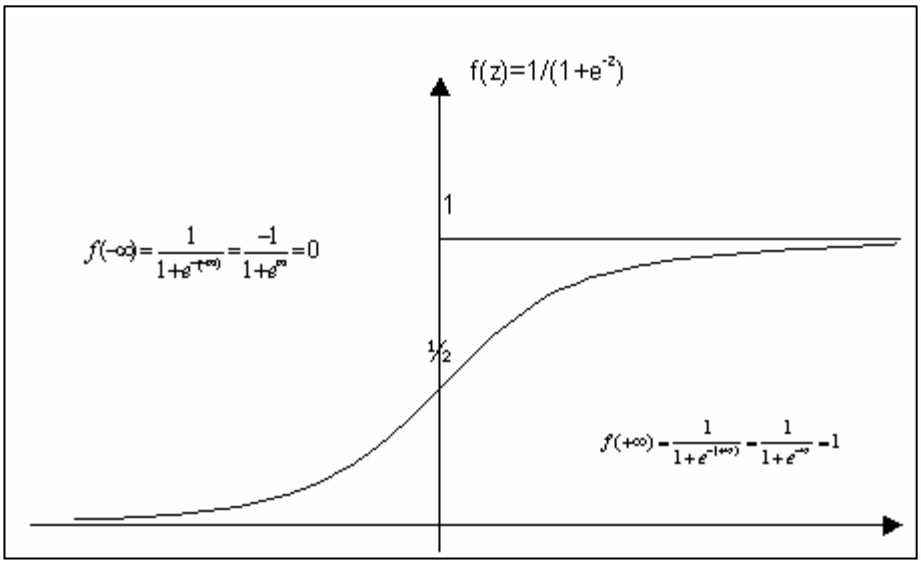
dir. z_i, (-∞,+∞) aralığında değişirken P_i, [0,1] aralığında değer almakta ve bu tip link fonksiyonunun kullanılmasıyla lojit model elde edilmektedir [45].

Lojistik regresyonun bahsedilen doğrusal olasılık modeline göre tercih edilmesinin nedeni lojistik modelin temelini oluşturan lojistik fonksiyondur. Bu sebeple lojistik regresyonun önemini açıklamak için, lojistik modelin matematiksel formunu tanımlayan lojistik fonksiyonu ifade etmek faydalı olacaktır.

$$f(z) = \frac{1}{1+e^{-z}} \qquad -\infty < z < +\infty \tag{3.16}$$

olmak üzere şekil 3.1'de görüldüğü gibi, Eş. 3.16 ile tanımlanan fonksiyonda z; $-\infty$ ile $+\infty$ arasında değer alırken, $z=-\infty$'da $f(z)$=0 ve $z=+\infty$'da $f(z)$=1 değerini almaktadır. Yani $f(z)$ lojistik fonksiyonu [0,1] kümesinin dışında değer almamaktadır $(0 \le f(z) \le 1)$. Bu lojistik modelin popüler olmasının birinci nedenidir çünkü model olasılığı tanımlamak için tasarlanmaktadır. Bu olasılık değeri bir bireyin ya da deney biriminin

probleme göre başarılı ya da başarısız olma olasılığını vermektedir. Bu diğer olasılık modelleri için her zaman doğru olmadığından, olasılık tahmin edilmek istendiğinde lojistik model genellikle öncelikli olarak tercih edilmektedir.



Şekil 3.1. Lojistik fonksiyon (S-Shape)

Lojistik modelin yaygın olarak kullanılmasının en önemli nedeni, lojistik fonksiyonun biçimi ile ilgilidir. Şekil 3.1 incelendiğinde, z=-∞'dan başladığında ve sağa doğru gidildiğinde, $f(z)$'nin değeri bir süre için sıfıra yakın olur, daha sonra etkileyici bir şekilde 1'e doğru yaklaşır ve son olarak z=+∞ değerini aldığında $f(z)$ fonksiyonu 1 değerine ulaşmaktadır. Yani, lojistik fonksiyonun grafiği S-biçiminde oluşmaktadır.

z değişkeninin birkaç risk faktörünün birleşmesi sonucunda bir kanıt olarak gözlemlenmesi ve verilen z değeri için riskin $f(z)$ ile gösterilmesi durumunda, araştırmacılar S-biçimindeki lojistik fonksiyona başvururlar. S-biçimli $f(z)$ fonksiyonunda, bir bireyin ya da deney birimine ilişkin riskin küçük z değerleri için, riskin eşik değerine (threshold value = 0.5) ulaşıncaya kadar düşük olarak meydana geldiğini, sonra orta z değerleri için riskin

arttığını ve son olarak z yeterince büyük değerler aldığında riskin 1 civarında oldukça yüksek olarak kaldığı gözlemlenmektedir.

Lojistik regresyonu yaygın kılan iki önemli noktayı özetlemek gerekirse, lojistik fonksiyonun tahminlerin 0 ile 1 aralığında olmak zorunda olduğu durumlarda bunu sağlaması ve lojistik fonksiyonun S şeklindeki grafiğinin risk faktörlerinin kombinasyonunun temsili olarak düşünülebilecek z değişkeninin $-\infty$'a yaklaşması durumunda 0 değerine, z değişkeninin $+\infty$'a yaklaşması durumunda ise 1 değerine yaklaşmasıdır. Bu durum da risk olasılığı olarak tanımlanabilen bağımlı değişken için önemli bir husustur [46].

3.4. Lojistik Regresyon Modeli ve Varsayımları

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, normallik varsayımının bozulması nedeniyle doğrusal regresyon analizine alternatif olmaktadır. Lojistik regresyonda bağımsız değişkenler ile iki ya da çok sınıflı kategorik bağımlı değişken arasındaki ilişkinin tanımlanması için matematiksel modelleme yapmak amaçlanmaktadır [46].

Doğrusal olasılık modeli başlığı altında bağımlı değişken değeri olarak ifade edilen olasılık değerinin çeşitli dönüşümlerle $-\infty$, $+\infty$ arasında tanımlı hale getirilmesi amacıyla yapılacak dönüşümlerden birinin lojit dönüşüm olduğundan söz edilmişti. Lojit dönüşümde ilk olarak;

$$E(y_i) = P(y_i = 1) = \sum_{k=0}^{p} \beta_k x_{ik} \qquad i = 1,\ldots,n \tag{3.17}$$

modelinde olasılık değerleri üzerinde $P/1-P$ dönüşümü yapılarak bağımlı değişkenin sınırları 0, $+\infty$ yapılmakta, daha sonra ise bu oran değerinin

logaritması alınarak bağımlı değişkenin sınırları -∞, +∞ yapılmaktadır. Bu dönüşümlerden sonra elde edilen yeni fonksiyon:

$$E(y_i) = P(y_i = 1) = L_i = ln(\frac{P_i}{1-P_i}) = \sum_{k=0}^{p} \beta_k x_{ik} \tag{3.18}$$

olarak yazılabilir. Bu modele "Lojistik model" ya da kısaca "Lojit" denmektedir. Ayrıca kullanılan $\ln(\frac{P}{1-P})$ dönüşümü de "lojit dönüşüm" adını almaktadır [1].

