FARKLI SINIFLANDIRMA YÖNTEMLERİNİN KARŞILAŞTIRILMASI VE BİR UYGULAMA
YÜKSEK LİSANS TEZİ

Ebru EKİCİ

Anabilim Dalı: İstatistik

Danışman: Yrd. Doç. Dr. Nurhan HALİSDEMİR

ELAZIĞ – 2012

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

FARKLI SINIFLANDIRMA YÖNTEMLERİNİN KARŞILAŞTIRILMASI VE BİR UYGULAMA

YÜKSEK LİSANS TEZİ

Ebru EKİCİ
(101133101)

Anabilim Dalı: İstatistik
Programı: Olasılık Süreçleri Ve Olasılık Teorisi

Danışman: Yrd. Doç. Dr. Nurhan HALİSDEMİR

ELAZIĞ – 2012

T.C
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

FARKLI SINIFLANDIRMA YÖNTEMLERİNİN KARŞILAŞTIRIMASI VE BİR UYGULAMA

YÜKSEK LİSANS TEZİ

Ebru EKİCİ

(101133101)

Tezin Enstitüye Verildiği Tarih : 16 Temmuz 2012
Tezin Savunulduğu Tarih : 06 Ağustos 2012

Tez Danışmanı : Yrd. Doç. Dr. Nurhan HALİSDEMİR (F.Ü)

Diğer Jüri Üyeleri : Doç. Dr. Cemil ÇOLAK (İ.Ü)

Yrd. Doç. Dr. Mehmet GÜRCAN (F.Ü)

TEMMUZ-2012

ÖNSÖZ

Çalışmamda bana büyük katkısı olan başta danışman hocam sayın Yrd. Doç. Dr Nurhan Halisdemir olmak üzere, bana vaktini ayırıp bilgi ve yardımlarını esirgemeyen diğer tüm bölüm hocalarıma ve arkadaşlarıma teşekkür ederim.

Ebru EKİCİ

ELAZIĞ-2012

İÇİNDEKİLER

Sayfa No

ÖZET

Bu çalışmada, veri madenciliği yöntemlerinden (modelleri) biri olan sınıflama ve regresyon modelleri içerisinde yer alan karar ağacı algoritmalarından CHAID analizi ile daha klasikleşmiş bir metot olan lojistik regresyonun sınıflama özelliklerinin karşılaştırılması amaçlanmaktadır. Bu kapsamda 2009-2011 yılları arasında Fırat Üniversitesi Tıp Fakültesinden alınan diyabet hastalarına ait verilerle bir uygulama yapılmış ve çalışmaya alınan veri seti için CHAID analizinin lojistik regresyon analizine göre daha iyi bir doğru sınıflandırma oranına sahip olduğu görülmüştür.

Anahtar Kelimeler: Lojistik regresyon analizi, CHAID analizi,

SUMMARY

AN APPLICATION ON THE COMPARISON OF VARIOUS CLASSIFICATION METHODS

In this study, it is aimed to compare the CHAID analysis which a decision tree algorythm within regression models and a classification method in data mining with the logistic regression which is the more a classical method and its classification features. In this context, it was applied this methods on the data which were obtained from diabets who came to Firat University Medical Center between 2009 and 2011. As a result of, it was determined that CHAID analysis have a beter classification rate than logistic regression.

Keywords: Logistic regression analysis, CHAID analysis,

ŞEKİLLER LİSTESİ

Sayfa No

TABLOLAR LİSTESİ

Sayfa No

KISALTMALAR LİSTESİ

CHAID :Otomatik Ki-Kare Etkileşim Belirleme (Chi-Squared Automatic Interaction Detector)

CART :Sınıflama Ve Regresyon Ağaçları (Classification And Regression Trees)

VM :Veri Madenciliği

AST :Aspartat Aminotransferaz

ALT :Alanin Aminotransferaz

KVS :Kalp-Damar Hastalıkları (Kardiyovasküler Sistem)

LRA :Lojistik Regresyon Analizi

QUEST :Hızlı, Yansız, Etkin İstatistiksel Ağaç

SD :Serbestlik Derecesi

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 [2].

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 tekniklerdir [2].

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 [2].

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 analizi değil aynı zamanda karar ağaçları da kullanılmaktadır [2].

Bağımlı değişkenin kategorik var-yok (0-1; binary) olduğu durumlarda kullanılan lojistik regresyon tekniği ile modellemeler yapılıp, modele ilişkin risk faktörleri (odds ratio) de tahminlenebilmektedir. CHAID analizinde ise, bunun yapılabilme gücünün daha yüksek olduğu gözlenebilmektedir. Çünkü CHAID analizinde bağımsız değişkenlerin en üst düzeydeki etkileşimlerini modele alan algoritma sayesinde benzer özellikleri taşıyan karakterler aynı homojen düğümlere taşınmaktadır. Böylece, elde edilen karar ağacında bütün detaylar net bir şekilde izlenebildiği gibi, elde edilen regresyon denklemine ait parametrelerin de daha güvenilir sonuçlar vermesi beklenmektedir [18].

2. MATERYAL VE METOT

2.1. Veri Madenciliği

2.1.1. Veri Madenciliğinin Tanımı

Bilgisayar sistemleri, her geçen gün ucuzlaması ve güçlerinin giderek artması nedeniyle yaşamın her alanına hızla girmektedir. İşlemcilerin hızlanması, disk kapasitelerinin artması, bilgisayar ağlarındaki ilerleme sonucu her bilgisayarın başka bilgisayarlardaki verilere ulaşması olanağı, bilgisayarların çok büyük miktardaki verileri saklayabilmesine ve daha kısa sürede işleyebilmesine olanak sağlamaktadır. [7].

Teknolojinin büyük hızla gelişmesi sonucu bu şekilde durmadan büyüyen ve işlenmediği sürece değersiz gibi görünen veri yığınları oluşmaktadır. Bu veri yığınlarını, içlerinde altın madenleri bulunan dağlara benzetmek mümkündür. Bu madenlere ulaşmak için kullanılan yöntem ise, temelinde istatistik uygulamaları yatan " veri madenciliğidir ".

Veri madenciliği birleşik verilerdeki gizli bilgileri bulmak ve is uzmanlığını arttırmak amacıyla yapılan yeni bir karar destek analiz işlemidir. Bazı anahtar kelimeler kullanılarak 4 aşamalı ayrıntılı VM tanımı şöyledir [7].

1. VM, bir süreçtir.
2. VM, karar destek araçlarının niteliğini yüceltir.
3. VM, gizlenmiş bilgileri bulur.
4. VM, is uzmanları için kavrayış dağıtıcı bir sistemdir

Veri madenciliği veri kümesi içerisinde keşfedilmemiş örüntüleri bulmayı hedefleyen teknikler kolleksiyonunu betimlemektedir. Veri madenciliği, William Frawley ve Gregory Piatetsky-Shapiro (1991) tarafından, '... verideki gizli, önceden bilinmeyen ve potansiyel olarak faydalı enformasyonun önemsiz olmayanlarının açığa çıkarılması...' biçiminde yapılan bilgi keşfi tanımını destekler [14].

Veri tabanının genel özellikleri

- Veritabanları, gerçek dünyanın belli bir açısını temsil eden daha küçük bir dünyadır.
- Veritabanı mantıksal çerçevede birbiriyle tutarlı bir veri topluluğudur. Bu haliyle rastgele toplanmış yani belirli bir sıralama veya gruplama yapılmamış bilgilere veritabanı demek doğru değildir.
- Veritabanı önceden belirlenmiş bir amaca hizmet etmek üzere tasarlanır ve yapılır.

- Veritabanı, herhangi bir büyüklükte ve karmaşıklıkta olabilir.
- Veritabanı elle veya bilgisayar ile oluşturulup idare edilebilir.

2.1.2. Veri Madenciliğinin Amaçları

Veri madenciliği uygulamalarında amaçlar, sınıflandırma, kümeleme, bağıntı kurma, görselleştirme ve tahminleme başlıkları altında ele alınabilir. Bu amaçlara hizmet eden algoritmalar ve teknikler açısından da, veri madenciliği teknikleri karar ağaçları (decision trees), yapay sinir ağları (neural networks), genetik algoritmalar ve istatistiksel analizler olarak sınıflandırılmaktadır [21].

Kümeleme, kişileri, nesneleri, gözlemleri, benzer eğilim ve özelliklerine göre gruplamaktır. Kümelemenin, sınıflandırmadan farkı grupların ayrıştırılırken önceden tanımlanmış özelliklere göre sınıflandırılmamalarıdır. Yani bir değişkene bağımlı kalarak sınıflandırma söz konusu değildir. Kendi içlerinde homojen olan, ancak birbirlerinden farklı özellikler taşıyan gruplar oluşur [21].

Bağıntı teknikleri, birbiriyle ilişkili olan değişkenlerin ortaya çıkarılması ve aralarındaki bağlantının ne derece kuvvetli olduğunun belirlenmesine yöneliktir. Bağıntı kurulan değişkenler, müşterinin iki veya daha fazla özelliği olabilir veya aldığı ürün ya da hizmet grupları arasında bağıntı olabilmektedir [21].

Bağıntı analizi esasına dayanan ve çok kullanılan yöntemlerden birisi sepet analizidir (market basket analysis – MBA). Sepet analizi özellikle işlemsel veriyi ilişkilendirir. A ürününün alınmasıyla B ürününün veya C ürününün alınması arasında bir bağlantı olup olmadığı, varsa, bu bağlantının kuvvet ve önem derecesi ortaya çıkarılır. Amaç, A ürününü alan müşteriye B ürününü de sunmaktır. Bu yöntemle çapraz satış (cross selling) ve üst seviye satış (upselling) imkanı doğmaktadır [21].

Tahminde bulunmada, veriler önce analiz edilerek aralarındaki ilişkiler ve ne derece birlikte hareket edip etmedikleri incelenir. Diğer tekniklerden başlıca farkı, tahmin edilecek bağımlı bir değişkenin, ona etki eden birçok bağımsız değişken kullanılarak isabetli bir biçimde tahmin edilmesidir. Bu modeller sayesinde bağımsız değişkenlerin alacağı değere göre bağımlı değişkenin alacağı değer tahmin edilebilmektedir. Dolayısıyla bağımsız değişkenlerdeki değişkenliğe göre, gelecekte bağımlı değişkenin ne olacağına dair tahminler yapılır [21].

Görselleştirmede ise veriler grafiklerle sunulmaktadır. Grafik ve haritalar yardımıyla, verideki eğilimler, değişkenlik veya homojenlik, nerelerde kümelendiği veya ayrıştığı, nasıl bir seyir izlediği hakkında fikir sahibi olmak çok daha kolaydır [21].

2.1.3. 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 [2].

Örnek birkaç kullanım alanı aşağıdadır:

İşletme alanındaki uygulamalar: Banka müşterilerinin kredi durumları ve ödemeleri incelenerek aralarında riskli olan müşterilerin tespit edilmesi ve aynı risk grubuna düşebilecek diğer müşterilerin önceden tahmin edilebilmesi [30].

Eğitim alanındaki uygulamalar*:* Öğrencilerin performans ve memnuniyetlerinin arttırılması [30].

Tıp alanındaki uygulamalar*:* Bir ilacın hangi yaş grup hastalarında nasıl etki yaratacağı konusunda tahminde bulunabilme. Yeni bulunan kanser tedavi yöntemi için en uygun adayı belirleme [30].

Spor alanındaki uygulamalar*:* Bir basketbol takımına alınacak oyuncunun gelecekteki performansı ve takıma sağlayacağı yararları tahmin edilebilme [30].

Kütüphanecilik alanındaki uygulamalar*:* Bir müşterinin aldığı kitabı ne zaman getireceği ve bir sonraki gelişinde hangi kitabı seçeceği konusunda tahminler yapılabilme [30].

Turizm alanındaki uygulamalar*:* Bir bölgeye turistlerin niçin geldiğini tespit ederek reklam kampanyası düzenlemek ve bir sonraki sezonda turist sayısını arttırmak [30].

Web alanındaki uygulamalar*:* Geçmişteki ve su anki veriler analiz edilerek geleceğe yönelik tahminlerde bulunulabilir. Bu durum özellikle karlılık, ciro, pazar payı, hava durumu gibi analizlerde çok rahat kullanılabilir [30].

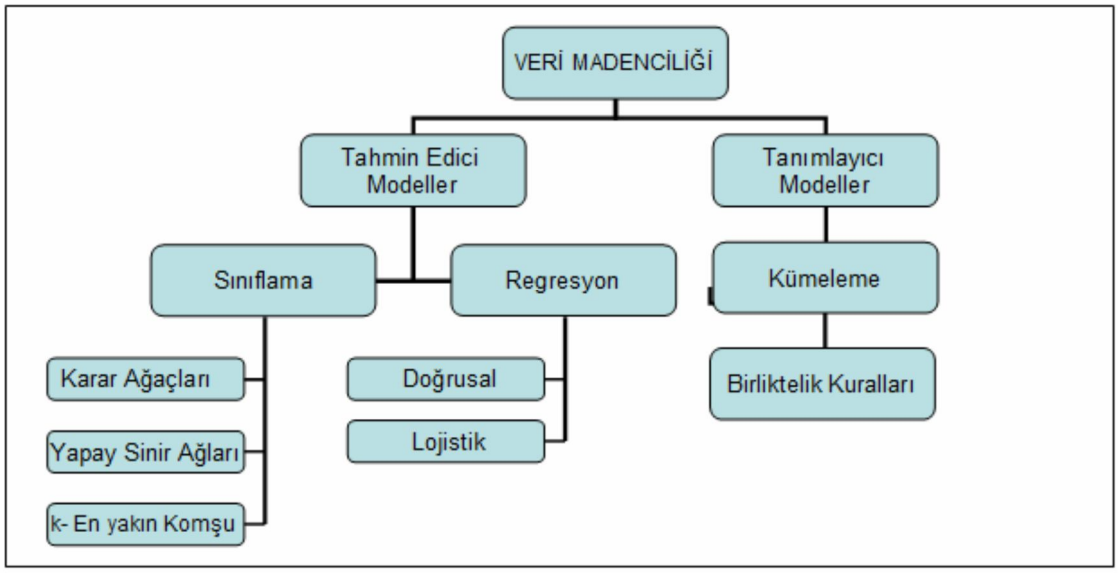
2.1.4. Veri Madenciliği Yöntem Ve Teknikleri (Modelleri)

Bir VM modeliyle aşağıdaki işlemlerden bir veya birkaçı gerçekleştirilebilir:

Sınıflama (Classification) ve Regresyon (Regression) Modelleri,

Kümeleme (Clustering) Modelleri,

Birliktelik Kuralları (Association Rules) ve Ardışık Zamanlı Örüntüler (Sequential Patterns). Sınıflama ve regresyon modelleri tahmin edici, kümeleme, birliktelik kuralları ve ardışık zamanlı örüntü modelleri tanımlayıcı modellerdir.



Şekil 2.1. Veri madenciliği modelleri[2].

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. Tanımlayıcı modellerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır [2].

2.1.4.1. Sınıflama Ve Regresyon Modelleri

Mevcut verilerden hareket ederek geleceğin tahmin edilmesinde faydalanılan ve veri madenciliği teknikleri içerisinde en yaygın kullanıma sahip olan Sınıflama ve regresyon, önemli veri sınıflarını ortaya koyan veya gelecek veri eğilimlerini tahmin eden modelleri

kurabilen iki veri analiz yöntemidir. Sınıflama kategorik değerleri tahmin ederken, regresyon süreklilik gösteren değerlerin tahmin edilmesinde kullanılır. Örneğin, bir sınıflama modeli banka kredi uygulamalarının güvenli veya riskli olmalarını kategorize etmek amacıyla kurulurken, regresyon modeli geliri ve mesleği verilen potansiyel müşterilerin bilgisayar ürünleri alırken yapacakları harcamaları tahmin etmek için kurulabilir [24-26].

Sınıflama ve regresyon modellerinde kullanılan baslıca teknikler,

- Karar Ağaçları (Decision Trees),
- Lojistik Regresyon (Logistic Regression),
- Yapay Sinir Ağları (Artificial Neural Networks),
- Genetik Algoritmalar (Genetic Algorithms),
- K-En Yakın Komsu (K-Nearest Neighbor),
- Bellek Temelli Nedenleme (Memory Based Reasoning),
- Naive-Bayes,

2.2. Veri Madenciliğinde Karar Ağaçları

Sınıflama ve regresyon modellerinin bir yöntemi olan karar ağaçları, kurulmasının ucuz olması, yorumlanmasının kolay olması, veritabanı sistemleri ile kolayca entegre edilebilmeleri ve güvenilirliklerinin iyi olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahip tekniktir. Karar ağacı, adından da anlaşılacağı gibi bir ağaç görünümünde tahmin edici bir tekniktir [1].

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 [2].

