# LOJİSTİK REGRESYON ANALİZİ İLE CART ALGORİTMASININ KARŞILAŞTIRILMASI: ÜLKEMİZDE MEYDANA GELEN TRAFİK KAZALARI ÜZERİNE BİR UYGULAMA

**Ahmet TEKİN**

**Yüksek Lisans Tezi**
**Ekonometri Ana Bilim Dalı**
**Prof. Dr. Mehmet Suphi ÖZÇOMAK**
**2023**

**T.C.**
**ATATÜRK ÜNİVERSİTESİ**
**SOSYAL BİLİMLER ENSTİTÜSÜ**
**EKONOMETRİ ANA BİLİM DALI**

**Ahmet TEKİN**

# LOJİSTİK REGRESYON ANALİZİ İLE CART ALGORİTMASININ KARŞILAŞTIRILMASI: ÜLKEMİZDE MEYDANA GELEN TRAFİK KAZALARI ÜZERİNE BİR UYGULAMA

**YÜKSEK LİSANS TEZİ**

**TEZ YÖNETİCİSİ**
**Prof. Dr. Mehmet Suphi ÖZÇOMAK**

**ERZURUM – 2023**

**SOSYAL BİLİMLER ENSTİTÜSÜ**
Graduate School of Social Sciences

# TEZ BEYAN FORMU

## SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜNE

## BİLDİRİM

***Atatürk Üniversitesi Lisansüstü Eğitim ve Öğretim Uygulama Esaslarının ilgili maddelerine*** göre hazırlamış olduğum "**LOJİSTİK REGRESYON ANALİZİ İLE CART ALGORİTMASININ KARŞILAŞTIRILMASI: ÜLKEMİZDE MEYDANA GELEN TRAFİK KAZALARI ÜZERİNE BİR UYGULAMA**" adlı tezin/raporun tamamen kendi çalışmam olduğunu ve her alıntıya kaynak gösterdiğimi taahhüt eder, tezimin/raporumun kâğıt ve elektronik kopyalarının aşağıda belirttiğim koşullarda saklanmasına izin verdiğimi onaylarım.

Gereğini bilgilerinize arz ederim *.

☐ Tezimin/Raporumun tamamı her yerden erişime açılabilir.

☐ Tezimin/Raporumun makale için **altı ay,** patent için **iki yıl** süreyle erişiminin ertelenmesini istiyorum.

**14/12/2024**

**Ahmet TEKİN**

---

* **LİSANSÜSTÜ TEZLERİN ELEKTRONİK ORTAMDA TOPLANMASI, DÜZENLENMESİ VE ERİŞİME AÇILMASINA İLİŞKİN YÖNERGE**

.....

**ÜÇÜNCÜ BÖLÜM**

**Çeşitli ve Son Hükümler**

**Lisansüstü tezlerin erişime açılmasının ertelenmesi MADDE 6– (1)** Lisansüstü tezle ilgili **patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda**, tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu **iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.**

**(2)** Yeni teknik, materyal ve metotların kullanıldığı, henüz **makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış** ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç imkanı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulunun gerekçeli kararı ile **altı ayı aşmamak üzere tezin erişime açılması engellenebilir**.

**Gizlilik dereceli tezler MADDE 7– (1)** Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, tezin yapıldığı kurum tarafından verilir. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, ilgili kurum ve kuruluşun önerisi ile enstitü veya fakültenin uygun görüşü üzerine üniversite yönetim kurulu tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.

**(2)** Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir.

SOSYAL BİLİMLER ENSTİTÜSÜ
Graduate School of Social Sciences

**TEZ KABUL TUTANAĞI**

**SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜNE**

Prof. Dr. Mehmet Suphi ÖZÇOMAK danışmanlığında, Ahmet TEKİN tarafından hazırlanan bu çalışma 14 / 12 / 2023 tarihinde aşağıda isimleri yazılı jüri tarafından. Ekonometri Anabilim Dalı'nda Yüksek Lisans Tezi olarak kabul edilmiştir.

| | | |
|---|---|---|
| **Başkan** | : Prof. Dr. Selahattin YAVUZ | İmza: |
| **Jüri Üyesi** | : Prof. Dr. Mehmet Suphi ÖZÇOMAK | İmza: |
| **Jüri Üyesi** | : Doç. Dr. Hakan EYGÜ | İmza: |

Prof. Dr. Sait UYLAŞ
Enstitü Müdürü

OF27b_V4_03.11.2019

# İÇİNDEKİLER

# ÖZET

## YÜKSEK LİSANS TEZİ

## LOJİSTİK REGRESYON ANALİZİ İLE CART ALGORİTMASININ KARŞILAŞTIRILMASI: ÜLKEMİZDE MEYDANA GELEN TRAFİK KAZALARI ÜZERİNE BİR UYGULAMA

**Ahmet TEKİN**

**Tez Danışmanı: Prof. Dr. Mehmet Suphi ÖZÇOMAK**

**2023, 88 Sayfa**

**Jüri: Prof. Dr. Mehmet Suphi ÖZÇOMAK**
**Prof. Dr. Selahattin YAVUZ**
**Doç. Dr. Hakan EYGÜ**

Bu çalışmanın temel amacı; son yıllarda literatürde yaygın olarak kullanılan lojistik regresyon analizi ile CART algoritmasının karşılaştırılmasıdır. Bu amaçla 2013 – 2022 yılları arasında ülkemizde meydana gelen trafik kaza verileri kullanılarak lojistik regresyon analizi ile CART algoritması karşılaştırılmıştır.

Çalışmada ülkemizde meydana gelen trafik kazalarına sebep olan faktörler ortaya konularak trafik kazalarının ölümlü ya da yaralamalı olma durumlarına göre söz konusu faktörler ayrı ayrı incelenmiştir.

Çalışma ilk iki bölüm teori ve son bölüm uygulama olmak üzere üç bölümden oluşmaktadır. Birinci bölümde Veri madenciliği ile karar ağaçları kısaca tanıtılmış ve CART algoritmasının gelişimi, uygulama alanı ve uygulama süreci anlatılmıştır. İkinci bölümde lojistik regresyon analizi ayrıntılı olarak incelenmiştir. Son bölüm olan üçüncü bölümde ise Emniyet Genel Müdürlüğü Trafik Başkanlığından alınan 2013 – 2022 yılları arasında meydana gelen ölümlü veya yaralamalı trafik kaza veriler kullanılarak lojistik regresyon analizi ile CART algoritması karşılaştırılmıştır.

Çalışmanın sonucunda; 2013 – 2022 yılları arasında meydana gelen ve yayaların karıştığı trafik kazaları bakımından iki metodun performansları karşılaştırıldığında; lojistik regresyon analizi ile CART algoritması arasında sınıflandırma açısından bir farkın olmadığı tespit edilmiştir. Sürücülerin karıştığı trafik kazaları bakımından iki metodun performansları karşılaştırıldığında ise CART algoritmasının sınıflandırma performansının lojistik regresyon analizine göre daha iyi olduğu tespit edilmiştir.

**Anahtar Kelimeler:** Karar Ağaçları, CART Algoritması, Lojistik Regresyon, Trafik Kazası

## ABSTRACT

### MASTER'S THESIS

### COMPARISON OF LOGISTIC REGRESSION ANALYSIS AND CART ALGORITHM: AN APPLICATION ON TRAFFIC ACCIDENTS THAT OCCURRED IN OUR COUNTRY

**Ahmet TEKİN**

**Advisor: Prof. Dr. Mehmet Suphi ÖZÇOMAK**

**2023, 88 Pages**

**Jury: Prof. Dr. Mehmet Suphi ÖZÇOMAK**
**Prof. Dr. Selahattin YAVUZ**
**Assoc. Prof. Dr. Hakan EYGÜ**

The main purpose of this study is to compare logistic regression analysis with the CART algorithm, which has been widely used in the literature in recent years. For this purpose, CART algorithm and logistic regression analysis were compared using traffic accident data that occurred in our country between 2013 and 2022.

In the study, the factors that cause traffic accidents in our country were revealed and these factors were examined separately according to whether the traffic accidents caused death or injury.

The study consists of three parts: the first two parts are theory and the last part is practice. In the first part, the development of the CART algorithm, its application area and application process are explained. In the second section, logistic regression analysis is examined in detail. In the third section, which is the last section, the CART algorithm and logistic regression analysis were compared using data on traffic accidents with fatalities or injuries occurring between 2013 and 2022, obtained from the Traffic Directorate of the General Directorate of Security.

As a result of the study; When the performances of the two methods are compared in terms of traffic accidents involving pedestrians that occurred between 2013 and 2022; It was determined that there was no difference in terms of classification between the CART algorithm and logistic regression analysis. When the performances of the two methods were compared in terms of traffic accidents involving drivers, it was determined that the classification performance of the CART algorithm was better than the logistic regression analysis.


**Keywords:** Decision Trees, CART Algorithm, Logistic Regression, traffic Accident

# TABLOLAR DİZİNİ

# ŞEKİLLER DİZİNİ

# ÖNSÖZ

"Lojistik Regresyon Analizi ve CART Algoritmasının Karşılaştırılması: Ülkemizde Meydana Gelen Trafik Kazaları Üzerine Bir Uygulama" adlı bu çalışma Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Anabilim Dalı Yüksek lisans Tezi olarak hazırlanmıştır. Bu tezin hazırlanması aşamasında yardımlarına ve tecrübelerine sıkılıkla başvurduğum danışman hocam Prof. Dr. Mehmet Suphi ÖZÇOMAK'a teşekkürlerimi bir borç bilirim.

Çalışmanın bütün aşamalarında yapıcı ve yönlendirici eleştirileri bana ekstra destek olan Prof. Dr. Selahattin YAVUZ hocama ve Doç. Dr. Hakan EYGÜ hocama ayrıca teşekkür ederim.

Akademik hayatım boyunca her daim yanımda olan, birbirimiz ile hep fikir alışverişinde olduğumuz ve eksikliğini bir an olsun hissetmediğim arkadaşım, kardeşim Arş. Gör. Hasan Hüseyin TEKMANLI'ya, çocukluk arkadaşım, Nazan DEMİRHAN'a da bana yol göstermesinden dolayı teşekkür ederim.

Ayrıca bu tez çalışması süresince, araştırmalarımda sağladıkları kolaylıklardan ve desteklerinden dolayı görevli bulunduğum şube müdürüm, sıralı amirlerim, iş arkadaşlarıma teşekkür ederim.

Son olarak desteklerini hiçbir zaman esirgemeyen babam, annem, abim, yengem, yeğenim ve tüm aileme sonsuz teşekkür ederim.

**Erzurum, 2023** **Ahmet TEKİN**

# GİRİŞ

On sekizinci yüzyılda sanayi devriminin de etkisiyle gelişen teknoloji beraberinde veri madenciliğini de etkileyerek birçok veri madenciliği algoritmasının gelişimine de katkı sağlamıştır. Veri madenciliğinin kavramsal olarak hayatımıza girişi 1960'lı yılları bulmuştur. Geleceği merak etme duygusuna yenik düşen insanoğlu eldeki mevcut veriler ile geleceği tahmin etmeye çalışmıştır. Bu çalışmalar iş hayatında maliyetlerin azalmasına ve daha çok anlamlı veriye ulaşmaya olanak sağlamıştır. Böylece ticari olarak iş hacmi genişlemiş ve araştırmacılar için daha çok veriye ulaşma imkânı doğmuştur.

Fakat teknolojinin bu denli hızlı ilerlemesi elde edilen verileri devasa hale getirmiş ve büyük veri depolarının yaygınlaşması beraberinde karmaşık verileri oluşturmuştur. Oluşan bu karmaşık durumdan kurtulabilmek için algoritmalar geliştirilmeye çalışılmış ve bu sayede veri madenciliği ortaya çıkmıştır. Verilerin bu denli hızlı ve her geçen gün artarak ilerleyişi beraberinde karmaşıklığı getirerek anlamlılığını zorlaştırmıştır. Bu karmaşık verilerden anlamlı sonuçlara ulaşabilmek için tahmin edici ve tanımlayıcı yöntemler kullanarak gelecek veriler için çıkarımlarda bulunmak veya verileri tanımlayarak çıkarımlarda bulunulmaya çalışılmıştır. Bu çıkarımların başında veri tabanlarında yapılan çalışmalar dikkat çekmektedir. Fakat verilerdeki olağanüstü artış karşısında çaresiz kalınmış ve verilerden faydalanılması oldukça güç bir duruma gelinmiştir. Bu durum karşısında eski yöntemler etkisiz kalmış ve araştırmacılar veri madenciliği başlığı altında yeni sınıflandırma metotları aramaya başlamıştır.

Bu keşfedilen metotların uygulamalar üzerinde başarı durumunu görebilmek için karşılaştırılması oldukça önemli bir durumdur. Çünkü her algoritmanın kullanım alanı ve kullandığı veri tipleri farklılık gösterdiğinden dolayı farklı sonuçlar verebilmektedir.

## I. ÇALIŞMANIN AMACI VE KAPSAMI

Karar ağacı yöntemlerinden biri olan CART algoritması veri sınıflandırma metotları içerisinde yaygın olarak kullanılan algoritmalardan biridir. 1984 yılında geliştirilmiştir. Görsel olarak kolay anlaşılabilen bir yöntem olan CART algoritması bölünerek sınıflandırma yapmaktadır. Bağımlı değişkenin kategorisine göre sınıflandırma ya da regresyon ağacı ismini almaktadır.

Bu çalışmada kullanılan bir diğer sınıflandırma yöntemi ise lojistik regresyon analizidir. Lojistik regresyon analizi de bağımlı değişkenin kategorik olarak incelendiği bir yöntemdir. CART algoritmasına göre daha eski bir yöntem olan lojistik regresyon analizi, bağımsız değişkenlerin sonuç üzerindeki etkilerini gösteren bir yöntemdir.

Çalışma üç bölümden oluşmaktadır Birinci bölümde Veri madenciliği ile karar ağaçları kısaca tanıtılmış ve CART algoritmasının gelişimi, uygulama alanı ve uygulama süreci anlatılmıştır. İkinci bölümde lojistik regresyon analizi ayrıntılı olarak incelenmiştir. Son bölüm olan üçüncü bölümde ise Emniyet Genel Müdürlüğü Trafik Başkanlığından alınan 2013 – 2022 yılları arasında meydana gelen ölümlü veya yaralamalı trafik kaza veriler kullanılarak lojistik regresyon analizi ile CART algoritması karşılaştırılmıştır. Ayrıca ülkemizde meydana gelen trafik kazaları; CART algoritması ve lojistik regresyon analizi ile irdelenerek trafik kazalarına sebep olan faktörler tespit edilmeye çalışılmıştır.

# BİRİNCİ BÖLÜM

# KARAR AĞAÇLARI VE CART ALGORİTMASI

## 1.1. VERİ MADENCİLİĞİ KAVRAMI

İnsanoğlunun varlığından günümüze kadar veri hayatımızın bir parçası olmuştur. Sanayi devriminin gelişmesiyle başlayan teknoloji ile elde edilen verinin hacmi gün geçtikçe artmıştır. Günümüzde ise elektronik cihazların günlük hayatımızda etkin bir şekilde kullanılması veri hacmini inanılmaz boyutlara ulaştırmıştır. Bu durumda dünya üzerinde var olan veri sayısının her yirmi ayda bir iki katına çıktığı tahmin edilmektedir (Olson ve Delen, 2008). Bu artış karşısında verilerin saklanması için kullanılan veri tabanları tek başlarına yetersiz olmaya başlamıştır (Şentürk, 2006). Bu büyük ölçekli verilerin anlamlı ve doğru analiz edilebilmesi veri madenciliği ile sağlanmaktadır. Ancak artan verilerden anlamlı bilgiler çıkarma işlemi; maliyetleri ve bunlar için harcanan zamanı artırmış ayrıca depolama sorunu ile birlikte işin içinden çıkılmaz bir hal almıştır (Özkan, 2016).

Bu açıklamalardan yola çıkarak veri madenciliği, dijital ortamda tutulan bu büyük verilerin birbirleri ile bağlantılı olacak şekilde eşleştirilerek oluşturulan algoritmalar ile gelecek hakkında yorumlarda bulunulması şeklinde ifade edilmiştir. Veri madenciliği literatürde birçok farklı şekilde tanımlanmıştır. Sumathi ve Sivanandam (2006) veri madenciliğini; "geniş veri tabanlarında önceden bilinmeyen değerli ve kullanışlı bilginin ortaya çıkarılması ve kritik iş kararlarında kullanılması süreci" olarak, Tang ve MacLennan (2005) "veri madenciliğini otomatik veya yarı otomatik biçimlerde verinin analiz edilerek gizli örüntülerin bulunması şeklinde" olarak, Cabena (1998) veri madenciliğini; "geniş veri tabanlarından bilgi çıkarımını hedeflemek için makine öğrenimini, örüntü tanıma, istatistik, veri tabanı ve görselleştirme tekniklerini bir araya getiren disiplinler arası bir alan" olarak ve Gartner Group (2007) ise veri madenciliğini; "veri ambarlarında depolanan büyük miktardaki verinin istatistiksel ve matematiksel tekniklerle birlikte örüntü tanıma teknolojilerinin de kullanılarak incelenmesi yoluyla anlamlı yeni ilişkiler, örüntüler ve eğilimler bulması süreci" olarak tanımlamıştır.

Bu tanımlamalardan yola çıkarak büyük veri setlerinde, veri tabanlarında veya veri ambarlarında bulunan veriler arasında var olan, bilinmeyen, klasik yöntemlerle görülemeyen ve sıradan olmayan ilişkileri, örüntüleri, belirli yapıları veya eğilimleri ortaya çıkarmak amacıyla istatistik, matematik, makine öğrenimi ve bilgisayar uygulamaları alanlarının birleşim ile istatistiki teknikler kullanılarak analiz edilmesi ve sonuçların anlamlı bir şekilde özetlenmesi ve görselleştirilmesi sürecidir (Irmak,2009).

## 1.2. VERİ MADENCİLİĞİNİN UYGULAMA ALANLARI

Geniş bir veri toplama alanı ve günümüz teknolojisi sayesinde hızlı, kolay ve anlaşılabilir olması, veri madenciliğinin oldukça yaygın olarak tercih edilmesinde etkin bir rol aldığı görülmüştür.

Veri madenciliği uygulaması, herhangi bir meslek kolu veya alanı ayırt etmeksizin, fazla miktarda veriye sahip olan ve bu verilerin işlenmesine imkân sağlayan tüm alanlarda kullanılabilir (Silahtaroğlu G, 2008). Teknolojinin gelişimi ve hayatımızda etkin bir şekilde rol alması dolayısıyla veri hacmi farklı farklı sektörlerde çok büyük boyutlara ulaştırmıştır. Biriken bu veri yığınları içerisinde veri işleme adımlarının bulunması veri madenciliğini istatistik, veri tabanı teknolojisi, makine öğrenmesi, bilgi bilimi ve görselleştirme teknikleri başta olmak üzere çeşitli disiplinlerle ilişkili hale getirmiştir (Han ve Kamber, 2001:31; Gülpınar, 2008:39).

Veri madenciliğinin bazı uygulama alanları Tablo 1.1'de gösterilmiştir.

**Tablo 1.1.** Veri Madenciliğinin Uygulama Alanları

| Kullanılan Sektörler | Kullanım Amaçları |
|---|---|
| Pazarlama | Market sepet analizi |
| | Pazarlama kampanyalarının planlanması |
| | Satış tahmini |
| | Çapraz satış |
| | Müşteri analizi |
| Bankacılık ve sigortacılık | Kredi talep değerlendirmesi |
| | Finansal tablo karşılaştırmasında korelasyonların tespiti |
| | Kredi risk analizlerinde |
| | Banka hesap bilgilerinin/kart bilgisi ile dolandırıcılığın tespiti |

**Tablo 1.1.** (Devamı)

| | |
|---|---|
| Borsa | Pazar analizi |
| | Alım-satım stratejilerinin en iyi duruma getirilmesinde |
| Telekomünikasyon | Mobil müşteri analizi |
| Perakendecilik sektörü | Satış noktası veri analizi |
| | Alışveriş sepeti analizi |
| | Alım stratejisi belirleme |
| | Raf düzeni ve mağaza yerleşimi belirleme |
| | Promosyon stratejisi belirleme |
| Sağlık ve ilaç | İlaçların geliştirilmesi |
| | Hastalıkların teşhisi |
| | Tedavi sürecinin belirlenmesi |
| Endüstri | Kalite kontrol tespitinde |
| | Lojistik ve üretim aşamalarının belirlenmesinde |
| Bilim, bilişim ve mühendislik | Hücre analizi |
| | İnternet dolandırıcılığı tespiti |
| | Bilgisayar ağlarına yetkisiz girilmesinin tespiti |
| | Parmak izi ve yüz şeklinden kimlik tespiti |
| | Uzay analizi |
| Biyoloji | Microarray veri analizi |

**Kaynak**: (Ezgi, 2021).

Oluşan devasa veri toplulukları birçok farklı sektörde de etkin bir şekilde veri madenciliği kullanımını artırmıştır. Çünkü firmalar ellerindeki bu devasa veriler ile pazar payında güçlü bir duruş ve rekabetlerini artırmayı hedeflenmektedir. Veri madenciliğinin kullanım alanlarına göre kullanım oranları Tablo 1.2'de gösterilmiştir (Gorunescu, 2011).

**Tablo 1.2.** Veri Madenciliğinin Kullanım Alanlarına Göre Kullanım Oranları

| **Kullanım Alanı** | **Kullanım Oranı (%)** |
|---|---|
| CRM/ Müşteri analitiği | 32,8 |
| Bankacılık | 24,4 |
| Direk pazarlama | 16,1 |
| Kredi puanlama | 15,6 |
| Telekomünikasyon | 14,4 |
| Dolandırıcılık tespiti | 13,9 |

**Tablo 1.2.** (Devamı)

| | |
|---|---|
| Satış | 11,7 |
| Sağlık | 11,7 |
| Finans | 11,1 |
| Bilim | 10,6 |
| Reklamcılık | 10,6 |
| E- Ticaret | 10 |
| Sigortacılık | 10 |
| Web madenciliği | 8,3 |
| Sosyal ağlar | 7,8 |
| İlaç | 7,8 |
| Biyoteknoloji | 7,8 |

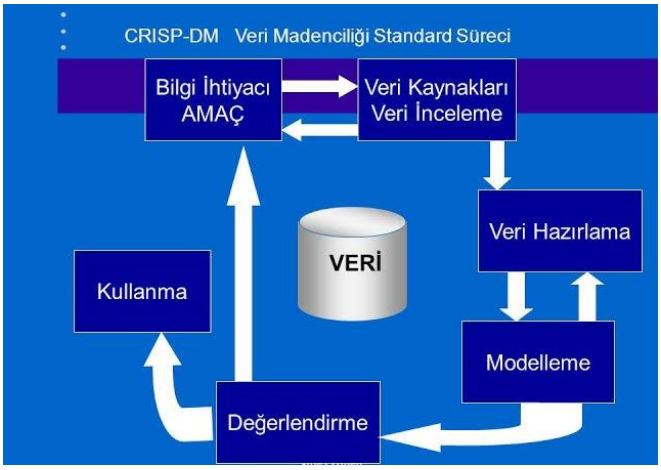
**Kaynak**: (Duygu, 2021).

Kullanım oranlarına bakıldığında en çok kullanım alanı olarak müşteri analitiği göze çarpmaktadır. Müşteri analitiği müşterilerin davranışlarını değerlendirerek ve satın alma niyetini tahmin ederek işletmelerin veri destekli karar almalarını sağlamaya yaramaktadır. Bu sektörleri sırasıyla bankacılık, direk pazarlama ve kredi puanlama alanları takip etmektedir.

## 1.3. VERİ MADENCİLİĞİNİN AŞAMALARI

Sektörel bazda farklı farklı alanlarda iş ve işlemlerini sürdürdükleri için firmaların amaçlarının net olması gerekmektedir. Firmalar amaçları doğrultusunda kullanacakları verileri de farklı kaynaklardan alarak kullanacakları için faydalanacakları kaynağın da net ve amaçlarına yönelik olması gerekmektedir (Duygu, 2021).

Veri madenciliği aşamalarında genel olarak CRISP-DM (Cros-Industry Standart Proces for Data Mining) “Endüstriler Arası Standart İşleme Süreci” anlamına gelen aşamalar uygulanmaktadır. Bu aşamalar işi anlama, verileri anlama, veri hazırlama, modelleme, değerlendirme ve uygulama olarak altı başlıktan oluşmaktadır (Helberg, 2022). Bu aşamalar Şekil 1.1’de gösterilmiştir.




**Şekil 1.1** Veri Madenciliği Standart Süreci (Helberg, 2022).

**A-İşi anlama:** Bu ilk aşama veri madenciliğinin en önemli aşamasıdır. Çünkü problemi veya amacın doğru anlaşılması ve analiz planlamasının yapıldığı aşamadır. Eğer bu aşamada problemin anlaşılmaması gereksiz bir veri analizi ve bu durumda iş gücü ve zaman kaybına neden olacaktır (Helberg, 2022).

Mevcut verilerin yeterliliği, dış kaynak ihtiyacına gerek olup olmadığı ve sonuçların raporlanma durumunun belirlenmesi, çalışmada karşılaşılacak riskler ayrıca hangi teknolojilerin kullanılacağına bu aşamada karar verilerek, maliyet hesaplaması yapılmaktadır (Helberg, 2022).