Lojistik fonksiyonun elde edildiği modelde kullanılan P_i olasılık değeri[2] ise:

$$P_i = \frac{\exp\left(\sum_{k=0}^{p} \beta_k x_{ik}\right)}{1+\exp\left(\sum_{k=0}^{p} \beta_k x_{ik}\right)} \tag{3.19}$$

biçiminde tanımlanmaktadır [47].

Bu modelde bağımlı değişkenin iki sınıflı olması sebebiyle hata terimi ε;

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_i x_{ij} + \varepsilon_i = 0 \;\Rightarrow \varepsilon_i = -\beta_0 - \sum_{j=1}^{k} \beta_i x_{ij} \tag{3.20}$$

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_i x_{ij} + \varepsilon_i = 1 \;\Rightarrow \varepsilon_i = 1 - \beta_0 - \sum_{j=1}^{k} \beta_i x_{ij} \tag{3.21}$$

---

[2] Eş. 3.19 ile verilen modelde kullanılan $\sum \beta_k x_{ik}$ ifadesi z_i ile ifade edilecek olursa ($z_i = \sum \beta_k x_{ik}$); $P_i = \frac{e^{z_i}}{1+e^{z_i}}$ ile belirtilen fonksiyon genellikle "link fonksiyon" olarak da bilinmektedir.

değerlerini almaktadır. Hata terimlerine ilişkin daha önce verilen Eş. 3.8 ve 3.9'dan yola çıkarak;

$$E(\varepsilon_i) = \Pr(y_i = 0)(-\beta_0 - \sum_{j=1}^{k} \beta_i x_{ij}) + \Pr(y_i = 1)(1 - \beta_0 - \sum_{j=1}^{k} \beta_i x_{ij})$$

$E(\varepsilon_i) = 0$ ve (3.22)

$$V(\varepsilon_i) = E(\varepsilon_i^{\,2})$$

$$= \Pr(y_i = 0)(-\beta_0 - \sum_{j=1}^{k} \beta_i x_{ij})^2 + \Pr(y_i = 1)(1 - \beta_0 - \sum_{j=1}^{k} \beta_i x_{ij})^2$$

$$= (1 - \Pr(y_i = 1))(\Pr(y_i = 1))^2 + \Pr(y_i = 1)(1 - \Pr(y_i = 1))^2$$

$$= \Pr(y_i = 1)(1 - \Pr(y_i = 1))$$

$$= P_i(1 - P_i) \quad (3.23)$$

varsayımları sağlanmaktadır. Yani hata terimi 0 ortalama ve $P(1-P)$ varyanslıdır. Hata terimi bu parametrelerle binom dağılımlı olup, analiz de bu teorik temele dayanmaktadır. Lojistik modele ilişkin varsayımlar kısaca şöyledir:

1) $y_i \in$ (0,1)

2) $P(y_i = 1 \mid x_i) = P_i$

3) $y_1, ..., y_n$ değerleri istatistiksel olarak bağımsızdır,

4) Bağımsız değişkenler olan x_k'lar birbirinden bağımsızdır [44].

Ayrıca modelin bağımlı değişkeninin sınırlarını genişletmek için kullanılan $\ln(\frac{P}{1-P})$ lojit dönüşümünün de bazı önemli özellikleri şunlardır:

- P arttıkça lojit (P) de artar,
- P, 0-1 arasında iken lojit (P) tüm reel sayı değerlerini alır,
- Eğer P < 0,5 ise lojit (P) < 0,
- Eğer P > 0,5 ise lojit (P) > 0 ve
- Eğer P = 0,5 ise lojit (P) =0'dır [44].

Bağımsız değişkenler üzerine herhangi bir kısıtlama getirilmeden lojistik regresyon analizinde bağımsız değişkenlerin durumuna göre farklı modeller kullanılabilir. Bu modeller:

a) Bağımsız değişkenlerin tümü kesikli ise;

$$\ln \frac{P_i}{1-P_i} = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \tag{3.24}$$

b) Bağımsız değişkenlerin tümü sürekli ise $\Pr(x_1,...,x_p)$ p bağımsız değişken üzerinde koşullu başarı olasılığı olmak üzere lojistik model;

$$\ln \frac{\Pr(x_1,...,x_p)}{1-\Pr(x_1,...,x_p)} = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \tag{3.25}$$

c) Bağımsız değişkenlerin bazılarının sürekli bazılarının kesikli olması durumunda çok değişkenli frekans dağılımı başarı durumu için f_1 $(x_1,...,x_p)$ ve başarısızlık durumu için f_0 $(x_1,...,x_p)$ biçiminde tanımlanmış iken lojistik model;

$$\ln \frac{\Pr(x_1,...,x_p) f_1(x_1,...x_p)}{(1-\Pr(x_1,...,x_p)) f_0(x_1,...,x_p)} = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \tag{3.26}$$

olarak tanımlanmaktadır [44].

Burada β katsayıları, gözlemleri f_0 ve f_1 fonksiyonlarına karşılık gelecek biçimde ayırma özelliğine sahip parametre değerleridir. Parametre tahmin değerleri ise en çok olabilirlik, yeniden ağırlıklandırılmış en küçük kareler ve minimum lojit Ki-Kare yöntemleri ile hesaplanabilir.

Söz konusu lojistik regresyon modellerine ait fonksiyonlar süreklidir ve x bağımsız değişkeni ile β parametre değerleri ne olursa olsun olasılık 0 ile 1 arasında değerler almaktadır.

3.5. Parametrelerin Anlamlılık Testleri ve Modelin Uyum İyiliği

Oluşturulan bir modelde katsayıların kestiriminden sonra yapılması gereken ilk şey modelin içerdiği parametrelerin anlamlılığının ölçülmesidir. Doğrusal regresyon analizinde tahmin edilen regresyon katsayısının anlamlılık testinde kullanılan en yaygın yöntemler varyans analizi ve t testidir. Lojistik regresyon analizindeki yaklaşım da benzerdir. Doğrusal regresyonda varyans analizi tekniğiyle genel kareler toplamı kendi içinde iki kısma ayrılmakta ve bunlar gözlemlerin regresyon doğrusundan sapmalarının kareleri toplamı için hata kareler toplamı ve regresyon modelinde tahmin edilen değerlerin kareleri toplamı için regresyon kareleri toplamı adını almaktadır. Öte yandan tahmin edilen regresyon katsayısı için t testi ile bu katsayı kendi standart hatasına oranlanmaktadır.