Karar ağaçları Bierman ve Friedman tarafından 1973 yılında önerilmiş olup değişkenleri parçalayarak bir ağaç oluşturmaya dayanmaktadır. Karar ağacında, tanımlanan sorunun cevabı gruplara ayrılmaktadır Soruya verilecek bir ölçüt belirlendikten sonra setler arasındaki riski maksimize edecek şekilde cevaplar bölünmektedir. En iyi bölünmeyi bulmak için her soruda bu işlem tekrar edilmektedir. Bir soru için grup oluşturulduktan ve gruplar arasındaki risk maksimize edildikten sonra oluşan iki grup için aynı işlemler devam

ettirilmektedir. Bu işlemler istatistik olarak anlamlı bir fark bulunana kadar devam ettirilip istatistik olarak anlamlı bir fark bulunmadığı durumlarda ise son verilmektedir. Ayrıştırma işlemi tamamlandıktan sonra ise o grup içerisinde yer alan gözlemlerin oranına göre grup değerlendirilmektedir [3,4].

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 [5,2].

Karar ağaçları oluşturulurken kullanılan algoritmanın ne olduğu önemli bir husustur. Kullanılan algoritmaya göre ağacın şekli değişebilir. Bu durumda değişik ağaç yapıları da farklı sınıflandırma sonuçları verecektir [2].

Karar ağacı karar düğümleri, dallar ve yapraklardan oluşur. Karar düğümü, gerçekleştirilecek testi belirtir. Bu testin sonucu ağacın veri kaybetmeden dallara ayrılmasına neden olur. Her düğümde test ve dallara ayrılma işlemleri ardışık olarak gerçekleşir ve bu ayrılma işlemi üst seviyedeki ayrımlara bağımlıdır. Ağacın her bir dalı sınıflama işlemini tamamlamaya adaydır. Eğer bir dalın ucunda sınıflama işlemi gerçekleşemiyorsa, o dalın sonucunda bir karar düğümü oluşur. Ancak dalın sonunda belirli bir sınıf oluşuyorsa, o dalın sonunda yaprak vardır. Bu yaprak, veri üzerinde belirlenmek istenen sınıflardan biridir. Karar ağacı işlemi düğümünden başlar ve yukarıdan aşağıya doğru yaprağa ulaşana dek ardışık düğümleri takip ederek gerçekleşir [1].

Ağaç oluşturmada yapılan işlemlerden bir tanesi de budama işlemidir. Budama ağaçta oluşmuş sonucu etkilemeyen ve sınıflanmaya herhangi bir katkısı olmayan dalların ağaçtan alınmasıdır. Bir bakıma gereksiz ayrıntıların sonuçtan çıkartılması işlemidir. Ağaçtan birçok düğüm ve dal oluşursa, ağacın alt dallara ve yapraklarına ulaşan veri sayısı da azalacaktır. Bu da ağacın hassasiyetini azaltacaktır [2,6].

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 [2,6].

- 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 [2,6].

Karar ağaçları yöntemlerinin güçlü yönleri ise şunlardır:

- Sınıflama yaparken çok küçük bir hesaplama gerektirir,
- Hem kategorik hem de sürekli değişkenler ile çalışabilir ayrıca karmaşık veri setlerine kolaylıkla uygulanabilir,
- Sınıflama ve tahmin için hangi değişkenlerin daha önemli olduklarını açık belirtiler ile gösterir,
- Parametrik olmayan bir model olduğu için varsayımları çok kısıtlıdır,
- Bağımlı ve bağımsız değişkenler arasındaki ilişki görsel sunuma sahip olduğundan, ağaç seklindeki model sonuçları çok fazla istatistik bilgisine gerek duyulmadan kolay bir şekilde yorumlanabilir.
- Tanımlanan bağımlı değişken için olabilecek bütün bağımsız değişkenleri ve onların tüm kombinasyonlarını modele katar ve mümkün olan en doğru sınıflandırmayı yapar. Değişken kombinasyonlarına da bakıldığı için interaksiyonlar da değerlendirilmiş olur.
- Hem bağımlı hem de bağımsız değişkenler için kayıp veya eksik değerler ile aşırı uç değerlerden etkilenmeyen bir metottur [5].

Karar ağacı temelli analizlerin yaygın olarak kullanıldığı alanlar şunlardır:

- Belirli bir sınıfın muhtemel üyesi olacak elemanların belirlenmesi (Segmentation),
- Çeşitli vakaların yüksek, orta, düşük risk grupları gibi çeşitli kategorilere ayrılması (Stratification),
- Gelecekteki olayların tahmin edilebilmesi için kurallar oluşturulması,
- Parametrik modellerin kurulmasında kullanılmak üzere çok miktardaki değişken ve veri kümesinden faydalı olacakların seçilmesi,
- Sadece belirli alt gruplara özgü olan ilişkilerin tanımlanması**,**
- Kategorilerin birleştirilmesi ve sürekli değişkenlerin kesikliye dönüştürülmesidir [7].

Karar ağacı temelli tipik uygulamalar ise,

- Hangi demografik grupların mektupla yapılan pazarlama uygulamalarında yüksek cevaplama oranına sahip olduğunun belirlenmesi (Direct Mail),
- Bireylerin kredi geçmişlerini kullanarak kredi kararlarının verilmesi (Credit Scoring),
- Geçmişte isletmeye en faydalı olan bireylerin özelliklerini kullanarak ise alma süreçlerinin belirlenmesi,
- Tıbbi gözlem verilerinden yararlanarak en etkin kararların verilmesi
- Hangi değişkenlerin satışları etkilediğinin belirlenmesi,
- Üretim verilerini inceleyerek ürün hatalarına yol açan değişkenlerin belirlenmesidir [5].

Karar Ağaçları yöntemleri, temel anlamda hedef (bağımlı) değişkeni, tahmin edici değişkenlere göre ayırma mantığına dayansa da; bünyesinde değişik amaçlara hizmet eden birbirinden farklı algoritmalara sahiptir. [5]

Karar ağacı oluşturmak için geliştirilen bu algoritmalar arasında CHAID (Chi-Squared Automatic Interaction Detector), Exhaustive CHAID, C&RT (Classification and Regression Trees), ID3, C4.5, QUEST (Quick, Unbiased, Efficient Statistical Tree), C5.0, yer almaktadır.

2.2.1. ID3

ID3 algoritması veritabanı bölünmeden önce doğru sınıflandırma yapmak için gelen bilgiyle, veritabanı bölündükten sonra doğru sınıflandırma için gelen bilgi arasındaki farkı kullanarak, öncelikli düğümü ve dallanmalara karar verir. Bu aradaki fark ise kazanım olarak adlandırılır. Gerçekten de veritabanı bölününce, yani dallanmalar oluştukça doğru sınıflandırma için gerekli bilgi sayısı da azalacaktır. [8].

Karar ağaçları yardımıyla sınıflandırma işlemlerini yerine getirmek üzere Quinlan tarafından birçok algoritma geliştirilmiştir. Bunlar arasında yer alan ID3 ve C4.5 algoritmalarıdır. Bu yöntemde karar ağacında hangi niteliğe göre dallanmanın yapılacağını belirlemek üzere entropi kavramına başvurulur. Bir sistemdeki belirsizliğin ölçüsüne "entropi "adı verilir [8]. Entropi kavramı, eldeki bilginin sayısallaştırılmasıdır. Entropi beklentisizliğin maksimumlaşmasıdır.

Verilerin ham halinin entropisi ile yani başlangıçtaki entropiyle her bir alt bölümün entropilerinin ağırlıklı toplamı arasındaki fark alınır. Bu fark hangi alt bölüm için büyükse o alt bölüme doğru dallanma yapılır.

ID3 algoritmasının yaptığı şey, ağacı doğru kurmaktır. Kurulu karar ağacının her seviyesinde geriye kalan bilgi gereksinimi (remaining information required) minimize edilmektedir. [8].

2.2.2. C4.5

C4.5 algoritması ID3 algoritmasına şu konular açısından üstünlük sağlamaktadır: Karar ağacı oluştururken kayıp veriler hesaba katılmaz. Yani, kazanım oranı hesaplanırken, sadece verileri eksik olmayan diğer kayıtlar kullanılır. C4.5 algoritması, kayıp verileri diğer veri ve değişkenler yardımıyla öngörerek kazanım oranın hesaplanmasında kullanır Böylece daha duyarlı ve daha anlamlı kurallar çıkartabilen bir ağaç üretilebilir. [9].

C4.5 ağacın budanması işlemi için iki ayrı yöntem kullanır: Alt ağaç yerleştirmesi ağaç içindeki alt-ağaçlar yapraklara dönüştürülür. Ancak, bu değişimin yapılabilmesi için konulacak yaprağın hatasının, alt ağacın hatasından düşük olması gerekir. İkinci yöntem ise bir alt-ağacın, bu ağacı en çok kullanan ağacın yerini almasıdır. Yine burada da hata oranlarının yapılan değişiklikten sonra düşmesi gerekmektedir. C4.5, tek değişkenli bir

karar ağacı algoritmasıdır. ID3 algoritmasında bazı eksiklikler ve sorunlar Quinlan'ın C4.5 geliştirdiği algoritmasıyla giderilmiştir. [9].

2.2.3. C5.0

C5.0 algoritması C4.5'in geliştirilmiş hali olup, özellikle büyük veri setleri için kullanılmaktadır. C5.0 algoritması doğruluğu arttırmak için boosting algoritmasını kullandığından boosting ağaçları olarak da bilinir C5.0 algoritması C4.5'e göre çok daha hızlı olup, hafızayı daha verimli kullanmaktadır [9].

2.2.4. QUEST

QUEST algoritması 1997 yılında Loh and Shih tarafından geliştirilmiştir. İkili karar ağacı yapısı kullanan bir sınıflandırma algoritmasıdır. İkili ağaç kullanılmasının sebebi, ikili ağaçlarda budama ve doğrudan durma kuralı gibi tekniklerin kullanılabilmesidir. QUEST algoritması ağacın oluşturulması sırasında değişken seçimi ve bölünmeyi eşzamanlı olarak yapan CHAID ve CART' ın aksine hepsi ile ayrı ayrı ilgilenir. QUEST algoritması, ağacın dallanması sırasındaki önyargılı seçimin daha genel hale getirilmesi ve hesaplama maliyetinin düşürülmesi amacıyla geliştirilmiştir. Fakat henüz sınıflandırmadaki doğruluk, ağacın büyüklüğü ve dallanmadaki değişiklik konularında diğerlerine açık bir üstünlük sağlayan sınıflandırma algoritması yoktur [9].

2.2.5. CART (C&RT)

CRT (CART ya da C&RT), sınıflama ve regresyon için kullanılan özyinelemeli bölme metodu Breiman tarafından 1984'te geliştirilmiştir. Ele alınan bağımlı değişkenin (hedef değişken, target variable) kategorik yapıda ise yöntem sınıflama ağaçları (Classification Tree, CT), sürekli ise regresyon ağaçları (Regression Tree, RT) olarak adlandırılmaktadır. Bağımsız değişken tektir ve sınıflayıcı, sıralayıcı ve sürekli türde; açıklayıcı değişken bir ya da daha fazla olabilir ve sınıflayıcı, sıralayıcı ve sürekli türde ölçülmüş olabilir [5].

CART ağaç modeli, tek değişkenli ikili kararların bir hiyerarşisini içerir. CART verileri iki alt kümeye ayırdığı için her bir alt küme içindeki durumlar, bir önceki alt kümeden daha homojen olacaktır. Bu ardışık süreç, homojenlik kriterine ulaşılıncaya veya

diğer bazı durma kriterleri sağlanıncaya değin kendini tekrar eder. Aynı kestirim değişkeni ağaçta farklı düzeylerde pek çok kez kullanılabilir. Ağacın yapısı önceden belirlenmemekte, verilerden türetilmektedir. CART, kök düğümünde, verilerin iki gruba bölünmesi için en iyi değişkenin seçilmesini sağlar ve farklı bölümlendirme (splitting) kriterleri kullanır. Bu bölümlendirme kriterlerinin tümü, her bir alt kümedeki sınıf etiketlerini mümkün olduğunca homojen olacak biçimde bölümlendirir [10].

Bölümlendirme prosedürü çocuk düğümlere (child node) veya alt düğümlerin her birine ardışık olarak uygulanır. CART ağaçları, kesin bir heterojenlik (impurity) ölçüsüne bağlı olarak düğümlere ayrılmış iki değerli (binary) ağaçlardır ve bu nedenle de sonuçta homojen dallar oluşmaktadır. Ağacın hedefi benzer veya aynı çıktı değerlerine sahip olma eğiliminde olan alt gruplar yaratmaktır. CART modelleri için bölünmelerin bulunmasında kullanılan dört farklı heterojenlik ölçüsü mevcuttur. Kategorik hedef değişkenler için Gini. Twoing veya (sıralayıcı hedef değişkenleri için) sıralı Twoing. sürekli hedef değişkenler için ise en küçük kareli sapma (LSD) hesaplamaları kullanılabilir [10]. Bu hesaplamalarda kar, maliyet değerleri ve değişken kategorileri arasındaki önceliklerin tanımlanabilmesi gibi sağlanan çeşitli esneklikler, CRT algoritmasının günümüzde yaygın olarak tercih edilmesine neden olmaktadır [3].

C&RT algoritmasının adımları su şekildedir:

1. İlk düğümden başlayarak bölünmeyi sağlayacak tüm mümkün adayların içinden bir tane ayıraç seçilir,
2. Hedef değişkenin tipine göre safsızlık ölçütü hesaplanır,
3. Açıklayıcı değişkenler safsızlık ölçütlerine göre karşılaştırılır,
4. Safsızlık ölçütünü maksimum olan değişkene göre ayrıştırma yapılır,
5. Ayıraç belirleme sürecine diğer açıklayıcı değişkenler içinde devam edilir,
6. Herhangi bir durdurma kuralına rastlayana kadar ağaç büyütülür.

Bu algoritma ile elde edilen maksimum ağaçtan optimum ağacı oluşturmak için budama yapılmalıdır. C&RT ile oluşan ağaçların budanmasında ayrı bir değerlendirme veri setine gerek vardır. Optimum ağaç hem hatalı ayırma riskini hem de ağacın karmaşıklığını ölçen bir indeks (maliyet-karmaşıklık ölçüsü) kullanarak bu indeks değerini minimum yapan ağaçtır [5].

CART analizi ağaç yapısına dayalı diğer sınıflama teknikleri ile kıyaslandığında çok sayıda avantaja sahiptir.

1. Parametrik olmayışıdır. Diğer bir söyleyişle ön kestirici veya aynı anlama gelmek üzere bağımsız değişken değerlerine ilişkin varsayımlar gerektirmemektedir. Bu nedenle CART analizinde kullanılacak değişkenler çok çarpık sayısal değişkenler olabileceği gibi, sınıflayıcı veya sıralayıcı yapıya sahip kategorik değişkenler de olabilir. Bu önemli bir özelliktir ve analizi yapacak araştırmacıya, normallik araştırma ve dönüşüm yapma gibi işlemler gerektirmediğinden zaman kazandırmaktadır.

2. CART analizi, ele alınan problem yüzlerce mümkün bağımsız değişken içerse bile, bölümlendirilecek tüm mümkün değişkenleri araştırma gücüne sahiptir.

3. Göreceli olarak otomatik bir makine öğrenim tekniği olması. Diğer bir söyleyişle analizin karmaşıklığı ile kıyaslandığında, araştırmacıya göreceli olarak az miktarda girdi gerekmektedir. Diğer çok değişkenli modelleme yöntemleri araştırmacılara çok fazla girdi gereksinimi yüklemekte, geçici sonuçların analizini gerektirmekte ve ilgili yöntemin modifikasyonu gerekmektedir.

4. CART, ele alınan veri kümesi eksik değerler içerdiğinde kullanışlı bir analizdir. Eksik değerler çok fazla olduğunda. Bu değerler bir vekil değişken olarak ağaç yapısında yer alırlar

5. İstatistikçi olmayanlar için bile yorumunun çok kolay olduğunu söylemek mümkündür.

Dezavantaj olarak ise; Çoklu değil de iki değerli ağaç tekniği olması dezavantaj olarak sayılabilir. Fakat değişken sayısı çok fazla olduğunda veya değişkenlerin çok fazla kategorisinin olması durumunda iki değerli ağaç yapısı daha yorumlanabilir sonuçlar üretebilir [10].

Bireyleri dallara/sınıflara atarken kullandıkları bölünme kriterlerine göre farklı karar ağaçları algoritmaları vardır. CART (Classification and Regression Tree) ve QUEST algoritmaları Gini indeksini, C4.5 ve C5 algoritması bilgi kazancını ve CHAID algoritması ise ki-kare istatistiğini kullanarak karar ağacını oluşturmaktadır [10].