**B-Verileri anlama:** Bu aşamada mevcut verilerin bir ön incelemeden süzülerek amaca uygunluğu değerlendirilir, verilerin özellikleri belirlenir, farklı türde veri ihtiyaçları tespit edilir ve yeterlilik incelemesi yapılır. Hatalı veriler üzerinde çalışılıp sorunlar giderilmeye çalışılır. Veriler derlenip toparlandıktan sonra amaca uygun olarak oluşturulacak hipotezler veriler üzerinde de kontrol edilir (Helberg, 2022).

**C-Veri hazırlama:** Günümüz dünyasında var olan veriler, çok büyük boyutlara ulaştığından homojen bir dağılım sergileyememekte bu nedenle gürültülü, eksik ve tutarsız verileri de içlerinde barındırmaktadır. Düşük kaliteli veriler beraberinde düşük kaliteli sonuçları getireceğinden veri madenciliğinin kalitesini artırmak için veriler bir takım ön işlemlerden geçirilmektedir (Han ve Kamber, 2001:78). Bu aşamada daha önce toplanan veriler analize hazır hale getirilmektedir. Analizde uygulanacak yönteme göre veriler en baştan tekrar kontrol edilir (Helberg, 2022).

Bu aşama aynı zamanda analizin en çok zaman harcandığı kısımdır. Çünkü bu aşamadan sonra modelleme aşamasında oluşabilecek herhangi bir yanlışlıkta tekrar bu aşamaya geri dönülmektedir. Veri hazırlama aşamaları sırasıyla aşağıdaki başlıklarda incelenebilmektedir (Helberg, 2022).

**1. Veriyi seçme:** Modelin amacına uygun olarak veri seti belirlenir. Burada gereğinden az ya da fazla veri kullanılmamasına dikkat edilmelidir. Az veri kullanımı veri kümesinin kalitesini ve yeterlilik seviyesini etkileyeceği için eksik kalabilir. Fazla veri ise veri kirliliğine sebep olacağı için iş yükünü artıracaktır (Helberg, 2022).

**2. Veriyi temizleme:** Amacımıza uygun olarak oluşturulmak istenen veriler istenilen niteliklere sahip olması gerekmektedir. Bu nitelikleri taşımayan veri setleri üzerinde eksik, tutarsız veya kayıt dışı bulunan verilerin çıkarılması sonucunda hata oranın düzenlenmesi işlemine "veri temizleme işlemi" adı verilmektedir. Bu eksik, tutarsız veya kayıt dışı olan bu verilere "gürültü" denilmektedir. Verilerin gürültülü olma nedenleri (Deloitte, 2014);

- Hatalı veri toplama gereçleri
- Veri giriş problemleri
- Veri girişi sırasında kullanıcıların hatalı yorumları
- Veri iletim hataları
- Teknolojik sınırlamalar
- Veri isimlendirmede veya yapısında uyumsuzluklar olabilir.

Veri setlerindeki gürültü ile başa çıkmanın yolları; eksik gözlemler çıkartılabilir. Kayıp veri oranının toplam veriye oranı 1/3 civarında olduğunda çalışma sonucunu olumsuz yönde etkileyecektir. Witten ve Frank'e göre (2005) bu gözlemleri veri setinden çıkarmak acımasızlıktır. Çünkü kayıp gözlem içeren birimler genellikle iyi bilgi sağlamaktadır. Kayıp değerler için genel sabit bir yeni gözlem değerleri de atanabilir, ancak bu yöntemde analiz için farklılıklar oluşturacağından ve yanıltıcı sonuçlar çıkartacaktır (Silahtaroğlu, 2016). Kayıp verilerin yerine diğer verilerden elde edilen ortalamalar yazılabilir veya veriler regresyon, karar ağacı vs. yöntemlerinden biri ile tahmin edilerek eksik verilerin yerine yazılabilir.

**3. Veriyi kurma:** Yapılan temizleme işlemleri sonrasında kullanıma uygun hale getirilen veri setlerinin güncellenmesi aşamasıdır (Helberg, 2022).

**4. Veriyi birleştirme:** Veri madenciliği çok geniş bir hacime sahip olduğu için toplanan veri setleri arasında uyumsuzluklar oluşmaktadır. Veri temizleme işlemleri yapılıktan sonra veri setleri tek tipe indirilerek analize uygun hale getirilmesi işlemidir (Helberg, 2022).

**5. Veri indirgeme ve biçimlendirme:** Veri indirgemede amaç fazlalık olan verilerden kurtularak amaç doğrultusunda yorumlama zorluklarını ortadan kaldırıp daha etkili bir sonuca ulaşmaktır (Helberg, 2022).

**6. Modelleme:** Toplanan veriler hazırlık aşamasından geçtikten sonra amaca uygun olacak metotlarda en iyi sonuca ulaşana kadar çözümlenmesidir (Helberg, 2022).

Bütün bu aşamalar tamamlandıktan sonra veri değerlendirilmeye hazır hale getirilmiş olur ve değerlendirme aşamasına geçilmektedir.

**C-Değerlendirme:** Bu aşamada kurulmuş olan modelin nihai olarak sunulmasından önce değerlendirilerek mevcut verilerin kalitesi, sonraki süreçte nasıl bir yol izleneceği, sonradan kullanılabilecek verilerin neler olabileceği, amaç doğrultusunda uygun olup olmadığının kontrol edildiği, bütçe üzerinde nasıl bir etki yarattığı ve modelin eksik yönlerinin tespit edilip giderildiği aşamadır (Helberg, 2022).

**D-Uygulama:** Uygulama aşamasında tamamlanan model tekrar gözden geçirilerek kullanım için uygunluğuna bakılır. Model uygulandıktan sonra elde edilen veriler sonucunda ortaya çıkan bilgiler kurum veya kişiler için rapor şeklinde düzenlenir (Helberg, 2022).

## 1.4. VERİ MADENCİLİĞİNDE KULLANILAN YÖNTEMLER

Veri madenciliği modelleri işlevsellikleri bakımından 2 temel başlık altında 4 gruba ayrılmaktadır. Bu gruplandırmaları ise; Sınıflandırma, istatiksel tahmin, kümeleme ve ilişki analizi teknikleri oluşturmaktadır.

Bu ayrımda tanımlayıcı modeller; veri gruplarının içinde gizlenmiş olarak bulunan bilgilere ulaşarak karar verme aşamasında kolaylık sağlamasını sağlarken, tahmin edici modeller ise daha önce oluşturulmuş modellerin, yeni verilere uygulanarak tahmin

edilmesini sağlamaktadır (Sevimli, 2015). Söz konusu yöntemler Tablo 1.3'te gösterilmiştir.

**Tablo 1.3.** Veri Madenciliği Yöntemleri

| **Veri Madenciliği Yöntemleri** | | | |
|---|---|---|---|
| **Tahmin Edici Yöntemler** | | **Tanımlayıcı Yöntemler** | |
| **Sınıflandırma Yöntemleri** | **İstatistiki Tahmin Yöntemleri** | **İlişki Analizi Yöntemleri** | **Kümeleme Analizi Yöntemleri** |
| • Karar ağaçları<br>• Yapay sinir ağları<br>• Genetik algoritmalar<br>• Bellek tabanlı sınıflandırma | • Regresyon analizi<br>• Diskriminant analizi<br>• Lojistik regresyon analizi<br>• Zaman serileri analizi | • Birliktelik kuralları<br>• Ardışık zamanlı örüntüler | • Hiyerarşik kümeleme<br>• Hiyerarşik olmayan kümeleme |

**Kaynak**: (Sevimli, 2015).

### 1.4.1. İlişki Analizi

Silahtaroğlu (2008) ilişki analizini veri setlerinde bulunan değişkenler arasındaki bağlantı durumlarını açıklamak için yapılan işlemler olarak tanımlamaktadır. Kullanım alanı olarak ilişki analizi yöntemleri genel olarak pazarlama, finans, eğitim gibi alanlarda kullanılmaktadır.

**Birliktelik Kuralları:**

Birliktelik kuralları veri setlerinde bulunan değerlerin aynı zamanda meydana getirdikleri olaylardan oluşmaktadır. Örnek olarak müşterinin bir ürün satın alırken onunla birlikte başka bir üründe satın alma eğilimini tespit ederek geleceğe dair müşteri davranışlarını tahmin etmek için kullanılır. Birliktelik kurallarını oluşturan iki ölçüt mevcuttur. Bu kurallardan biri destek ölçütü diğeri ise güven ölçütüdür. Destek ölçütü incelenen olayların ne kadar tekrarlandığını göstermektedir. Destek ölçütü, formül 1.1'de gösterildiği gibi hesaplanmaktadır. (Yaşasınoğlu, 2019)

$$\text{Destek (A}\rightarrow\text{B)} = \frac{Sayı(A,B)}{N} \tag{1.1}$$

Güven ölçütü ise birinci olayı gerçekleştiren değerlerin ikinci bir olayı gerçekleştirme olasılığını göstermektedir. Güven ölçütü, formül 1.2'de gösterildiği gibi hesaplanmaktadır. (Yaşasınoğlu, 2019)

Güven (A—›B) =$\frac{Sayı(A,B)}{Sayı(A)}$ (1.2)

**Ardışık Zamanlı Örüntüler:**

Ardışık zaman örüntüleri birbirleri ile bağlantılı olarak farklı zaman dilimlerinde meydana gelmiş olan olayların birbirleri ile ilişkilerinin belirlenmesi için kullanılan bir yöntemdir. (Şimşek Gürsoy, 2009)

### 1.4.2. Kümeleme Yöntemleri

Kümele yöntemlerinde amaç benzer özelliklere sahip veya birbirlerine mesafe olarak yakın değişkenlerin gruplandırılarak ayrıştırılmasını sağlamaktır. Kümelemenin belirgin olabilmesi için veriler arasındaki benzerlik ve farkın oldukça büyük olması gerekmektedir. Sınıflandırma yöntemleri farklı olarak kümeleme yöntemleri belirgin küme sayısı yoktur. Kümeleme yöntemleri verilerin karakteristik özelliklerine ve uzaklıklarına göre küme sayılarını belirlerken, sınıflandırma da önceden belirlenmiş sınıflar mevcuttur. Kümeleme yöntemleri genel olarak ekonomi, piyasa, tıp, biyoloji ve antropoloji alanlarında yaygın olarak kullanılmaktadır. Verilerin birbirleri ile olan uzaklıkları formül 1.3, 1.4 ve 1.5'te gösterildiği gibi hesaplanmaktadır (Tan, 2005).

$x = (x_1, x_2, \ldots\ldots, x_n)$.  $y = (y_1, y_2, \ldots\ldots, y_n)$

Öklid Uzaklığı

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1.3)$$

Manhattan Uzaklığı

$$d(x, y) = \sum_{i=1}^{n}|x_i - y_i| \quad (1.4)$$

Minkowski Uzaklığı

$$d(x, y) = [\sum_{i=1}^{n}|x_i - y_i|^m]^{1/m} \quad (1.5)$$

**Hiyerarşik Kümeleme Yöntemleri:**

**1. K-En Yakın Komşu Algoritması**

Bu algoritma verilerden oluşan kümelerin arasındaki mesafenin en az olduğu durumlarda kullanılmaktadır. Kümeler arasındaki mesafeyi ise yukarıda 1.3, 1.4 ve 1.5

formüllerinden herhangi biri ile kullanılarak bulunabilir. Bu formüllerden biri ile hesaplanan min d(x, y) değeri birbirine yakın olan iki kümenin mesafesini belirlemektedir. Bütün veriler için bu işlem tekrarlandıktan sonra oluşturulacak olan benzerlik, uzaklık, tablosunda birbirine benzeyen gözlemler tekrar bir küme oluşturarak yeniden bir benzerlik, mesafe ölçüsü belirlenmektedir. Bu durumda algoritma da kullanılacak veri setinde gözlem değerlerinin fazla olması büyük bir zaman kaybına neden olmaktadır (Dunham, 2002).

**2. K-En Uzak Komşu Algoritması**

Bu algoritmada ise oluşturulan kümelerin birbirlerine en uzak olan mesafelerde bulunan değerleri seçilir (Özkan, 2008). K-en yakın komşu algoritmasında olduğu gibi değişkenler arasındaki uzaklıklardan yola çıkarak en yakın olanlar ile ilk küme oluşturulduktan sonra tekrar uzaklıklar hesaplanarak en uzak mesafelerdeki değerler kullanılarak k değerleri oluşturulur ve bu değerler içerisinde en uzak değere sahip k değeri kümeler arasındaki uzaklık olarak kabul edilir. (Kuru, 2019)

Veri setindeki uç değerlere karşı oldukça hassas olan bu algoritma büyük veri setlerini parçalarına ayırdıktan sonra kümeleme yapmaktadır (Guha, Rastogi ve Shim, 1998).

**Hiyerarşik Olmayan Kümeleme Yöntemleri:**

**K-Ortalamalar Yöntemi:**

Veri setindeki bütün değişkenlerin bağımsız değişken olarak varsayıldığı bir tekniktir. Bütün verileri daha önce belirlenen k adet kümeye dahil olmaktadır (Akküçük, 2011).

K-Ortalamalar yöntemi uygulama aşamaları:

- K adet birimin rassal olarak küme merkezi seçilmelidir. Küme merkezi seçimi Formül 1.6 ile hesaplanmaktadır.

  $M_k = \frac{1}{N_k} \sum_{i=1}^{N_k} x_{ik}$ (1.6)

- Kümeler içi değişim için formül 1.7 kullanılır.

  $e_i^2 = \sum_{i=1}^{N_k} (x_{ik} - M_k)^2$ (1.7)

- Kümeler uzayını hesaplamak için ise kümeler içi değişimlerin hata kareleri toplamı alınarak hesaplanır. Formül 1.8 de gösterilmiştir.

$$E_k^2 = \sum_{k=1}^{K} e_k^2 \quad (1.8)$$

Veri setindeki bütün değerler en yakın kümeye atanır ve küme merkezleri tekrar hesaplanır. Bu işlem önceden hesaplanan merkez değerinin aynısı çıkana kadar tekrarlanır. (Sarıman, 2011)

### 1.4.3. Sınıflandırma Yöntemleri

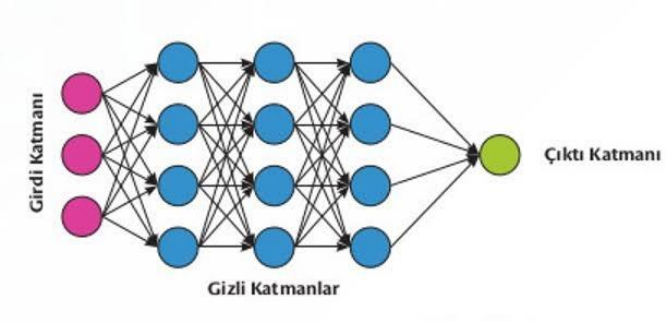
Sınıflandırma yöntemleri verilerin ortak özelliklerine göre hangi sınıfta olacağını tahmin etmek için kullanılan yöntemlerdir. Genel olarak sade yapıda bir model olarak oluşturan sınıflandırma yöntemleri gruplandırılmış veriler hakkında bir yorum yapabilme fırsatı verirken aynı zamanda bağımsız değişkenlerin bilinmesi halinde de bağımlı değişken hakkında öngörü yapabilme fırsatı da vermektedir. Birçok alanda kullanılan sınıflandırma teknikleri veri yığınlarının giderek artması ve gelişen makine ve istatistiki teknikler ile birlikte sınıflandırmanın haricinde bilim insanları tarafından geleceği tahmin etmek için veya karşılaşılabilecek vakalardan en faydalı sonuca ulaşmak için de kullanılmıştır (Ye, 2003).

Sınıflandırma yöntemlerinde mevcut verilerdeki değişkenler dikkate alınarak farklı farklı alanlarda bulunan aynı özelliklerdeki veriler birleştirilmeye çalışılmıştır. Bu yöntem ilk keşfedildiğinde büyük veri haznelerinden mantıklı bir sonuç çıkarmak için kullanılmıştır. Zamanla yapay zekânın da gelişmesi ile araştırmacılar eldeki veriler ile gelecek hakkında öngörüde bulunmaya başlamış ve amaçları doğrultusunda hareket etmek için kullanmışlardır. Fakat öngörüde sınıflandırma, gelecek için tahmin edilen belirli bir davranışa ya da belirli bir değere göre yapılmıştır. Fakat öngörü işleminde yapılan sınıflandırmanın doğru olup olmadığını test etmenin tek yolu “bekle ve gör” prensibine dayanmaktadır. Örnek olarak deprem tahmini ve turizm şirketi müşterilerinden hangilerinin bu yaz yurt dışında tatil yapmak isteyeceğinin belirlenmesi verilebilir. Sınıflandırma modelleri ile regresyon modelleri arasındaki temel farklığı ise bağımlı değişkenin kategorik veya sürekli olması oluşturmaktadır. Çünkü bağımlı değişken sürekli ise analize regresyon modeli uygulanmalıdır (Farboudi, 2009).

**Yapay Sinir Ağları:**

İnsan beyninin yapısından hareketle 1943 yılında ilk çalışma McCulloch ve arkadaşları tarafından yapılmıştır. Yapay sinir ağları kendini yenileyebilen algoritmalardır ve sınıflandırma, optimizasyon, ilişki kurabilme, eksik veri ile çalışabilme ve örnek veriler sayesinde öğrenme becerisi kazanabilme özelliklerine sahiptir. Bu yöntem model tahmin etmede kullanıldığı için net sonuçlar vermemektedir. Diğer modeller ile karşılaştırıldığında ise yapay sinir ağlarında kullanılan veri setlerinde daha fazla değişen yer almaktadır. Bu özelliği sayesinde daha karmaşık verileri daha basit bir şekilde modelleyebilmektedir.

Girdi katmanı, gizli katmanı ve çıkış katmanı olmak üzere birbiri ile bağlantılı 3 katmandan oluşan yapay sinir ağlarının katmanları Şekil 1.2'de gösterilmiştir.




**Şekil 1.2.** Örnek Bir Yapay Sinir Ağı Yapısı
**Kaynak**: Patika Akademi, https://academy.patika.dev/blogs/detail/yapay-sinir-aglariys

Girdi katmanında elde bulunan veriler ilgili ağırlıkları ile çarpıldıktan sonra sinir ağları ile eşlenerek sisteme dahil edilir. Gizli katman girdi ile çıktı arasında ilişki olmadığında aralarındaki ilişkiyi açıklamada kullanılmaktadır. Bazen bir bazen de birden fazla katmandan oluşabilmektedir. Bu katmanda işlenen veriler çıktı katmanına gönderilir (Çelik, 2009).

**Genetik Algoritmalar:**

John Holland ve arkadaşları zor ve karmaşık olan problemlerde kolay bir çözüme ulaşmak için 1960'lı yıllarda Darwin'in evrim teorisinden esinlenerek arama ve optimizasyon tekniği olan bu algoritmayı geliştirmişlerdir. Bu algoritmada amaç canlı varlıklardaki özellikleri bilgisayar ortamında taklit ederek model oluşturmaktır. Birçok uygulama alanında yer bulan bu algoritma otomatik programlama, ekonomi, ekoloji,

üretim hattı yerleşimi ve öğrenme kabiliyetli makineler alanlarında sıkça kullanılmaktadır (Duygu, 2021).

**Bellek Tabanlı Sınıflandırma:**

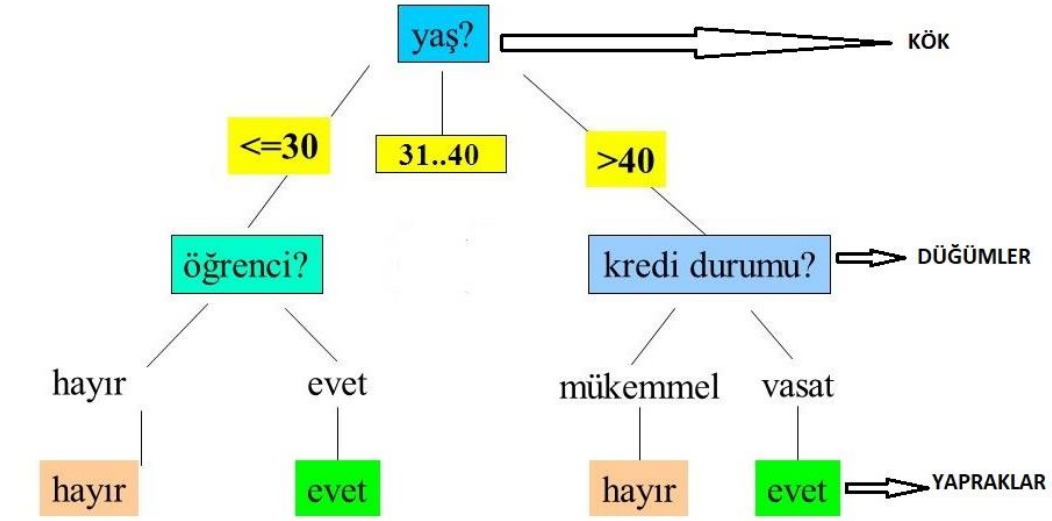
Bellek tabanlı sınıflandırma teknolojinin gelişimi üzerine bilgisayar tabanlı algoritmalarında gelişmesi üzerine önem kazanmış bir madencilik yöntemidir. Sistem performans yönetimi, işgücü kullanım stratejisi, kredi skorlama sistemleri ve performans tahminlime vb. işlemlerde kullanılmaktadır. (Duygu, 2021).

## 1.5. VERİ MADENCİLİĞİNDE KARAR AĞAÇLARI

Karar Ağaçları modelleri parametrik olmayan, fazlaca hesap gerektirmeyen ayrıca sınıflama ve regresyon yöntemlerinde kullanılabilen bir tekniktir. Karar ağaçları hangi sınıfa ait olduğu bilinen bir veriden tümevarım uygulamasıyla öğrenilen ağaç modelinde bir karar yapısıdır (Sun J. And Li H., 2008). Bu yöntem karmaşık ve büyük verilere karar kuralları uygulayarak bölme işlemi gerçekleştirilip benzer verilerin bir araya getirilerek gruplandırılmasıdır. Karar ağaçlarında genel amaç genelleme hatasını minimize ederek en uygun karar ağacını elde etmektir (Ercan, 2017). Ağaç yapısına benzerliğinden dolayı kolay anlaşılan bir yapıya sahiptir. Tahmin edici ve tanımlayıcı yöntemler için de kullanılabilmesinden dolayı en çok tercih edilen algoritmalar arasındadır.

Karar ağacının yapısı kök ve düğümlerden başlayarak özelliklerine göre alt dallara dağılım göstermektedir. Dağılım sonucunda en uygun çözüme kadar bu işlem sürdürülüp analiz tamamlanmaktadır (Ercan, 2017).

Şekil 1.3 de örnek bir karar ağacı üzerinde kök düğüm dallar ve yaprakları gösterilmektedir




**Şekil 1.3.** Örnek Karar Ağacı (H., F. Dalkılıç, 2015)

Yukarıdaki örnek karar ağacından da anlaşılacağı üzere veri setindeki tüm gözlemleri kapsayan bir değişken kök değişken olarak seçilir. Alt gruplara ayrılarak boğumlar (düğüm) vasıtasıyla aşağıya doğru inilir. Bu ara düğümler "internal node/non-terminal node" olarak adlandırılmaktadır (Hssina, vd., 2014). Aşağıya doğru inilirken her düğümde test edilir ve ağacın veri kaybetmesi önlenir. Sonuç olarak hiçbir bölünmenin gerçekleşmediği veya bölünmelerin herhangi bir iyileşmeye yol açmadığı adımda durur ve yapraklar ile ağaç formasyonu son bulur (Berry, 2004).

Karar ağaçlarında bulunan her ara düğümün kök düğüme olan uzaklığına düzey, mertebe veya derinlik denilmektedir (Bayhan, 2021). Kök düğümün derinliği ise bire eşittir (Sevimli, 2015). Derinlik veya düzey sayılarındaki artış ağacın karmaşıklığını da artırmaktadır (Dahan vd., 2014).

### 1.5.1. Karar Ağaçlarının Avantaj ve Dezavantajları

Karar ağaçlarının avantajları aşağıdaki gibi sıralanabilir;

1. Kolay analiz edilmesi,
2. Parametrik istatistiki yöntemlerin zorunlu şartlarını (normal dağılım, varyansların homojenliği vb.) sağlamasına gerek yoktur (Bayhan 2021).
3. Maliyetlerinin düşük olması,
4. Güvenilirlik seviyesinin yüksek olması,

5. Veri tabanı sistemleri ile kolayca entegre edilmesi,
6. Güvenilirliklerinin iyi olması,
7. Kolay yorumlanabilmesi.

gibi etmenlerinden dolayı yaygın bir kullanıma sahiptir. Öğrenilebilmesi açısından da basit bir yapıda olması uygulanabilirlik açısından avantaj sağlamaktadır. Kullanım alanları göz önüne alındığında çok geniş bir alana sahiptir. Özellikle büyük veri yığınlarının sınıflandırılmasında ve hatalı verilerin bulunduğu durumlarda çözüm sağlamaktadır (Aitkenhead, 2008).

İstatistiki yöntemlerde veri setinin türüne göre kullanılacak yöntem değişkenlik gösterirken karar ağaçlarında kullanılan bazı yöntemlerde (CART, Chaid, Mars vb.) değişkenin türü kullanılacak yöntemi değiştirmemektedir (Sevimli Saitoğlu, 2015).

Değişkenlerin aşırı farklılık içerdiği veri setlerinde bölünme de fazla olacağından dolayı yapraklar da (uç düğüm) fazla olacaktır. Bu farklılıktan kaynaklı olarak uç düğümlerde bulunan değişken sayısı da az olacağından sınıflandırmanın güvenilirliği de daha az olabilmektedir (Gülpınar, 2008).