Lojistik modelde normallik varsayımı kısıtı olmadığından uyum iyiliği testlerinde t ve F gibi parametrik testler yerine Ki-Kare, olabilirlik oran testi gibi (likelihood ratio test) ölçütlerden yararlanılmaktadır [44].

3.5.1. Olabilirlik oran testi

Lojistik regresyonda kurulan modelin önemliliğinin test edilmesinde, diğer bir ifadeyle modele girmesi gereken bağımsız değişkenlerin belirlenmesinde, lojistik regresyondaki beklenen değerlerle gözlenen değerlerin karşılaştırılması, Eş. 3.27'de tanımlanan logaritmik olabilirlik fonksiyonundan faydalanılarak elde edilir.

$$L(\beta) = ln(\beta) = \sum_{i=1}^{n} [(y_i lnP(x_i)) + (1 - y_i) ln(1 - P(x_i))] \tag{3.27}$$

Olabilirlik fonksiyonu kullanılarak gözlenen ve beklenen değerlerin karşılaştırılması aşağıdaki eşitliğe dayanmaktadır.

$$D = -2 \ln \left(\frac{\text{Tahmin edilen modelin olabilirliği}}{\text{Doymuş modelin olabilirliği}} \right) \tag{3.28}$$

Eş. 3.28'de tahmin edilen model; sadece önemli olduğu düşünülen değişkenleri içeren modeli, doymuş model ise tüm değişkenleri içeren modeli göstermektedir. Parantez içindeki ifade de olabilirlik oranı (likelihood ratio) olarak adlandırılmakta ve bu teste olabilirlik oran testi (likelihood ratio test) adı verilmektedir.

Eş. 3.27 ve Eş. 3.28'den yararlanarak,

$$D = -2 \sum_{i=1}^{n} [y_i \ln(\frac{\hat{P}(x_i)}{y_i}) + (1 - y_i) \ln(\frac{1 - \hat{P}(x_i)}{1 - y_i}) \tag{3.29}$$

olarak tanımlanmaktadır. Eş. 3.29'da verilen D istatistiği sapma (deviance) olarak da bilinmekte ve uyum iyiliği değerleri için bazı yaklaşımlarda önemli rol oynamaktadır. Lojistik regresyon için D istatistiği, doğrusal regresyon için

hata kareler toplamı görevini görmektedir [1]. $(n-p)$ serbestlik derecesi ile Ki-Kare dağılımı gösteren bu ölçüt çeşitli modeller arasından seçim yapmak için kullanılmakta ve en küçük sapmaya sahip olan model en iyi model olarak alınmaktadır [47].

Bağımsız değişkenin anlamlılığının testi için, eşitlikte bağımsız değişken var ve yok (sadece sabit terimin bulunduğu) iken elde edilen D değerleri karşılaştırılmaktadır. Bu durumda D'de meydana gelen değişim, modelin içerdiği bağımsız değişkenlere bağlı olmakta ve D'deki değişme o bağımsız değişkenin modele katkısını göstermektedir. Bu durum için,

$$G = D\left(\text{Bağımsız değş. içermeyen model}\right) - D\left(\text{Bağımsız değş. içeren model}\right) \quad (3.30)$$

yazılabilir.

Eş. 3.30'da her iki D değeri için de doymuş modelin olabilirliği aynı olacağından G değeri,

$$G = -2\ln\left(\frac{\text{Değişkensiz modelin olabilirliği}}{\text{Değişkenli modelin olabilirliği}}\right) \quad (3.31)$$

kullanılarak hesaplanabilir. Lojistik regresyonda bu istatistik değeri, doğrusal regresyondaki F testi gibi işlem görmektedir.

Tek bağımsız değişkenli bir model için, $H_0 : \beta_1 = 0$ hipotezi altında G istatistiği; 1 serbestlik dereceli Ki-Kare dağılımı, çok değişkenli modellerde ise p (parametre sayısı) serbestlik derecesi ile Ki-Kare dağılımı izlemektedir.

Olabilirlik oran testine istatistiksel olarak benzer testler de mevcuttur. Bu testler Rao (1973) tarafından ortaya atılan, Wald test ve Score test olarak bilinmektedirler [1].

3.5.2. Wald ve score test

Wald test, β_1 eğim parametresinin en çok olabilirlik tahmininin bu parametrenin standart hata tahminine oranlanmasıyla elde edilmektedir. Elde edilen bu oran $H_0 : \beta_1 = 0$ hipotezi altında standart normal dağılıma yaklaşmaktadır. Wald test istatistiği,

$$W = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \tag{3.32}$$

eşitliği ile hesaplanmaktadır.

Wald testi sonucu bulunan W değeri z dağılımı göstermekte ve standart normal dağılıma ait tablo değeri ile karşılaştırılmaktadır. Wald test istatistiği değerleri kullanılarak modeldeki bağımsız değişkenlerin ayrı ayrı anlamlı olup olmadıkları belirlenmektedir.

Büyük β_1 değerleri için standart hataların tahmininin arttırılması Wald istatistiğinin dezavantajıdır. Bu durum H_0 hipotezi yanlış iken reddedilmesi konusunda yanılgıya neden olmaktadır. Hem olabilirlik oran testi G, hem de Wald testi W, β_1 için en çok olabilirlik tahmininin hesaplanmasına gerek duymaktadır.

Score testi ise, bu şekilde hesaplamalar gerektirmeyen bir yöntemdir. Ancak bu testin hazır paket programlarda fazla bulunmaması kullanılma sıklığını kısıtlamaktadır. Log-olabilirliklerinin türevlerinin dağılım teorilerine dayanan Score testi genel olarak, matris hesaplamalarını gerektiren çok değişkenli bir testtir [1].

Score test için test istatistiği (ST):

$$ST = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y}) \sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{3.33}$$

İle ifade edilmektedir.

3.5.3. Modelin uygunluğunun belirlenmesinde kullanılan diğer uyum iyiliği testleri

Modelde bulunması gereken tüm değişkenlerin modele dahil edilmesinden sonra, modelin bağımlı değişkeni açıklamadaki etkinliğinin araştırılması uyum iyiliği kontrolü olarak bilinmektedir. Lojistik regresyon modelinin uygunluğunu açıklamada kullanılan bazı test istatistiklerine aşağıda yer verilmiştir.