2.2.6. CHAID ANALİZİ

CHAID yöntemi; bir populasyonu, bağımlı değişkendeki varyasyonu gruplar içi minimum, gruplar arası maksimum olacak şekilde farklı alt gruplara veya bölümlere

tekrarlı olarak ayıran ve değişkenler arasındaki etkileşim veya kombinasyonları bulan bir yöntemdir.

Bilimsel çalışmalardaki önemli istatistiksel konulardan biri, üzerinde durulan olayı önemli derecede etkileyen faktörlerin yanı sıra, bu faktörlerin hangi seviyesindeki etkinin yüksek olduğunu belirlemektir. Bağımlı değişkenlerdeki değişimi önemli derecede etkileyen faktörleri irdeleyen CHAID analizi, model içerisinde yer alan değişkenlerin etkileşim ve genel olarak ortak düzeydeki kombinasyonlarını da tespit etmeye çalışmaktadır [18].

CHAID metodu 1980'de Kaas tarafından en iyi bölmeyi hesaplamak için istatistik olarak anlamlı bir farklılığın olmadığı, hedef değişkene uyan çiftlerde tahmin değişkeninin olası kategori çiftini birleştirmesiyle oluşturulmuştur. C&RT algoritmasına benzemektedir; fakat veriyi bölümlere ayırırken farklı bir yol kullanmaktadır. En uygun bölümleri seçmek için kullanılan entropy veya gini metrikleri yerine chi-square testi kullanılmaktadır. Herhangi bir düğümdeki en iyi bölmeyi hesaplamak için tahmin değişkenleri kategorisinin herhangi kullanılabilir parçası hedef değişkene uyan bir çiftin içinde istatistik olarak anlamlı bir fark kalmayıncaya kadar birleştirilmektedir [3].

Bilindiği üzere tanımlayıcı istatistikler evren hakkında kısa özetler sunarken; süreç içerisinde yaşanabilecek değişimler hakkında araştırmacıya bir ön bilgi vermektedir. Buna karşın, analitik istatistiksel yöntemlerde kurulan model ile bir regresyon denklemi elde edilebilmektedir. Genel anlamda regresyon, modelde mevcut değişkenlerin ilişki yapılarını, bağımsız değişkenlerin bağımlı değişken üzerindeki teferruatlı etkilerini araştırmacıya bir tahminleme metodu olarak sunmayı hedeflemektedir [18].

CHAID, regresyon problemlerinde kullanılabileceği gibi karar ağaçlarının oluşturulmasında etkilidir. CHAID analizi ile elde edilecek bir regresyon denklemi, bilinen klasik varsayımlardan (normallik, doğrusallık, homojenlik vb.) bağımsız tutulmaktadır. Çünkü güçlü bir öteleme algoritması (iteration algorithm) ile bütün olan evren kararlı alt düğümlere (node) bölünebilmektedir. Bu işlem ayrıca verilerin dağılımında normalliği ve homojenliği sağlayabilmektedir. Değişkenler arasındaki ilişki lineer yapıdan daha karmaşık ise veride gizli olan bu ilişkiyi bulmak için verinin belli kısımlarını eleme yöntemi olan CHAID kullanılır."Ki-kare" ismini almasının nedeni algoritmasında birçok çapraz tablonun kullanılması ve istatistiksel önem oranları ile çalışmasıdır [5,11,20].

İnceleme altına alınan örneklem büyüklüğünün fazla olduğu durumlarda; evren içerisindeki homojenlik ilkesinin ihlalinden dolayı, gelişigüzel kurulan bir regresyon

modeline ait tahminleme (estimation) gerçeği yansıtmayabilir. Bundan dolayı bütünü parçalara bölmek ve sınıflandırılmış olası alt gruplar ile incelemeyi sürdürmek daha gerçekçi bir zemin oluşturabilmekte ve bu yolla istatistikte önemsenen homojenlik şartı da yerine gelebilmektedir. Yine bilindiği üzere, klasik regresyon denkleminde (parametrik durumlar için) normallik, doğrusallık, homojenlik ve toplanabilirlik gibi varsayımlar gerçekleşmeden bir analizin yapılması mümkün olamamaktadır. Bu anlamda CHAID analizi ile elde edilecek bir regresyon denklemi, bilinen klasik varsayımlardan muaf tutulmaktadır. Çünkü güçlü bir öteleme algoritması (iteration algorithm) ile bütün olan evren kararlı alt düğümlere (node) bölünebilmektedir. Bu işlem beraberinde, verilerin dağılımında normalliği ve homojenliği sağlayabilmektedir. Yine aynı şekilde CHAID analizi sürekli ve kategorik verileri, aynı anda modele alabilmeye olanak tanımaktadır. Bu nedenle CHAID analizi parametrik ve parametrik olmayan ayrımını kaldırmakta, yöntem algoritmasında istatistiksel olarak yarı parametrik (semi-parametric) özellik taşımaktadır[18].

CHAID ile diğer karar ağacı modelleri arasındaki en önemli farklılıklarından birisi, ağaç türetimidir. ID3, C4.5 ve CART ikili ağaçlar türetirken, CHAID ikili olmayan çoklu ağaçlar türetir. CHAID sürekli ve kategorik tüm değişken tipleriyle çalışabilmektedir. Bununla beraber, sürekli tahmin edici değişkenler otomatik olarak analizin amacına uygun olarak kategorize edilmektedir. CHAID, Ki-Kare metriği vasıtasıyla, ilişki düzeyine göre farklılık rastlanan grupları ayrı ayrı sınıflamaktadır. Dolayısıyla, ağacın yaprakları, ikili değil, verideki farklı yapı sayısı kadar dallanmaktadır [12].

CHAID Analizi, diğer metotlara göre kategorik ve sürekli değişkenler üzerinde çalışabilmesi, ağaçta her düğümü ikiden fazla alt gruba ayırabilmesi gibi avantajlarından dolayı günümüzde en yaygın tercih edilen metottur.

CHAID Analizi;

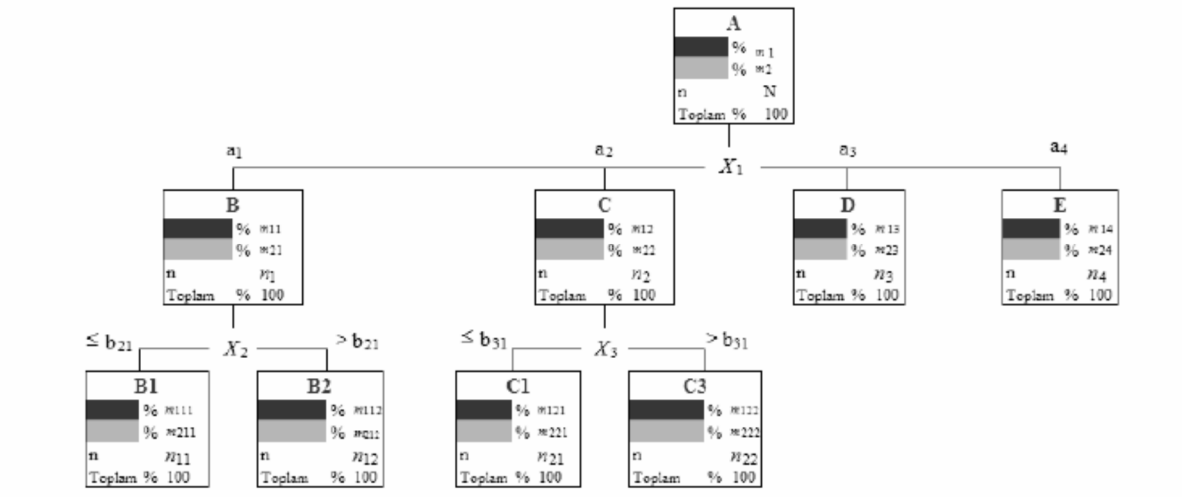
- Sınıflama ölçme düzeyinde ölçülmüş bir bağımlı değişkeni en iyi şekilde açıklamak için kullanılır,
- Açıklayıcı değişkenler sınıflayıcı, sıralayıcı ve aralıklı ölçek ile ölçülmüş olabilir, Bağımsız değişkendeki veri yapısının aynı tip ölçekle ölçülmüş olmasına gerek yoktur,
- Kayıp verileri yeni bir kategori gibi davranır ve bu kategoriyi p-değeri hesaplamalarına dahil eder,
- Kategorileri sıralanabilen ya da sıralanmayan, açıklayıcı değişkenlerin yer aldığı veri kümesini, bağımlı değişkene göre detaylı alt kümelere böler,

- Bu bölünme işlemini gerçekleştirirken, açıklayıcı değişkenlere ait kategorileri, bağımsız olarak yeniden düzenler, yani kategorileştirir
- Daha sonraki her bölünmeyi yeniden bağımsız olarak gerçekleştirir [5].

Yani CHAID analizi, çok kategorili değişkenlerin yer aldığı büyük bir veri kümesini, benzer kategorileri birleştirerek, önemli sayılan değişkenlere göre bölerek, bir bakıma önceki durumuna oranla özet şekilde tanımlamış olur. Her bir açıklayıcı değişken için kategorilerin anlamlı bir şekilde birleştirilmesinden sonra, bağımlı değişkene göre kontenjans tabloları oluşturularak, Bonferronni p değerleri ile Ki-Kare istatistikleri hesaplanır. Açıklayıcı değişkenler birbirleri ile karşılaştırılıp, en küçük Bonferronni p değerine sahip olan açıklayıcı değişkenin kategorilerine göre, veriler alt gruplara ayrılır [5].

CHAID analizinde her bir açıklayıcı değişken için en iyi bölünme bulunur. Daha sonra açıklayıcı değişkenler en iyi seçilene kadar karşılaştırılır ve seçilen en iyi açıklayıcı değişkene göre yeniden bölünmeler yapılır. Tüm alt bölümler bağımsız olarak yeniden analiz edilir. Her bir açıklayıcı değişken kategorilerini izin verdiği mümkün bölünmeler gerçekleştirilerek Ki-kare testindeki önem derecesine göre kontenjans tabloları oluşturulur. Buradan yola çıkarak CHAID analizi ki-kare istatistiklerini, Bonferronni yaklaşımını ve kategori birleştirme algoritmasını kullanarak araştırmacının ağaç diyagramı ile en önemli açıklayıcı değişkenleri ve bağımlı değişken ile olan etkileşimleri elde etmesini sağlar [5].

Y'nin hedef değişken olduğu varsayılarak, Sekil 2.2 Karar Ağacı Değişken ilişkisi dikkate alındığında [14].



Şekil 2.2. Karar ağaç modeli örneği [14]

CHAID algoritması hedef değişken ve tahmin edici değişkenler arasında Ki-Kare Bağımsızlık Testi uygulayarak istatistiksel açıdan en güçlü ilişkiye sahip değişkenden başlayarak; hedef değişkenden, tahmin edici değişkenleri dallandırır.

Şekil 2.2'deki karar ağacı örneğine bakıldığında, Y hedef değişkeni ile en güçlü ilişkiye *X*1 tahmin edici değişkeni sahiptir. Şekil 2.2'den de görüldüğü gibi CHAID, diğer karar ağacı algoritmalarının aksine ikili değil çoklu olarak dallanmaktadır. Böylece, veri içerisindeki bütün önemli detayları ve ayrışmaları tespit etmek mümkün olmaktadır. Dolayısıyla, çalışma özünde bütün farklı risk profillerini tanımlamaktadır.

- X1, X2 ve X3 olmak üzere sadece 3 değişken hedef Y değişkeniyle istatistik olarak önemli ilişkiye sahiptir.
- X1 değişkeni Y hedef değişkeniyle istatistik olarak en önemli ilişkiye sahiptir.
- X2 değişkeni X1 değişkeniyle X1=a1 olması koşuluyla istatistik olarak önemli ilişkiye sahiptir.
- X3 değişkeni X1 değişkeniyle X1=a2 olması koşuluyla istatistiksel açıdan önemli ilişkiye sahiptir [14].

2.2.6.1. CHAID Algoritması

Karar ağacında uygulanan CHAID algoritması aşağıda verilmektedir: [12,15]

1) Her bir tahmin edici değişken X için, X'in, Y hedef değişkenini dikkate alan en az öneme sahip kategori çiftini bul (bu, en büyük p değerine sahip olandır). Yöntem, Y'nin ölçüm düzeyine bağlı olarak p değerlerini hesaplayacaktır.

a) Eğer Y sürekli ise F testini kullan.

b) Eğer Y isimsel ise X'in kategorileri satırlarda ve Y'nin kategorileri sütunlarda olacak biçimde iki yönlü çapraz tablo düzenle. Pearson ki-kare testini veya olabilirlik oranı testini kullan.

c) Eğer Y sıralı ise bir Y birliktelik modeli uydur. Olabilirlik oranı testini kullan.

2) En büyük p değerine sahip X'in kategori çifti için, p değerini önceden belirlenmiş alfa düzeyi α birleş ile kıyasla.

a) Eğer p değeri α birleş 'den büyük ise bu çifti bir tek kategori altında birleştir. X' in yeni kategori kümesi için süreci Adım 1'den başlat.

b) Eğer p değeri α birleş 'den küçük ise Adım 3'e git.

3) X' in ve Y' nin kategori kümesi için uygun Bonferroni düzeltmesini kullanarak, düzeltilmiş p değerini hesapla.

4) En küçük düzeltilmiş p değerine sahip X tahmin edici değişkenini seç (en önemli olan). Bunun p değerini önceden tanımlanmış alfa düzeyi α böl ile kıyasla.

a) Eğer p değeri, α böl değerinden küçük veya eşit ise düğümü X'in kategori kümesini temel alarak böl.

b) Eğer p değeri, α böl değerinden büyük ise düğümü bölme. Bu düğüm uç düğümdür.

c) Ağaç büyütme sürecini durma kuralları görülene kadar sürdür [12,15].

Değişkenlerin bölünmeye uygun olup olmadığına, Bonferroni düzeltilmiş p değeri kullanılarak karar verilir. Bonferroni yaklaşımı, her bir grubun ortalama vektörlerinin genel ortalama vektöründen farkları bulunduktan sonra bu farkların sıfır olup olmadığını araştırmaya dayanır.

Genel ortalama vektörü $\bar{x}$ ve her grubun j. Değişkene göre ortalama vektörleri $\bar{x}_g$ aşağıdaki gibi gösterilir;

$$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} \quad \bar{x}_1 = \begin{bmatrix} \bar{x}_{11} \\ \bar{x}_{22} \\ \vdots \\ \bar{x}_{p1} \end{bmatrix} \quad \bar{x}_2 = \begin{bmatrix} \bar{x}_{12} \\ \bar{x}_{22} \\ \vdots \\ \bar{x}_{p2} \end{bmatrix} \ldots\ldots\ldots\ldots \bar{x}_g = \begin{bmatrix} \bar{x}_{1g} \\ \bar{x}_{2g} \\ \vdots \\ \bar{x}_{pg} \end{bmatrix} \tag{2.1}$$

Her bir grubun ortalama vektörünün, genel ortalama vektöründen farkları değişkenlere göre aşağıdaki gibi belirlenir.

$$d_1 = \bar{x}_1 - \bar{x} = \begin{bmatrix} \bar{x}_{11} \\ \bar{x}_{22} \\ \vdots \\ \bar{x}_{p1} \end{bmatrix} - \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} \ldots\ldots\ldots\ldots d_g = \bar{x}_g - \bar{x} = \begin{bmatrix} \bar{x}_{1g} \\ \bar{x}_{2g} \\ \vdots \\ \bar{x}_{pg} \end{bmatrix} - \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} \tag{2.2}$$

K grup ile l grup i. Değişken ortalamaları arasındaki ortalama farkları arasındaki 1-$\propto$ güven aralığı şöyle hesaplanır.

$$(d_{ki} - d_{li}) = (\bar{x}_{ki} - \bar{x}_{li}) \pm t\left(\frac{a}{pg(g-1)}, (N-g)\right)\sqrt{(\frac{1}{n_k} + \frac{1}{n_l})\frac{w_{ij}}{N-g}} \tag{2.3}$$

N=n_1+n_2+…+ng, p değişken sayısı, g grup sayısı ve wij, w matrisinin köşegen elemanıdır. W matrisi, gruplar içi değişimi gösterir ve;

$$W = \sum_{i=1}^{g} \sum_{j=k}^{n_j} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_j) \tag{2.4}$$

g =grup sayısı

nj=i. Gruptaki veri sayısı

İfadesi ile hesaplanır. Her bir değişken için, gruplar ikişerli olarak dikkate alınır ve eşitlik (2.3) kullanılarak i. değişken için elde edilen aralığın sıfır değerini içerip içermediği kontrol edilir. Eğer sıfır değeri belirlenen aralıkta yer alıyorsa, ilgili gruplar arasında anlamlı bir farklılık olmadığı, aksi durumda grupların farklı oldukları seklinde yorumlanır.