Karar ağaçlarında kullanılan veri setlerindeki değişim karşısında ağaç yapısı tümden değişim göstereceği için öngörü yapabilmek diğer yöntemlere göre daha zayıftır (James, vd., 2013).

### 1.5.2. Karar Ağaçlarının Türleri

#### 1. ID3 Algoritması:

J. Ross Quinlan tarafından 1970'lerin sonunda geliştirilmiş olan algoritma daha kısa hesaplar ile büyüme ve budama olmak üzere sade bir ağaç oluşturmayı amaçlamaktadır. Veri setindeki mevcut verilerin birbiriyle olan farklılıklarından yararlanarak çalışan bir algoritmadır. Bu farklılıkların ölçümüne entropi ölçüsü (belirsizlik ölçüsü) denilmektedir.

Bu algoritmada karar ağaçları, özellik ve amaca göre düğüm ve yapraklardan oluşan bir sınıflandırma algoritmasıdır. Bhardwaj ve Vatta (2013) yılında yaptığı sürekli değişkenler üzerine çalışmasında kabul edilebilir bir sonuç vermediğini saptamıştır. Yani anlamlı bir sonuca ulaşmak için bağımlı ve bağımsız değerlerin kategorik değişkenlerden oluşması gerekmektedir (Sevimli, 2015). Sınıflandırma temelli işlem yürüttüğü için bu

algoritmanın zayıf yönlerinden biridir. Sınıflamadaki amacı ise en ayırıcı özelliğe sahip değişkeni bulmaya çalışmaktadır. Bu değişkeni bulurken ise entropi değerinden faydalanmaktadır. Ayrıca kayıp gözlem ve gürültü içeren veri setlerinde performansının zayıf olması, dallanma da kullanılan entropi kriterinin doğru bir sınıflandırma yapamayacağı ve değişkenlerdeki kategorilerin artması entropi değerini düşürdüğü için güvenilir karar verilememektedir. Büyük ve ayrıntılı verilerin olduğu çalışmalarda ise doğru sonuç veremeyeceği için ön işleme işleminin daha detaylı yapılması gerekmektedir. Görülen bu eksiklikler üzerine ID3 algoritması bilim çevreleri tarafından tam olarak benimsenemediği görülmüştür. Fakat ID3 algoritmasının eksikliklerinin düzeltilmesi ve geliştirilmesi amacıyla birçok algoritmanın da temeli atılmıştır (Sevimli, 2015).

**2. C4.5 Algoritması:**

J. Ross Quinlan tarafından ID3 algoritmasındaki eksikleri gidererek daha doğru bir sınıflandırma yapmak için geliştirilmiştir. Bu algoritmada sürekli değişkenler ön işlemlerden geçerek bölünme için uygun hale getirilip kullanıldığı, ayrıca kayıp gözlem değerlerini de öngörerek analizde kullanabilmektedir. Bu açıdan ID3 algoritmasına göre daha anlamlı ve duyarlıdır.

Ayrıca bu algoritma entropi değerini hesaplarken kayıp değerleri dışarıda bırakarak işlem yapmakta olup veri setinin homojenliği konusunda avantaj sağlamaktadır (Sing ve Gupta, 2014).

**3. Quest Algoritması:**

İkili bir karar ağacı yapısı kullanan bir algoritmadır (W. Y. Loh and Y. S. Shih, 1997). Tek değişkenli ve doğrusal kombinasyonlu bölünmeleri desteklemekte olup, bölünme aşamalarında değişkenler arasındaki ilişki ve değişken türüne bağlı olarak Varyans analizi, Levene testi veya Ki-Kare testi ile hesaplamalar yapılmaktadır. Oluşumunda değişken seçimi ve bölünmeyi yapamadığı için ayrı ayrı ele almaktadır. Maliyeti düşürdüğü için ve yanlı seçimleri genel hale getirdiği için geliştirilmiştir (Coşkun ve Bülbül, 2019).

### 1.5.3. Karar Ağaçlarının Oluşum Süreçleri

Karar ağaçları oluşum süreçleri olarak temelde iki adımda meydana gelmektedir. Birinci adım öğrenme basamağıdır. Bu adımda daha önceden bilinen eğitim verisinin (örnek veri) model oluşturmak için sınıflama algoritmaları tarafından çözümlenmesidir. İkinci adım ise karar ağacının veya sınıflama algoritmasının doğruluğunun test edildiği sınıflamadır. Diğer bir ifade ile karar ağaçları ağacın oluşturulması ve budama aşamalarından oluşmaktadır. Ağacın oluşturulması için ilk işlem olarak en iyi bölünmeyi sağlayacak girdi değişkenine karar vermektir. Bu ilk bölünme ağacın formasyonunu çok etkilediği için en önemli aşamasıdır. Kök düğümden ayrılan ara düğümler, veri sınıflandırmasını daha homojen gruplara ayırarak devam eder. Bölünmenin gerçekleşmediği durumda ise sona erer (Güner, 2014).

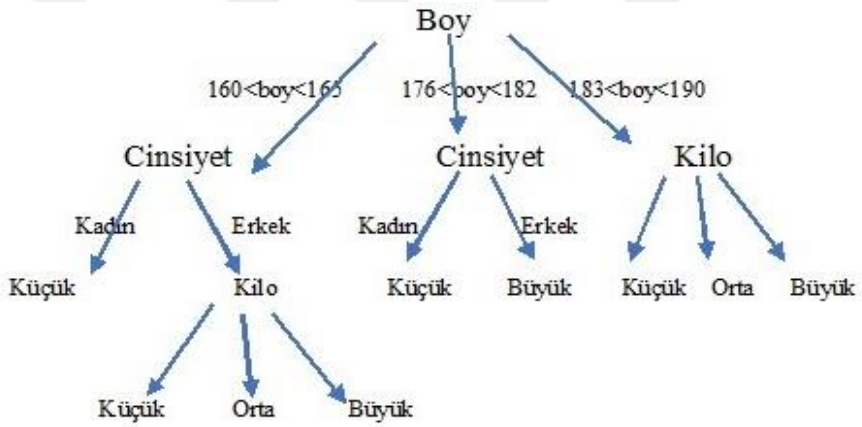
Bir örnek üzerinden karar ağacı algoritması ilkelerini inceleyelim.

**Tablo 1.4.** Örnek Veri Seti

| **Cinsiyet** | **Kilo** | **Boy** | **Beden** |
|---|---|---|---|
| K | 47 | 170 | Orta |
| E | 48 | 150 | Küçük |
| K | 52 | 158 | Orta |
| K | 56 | 164 | Orta |
| E | 58 | 160 | Küçük |
| E | 60 | 159 | Orta |
| K | 61 | 162 | Küçük |
| K | 62 | 174 | Orta |
| E | 67 | 168 | Orta |
| K | 68 | 177 | Büyük |
| K | 71 | 170 | Orta |
| E | 73 | 165 | Küçük |
| E | 84 | 175 | Orta |
| E | 85 | 190 | Büyük |
| E | 97 | 180 | Büyük |

**Kaynak**: (Silahtaroğlu, 2020)

Bu örnek veri setimizde beden değişkeni bağımlı değişken olarak düşünelim. Diğer değişkenlerimiz (boy, kilo, cinsiyet) ise düğümlerimizi temsil edecektir. Karar ağacına başlarken yukarıda da dediğimiz gibi en önemli nokta kök düğümü tespit etmektir. Çünkü kök düğüme göre ağaç formasyonu değişmektedir. Bu aşamada kendimize soracağımız soru şu olmalı; öyle bir değişken seçmeliyim ki verilen yanıt ne olursa olsun, diğer değişkenlerle kıyaslandığında elimizdeki veri tabanını kabaca iki eşit parçaya bölsün (Silahtaroğlu, 2020). Seçilen kök düğümüne bağlı olarak ağaç türlerinde fark olması muhtemeldir. Amacımız anlaşılabilir ve doğru olan sınıflandırmayı oluşturmaktır. Bazen veri setimizdeki değişken sayısının fazla olması ağaç formasyonların da benzerlik gösterebilir (Silahtaroğlu, 2020). Şekil 1.4'te kök düğüm olarak boy değişken seçildiğinde oluşacak karar ağacı formasyonu gösterilmektedir.

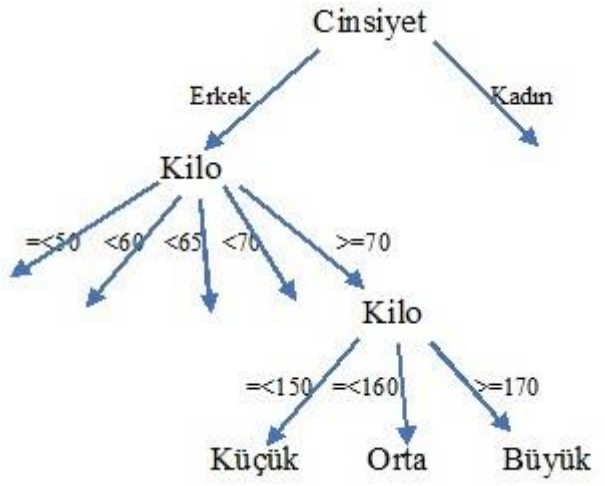



**Şekil 1.4.** Karar Ağacı Formasyonu (Silahtaroğlu, 2020)

Şekil 1.4'te görüldüğü üzere boy kök düğümü ile başlanıldığı için cinsiyet ve kilo soruları birden fazla ara düğüme sorulmaktadır (Silahtaroğlu, 2020).

Ağaç formasyonu oluşturulurken dallanma kriterleri bu aşamada çok önemlidir. Cinsiyet düğümü sadece iki yöne ayrılırken boy düğümünden ayrılacak olan dalların sayısının kaç olacağı araştırmacıyı düşündürmektedir. Örnek olarak bütün boy uzunlukları için dallanma mı olacak? Yoksa aralıklara mı bölünecek? Aralılara bölündüğü durumlarda kadın ve erkek için bu farklı olacaktır (Silahtaroğlu, 2020).

Fakat kök düğüm için cinsiyet değişkeninden başlanırsa boy ve kilo soruları her cinsiyet için sadece bir defa sorularak dallara ayrılacaktır (Silahtaroğlu, 2020).




**Şekil 1.5.** Karar Ağacı Formasyonu (Silahtaroğlu, 2020)

Şekil 1.5'te görüldüğü üzere kök düğümün farklılaşması sonucunda ağaç formasyonu değişmektedir. Daha anlaşılır ve basit bir formasyon oluşturmak için kök düğüm daha genel bir ayrıma tabi olabilecek bir değişken olmalıdır (Silahtaroğlu, 2020).

### 1.5.4. Karar Ağaçlarının Budanması

Karar ağaçları homojen alt gruplara ayrılmak için fazlaca bölünme gerçekleştirmektedir. Veri sayısının az olduğu durumlarda bu bölünmelere bağlı olarak aşırı uyum ve aşırı öğrenme problemleri karşımıza çıkmaktadır (Kretowski, 2019). Aşırı uyum ve aşırı öğrenmenin olması yeni eklenecek değişkenlerin ağaç yapısı içinde kötü bir performans sergilemesine yol açacaktır (Strobl, vd., 2009).

Bu durumda karar ağaçlarının güvenilir ve gerçekçi olabilmesi için ağaç oluşturulurken ya da ağaç oluşturulduktan sonra yapılması gereken işlemlerden bir tanesi de budama işlemidir. Budama işlemi kısaca ağaçta oluşmuş sonucu etkilemeyen ve sınıflandırmaya herhangi bir katkısı olmayan dalların ağaçtan alınmasıdır. Yani gereksiz ayrıntıların sonuçtan çıkarılması işlemidir (Silahtaroğlu, 2020). Veri sayısı ve değişken sayılarının fazla olması ağaç formasyonu üzerinde birçok düğüm ve dallanmaya neden olacaktır. Bu da yapraklarda oluşan veri sayısında azalmaya neden olacaktır. Ayrıca karar ağacının hassasiyetini de azaltacaktır (Cabena, 1998).

Budama işlemleri iki yerde yapılmaktadır:

1- Ağaç oluşum aşamasında iken daha çok düğüm ve dallanma oluşumunu engellemek için yapılmaktadır. Bu duruma ön budama (pre-pruning) işlemi denilmektedir.
2- Ağaç yapısına hiç dokunulmadan süreç tamamlanır ve sonrasında budama işlemi yapılmaya başlanır ise bu durma son budama (post-pruning) denilmektedir.

Oluşturulan karar ağacının minimum derinlikte maksimum bilgiyi verdiğini test etmek için model doğrulama ölçütünün olması gerekmekte olup, bu model kök, düğüm ve dallanma kriterlerindeki seçimlerde izledikleri yol bakımından birbirlerinden ayrılmaktadırlar. Karar ağaçlarında kullanılan veriler genellikle non-parametrik olduğu için karar ağaçlarının doğruluğu ve sınıflandırma gücünü ölçmek için ölçütler geliştirilmiştir (Strobl, ve diğ., 2009; De'ath ve Fabricius, 2009). CART algoritması için oluşturulan ölçütler Tablo 1.5'te gösterilmektedir.

**Tablo 1.5.** Model Doğrulama Ölçütleri

| Model Doğrulama Ölçütü | Sınıflandırma Ağaçları | Regresyon Ağaçları |
|---|---|---|
| En Küçük Maliyet-Karmaşıklık Ölçütü | ✓ | |
| En Küçük Hata-Karmaşıklık Ölçütü | | ✓ |
| Test Örneği Ölçütü | ✓ | ✓ |
| Çapraz Geçerlilik Ölçütü | ✓ | ✓ |

**Kaynak**: Sevimli Saitoğlu, Y. (2015).

Tablo 1.5 incelendiğinde CART algoritmasında kullanılabilecek bazı model doğrulama ölçütlerinin sınıflandırma ağaçları için, bazılarının regresyon ve bazılarının ise hem sınıflandırma hem de regresyon ağaçları için olduğu görülmektedir. Bu doğrulama ölçütlerinde ilk aşamada maksimum ağaç yapısına ulaşılır ve akabinde budama işlemleri uygun model türüne göre uç düğümlerden başlayacak şekilde budama işlemi başlatılır. Optimum ağaca ulaştığında ağaç formasyonu sonlandırılır. Optimum budama seviyesine ulaşıldığını anlayabilmek için ise test örneği ölçütü veya çapraz geçerlilik ölçütü kullanılmaktadır. (Bayhan, 2021).

## 1.6. CART ALGORİTMASI

CART algoritması; bilimsel araştırmalarda kullanılan istatistiki analiz ve sınıflama tekniklerinin içerisindeki varsayımların fazla olmasından dolayı ve bu varsayımların analiz imkânlarını kısıtlamasından dolayı geliştirilen alternatif bir yöntemdir (Temel, Çamdeviren, 2005). Yani parametrik olmayan bir istatistiksel metot olup, Breiman ve arkadaşları tarafında 1984 yılında geliştirilmiştir. Genel bir tanım yapılacak olursa CART algoritması; bölünme kurallarına fazlaca değişken dâhil ederek iki uç düğüme bölünmeyi sağlayıp tekrarlanabilir bir şekilde sınıflandırma ve regresyon yöntemidir (Denison, vd., 1998).

Bu yöntem ile incelenecek bir çalışmada, bağımlı değişkenin kategorik olması durumunda sınıflama ağaçları, bağımlı değişkenin sürekli olması durumunda ise bu yöntem regresyon ağaçları olarak isimlendirilir. Temel amaç olarak CART algoritması homojen alt gruplar oluşturmak ve güçlü tahminler yapabilmektir (Deconinck, Hancock, 2005).

CART algoritması oluşum kısımlarına bakıldığında ise üç kısımdan oluşmaktadır. Bunlar max ağacın oluşturulması, uygun ağaç genişliğinin seçilmesi ve oluşturulan ağaçtan hareketle yeni verilerin sınıflandırılmasıdır. Önceki bölümlerde bahsedildiği üzere karar ağaçlarında budama işlemi ön budama ve son budama olmak üzere iki türde yapılmaktadır. Bu durumda CART algoritmasında ilk aşama da ağaç max olgunluk seviyesine çıkarılır ve sonra budama işlemi gerçekleştirilir. Bu sayede aşırı uyum sorunundan kaçınılmış olmaktadır (Timofeev, 2004).

### 1.6.1. CART Algoritmasının Avantaj ve Dezavantajları

CART algoritmasının avantaj ve dezavantajları aşağıdaki gibi sıralanabilir;

- CART algoritması diğer istatistik analizlerinden farklı olarak bağımlı değişken ile bağımsız değişkenler arasındaki ilişki yapısını incelemekle yetinmeyip bağımsız değişkenlerin birbirleri ile olan ilişkilerini de inceleyen bir algoritmadır.

- Yapılacak analizin konusuna göre ağaç yapısını şekillendirdiğinden ayrıca kategorik ve sürekli değişkenler için uygulanabilir olmasından dolayı herhangi bir veri setine rahatlıkla uygulanabilmektedir.
- Bir diğer avantajı ise ağaç formasyonuna benzerliğinden basit anlaşılabilir bir yapıda olup kapladığı alan olarak da az yer kaplamaktadır.

Basit anlaşılabilir bir yapıda olması bağımsız değişkenlerin kademeli olarak ele alınmasından kaynaklı olarak ağacın karmaşıklık yapısı azalmaktadır. Yani her düğümde bölünmeler önem sırasına, bölümlerine en önemli verisine göre yapılmaktadır. (Bayhan, 2021).

- Standart bir veri setine sahip değişkenler üzerinde yapılacak değişmeler karşısında bölünme kuralı bu değişmeden etkilenmemektedir. Yani veri setindeki değerin karesi alınarak işlem yapıldığı zaman model sonucunu etkilememektedir (Yohannes ve Webb, 1999).
- Algoritmanın belki de en büyük avantajlarından biri, büyük veri kümelerini ve bunun sonucunda ortaya çıkan yüksek boyutluluk problemlerini kolayca ele almasıdır. Sonuç olarak, analize dahil edilen birçok değişken arasından birkaç önemli değişken kullanılarak faydalı sonuçlar üretilebilir. (Yohannes ve Webb, 1999).
- CART algoritması parametrik olmayan bir yöntem olduğu için parametrik yöntemlerin getirdiği varsayımlar olmadan da analiz yapma imkanı sunmaktadır. Ancak bu durumun avantaj mı dezavantaj mı olduğu konusunda birçok tartışma söz konusudur (Shao ve Lunetta, 2012).

Çünkü bu durumda örneklemin küçük olması ve popülasyon hakkında ön bilgi olmaması durumunda parametrik olmayan yöntemler kullanılmasında avantaj sağladığı, ancak parametrik yöntemler ile benzer koşullar altında kullanıldığında ise dayanıksız ve zayıf bir analiz çıktısı vermesinden dolayı dezavantajlı olduğu gözlemlenmiştir (Karagöz, 2010). Bu açıdan bakıldığında, parametrik varsayımları sağlayan örnek bir veri setine sahip olunması durumunda, CART algoritmasının kullanılması önerilmemektedir (Yohannes ve Webb, 1999).

Çalışılan örneklemdeki veri sayısının az olması durumunda analizden çıkarılan veya analize dahil edilen birimlerin bazı durumlarda sonuçları büyük oranda etkilediği

sonucuna varılmıştır. Bu durumda CART algoritmasının kullanılması, özellikle küçük bir örnek setinin olduğu durumlarda, sonuçlar taraflı olduğu için önerilmez (Yohannes ve Webb, 1999).

- CART algoritması bir tanımlayıcı metot olduğu için güven aralığı ve anlamlılık değerlerini elde etmek zordur. Bu yüzden dolayı yapılan işlemlerin diğer istatistiki yöntemler ile karşılaştırılması önerilmektedir.
- Ayrıca CART algoritmalarında yalnızca bağımsız değişkenlerin halihazırdaki değerlerine göre tahminler yapmakta olup, böylece bağımsız değişkenin anlamlılık değerleri ve güven aralıkları elde etmek zorlaşmaktadır. Bu sebeple CART algoritması ile yapılan hesaplamaların diğer istatistiki metotlarla kıyaslanması önerilmiştir (Speybroeck, 2012).

## 1.7. CART ALGORİTMASINDA SINIFLANDIRMA AĞAÇLARI

CART algoritması daha önce de ifade edildiği üzere ikili ağaç formasyonundan oluşan bir yapıya sahiptir. Bu ayrımı ise bağımsız değişken safsızlığı yani değişim ölçülerindeki değişkenlik kullanılarak hesaplanan oran ile seçilmektedir. Değişken ölçütlerinde gini, twoing, en küçük kareler sapması gibi birçok dallanma ölçütü mevcuttur. Yalnız CART algoritması için kullanılan dallanma ölçütleri gini ve twoing ölçütleridir (Bayhan, 2021).

### 1.7.1. Gini Dallanma Ölçütü

Gini dallanma ölçütü kök düğümden başlanmak üzere her düğümde en iyi bölünmeyi aramaktadır. Bu bölünme öncelikle kök düğüm için yapılarak devam eden düğümlerde de tekrarlanarak uç düğümlere kadar devam ettirilmektedir. CART algoritmasında gini dallanma ölçütüne göre sınıflandırma yapılabilmesi için ilk aşamada nitelik değerleri ikili olacak şekilde sağ ve sol olarak gruplandırılmalıdır. İkinci aşamada ise nitelikleri ile ilgili olarak gini sol ve gini sağ değerleri hesaplanmalıdır. Gini sol ve sağ değerlerinin hesaplanması formül 1.9 ve 1.10'da gösterilmektedir (Kuyucu, 2012).

$$\text{Gini}_{\text{sol}} = 1 - \sum_{i=1}^{k} \left(\frac{L_i}{T_{sol}}\right)^2 \tag{1.9}$$

$$Gini_{sağ} = 1-\sum_{i=1}^{k}\left(\frac{R_i}{T_{sağ}}\right)^2 \quad (1.10)$$

k = Sınıf Sayısı,

T = Bir Düğümdeki Örneklem,

$T_{sol}$ = Sol Düğümdeki Örneklerin Sayısı,

$T_{sağ}$ = Sağ Düğümdeki Örneklerin Sayısı,

$L_i$ = Sol Düğümde i Kategorisindeki Örneklerin Sayısı,

$R_i$ = Sağ Düğümde i Kategorisindeki Örneklerin Sayısını ifade etmektedir.

Üçüncü aşama olarak gini sol ve sağ değerleri hesaplandıktan sonra gini indeks değeri hesaplanmaktadır.

Gini indeks değerinin hesaplanması formül 1.11'de gösterilmektedir.

$$Gini_j = \frac{1}{n}(T_{sol} \text{ x } Gini_{sol} + T_{sağ} \text{ x } Gini_{sağ}) \quad (1.11)$$

n = Veri Setindeki Satır Sayısı

Herbir değişken için bu işlemler tekrarlandıktan sonra gini indeks değeri düşük olan değişkenden dallara ayrılır. Her düğümde bu işlem tekrarlanarak ağaç formasyonu devam ettirilir. Kesin bir sınıflandırma yapıldığı zaman süreç durdurulur (Kuyucu, 2012).

### 1.7.2. Twoing Dallanma Ölçütü

Twoing dallanma ölçütünde ilk aşamada aday bölünme tablosu oluşturularak değişkenlere iki sınıf atanmaktadır. Yapılacak bu işlemde amaç, veri setini her aşamada ikiye bölerek ağacı devam ettirmektir. Aday bölünme tablosunu açıklayacak olursak karar ağaçlarında dallanmanın iki taraflı olduğu ifade etmiştik ama her zaman verilerimizin nitelikleri iki parçadan oluşmamaktadır. Bu aşamada verilerimizin niteliklerine göre sol ve sağ olmak üzere iki parçaya bölerek aday bölünme tablosunu oluşturulmaktadır (Lowe ve Kulkarni, 2015).

İkinci aşamada oluşturulan bu aday bölünme tablosunun her satırı için $P_L$, $P(C_j/t_L)$ ve $P_R$, $P(C_j/t_R)$ değerleri hesaplanır.

$P_L$ ve $P_R$ = aday bölünme satırındaki niteliğin olasılık dağılımını ifade etmektedir.

$P(C_j/t_{L)}$ ve $P(C_j/t_R)$ ise hedef sınıfımız için sınıflandırılan değişkenlerin sol ve sağ sınıf için olasılık dağılımını ifade etmektedir.

$C_j$ = hedef sınıf değeri,

$t_L$, $t_R$ = hedef sınıftaki toplam değeri,

t = dallanmanın yapılacağı düğümü ifade etmektedir.

Üçüncü aşamada ise her satır için uygunluk ölçütü hesaplanır. Uygunluk ölçütü formül 1.12'de gösterilmiştir.

$$\Psi(s/t) = 2P_L P_R \sum_{J=1}^{M} \left| P(C_J/t_L) - P(C_J/t_R) \right| \quad (1.12)$$

Bu işlem tüm aday bölünme satırları için uygulanmaktadır.

Dördüncü aşamada ise uygunluk ölçütü en büyük olan satır seçilerek kökten sonra dallanacak değişken olarak seçilir. Her düğümde bu işlem tekrarlanarak ağaç formasyonu devam ettirilir. Kesin bir sınıflandırma yapıldığı zaman süreç durdurulur (Lowe ve Kulkarni, 2015).

### 1.7.3. CART Algoritmasının Teorik Uygulaması

Araştırmacı tarafından rastgele oluşturulan Tablo 1.6'daki veriler kullanılarak gini ve twoing safsızlık ölçütleri kullanılarak CART algoritması karar ağacı oluşturulacaktır.