– 2 LogL istatistiği

$-2\,LogL$ istatistiği bağımlı değişkendeki açıklanmayan varyansın anlamlılığını göstermektedir. Log olabilirlik değeri 0–1 aralığında değerler almaktadır. Bu oran, bağımlı değişkenin bağımsız değişkenler tarafından tahmin edilme olasılığını göstermektedir. 1'den küçük sayıların logaritması 0 ile −∞ arasındaki $LogL$ istatistiği, en çok olabilirlik algoritması ile tahmin edilmektedir. $-2\,LogL$ istatistiği yaklaşık olarak Ki-Kare dağılımına uyduğundan, lojistik regresyon analizindeki $-2\,LogL$ istatistiği, regresyon analizindeki hata kareleri toplamına benzemektedir. Yani olabilirlik oranı 1 ise, $-2\,LogL$ istatistiği sıfıra eşit olmaktadır. Model, verileri tam olarak temsil ederse olabilirlik oranı 1 ve dolayısıyla $-2\,LogL$'nin sıfır olması demektir, bu durum da daha küçük $-2\,LogL$ istatistiğinin her zaman daha iyi bir modeli göstermekte olduğu sonucunu ortaya çıkarmaktadır [48].

Modele ilişkin Ki-Kare testi (C istatistiği)

Bu istatistik lojistik regresyon modelini genel olarak test etmekte yani sabit terim dışındaki tüm β katsayılarının sıfır olup olmadığını kontrol etmektedir. Bu haliyle de doğrusal regresyondaki F testine benzemektedir.

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$

H_1 :En az bir β_k sıfırdan farklıdır.

Modele ilişkin hesaplanan bu Ki-Kare istatistiği temel olarak olabilirlik oranı prensibine dayanmaktadır.

$L_0 : \beta_0$ dışındaki bütün β katsayılarının 0 olduğu durumdaki olabilirlik değeri,

L_1 : Uygun modelin olabilirlik fonksiyonu değeri olmak üzere;

$$C = -2\log(\frac{L_0}{L_1}) \tag{3.34}$$

biçiminde tanımlanmaktadır.

Hipotezde β_0 katsayısının sıfıra eşitliği sınanamadığından C istatistiği, incelenen modelin parametre sayısı ile yalnız sabit terimli modelin parametreleri arasındaki farka eşit $(k-1)$ serbestlik derecesi ile Ki-Kare dağılımına uymaktadır [49].

Doğru sınıflandırma oranı (Correct Classification Rate)

Lojistik modellerin sınıflandırma çizelgelerinden elde edilen doğru sınıflandırma oranı da bir uyum iyiliği testi olarak kullanılmaktadır. Sınıflandırma çizelgelerinde bağımlı değişkenin gerçek değerleri ile tahmin

değerleri çaprazlanmakta ve 0,5 eşik değerine göre sınıflama yapılmaktadır. Gözlemler 0,5 tahmin değerini aşarsa 1, diğer hallerde ise 0 grubuna atanmaktadır.

Doğru sınıflandırma oranını uyum iyiliği testi olarak kullanmanın bazı dezavantajları söz konusudur. Sürekli olarak tahmin edilen bağımlı değişken, kritik değerleri yardımı ile kesikli hale getirilmektedir. Pratikte P = 0,48 ve P = 0,52 arasında çok az fark olmasına rağmen, 0,5 kritik değeri ile karşılaştırıldığında yakın olan bu değerler faklı gruplara atanacaktır. Öte yandan sınıflandırma çizelgesinde daha çok gözlemli gruba atama yapılmakta olup bu da model uyumundan bağımsızdır. Amacı sınıflandırma olan bu çizelgenin kullanımı uyum iyiliği testine sadece bir katkı sağlamaktadır [50].

3.6. Lojistik Regresyon Modelinin Katsayılarının Açıklanması

Önceki bölümde anlatılan lojistik modelin parametrelerinin anlamlılık testleri ve uygun modelin belirlenmesinde kullanılan uyum iyiliği testlerinden sonra uygun model için elde edilen katsayı tahminlerinin yorumlanması gerekmektedir. Bilindiği üzere 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 lojit dönüşümdür ve Eş. 3.18 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 [1].

Çizelge 3.1. 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 (π) ise, diğer olayın olma olasılığı (1-π) olacaktır. Odds w ile gösterilecek olursa Odds oranı bu iki olasılığın oranlanması ile bulunmaktadır.

$$w = \frac{\pi}{1-\pi} \tag{3.35}$$

Çizelge 3.1'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;

$$x\text{=1 için: } w_1 = \frac{P(1)}{1-P(1)}, \qquad x\text{=0 için: } w_0 = \frac{P(0)}{1-P(0)} \tag{3.36}$$

olarak tanımlanabilir. Bu durumda Odds oranı ise;

$$\Psi = \frac{w_1}{w_0} = \frac{\dfrac{P(1)}{1-P(1)}}{\dfrac{P(0)}{1-P(0)}} \tag{3.37}$$

olacaktır. Odds oranı bağımlı değişkenin y=1 görülme olasılığı bakımından, x=1 olanları, x=0 olanlarla karşılaştırmada kullanılır. Çizelge 3.1 için odds oranına bakılacak olursa:

$$\psi = \frac{(\dfrac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}})(1+e^{\beta_0+\beta_1})}{(\dfrac{e^{\beta_0}}{1+e^{\beta_0}})(1+e^{\beta_0})} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1} \tag{3.38}$$

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

$$\ln \psi = \ln e^{\beta_1} = \beta_1 \tag{3.39}$$

katsayısına eşittir.

Odds oranı, nispi risk[3] 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 gösteriyorsa 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;

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

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

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.

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 [51]. Özet olarak, lojistik regresyon katsayısı ve olasılık oranı arasındaki ilişki, lojistik regresyon sonuçlarını açıklamamız için bir temel teşkil etmektedir.

4. UYGULAMA

Çalışmanın bu bölümünde, çok geniş bir çalışma alanı olan veri madenciliğinin sınıflama ve regresyon modellerine ilişkin teknikleri ile sınırlandırılan, Sosyal Güvenlik Kurumu 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, 2007–2009 yılları arası Sosyal Güvenlik Kurumu ilaç 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[4] 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.

[4] 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[5] 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:

[5] 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:

i. Özel Hastaneler: Özel Hastaneler Yönetmeliğine göre ruhsat almış hastaneler, “Ayakta Teşhis ve Tedavi Yapılan Özel Sağlık Kuruluşları Hakkında Yönetmelik” kapsamında açılan özel poliklinikler ve tıp merkezleri ile “Ayakta Teşhis ve Tedavi Yapılan Özel Sağlık Kuruluşları Hakkında Yönetmelik”in geçici ikinci maddesine göre faaliyetlerine devam eden tıp merkezleri ve dal merkezleridir.

ii. 2. Basamak Sağlık Bakanlığı Hastaneleri: Eğitim ve araştırma hastanesi olmayan devlet hastaneleri ve dal hastaneleri ile bu hastanelere bağlı semt poliklinikleri, entegre ilçe hastaneleri, belediyelere ait hastaneler ile kamu kurumlarına ait tıp merkezleri ve dal merkezleridir.

iii. 3. Basamak Sağlık Bakanlığı Hastaneleri: Sağlık Bakanlığına bağlı eğitim araştırma hastaneleri ve özel dal eğitim ve araştırma hastaneleri ile bu hastanelere bağlı semt polikliniklerdir.

iv. Üniversite Hastaneleri: Üniversite hastaneleri ile bu hastanelere bağlı sağlık uygulama ve araştırma merkezleri, enstitüler ve semt polikliniklerdir.