CHAID algoritması, tahmin edici değişkenin tüm değerlerini dikkate alarak analiz yapar. Hedef değişkeni dikkate alarak istatistik olarak benzer olan değişkenleri birleştirir. Ve farklı olan değişkenle işlemi sürdürür. Daha sonra karar ağacının ilk dalını oluşturmak için en iyi tahmin edici değişkeni seçer. Her bir düğüm seçilen değişkenin benzer değerlerinden oluşur. Bu süreç ağaç tamamıyla büyüyene kadar tekrarlanarak devam eder. Yapılacak testler hedef değişkenin türüne göre değişmektedir. Eğer değişken sürekli bir değişken ise F testi, kategorik (nominal/ordinal) bir değişken ise ki-kare kullanılır.

CHAID analizinin dışında sınıflandırma ya da ağaç şeklinde hiyerarşik diyagram modelleri kuran diğer teknikler ise: "Sınıflandırma ve Regresyon Ağaçları (Classification and Regression Trees; C&RT)" ve "Hızlı-Yansız-Etkili-İstatistik Ağacı (Quick-Unbiased-Efficient-Statistical Tree; QUEST)" şeklindedir. CHAID analizinde bağımlı ve bağımsız değişkenlerin tipi model kurma aşamasında bir sınırlılık getirmediği halde, C&RT ve QUEST sınıflandırma teknikleri ise, CHAID analizinden farklı olarak ele alınan değişken türlerine (kategorik, sınıflama, sıralama) bir sınırlılık getirmektedir. Bu yönüyle CHAID analizi, önemli avantajlara sahiptir [18].

2.3. Lojistik Regresyon Analizi

Lojistik regresyon, yanıt değişkeninin kategorik olarak; ikili veya çoklu kategoriler halinde gözlendiği durumlarda açıklayıcı değişkenlerle neden-sonuç ilişkisini belirlemede yararlanılan bir yöntemdir. Açıklayıcı değişkenlere göre cevap değişkenin beklenen değerinin, olasılık, olarak elde edildiği bir regresyon yöntemidir [37]. Aynı zamanda

açıklayıcı değişkenlerin etkilerine dayanarak verilerin sınıflandırılmasında da kullanılabilmektedir [36].

Lojistik regresyon analizi diskriminant analizi ile birlikte sınıflandırma amacıyla kullanılan istatistik yöntemlerden biridir. Ancak diskriminant analizi tüm öngörü değişkenlerinin normal dağıldığı varsayımı, nominal öngörü değişkenlerinin kullanılamaması ve sınıfları ayıran sınırların doğrusal olması gibi sebeplerle veri madenciliğinde pek yaygın olarak kullanılmamaktadır [44].

Lojistik regresyon bağımlı değişkenin binom dağılımına sahip olması durumunda kullanılan bir analiz yöntemidir. Lojistik model ‘’Genelleştirilmiş Doğrusal Model’’ olarak bilinen çok değişkenli model ailesinin bir üyesidir. Modellerde sonuç değişkeni açıklayıcı değişkenlere doğrusal bir yapı ile bağlıdır [38].

Bu yöntemde, açıklayıcı değişkenlerin bağımlı değişkenler üzerindeki etkileri olasılık olarak hesaplanarak risk faktörlerinin olasılık olarak belirlenmesi sağlanır [40,41].

Bu tekniğin yaygın olarak kullanılmasının nedenleri arasında yorumlanmasının kolay olması ve bağımsız değişkenler üzerinde herhangi bir ön şart gerektirmemesi gösterilebilir. Lojistik regresyon modelinin bu esnekliği sağlıktan sosyal bilimlere kadar her alanda yaygın olarak kullanılmasını sağlamıştır [36].

Doğrusal regresyon analizinde bağımlı değişkenin değeri tahmin edilirken, LRA da bağımlı değişkenin alacağı değerlerden birinin gerçekleşme olasılığı tahmin edilir[39].

Lojistik Regresyonun amacı da, bir veya birden çok bağımsız değişken ile sonuç değişkeni arasında bir model kurmaktır. Farkı ise diğer regresyon yöntemlerinde sonuç değişkeni sürekli değerler alırken, lojistik regresyonda ise sonuç değişkeninin kesikli iki veya daha çok değer aldığı durumlarda kullanılıyor olmasıdır [40].

Lojistik regresyonda, doğrusal regresyon analizinde olduğu gibi bazı açıklayıcı değişken değerlerine dayanarak tahmin yapılmaya çalışılır. Ancak bu iki yöntem arasında üç önemli fark vardır [43].

1. Doğrusal regresyon analizinde tahmin edilecek olan yanıt değişken sürekli iken, lojistik regresyon analizinde yanıt değişken kesikli bir değer almaktadır.

2 Doğrusal regresyon analizinde yanıt değişkenin değeri tahmin edilirken lojistik regresyon analizinde ise yanıt değişkenin alabileceği değerlerin gerçekleşme olasılığı tahmin edilir.

3 Doğrusal regresyon analizinde sonuç çıkarım için yanıt değişkenin normal dağılım göstermesi şartı aranırken, lojistik regresyon analizinde böyle bir şart yoktur [43].

Lojistik regresyon denkleminde P incelenen olayın gözlenme olasılığını göstermektedir. İncelenen bir olayın olasılığının kendi dışında kalan diğer olayların olasılığına oranına *ODDS* değeri denir. İncelenen iki farklı olayın ODDS değerlerinin birbirine oranına ise *ODDS* Oranı denir. Lojistik regresyon denkleminde ODDS Oranı, $Exp(\beta)$ olarak ifade edilir. Olasılık oranı (Odds), bir olayın meydana gelme olasılığının meydana gelmeme olasılığına oranı olduğuna göre; $Exp(\beta_p)$ Y değişkeninin x_p değişkeninin etkisi ile kaç kat daha fazla ya da % kaç oranında fazla gözlenme olasılığına sahip olduğunu belirtir [45].

2.3.1. Lojistik Regresyon Analizinde Değişken Seçimi

Lojistik Regresyon Analizi; sürekli, kesikli, ikili ya da bunların bir karışımı olan veri setlerinden kategorik bir sonucu tahmin etmeye olanak sağlar. Lojistik regresyon modellerinde kategorik bağımsız değişken/değişkenler, sadece sürekli bağımsız değişken/değişkenler veya hem kategorik hem de sürekli bağımsız değişkenler kullanılabilir [49].

2.3.2. İkili (Binary)Lojistik Regresyon Modeli

Çeşitli gösterim biçimleri olan genel doğrusal regresyon modeli,

$$E(y / x_{i1}, \ldots\ldots, x_{ip}) = \sum_{k=0}^{p} \beta_k x_{ik}; i = 1, .., n \; için \tag{2.5}$$

biçiminde koşullu beklenen değer olarak da yazılması mümkündür. Bu modelde açıklayıcı değişkenler üzerinde kısıt yok iken, y bağımlı değişkeninin sürekli olması koşulu vardır. Herhangi bir i'inci gözlem için,

$$y_i = \sum_{k=0}^{p} \beta_k x_{ik} + u_i \tag{2.6}$$

biçiminde ifade edilen modelde açıklayıcı değişkenler üzerinde bir kısıt olmadığından y_i sonuç değeri -∞ ile +∞ arasında tüm değerleri alabilmektedir. Bağımlı değişkenin 0,1 gibi değerler aldığı durumda bu kural bozulmakta ve $P(y_i = 1)$, i'inci gözlemin 1 değerini alma olasılığı olmak üzere, beklenen değer,

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

olarak bulunur. Sol tarafı 0-1 arasında değerleri alan bu denkleme doğrusal olasılık modeli adı verilmektedir Açıklayıcı değişkenlerin sınırsız değerler alması nedeniyle söz konusu eşitlik her zaman sağlanamamaktadır. Bu sebeple çeşitli dönüşümler yapılmaktadır. Bu dönüşümlerden en yaygın olarak kullanılan iki tanesi logit ve probit dönüşümlerdir.

Logit dönüşümde doğrusal olasılık modelinde olasılık değerleri üzerinde $P/(1-P)$ dönüşümü yapılarak sonuç değişkeninin sınırları 0, +∞ yapılmakta, daha sonra ise bu oran değerinin doğal logaritması alınarak sonuç değişkenin sınırları -∞, +∞ yapılmaktadır. Bu dönüşümlerden sonra elde edilen yeni fonksiyon,

$$E(y_i) = L_i = \log\left(P_i / (1 - P_i)\right) = \sum_{k=0}^{p} \beta_k x_{ik} \tag{2.8}$$

olarak yazılmaktadır. Lojistik model ya da kısaca logit olarak bilinen bu modelde P_i olasılık değeri,

$$P_i = exp\left(\sum_{k=0}^{p} \beta_k x_{ik}\right) / 1 + \left(exp\left(\sum_{k=0}^{p} \beta_k x_{ik}\right)\right) \tag{2.9}$$

biçiminde tanımlanmakta ve lojistik fonksiyon adını almaktadır. Bu modelde sonuç değişkeninin iki değer alması nedeni ile hata terimi sıfır ortalama ve $P/(1-P)$ varyanslıdır. Hata terimi bu parametrelerle binom dağılımlı olup, analiz bu teorik temele dayanmaktadır.

Logit fonksiyonu aynı zamanda şu şekilde de gösterilmektedir:

$$P_i = \frac{1}{1 + e^{-(\alpha+\beta_x)}} \tag{2.10}$$

Bu eşitliğe lojistik dağılım fonksiyonu adı verilir. α+βx=Z olarak kabul edilirse bu durumda,

$$\frac{P_i}{1-P_i} = \frac{1+e^{z_i}}{1+e^{-z_i}} = e^{z_i} \tag{2.11}$$

eşitliğine ulaşılır. Bu eşitlik odds (bahis) oranı olarak adlandırılır. Odds oranı daha özet bir ifadeyle olayın gerçekleşme olasılığının olayın gerçekleşmeme olasılığına olan oranını ifade etmektedir. Odds oranından genellikle ikili değişken arasındaki ilişkinin ölçümlenmesinde yararlanılır. Etki katsayısı veya etki büyüklüğü olarak tanımlanan Exp(β) aynı zamanda Odds oranını vermektedir ve bu değer açıklayıcı değişkenlerin etkisinin kolayca yorumlanabilmesi açısından önemlidir. Odds oranının doğal logaritması alınırsa Logit'e ulaşılır. Yani odds oranının logaritması katsayı tahminleri bakımından yalnız X'e göre değil ana kütle katsayılarına göre de doğrusaldır. Ayrıca odds oranları, x'in arttığı her birim için e^{β} 'nın katları kadar artar. Böylece odds oranının logaritması alınmak suretiyle doğrusal olmayan ilişki logit fonksiyonu yardımıyla doğrusal hale getirilmiştir [49].

2.3.3. Logit Modelin Özellikleri

Logit modeller normal dağılım, kovaryans matrislerinin eşitliği gibi kısıtlayıcı varsayımlara sahip olmadığından diğer yöntemlere göre avantaja sahiptir. Ayrıca bağımlı değişkenin kesikli olması yöntemin uygulanabilirliği üzerinde bir etki yaratmamaktadır. Son olarak model parametreleri logaritmik odds oranları kullanılarak kolayca izah edilebilir ve yorumlanabilir [49].

2.3.4. Modelin Parametre Tahmini

Modelin katsayılarının tahmininde En Çok Olabilirlik Yöntemi, Yeniden Ağırlıklandırılmış İteratif En Küçük Kareler Yöntemi, Minimum Logit Ki-Kare Yöntemi kullanılmaktadır. Açıklayıcı değişkenlerin hepsi sürekli ise minimum logit kikare yöntemi, değişkenlerin hepsi kesikli ise en çok olabilirlik yöntemi, hem sürekli hem de kesikli ise ağırlıklandırılmış iteratif en küçük kareler yöntemi kullanılmaktadır [49].

2.3.4.1. En çok olabilirlik yöntemi

Lojistik regresyon çözümlemesinde bağımlı değişken ile bağımsız değişkenler arasındaki ilişki doğrusal olmadığı için model parametreleri en küçük kareler yöntemi ile tahmin edilemez. Başarı olasılığı $P_i = P(y_i=1/x)$, başarısızlık olasılığı $1 - P_i$ olduğunda i'inci gözlem için olasılık,

$$P(y_i \ / \ x_i) = P_i^{y_i}(1 - P_i)^{1-y_i}; i = 1,..,n \tag{2.12}$$

için biçiminde yazılacak olursa, bu olasılık n gözlem için olabilirlik fonksiyonu olarak,

$$L(y \ / \ x) = P(y \ / \ x) = \sum_{i=1}^{n} P_i^{y_i}(1 - P_i)^{1-y_i} \tag{2.13}$$

biçiminde ifade edilebilir. Bilindiği gibi en çok olabilirlik yöntemi p açıklayıcı değişkene ilişkin β'ların kestirimini, sonuç değişkeni y'nin gözlenme olabilirliğini maksimum kılacak biçimde bulmayı amaçlamaktadır. Yani $L(y/x,\beta)$ olabilirlik fonksiyonunu maksimum yapacak $\hat{\beta}$ katsayılar vektörünü belirlemek ana hedeftir. Bu durumda yukarıdaki eşitliklerden yararlanılarak lojistik modelin olabilirlik fonksiyonunun logaritması,

$$logL(y/x,\beta) = \sum_{i=1}^{n} (y_i logP_i + (1 - y_i)\log\,(1 - P_i)) \tag{2.14}$$

biçiminde olup, bunun β' ya göre birinci türevi,

$$\sum_{i=1}^{n} (y_i - P_i)x_{ij} = 0; j = 1, \dots, p \tag{2.15}$$

için olabilirlik denklemini vermektedir. Bu denklemin çözümünde ise $\hat{\beta}$ kestirim değerleri bulunmaktadır. Logit modelde gösterilen P_i' nin β' larda doğrusal olmaması nedeniyle en çok olabilirlik yönteminden iteratif yolla çözüme gidilir. İteratif çözümlemede β' lara herhangi bir başlangıç değerleri verilerek elde edilen kestirimlerden, her adımda δ kadar eksiltme ya da artırma yapılıp türevler alınarak sonuca ulaşılır. Sonuca ulaşmanın

göstergesi yakınsamanın sağlanmasıdır. Yakınsama ise iterasyonlar arasında fark olmaması durumunda sağlanmaktadır. [49].

2.3.4.2. Yeniden ağırlıklandırılmış iteratif en küçük kareler yöntemi

Gruplandırılmış verilerde *J* grubunun her birinde n_j denemeden r_j başarı elde edildiğinde başarı oranı $P_j = r_j/n_j$ olarak tanımlanabilir. $var(r_j/n_j) = P_j \quad (1 - P_j)/n_j$ olduğundan, her binom dağılımlı gözlem için varyans değişmektedir. Bu durumda logit (r_j/n_j) ' nin açıklayıcı değişkenler üzerinde $w_j = n_j/P_j(1 - P_j)$ ağırlığı ile ağırlıklandırılmış regresyon uygulanmalıdır. Ancak w_j ağırlık değerleri de P_j' nin bir fonksiyonu olduğu için en küçük kareler yöntemi iteratif olarak uygulanacak ve ağırlık değerleri her adımda (kestirim değerlerine bağlı olarak) yeniden elde edilecektir [49].

2.3.4.3.Minimum logit ki-kare yöntemi

Ağırlıklı en küçük kareler kestirim yönteminin özel bir biçimi olan ve Berkson tarafından geliştirilen bu yöntemde $2 \times J$ çapraz tablolarındaki beklenen ve gözlenen logit değerleri arasındaki farktan yararlanılmaktadır. Yöntem tekrarlı veriler olması durumunda kullanılmaktadır. Bir önceki yöntemde verilen P_j olasılığı üzerinden yapılan logit dönüşümü, bu yöntemde sonuç değişkenini oluşturmaktadır. Kestirimde kullanılan ağırlık değerleri $n_j/P_j(1 - P_j)$ olarak elde edilmektedir. Bu bilgiler ışığında yöntem logit değeri olarak tanımlanan sonuç değişkeninin, açıklayıcı değişkenler ile (tanımlanan ağırlık değerleri ile ağırlıklandırılmış) regresyonundan en küçük kareler kestirimlerini elde etmeye dayanmaktadır. Buradan tek adımda bulunan ağırlıklı en küçük kareler kestirimleri minimum logit ki-kare kestirimleri adını almaktadır [49].

2.3.5. Modelin Katsayılarının Testi ve Yorumlanması

Modelin verilere uyumunun belirlenmesindeki önemli adımlardan biri, uyumun iyiliği diğer bir deyişle, modelin gözlenen verileri ne kadar iyi tanımlanabildiğinin incelenmesidir. Bağımsız değişkenlerin modele eklenmesi veya çıkarılması ile ilgili olarak

yapılan analitik çalışma burada ele alınacaktır. Bu analiz ile modelde kullanılacak katsayıların önem kontrolü yapılmış olacaktı [49].