**Tablo 1.6.** Örnek Veri Seti

| No | Eğitim | Yaş | Cinsiyet | Kabul |
|---|---|---|---|---|
| **1** | Orta | Yaşlı | Erkek | Evet |
| **2** | İlk | Genç | Erkek | Hayır |
| **3** | Yüksek | Orta | Kadın | Hayır |
| **4** | Orta | Orta | Erkek | Evet |
| **5** | İlk | Orta | Erkek | Evet |
| **6** | Yüksek | Yaşlı | Kadın | Evet |
| **7** | İlk | Genç | Kadın | Hayır |

**Gini Dallanma Ölçütü Teorik Uygulaması**

**1. Adım:** İkili gruplandırma tablosu oluşturulur.

**Tablo 1.7.** İkili Gruplandırma Tablosu

| **Kabul** | **Eğitim** | | **Yaş** | | **Cinsiyet** | |
|---|---|---|---|---|---|---|
| | **İlk** | **Orta, Yüksek** | **Genç** | **Orta, Yaşlı** | **Kadın** | **Erkek** |
| **Evet** | 1 | 3 | 0 | 4 | 1 | 3 |
| **Hayır** | 2 | 1 | 2 | 1 | 2 | 1 |
| **Toplam** | 3 | 4 | 2 | 5 | 3 | 4 |

**2. Adım:** Her değişken için (1.9) ve (1.10)'daki formüller kullanılarak $Gini_{sol}$ ve $Gini_{sağ}$ değerleri hesaplanır.

Eğitim değişkeni için;

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

$= 1-[(İLK_{EVET}/İLK)^2+[(İLK_{HAYIR}/İLK)^2] = 1-[(1/3)^2+(2/3)^2] = 0.44$

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

$=1-[(ORTA, YÜKSEK_{EVET}/ORTA,YÜKSEK)^2+(ORTA,YÜKSEK_{HAYIR}/ORTA,YÜKSEK)^2]$

$=1-[(3/4)+(1/4)] = 0.38$

Yaş değişkeni için;

$Gini_{sol} = 1-[(GENÇ_{EVET}/GENÇ)^2+[(GENÇ_{HAYIR}/GENÇ)^2] = 1-[(0/2)^2+(2/2)^2] = 0$

$Gini_{sağ} = 1-[(ORTA,YAŞLI_{EVET}/ORTA,YAŞLI)^2+(ORTA,YAŞLI_{HAYIR}/ORTA,YAŞLI)^2]$

$= 1-[(4/5)^2+(1/5)^2=0.32$

Cinsiyet değişkeni için;

$Gini_{sol} = 1-[(KADIN_{EVET}/KADIN)^2+[(KADIN_{HAYIR}/KADIN)^2] = 1-[(1/3)^2+(2/3)^2] = 0.44$

$Gini_{sağ} = 1-[(ERKEK_{EVET}/ERKEK)^2+(ERKEK_{HAYIR}/ERKEK)^2] = 1-[(3/4)^2+(1/4)^2] = 0.38$

**3. Adım:** Gini indeks değeri hesaplanmalıdır. Her değişken için (1.11)'deki formül kullanılarak indeks değeri hesaplanmaktadır.

$Gini_j = \frac{1}{n}(T_{sol} \text{ x } Gini_{sol} + T_{sağ} \text{ x } Gini_{sağ})$

Eğitim değişkeni için;

$Gini_{eğitim} = 1/7[(3*0.44)+(4*0.38)] = 0.41$

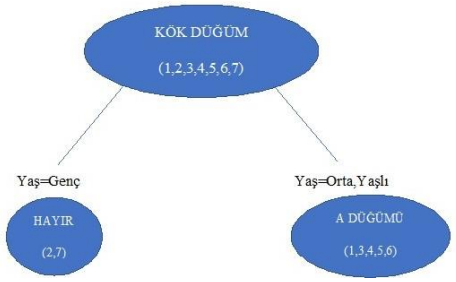
Yaş değişkeni için;

$Gini_{yaş} = 1/7[(2*0)+(5*0.32)] = 0.23$

Cinsiyet değişkeni için;

$Gini_{cinsiyet} = 1/7[(3*0.44)+(4*0.38)] = 0.41$

**4. Adım:** En küçük gini indeks değeri seçilir ve kök düğümden başlanılarak bu değişken niteliklerinde ikiye ayrılarak ağaç formasyonu oluşmaya başlar.




**Şekil 1.6.** Gini Karar Ağacı Formasyonu 1. Düğüm

Şekil 1.6'da yaş değişkeni en küçük indeks değerine sahip olduğu için sınıflandırılmıştır. Yaş değişkeninin genç niteliklerine bakıldığı zaman (2 ve 7) tümü için hedef değişkenleri yani kabul durumları hayır olduğu için kesin bir sınıflama meydana geldiğinden ağaç formasyonu orta,yaşlı sınıflandırmasından devam ettirilecektir. Devam ettirilirken $Yaş_{genç}$ satırları çıkarılarak yukarıda belirtilen adımlar tekrar yapılacaktır.

**Tablo 1.8.** 2. ve 7. Satırların Çıkarılması İle Oluşan Örnek Veri Seti

| No | Eğitim | Yaş | Cinsiyet | Kabul |
|---|---|---|---|---|
| **1** | Orta | Yaşlı | Erkek | Evet |
| **3** | Yüksek | Orta | Kadın | Hayır |
| **4** | Orta | Orta | Erkek | Evet |
| **5** | İlk | Orta | Erkek | Evet |
| **6** | Yüksek | Yaşlı | Kadın | Evet |

Yeni şeklini alan veri setimiz için tekrar yukarıda bahsedilen adımlar gerçekleştirilerek ikili gruplandırma tablosu ve $gini_{sol}$ ve $gini_{sağ}$ değerleri hesaplanmaktadır.

**Tablo 1.9.** 2. ve 7. Satırların Çıkarılması Meydana Gelen İkili Gruplandırma Tablosu

| **Kabul** | **Eğitim** | | **Yaş** | | **Cinsiyet** | |
|---|---|---|---|---|---|---|
| | **İlk** | **Orta, Yüksek** | **Genç** | **Orta, Yaşlı** | **Kadın** | **Erkek** |
| EVET | 1 | 3 | 2 | 2 | 1 | 3 |
| HAYIR | 0 | 1 | 1 | 0 | 1 | 0 |
| TOPLAM | 1 | 4 | 3 | 2 | 2 | 3 |

Tekrar bütün değişkenlerimiz için $gini_{sol}$ ve $gini_{sağ}$ değerleri,

Eğitim değişkeni için;

$Gini_{sol} = 0$ $Gini_{sağ} = 0.38$

Yaş değişkeni içim;

$Gini_{sol} = 0.45$ $Gini_{sağ} = 0$

Cinsiyet değişkeni içim;

$Gini_{sol} = 0.5$ $Gini_{sağ} = 0$

Bütün değişkenlerimiz için indeks değerleri,

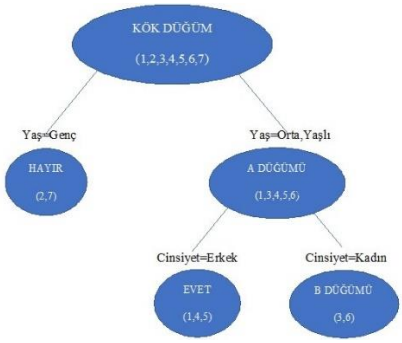
$Gini_{eğitim} = 0.30$

$Gini_{yaş} = 0.27$

$Gini_{cinsiyet} = 0.20$

olarak hesaplanmaktadır.

Yeni hesaplanan gini indeks değerlerine bakıldığında en küçük indeks değerine sahip olan cinsiyet değişkeninin niteliklerine göre dallanmanın gerçekleşeceğini görmekteyiz.




**Şekil 1.7.** Karar Ağacı Formasyon 2. Düğüm

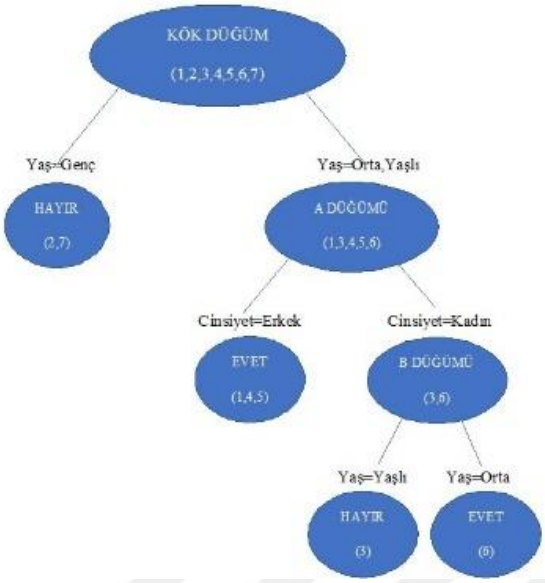
Cinsiyet değişkeninin erkek niteliklerine bakıldığı zaman (1,4,5) tümü için hedef değişkenleri yani kabul durumları evet olduğu için kesin bir sınıflama meydana geldiğinden ağaç formasyonu kadın sınıflandırmasından devam ettirilecektir. Devam ettirilirken $\text{Cinsiyet}_{\text{erkek}}$ satırları çıkarılarak yukarıda belirtilen adımlar tekrar yapılacaktır.

**Tablo 1.10.** 1.4. ve 5. Satırların Çıkarılması Meydana Gelen Örnek Veri Seti

| No | Eğitim | Yaş | Cinsiyet | Kabul |
|---|---|---|---|---|
| **3** | Yüksek | Orta | Kadın | Hayır |
| **6** | Yüksek | Yaşlı | Kadın | Evet |

Yeni şeklini alan veri setimiz için tekrar yukarıda bahsedilen adımlar gerçekleştirilerek ikili gruplandırma tablosu ve $\text{gini}_{\text{sol}}$ ve $\text{gini}_{\text{sağ}}$ değerleri hesaplanmaktadır.

Tekrar hesaplamalar yapıldıktan sonra yaş değişkeni en küçük indeks değerine sahip olmaktadır. Yaş değişkeninin de kabul değişkeninde evet ve hayır diye ikiye ayrılmasından dolayı kesin bir sınıflandırma oluşturmasından ağaç formasyonu burada durmaktadır. Ağaç formasyonun son hali Şekil 1.8'de gösterilmektedir.




**Şekil 1.8.** Karar Ağacı Formasyon Son Hali

**Twoing Dallanma Ölçütü Teorik Uygulaması**

Araştırmacı tarafından rastgele oluşturulan Twoing dallanma ölçütü için tekrar Tablo 1.6'daki örnek veri seti kullanılarak karar ağacı oluşturulacaktır.

1. **Adım:** Aday bölünme tablosu oluşturulur.

**Tablo 1.11.** Aday Bölünme Tablosu

| Aday Bölünme | $t_L$ | $t_R$ |
|---|---|---|
| **1** | $Eğitim_{ilk}$ | $Eğitim_{Orta,Yüksek}$ |
| **2** | $Eğitim_{Orta}$ | $Eğitim_{İlk,Yüksek}$ |
| **3** | $Eğitim_{Yüksek}$ | $Eğitim_{İlk,Orta}$ |
| **4** | $Yaş_{Genç}$ | $Yaş_{Orta,Yaşlı}$ |
| **5** | $Yaş_{Orta}$ | $Yaş_{Genç,Yaşlı}$ |
| **6** | $Yaş_{Yaşlı}$ | $Yaş_{Genç,Orta}$ |
| **7** | $Cinsiyet_{Erkek}$ | $Cinsiyet_{Kadın}$ |
| **8** | $Cinsiyet_{Kadın}$ | $Cinsiyet_{Erkek}$ |

2. **Adım:** Oluşturulan aday bölünme tablosunun her satırı için $P_L$, $P(C_j/t_L)$ ve $P_R$, $P(C_j/t_R)$ değerleri hesaplanır. Yukarıda bahsedildiği üzere $P_L$ 1. Aday bölünme satırındaki niteliğin tüm değişkenlerimize oranını ifade etmektedir. Yani;

$P_L=3/7=0,42$ $P_R=4/7=0,57$

$P(C_j/t_L)$ ise hedef sınıfımız için sınıflandırılan değişkenlerin sol ve sağ sınıf için olasılık dağılımını ifade etmektedir. Yani her değişken hedef sınıfımızda iki niteliğe sahip ya evet yada hayır, j 'de bizim bu değerlere karşılık gelen sayısal ifadelerdir.

$t_L = 3$ $P_{Evet}/T_L = 1/3 = 0,33$ $P_{Hayır}/T_L = 2/3 = 0,66$

Aynı işlemler sağ taraf içinde tekrarlanır.

$t_R = 4$ $P_{Evet}/T_R = 3/4 = 0,75$ $P_{Hayır}/T_R = 1/4 = 0,25$

Bu işlemler bütün aday bölünme satırları için uygulanıp $t_L$ ve $t_R$ tabloları oluşturulacaktır.

**Tablo 1.12.** tL Tablosu

| Aday Bölünme | $t_L$ Kat Sayısı | $P_L$ | Evet Sayısı | Hayır Sayısı | $P_{Evet}/T_L$ | $P_{Hayır}/T_L$ |
|---|---|---|---|---|---|---|
| **1** | 3 | 0,42 | 1 | 2 | 0,33 | 0,66 |
| **2** | 2 | 0,28 | 2 | 0 | 1 | 0 |
| **3** | 2 | 0,28 | 1 | 1 | 0,5 | 0,5 |
| **4** | 2 | 0,28 | 0 | 2 | 0 | 1 |
| **5** | 3 | 0,42 | 2 | 1 | 0,66 | 0,33 |
| **6** | 2 | 0,28 | 2 | 0 | 1 | 0 |
| **7** | 4 | 0,57 | 3 | 1 | 0,75 | 0,25 |
| **8** | 3 | 0,42 | 1 | 2 | 0,33 | 0,66 |

**Tablo 1.13.** tR Tablosu

| Aday Bölünme | $t_R$ Kat Sayısı | $P_R$ | Evet Sayısı | Hayır Sayısı | $P_{Evet}/T_R$ | $P_{Hayır}/T_R$ |
|---|---|---|---|---|---|---|
| **1** | 4 | 0,57 | 3 | 1 | 0,75 | 0,25 |
| **2** | 5 | 0,71 | 2 | 3 | 0,4 | 0,6 |
| **3** | 5 | 0,71 | 3 | 2 | 0,6 | 0,4 |
| **4** | 5 | 0,71 | 4 | 1 | 0,8 | 0,2 |
| **5** | 4 | 0,57 | 2 | 2 | 0,5 | 0,5 |
| **6** | 5 | 0,71 | 2 | 3 | 0,4 | 0,6 |
| **7** | 3 | 0,42 | 1 | 2 | 0,33 | 0,66 |
| **8** | 4 | 0,57 | 3 | 1 | 0,75 | 0,25 |

**3. Adım:** Tablo 1.11 ve 1.12' de hesaplanan değerler ile her aday bölünme satır için Formül 1.12 ile uygunluk ölçütü hesaplanır.

$$\Psi(s/t) = 2P_L P_R \sum_{J=1}^{M} \left| P(C_J/t_L) - P(C_J/t_R) \right|$$

1. aday bölünme satırı için;

$\Psi(1/t) = 2*0,42*0,57*[\,|\,(0,33-0,75)\,|+|\,(0,66-0,25)\,|\,]= 0,39$

2. aday bölünme satırı için;

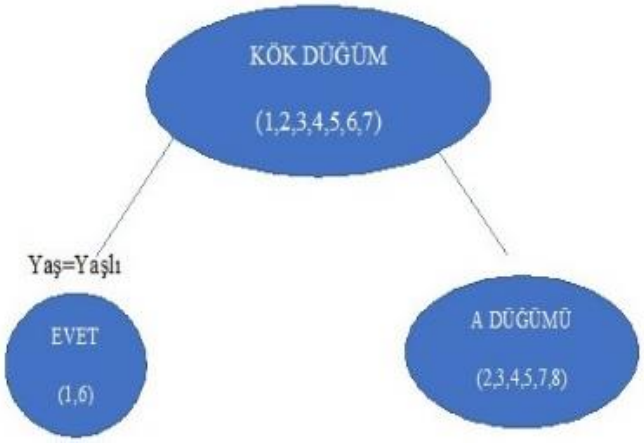
$\Psi(2/t) = 2*0,28*0,71*[\,|\,(1-0,4)\,|+|\,(0-0,6)\,|\,]= 0,47$

Bütün aday bölünme satırları için yapılan hesaplamaları Tablo 1.14'te gösterilmiştir.

**Tablo 1.14.** Uygunluk Ölçütü Tablosu

| **Aday Bölünme** | **Ψ(s/t)** |
|---|---|
| **1** | 0,39 |
| **2** | 0,47 |
| **3** | 0,07 |
| **4** | 0,63 |
| **5** | 0,15 |
| **6** | 0,83 |
| **7** | 0,39 |
| **8** | 0,39 |

4. **Adım:** uygunluk ölçütü en büyük olan aday bölünme satırı kökten sonra dallanacak olan nitelikleri oluşturur.




**Şekil 1.9.** Twoing Karar Ağacı Formasyonu 1. Düğüm

Şekil 1.9'da yaş değişkeni en büyük uygunluk ölçütüne sahip olduğu için sınıflandırılmıştır. Yaş değişkeninin yaşlı niteliklerine bakıldığı zaman (1 ve 6) tümü için

hedef değişkenleri yani kabul durumları evet olduğu için kesin bir sınıflama meydana geldiğinden ağaç formasyonu genç, orta sınıflandırmasından devam ettirilecektir. Devam ettirilirken $Yaş_{yaşlı}$ satırları çıkarılarak yukarıda belirtilen adımlar tekrar yapılacaktır.

Yukarıda belirtilen adımlar tekrarlandığı zaman oluşan aday bölünme tablosu ve uygunluk ölçüsü tablosu gösterilmiştir.

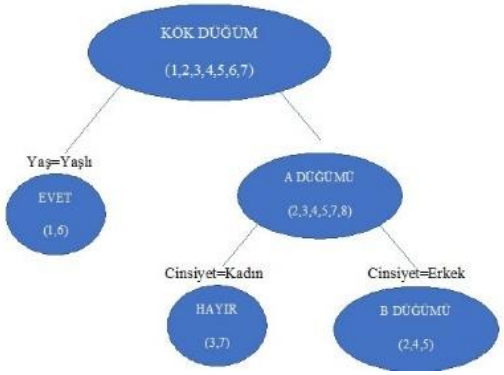
**Tablo 1.15.** Aday Bölünme Tablosu

| **Aday Bölünme** | $\mathbf{t_L}$ | $\mathbf{t_R}$ |
|---|---|---|
| **1** | $Eğitim_{ilk}$ | $Eğitim_{Orta,Yüksek}$ |
| **2** | $Eğitim_{Orta}$ | $Eğitim_{İlk,Yüksek}$ |
| **3** | $Eğitim_{Yüksek}$ | $Eğitim_{İlk,Orta}$ |
| **4** | $Yaş_{Genç}$ | $Yaş_{Orta}$ |
| **5** | $Yaş_{Orta}$ | $Yaş_{Genç}$ |
| **7** | $Cinsiyet_{Erkek}$ | $Cinsiyet_{Kadın}$ |
| **8** | $Cinsiyet_{Kadın}$ | $Cinsiyet_{Erkek}$ |

**Tablo 1.16.** Uygunluk Ölçütü Tablosu

| **Aday Bölünme** | **Ψ(s/t)** |
|---|---|
| **1** | 0,15 |
| **2** | 0,48 |
| **3** | 0,32 |
| **4** | 0,63 |
| **5** | 0,63 |
| **7** | 0,63 |
| **8** | 0,63 |

Tablo 1.16'da görüldüğü gibi 4, 5, 7, 8 aday bölünme satırlarının uygunluk ölçüleri eşit çıkmaktadır. Bu gibi durumlarda herhangi biri seçilerek sınıflandırma işlemi yapılıp değişkenin niteliklerine göre dallanma gerçekleştirilir.




**Şekil 1.10**. Twoing Karar Ağacı Formasyonu 2. Düğüm

Şekil 1.10'da cinsiyet değişkeni tercih dilmiş olup sınıflandırılmıştır. Cinsiyet değişkeninin kadın niteliklerine bakıldığı zaman (3 ve 7) tümü için hedef değişkenleri yani kabul durumları hayır olduğu için kesin bir sınıflama meydana geldiğinden ağaç formasyonu erkek sınıflandırmasından devam ettirilecektir. Devam ettirilirken $Cinsiyet_{kadın}$ satırları çıkarılarak yukarıda belirtilen adımlar tekrar yapılacaktır.

Yukarıda belirtilen adımlar tekrarlandığı zaman yeniden oluşan aday bölünme tablosu ve uygunluk ölçüsü tablosu gösterilmiştir.

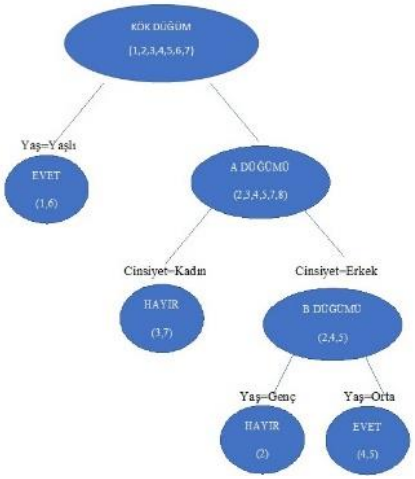
**Tablo 1.17.** Aday Bölünme Tablosu

| **Aday Bölünme** | $\mathbf{t_L}$ | $\mathbf{t_R}$ |
|---|---|---|
| **2** | $Eğitim_{İlk}$ | $Eğitim_{Orta}$ |
| **4** | $Yaş_{Genç}$ | $Yaş_{Orta}$ |
| **5** | $Cinsiyet_{Erkek}$ | |

**Tablo 1.18.** Uygunluk Ölçütü Tablosu

| **Aday Bölünme** | **Ψ(s/t)** |
|---|---|
| **2** | 0,43 |
| **4** | 0,87 |
| **5** | 0 |

Uygunluk ölçütü en büyük olan aday bölünme satırının değişken niteliklerine göre dallarına ayrılır.




**Şekil 1.11.** Twoing Karar Ağacı Formasyonu Son Hali

Uygunluk ölçüsü en büyük olan aday bölünme satırına karşılık gelen değişkenin niteliklerine bakıldığı zaman kabul değişkeninde evet ve hayır diye ikiye ayrılmasından dolayı kesin bir sınıflandırma oluşturulmuş ve ağaç formasyonu burada durmaktadır. Ağaç formasyonun son hali şekil 1.11'de gösterilmektedir.

### 1.7.4. Gini ve Twoing Katsayılarının Karşılaştırılması

Karar ağacı algoritmalarında kullanılacak olan dallanma ölçütü ağacın performansını etkilediği için büyük bir öneme sahiptir. Dallanma ölçütü seçilirken sınıflandırma hatasının en az olduğu ölçütün seçilmesi gerekmektedir.

Literatürde yapılan çalışmalarda iki ölçütünde hem benzer hem de farklı yönlerinin olduğu ifade edilmiştir.

Roman (2004) çalışmasında iki ölçütünde benzer sonuçlar verdiğini ve farklılıkların kök düğüme yakın düğümlerde gözlenmediğini uç değerlerde gözlendiğini ifade etmiştir. Değişken sayılarının ve bağımlı değişkenin niteliklerindeki azlık ve çokluk iki ölçüt arasında temel farklılığı oluşturmaktadır. Değişken sayılarının az olması durumunda iki ölçüt arasında benzer sonuçların gözlendiği, aksi durumda farklılıklar meydana gelmektedir.

Literatür taramasında da anlaşıldığı üzere yapılan bazı karşılaştırmalarda bağımlı değişkenin sınıf değerlerinin az olması durumunda gini dallanma ölçütünün daha iyi sonuçlar verdiğini, fakat sınıf ve değişken sayılarının artması durumunda twoing

dallanma ölçütünün daha başarılı olduğu ifade edilmiştir. Bunun temel sebebini ise Reddy ve Vasu (2013) twoing dallanma ölçütünde verilerin %50'sini ikiye ayırarak kullanmasının etkili olduğunu ileri sürmektedir.

Yapılan literatür araştırmasından ve yukarıda örnek üzerinden de görüldüğü üzere değişken sayısının ve bağımlı değişkendeki nitelik sayısının az olması durumunda benzer sonuçların meydana geldiği fakat değişkenlerin artması durumunda ise twoing dallanma ölçütünün kullanılmasının daha doğru olacağı yapılan literatür araştırmasından anlaşılmaktadır.

# İKİNCİ BÖLÜM

# LOJİSTİK REGRESYON ANALİZİ

## 2.1. LOJİSTİK REGRESYON ANALİZİ

Lojistik regresyon analizi (1944) yılında Berkson tarafından ileri sürülmüş bir analiz metodudur. Bu metotta kullanılacak verilerin ölçek düzeyleri (nominal, ordinal, aralık ve oran) bilinmesi gerekmektedir. Çünkü belirlenen ölçek durumuna göre farklı testler uygulanmaktadır ve analize uygun olmayan değerleri ayırt etmemize yaramaktadır. Lojistik regresyon yöntemlerinde kategorik bağımlı değişkenler ile kategorik/sayısal bağımsız değişkenlerin aralarındaki ilişkiyi açıklamaya yardımcı olacak şekilde en uygun modeli kurmak amaçlanmaktadır. Kullanılan verilerin kategorik veya kesikli değerlerden oluşması lojistik regresyonu diğer regresyon yöntemlerine kıyas ile daha çok kullanılmasına sebep olmaktadır. Bu doğrultuda regresyon modeli oluşturulurken çok değişkenli normallik, süreklilik ve eşvaryanslılık gibi varsayımları da dikkate almaksızın analiz yapılabilmektedir (Bayram, 2015).