Bu grup için kategorilerin değer atamaları ise:

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- 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 sınıflandırılmıştır. Bu sınıflandırma için CART ve lojistik regresyon teknikleri kullanılmış ve sonuçta iki yöntemden elde edilen doğru sınıflandırma oranları karşılaştırılmıştır. Çalışmaya alınan 6.772.313 hastaya ait tanımlayıcı bilgiler aşağıdaki çizelgelerde verilmiştir.

Çizelge 4.1. Hastaların penisilin kullanma durumlarına göre dağılımı

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

Çizelge 4.1 incelendiğinde; solunum sistemi hastalıkları için antibiyotik kullanan 6.772.313 hasta içerisinden 2.484.352 hastanın penisilin grubu antibiyotik kullandığı, 4.287.961 hastanın ise penisilin grubu antibiyotik kullanmadıkları görülmektedir.

Çizelge 4.2. Hastaların tanı koduna göre dağılımı

Bağımsız Değişken		Sayı	Yüzde (%)
Tanı Kodu	711	2.620.561	38,7
	712	116.141	1,7
	713	7.366	0,1
	714	972.023	14,4
	715	24.997	0,4
	718	6.082	0,1
	720	29.457	0,4
	722	14.728	0,2
	724	2.729.664	40,3
	728	201.805	3.0
	744	49.489	0.7

Seçilen tanılar içerisinde ise 724 tanı kodlu akut bronşit % 40,3 oranı ile diğer tanı gruplarına göre hasta sayısının en fazla olduğu gruptur. Ayrıca 711 tanı koduna karşılık gelen akut tonsilitin % 38,7 oranı ile hasta sayısı bakımından ikinci büyük grup olduğu gözlenmektedir.

Çizelge 4.3. Antibiyotik kullanan hastaların hastane türüne göre dağılımı

Bağımsız Değişken		Sayı	Yüzde (%)
Hastane Grubu	Özel Hastaneler (1)	1.886.064	27,8
	2. Basamak Sağlık Bak. Hastaneleri (2)	4.001.971	59,1
	3. Basamak Sağlık Bak. Hastaneleri (3)	759.926	11,2
	Üniversite Hastaneleri (4)	124.352	1,8

Reçeteye solunum sistemi hastalıklarında yazılan antibiyotikler hastane türü bakımından değerlendirildiğinde; en fazla antibiyotiğin % 59,1 oranıyla 2. Basamak Sağlık Bakanlığı hastanelerinde, en az antibiyotiğin ise % 1,8 oranı ile üniversite hastanelerinde yazıldığı gözlenmektedir. Ayrıca çizelge 4.4'te de yine en fazla antibiyotiğin % 53,2 oranı ile kadın hastalara yazıldığı görülmektedir.

Çizelge 4.4. Hastaların cinsiyete göre dağılımı

Bağımsız Değişken		Sayı	Yüzde (%)
Cinsiyet	Kadın (1)	3.605.511	53,2
	Erkek (2)	3.166.802	46,8

Çizelge 4.5. Hastaların kullandığı ilacın fiyata göre dağılımı

Bağımsız Değişken		Sayı	Yüzde (%)
İlacın Fiyatı (Fiyat_G)	<= 5 TL (1)	3.007.899	44,4
	5 TL - 25 TL (2)	1.555.873	23
	> 25 TL (3)	2.208.541	32,6

Solunum sistemi hastalarının kullandığı antibiyotiğin fiyata göre dağılımına bakıldığında; % 44,4 oranı ile düşük fiyatlı (5 TL'den az) antibiyotiklerin reçeteye en fazla yazılan antibiyotik olduğu görülmektedir. Ancak fiyatı 25

TL'den fazla olan yüksek fiyatlı antibiyotiklerin de % 32,6 gibi oldukça yüksek bir oranda yazıldığı çizelge 4.5'ten anlaşılmaktadır.

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.

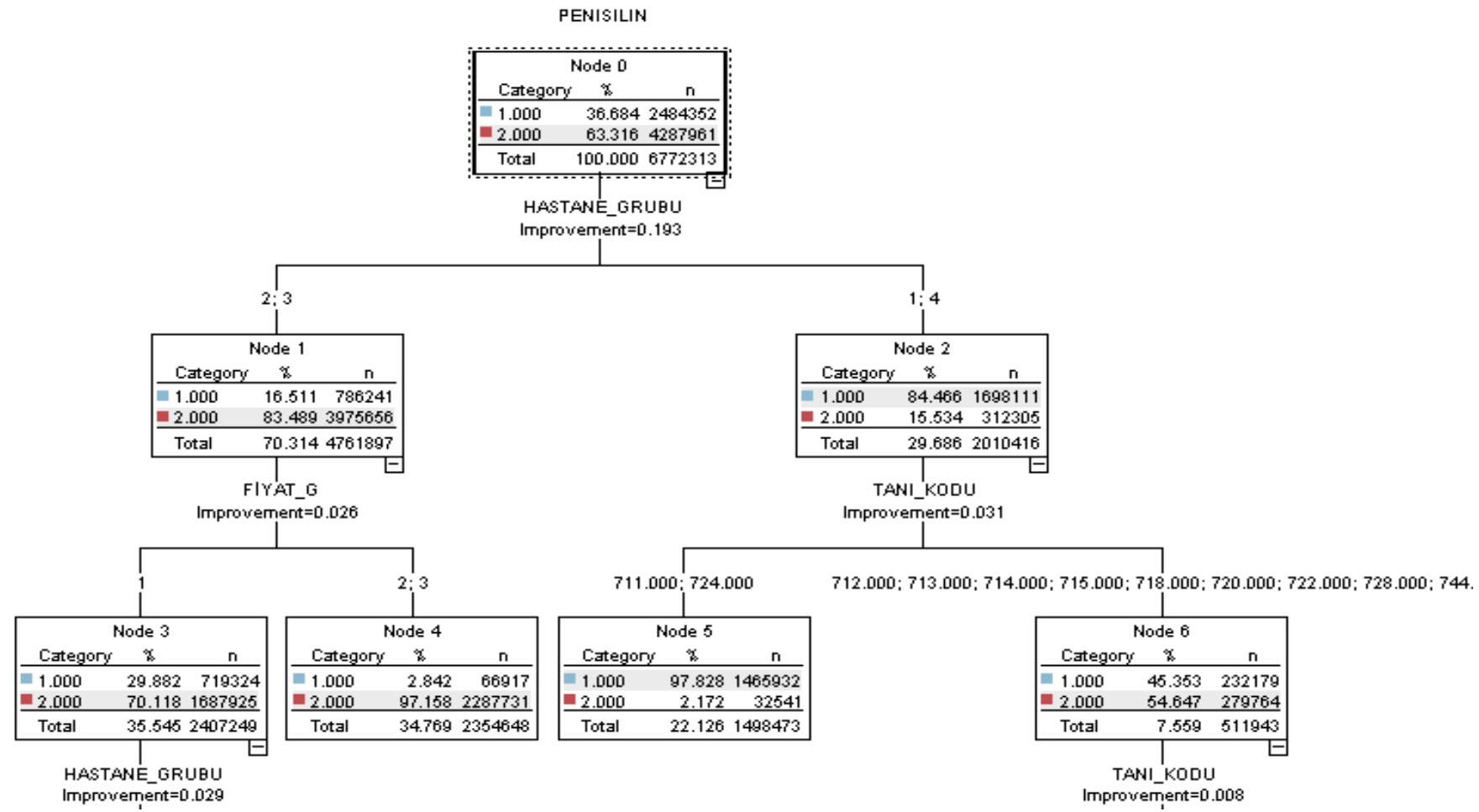
Bu doğrultuda analizlerde tüm verinin % 70'i model oluşturmak amacı ile eğitim verisi, geri kalan % 30'u ise sınıflama kurallarının doğruluğunu test etmek amacıyla test verisi olarak kullanılmıştır. Böylece sınıflandırma modeli öğrenme kümesi üzerinde geliştirilmiş olup test verisinden oluşan sınama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