2.3.5.1. Olabilirlik oran testi

Doğrusal regresyonda regresyon kareler toplamı ne kadar büyük olursa bağımsız olmakla birlikte bu yöntemde gözlemlenen değerlerin tahmin edilen değerlerle karşılaştırılması log olabilirlik ile yapılır.

Burada $H_o: \beta_1 = 0$ hipotezi test edilmektedir. Geçerli model sadece önemli olan değişkenleri içeren model, doymuş model ise değişken sayısı kadar parametre içeren model olmak üzere;

$D = -2ln$(*Geçerli Modelin Benzerliği/Doymuş Modelin Benzerliği)*

olarak hesaplanır. Parantezin içerisindeki ifade benzerlik ya da olabilirlik oranı likelihood ratio) olarak ifade edilir ve aşağıdaki test istatistiği elde edilir.

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\pi_i}{1-y_i}\right)\right] \tag{2.16}$$

Bir değişkenin modeldeki etkisini ölçmek için değişken modelde yer alırken ve modelden çıkartıldığında elde edilen D değerleri arasındaki farka bakılır.

$$\begin{aligned} G = D(&Dğişken\ içermeyen\ model) - D(Değişken\ içeren\ model) \\ &= -2\ln\,(Değişken\ içermeyen\ modeli \\ &/Değişken\ içeren\ modelin\ olabilirliği) \end{aligned} \tag{2.17}$$

şeklinde bulunur. Burada bulunan G değeri Ki-kare dağılımına uymaktadır [49].

2.3.5.2. Wald testi

Wald istatistiği β parametresi ile standart hatasının oranıdır ve Z dağılımı göstermektedir. Doğrusal regresyondaki t testinin alternatifidir. Wald X^2 istatistiği t değerlerinin karesine eşittir.

$$W = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \tag{2.18}$$

Wald istatistiği standart normal bir değişkendir. Karesi 1 serbestlik derecesi ile X^2 dağılır. Bu durumda;

$$W = \left[\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}\right]^2 \tag{2.19}$$

istatistiği tanımlanabilir. Büyük β değerleri için tahmin edilen standart hatalar da büyük çıkmaktadır. Bu durum H_0 hipotezi yanlış iken kabul edilmesi olasılığını artırmaktadır. Modelin katsayılarının yorumlanması için lojistik regresyon modelinde aşağıdaki eşitlikten yararlanılır.

$$\ln\left(\frac{P_i}{1-P_i}\right) = d = X'\beta \tag{2.20}$$

$\left(\frac{P_i}{1-P_i}\right)$' nin logaritması modelin katsayılarının doğrusal bir şekilde yorumlanabilmesini sağlamaktadır. Bu şekilde açıklayıcı değişkendeki 1 birim değişmenin olasılık üzerindeki etkisi görülmektedir [49].

2.3.5.3. Pearson ki-kare testi

Karl Pearson tarafından 1900 yılında bulunan ve değişik kullanım amaçları olmasına karşılık, var olan veya olması gereken frekanslar arasındaki farklılıkların anlamlılığının test edilmesi temeline dayanan bir başka testte Pearson ki-kare testidir.

$$r_i = \frac{y_i - \pi(x_i)}{\sqrt{\pi(x_i)(1-\pi(x_i)}} \tag{2.21}$$

formülü ile hesaplanan bu istatistiğinin değerinin büyük olması yani anlamlı çıkmaması modelin verilere uyumunun başarısız olduğunu göstermektedir. Ki-kare dağılımına uyduğu ifade edilse de bazı şartlar sağlanmadıkça tam bir uyum ölçütü olarak bu istatistiğin kullanılamayacağı düşünülmektedir.

Conover (1999: 241) istatistiğin ki-kare dağılımına uyması için Koehler ve Larntz'ın

- toplam gözlem sayısının $n \geq 10$
- sınıf sayısının $c \geq 3$
- beklenen değerlerin hepsinin $E \geq 0{,}25$

şeklinde önerdiği koşulların sağlanması veya bir başka yol olarak çok sayıda küçük beklenen değerlerin bir araya getirilmesi gerektiğini belirtmiştir [49].

2.3.6. Modelin Uyum İliğinin Ölçülmesi

Katsayılarının bulunması ve önem kontrolünden sonra modelin aşağıda verilen durumlara karşı uyum iyiliğinin test edilmesi gerekmektedir.

- Logaritmik dönüşüm yerine başka bir dönüşüm daha iyi olabilir,
- Logaritmik dönüşüm uygun olsa bile modeldeki açıklayıcı değişkenlerin bir kısmı uygun olmayabilir ya da bazı etkileşim terimlerinin de modele katılması gerekebilir,
- Değişkenlerin modelde bulunması uygundur, ancak ölçek yanlış olabilir,
- Veriler arasında aykırı değer olabilir

Bu gibi durumlara karşı lojistik modelin uyum iyiliğini araştırmada kullanılan ölçütlerden önemli olanları şunlardır:

- Tüm değişkenleri içeren model ile kestirilen modele ilişkin olabilirlik oran değerlerinin farkına dayanan (hata kareler toplamına benzer) ölçütlerin ki-kare dağılacağı düşüncesinden hareketle, kurulan modelin geçerliliği sınanmaktadır. Bu yolla modele girecek açıklayıcı değişkenlere ve eklenecek karesel terimlere karar verilmektedir.
- Hata terimlerinin, x değerlerine ya da olasılık değerlerine karşı çizimi ile aykırı değer araştırması yapılmaktadır.
- Hata kareler toplamı ve olabilirlik oranına dayalı R^2 türü ölçütler de modelin uyumunu test etmede kullanılmaktadır.
- Lojistik model ayrımsama amacıyla kullanıldığında modelin doğru sınıflandırma oranı da bir uyum iyiliği ölçütüdür.

Lojistik modellerin uyum iyiliğinin belirlenmesinde aşağıdaki hipotez testi kullanılmaktadır.

H_0: Model uygundur.

H_1: Model uygun değildir.

Bu hipotezde H_0 hipotezinin kabul edilmesi modelin anlamlı olacağını göstermektedir. Lojistik modellerde normallik varsayımının bulunmaması sebebi ile parametrik testler kullanılmamakta Ki-kare ve G2 gibi parametrik olmayan ölçütlerden yararlanılmaktadır.

Ki-kare ve G2 bilinen en basit parametrik olmayan ölçütlerdir. Çünkü O gözlenen, E-beklenen değerleri, OlogO ve OlogE sırasıyla gözlenen ve beklenen olabilirlikleri göstermek üzere bu ölçütler,

$$X^2 = \sum \frac{(O-E)^2}{E}$$
$$G^2 = -2\sum Olog(E/O) = 2\sum Olog(O/E)$$
$$G^2 = -2\left(\sum Olog(E) - \sum Olog(O)\right) = -2(\log(E) - \log(O)) \qquad (2.22)$$

biçiminde tanımlanmaktadır. Tekrarlı veriler için ki-kare ölçütü; $\hat{P}_i$ olasılık kestirimi değeri olmak üzere benzer biçimde tanımlanmaktadır.

$$X^2 = \sum_{i=1}^{j} \frac{(y_i - n_i\hat{P}_i)^2}{n_j P_j(1-\hat{P}_i)} \qquad (2.23)$$

Lojistik regresyon analizinde, kurulan modelin önemliliğini test etmede ve bir anlamda modele girmesi gereken açıklayıcı değişkenleri belirlemede yine G^2 yaklaşımı kullanılmaktadır. Bu amaçla önerilen sapma ölçütü, doymuş model; değişken sayısı kadar parametre içeren model, kestirilmiş model; sadece önemli olduğu düşünülen değişkenleri içeren model olmak üzere,

$$D = -2log\frac{(Geçerli\ Modelin\ Olabilirliği)}{(Doymuş\ Modelin\ Olabilirliği)} \qquad (2.24)$$

biçiminde tanımlanmaktadır.

Olabilirlik fonksiyonunun yazılması ile,

$$D = \sum_{i=1}^{n} {d_i}^2 = -2 \sum_{i=2}^{n} y_i \log\left(\hat{P}_i / y_i\right) + (1 - y_i) \log\left((1 - \hat{P}_i)/(1 - y_i)\right) \tag{2.25}$$

biçimine dönüşen sapma ölçütü, *p* modeldeki parametre sayısını göstermek üzere *(n-p)* serbestlik dereceli Ki-kare tablo değeri ile karşılaştırılmaktadır. Ayrıca sapma değeri minimum olan model en iyi model olarak bilinmektedir.

Tekrarlı veriler için sapma ölçütü ise,

$$d_j = \mp\sqrt{2}\left(y_j \log\left((y_i / n_j \hat{P}_j\right) + \left(n_j - y_j\right) \log\left((n_j - y_j)/n_j(1 - \hat{P}_j)\right)\right)^{1/2} \tag{2.26}$$

iken

$$D = \sum_{j=1}^{J} {d_j}^2 \tag{2.27}$$

eşitliğinden bulunarak *(J-p-1)* serbestlik dereceli Ki-kare tablo değeri ile karşılaştırılmaktadır.

Çoklu doğrusal regresyonda katsayıların anlamlılığına ilişkin tümel F testine karşılık gelebilecek benzer bir test, lojistik regresyon analizi için de geliştirilmiştir. L_0 sadece sabit terimden oluşan modelin olabilirlik değeri, L_1 elde edilen modelin olabilirlik değeri olmak üzere,

$$C = -2 \log\left(L_0 / L_1\right) = -2(\log L_0 - \log L_1) \tag{2.28}$$

olarak tanımlanan ölçüt *(p-1)* serbestlik dereceli Ki-kare dağılımı göstermektedir.

Uyum iyiliği testi yaparken ayrıca Hosmer Lemeshow testi de kullanılmaktadır. Hosmer ve Lemeshow gözlenen verilerin beklenen verilere uyum iyiliğini test etmek maksadıyla 7 adet istatistik geliştirmiştir. Ancak bunlardan sadece ikisi kullanıldığından burada C ve H ile ifade edilen bu iki istatistiğe yer verilecektir. C istatistiği verileri tahmin edilen olasılıklara göre ayırırken, H istatistiği verileri belirlenen bir kesim noktasına (cut-off point) göre ayırır. Gruplara ayrılan verilere, daha sonra ki-kare uyum iyiliği testi uygulanmaktadır. Dolayısıyla her grup için beklenen ile gözlenen değerlerin bir karşılaştırması yapılmaktadır. Hesaplanan p değeri 0,05'den büyük ise gözlenen ve

beklenen değerler arasında fark olmadığını gösteren sıfır hipotezi reddedilir. Bu durum modelin mevcut veriye olan uyumunun iyi olduğunu gösterir.

Bu testte aşağıdaki istatistik kullanılmaktadır.

$$G_{HL}{}^2 = \sum_{j=1}^{10} \frac{(O_j - E_j)^2}{E_j(1 - E_j/n_j)} \tag{2.29}$$

Belirtilen istatistikle eşit sayıda tahmin edilen olasılıklara göre 10 grup oluşturulmakta ve bu dağılımın Ki-kare dağılımına uyduğu düşünülmektedir.

Modelin uyum iyiliğinin belirlenmesinde R^2 istatistiklerinden de yararlanılmaktadır. Doğrusal regresyon analizindeki R^2'ye benzer şekilde yorumlanan ve kullanılan çeşitli R^2 istatistikleri mevcuttur. Modelin uyum iyiliğini belirlemede kullanılan istatistiklerden biri McFadden'in R^2 uyum iyiliği kriteridir. $-2lnL_0$ logit modelde yer alan sabiti ve $-2lnL_1$ modelin tüm açıklayıcı değişkenlerini içermek kaydıyla McFadden R^2 formülü şu şekilde hesaplanır:

$$McFaddenR^2 = 1 - \frac{-2lnL_1}{-2lnL_0} = 1 - \frac{lnL_1}{lnL_0} \tag{2.30}$$

R^2 istatistiği doğrusal regresyondaki belirtme katsayısına benzer, ancak onun gibi değerlendirilmez. Değişkenler arasındaki ilişkinin gücü açısından belirtme katsayısının mümkün olduğunca yüksek çıkması istenir. Aksi takdirde modelin bağımlı değişkeni açıklamada yetersiz kaldığı düşünülebilir. Oysa McFadden R^2 'nin değeri genellikle düşük çıkar ve bu modelin iyi olmadığı anlamına gelmez. Yapılan uygulamalarda R^2 'nin 0,2 ile 0,4 arasında değerler aldığı gözlemlenmiştir.

Bir başka R^2 istatistiği ise Cox ve Snell R^2 istatistiğidir. Cox ve Snell R^2;

$$Cox - SnellR^2 = 1 - \left[\frac{L_0}{L}\right]^{\frac{2}{I}} \tag{2.31}$$

formülü ile gösterilir ve I gözlem sayısını göstermektedir. Formülden de anlaşılacağı üzere R^2 'nin en yüksek değerinin bile birden küçük çıkması diğer bir deyişle maksimum değerinin hiçbir zaman 1'e ulaşamaması sonucun yorumlanmasında sorun yaratmaktadır.

Diğer bir test istatistiği ise Nagelkerke R^2 istatistiğidir. Nagelkerke R^2;

$$Nagelkerke R^2 = \frac{Cox - Snell}{1 - \left[\frac{L_0}{L}\right]^2} \tag{2.32}$$

Bu ölçümün maksimum değeri 1 olabilir ve doğrusal regresyondaki belirtme katsayısı gibi yorumlanabilir. Nagelkerke R^2, Cox-Snell R^2'den her zaman büyük ancak doğrusal regresyondaki R^2 değerinden genellikle küçük olmaktadır [49].

3. BULGULAR

3.1. Uygulamada Kullanılacak Değişkenler Hakkında Genel Bilgiler

Nöropati, diyabet hastalarında en sık görülen sorunlardandır, ve yaşamları süresince %25-50 karşılaşılabilir. Diyabet olgularında periferik (eller ayaklar) sinir fonksiyon bozuklukları belirtileri diyabetik nöropati adını alır.

Nöropatinin belirtileri; ayaklarda ve ayak parmaklarında hissizlik-uyuşukluk yanma karıncalanma diabetik nöropati bulgusu olabilir. Bu hissizlik nedeniyle diabetik nöropati hastaları küçük kesikleri ve travmaları hissetmeyebilirler. Özellikle şu bulguların bulunması diabetik nöropatiyi düşündürür: ayaklarda yanıcı tarzda, özellikle geceleri artan ağrı. Termografiyle ayaklarda dolaşım bozukluğu gösterilebilir [52].

Ölçülen Hb1ac değerleri %9 üzerinde olduğu zaman sinirlerde ileti hızında yavaşlama gözlenir. Bundan sonra sinir hücrelerinde yoğunluk kaybı gelişir, nöropatik ağrı oluşur. Diabetik periferik nöropatik ağrı, açlık kan şekeri yüksek-glukoz tolerans testi bozulmuş hastalarda gelişebilir. Kalpte gelişen otonom nöropati, diyabet'li olgularda görülen kalp enfartüsünün en önemli sebebidir. Tip 2 diyabet olgularında nöropati en yüksek oranda hastalık süresi 25 yılı aşanlarda görülmektedir. Nöropati gelişen olgularda ölüm riski artmıştır, en sık sebep koroner kalp hastalığıdır [49].

Diyabetik nöropati gelişiminde etkin faktörler;

Kötü glisemik kontrol, Uzun diabet süresi (hastalığın yaşı), ileri yaş, erkek cinsiyet, Boy, alkol, hipertansiyon, sigara, hiperlipidemi, APOE genotipi Aldoz redüktaz gen hiperaktivitesi, anjotensin-converting enzim genotipi [50].

Genel olarak diyabet süresi uzadıkça nöropati olasılığı da artar. Genel kabule göre hastalığı 25 yılı aşmış diyabetlilerden yaklaşık yarısında nöropati gelişmesi beklenir. Şeker ayarı bozuk giden, zaman zaman aşırı yükselme ve düşmeleri olduğu kişilerde bu tehdit daha fazladır. Ayrıca koşulların olumlu gittiği bir diyabetlide araya giren ve ağır geçen bir ateşli hastalık, açlık gibi durumlar birden hem şeker ayarını bozar hem de nöropati ortaya çıkışını kolaylaştırabilir. Uzun süre ve makul denebilecek ölçülerin dışına çıkan alkol kullanımı da nöropatilerin hem oluşumunu kolaylaştırır, hem de olan nöropatiyi ağırlaştırır. Sigara ise daha çok damarlara etki eder. Bacak damarlarında daralmaların ortaya çıkması ise diğer dokuların yanı sıra sinirlerin de beslenmesini, ve oksijen almasını bozarak yine nöropatiyi kolaylaştırır [51].