Yukarıda belirtildiği üzere lojistik regresyon analizinin diğer regresyon analizlerinde kullanılan genel varsayımları dikkate almadan analiz yapabilmektedir. Fakat lojistik regresyonun kendi özelinde belli birtakım varsayımları bulunmaktadır. Bu varsayımlar:

1. Veri setinde eksik veya aykırı değerler varsa bu değerler tespit edilerek gerekli düzenleme işlemi yapılarak analiz sonucunu etkilemesi önlenir.
2. Kategorik değişkenlerde verilerin frekans grupları 1-5 arasında olmalıdır.
3. Lojistik regresyon analizi bağımsız değişkenler ile sonuç değişkeninin logit değeri arasında doğrusal ilişki olduğunu varsaymaktadır (Taşkın, 2022).
4. Gözlem değerleri arasında çoklu doğrusal bağlantı olmamalıdır.

Lojistik regresyon analizinin çok fazla tercih edilmesini maddeler halinde sıralayabiliriz.

- Bağımsız değişkenin kesikli/sayısal olması analiz için bir engel oluşturmamaktadır.
- Analiz sonucunda parametrelerin yorumlanması oldukça rahattır.

- Lojistik regresyon analizini yapabilmek için birçok istatistiki uygulama mevcuttur. (Spss, Sas, R, Stata vb.)
- Bağımsız değişkenlerin olasılık fonksiyonlarının dağılımıyla alakalı bir sınırlaması yoktur (Çatal, 2021).
- Lojistik regresyon analizindeki bütün olasılıklar sıfır (0) ile bir (1) arasında değerler almaktadır.
- Bağımlı ve bağımsız değişkenler arasında illaki doğrusal ilişki olacak diye bir şart aranmamaktadır. Logaritmik dönüştürmeler yaparak ilişkinin şeklini doğrusal hale getirip işlemlerine devam eder.

Lojistik regresyonun kendine ait varsayımlarının var olmasına rağmen doğrusal regresyon ile arasında farklarda bulunmaktadır. Bu farkları gösterir tablo 2.1 de gösterilmektedir.

**Tablo 2.1.** Doğrusal Regresyon Modeli ile Lojistik Regresyon Analizi Arasındaki Fark Tablosu

| | **Doğrusal Regresyon Modeli** | **Lojistik Regresyon Analizi** |
|---|---|---|
| Bağımlı Değişkenin Sürekli Olması | ✓ | X |
| Bağımlı Değişkenin Kesikli Olması | X | ✓ |
| Bağımlı Değişken Değerinin Tahmin Edilmesi | ✓ | X |
| Bağımlı Değişkenin Alabileceği Değerlerden Birinin Gerçekleşme Olasılığının Tahmin Edilmesi | X | ✓ |
| Bağımsız Değişkenin Çoklu Normal Dağılım Göstermesi Şartı | ✓ | X |

**Kaynak**: Selin Taşkın, Y. (2022).

Lojistik regresyon analizi oluşturulduktan sonra model ki-kare test istatistiği ile anlamlılığı test edilmektedir. Model bütün olarak anlamlılığının yanı sıra parametre katsayılarının da anlamlılığı Wald istatistiği ile test edilmektedir (Kalaycı, 2014)

### 2.1.1. Lojistik Regresyon Analizinin Kullanıldığı Alanlar

Genellikle bağımlı değişkenin iki kategoriden, neden/sonuç ilişiklerinin oluştuğu analizlerde kullanılmaktadır. Ayrıca doğrusal regresyon analizlerinde bağımlı değişkenin

sürekli (nicel) verilerden oluşması ve normallik varsayımına dayandırılması, lojistik regresyon analizlerinde de bu varsayımların gerekli olmaması analizlerde kolaylık sağlamaktadır.

Zamanla gelişen istatistik paket programları da lojistik regresyon analizlerinin uygulanabilirlik alanlarını etkilemiş ve bununla beraber yorumlanmasının kolay ve anlaşılabilir bir şekilde özellikle tıp alanında ve diğer disiplinlerde (sosyal bilimler, fen bilimleri, veterinerlik, ekonomi, taşımacılık, pazarlama, meteoroloji, askeri konular vb.) çok geniş alanlara yayılmasına neden olmuştur.

### 2.1.2. Lojistik Regresyonda Kullanılan Bazı Kavramlar

**Lojistik Fonksiyonu:**

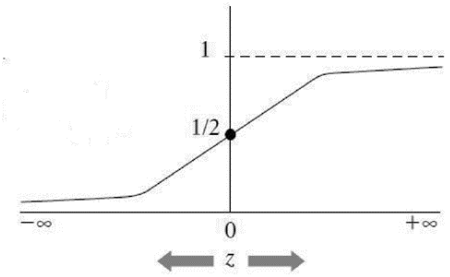
Lojistik fonksiyonu; lojistik regresyon analizinde olasılığın nasıl olduğunu modellemek için kullanılmaktadır. Herhangi bir z değerine bağlı olarak f(z) biçiminde tanımlandığında, f(z) fonksiyonu lojistik modele bağlı olan lojistik fonksiyonu ifade etmektedir. Lojistik fonksiyon,

$$f(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(b_0.b_1x_1+\cdots\ldots+b_k.x_k+a)}} \quad (2.1)$$

$$z = \beta_0 + \beta_1X_1 + \beta_2X_2 + \ldots\ldots + \beta_kX_k$$

denklem 2.1'deki gibi hesaplanmaktadır (Taşkın,2022).

Grafiksel gösterimi Şekil 2.1'deki gibidir.




**Şekil 2.1.** Lojistik Fonksiyon Grafiği (D. G. Klienbaum ve M. Klein, 2002)

Grafik incelendiği zaman z değeri -∞ ile +∞ arasında değerler almaktadır. z değeri -∞'a giderken f(z) fonksiyonu sıfır (0) değerini +∞' sonsuza giderken ise bir (1) değerini

almaktadır. f(z) fonksiyonunun bu durumu lojistik modelin yaygın olarak kullanılmasını kolaylaştırmaktadır (Kleinbaum, 2022).

**Odds Oranı:**

Odds oranı bir olayın gerçekleşme ve gerçekleşmeme oranı üzerinde duran bir olasılık yöntemidir. Yani odds oranı bağımsız değişkenin bağımlı değişken üzerindeki etkisini ifade etmektedir. İkili veya ikiden fazla kategorili değişkenlerin olması durumuna göre farklı olarak yorumlanmaktadır. İkili bağımlı değişken olması durumunda birbirlerine göre olasılıksal olarak kıyaslamalarını, ikiden fazla bağımlı değişken olması durumda ise grupların ikili olarak tekrarlanan çözümlerinin birbirlerinden daha yüksek olasılıkla olma durumlarını ortaya koymaktadır. Odds oranı yorum açısından $R^2$ değerine benzemektedir. Odds oranı arttıkça değişkenin açıklama oranı da artmaktadır (Özdemir, 2011).

$$O = \frac{P}{1-P} = \frac{Bir\ Olayın\ Gerçekleşme\ Olasılığı}{Bir\ Olayın\ Gerçekleşmeme\ Olasılığı} \tag{2.2}$$

Denklem 2.2'deki gibi hesaplanmaktadır. Odds oranı olasılık tabanlı olduğundan dolayı alt sınırı sıfır olmakta fakat üst sınır için bir limit bulunmamaktadır (Allison, 2001)

**Logit model:**

Logit model odds oranının logaritmasının alınması ile oluşturulan bir modeldir ve denklem 2.3'teki gibi formüle edilmektedir.

$$g(x) = \ln\left(\frac{p(x)}{1-p(x)}\right) = \ln e^{(\beta_0+\beta_1 x)} = \beta_0+\beta_1 x \tag{2.3}$$

Oluşturulan bu model diğer olasılık fonksiyonları gibi sıfır ile bir arasında değer almayıp -∞ ile +∞ arasında değer almaktadır. Ayrıca modelin bağımlı değişkeni ile bağımsız değişkeni arasında doğrusal bir ilişkiye karşın hesaplanan olasılıklar ile bağımlı değişkenler arasında doğrusal bir ilişki bulunmamaktadır (Albayrak, 2006)

Logit dönüşümünün önemli bir özelliği ise diğer regresyon modellerinde istenen birçok özelliği barındırmasıdır. Parametreler bakımından logit doğrusal ve süreklidir. Diğer bazı özellikleri ise şöyle sıralanabilir;

- p(x) arttıkça g(x) de artar.

- p(x) sıfır ile bir arasında bir değer alırken g(x) tüm gerçel doğru üzerindeki değerleri alır.
- Eğer p(x)<0,5 ise g(x)<0, p(x)>0,5 ise g(x)>0 dır (Orhan, 2021).

## 2.2. LOJİSTİK REGRESYON ANALİZİNİN TAHMİN YÖNTEMLERİ VE ÖNEM TESTİ

Lojistik regresyon analizinin değerlendirilmesi üç aşamadan oluşmaktadır. Bunlar:

1- Modelin genel yeterliliğinin değerlendirilmesi,
2- Bağımsız değişkenlerin anlamlı olup olmadıklarının değerlendirilmesi,
3- Modelin lojistik regresyon varsayımlarını karşılayıp karşılamadığına

bakılır.

Lojistik regresyon analizinin tahmin yöntemlerinde, en çok olabilirlik yöntemi, en küçük kareler yöntemi ve ağırlıklandırılmış en küçük kareler yöntemi yaygın bir kullanıma sahiptir.

### 2.2.1. Lojistik Regresyon Analizi İçin Değişken Seçimi

Lojistik regresyon analizlerinde, verilerin net bir modele ulaşabilmesi, açık ve karmaşık olmaması için belirli bir değişken seçim stratejisinden geçmesi gerekmektedir. Burada temel amaç modeli olumsuz yönde etkileyen bağımsız değişkenlerin elenmesidir. Değişken sayılarının azlık veya çokluk durumlarına göre değişkenler arasında etkileşim ve açıklayıcı değişkenlerin sayısı da etkileneceği için değişken sayılarının az olması durumlarında bütün modellerin oluşturulması tavsiye edilmektedir. Bu modeller seçilirken Akaike Bilgi Kriteri, Bayesci Bilgi Kriteri ve Doğru Sınıflama Oranları Kriterleri içerisinde uyumu en yüksek model seçilmelidir (Yasinoğlu, 2022).

Değişken sayılarının fazla olması durumlarında ise açıklayıcı değişkenlerin sayıları da fazla olacağından adımsal yöntemler ve enter yöntemi tercih edilir. Adımsal yöntemler ise kendi içerisinde üç grupta incelendiği bilinmektedir. Bu yöntemler;

1- İleriye doğru seçim yöntemi,
2- Geriye doğru seçim yöntemi,
3- Adımsal Seçim Yöntemidir.

**İleriye Doğru Seçim Yöntemi:**

Bu yöntem değişkenlerin modele eklenmesi mi gerekli veya çıkarılması mı gerekli sorusuna yanıt aramaktadır. Denkleme ilk girecek olan değişken ise gruplar arasındaki ayrımı en iyi sağlayan değişken olmaktadır. Yalnızca sabit terimin bulunduğu bağımsız değişkenlerin bulunmadığı modeller ile işleme başlamaktadır. Modele ilave edildiğinde bağımsız değişkenler içerisinden logaritmik olarak en fazla katkıyı sağlayan değişken modele ilave edilir ve ilave edilen değişkenin modele olan katkısı önemsiz dereceye kadar düşene dek modele bağımsız değişkenler ilave edilmeye devam edilir (Alpar, 2017).

**Geriye Doğru Seçim Yöntemi:**

Bu yöntem başlangıçta tüm potansiyel değişkenlerin modele dahil edilmesi ile başlamaktadır. Daha sonra sırasıyla modele anlam katmayan değişkenler istatistiksel kriterler yardımıyla (Akaike Bilgi kriteri, Bayes Bilgi kriteri gibi) modele en az katkı sağlayan değişken modelden çıkarılır ve model tekrar değerlendirilir. En sonda modelin tahmin gücünü tam olarak yansıtan değişken modelde yerini alır ve yöntem tamamlanmış olur (Alpar, 2017).

**Adımsal Seçim Yöntemi:**

Bu yöntem hem ileriye doğru seçim hem de geriye doğru seçim yöntemlerinin içerisinde barındıran ortak bir yöntemdir. Modeldeki bağımsız değişkenler adım adım istatistiksel kriterler yardımıyla (Akaike Bilgi kriteri, Bayes Bilgi kriteri gibi) eklenir veya çıkarılır. Yani başlangıç aşamasında Wald ile β katsayılarının anlamlılıklarına göre modele dahil edilen değişken daha sonra geriye doğru seçim yöntemi ile elenerek modelden çıkarılabilir (Alpar, 2017).

**Enter Yöntemi:**

Bu yöntemde adımsal seçim yöntemi gibi içerisinde hem ileriye doğru seçim yöntemini hem de geriye doğru seçim yöntemini barındırmaktadır. Bu yöntemin adımsal yöntemden farkı ise burada modelde hiçbir değişken yokken başlar ve her adımda değişken eklenir veya çıkarılır. Yöntem eklenecek veya çıkarılacak değişken kalmayana kadar devam etmektedir. Modele değişken eklendikçe istatistiksel değerlendirmelerle modelin iyileştirilmesine odaklanır. Bu yöntemde dikkat edilmesi gereken husus ise

gereksiz değişkenlerin olması durumunda modelin karmaşıklığı artabilmektedir (Hatip Ünlü, 2022).

### 2.2.2. Lojistik Regresyon Analizinde Aykırı Değer Tespiti ve Normalizasyon Teknikleri

Veri setinde aykırı gözlemlerin bulunması parametre tahminlerini ve yorumlamalarının yanlış olmasına neden olmaktadır.

Yavuzkanat (2011) bu durumu; aykırı gözlemler, bağımlı değişkenin 0-1 değerlerini alma olasılıklarını modellediğinden, gerçekte başarılı sonuç alan gözlemin bu değeri alma olasılığının küçük çıkmasına ya da gerçekte başarısız sonuç alan gözlemin bu değeri alma olasılığının yüksek çıkmasına sebep olabilirler. Böylece lojistik regresyonda aykırı gözlemler modelden elde edilen başarı olasılığı küçük olduğu halde gözlemlenen durumu başarılı olan ya da modelden elde edilen başarı olasılığı yüksek olduğu halde gözlemlenen durumu başarısız olan gözlemler olarak ifade etmiştir. Bu aykırılığı tespit etmek ve ortadan kaldırmak için bazı teknikler aşağıda ifade edilmektedir Yavuzkanat (2011).

**Min-Max Normalizasyonu:**

Verilerin belirli bir aralığa dönüştürülmesi için kullanılan bir ölçekleme yöntemidir. Bu yöntem çıktıları doğrusal olarak dönüştürmeyi amaçlayarak değişkenleri 0 ile 1 veya -1 ile 1 arasında ölçeklendirir. Bu sayede, verilerin farklı ölçeklerde olmasından kaynaklanan etkilerin dengelemesi sağlanır (Adeyemo, 2018).

$$\text{Doğrusal Dönüşüm} = \frac{(X - Min(x))}{(Max(x) - Min(x))} \tag{2.1}$$

Formül 2.1'de X gözlemlenen değerler kümesindeki her bir veriyi, Min ve max değerleri ise gözlemlenen değerler kümesindeki en küçük ve en büyük değerleri ifade etmektedir.

**Z Skoru Normalizasyonu:**

Bütün girdi değişkenlerini standart sapması bir ve ortalaması sıfır olan ortak bir değere dönüştürür. Her değişken için ortalama ve standart sapma değeri hesaplanır. Böylece verilerin farklı ölçeklerde olmasından kaynaklanan etkilerin dengelemesi yapılır.

Hesaplanan bu standart sapma ve ortalama değerleri her bir girdiyi normalleştirmek için kullanılır (Adeyemo, 2018). Ayrıca bu yöntem, verilerin dağılımını korurken aynı zamanda verilerin karşılaştırılabilir hale gelmesini sağlar. Fakat Z skoru normalizasyonu, verilerin dağılımına bağlı olarak aykırı değerlerin etkisini azaltmaz (Ünlü, 2022).

$$z = \frac{x-\mu}{\sigma} \tag{2.2}$$

Burada μ değişkenin ortalamasını, σ simge ise değişkenin standart sapmasını ifade etmektedir. (Ünlü, 2022).

**Robust Yöntemi:**

Bu normalizasyon yönteminde ise medyan değeri kaldırılır ve veriler nicel aralığa göre ölçeklendirilmektedir (Adeyemo, 2018). Medyan kullanılması, verilerin ortalamasına karşı daha dirençli olmasını sağlar. Ayrıca verileri aykırı değerlerin etkisinden korumak için kullanılan yöntemlerden biridir ve veri setinde aykırı değerlerin bulunduğu durumlarda kullanışlıdır (Ünlü, 2022).

$$x' = x - \frac{median(x)}{Q_3-Q_1} \tag{2.3}$$

### 2.2.3. Lojistik Regresyonda Analizinde Tahmin Yöntemleri

Lojistik regresyon analizinde parametre tahminlerinde farklı yöntemler kullanılmaktadır. Bunlardan bazıları ise en çok olabilirlik yöntemi, minimum logit Ki-Kare yöntemi ve ağırlıklandırılmış ki-kare yöntemidir.

**En Çok Olabilirlik Yöntemi:**

Bu yöntem bağımlı değişkenin olasılık dağılımından yola çıkarak gözlemlenen değerler içerisinden en büyük olasılığa sahip parametreyi bulmayı amaçlamaktadır. En çok olabilirlik yöntemi kullanılırken parametre, olasılık fonksiyon denklemi biçiminde gösterilir ve logaritması alınır. Bu işlev, gözlemlenen verilerin olasılıklarının bilinmeyen parametrelerin bir işlevi olarak tanımlanmasıdır. Fonksiyonu max yapacak biçimde yani gözlemlenen verilere yakın olan fonksiyon seçilir. Model tercihi yapıldıktan sonra pearson Ki-Kare ve $R^2$ değerlerine bakılarak $R^2$ değeri hangisinin yüksek olmasına göre modelin analizinin daha iyi tahmin ettiğine karar verilir (Ayhan, 2006).

Yapılan çalışmadaki parametre ve denek sayıları kullanılacak olan en çok olabilirlik yöntemini etkileyebilmektedir. Parametre sayısının denek sayısına göre fazla olması durumunda koşullu en çok olabilirlik yöntemi, parametre sayısının denek sayısına göre daha az olması durumunda koşulsuz en çok olabilirlik yöntemi tercih edilmektedir. Bu iki yöntem arasındaki temel fark ise koşullu en çok olabilirlik yöntemi tarafsız (yansız) olması koşulsuz en çok olabilirlik yönteminin ise taraflı (yanlı) sonuçlar vermesidir. Lakin hangi yöntemin kullanılmasına tam olarak karar verilemeyen durumlarda koşullu en çok olabilirlik yönteminin tercih edilmesi önerilmektedir (Kleinbaum ve Klein, 2002).

Denklemsel gösterimi;

y'nin sıfır (0) yada bir (1) değerlerini aldığı, ($x_i$ , $y_i$) ikilileri başarı olasılığı $P_i=P(y_i=1 \mid x)$ olarak tanımlandığında başarısızlık olasılığı 1- $P_i$ ve i = 1,2,3,……,n olup;

$$P(y_i \mid x_i) = P_i^{y_i}(1 - P_i)^{1-y_i} \tag{2.4}$$

2.4'teki ifade bütün gözlem değerlerini ifade edecek şekilde gösterildiğinde;

$$P(Y \mid X) = \prod_{i=1}^{n} P_i^{y_i}(1 - P_i)^{1-y_i} \tag{2.5}$$

Şeklinde en çok olabilirlik fonksiyonu tanımlanmaktadır. Fonksiyon ile daha rahat işlem yapılabilmesi için logaritması alınır.

$$\ln(y \mid x, \beta) = \sum_{i=1}^{n}\{y_i ln(P_i) + (1 - y_i) ln(1 - P_i)\} \tag{2.6}$$

Bu adımdan sonra β değerlerine göre birinci türevleri alınır ve sıfıra (0) eşitlenir.

$$\sum_{i=1}^{n}(y_i - P_i)x_{ij} \tag{2.7}$$

j=1,2,3,…..,p

oluşan bu eşitliklere olabilirlik eşitliği denilmektedir. Bu denklemlerden elde edilen B değerlerine En çok olabilirlik kestirimi denilmektedir (Uruk, 2007).

**Ağırlıklandırılmış Ki-kare Yöntemi:**

Ki-kare testini kullanarak değişkenler arasındaki ilişkiyi değerlendiren bir istatistiksel yöntemdir. i gruba ayrılmış verileri bütünü için $n_i$ deneme yaparak bu denemeler sonucunda $r_i$ başarı elde edildiği varsayılarak başarı oranı $p_i = {r_i}/{n_i}$ tanımlanmaktadır. Bu başarı oranın varyansı alınır.

$$\mathrm{Var}\left(\frac{r_i}{n_i}\right) = \frac{p_i(1-p_i)}{n_i} \tag{2.8}$$

Bağımsız değişkenler için;

$$w_i = \frac{n_i}{p_i(1-p_i)} \tag{2.9}$$

ağırlığı ile ağırlıklandırılmış regresyon tahmini yapılır. Ancak bağımsız değişkenler için elde edilen tahmin ($w_i$) $p_i$ ' nin bir fonksiyonu olduğu için en küçük kareler yöntemi tekrarlanarak uygulanabilmekte ve ağırlık değerleri her aşamada tekrar elde edilmektedir (Tatlıdil, 1996)

**Minimum Logit Ki-kare Yöntemi:**

Ağırlıklarındırılmış ki-kare yönteminin farklı bir versiyonu olan minimum logit Ki-Kare yöntemi 2xJ çapraz tablolarındaki beklenen ve gözlenen değerlerin aralarındaki farklardan yararlanılarak kullanılan bir metottur (Taşkın, 2022).

Bu yöntemin temel amacı gözlemlenen verileri maksimum yapacak şekilde belirlenmeyen parametrelerin değerlerini tahmin etmektir. Genel itibari ile bu yöntem tekrarlı veri setlerinin bulunduğu analizlerde kullanılmaktadır.

Yeniden ağırlıklandırılmış en küçük kareler yönteminde bahsedilen $P_i$ olasılığı ile yapılan logit dönüşümü bağımlı değişkenini oluşturmaktadır. Minimum logit ki-kare yöntemi logit değeri olan bağımlı değişkenin ağırlıklandırılmış regresyonundan en küçük kareler kestirimlerini bulmaya dayanmaktadır. Tek adımda ağırlıklı en küçük kareler kestirimleri minimum logit ki-kare kestirimleri olmaktadır (Öz, 2015).

### 2.2.4. Lojistik Regresyon Analizinin Çeşitleri

Lojistik regresyon analizi; bağımlı değişkenin kategori sayısına ve ölçek durumuna göre üç ayrı başlık altında incelenmiştir.

**İkili Lojistik Regresyon Modeli:**

Bağımlı değişkenin kategorik olduğu durumlarda bağımlı değişken ile bağımsız değişken arasındaki neden sonuç ilişkilerini inceleyen bir yöntemdir. Bağımlı değişken başarılı/başarısız, evet/hayır, olumlu/olumsuz gibi niteliklerden oluşmaktadır.

Yukarıda anlatıldığı bir olayın meydana gelme olasılığı P meydana gelmeme olasılığı 1-P olmak üzere odds oranı eşitlik 2.2 deki gibi hesaplarak sonuçların [0;+∞) arasında olması sağlanır. Daha sonra eşitlik 2.3'teki gibi logit dönüşümü yapılarak sonuçların (-∞;+∞) arasında olması sağlanır ve süreklilik oluşturulmuş olur (Özsürünç, 2022).

Lojistik regresyon modeli tahmin edilen olasılık değeri ise 2.1'deki eşiklik ile elde edilir. Model tahmin edildikten sonra katsayıların anlamlılıklarının test edilmesi ve yorumlanması gerekmektedir. Hipotezleri çoklu doğrusal regresyon modeline benzer kurulmaktadır.

$H_0 : \beta=0$

$H_1 : \beta\neq0$

Ancak lojistik regresyon analizinde logit değeri kullanıldığı için bağımsız değişkene ait katsayının bir (1) olup olmadığı veya bağımlı değişkendeki bir gruba dahil olma olasılığının 0,5 olup olmadığı test edilmektedir (Özsürünç, 2022).

Katsayıların anlamlılığını test etmek için wald testi, olabilirlik oran testi ve skor testi kullanılan yöntemlerden biridir. Bu test istatistikleri ile ilgili bilgiler ilerleyen başlıklarda verilecektir. Fakat tek cümle ile bahsetmek gerekirse hesaplanan bütün katsayıların anlamlılığını test etmektedir.

Lojistik regresyon analizinde katsayıların yorumlanması da farklılık içermektedir. Çünkü bağımlı değişken değeri 0-1 arasında bir değer almaktadır. Bu durumda orijinal ve hesaplanmış olmak üzere iki farklı değerin yorumlanması gerekmektedir. Orijinal değerler ilişkinin yönünü belirlemede kullanılırken etki büyüklüğünü ise bağımsız değişken katsayılarının $e^{\beta_1}$, $e^{\beta_2}$, ….. $e^{\beta_n}$ şeklinde üstel değerleri hesaplanmalıdır. Yüksek olan değerin katsayısının bağımlı değişkene olan etkisi de yüksek olmaktadır (Özsürünç, 2022).