4.1. CART Analizi Uygulaması

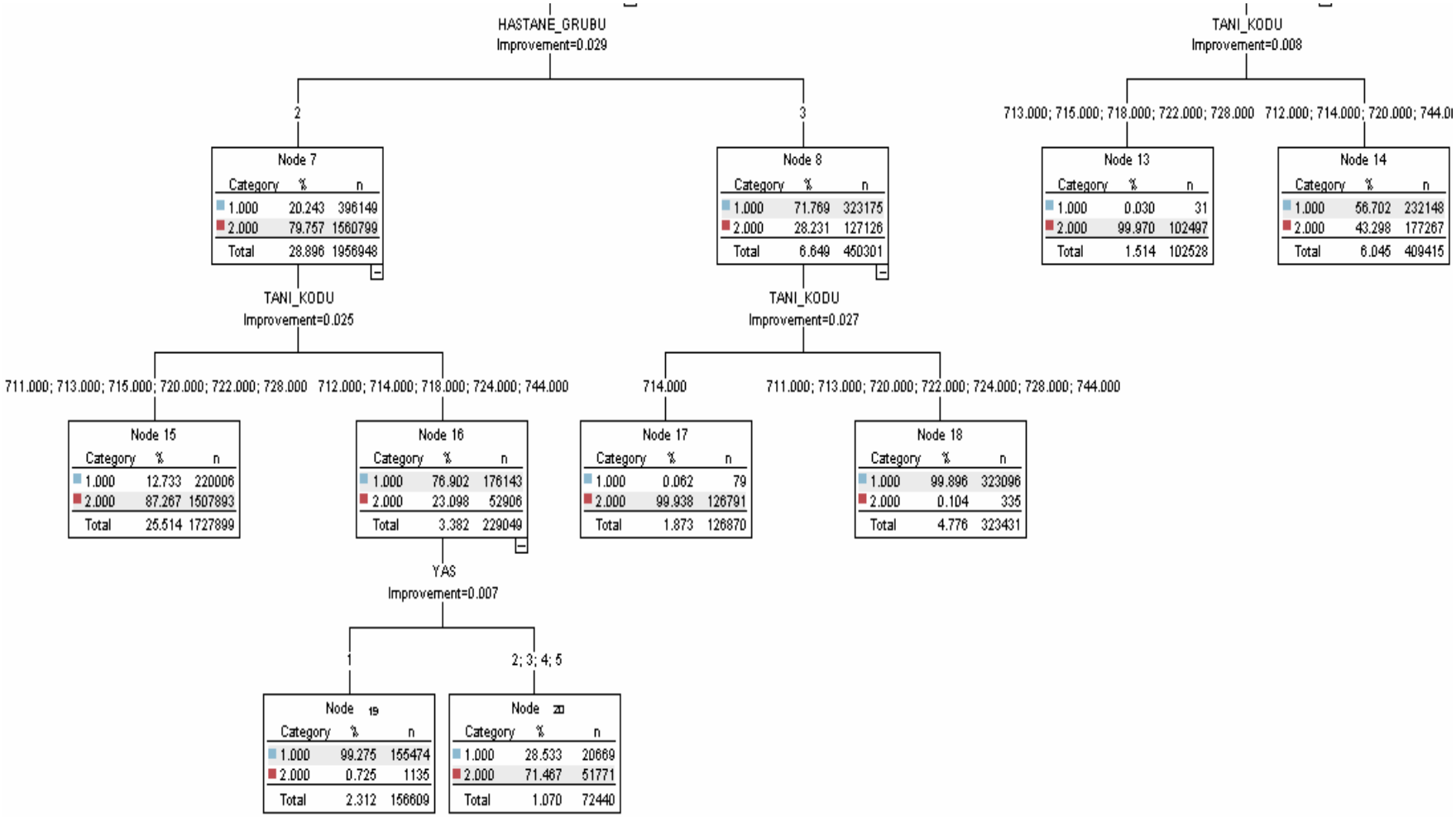
Bu tekniğin uygulanmasında öncelikle her değişken için hazırlanmış olan veri kümeleri clementine 12.0 yazılımına yüklenmiş, değişkenlerin seçimi “select” düğümü ile yapılmış ve modele girmesi istenmeyen değişkenler de “filter” düğümü ile elenmiştir. Ayrıca yaş ve fiyat gibi değişkenler “derive” düğümü 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 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 4.1’de gösterilmiştir.



Şekil 4.1. CART analizi sonucu karar ağacı



Şekil 4.1. (Devam) CART analizi sonucu karar ağacı

Şekil 4.1 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 4.1 ve çizelge 4.6'dan yararlanarak aşağıda açıklanmaktadır.

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.

Çizelge 4.6. 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	

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.

Fiyatı 5 TL'nin altında, Sağlık Bakanlığı 2. basamak hastanelerinde ve 711 tanı kodlu akut tonsillit, 713 tanı kodlu akut obstrüktif larenjit ve epiglottit, 715 tanı kodlu belirlenmiş influenza virüsüne bağlı influenza, 720 tanı kodlu bakteriyel pnömoni, 722 tanı kodlu başka yerde sınıflanmamış hastalıklarda bulunan Pnömoni ve 728 tanı kodlu kronik rinit, nazofarenjit ve farenjit hastalarına yazılan antibiyotikler de profil 3'ü oluşturmaktadır. Profil 3'te yer alan toplam 1.727.899 antibiyotik yazılan hasta içerisinden penisilin grubu antibiyotik kullananların sayısı 220.006 olup bunların oranı % 13'tür.