Nöropati iki ana grupta incelenir, sensörimotor ve otonomik olarak en sık görülen periferik nöropatidir. Ayaklar, bacaklar, eller ve kollarda ağrı veya his kaybı ile ortaya çıkar. Nöropati tedavi yaklaşımı ise etkin bir tedavisi olmamakla birlikte diyabete bağlı nöropati tedavisinde ister ağrılı ister ağrısız olsun, kan şekeri yüksekliğinin düzeltilmesi gerekir [49]. Ayrıca kan yağlarının düşürülmesi, alkol ve sigara tüketiminin bırakılması da bunlardan bazılarıdır [50].

Nöropati gelişimini önlediği öne sürülen bazı ilaçlar vardır. Bu gün için bunların hiç biri ile oluşmuş ağır bir nöropatinin geri dönmesi söz konusu değildir. Buna karşılık nöropati belirtilerinin bazıları bir dereceye kadar tedavi edilebilir. Ağrıya etkili olan bir çok ilaç vardır. Bunlardan herhangi biri mutlaka etkili olacaktır, ağrıyı önlemede kullanılan ilaçların nöropatinin gelişim ve ilerlemesini önleyici bir yanı yoktur. Diabetik nöropatiyi önlemenin en kestirme yolu diyabet tanısı alındığı andan itibaren onun kurallarına göre yaşamaktır [51].

Analizde kullanılan değişkenler hakkında bilgiler;

Diyabet: Diyabet, başta karbonhidratlar olmak üzere protein ve yağ metabolizmasını ilgilendiren bir metabolizma hastalığıdır. Kendisini kan şekerinin sürekli yüksek olması ile gösterir.

Nöropati: (diyabet sinir hasarı) periferik ve otonom sinirlerin tahribi sonucu ortaya çıkan durum.

Yaş: hastanın yaşı, diabetik nöropati ile yakın ilişkisi gözlenmiştir.

Cinsiyet: erkeklerde kadınlara oranla daha çok gerçekleştiği gözlenmiştir.

Soy geçmişi: Ailede diyabet öyküsü

Hastalığın yaşı: Hastanın kaç yıldır diyabet hastası olduğunu gösterir. Uzun diyabet süresinin nöropati ile yakın ilişkisi gözlenmiştir.

İnsülin kullanma durumu: hastanın seker hastalığa bağlı olarak vücutta eksik olan insülini yerine koyarak kan şekeri kontrolünü sağlamasıdır. İnsülin kullanımının nöropatiyi geciktirdiği ileri sürülmüştür.

KVS (kardiyovasküler sistem): (kalp-damar hastalıkları) İçinde kan ve lenfin dolaştığı anatomik yapıların bütünü; bu dolaşımın amacı, tüm dokulara oksijen ve metabolizma etkinlikleri için gerekli maddelerin ulaştırılması ve hücre metabolizması artıklarının dokulardan uzaklaştırılmasıdır. Diyabetin kardiyovasküler hastalık riskini artırdığı ileri sürülmüştür.

Trigliserit: Trigliseritler vücudumuzda besin ve enerjinin depo şeklidir. Bu maddeler vücuda alınan ancak yakılamayan besinlerin fazlalarından, organların etrafında ve deri altında biriktirilerek oluşturulurlar. Kısaca trigliseritler bağırsaktan emilen sindirilmiş besin maddelerinin esterleşmesiyle (yağlaşmasıyla) oluşur. Sinir hasarı trigliseriti yüksek kişilerde görülmektedir.

AST Ve ALT: Karaciğere özgü olan ve karaciğer hasarını belirlemek için sıklıkla kullanılan enzimler aminotanferazlardır. Bunlar aspartat aminotransferaz (AST) ve alanin aminotransferaz (ALT) dir. AST ve ALT enzimlerinde görülen yükselmeler alkol kullanımı, şeker hastalığı, kronik hepatit C ve şişmanlığa bağlı karaciğer yağlanmasıdır.

Glukoz: Glukoz bir karbonhidrattır ve vücutta en çok ve en kolay kullanılan enerji hammaddesidir. Kan glikozunun düzenlenmesindeki bozukluk şeker hastalığını meydana getirir.

Kolesterol: Karaciğer tarafından üretilen ve hormonların üretilmesinde rol alan bir çeşit yağdır. Yüksekliği diyabet hastalağını etkilemektedir

Hiperlipidemi: Kanda dolaşan yağ (lipit) miktarı artışının bilimsel olarak tanımlanmasıdır. hiperlipidemi lipit tanımı içinde kolesterol, fosfolipitler ve trigliseritlerden bahsedilmektedir. Kanda çeşitli yağların yüksekliğini ifade etmek için kullanılan bir terimdir. Bu çeşitli yağ tipleri kolesterol, trigliserit, LDL-kolesteroldur. Tüm vücuttaki önemli atardamarlarda tıkanıklıklara sebep olur. Hiperlipidemi genetik faktörlere, şişmanlığa, aşırı tütün kullanımına bağlı olarak çıkabilir. Nöropatiyi etkileyen faktörler arasında yer almaktadır.

Hipertansiyon: Yüksek kan basıncına hipertansiyon denilmektedir. Hastalarda sıklıkla hipertansiyon ve diyabet bir arada olur.

Hb1ac: Kırmızı kan hücrelerinde glikozun bağlı olduğu hemeglobin yüzdesini gösteren bir ölçü birimidir. Diyabetin takibinde en önemli tahlillerden birisidir. Hb1ac 3 aylık kan şekeri bilançosudur. Ve 3 aylık kan şekerinin ortalamasını verir. Kan şekerinin iyi kontrol edilip edilmediğini gösterir. Hb1ac değeri %9 ve üzeri ise hasta şekerini kontrol edemiyor. Diyabete bağlı hastalıklar çıkabilir risk yüksektir anlamına gelir.

3.2. Uygulama

Lojistik regresyon ve CHAID analizi tekniklerinin sınıflama etkinliklerinin karşılaştırılması amacıyla, gereksinim duyulan veri seti Fırat Üniversitesi Tıp Fakültesi Endokrinoloji ve Nöroloji Anabilim Dalı Poliklinikliği veri tabanından sağlanmıştır. Yapılan çalışmanın veri setinden çalışmamıza uygun bağımsız değişken sayının 10 katından fazla birey üzerinde araştırma yapılmaya çalışılmış bunun üzerine 450 diyabet (şeker hastalığı) hastası alınmıştır, çalışmamızın konusu kapsamında nöropati değişkenini etkilediğini düşünülen 14 tane bağımsız değişken veri setine dahil olmak üzere seçilmiştir.

Bağımlı değişken, nöropati değişkeni iken nöropatiyi etkilediği düşünülen diğer etkenler ise bağımsız değişkenler olarak alınmış ve veri setini oluşturan tüm değişkenler aşağıdaki tabloda belirtilmiştir.

Tablo 3.1. Değişkenlerin listesi

	Değişkenler	Birim
Y	Nöropati olgusu	(Var, Yok)
X_1	Yaş	(yıl olarak)
X_2	Cinsiyet	(Erkek, Kadın)
X_3	Soy Geçmişi	(Var, Yok)
X_4	Hastalığın Yaşı	(yıl olarak)
X_5	İnsülin Kullanma Durumu	(Var, Yok)
X_6	KVS (kardiyovasküler sistem)	(Var, Yok)
X_7	Trigliserit	Mg/dL
X_8	AST (aspartat aminotransferaz)	U/L
X_9	ALT (alanin aminotransferaz)	U/L
X_{10}	Glukoz	Mg/dL
X_{11}	Kolesterol	Mg/dL
X_{12}	Hiperlipidemi	(Var, Yok)
X_{13}	Hipertansiyon	(Var, Yok)
X_{14}	Hb1ac	%

Analizler için SPSS 15.0 istatistiksel analiz paket programları kullanılmıştır

Uygulamaya dâhil edilen 450 diyabet hastasında nöropati olgusunun gerçekleşip gerçekleşmemesine göre "var" ve "yok" olarak sınıflandırılmıştır. Bu sınıflandırma için Lojistik regresyon analizi ile CHAID analizi teknikleri kullanılarak iki yöntemden elde edilen doğru sınıflandırma oranları karşılaştırılmıştır.

Çalışmaya alınan 450 deneğe ait tanımlayıcı bilgiler aşağıdaki tablolarda verilmiştir.

Tablo 3.2. Deneklerin Nöropati olgularına göre dağılımları

Değişken	**Sayı**	**Yüzde (%)**
Nöropati (var)	184	40.9
Nöropati (yok)	266	59.1
Toplam	450	100

Tablo 3.3. Deneklerin soy geçmişlerine göre dağılımları

Değişken	**Sayı**	**Yüzde (%)**
Soy geçmişi (var)	128	28.4
Soy geçmişi (yok)	322	71.6
Toplam	450	100

Tablo 3.4. Deneklerin KVS(Kardiyovasküler sistem) olgularına göre dağılımları

Değişken	**Sayı**	**Yüzde (%)**
KVS (var)	335	74.4
KVS (yok)	115	25.6
Toplam	450	100

Tablo 3.5. Deneklerin Cinsiyetlerine göre dağılımları

Değişken	**Sayı**	**Yüzde (%)**
Erkek	179	39.8
Kadın	271	60.2
Toplam	450	100

Tablo 3.6. Deneklerin Hipertansiyon olgularına göre dağılımları

Değişken	**Sayı**	**Yüzde (%)**
Hipertansiyon (var)	251	55.8
Hipertansiyon (yok)	199	44.2
Toplam	450	100

Tablo 3.7. Deneklerin insülin kullanmalarına göre dağılımları

Değişken	**Sayı**	**Yüzde (%)**
İnsülin (var)	220	48,9
İnsülin (yok)	230	51.1
Toplam	450	100

Tablo 3.8. Deneklerin Hiperlipidemi olgularına göre dağılımları

Değişken	**Sayı**	**Yüzde (%)**
Hiperlipidemi (var)	208	46.2
Hiperlipidemi (yok)	242	53.8
Toplam	450	100

Tablo 3.9. Deneklerin sürekli değişkenlere göre dağılımları

	Ortalama	(Max-Min) Değer
Yaş	53.65	98-17
Hastalığın yaşı	9.75	35-1
Trigliserit	197.70	974-33
AST	24.14	108-8
ALT	26.60	251-5
Glukoz	243.81	668-41
Kolesterol	196.04	432-58
Hb1ac	10.38	17.30-4.60

Lojistik regresyon ile sınıflandırma

Nöropati olarak belirlenen Y bağımlı değişkeni, nöropatisi olması durumunda 1, olmama durumunda ise 0 olarak kodlanmıştır. LRA ile test edilecek olan hipotezler ise aşağıdaki şekilde kurulmuştur.

SPSS de geriye doğru olabilirlik oranı seçim kriteri (backward LR) kullanılarak ele alınan 14 değişkenin içerisinden önemli olanları belirlenmeye çalışılıp sınıflandırma tablosu oluşturulmuştur. Önemli değişkenleri belirlemede genellikle geriye dönük eleme yöntemi yeğlenmektedir. Geriye dönük eleme yönteminde, ilk aşamada tüm değişkenler modele alınır. Her adımda modelden çıkartıldığında sapmada en küçük artışa neden olan değişken modelden çıkarılır. Bu işleme, modelden çıkartıldığında sapmada önemli değişime neden olan değişken bulunana kadar devam edilir [46].

Çoklu lojistik regresyon çözümlenmesinde amaç; en az değişkeni kullanarak açıklayıcılık oranı yüksek olan modeli elde edebilmektir. Bu nedenle, lojistik regresyonda tek değişkenli analizler yapılarak *p* değeri 0,25'e kadar bulunan bağımsız değişkenler olası risk faktörleri olarak çoklu lojistik regresyon modeline alınır. *p* değeri 0,25'ten büyük olanlar çözümlemeye alınmazlar [46].

Bu amaçla yapılan tek değişkenli lojistik regresyon analizi şekilde görüldüğü gibidir.

Tablo 3.10. Tek Değişkenli Lojistik Regresyon Analizi sonuçları

	β_j	$S(\beta_j)$	Wald	Sd	P	Odds oranı
Yaş	0,027	0,006	18,331	1	0,000	1,027
Cinsiyet	0,007	0,196	0,001	1	0,970	1,007
Soy geçmiş	0,244	0,215	1,284	1	0,257	1,276
Hastalığın yaşı	0,141	0,016	74,928	1	0,000	1,152
İnsülin	0,822	0,196	17,606	1	0,000	2,276
KVS	0,334	0,218	2,344	1	0,126	1,396
Hipertansiyon	0,896	0,201	19,948	1	0,000	2,450
Trigliserit	0,001	0,001	4,088	1	0,043	1,001
AST	-0,001	0,007	0,009	1	0,925	0,999
ALT	0,005	0,004	2,073	1	0,150	1,005
Glukoz	0,004	0,001	13,350	1	0,000	1,004
Kolesterol	0,001	0,002	0,086	1	0,770	1,001
Hiperlidemi	1,052	0,198	28,181	1	0,000	2,865
Hb1ac	0,127	0,037	12,120	1	0,000	1,136

Tablo 3.10' da yer alan Wald testi ile her bir bağımsız değişkene ilişkin katsayının önemli olup olmadığı incelenir. Wald testi için anlamlılık değeri 0,25 alınmıştır. Verilen sonuçlar incelendiğinde modele alınan 14 değişkenden sadece 4'ünün katkısının anlamsız olduğu görülmektedir. Bu değişkenler cinsiyet, soy geçmiş, AST ve kolesterol değişkenleridir.

Tablo 3.10'un verilen ihtimal düzeyi (p) 0.25'in altında (p<0.25) bulunan değişkenler çok değişkenli model için aday değişkenler olarak belirlenmiştir.

Aday değişkenlerle kurulan çok değişkenli modele ilişkin sonuçlar Tablo 3.11' de verilmiştir.

Tablo 3.11. Tek Değişkenli Modelde Aday Değişken Olarak Alınan Değişkenleri Kapsayan Geriye Dönük Eleme Yöntemi İle Elde Edilen Çok Değişkenli Model Sonuçları

		β_j	$S(\beta_j)$	Wald	Sd	P	Odds oranı	G.A
Adm 5	Yaş	0,016	0,008	4,283	1	0,038	1,016	1,001-1,031
	Hastalığın yaşı	0,151	0,018	68,526	1	0,000	1,163	1,122-1,206
	KVS	0,526	0,273	3,720	1	0,054	1,693	0,992-2,889
	ALT	0,014	0,005	8,611	1	0,003	1,014	1,005-1,023
	Hiperlipidemi	0,939	0,232	16,413	1	0,000	2,557	1,623-4,026
	Hb1ac	0,140	0,045	9,498	1	0,002	1,150	1,052-1,257
	Sabit	-4,712	0,752	39,267	1	0,000	0,009	

SPSS de geriye doğru olabilirlik oranı seçim kriteri (backward LR) kullanılarak Adım 5 de sonuca ulaşılmıştır. Son aşamada, modelde yaş, hastalığın yaşı, ALT, Hiperlipidemi, Hb1ac değişkenleri kalmıştır.

Lojistik regresyon analizi ile test edilecek hipotez aşağıdaki gibi kurulmuştur.

H_0: Yaşının, hastalığın yaşının, insülin kullanma durumunun, KVS durumunun, hipertansiyon değerinin, trigliserit değerinin, ALT değerinin, glukoz değerinin, kolesterol değerinin, hiperlipidemi durumunun, hb1ac değerinin nöropati değişkeni üzerinde bir etkisi yoktur.

H_1: En az bir değişkenin nöropati değişkeni üzerinde etkisi vardır.

Tablo 3.11' de yer alan modelde, değişkenlere ait katsayıların anlamlılığını test eden Wald istatistiğinin (H_0: β_i=0 ve H_1: $\beta_i \neq 0$) anlamlılık düzeyi olan *p* değerlerine bakılıp her bir değişkenin anlamlılık testinin yapılması sonucunda, KVS dışında kalan diğer bütün değişkenlerin bağımlı değişkenle istatistiksel olarak anlamlı bir ilişki içinde olduğu görülmektedir ($p < 0,05$, KVS için $p = 0,054 > 0,05$)

Analizde bağımlı değişken olarak alınan Nöropati 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 tablo 11' de gösterilmiştir.

Tablo 11' de elde edilen odds oranı değerlerine bakıldığında; Hiperlipidemisi olanların nöropati olma olasılığı olmayanlara göre 2,557 kat daha fazladır. Diğer bir ifade ile nöropati görülme olasılığı hiperlipidemisi olanlarda 2,557 kat artmaktadır. (%95 güven aralığı 1,623-4,026) Yaş değişkenindeki bir birim artış nöropati görülme olasılığını 1,016 kat artırmaktadır. (%95 güven aralığı 1,001-1,031) Hastalığın yaşı değişkenindeki bir birim artış nöropati görülme olasılığını 1,163 kat artırmaktadır. (%95 güven aralığı 1,122-1,206) Alt değişkenindeki bir birim artış nöropati görülme olasılığını 1,014 kat artırmaktadır. (%95 güven aralığı 1,005-1,023) Hb1ac değişkenindeki bir birim artış nöropati görülme olasılığını 1,150 kat artırmaktadır. (%95 güven aralığı 1,052-1,257)

LRA sonucunda lojistik regresyon modeli:

$$\hat{Y} = \frac{1}{1 + e^{-(-4,712+0,16X_1+0,151X_4+0,526X_6+0,14X_9+0,939X_{12}+0,14X_{14})}}$$

olarak elde edilmiştir.