**Multinominal Lojistik Regresyon Modeli:**

Bağımlı değişkenin kategori sayısının ikiden fazla olduğu durumlarda kullanılan ve bağımlı değişken ile bağımsız değişken arasında neden sonuç ilişkisinin inceleyen bir metottur. İkili lojistik regresyon modelinin geliştirilmiş versiyonu olan metotta da bağımlı değişken nitel veya nicel olabilmektedir. Multinominal lojistik regresyon modelinde

öncelikle bir referans kategori belirlenir ve diğer kategoriler bu referans kategorisi ile karşılaştırılır. Referans olarak seçilecek kategorinin belirli bir düzeni olamayıp araştırmacının tercihine bırakılmıştır. Bağımlı değişkenin binom yerine multinom olmasından dolayı kategori sayısının bir (1) eksiği kadar denklem oluşmaktadır (Özsürünç, 2022).

Örneğin üç (3) kategorili (k = 0, 1, 2) bağımlı değişkene sahip bir çalışmada multinominal lojistik regresyon modeli denklem 2.10 ve 2.11'de gösterilmiştir.

$$k_1(x) = \ln\left(\frac{P(Y=1 \mid x)}{P(Y=0 \mid x)}\right) = \beta_{10}+\beta_{11}X_1+\beta_{12}X_2+ \ldots +\beta_{1p}X_p \quad (2.10)$$

$$k_2(x) = \ln\left(\frac{P(Y=2 \mid x)}{P(Y=0 \mid x)}\right) = \beta_{20}+\beta_{21}X_1+\beta_{22}X_2+ \ldots +\beta_{2p}X_p \quad (2.11)$$

Genel bir gösterim ile ifade etmek gerekirse;

$$P(Y = j \mid x) = \frac{e^{g_j(x)}}{\sum_{k=0}^{2} e^{g_k(x)}} \quad (2.12)$$

Multinominal lojistik regresyon modelinde de katsayıların anlamlılığı test edilip yorumlanması gerekmektedir. Bu modelde en az üç grup olduğundan en az iki lojistik regresyon modeli oluşmaktadır. İkili lojistik regresyon modelinde olduğu gibi her bir regresyon modeli için orijinal ve hesaplanmış değerlerin yorumlanması gerekmektedir. Multinominal lojistik regresyon modelinde de orijinal değerler ilişkinin yönünü belirlemede kullanılırken etki büyüklüğünü ise bağımsız değişken katsayılarının $e^{\beta_1}$, $e^{\beta_2}$, ….. $e^{\beta_n}$ şeklinde üstel değerleri hesaplanmalıdır. Yüksek olan değerin katsayısının bağımlı değişkene olan etkisi de yüksek olmaktadır (Özsürünç, 2022).

**Ordinal Lojistik Regresyon Modeli:**

Kategorik bağımlı değişkenin ordinal bir ölçeğe haiz olduğu ve bağımlı değişkenlerin üç veya daha fazla olduğu durumlarda kullanılan bir analiz yöntemidir. K gruplu verilerin bulunduğu ve grupların sıralı olduğu durumlarda kategori sayılarının ikili kombinasyonları kadar model çıkmasıdır. Sıralanabilir olmasından kastedilen kendi içinde bağımlı değişkenin belirli bir önem sırasının olmasından kaynaklanmaktadır. Multinominal lojistik regresyon modeline göre yorumlanması daha kolaydır. Çünkü tahmin edilen model sayısı ordinal lojistik regresyonda tekdir. Fakat dezavantajı da bulunmaktadır. Buda verilere uygun olmayan kısıtlamalar getirmesidir. Ordinal lojistik

regresyonunda paralellik varsayımı (orantılı odds varsayımı) bulunmaktadır. Bu varsayım sıralı olarak farklı bölünmelerden bağımsız olarak bütün kategoriler için bağımsız değişkenlerin oddsları üzerinde aynı etkiye sahip olduğunu ileri sürmektedir. Yani ayrı ayrı lojistik regresyonlar olsa bile tek bir odds değerini ifade etmektedir (Esenyel, 2022).

Ordinal lojistik regresyon modelinde de diğer modellerde olduğu gibi bütün katsayıların anlamlılığı test edilir. Anlamlı ise hesaplanan olasılık değerini nasıl etkilediği yani grup atamalarının tahminini nasıl etkilediği yorumlanabilir. Yukarıda belirtildiği üzere ikili kombinasyonları kadar oluşturulan modeller için ortak bir olasılık oranı gözlenecektir (Özsürünç, 2022).

Bağımsız değişken katsayısının pozitif olması durumunda odds oranının birden yüksek çıkması, bağımsız değişken değeri bir birim arttıkça veya bağımsız değişken kategorisi referans kategorisinden karşılaştırma yapılacak kategoriye geçtikçe bağımlı değişken kategorileri için düşük kategoriden yüksek kategoriye geçme olasılığının da odds oranı kadar artacağı söylenebilir. Bağımsız değişken katsayısının negatif olduğu durumda ise tam tersi şeklinde yorumlanır. Ordinal lojistik regresyon modelinde de orijinal değerler ilişkinin yönünü belirlemede kullanılırken etki büyüklüğünü ise bağımsız değişken katsayılarının $e^{\beta_1}$, $e^{\beta_2}$,... $e^{\beta_n}$ şeklinde üstel değerleri hesaplanmalıdır. Yüksek olan değerin katsayısının bağımlı değişkene olan etkisi de yüksek olmaktadır (Özsürünç, 2022).

## 2.3. LOJİSTİK REGRESYONDA PARAMETRELERİN ANLAMLILIK TESTİ

Yukarıda anlatılan lojistik regresyon çeşitlerinden biri kullanılarak modelleştirildikten sonra parametrelerin anlamlılıkları test edilmelidir. Bu test yöntemlerindeki amaç parametrenin modelimizde bulunduğunda verdiği bilgi ile modelimizde bulunmadığı zaman verdiği bilgiyi karşılaştırmaktır (Alpar, 2017). Literatürde kullanılan bazı test yöntemleri aşağıda ifade edilmektedir.

### 2.3.1. Olabilirlik Oran Testi

Olabilirlik oran testinde parametrenin/parametrelerin modelde bulunduğu zaman ile modelde bulunmadığı zamandaki değerlerini kıyaslamaktır. Bir başka değişle

gözlemlenen veri hacmini tahmin etmenin ihtimalini en iyi yapabilecek olan bilinmeyen parametrelerin değerini vermektir.

$$G = -2\ln\left[\frac{L(Parametre\ Modelde\ Bulunmadığında)}{L(Parametre\ Modelde\ Bulunduğunda)}\right] \quad (2.13)$$

Denklem 2.13'te gösterildiği gibi istatistik değeri hesaplanmaktadır. Serbestlik derecesi ise tahmin edilen iki modelde elde edilen parametre sayılarının arasındaki farka eşittir (Alpar, 2017).

Olabilirlik oran testinde öncelikle sadece sabit terimin bulunduğu model oluşturulur. Sonrasında ise sabit terim ile bağımsız değişkenlerin bulunduğu model oluşturularak -2 ile çarpılıp oran test değeri elde edilmiş olur (Alpar, 2017).

Bu test yönteminde elde edilen istatistik değerinin küçük olması modele eklenen değişkenlerin modele önemli bir katkı sağlamadığını ve modelde bulunmasına gerek olmadığını göstermektedir. Bu yöntemin uygulana bilmesi için gözlem sayısının yeterince büyük olması gerekmekte ve gözlem değerleri eksiksiz bir şekilde girilmiş olması gerekmektedir. Aksi takdirde eksik gözlem içeren bağımsız değişken modelden atılarak model oluşturulmaya çalışılır (Alpar, 2017).

### 2.3.2. Wald Testi

Tahmin edilen β katsayısının anlamlılığını ifade eden bir ölçü olan wald testi bağımsız değişkenin modele katkısını ifade etmektedir. Bir başka değişle regresyon modelinde bulunan parametrelerin test edilmesi için kullanılan t istatistik değerinin logit modeldeki karşılığıdır. Yani parametrelerin istatistiksel olarak anlamlı olup olmadığını değerlendirmek için kullanılan bir istatistiksel test yöntemidir (Astar, 2009).

Wald test istatistiği, β'nın tahmin edilen değerlerinin kendi standart hatasına oranlanması ile hesaplanmaktadır.

$$W = \frac{\beta_1}{S(\beta_1)} \quad (2.14)$$

Eşitlik 2.14'ün pay ve paydalarının kareleri alındığı zaman istatistik değeri Ki-Kare dağılımına uymaktadır (Kahraman, 2021).

### 2.3.3. Skor Testi

İstatistiksel modellerin parametrelerinin anlamlılığını değerlendirmek için kullanılan bir istatistiksel test yöntemidir. Bu test, genellikle maksimum olabilirlik yöntemiyle tahmin edilen parametrelerin istatistiksel olarak anlamlı olup olmadığını belirlemek için kullanılır. Skor testi diğer test yöntemlerinden farklı olarak en çok olabilirliğin hesaplamasına gerek duymadan hangi parametrenin modele dahil olacağına veya hangisinin modelden çıkarılması gerektiğine yardımcı olan bir test istatistiğidir. Skor testi, parametrenin log-olabilirlik fonksiyonunun birinci türevidir. Tahmin edilen parametrenin değerlendirilmesi ve standart hatalarının hesaplanması için kullanılır. Bu test istatistiğinde matematiksel işlemler diğer yöntemlere nazaran daha azdır. Skor test istatistiği;

$$ST = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sqrt{y_i(1 - \bar{y})\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{2.15}$$

şeklinde hesaplanmaktadır (Kahraman, 2021).

## 2.4. LOJİSTİK REGRESYON MODELİNİN UYUMUN İYİLİĞİNİN BELİRLENMESİ

Lojistik regresyon modelinde değişkenlerin anlamlılığı kadar model uyumu da önemli olmaktadır. Bunun için model oluşturulduktan sonra modele dâhil edilmek istenen değişkenlerin bağımlı değişkenin üzerindeki etkisinin de belirlenmesi gerekmektedir. Uyumun iyiliğinin belirlenmesi için kullanılan bazı yöntemler arasında Pearson Ki-Kare İstatistiği, Sapma Ölçüsü, Hosmer-Lemeshow Testi ve Sınıflandırman Tablosu bulunmaktadır (Taşkın, 2022).

### 2.4.1. Pearson Ki-kare İstatistiği

Pearson ki-kare istatistiği, araştırılan değerler için gözlenen ve tahmin edilen olasılıklarının kıyaslanması ile uyum iyiliğini incelemektedir.

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_j\hat{\pi}_j(1-\hat{\pi}_j)} \tag{2.16}$$

Pearson ki-kare istatistiği eşitlik 2.16'daki gibi hesaplanmaktadır. Burada;

n = başarıların beklenen sayısını,

1-n = başarısızlıkların beklenen sayısını ifade etmektedir.

2.16'da hesaplanan Ki-kare değeri ile Ki-kare tablosunda serbestlik derecesine karşılık gelen tablo değeri ile karşılaştırılır. Pearson istatistiği tablo değerinden büyük ise gözlenen değişkenin beklenen dağılımına uygun olduğu söylenebilmektedir (Taşkın, 2022).

### 2.4.2. Sapma Ölçüsü

Uyumun iyiliğinin belirlenmesi için kullanılan yöntemlerden biri de gözlenen değerlerden tahmin edilen değerler arasında sapmanın ne kadar olduğunu gösteren sapma ölçüsüdür (Taşkın, 2022). Model uyumu için hesaplanan sapma ölçüsünün küçük olması gerekmektedir.

$$D = 2\sum_{i=1}^{n} O_{ij} ln\left(\frac{O_{ij}}{E_{ij}}\right) \quad (2.17)$$

Sapma ölçüsü eşitlik 2.17'deki gibi hesaplanmaktadır. Burada;

$O_{ij}$ = Gözlemlenmiş değerleri,

$E_{ij}$ = Beklenen Değerleri ifade etmektedir.

### 2.4.3. Hosmer-Lemeshow Testi

Hosmeer-Lemeshow testi modeli bir bütün olarak test etmesinden dolayı model Ki-Kare testi olarak ta bilinmektedir. Genellikle risk tahmin modellerinde kullanılmaktadır. Sabit terim haricindeki bütün logit bağımsız değişkenlerin katsayılarının sıfıra eşit olup olmadıklarının test edildiği bir yöntemdir (Taşkın, 2022).

Hosmer-Lemeshow test istatistiği;

$$C = \sum \frac{(G-B)^2}{B} \quad (2.18)$$

G: Gözlenen Değer,

B: Beklenen Değer

Eşitlik 2.18'deki gibi ifade edilmektedir.

k-2 serbestlik dercesine sahip Ki-Kare dağılımına göre kıyaslanan Hosmer-Lemeshow testi için en az 400 gözlem önerilmektedir.

### 2.4.4. Sınıflandırma Tablosu

Lojistik regresyon analizinin temel açlarında biri olan sınıflandırma ve gruplandırma olduğundan dolayı uygulanan sınıflandırmanın test edilmesi de uyumun iyiliği için önemli bir durumdur. Bu tablo bağımlı değişkene sıfır ve bir rakamları atanarak gözlemlenen değerler ile beklenen değerlerin çaprazlanması ile oluşturulmaktadır. Bu çaprazlamada ikili bağımsız değişken belirleyebilmek için kesim noktası belirlenmektedir. Bu nokta genellikle 0.5 olarak belirlenmektedir. Kestirilen değerler Y eksen noktasını 0.5'in üzerinde kesen değişkenler 1'e 0.5 noktasının altında kesen değişkenler ise 0'a atanır (Taşkın, 2022).

Eşitlik 2.19'da gözlem sonuçlarının ayarlanan gruplardan birine atanması için kullanılmaktadır.

$$\left[\frac{1}{\left(1+e^{-\left(x_i^j\beta\right)}\right)}\right] \tag{2.19}$$

## 2.5. LOJİSTİK REGRESYON MODELİNİN DEĞERLENDİRİLMESİ

Oluşturulan modeller karşılaştırılarak hangi modelin amacımızı daha iyi ifade ettiğini belirmek için hesaplamalara ihtiyaç vardır. Genel olarak model uygunluğunun değerlendirilmesinde oluşturulan model ile gerçek değerler arasındaki standart farklara bakılmaktadır (Ürük, 2007).

### 2.5.1. Standart Olmayan Hatalar

Bu hata türünde gerçek değerler ile tahmin edilen modelin değerleri arasındaki fark göz önüne alınmaktadır. Bu farkın eşit olduğu ifade edilmektedir (Ürük, 2007).

$$e_i = \frac{e_i}{P_i(1-P_i)} \tag{2.20}$$

i. Birimin standart olmayan hatası eşitlik 2.20'deki gibi hesaplanmaktadır.

### 2.5.2. Standart Hatalar

Parametre tahminlerinin hassasiyetini gösteren istatistiksel bir ölçüdür. Ayrıca parametre tahminlerinin değişkenliğini ve belirsizliğini yansıtır. Düşük olan standart hata değerleri tahminin daha güvenilir olduğunu, yüksek olan standart hatalar ise tahminin güvenilirliğinin daha az olduğunu ifade eder. Sonuç olarak parametre tahminlerinin güvenilirliğini ve istatistiksel anlamlılığını belirlemek için önemli bir ölçüdür. Standart hata değerleri eşitlik 2.21'deki gibi standart olmayan hata değerlerinin kendi standart sapma değerlerine bölünmesiyle elde edilmektedir. Genel olarak örnek hacimlerinin büyük olduğu analizlerde normal dağılım göstermektedir (Kahraman, 2022).

$$Z_i = \frac{e_i}{\sqrt{P_i(1-P_i)}} \tag{2.21}$$

### 2.5.3. Sapma Değerleri

Sapma değeri, modelin veriyle uyumu veya uyumsuzluğu hakkında bilgi sağlamaktadır. Gözlemlenen verilerin model tarafından tahmin edilen değerlerden ne kadar farklı olduğunu ölçer. Sapma değerleri yukarıda standart hatalarda da bahsedildiği üzere daha düşük sapma değerleri, modelin veriyle uyum sağladığını ve daha iyi tahminler yaptığını, yüksek sapma değerleri ise modelin veriye uyum sağlamada zayıf olduğunu işaret eder (Kahraman, 2022). Sapma değerleri eşitlik 2.22 ve 2.23'teki gibi hesaplanmaktadır.

$$\text{Sapma} = \sqrt{-2\ln(P_i)} \tag{2.22}$$

$$\text{Sapma} = -\sqrt{-2\ln(1-P_i)} \tag{2.23}$$

### 2.5.4. Uzaklık Değerleri

Tahmin edilen model üzerinde etkisi olan gözlemlerin belirlenmesi için kullanılmaktadır. Bu değerler sıfır (0) ve bir (1) değerleri aralığında değerler almakta olup, sıfıra yaklaştığında etkisi azalmakta bire yaklaştığında ise etkisinin arttığı şeklinde yorumlanmaktadır.

Uzaklık değerleri tahmin edilen modelin uzaklık değer ortalaması ile karşılaştırılmaktadır. Uzaklık değerlerin ortalaması ise P/n oranlaması ile hesaplanmaktadır. Burada P parametre sayısını ve n ise örnek hacmi ifade etmektedir (Kahraman, 2022).

### 2.5.5. Cook Uzaklığı

Cook uzaklığı, her bir gözlemin model üzerindeki etkisini ölçmektedir. Genellikle gözlem etkisi ve aykırı değerlerin belirlenmesinde kullanılmaktadır. Yüksek Cook uzaklığı değerine sahip olan gözlemler, model üzerinde büyük bir etkiye sahip olabilir ve modelin doğruluğunu etkileyebilir. Yani, bu gözlemler dikkate alınarak modelin yeniden değerlendirilmesi veya gözlemlerin incelenmesi gerekmektedir. Bu uzaklık metodunun farkı ise tahmin edilen herhangi bir parametrenin model üzerindeki etkisini belirlemek için ve tahmin edilen katsayının modelden çıkarılması halinde modeldeki diğer parametrelerin hangi oranda değişim göstereceğini göstermektedir (Kahraman, 2022).

$$CU = Z_i^2 \left(\frac{h_i}{1-h_i}\right) \tag{2.24}$$

$Z_i$ = Standartlaştırılmış Hata,

$h_i$ = Uzaklık Değerini ifade etmektedir.

Cook uzaklık değeri eşitlik 2.24'teki gibi hesaplanmaktadır.

### 2.5.6. DfBeta Değerleri

Dfbeta değerleri, her bir gözlemin modelin parametre tahminlerine olan etkisini ölçen istatistiksel yöntemdir. Bu değer gözlemin parametre tahminlerini ne kadar değiştirdiğini ve bu değişimin modelin çıktılarını nasıl etkilediğini göstermektedir. Büyük değerler gözlemin parametre tahminlerini daha fazla değiştirdiğini ve çıktıları önemli ölçüde etkilediği göstermektedir. Bu değerlerin hesaplanmasında ise tahmin edilen modelden çıkarılan parametre sonucunda modelde kalan parametre katsayılarındaki değişimi göstermektir (Kahraman, 2022). DfBeta değerleri eşitlik 2.25 teki gibi hesaplanmaktadır.

$$DfBeta = B_j - B_j^i \tag{2.25}$$

$B_j$ = Bütün Birimlerin Modele Dahil Edilmesi Halinde Parametreleri,

$B_j^i$ = i. Birimin Modelden Çıkarılması Sonucunda Hesaplanan Parametreleri ifade etmektedir.

# ÜÇÜNCÜ BÖLÜM

# ÜLKEMİZDE MEYDANA GELEN TRAFİK KAZA VERİLERİ KULLANILARAK LOJİSTİK REGRESYON ANALİZİ İLE CART ALGORİTMASININ KARŞILAŞTIRILMASI

## 3.1. ARAŞTIRMANIN AMACI VE ÖNEMİ

Çalışmada, 2013-2022 yılları arasında Ülkemizde meydana gelen trafik kazaları verileri kullanılarak lojistik regresyon analizi ile CART algoritmasının karşılaştırılması amaçlanmaktadır. Çalışmada ayrıca Ülkemizde 2013-2022 yılları arasında meydana gelen trafik kazalarının yaralamalı veya ölümlü olma durumları kullanılarak iki yöntemin sınıflandırma performansları karşılaştırılmıştır. Böylece hangi yöntemin sınıflandırma performansının daha iyi olduğu belirlenmeye çalışılmıştır. Çalışmada trafik kazalarının yaralamalı veya ölümlü olma durumunu etkileyen faktörler tespit edilip toplumun bilinçlendirilmesi amaçlanmaktadır.

## 3.2. ARAŞTIRMA VERİSİ

Araştırmada kullanılan veriler, Emniyet Genel Müdürlüğü Trafik Başkanlığı tarafından tutulan 2013-2022 yılları arasında meydana gelen yaralamalı veya ölümlü trafik kaza verilerinden oluşturmaktadır. Trafik kazalarını etkileyebilecek birçok etken göz önüne alındığında araştırmada kullanılacak veriler oldukça büyük boyutta olmaktadır. Trafik kazalarını etkilediği düşünülen değişkenler dikkate alınmıştır. Çalışmada bağımsız değişken olarak; kazanın meydana geldiği yolun tipi, yolun kaplaması, yolun sınıfı, gün durumu, hava durumu, yolun yüzeyi, kazaya karışan araç sayısı, sürücü yaşı, yolcu kusurları ve sürücü kusurları dikkate alınmıştır.

## 3.3 ARAŞTIRMA METODOLOJİSİ

Çalışmada lojistik regresyon analizi ile CART algoritması yardımıyla trafik kaza durumları sınıflandırmak istenmiştir. İki analizde de bağımlı değişken “1” ve “0” olarak kodlanmıştır. “1” trafik kazasının ölümlü sonuçlandığını ve “0” ise trafik kazasının yaralamalı sonuçlandığını ifade etmektedir.

CART algoritması ve lojistik regresyon analizi sonucunda sürücü ve yaya için kazanın yaralamalı mı ölümlü mü olma durumu için değişkenler kullanılarak sınıflandırma tahminleri yapılmış devamında ise anlamsız olduğu görülen bağımsız değişkenler analiz ve algoritmadan çıkarılarak tekrar sınıflandırma değerleri tahmin edilmiş ve karşılaştırılmıştır. Bu karşılaştırma için phyton paket programı kullanılmıştır.

## 3.4. ANALİZ VE BULGULAR

Çalışmada gini dallanma ölçütü kullanılarak maksimum ağaç modeli ile CART algoritması ve lojistik regresyon analizi için sınıflandırma değerleri oluşturulmuştur. Devam eden süreçte anlamsız olan bağımsız değişkenler modelden çıkarılarak lojistik regresyon analizi ve CART algoritması için hesaplanmalar yeniden yapılmıştır. Daha sonra iki yöntemin sınıflandırma başarıları karşılaştırılmıştır.

## 3.5 TANIMLAYICI İSTATİSTİKLER

Çalışmada ilk olarak analizlerde kullanılan değişkenler için frekans tabloları oluşturulmuştur. Kazaların illere göre dağılımı Tablo 3.1'de verilmiştir.

**Tablo 3.1.** Kazaların İllere Göre Dağılımı

| İller | Frekans | Yüzde | İller | Frekans | Yüzde |
|---|---|---|---|---|---|
| Adana | 43734 | 2,9 | Konya | 55886 | 3,7 |
| Adıyaman | 9612 | 0,6 | Kütahya | 10082 | 0,7 |
| Afyonkarahisar | 15089 | 1,0 | Malatya | 13759 | 0,9 |
| Ağrı | 5982 | 0,4 | Manisa | 34834 | 2,3 |
| Amasya | 8787 | 0,6 | Kahramanmaraş | 20388 | 1,3 |
| Ankara | 111213 | 7,4 | Mardin | 8440 | 0,6 |
| Antalya | 65349 | 4,3 | Muğla | 36552 | 2,4 |
| Artvin | 2424 | 0,2 | Muş | 2841 | 0,2 |
| Aydın | 30263 | 2,0 | Nevşehir | 7684 | 0,5 |
| Balıkesir | 30377 | 2,0 | Niğde | 7924 | 0,5 |
| Bilecik | 4758 | 0,3 | Ordu | 13327 | 0,9 |
| Bingöl | 3990 | 0,3 | Rize | 6261 | 0,4 |
| Bitlis | 4721 | 0,3 | Sakarya | 22841 | 1,5 |

**Tablo 3.1.** (Devamı)

| | | | | | |
|---|---|---|---|---|---|
| Bolu | 6765 | 0,4 | Samsun | 28515 | 1,9 |
| Burdur | 8106 | 0,5 | Siirt | 4259 | 0,3 |
| Bursa | 55855 | 3,7 | Sinop | 3901 | 0,3 |
| Çanakkale | 10839 | 0,7 | Sivas | 13407 | 0,9 |
| Çankırı | 4376 | 0,3 | Tekirdağ | 18623 | 1,2 |
| Çorum | 13564 | 0,9 | Tokat | 13522 | 0,9 |
| Denizli | 26935 | 1,8 | Trabzon | 13107 | 0,9 |
| Diyarbakır | 19745 | 1,3 | Tunceli | 1246 | 0,1 |
| Edirne | 6928 | 0,5 | Şanlıurfa | 23292 | 1,5 |
| Elazığ | 11037 | 0,7 | Uşak | 10156 | 0,7 |
| Erzincan | 6663 | 0,4 | Van | 12590 | 0,8 |
| Erzurum | 10614 | 0,7 | Yozgat | 8870 | 0,6 |
| Eskişehir | 18038 | 1,2 | Zonguldak | 8828 | 0,6 |
| Gaziantep | 32792 | 2,2 | Aksaray | 10752 | 0,7 |
| Giresun | 7149 | 0,5 | Bayburt | 1373 | 0,1 |
| Gümüşhane | 2440 | 0,2 | Karaman | 6617 | 0,4 |
| Hakkari | 1558 | 0,1 | Kırıkkale | 8153 | 0,5 |
| Hatay | 34242 | 2,3 | Batman | 5766 | 0,4 |
| Isparta | 12182 | 0,8 | Şırnak | 4742 | 0,3 |
| Mersin | 52580 | 3,5 | Bartın | 3180 | 0,2 |
| İstanbul | 166483 | 11,0 | Ardahan | 904 | 0,1 |
| İzmir | 91014 | 6,0 | Iğdır | 2630 | 0,2 |
| Kars | 3225 | 0,2 | Yalova | 6133 | 0,4 |
| Kastamonu | 6710 | 0,4 | Karabük | 4602 | 0,3 |
| Kayseri | 33540 | 2,2 | Kilis | 5113 | 0,3 |
| Kırklareli | 6492 | 0,4 | Osmaniye | 15706 | 1,0 |
| Kırşehir | 5652 | 0,4 | Düzce | 8726 | 0,6 |
| Kocaeli | 35299 | 2,3 | **Toplam** | **1512654** | **100** |

Tablo 3.1 incelendiğinde; en çok kaza oranın yüzde 11 ile İstanbul ilinde meydana geldiği görülmüştür. İstanbul ilinin trafikte bulunan araç sayısı ve nüfusunun fazla olması bu durumu destekler niteliktedir. Tunceli, Ardahan ve Bayburt gibi nüfusu düşük illerin ise kaza durumlarının düşük olması nüfus ve araç sayılarına göre muhtemel bir durumdur.