Birinci temel profilde Sağlık Bakanlığı 2. ve 3. basamak hastaneleri için yapılan temel sınıflandırma içerisinden fiyatı 5 TL'nin altında, 3. Basamak hastanelerde ve 714 tanı kodlu akut üst solunum yolu enfeksiyonları birden fazla olan hastalara yazılan toplam 126.870 antibiyotiğin % 0,06'sını oluşturan penisilin grubu antibiyotikler profil 4'ü (Node 17) oluşturmaktadır.

Ayrıca aynı değişkenler için yapılan sınıflamada, 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.

Solunum sistemi hastalıkları için Sağlık Bakanlığı 2. ve 3. basamak hastanelerinde yazılan ve fiyatı 5 TL'nin üzerinde olan antibiyotikler de profil 6'yı (Node 4) oluşturmakta ve bu profilde yer alan toplam 2.354.648 antibiyotik kullanan hastanın yaklaşık % 3'ünü penisilin grubu antibiyotik kullanan hastalar oluşturmaktadır.

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 penisilin grubu antibiyotiklerin 712 tanı kodlu akut larenjit, 713 tanı kodlu akut obstrüktif larenjit ve epiglottit, 714 tanı kodlu akut üst solunum yolu enfeksiyonları birden fazla olan, 715 tanı kodlu belirlenmiş influenza virüsüne bağlı influenza, 718 tanı kodlu streptococcus pneumoniae'ye bağlı Pnömoni, 720 tanı kodlu bakteriyel pnömoni, 722 tanı kodlu başka yerde sınıflanmamış hastalıklarda bulunan Pnömoni, 728 tanı kodlu kronik rinit nazofarenjit ve farenjit ile 744 tanı kodlu bronşiektazi hastalarına yazıldığı görülmektedir.

Bu iki profil arasından 713 tanı kodlu akut obstrüktif larenjit ve epiglottit, 715 tanı kodlu belirlenmiş influenza virüsüne bağlı influenza, 718 tanı kodlu streptococcus pneumoniae'ye bağlı Pnömoni, 722 tanı kodlu başka yerde

sınıflanmamış hastalıklarda bulunan Pnömoni ve 728 tanı kodlu kronik rinit, nazofarenjit ve farenjit hastalarına yazılan antibiyotikler profil 7'yi oluşturmaktadır. Profil 7'de yer alan toplam 102.528 antibiyotik yazılan hasta içerisinde penisilin grubu antibiyotik kullananların sayısı oldukça düşük olup bu hastaların oranı % 0,03'tür.

Benzer şekilde profil 8'i (Node 14) oluşturan 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 ise penisilin grubu antibiyotikleri kullananların oranının % 57 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 açıklanan profillerin oluşturulduğu 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 çizelge 4.7 ve 4.8'de söz konusu oranların sunulduğu sınıflandırma tabloları ortaya çıkmıştır.

Çizelge 4.7. 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%**

Çizelge 4.7'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 2.176.650'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.

Çizelge 4.8. Test verisi üzerinden elde edilen doğru sınıflandırma oranı tablosu

			Tahmin Değerleri			
			Penisilin			
			Kullananlar	Kullanmayanlar	Toplam	**Doğru Sınıflandırma Oranı**
Gerçek Değerler	Penisilin	Kullananlar	932.928	131.922	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.

4.2. 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 çizelge 4.9'da gösterilmiştir.

Çizelge 4.9'da 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 $p = 0,103 > 0,05$).

Çizelge 4.9. 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)	.	.	.	.	.	.	.	.

Çizelge 4.9'da 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.

Çizelge 4.10. 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

Oluşturulan lojistik modelde bağımlı değişken ile bağımsız değişkenler arasında ilişkiyi gösteren Nagelkerke R^2 değeri ise 0,729 olarak bulunmuştur. Bu da modelde kullanılan bağımsız değişkenlerin bağımlı değişkeni açıklama oranını yani kurulan lojistik modelin kullanılan değişkenlerle açıklanma oranının % 79,2 olduğunu göstermektedir. Bu açıklama oranının yeterince yüksek ve modelin anlamlı olması dışarıda kalan değişkenlerin çok fazla önemli olmadığını, modele alınan değişkenlerin yeterli olduğunu göstermektedir.

Çizelge 4.11'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 çizelge 4.12'de gösterilmiştir.

Çizelge 4.11. 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%**

Çizelge 4.11'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.

Çizelge 4.12. 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	2.904.039	**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 çizelge 4.9'da 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 çizelge 4.13'de gösterildiği gibi hesaplanabilmektedir.

Çizelge 4.13. 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 çizelge 4.13'de görüldüğü üzere, söz konusu özellikleri taşıyan bir hasta için reçeteye yazılan bir antibiyotiğin penisilin grubu antibiyotik olma olasılığı % 99,8 olarak hesaplanmıştır.

4.3. 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. Çizelge 4.14'te her bir modelin üretilmesi için gerekli işlem süresi, modele ilişkin sınıflandırma ve öngörü başarısı yer almaktadır.

Çizelge 4.14. 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.

Çizelge 4.14'te 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 çizelge 4.14 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.

5. SONUÇ VE ÖNERİLER

Son yıllarda yaygın olarak kullanılmaya başlanan ve büyük veri kümeleri içinde saklı durumda bulunan işlenmemiş bilgiyi 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 2007–2009 yılları arası Sosyal Güvenlik Kurumu 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.

Çalışma kapsamında oluşturulan lojistik modelde, bağımlı değişken ile bağımsız değişkenler arasında ilişkiyi gösteren Nagelkerke R^2 değeri 0,729 olarak bulunmuştur. Bu da modelde kullanılan bağımsız değişkenlerin bağımlı değişkeni açıklama oranını yani kurulan lojistik modelin kullanılan

değişkenlerle açıklanma oranının % 79,2 olduğunu ve bu oranın yeterince yüksek ve modelin anlamlı olması da dışarıda kalan değişkenlerin çok fazla önemli olmadığını, modele alınan değişkenlerin yeterli olduğunu göstermiştir.

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 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.

KAYNAKLAR

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

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

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

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

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

6. Koyuncugil, A. S., "Borsa Şirketlerinin Sektörel Risk Profillerinin Veri Madenciliğiyle Belirlenmesi", ***Sermaye Piyasası Kurulu Araştırma Raporu, Ankara,*** 1-29 (2007).

7. 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).