Elde edilen modelin geçerliliği Hosmer Lemeshow testi ile sınanmıştır.

H_0: Tahmin denklemi anlamlıdır.

H_1: Tahmin denklemi anlamlı degildir.

Elde edilen tahmin modeli için Hosmer Lemeshow ki-kare değeri 7.227 olarak hesaplanmıştır. $p = 0.512 > \alpha = 0.05$ olarak elde edilmiş ve modelin uygun olduğuna dair H_0 hipotezi kabul edilmiştir. Buna göre lojistik regresyon sonucunda elde edilen bağımsız değişkenlerden en az birinin katsayısı sıfırdan farklı çıkmıştır. Bulunan tahmin denklemi anlamlıdır.

Tablo.3.12. Model Katsayılarının Omnibus Testleri

		Ki-Kare	Sd	p
Adım-5	Step	-2.515	1	0,113
	Blok	140,595	6	0,000
	Model	140,595	6	0,000

$H_{0:}\ B_1 = B_2 = \ldots = B_k = 0$

$H_{1:}\ B_1 \neq B_2 \neq \ldots \neq B_k \neq 0$

Bu hipotezler Ki-Kare fark testleri kullanılarak sınanmaktadır. Sabit terimli ve bağımsız değişkenleri içeren modellerin -2LogL istatistikleri arasındaki fark istatistiği, modellerin serbestlik dereceleri arasındaki farkla Ki-Kare dağılımına uymaktadır.

Tablo 3.12.'de Elde edilen sonuçlara göre beşinci adımda iki modelin -2LogL istatistikleri arasındaki fark, 6 serbestlik derecesiyle 140,595 dir. Tabloda anlamlılık düzeyini gösteren $p = 0.000 < \alpha = 0.05$ olarak elde edilmiş ve H_0 hipotezi red edilmiştir. Lojistik regresyon katsayılarının hepsi aynı anda sıfıra eşit değildir.

Tablo 3.13. Model Özeti

Adım	-2 log likelihood	Cox & Snell R kare	Nagelkerke R kare
5	468,212	,268	,362

Modele ait Nagelkerke R kare =0,362 olarak bulunmuştur. Yani kurulan lojistik modelin kullanılan değişkenlerce açıklanma oranı %36'dir. Modelin anlamlı olduğu söylenebilir.

Mcfaden R^2, Cox-Snell ve Nagelkerke R^2 bu katsayılar doğrusal regresyondan farklı olarak genel olarak küçük çıkma eğiliminde olduğundan, model uyumunu değerlendirmekten çok model oluşturma aşamasında faklı modellerin performansını değerlendirmek için kullanılmaları önerilmektedir [46].

Tablo 3.14. Lojistik Regresyon Sınıflandırma Tablosu

Adım-5			Tahmin Değerleri			
			Nöropati			
			Yok	Var	Toplam	Doğruluk (%)
Gerçek Değerler	Nöropati	Yok	210	56	266	% 78,9
		Var	72	112	184	% 60,9
Toplam			282	168	450	% 71,6

Tablo 3.14' de görüldüğü gibi, 450 hastadan gerçekte 266 tanesinde nöropati görülmezken, Geriye kalan 184 hastada görülmektedir. Nöropati görülmeyen 266 hastanın 210 tanesi LRA ile yapılan sınıflandırma işleminde doğru, 56 tanesi ise hatalı olmak üzere % 78,9 'lük bir doğruluk yüzdesi ile sınıflandırılmıştır. Nöropati görülen 184 hastanın ise 112 tanesi LRA ile yapılan sınıflandırma işleminde doğru, 72 tanesi ise hatalı olmak üzere % 60,9'lik bir doğruluk yüzdesi ile sınıflandırılmıştır. LRA ile yapılan sınıflandırma işleminde genel doğruluk değeri ise 450 hastanın 322 tanesi doğru sınıflandırılarak % 71.6 olarak hesaplanmıştır.

CHAID analizi ile sınıflandırma

Karar ağaçları algoritmaları uygulama bakımından bir hedef değişken (bağımlı değişken) ve hedef değişkeni açıklamaya yönelik kullanılacak açıklayıcı değişkenler (bağımsız değişkenler) olmak üzere iki grup değişken ile gerçekleştirilmektedir.

Risk değeri, sonuç modeli ve modelleri oluşturmak için bilgi sağlayan CHAID model tablosunun en uygun ağaç yapısını belirlemede bir göstergedir. Risk değerleri ise standart hata, yerine koyma (resubstitution) ve çapraz geçerlilik (crossvalidation) yüzdelerinden oluşmaktadır. En uygun modellerin seçiminde yerine koyma, çapraz geçerlilik ölçütlerinden yararlanılmıştır. Hedef değişken olarak nöropati değişkenini tanımlayan modelde riskin ölçüsü yanlış sınıflandırma oranını göstermektedir. Modelin uygulanması sırasında çapraz geçerlilik kuralı seçilmiş, nöropati bağımlı değişkeni için yerine koyma ve çapraz geçerlilik risk değerleri birbirine en yakın olan ağaç yapısı dikkate alınarak, CHAID analizinde büyük örneklem büyüklükleri için çocuk düğüm sayısı 10 ebeveyn düğüm sayısı 20, Aksine küçük örneklemler için çocuk düğüm sayısı 5 ebeveyn düğüm sayısı 10 olarak alınabilir [47] ifadesi ile çapraz geçerlilik ve yerine koyma risk

değerleri göz önünde tutularak çocuk düğüm sayısı 10 ve ebeveyn düğüm sayısı 20 olması durumunda ağacın durdurulmasına karar verilmiştir. Ayırma ve birleştirme için 0,05 alfa değeri seçilmiştir. çapraz geçerlilik ve yerine koyma, standart hata tablo 3.15 şekilde gösterildiği gibidir.

Tablo 3.15. Risk ve Standart Hata Tablo Değeri

Metod	Risk (%)	Standart Hata (%)
Yerine koyma (resubstitution)	,256	,021
Çapraz geçerlilik (crossvalidation)	,278	,021

Tablo 3.16. Çözümlenen modelin özeti

Model Özeti		
Tanımlar	Büyüme Tekniği	CHAID
	Bağımlı Değişken	Nöropati (Yok-Var)
	Bağımsız Değişkenler	Yaş
		Cinsiyet
		Soy Geçmişi
		Hastalığın Yaşı
		İnsülin Kullanma Durumu
		KVS
		Trigliserit
		AST
		ALT
		Glukoz
		Kolesterol
		Hiperlipidemi
		Hipertansiyon
		Hb1ac
	Maksimum Ağaç Uzunluğu	3
	Ana Düğümdeki Minimum Durum Sayısı	20
	Alt Düğümdeki Minimum Düğüm Sayısı	10
Sonuçlar	Modele Giren Bağımsız Değişkenler	Hastalığın Yaşı
		Yaş
		Hiperlipidemi
		KVS
	Düğüm Sayısı	16
	Terminal Düğüm Sayısı	9
	Uzunluk	3

Tablo-3.16 model hakkında genel bilgileri içermektedir.

Tablo 3.17. Kazanç Tablosu

Kazanç Tablosu						
Düğümler	**Düğüm**		**Kazanç**		**Cevap**	**İndeks**
	N	%	N	%	%	%
15	20	4,4	20	10,9	100,0	244,6
14	75	16,7	51	27,7	68,0	166,3
13	40	8,9	25	13,6	62,5	152,9
8	76	16,9	44	23,9	57,9	141,6
6	78	17,3	21	11,4	26,9	65,8
12	17	3,8	4	2,2	23,5	57,5
5	56	12,4	13	7,1	23,2	56,8
11	35	7,8	6	3,3	17,1	41,9
10	53	11,8	0	,0	,0	,0

Kazanç Tablosu'nda;

Düğüm (Node) : Düğümlerin sıra sayısı

Düğüm (n) : Hedef sınıfındaki örnek durumların sayısı

Düğüm (%) : Hedef sınıfına düsen toplam örnek durumlarının yüzdesi

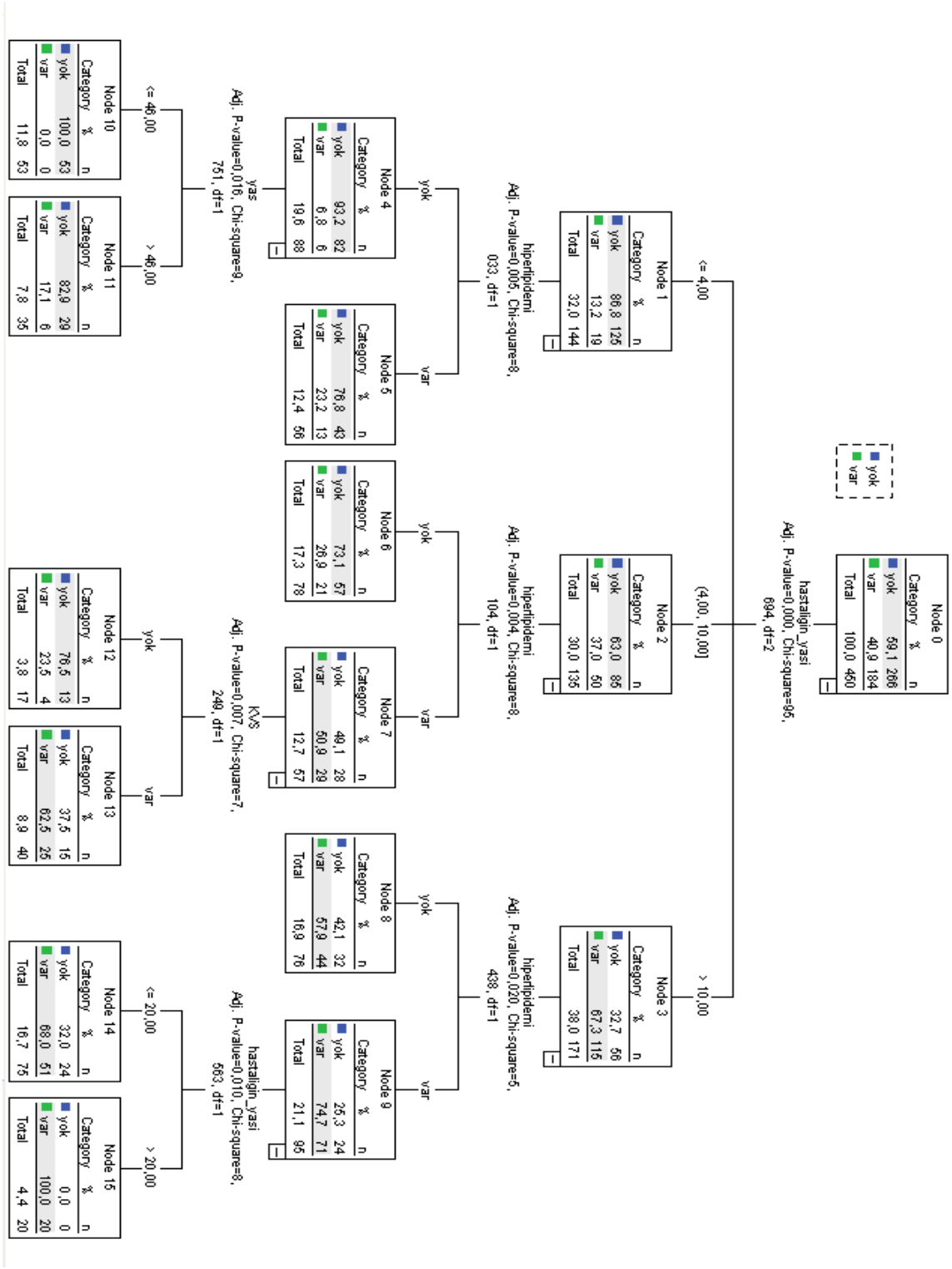
Kazanç (n) : Grup için kazanç değeri (Gruptaki nöropati olanların sayısı)

Kazanç (%) : Grup için kazanç yüzdesi (Gruptaki nöropati olanların oranı)

Cevap (%) : Grup için cevap yüzdesi

İndeks (%): Verilerin hepsi için, grup kazancının kazanca oranı [48].

Kazanç Tablosu (Gain Summary) hedef kategorideki (burada nöropatisi olanlar seçilmiştir) hangi düğümün en düşük ya da en yüksek orana sahip olduğunu göstermektedir. Hangi düğümün nöropatisi olanları belirlemede en çok etkili olduğu bilinmek istenir. Bunun için kazanç tablosu incelendiğinde, nöropatisi olanları belirlemede onbeşinci düğümün en belirleyici düğüm olduğu gözlemlenmektedir (%100,0).En az belirleyici olan düğüm ise %0 oranıyla onuncu düğümdür. Cevap sütunu ise düğümdeki nöropatisi olanların oranını göstermektedir. İndeks sütunundaki değerler cevap değerinin örneklemede nöropatisi olanların oranına bölümü şeklinde hesaplanır. Ve indeks değeri yüksek olan düğümlerdeki özellikler riskli olan grupları tanımlar. Burada onbeşinci düğümdeki nöropati olgusu incelendiğinde, Hiperlipidemisi olan ve hastalığın yaşı 20 yaş ve üzeri olan riskli gruplardır, yorumu yapılabilinir.



Şekil 3.1. Analiz sonucunda elde edilen karar ağacı

CHAID analizi sonucu oluşturulan bu ağaçta, 450 tane diabet hastasının %40,9' da nöropati görülürken %59.1'de görülmemektedir. Nöropati değişkeninin belirleyicileri arasında hastalığın yaşı, hiperlipidemi, KVS ve yaş değişkenleri yer almaktadır. Nöropati olma riskinin en iyi belirleyicisi olarak hastalığın yaşı değişkeninin olduğu görülmektedir. Hastalığın yaşı arttıkça nöropati görülme oranlarında artış yaşanmaktadır. Hastalığın yaşı 4 ve altında olanlarda nöropati görülme oranı %13,2 hastalığın yaşı 4 ile 10 arasında olanlarda nöropati görülme oranı %37 iken hastalığın yaşı 10'un üzerinde olanlarda bu oran %67,3 tür.

Nöropati değişkenini ikinci en iyi tahmin eden değişken ise hiperlipidemi değişkenidir. Hastalığın yaşı 4 ve altında olan grup için hiperlipidemisi olmayanlarda nöropati görülme oranı %6,8 iken, hiperlipidemisi olanlarda bu oran %23,2 ve % 7,1 kazanç oranına sahiptir.

Hastalığın yaşı 4 ile 10 arasında olan hasta grubundan hiperlipidemisi olmayanlarda nöropati görülme oranı %26,9 ve %11,4 kazanç oranına sahipken, hiperlipidemisi olanlarda bu oran %50,9'dur. Hastalığın yaşı 10'un üzerinde olan hastalardan hiperlipidemisi olmayanlarda nöropati görülme oranı %57,9 ve %23,9 kazanç oranına sahip iken hiperlipidemisi olanlarda nöropati görülme oranı %74,7'dir.

Hastalığın yaşı 4 ve altında olup, hiperlipidemisi olmayan hasta grubu için yaşı 46 ve üzerinde olanlarda nöropati oranı %17,1 ve %3,3 kazanç oranına sahip iken yaşı 46 ve altında olanlarda nöropati görülme oranı %0 ve % 0 kazanç oranıyla en risksiz gruptur.

Hastalığın yaşı 4 ile 10 arasında olup, hiperlipidemisi olan hastalık grubu için, KVS'si olmayanlarda nöropati görülme oranı %23,5 ve %2,7 kazanç oranına sahipken KVS'si olanlarda nöropati görülme oranı %62,5 ve %13,6 ile en riskli gruptur.

Hastalığın yaşı 10'un üzerinde olup, hiperlipidemisi olan hastalık grubu için, hastalığın yaşı 20 ve altında olanlarda nöropati görülme oranı %68 ve %27,7 kazanç oranına sahipken, hastalığın yaşı 20 ve üzerinde olanlarda nöropati görülme oranı %100 ve kazanç oranı%10,9 olup, bu grup ve tüm düğümler içerisindeki en riskli hasta grubu olmaktadır.