Kazaların aylara göre dağılımı Tablo 3.2'de verilmiştir.

**Tablo 3.2.** Kazaların Aylara Göre Dağılımı

| Aylar | Frekans | Yüzde | Aylar | Frekans | Yüzde |
|---|---|---|---|---|---|
| Ocak | 95930 | 6,3 | Ağustos | 156699 | 10,4 |
| Şubat | 91493 | 6,0 | Eylül | 146669 | 9,7 |
| Mart | 110129 | 7,3 | Ekim | 140167 | 9,3 |
| Nisan | 114063 | 7,5 | Kasım | 124483 | 8,2 |
| Mayıs | 126617 | 8,4 | Aralık | 111784 | 7,4 |
| Haziran | 139350 | 9,2 | **Toplam** | **1512654** | **100** |
| Temmuz | 155270 | 10,3 | | | |

Tablo 3.2 incelendiğinde; en çok kazanın gerçekleştiği ayın %10,4 ile ağustos ayı, ağustos ayını ise yüzde 10,3 ile temmuz ayı takip etmektedir. Yaz döneminde nüfus hareketliliğinin de fazla olması kazaların esas nedenleri arasında olduğu söylenebilir.

Kazaların gerçekleştiği yolun tipine göre dağılımı Tablo 3.3'te verilmiştir.

**Tablo 3.3.** Kazaların Gerçekleştiği Yolun Tipi

| Yolun Tipi | Frekans | Yüzde |
|---|---|---|
| Bölünmüş Yol | 868815 | 57,4 |
| Tek Yönlü yol | 122900 | 8,1 |
| İki Yönlü Yol | 503506 | 33,3 |
| Diğer | 17433 | 1,2 |
| **Toplam** | **1512654** | **100** |

Tablo 3.3 incelendiğinde kazaların %57,4'ünün bölünmüş yolarda meydana geldiği görülmektedir. Buradan hareketle yaralamalı ya da ölümlü kazaların azaltılabilmesi için yapılacak yolların tek yönlü veya iki yönlü yol olmasına dikkat edilmelidir.

Kazaların gerçekleştiği yolun kaplaması ile ilgili değerler Tablo 3.4'te verilmiştir.

**Tablo 3.4.** Kazaların Gerçekleştiği Yolun Kaplaması

| Yolun Kaplaması | Frekans | Yüzde |
|---|---|---|
| Asfalt | 1402104 | 92,7 |
| Sathi kaplama | 12610 | 0,8 |
| Beton | 6793 | 0,4 |
| Parke | 81675 | 5,4 |
| Stabilize | 5550 | 0,4 |
| Toprak | 3922 | 0,3 |
| **Toplam** | **1512654** | **100** |

Tablo 3.4 incelendiğinde; kazaların %92,7'lik kısmının asfalt tipi kaplamalı yollarda meydana geldiği görülmektedir. Sürücülerin bu tip kaplamaya sahip yollarda diğer kaplamalı yollara göre daha dikkatsiz ve kontrolsüz hareket ettiğinden kazalar bu kadar yüksek orana sahip olabilir.

Kazaların gerçekleştiği yolun sınıfı ile ilgili değerler Tablo 3.5'te verilmiştir.

**Tablo 3.5.** Kazaların Gerçekleştiği Yolun Sınıfı

| Yolun Sınıfı | Frekans | Yüzde |
|---|---|---|
| Cadde | 906191 | 59,9 |
| Sokak | 181819 | 12,0 |
| Otoyol | 40918 | 2,7 |
| Devlet Yolu | 303720 | 20,1 |
| İl Yolu | 18593 | 1,2 |
| Köy Yolu | 7218 | 0,5 |
| Orman Yolu | 395 | 0,0 |
| Servis Yolu | 7403 | 0,5 |
| Bağlantı Yolu | 7311 | 0,5 |
| Park Alanı | 2405 | 0,2 |
| Tesis (mülk) Önü veya İçi | 4431 | 0,3 |
| Su Yolu Taşıtı | 78 | 0,0 |
| Diğer | 32172 | 2,1 |
| **Toplam** | **1512654** | **100** |

Tablo 3.5 incelendiğinde; kazaların %'59,9'luk kısmının cadde sınıfı yollarda meydana geldiği görülmektedir. Cadde sınıfı yolların şehir içlerinde daha etkin kullanıldığı için ve insan hareketliliğinde bu kısımda fazla olmasında kaynaklı olarak en çok kaza oranı bu sınıf yollarda meydana gelmektedir.

Kazaların gerçekleştiği gün durumu ile ilgili değerler Tablo 3.6'da gösterilmiştir.

**Tablo 3.6.** Kazaların Gerçekleştiği Gün Durumu

| **Gün Durumu** | **Frekans** | **Yüzde** |
|---|---|---|
| Gündüz | 1009527 | 66,7 |
| Gece | 462351 | 30,6 |
| Alacakaranlık | 40776 | 2,7 |
| Toplam | 1512654 | 100 |

Geceye göre hareketliliğin daha çok yaşandığı göz önüne alınarak gündüz kaza oranlarının yüksek çıkması normal bir durumdur. Buradan hareket ile gündüz trafiğe çıkan kişilerin daha dikkatli olması gerekmektedir.

Kazaların gerçekleştiği hava durumu ile ilgili değerler Tablo 3.7'de verilmiştir.

**Tablo 3.7.** Kazaların Gerçekleştiği Hava Durumu

| **Hava Durumu** | **Frekans** | **Yüzde** |
|---|---|---|
| Açık | 1328331 | 87,8 |
| Sis/duman | 12821 | 0,8 |
| Yağmur | 126979 | 8,4 |
| Kar | 13633 | 0,9 |
| Sulu sepken | 3187 | 0,2 |
| Dolu | 393 | 0,0 |
| Tipi | 598 | 0,0 |
| Kuvvetli rüzgâr | 1050 | 0,1 |
| Toz/ kum fırtınası | 174 | 0,0 |
| Bulutlu | 25488 | 1,7 |
| **Toplam** | **1512654** | **100** |

Tablo 3.7 incelendiğinde; kazaların %87,8'lik kısmının açık havada gerçekleştiği görülmektedir. Bu durumda sürücülerin diğer hava durumlarında araçlarını kullanma durumlarını tercih etmediği ayrıca diğer hava durumlarında araçlarını kullanan sürücülerin daha dikkatli davrandıkları düşünülebilir.

Kazaların gerçekleştiği yolun yüzeyi ile ilgili değerler Tablo 3.8'de verilmiştir.

**Tablo 3.8.** Kazaların Gerçekleştiği Yolun Yüzeyi

| **Yolun Yüzeyi** | **Frekans** | **Yüzde** |
|---|---|---|
| Kuru | 1293709 | 85,5 |
| Islak / nemli | 196036 | 13,0 |
| Karlı | 9456 | 0,6 |
| Buzlu | 8831 | 0,6 |
| Sel / su birikintili | 1828 | 0,1 |
| Diğer kaygan yüzey | 2794 | 0,2 |
| **Toplam** | **1512654** | **100** |

Tablo 3.8 incelendiğinde kazaların %85,5'lik kısmının yolun yüzeyinin kuru olduğu durumlarda gerçekleştiği görülmektedir. Bu durumda bir diğer istatistiğimizi destekler nitelikte olup, ıslak karlı vb. zeminlerde sürücülerin araçlarını kullanmayı daha az tercih ettiklerini ve kullanan kişilerin ise daha dikkatli hareket ettiklerini desteklemektedir.

Kazalarda sürücü kusurları ile ilgili değerler Tablo 3.9'da verilmiştir.

**Tablo 3.9.** Kazalarda Sürücü Kusurları

| **Sürücü Kusurları** | **Frekans** | **Yüzde** |
|---|---|---|
| Alkollü araç kullanmak | 10034 | 0,6 |
| Araç hızını yol hava ve trafiğin gerektirdiği şartlara uydurmamak | 630528 | 37,7 |
| Araçta zorunlu belge ve gereçleri bulundurmamak | 450 | 0,0 |
| Arkadan çarpmak | 143195 | 8,5 |
| Aşırı hızla araç kullanmak | 17071 | 1,0 |
| Bisiklet ve motosikletin kurallarına uyulmadan kullanılması | 5226 | 0,3 |
| Dönüş kurallarına uymamak | 124760 | 7,4 |

**Tablo 3.9.** (Devamı)

| | | |
|---|---|---|
| Eksik, bozuk veya uygun olmayan araç donanımı ile araç kullanmak | 3273 | 0,1 |
| Geçme yasağı olan yerlerden geçmek | 11006 | 0,6 |
| Hatalı veya yasak olan yerlerden geçmek | 8537 | 0,5 |
| Kavşak geçiş önceliğine uymamak | 13379 | 0,8 |
| Kavşak geçit ve kaplamanın dar olduğu yerlerde geçiş önceliğine uymamak | 239652 | 14,3 |
| Kaza mahallinde durmamak gerekli tedbirleri almamak | 1254 | 0,0 |
| Kırmızı ışık veya görevlinin dur işaretine uymamak | 44739 | 2,6 |
| Koruyucu tertibat (kemer, kask) kullanmamak | 329 | 0,0 |
| Kurallara uygun park etmiş araçlara çarpmak | 1837 | 0,1 |
| Manevraları düzenleyen genel şartlara uymamak | 64683 | 3,8 |
| Saygısızca araç kullanmak ve araçtan bir şeyler atmak | 274 | 0,0 |
| Seyir halinde telefonla konuşmak | 60 | 0,0 |
| Şerit izleme ve değiştirme kurallarına uymamak | 167577 | 10,0 |
| Taşıma sınırı üzerinde yolcu taşıma | 269 | 0,0 |
| Taşıt giremez işareti bulunan yerlere girmek | 48885 | 2,9 |
| Taşıt kullanma sürelerine uymamak | 120 | 0,0 |
| Tehlikeli veya aşırı şekilde yükleme yapmak | 1962 | 0,1 |
| Trafiği aksatacak şekilde yavaşlamak veya durmak | 2213 | 0,1 |
| Uygun olmayan veya uygunsuz şekilde ışık kullanmak | 220 | 0,0 |
| Yaya ve okul geçitlerinde yavaşlamamak, geçit hakkı vermemek | 14856 | 0,8 |
| Yolcu indirme ve bindirme kurallarına uymamak | 5603 | 0,4 |
| Trafik güvenliği ile ilgili diğer kurallara uymamak | 43453 | 2,6 |
| Diğer | 64392 | 3,8 |
| **Toplam** | **2775700** | **100** |

Tablo 3.9 incelendiğinde; kazaların %37,7'lik kısmının sürücülerin araç hızlarını hava yol ve trafiğin gerektirdiği şartlara uydurmadan kullandıkları görülmektedir. Bunu takiben kavşak geçit ve kaplamanın dar olduğu yerlerde geçiş önceliğine uymamaktan kaynaklandığı görülmektedir.

Kazaya karışan sürücülerin yaşları Tablo 3.10'da verilmiştir.

**Tablo 3.10.** Kazaya Karışan Sürücülerin Yaş Grupları

| Yaş Grupları | Frekans | Yüzde |
|---|---|---|
| 1-14 yaş arası | 28747 | 1,0 |
| 15-19 yaş arası | 166215 | 6,0 |
| 20-24 yaş arası | 350054 | 12,6 |
| 25-29 yaş arası | 360132 | 13,0 |
| 30-34 yaş arası | 331461 | 11,9 |
| 35-39 yaş arası | 300911 | 10,8 |
| 40-44 yaş arası | 256556 | 9,2 |
| 45-49 yaş arası | 206323 | 7,4 |
| 50-54 yaş arası | 166148 | 6,0 |
| 55-59 arası yaş | 126299 | 4,6 |
| 60-99 yaş arası | 85820 | 3,1 |
| Toplam | 2378666 | 85,7 |
| Eksik Gözlem | 397034 | 14,3 |
| **Toplam** | **2775700** | **100** |

Tablo 3.10 incelendiğinde; kazaya karışanların yaş dağılımına bakıldığında sürücülerin çoğunluğunun 20-44 yaşları arasında dağıldığı ağırlıklı olarak 20-29 yaşlar arasındaki genç yaş sürücülerin daha fazla ölümlü ya da yaralamalı kazaya karıştığı görülmektedir. Bu durumda genç sürücülerin trafik ve yol durumunun gerektirdiği ölçüde araçlarını genel itibari ile kullanmadıkları görülmektedir.

## 3.6. CART ALGORİTMASI İLE ELDE EDİLEN BULGULAR

Python uygulaması kullanılarak CART algoritması yardımıyla 2013-2022 yılları arasında Ülkemizde meydana gelen yaralamalı ya da ölümlü kazaları etkileyen faktörler sınıflandırılmak istenmiştir. Modelde bağımlı değişken olarak trafik kaza sonucu; ölümlü olması durumunda "1" yaralamalı olması durumu "0" olarak kodlanmıştır. Bağımsız değişken olarak ta modele alınan değişkenler sırasıyla; kaza yerleşim yeri, yol tipi, hava ve gün durumu, yolun yüzeyi, kaza oluşumu, yaş, yolcu ve sürücü kusuru, aracın cinsi ve cinsiyettir. Çalışmada yaya ve sürücüler için elde edilen CART algoritmasına ait karmaşıklık matrisi Tablo 3.11'de verilmiştir.

**Tablo 3.11.** CART Algoritmasına Ait Karmaşıklık Matrisi

| Gözlem | Tahmin Edilen Sınıf | | | |
|---|---|---|---|---|
| | Yaralamalı | Ölümlü | Toplam | Doğruluk oranı (%) |
| Yaya | 1495985 | 16669 | 1512654 | 98,90 |
| Sürücü | 1340045 | 172609 | 1512654 | 88,57 |

Tablo incelendiğinde 1512654 kaza verisi içerisinde oluşturulan model ile kaza sonucunda yayaların yaralanması %98,90 ile sürücülerin yaralanması ise %88,57 doğru olarak tahmin edilmiştir. Bu durumda oluşturulan modelin genel olarak doğru sınıflandırma oranı %93,73 olduğu söylenebilir.

CART algoritmasında anlamsız olarak tahmin edilen bağımsız değişkenler çıkarılarak tekrar karmaşıklık matrisi hesaplanmıştır. Oluşturulan karmaşıklık matrisi Tablo 3.12'de gösterilmiştir.

**Tablo 3.12.** CART Algoritmasına Ait Karmaşıklık Matrisi (Anlamlı Bulunan Değişkenler İçin)

| Gözlem | Tahmin Edilen Sınıf | | | |
|---|---|---|---|---|
| | Yaralamalı | Ölümlü | Toplam | Doğruluk oranı (%) |
| Yaya | 1101758 | 14085 | 1115843 | 98,74 |
| Sürücü | 1106858 | 8985 | 1115843 | 99,19 |

Tablo incelendiğinde anlamsız değerler modelden çıkarılınca sürücü doğruluk oranında net bir artış görüldüğü, bu durumda modelin genel olarak sınıflandırma oranı %98,96 olduğu söylenebilir.

## 3.7. LOJİSTİK REGRESYON ANALİZİ İLE ELDE EDİLEN BULGULAR

Bu aşamada 2013-2022 yılları arasında Ülkemizde meydana gelen trafik kazalarının ölümlü ya da yaralamalı olma durumunu etkileyen faktörler; Python uygulaması kullanılarak Lojistik regresyon analizi ile sınıflandırılmaya çalışılmıştır. Lojistik regresyon analizine ait karmaşıklık matrisi Tablo 3.12'de verilmiştir.

**Tablo 3.13.** Lojistik Regresyona Analizine Ait Karmaşıklık Matrisi

| Gözlem | Tahmin Edilen Sınıf | | | |
|---|---|---|---|---|
| | Yaralamalı | Ölümlü | Toplam | Doğruluk oranı (%) |
| Yaya | 1500458 | 12196 | 1512654 | 99,20 |
| Sürücü | 929958 | 582696 | 1512654 | 61,75 |

Tablo incelendiğinde oluşturulan model ile kaza sonucunda yayaların yaralanması %99,20 ile sürücülerin yaralanması ise %61,75 doğru tahmin edilmiştir. Bu durumda oluşturulan modelin genel olarak doğru sınıflandırma oranı %80,47 olduğu söylenebilir.

Lojistik regresyon analizinde anlamsız olan bağımsız değişkenler çıkarılarak tekrar karmaşıklık matrisi hesaplanmıştır. Oluşturulan karmaşıklık matrisi Tablo 3.14'de gösterilmiştir.

**Tablo 3.14.** Lojistik Regresyon Analizine Ait Karmaşıklık Matrisi (Anlamlı Olan Değişkenler İçin)

| Gözlem | Tahmin Edilen Sınıf | | | |
|---|---|---|---|---|
| | Yaralamalı | Ölümlü | Toplam | Doğruluk oranı (%) |
| Yaya | 1239010 | 9346 | 1248356 | 99,19 |
| Sürücü | 1010288 | 238068 | 1248356 | 80,92 |

Tablo incelendiğinde anlamsız değişkenler modelden çıkarıldığında sürücü doğruluk oranında net bir artış görüldüğü, bu durumda modelin genel olarak sınıflandırma oranının %90,06 olduğu söylenebilir.

Meydana gelen trafik kazalarının ölümlü mü yoksa yaralamalı mı olduğunu etkileyen bağımsız değişkenlere ait Lojistik regresyon analizine ait sonuçlar aşağıdaki tablolarda gösterilmiştir.

**Tablo 3.15.** Lojistik Regresyon Analizine Ait Tablolar

Yolun tipini gösterir tablo

| Yolun Tipi | B | S.E. | Wald | s.d. | p | Exp(B) |
|---|---|---|---|---|---|---|
| Diğer | | | 76,549 | 3 | 0,000 | |
| Bölünmüş yol | -,352 | ,086 | 16,573 | 1 | 0,000 | ,703 |
| Tek yönlü yol | -,674 | ,095 | 50,864 | 1 | 0,000 | ,510 |
| İki yönlü yol | -,404 | ,087 | 21,696 | 1 | 0,000 | ,668 |

Yolun türünü gösterir tablo

| Yolun Türü | B | S.E. | Wald | s.d. | p | Exp(B) |
|---|---|---|---|---|---|---|
| Toprak | | | 231,776 | 5 | 0,000 | |
| Asfalt | -1,367 | 0,105 | 168,881 | 1 | 0,000 | 0,255 |
| Sathi Kaplama | -1,255 | 0,135 | 86,799 | 1 | 0,000 | 0,285 |
| Beton | -1,354 | 0,190 | 50,539 | 1 | 0,000 | 0,258 |
| Parke | -1,688 | 0,122 | 190,990 | 1 | 0,000 | 0,185 |
| Stabilize | -0,608 | 0,147 | 17,151 | 1 | 0,000 | 0,544 |

Yolun sınıfını gösterir tablo

| Yolun Sınıfı | B | S.E. | Wald | s.d. | p | Exp(B) |
|---|---|---|---|---|---|---|
| Diğer | | | 9989,263 | 12 | 0,000 | |
| Cadde | -0,566 | 0,071 | 62,754 | 1 | 0,000 | 0,568 |
| Sokak | -0,955 | 0,084 | 130,637 | 1 | 0,000 | 0,385 |
| Otoyol | 1,792 | 0,075 | 567,811 | 1 | 0,000 | 6,000 |
| Devlet yolu | 1,385 | 0,071 | 378,120 | 1 | 0,000 | 3,997 |
| İl yolu | 1,003 | 0,089 | 127,802 | 1 | 0,000 | 2,727 |
| Köy yolu | 0,623 | 0,116 | 29,008 | 1 | 0,000 | 1,864 |
| Orman yolu | 0,705 | 0,309 | 5,191 | 1 | 0,023 | 2,023 |
| Bağlantı yolu | 0,863 | 0,119 | 52,982 | 1 | 0,000 | 2,370 |
| Tesis (mülk) Önü veya İçi | -0,524 | 0,222 | 5,575 | 1 | 0,018 | 0,592 |

**Tablo 3.15.** (Devamı)

Gün durumunu gösterir tablo

| Gün Durumu | B | S.E. | Wald | s.d. | p | Exp(B) |
|---|---|---|---|---|---|---|
| Alacakaranlık | | | 1069,515 | 2 | 0,000 | |
| Gündüz | -0,510 | 0,043 | 138,368 | 1 | 0,000 | 0,600 |
| Gece | 0,041 | 0,044 | ,902 | 1 | 0,342 | 1,042 |

Hava durumunu gösterir tablo

| Hava Durumu | B | S.E. | Wald | s.d. | p | Exp(B) |
|---|---|---|---|---|---|---|
| Bulutlu | | | 22,123 | 9 | 0,008 | |
| Açık | -0,050 | 0,066 | 0,585 | 1 | 0,444 | 0,951 |
| Sis/Duman | 0,174 | 0,097 | 3,239 | 1 | 0,072 | 1,190 |
| Yağmur | -0,133 | 0,074 | 3,207 | 1 | 0,073 | 0,876 |
| Kar | -0,318 | 0,126 | 6,368 | 1 | 0,012 | 0,728 |
| Sulu Sepken | -0,093 | 0,182 | 0,258 | 1 | 0,611 | 0,912 |
| Dolu | 0,165 | 0,318 | 0,271 | 1 | 0,602 | 1,180 |
| Tipi | 0,002 | 0,275 | 0,000 | 1 | 0,996 | 1,002 |
| Kuvvetli Rüzgar | 0,011 | 0,223 | 0,002 | 1 | 0,962 | 1,011 |
| Toz/Kum Fırtınası | 0,164 | 0,592 | 0,076 | 1 | 0,782 | 1,178 |

Sürücü yaş durumunu gösterir tablo

| Yaş Grupları | B | S.E. | Wald | s.d. | p | Exp(B) |
|---|---|---|---|---|---|---|
| 60-99 | | | 14,261 | 10 | 0,161 | |
| 1-14 | -0,121 | 0,085 | 2,058 | 1 | 0,151 | 0,886 |
| 15-19 | -0,121 | 0,051 | 5,603 | 1 | 0,018 | 0,886 |
| 20-24 | -0,118 | 0,046 | 6,645 | 1 | 0,010 | 0,889 |
| 25-29 | -0,094 | 0,046 | 4,273 | 1 | 0,039 | 0,910 |
| 30-34 | -0,085 | 0,046 | 3,405 | 1 | 0,065 | 0,918 |
| 35-39 | -0,082 | 0,046 | 3,128 | 1 | 0,077 | 0,921 |
| 40-44 | -0,040 | 0,047 | 0,740 | 1 | 0,390 | 0,960 |
| 45-49 | -0,117 | 0,049 | 5,705 | 1 | 0,017 | 0,889 |
| 50-54 | -0,082 | 0,051 | 2,595 | 1 | 0,107 | 0,921 |
| 55-59 | -0,128 | 0,054 | 5,730 | 1 | 0,017 | 0,880 |

Meydana gelen kazaların yaralamalı veya ölümlü olma durumları SPSS İstatistik paket programında incelenmiştir. Elde edilen sonuçlar ışığında meydana gelen kazaları yaralamalı veya ölümlü olma durumlarını etkilen değişkenler yolun tipi, yolun türü, yolun sınıfı, gün durumu, hava durumu ve sürücü yaşının anlamlı sonuçlar verdiği görülmektedir.