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

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

10. Coulter, D. M., Bate, A., Meyboom, R. H. B., Lindquist, M., Edwards, I. R., "Antipsychotic drugs and heart muscle disorder in international pharmacovigilance: Data Mining Study", ***BMJ,*** 322 (7296): 1207–1209 (2001).

11. Honigman, B., Light, P., Pulling, R. M., Bates, D. W., "A computerized method for identifying incidents associated with adverse drug events in outpatients", ***International Journal of Medical Informatics***, 61 (1): 21-32 (2001).

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

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

14. Alpaydın, E., "Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri", ***Bilişim 2000 Eğitim Semineri,*** İstanbul, 1-10 (2000).

15. 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).

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

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

18. 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).

19. 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).

20. 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).

21. Quinlan, J. R., "C4.5: Programs for Machine Learning", ***Morgan Kaufman***, USA, 1-4 (1993).

22. Mehta, M., Agrawal, R., Rissanen. J., "SLIQ: A Fast Scalable Classifier for Data Mining", ***5th International Conference on Extending Database Technology,*** Fransa, 18-32 (1996).

23. Tunay, K. B., "Türkiye'de Paranın Gelir Dolasım Hızlarının Mars Yöntemiyle Tahmini", ***ODTÜ Gelişme Dergisi,*** 28 (2): 431-454 (2001).

24. 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)

25. 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).

26. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., "Classification and Regression Trees", ***Wadsworth International Group***, California, 1-15 (1984).

27. Haughton, D., Oulabi S., "Direct marketing modeling with CART and CHAID", ***Journal of Direct Marketing***, 7 (3): 16-26 (2006).

28. Lemon, S., C., Roy, J., Clark, M., A., Friedmann, P., D., Rakowski, W., " Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison With Logistic Regression", ***Annals of Behaivoral Medicine***, 26 (3) :172-81 (2003).

29. Çamdeviren, H., Yazıcı, A., C., Akkuş, Z., Buğdaycı, R., Sungur, M., A., "Comparison of Logistic Regression Model and Classification Tree: An Application to Postpartum Depression Data", ***Expert Systems with Applications***, 32: 987–994 (2007).

30. Türe, M., Tokatlı, F., Kurt, İ., "Using Kaplan-Meier Analysis Together With Decision Tree Methods (C&RT, CHAID, QUEST, C4.5 and ID3) in Determining Recurrence-Free Survival of Breast Cancer Patients", ***Expert Systems With Applications***, 36 (2): 2017-2026 (2009).

31. Kayri, M., Boysan, M., "Bilişsel Yatkınlık İle Depresyon Düzeyleri İlişkisinin Sınıflandırma ve Regresyon Ağacı Analizi İle İncelenmesi", ***Hacettepe Üniversitesi Eğitim Fakültesi Dergisi***, 34: 168-177 (2008).

32. Albayrak, A., S., Akbulut, R., "Sermaye Yapısını Belirleyen Faktörler: İMKB Sanayi Ve Hizmet Sektörlerinde İşlem Gören İşletmeler Üzerine Bir İnceleme", ***Dumlupınar Üniversitesi Sosyal Bilimler Dergisi***, 22 (2008).

33. 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).

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

35. Freeman, D. H., "Logistic Regression", Applied Categorical Data Analysis, ***Marcel Dekker Inc.***, New York, 238 (1987).

36. Bircan, H., “Lojistik Regresyon Analizi: Tıp Verileri Üzerine Bir Uygulama”, ***Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi***, 2 : 185-208 (2004).

37. Anderson, J.A., “Robust Inference Using Logistic Models”, ***Bulletin of the International Statistical Institute***, 35-53 (1983).

38. Lesaffre, E., Albert, A., “A Multiple Group Logistic Regression Diagnostics”, ***Applied Statistics***, 38 (3): 125-440 (1989).

39. Agresti, A., “Categorical Data Analysis 2nd ed.”, ***John Wiley & Sons***, Florida, (2002).

40. İnternet : King, G., Zeng, L., “Logistic Regression in Rare Events Data”, http://gking.harvard.edu/files/0s.pdf (2001).

41. Bergh, H., Baigi, A., Mansson, J., Mattsson, B., Marklund, B., “Predictive factors for long-term sick leave and disability pension among frequent and normal attenders in primary health care over 5 years”, ***Public Health***, 121 (1): 25–33 (2007).

42. Temple, J., “Explaining the private health insurance coverage for older Australians”, ***People and Place***, 12 (2): 13-23 (2004).

43. Tezcan, B., “Lojistik Regresyon Analizi Ve Sigortacılık Sektöründe Bir Uygulama”, Yüksek Lisans Tezi, ***Marmara Üniversitesi Bankacılık ve Sigortacılık Enstitüsü***, İstanbul, 2 (2006).

44. Tatlıdil, H., “Uygulamalı Çok Değişkenli İstatistiksel Analiz”, ***Cem Web Ofset***, Ankara, (1996).

45. Green W.H., “Econometric Analysis 2nd ed.”, ***Macmillan Publication Company***, New York, (1993).

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

47. Collet, D., “Modelling Binary Data”, ***Chapman & Hall***, Florida, (2003).

48. ÜRÜK, E., “İstatistiksel Uygulamalarda Lojistik Regresyon Analizi”, Yüksek Lisans Tezi, ***Marmara Üniversitesi Fen Bilimleri Enstitüsü***, İstanbul, 45-46 (2007).

49. Aldrich, J. H., Nelson, F. D., Sullivan, J. L., “Linear Probability, Logit and Probit Models”, ***Sage Publications***, California, 56 (1984).

50. Ünvan, Y. A., “Kosullu Lojistik Regresyon Çözümlemesi Ve Avrupa Birliği Verisi Üzerine Bir Uygulama”, Doktora Tezi, ***Hacettepe Üniversitesi Fen Bilimleri Enstitüsü***, 11-12 (2006).

51. Allison, D. P., “Logistic Regression Using The SAS System 2nd ed.”, ***SAS Institute***, (2000).

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : KIRAN, Zeynep Burcu

Uyruğu : T.C.

Doğum tarihi ve yeri : 30.09.1981, Nevşehir

Medeni hali : Bekar

Eğitim

Derece	**Eğitim Birimi**	**Mezuniyet tarihi**
Yüksek lisans	Gazi Üniversitesi /İstatistik Bölümü	2010
Lisans	Gazi Üniversitesi/ İstatistik Bölümü	2003
Lise	Yüce Fen Lisesi	1998

İş Deneyimi

Yıl	**Yer**	**Görev**
2005-2006	Trakya Üniversitesi	İstatistikçi
2006-2009	Sosyal Güvenlik Kurumu	Sos. Güv. Uz. Yrd.
2009-	Sosyal Güvenlik Kurumu	Sos. Güv. Uzmanı

Yabancı Dil

İngilizce