Tablo 3.18. CHAID Analizinin Sınıflandırma Tablosu

Adım-5			Tahmin Değerleri			
			Nöropati			
			Yok	Var	Toplam	Doğruluk (%)
Gerçek Değerler	Nöropati	Yok	195	71	266	% 73,3
		Var	44	140	184	% 76,1
Toplam			239	211	450	% 74,4

Tablo–3.18' de görüldüğü gibi, 450 hastadan gerçekte 266 tanesinde nöropati görülmezken, Geriye kalan 184 hastada görülmektedir. Nöropati görülmeyen 266 hastanın 195 tanesi CHAID Analizi ile yapılan sınıflandırma işleminde doğru, 71 tanesi ise hatalı olmak üzere % 73,3'lük bir doğruluk yüzdesi ile sınıflandırılmıştır. Nöropati görülen 184 hastanın ise 140 tanesi CHAID Analizi ile yapılan sınıflandırma işleminde doğru, 44 tanesi ise hatalı olmak üzere % 76,1'lik bir doğruluk yüzdesi ile sınıflandırılmıştır. CHAID Analizi ile yapılan sınıflandırma işleminde genel doğruluk değeri ise 450 hastanın 335 tanesi doğru sınıflandırılarak % 74,4 olarak hesaplanmıştır.

4. SONUÇ VE TARTIŞMA

Günümüzde veri miktarındaki artış, verilerin bilgiye dönüştürülmesini zorunluluk haline getirmiştir. Bu çalışmada veri madenciliği tekniklerinden biri olan karar ağaçlarından yararlanılmıştır.

Bağımlı değişkenin kategorik var-yok (0-1; binary) olduğu durumlarda kullanılan lojistik regresyon tekniği ile modellemeler yapılıp, modele ilişkin risk faktörleri (odds ratio) de tahminlenebilmektedir. CHAID analizinde ise, bunun yapılabilme gücünün daha yüksek olduğu bildirilmektedir. Çünkü CHAID analizinde bağımsız değişkenlerin en üst düzeydeki etkileşimlerini modele alan algoritma sayesinde, benzer özellikleri taşıyan karakterler aynı homojen düğümlere taşınmaktadır. Böylece, elde edilen karar ağacında bütün detaylar net bir şekilde izlenebildiği gibi, elde edilen regresyon denklemine ait parametrelerin de daha güvenilir sonuçlar vermesi beklenmektedir [18].

Bu nedenle genellikle verilerin sınıflandırılmasında kullanılan geleneksel yöntemlerden biri olan lojistik regresyon analizi ile yeni bir yöntem olan karar ağacı algoritmalarından CHAID analizinin sınıflama özellikleri karşılaştırılmıştır. Karşılaştırma için Fırat Üniversitesi Hastanesinden dosya tarama işlemiyle 450 tane diyabet hastası ele alınmıştır. Bu hastaların nöropati olup olmamalarına göre sınıflandırma işlemi gerçekleştirilmiştir. Oluşturulan modellerin veri kümesi üzerinde sınıflandırma başarısının bir ölçüsü olan doğru sınıflandırma oranları açısından büyük bir faklılık göstermeyip, yakın sonuçlar verdiği gözlenmiştir. CHAID analizi sonucunda kurulan modelin genel sınıflandırma başarısı % 74,4 ve lojistik regresyon sonucunda kurulan modelin genel sınıflandırma başarısının % 71,6 olduğu görülmektedir. Nöropatisi olanlarını sınıflamada CHAID analizi %76,1 oranla bulmuşken lojistik regresyon analizinde bu oran %60,9 dur. Bu sonuçlarla birlikte CHAID analizinin daha yüksek sınıflandırma başarısına sahip olduğu görülmektedir.

Her iki uygulamanın da sonuçları baz alınarak yapılan inceleme sonucunda, nöropati olma durumunu etkileyen bazı değişkenler; Yaş, Hb1ac, ALT, KVS, Hiperlipidemi, hastalığın yaşı şeklinde sıralanmaktadır. İlgili veri setine göre her iki modele de giren hiperlipidemi, yaş ve hastalığın yaşı değişkenleri bu çalışma için nöropatiyi belirlemede etkili değişkenler olarak gözlemlenmişlerdir.

Hastalığın yaşı 4 ve altında olup hiperlipidemisi olmayan grupta 46 yaş altı grubun en risksiz grup olduğu kararı alınabilir. Ayrıca hastalığın yaşı 10'un üzerinde olup

hiperlipidemisi olanlarda hastalığın yaşı 20 ve üzerinde olanlar en riskli grup olduğu söylenebilir.

Bu çalışma sonuçları açısından faydalı bilgiler sunmaktadır. Ancak bulunan etkenlerin kesinlikle kullanılabilir olduğunu kanıtlayabilmek ve çesitlendirebilmek için örneklem kümesinin büyütülüp başka analizlerle de sınanmasında fayda vardır.

Sonuçların görselliği sınıflandırmanın daha geniş olması ve kolay yorumlanabilir olması nedeniyle karar ağacı tekniklerinin cevapları ortaya çıkarmakta etkili olduğun göstermektedir.

Sonuç olarak yapılan bu çalışma, heterojen veri setlerinde homojen düğümler oluşturması, parametrik olmayan yöntemler arasında olması nedeniyle diğer çok değişkenli tekniklerde sağlanması gereken istatistik varsayımların söz konusu olmaması, sürekli ve kategorik verilerin aynı anda modele dahil olması, bağımlı ve bağımsız değişkenler arasındaki ilişkilerin yönünü, önem sırasını değerlendirilerek görsellik sunması gibi özellikleriyle CHAID analizinin her alanda olduğu gibi sağlık bilimleri alanında da verimli bir şekilde kullanılabileceğine yönelik bir örnek oluşturduğu söylenebilir.

KAYNAKLAR

[1]. Ayık, Y. Z., Özdemir A., Yavuz U., 2007. Lise Türü Ve Lise Mezuniyet Başarısının, Kazanılan Fakülte İle İlişkisinin Veri Madenciliği Tekniği İle Analizi, *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi.*

[2]. Kıran, Z.B., 2010. Lojistik regresyon ve Cart analizi teknikleriyle Sosyal Güvenlik Kurumu ilaç provizyon sistemi verileri üzerinde bir uygulama, *Yüksek Lisans Tezi*, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

[3]. Yılmaz, Ş.K., 2008. Veri Madenciliği:İstanbul Menkul Kıymetler Borsası örneği, *Yüksek Lisans Tezi*, Zonguldak Karaelmas Üniversitesi Sosyal Bilimler Enstitüsü, Zonguldak.

[4]. Thomas, Lyn. C., 2000. A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Consumer", *International Journal of Forecasting*.

[5]. Pehlivan, G., 2006. CHAID Analizi ve Bir Uygulama, *Yüksek Lisans Tezi*, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

[6]. Temel, G. O., Çamdeviren, H., Akkuş, Z., 2005. Sınıflama Ağaçları Yardımıyla Restless Legs Syndrome (RLS) Hastalarına Tanı Koyma, *İnönü Üniversitesi Tıp Fakültesi Dergisi.*

[7]. Albayrak A.S., Yılmaz Ş.K., 2009. Veri Madenciliği: Karar Ağacı Algoritmaları ve İMKB verileri üzerine bir uygulama, *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi,* 31-52.

[8]. Özkan Y., 2008. Veri Madenciliği Yöntemleri, Papatya Yayıncılık Nisan.

[9]. Sancak, S., 2008. Saldırı Tespit Sistemleri Tekniklerinin Karşılaştırılması, *Yüksek Lisans Tezi*, Gebze Yüksek Teknoloji Enstitüsü, Sosyal Bilimler Enstitüsü, Gebze.

[10]. Oğuzlar, A., 2004. CART Analizi İle Hanehalkı İşgücü Anketi Sonuçlarının Özetlenmesi, *Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*.

[11]. Hoare, R., 2004. Using CHAID for Classification Problems, New Zealand Association 2004 Conference, Wellington, s.1

[12]. Yerge N., Kara G., 2009. Kriz Döneminde Kaliteden Kaçış, Eskişehir Osmangazi Üniversitesi Mühendislik Ve Mimarlık Fakültesi Endüstri Mühendisliği Bölümü, Şubat

[13]. SPSS Classification trees 15.0, SPSS Inc. USA

[14]. Koyuncugil, A.S., 2007. Borsa şirketlerinin Sektörel Risk Profillerinin Veri Madenciliğiyle Değerlendirilmesi, *Sermaye Piyasası Kurulu Arastırma Raporu*, Arastırma Dairesi, Ankara.

[15]. Doğan İ., 2003. Holştayn Irkı İneklerde Süt Verimine Etki Eden Faktörlerin CHAID Analizi İle İncelenmesi *Ankara Üniversitesi Veterinerlik Fakültesi Dergisi* 50,65-70.

[16]. Satıcı Ö., Akkuş Z., Alp A., 2009. Tıp Fakültesi Öğretim Elemanlarının Teknolojiye İlişkin Tutumlarının CHAID Analizi İle İncelenmesi*, Dicle Üniversitesi Tıp Fakültesi Dergisi.*

[17]. McCarty John A., Hastak M., 2007 . Segmentation Approaches in Data-Mining: A Comparison of RFM, CHAID, And Logistic Regression , *Journal of Business Research*

[18]. Kayrı M., Boysan M., 2007. Araştırmalarada CHAID Analizinin Kullanımı Ve Baş Etme Stratejileri , *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*

[19]. Koyuncugil A.S., Özgülbaş N., 2008. İMKB' DE İşlem Gören Kobi'lerin Güçlü Ve Zayıf Yönleri: CHAID Karar Ağacı Uygulaması, *Dokuz Eylül Üniversitesi İktisadi Ve İdari Bilimler Fakültesi Dergisi*

[20]. Üngüren D., Doğan H., 2010. Beş Yıldızlı Konaklama İşletmelerinde Çalışanların İş Tatmin Düzeylerinin CHAID Analiz yöntemiyle Değerlendirilmesi, C.Ü. *İktisadi Ve İdari Bilimler Dergisi*.

[21]. Karedeniz, N., 2008. Müşteri İlişkileri Yönetimi Açısından Veri Madenciliği Yöntemi Ve Hizmet Sektörü Üzerine Bir Uygulama, *Yüksek Lisans Tezi,* Marmara Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Anabilim Dalı ,İstanbul.

[22]. G.Piatetsky-Shapiro, and W.J.Fawley., 1991. Knowledge Discovery in Databases**,** AAAI/MIT Pres,

[23]. Koyuncugil, A.S., ve Özgülbaş, N., 2009. Veri Madenciliği: Tıp Ve Sağlık Hizmetlerinde Kullanımı Ve Uygulamaları,

[24]. Berson, A., Smith, S.,2000, Thearling, K.: Building Data Mining Applications for CRM, McGraw-Hill Professional Publishing, New York, USA.

[25]. Chaudhuri, S., 1998. Data Mining and Database Systems : Where is the Intersection?, IEEE Bulletin of the Technical Committee on Data Engineering, Vol.21 No.1, 4-8.

[26]. Özekes S., Çamurcu A.Y., 2002. Veri Madenciliğinde Sınıflama Ve Kestirim Uygulaması, *Marmara Üniversitesi Fen Bilimleri Dergisi* 18.

[27]. Akpınar, H., 2000. Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği, İstanbul Üniv. *İşletme Fakültesi Dergisi,* C:29 S: 1 Nisan.

[28]. Özekes, S., Veri Madenciliği Modelleri Ve Uygulama Alanları *İstanbul Ticaret Üniversitesi Dergisi*

[29]. www.deu.edu.tr/userweb/k.yaralioğlu/dosyalar/ver_mad.doc

[30]. Kayaalp, K., 2007. Asenkron Motorlarda Veri Madenciliği İle Hata Tespiti, *Yüksek Lisans Tezi,* Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü,Isparta.

[31]. L.L.Harlow., 2005 The Essence Of Multivariate Thinking : Basic Themes And Methods, LEA Publishers.

[32]. Kasap. E., 2007 Sigortacılık Sektöründe Müşteri İlişkileri Yönetimi Yaklaşımıyla Veri Madenciliği Teknikleri Ve Bir Uygulama, *Yüksek Lisans Tezi,* Marmara Üniversitesi Bankacılık Ve Sigortacılık Enstitüsü Sigortacılık Bölümü İstanbul.

[33]. Argüden. Y, Erşahin B., 2008. Veri Madenciliği veriden bilgiye, masraftan değere ARGE Danışmanlık A.Ş

[34]. Koyuncugil, A.S., 2006. Bulanık Veri Madenciliği Ve Sermaye Piyasalarına Uygulanması, *Doktora Tezi* Ankara Üniversitesi Fen Bilimleri Enstitüsü Ankara.

[35]. Albayrak. S., Sınıflama Ve Kümeleme Yöntemleri, Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü

[36]. Kaşko. Y., Çoklu Bağlantı Durumunda İkili(binary) Lojistik Regresyon Modelinde Gerçekleşen 1.Tip Hata ve Testin Gücü, *Yüksek Lisans Tezi*, Ankara Üniversitesi, Fen Bilimleri Enstitüsü, Zootekni Anabilim Dalı.

[37]. Çolak E., 2002. Koşullu Ve Sınırlandırılmış Lojistik Regresyon Yöntemlerinin Karşılaştırılması Ve Bir Uygulama, *Yüksek Lisans Tezi*, Osman Gazi Üniversitesi, Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalı.

[38]. Kartalkanat A., 2006. Ziraat İle İlgili Çalışmalarda Elde Edilen Kategorik Verilere İki Seviyeli Lojistik Regresyon Analizinin Uygulanması, Yüzüncü Yıl Üniversitesi, Fen Bilimleri Enstitüsü, Zootekni Anabilim Dalı.

[39]. Ocakoğlu G., 2006. Lojistik Regresyon Analizi Ve Yapay Sinir Ağı Tekniklerinin Sınıflama Özelliklerinin Karşılaştırılması Ve Bir Uygulama, *Yüksek Lisans*

Tezi, Uludağ Üniversitesi, Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalı.

[40]. Ürük. E., 2007. İstatistiksel Uygulamalarda Lojistik Regresyon Analizi, *Yüksek Lisans Tezi*, Marmara Üniversitesi, Fen Bilimleri Enstitüsü

[41]. Özdamar. K., 2004. Paket Programlar İle İstatistiksel Veri Analizi, Kaan Kitabevi, Eskişehir.

[42]. Hosmer. D.W., Lemeshow. Jr. S., 1989. Applied Logistic Regression, John Wiley& Sons

[43]. Arıcan. E., 2010, Nitel Yanıt Değişkene Sahip Regresyon Modellerinde Tahmin Yöntemleri, *Yüksek Lisans Tezi*, Çukurova Üniversitesi, Fen Bilimleri Enstitüsü.

[44]. Çakır. Ö., 2008. Veri Madenciliğinde Sınıflandırma Yöntemlerinin Karşılaştırılması Bankacılık Müşteri Veri Tabanı Üzerinde Bir Uygulama, Doktora Tezi, Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, İşletme Anabilim dalı, Sayısal Yöntemler Bilim Dalı.

[45]. Girginer. N., Cankuş. Bülent., 2008. Tramvay Yolcu Memnuniyetinin Lojistik Regresyon Analiziyle Ölçülmesi Estram Örneği, Celal Bayar Üniversitesi,*Yönetim Ve Ekonomi Dergisi* C:15 S1.

[46]. ALPAR. R., 2011. Uygulamalı çok değişkenli istatistiksel yöntemler detay yayıncılık Ankara .

[47]. Collins. K.M, Onwuegbuzie. A.J., Jiao. Q.G., Toward a Broader Understanding Of Stress and Coping

[48]. Turangil. N., 2006 Kredi Skorlamada Kullanılan Yöntemler Ve Uygulamaları, *Yüksek Lisans Tezi*, Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü

[49]. Burmaoğlu. S., 2009. Birleşmiş Milletler Kalkınma programı beşeri kalkınma endeksi verilerini kullanarak diskriminant analizi, lojistik regresyon analizi ve yapay sinir ağlarının sınıflandırma başarılarının değerlendirilmesi, *Doktora Tezi*, Atatürk Üniversitesi, Sosyal Bilimler Enstitüsü, İşletme Ana Bilim Dalı.

[50]. http://www.diyabetikcocuk.com/index.php/diyabet-ve-noropati.html

[51]. http://www.burhanettinuludag.com.tr/yazilar/hekimler/styled-5/files/dpn-bilinmeyenler.pdf

[52]. http://www.diabetvakfi.org/inf.php?partid=5&catid=5&pid=31

[53]. http://www.ailem.com/templates/library/1759.asp?id=13132

ÖZGEÇMİŞ

1986 yılında Elazığ'da doğan Ebru EKİCİ, orta öğrenimini, 2003 yılında tamamlamış, yükseköğrenimini ise 2006-2010 yılları arasında Fırat Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümünde tamamlamıştır. 2010 yılında Fırat Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, Olasılık Süreçleri ve Olasılık Teorisi Bilim Dalında yüksek lisans eğitimine başlamıştır ve halen eğitimini sürdürmektedir.