Tablolar incelendiğinde meydana gelen kazaların tek yönlü yolda meydana gelmesi durumunda oluşan kazanın yaralamalı olma durumu 0,674 oranında daha düşük olduğu, parke türü yolda meydana gelen trafik kazasının yaralamalı olma durumu diğer yol kaplamalarına göre yaralamalı olma durumu 1,688 oranında daha düşük olduğu, otoyollarda meydana gelen kazanın diğer yollara göre yaralamalı olma durumu 1,792 oranında daha fazla olduğu, gündüz meydana gelen kazanın yaralamalı olma durumunu 0,510 oranında azalttığı, karlı havalarda meydana gelen kazanın yaralamalı olma durumunu diğer hava şartlarına göre 0,318 kat azalttığı ve kazaya karışan 55-59 yaş aralığında olan sürücülerin diğer yaş gruplarına göre yaralamalı olma durumunu 0,128 oranında düşük olduğu görülmüştür.

# SONUÇ

Bu çalışmada amaç lojistik regresyon analizi ile CART algoritmasının karşılaştırılmasıdır. Bu karşılaştırmanın yapılabilmesi amacıyla Emniyet Genel Müdürlüğü Trafik Daire Başkanlığından alınan 2013-2022 yılları arasında Ülkemizde meydana gelen trafik kaza istatistikleri kullanılmıştır.

İncelenen dönem içerisinde ülkemizde meydana gelen 1512654 ölümlü ya da yaralamalı kaza verileri üzerinde çalışılmıştır. İl bazında bakıldığında İstanbul ili kaza oranı olarak %11'lik bir pay ile diğer illere göre daha çok kazanın yapıldığı il olarak görülmektedir. İstanbul ilini takiben diğer büyük şehirlerimizden olan Ankara ve İzmir takip etmektedir. En az kaza oranları incelendiğinde ise Tunceli, Bayburt ve Ardahan illerinin olduğu tespit edilmiştir.

Yine aynı yıllar arasındaki değerler incelendiğinde; en çok kazanın insan hareketliliğinin en çok olduğu yaz aylarında olduğu tespit edilmiştir. Kazaları etkileyen birçok etken olabileceği dikkate alınarak sürücülerden kaynaklı kusurlara bakıldığı zaman ülkemizde en çok araç hızlarını yol hava ve trafiğin şartlarına uydurmamadan kaynaklı olduğu görülmektedir. Sürücü kusurları haricinde yolun tipi, sınıfı, kaplaması, yüzeyi, hava durumu ve gün durumunun da etkili olabileceği düşünüldüğünde meydana gelen kazaların %57,4 ünün bölünmüş yollarda, %66,7 sinin ise gündüz ve %85,5 inde ise kuru yüzeyde gerçekleştiği görülmektedir. Buradan hareketle ülkemizde meydana gelen ölümlü ya da yaralamalı kazalar trafiğin gerektirdiği şartlarda hızını ayarlamayan sürücülerin dikkatsiz ve aşırı özgüvenli hareket ettiklerini göstermektedir.

Trafik kaza verilerinin bazı değişkenlerin benzer veya birbiri ile ilişkili olmasından dolayı korelasyon matrisine bakılarak değişkenler azaltılmıştır. 209.407 gözlem ile yaya kaza sonucu ve sürücü kaza sonucu durumu iki farklı hedef değişken üzerinden algoritmalar çalıştırılmıştır. Algoritmaların her ikisinde de eğitim için kullanıma ayrılan veri seti %75, test için ise %25'tir.

Bağımsız değişkenlerin tamamı kullanılarak CART algoritması ile hedef değişkenlere göre doğruluk oranları; yaya kaza sonucu durumu için %98,901, sürücü kaza sonucu durumu için %88, 57 olarak hesaplanmıştır. Lojistik regresyon analizi için hedef değişkenlere göre doğruluk oranları; yaya sonuç durumu için %99,20 sürücü durumu için ise %61,75 olarak hesaplanmıştır. Bağımsız değişkenler içerisinde anlamsız olarak tespit

edilenler çıkarıldığında ise CART algoritması ile hedef değişkenlere göre doğruluk oranları; yaya kaza sonucu durumu için %98,74, sürücü kaza sonucu durumu için %88,51 olarak hesaplanmıştır. Lojistik regresyon analizi için hedef değişkenlere göre doğruluk oranları; yaya sonuç durumu için %99,19 sürücü durumu için ise %80,92 olarak hesaplanmıştır. Bu sonuçlar ışığında CART algoritmasının sınıflama gücünün lojistik regresyon analizine göre daha doğru sınıflandırdığı görülmektedir.

Yapılan literatür çalışmalarında Bartfay ve arkadaşları (2006) yapay sinir ağları ve lojistik regresyon analizini kullanarak sınıflandırma oranlarını karşılaştırmış ve lojistik regresyon analizinin sınıflandırma oranının %65 yapay sinir ağlarının ise %67 olduğunu hesaplamıştır. Kuyucu (2012) ise yaptığı çalışmada yapay sinir ağlarının sınıflandırma oranının %87,29 ve AUC değerinin 0,929 olduğu lojistik regresyon analizinin sınıflandırması ise %81,78, AUC değerinin ise 0,828 olarak hesaplanmış ve CART algoritmasının lojistik regresyon analizinin sınıflandırmasına göre daha doğru olduğu görülmüştür. Güner (2014) yaptığı çalışmada ise CART algoritmasının sınıflandırma oranını %92,31 ve lojistik regresyon analizinin ise sınıflandırma oranını %91,36 olarak tespit etmiştir.

Genel itibariyle yapılan çalışmalarda az bir farkla da olsa CART algoritmasının sınıflandırma açısından daha doğru sınıflandırdığı görülmektedir. Bu çalışmada CART algoritmasının yaya sonuç durumu ile lojistik regresyon analizinin yaya sonuç durumu arasında belirgin bir farkın olmadığı fakat sürücünün kaza sonucu durumuna bakıldığı zaman CART algoritması daha doğru sınıflandırma yaptığı tespit edilmiştir. Bu durumda CART algoritmasının sınıflandırma çalışmalarında kullanılmasının daha uygun olacağı söylenebilir.

## KAYNAKLAR

A. M. Elsayad and Elsalamony, H. A. (2013). "Diagnosis of Breast Cancer Using Decision Tree Models and SVM". *International Journal of Computer Applications*, 83(5), 19-29.

Adeyemo, A. (2018). "Effects of Normalization Techniques on Logistic Regression in Data Science", *Proceedings of the Conference on Information Systems Applied Research Norfolk*, Virginia, 11, 4813.

Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., & Honrao, V. (2013). "Predicting Students' Performance Using ID3 and C4.5 Classification Algorithms". *International Journal of Data Mining & Knowledge Management Process, 3*(5). 39-52.

Ağlarcı, A. V. (2022). *Kronik Hepatit c Hastalığı Risk Belirlenmesinde Sıralı Lojistik Regresyon ve Makine Öğrenme Algoritmalarının Sınıflama Performanslarının Karşılaştırılması*. Doktora Tezi. Eskişehir: Osmangazi Üniversitesi Sağlık Bilimleri Enstitüsü.

Aitkenhead M. J. (2008). "A Co-Evolving Decision Tree Classification". *Expert Systems With Applications*, 34(1), 18–25.

Akküçük, U. (2011). *Veri Madenciliği Kümeleme ve Sınıflama Algoritmaları*. İstanbul: Yalın Yayıncılık.

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

Alanlar Ezgi (2021). *Pazar Sepeti Analizi İle Birliktelik Kurallarının Belirlenmesi: Perakende Sektöründe covid-19* Etkisi. (Yayımlanmamış Yüksek lisans Tezi), Karabük: Karabük Üniversitesi, Lisansüstü Eğitim Enstitüsü.

Albayrak, A. S. (2006). *Uygulamalı Çok Değişkenli İstatistik Teknikleri*. Ankara: Asil Yayın Dağıtım.

Allison, P. (2001). *Logistic Regression Using the SAS System: Theory and Application*. USA: SAS Institute Inc.

Alpar, R. (2017). *Uygulamalı Çok Değişkenli İstatiksel Yöntemler*. Ankara: Detay Yayıncılık.

Astar, M. (2009). *OECD Ülkelerinde Taylor Kural'ının Geçerliliğinin Logit Modelleri İle İncelenmesi.* (Yayınlanmış Yüksek Lisans Tezi), İstanbul: Marmara Üniversitesi, Sosyal Bilimler Enstitüsü.

Ayhan S. (2006) *Sıralı Lojistik Regresyon Analiziyle Türkiye'deki hemşirelerin iş bırakma niyetini etkileyen faktörlerin belirlenmesi*. Yüksek Lisans Tezi. Eskişehir: Osmangazi Üniversitesi, Fen Bilimleri Enstitüsü.

Barros, R., Carvalho, A., & Freitas, A. (2015). *Automatic Design of Decision-Tree Induction Algorithms.* New-York: Springer Briefs in Computer Science.

Bartfay E., Macklillop W., & Pater L. J. (2006). "Comparing the Predictive Value of Neural Network Models to Logistic Regression Models on the Risk of Death for Small-Cell Lung Cancer Patients", *European Journal of Cancer Care, 15*, 115–124.

Bayhan, Y. (2021). *Sınıflandırma ve Regresyon Ağaçları (CART) Algoritması: Borsa İstanbul Üzerine Bir Uygulama.* Yüksek Lisans Tezi, Erzurum: Atatürk Üniversitesi Sosyal Bilimler Enstitüsü.

Bayram, N. (2015). *Sosyal Bilimlerde SPSS İle Veri Analizi*. Bursa: Ezgi Kitapevi.

Berry, M. J. A., & Linoff, G. S. (2004). *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management (Second Edition).* Indianapolis, Indiana: Wiley Publishing Inc.

Bhardwaj, R., & Vatta, S. (2013). "Implementation of ID3 Algorithm". *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(6), 845-851.

Cabena Peter, (1998). *Discovering Data Mining: From Consept to Implementation*. USA: International Business Machines Corporation.

Clay Helberg, (2022), Data Mining With Confidence. England: SPSS.

Cook, D., Dixon, P., Duckworth, WM., Kaiser, M.S., Koehler, K., Meeker, W.Q., Stephenson W.R. (2001). *Binary Response and Logistic Regression Analysis*.

Coşkun M. and H. İ. Bülbül (2019). "Hane Halkı İnternet Hizmeti Sahipliğini Etkileyen Faktörlerin Karar Ağaçlari İle İncelenmesi". *Türk Bilim Araştırma Vakfı*, 12(2), 1-17.

Çalış, A., Kayapınar S. ve Çetinyokuş T. (2019). "Veri Madenciliğinde Karar Ağacı Algoritmaları ile Bilgisayar ve İnternet Güvenliği Üzerine Bir Uygulama", *Gazi Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği Dergisi*, 25, 3-4.

Çatal, B. (2021). *Lojistik Regresyon Analizi ve Sağlık Alanında Bir Uygulama*, Yüksel Lisans Tezi. Isparta: Süleyman Demirel Üniversitesi, Sağlık Bilimleri Enstitüsü.

Çelik, M. (2009). *Veri Madenciliğinde Kullanılan Sınıflandırma Yöntemleri ve Bir Uygulama*, Yüksek Lisans Tezi. İstanbul: İstanbul Üniversitesi Sosyal Bilimler Enstitüsü.

Dahan, H., Cohen, S., Rokach, L., & Maimon, O. (2014). *Proactive Data Mining with Decision Trees*. London: Springer Briefs in Electrical and Computer Engineering.

Deconinck, E., Hancock, T., Coomans, D., Massart, D.L., Heyden, Y.V. (2005). "Classification of Drugs in Absorption Classes Using the Classification and Regression Trees (CART) Methodology", *Journal of Pharmaceutical and Biomedical Analysis*, 39, 91–103.

Denison, D., Mallick, B., & Smith, A. (1998). "A Bayesian CART Algorithm". *Biometrika*, 85(2), 363-367.

Dunham, M. H. (2003). *Data Mining Introductory and Advanced Topics*. New Jersey: Pearson Education.

Ekici, E. (2012). *Farklı Sınıflandırma Yöntemlerinin Karşılaştırılması ve Bir Uygulama*, Yüksek Lisans Tezi. Elâzığ: Fırat Üniversitesi.

Ersöz, F. (2019). *Veri Madenciliği Teknikleri ve Uygulamaları*. Ankara: Seçkin Kitabevi.

Esen, M .F. (2009). *Veri tabanlarında Bilgi Keşfi: Veri Madenciliği ve Bir Sağlık Uygulaması*, Yüksek Lisans Tezi, İstanbul: İstanbul Üniversitesi Sosyal Bilimler Enstitüsü.

Esenyel İçen, N. M. (2022). *Bayesyen Multinominal Lojistik Regresyon İle İşletmelerin Finansal Başarısızlığının Değerlendirilmesi: İmalat Sanayi Uygulaması*, Doktora Tezi, İstanbul: İstanbul Üniversitesi Sosyal Bilimler Enstitüsü.

Fowdar, J., Bandar, Z., & Crockett, K. (2004). "Inducing Fuzzy Decision Trees in Non-Deterministic Domains using CHAID". *FLAIRS Conference* Manchester: American Association for Artificial Intelligence.

Guha S., Rastogi R., Shim K. (1998). "CURE: An Efficient Clustering Algorithm for Large Databases", *Newsletter ACM SIGMOD Record*, 27(2), 73-84.

Güner, Z. B. (2014). "Veri Madenciliğinde CART ve Lojistik Regresyon Analizinin Yeri: İlaç Provizyon Sistemi Verileri Üzerine Bir Uygulama". *Sosyal Güvence Dergisi*, (6), 53-99.

Hadi, A.S., Imon, R.A.H.M. (2008). "Identification of Multiple Outliers in Logistic Regression". *Communications in Statistics Theory and Methods*, 37, 1697–1709.

Hakan Dalkılıç, Feriştah Dalkılıç (2015). *Akademik Bilişim. https://ab.org.tr/ab15/kitap/318.pdf*

Han, J., Kamber, M. & Pei, J. (2011). *Data Mining: Concepts and Techniques.* Burlington: Elsevier Science.

Hatip Ünlü, E. (2022). *Major Yanıklı Kritik Hastalardaki Ürünleri Transfüzyonunun Hasta Sonuçlarına Etkisinin Ridge lojistik Regresyon modeli İle Değerlendirilmesi. Yüksek Lisans Tezi.* İstanbul: İstanbul Üniversitesi Lisansüstü Eğitim Enstitüsü.

Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). "A Comparative Study of Decision Tree ID3 and C4.5". *International Journal of Advanced Computer Science and Applications*, 4(2), 13-19.

Irmak, S., & Ercan, U. (2017). "Karar Ağaçları Kullanarak Türkiye Hane halkı Zeytinyağı Tüketimi Görünümünün Belirlenmesi". *Journal of Management Economics and Business*, 13(3). 553-564

James, G., Witten, D., Hastie, T., & Robert, T. (2013). *An Introduction to Statistical Learning*. New York: Springer Science.

Kahraman, F. (2021). *Moralite Skorlamalarının Koroner Yoğun Bakım Hastalarında Lojistik Regresyon Analizi İle Değerlendirilmesi.* Yüksek Lisans Tezi. Isparta: Süleyman Demirel Üniversitesi Sağlık Bilimleri Enstitüsü.

Kalaycı, Ş. (2014). *SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri*. Ankara: Asil Yayın Dağıtım.

Karagöz, Y. (2010). "Nonparametrik Yöntemlerin Güç ve Etkinlikleri". *Elektronik Sosyal Bilimler Dergisi*, 9(33), 18-40.

Karakış, R. (2009). *Yapay Sinir Ağları ve Lojistik Regresyon Yöntemleri ile Meme Kanseri Koltuk Altı Lenf Durumunun Belirlenmesi*. Yüksek Lisans Tezi. Ankara: Gazi Üniversitesi Bilişim Enstitüsü.

Kass, G. V. (1980). "An Exploratory Technique for Investigating Large Quantities of Categorical Data". *Applied Statistics, 29*(2), 119-127.

Kayaalp, G., Çelik Güney, M. ve Cebece, Z. (2015). "Çoklu Doğrusal Regresyon Modelinde Değişken Seçiminin Zootekniye Uygulanışı". *Çukurova Üniversitesi Ziraat Fakültesi Dergisi,* 30(1), 1-8.

Kayri, M., & Boysan, M. (2007). "Araştırmalarda CHAİD Analizinin Kullanımı ve Baş Etme Stratejileri İle İlgili Bir Uygulama". *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, 40(2), 133-149.

Kayri, M., & Kayri, İ. (2015). "The Comparison of Gini and Twoing Algorithms in Terms of Predictive Ability and Misclassification Cost in Data Mining: An Empirical Study". *International Journal of Computer Trends and Technology (IJCTT),* 27(1), 21-30.

Kılıçalan, M.B. (2018). *Hanehalkı İşgücü Araştırma Verileri İle Veri Madenciliği Yöntemlerinin Uygulanması ve Modellerin Karşılaştırılması.* Yüksek Lisans Tezi. Ankara: Hacettepe Üniversitesi, İstatistik Anabilim Dalı.

Kleinbaum, DG, Klein, M. (2002). *Logistic Regression: A Self-Learning Text*. 2nd Ed., New York: Springer.

Kleppin, L., Roland, P., & Winfried, S. (2008). "CHAID models on boundary conditions of metal accumulation in mosses collected in Germany in 1990, 1995 and 2000". *Atmospheric Environment Elsevier*, 42, 5222.

Klienbaum, D. G. ve Klein, M. (2002). Springer-Verlag New York Inc, 3. Baskı, s.5 tarafından yayımlanmış olan "*Logistic Regression: A Self Learning Text*".

Köktürk F., Ankaralı H., Sümbüloğlu V. (2009). "Veri Madenciliği Yöntemlerine Genel Bakış", *Türkiye Klinikleri J Biostat*, 1(1), 20-25.

Kretowski, M. (2019). *Evolutionary Decision Trees in Large-Scale Data Mining*. Switzerland: Springer.

Kuru, Ö. (2019). *Türkiye'deki Mevduat Bankalarının Etkinliklerinin Veri Zarflama Analizi (VZA). C5.0 ve CART Algoritması İle İncelenmesi.* Yüksek Lisans Tezi. Osmaniye: Osmaniye Korkut Ata Üniversitesi Sosyal Bilimler Enstitüsü.

Kuyucu, Y. E. (2012). *Lojistik Regresyon Analizi (LRA). Yapay Sinir Ağları (YSA) ve Sınıflandırma ve Regresyon Ağaçları (CART) Yöntemlerinin Karşılaştırılması ve Tıp Alanında Bir Uygulama*, Yüksek Lisans Tezi. Tokat: Gaziosmanpaşa Üniversitesi Sağlık Bilimleri Enstitüsü.

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. Hoboken: John Wiley & Sons.

Loh, W. Y. and Y. S. Shih (1997). "Shih, Split Selection Methods for Classification Trees", Statis*tica Sinica*, 815-40.

Lowe, B., & Kulkarni, A. (2015). "Multispectral Image Analysis Using Random Forest". *International Journal on Soft Computing (IJSC),* 6(1), 1-14.

McCulloch, W. S. and Pitts, W. (1943). "A Logical Calculus of the İdeas İmmanent in Nervous Activity", *Bull. Math. Biophys*, 5(4), 115-33, https://doi.org/10.1007/bf02478259

Menard, S. (2002). *Applied Logistic Regression Analysis*. 2nd Edition, London: SAGE Publications.

Montgomery, D.C., Peck, A.E., Vining, G.G. (2001). *Introduction to the Linear Regression Analysis.* USA: John Willey&Sons.

Öz, Kara (2015). *Lojistik Regresyon Analizi ve Kadın İşgücü Üzerine Bir Uygulama*. Yükek Lisans Tezi. Bursa: Uludağ Üniversitesi Sosyal Bilimler Enstitüsü, Ekonometri Anabilim Dalı.

Özkan, Y. (2008). *Veri Madenciliği Yöntemleri.* İstanbul: Papatya Yayıncılık Eğitim.

Özsürünç, R. (2022). *Veri Madenciliğinde Lojistik Regresyon Modellerinin İncelenmesi.* Doktora Tezi. İstanbul: Üniversitesi Sosyal Bilimler Enstitüsü.

Öztürk Eren (2022). Patika Akademi, *Patika Akademi*, *https://academy.patika.dev/blogs/detail/yapay-sinir-aglariys* adresinden alındı.

Quinlan, J. R. (1992). *C4.5: Programs for Machine Learning.* San Mateo, California: Morgan Kaufmann Publishers.

Roman, T. (2004). *Classification and Regression Trees (CART) Theory and Applications*. Berlin: Humboldt University.

Sevimli Saitoğlu, Y. (2015). *Sınıflama ve Regresyon Ağaçları*. (Yayımlanmamış Doktora Tezi), İstanbul: Marmara Üniversitesi Sosyal Bilimler Enstitüsü.

Shafique, U., Qaiser, H. (2014). "A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)". *International Journal of Innovation and Scientific Research*, 12(1), 219.

Sharma, R., Ghosh, A., & Joshi, P. (2013). "Decision Tree Approach for Classification of Remotely Sensed Satellite Data Using Open Source Support". *Journal of Earth System Science*, 122(5), 1237-1247.

Silahtaroğlu, G. (2020). *Veri Madenciliği Kavram ve Algoritmaları.* İstanbul: Papatya Yayıncılık.

Simar, W. H. (2007). *Leopold, "Applied Multivariate Statistical Analysis*". Second Edition, Berlin Almanya: Springer.

Singh, S., Gupta, P. (2014). "Comparative Study ID3, CART and C4.5 Decision Tree Algorithm: A Survey", *International Journal of Advanced Information Science and Technology*, 27(27), 97-103.

Speybroeck, N. (2012). "Classification and Regression Trees". *Int J Public Health*, 57(1), 243-246.

Strobl, C., Malley, J., & Tutz, G. (2009). "An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests". *NIH Public Access*, 14(4), 323-348.

Sun, J., and Li, H. (2008). "Data Mining Method for Listed Companies, Financial Distress Prediction", *Knowledge-Based Systems,* 21(1), 1–5.

Şatir E., Azboy, F., Aydin, A., Arslan, H. ve Haciefendioğlu, Ş. (2016). "Veri İndirgeme ve Sınıflandırma Teknikleri İle Glokom Hastalığı Teşhisi". *EL-Cezeri Journal of Science and Engineering*, 3(3), 485-97.

Şimşek Gürsoy, U. T. (2009). *Veri Madenciliği ve Bilgi Keşfi*. Ankara: Pegem Akademi.

Tabachnick, Barbara G., Fidell, Linda S. (1996). *Using Multivariate Statistics*. Harper Collins College Publishers.

Tan, P. N., Steinbach, M., Kumar, V. (2005). *Introduction to Data Mining*. Addison Wesley: Pearson International Edition.

Taşkın, S. (2022). *Edirne Turizminin Lojistik Regresyon İle Analizi.* Yüksek Lisans Tezi. Bursa: Uludağ Üniversitesi Sosyal Bilimler Enstitüsü.

Tatlıdil, H. (1996). *Uygulamalı Çok Değişkenli İstatistiksel Analiz*. Ankara: Cem Ofset.

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

Timofeev, R. (2004). *Classification and Regression Trees (CART) Theory and Applications.* (Master Thesis), Berlin: Humboldt University.

Uruk, E. (2007). *İstatistiksel Uygulamalarda Lojistik Regresyon Analizi.* (Yüksek Lisans Tezi), İstanbul: Marmara Üniversitesi Fen Bilimleri Enstitüsü.

Yasinoğlu, S. (2022). *Ergen Yaş Grubunu Etkileyen Sosyo-Kültürel Kriterlerin Multinominal Lojistik Regresyon Analizi İle İncelenmesi. Yüksek Lisans Tezi.* Kütahya: Dumlupınar Üniversitesi Lisansüstü Eğitim Enstitüsü.

Yaşasınoğlu, N. (2019) *Türkiye'de Anlaşmalı Boşanma İle Sosyal Güvenlik Kurumundan Aylık Alanlar Üzerine Veri Madenciliği CART Algoritması Uygulaması. Yüksek Lisans Tezi.* Ankara: Gazi Üniversitesi Fen Bilimleri Enstitüsü.

Yavuzkanat, M. (2011). *Lojstik Regresyonda Çoklu Aykırı Gözlemlerin Belirlenmesi ve Etkilerinin İncelenmesi.* Yüksek Lisans Tezi. Ankara: Gazi Üniversitesi Fen Bilimleri Enstitüsü.

Ye, N. (2003). *The Handbook of Data Mining*. New Jersey: Lawrence Erlbaum Associates, Publishers.

Yıldırım, B. (2019). *Modern Perakendecilik Sektöründe Veri Madenciliği ve Tekniklerinin Uygulanması*. (Yüksek Lisans Tezi). İstanbul: İstanbul Üniversitesi, Endüstri Mühendisliği Anabilim Dalı.

Yohannes, Y., & Webb, P. (1999*). Classification and Regression Trees, CART: a User Manual for Identifying Indicators of Vulnerability to Famine and Chronic Food Insecurity (Cilt vol 3).* Washington: International Food Policy Research Institut.

# ÖZGEÇMİŞ

| | |
|---|---|
| **Kişisel Bilgiler** | |
| Adı Soyadı | Ahmet TEKİN |
| Doğum Yeri ve Tarihi | |
| **Eğitim Durumu** | |
| Lisans Öğrenimi | |
| Y. Lisans Öğrenimi | |
| Bildiği Yabancı Diller | |
| Bilimsel Faaliyetleri | |
| **İş Deneyimi** | |
| Stajlar | |
| Projeler | |
| Çalıştığı Kurumlar | |
| **İletişim** | |
| E-Posta Adresi | |
| **Tarih** | |