# SIRALI LOJİSTİK REGRESYON VE SINIFLANDIRMA AĞAÇLARININ PERFORMANS KARŞILAŞTIRMASI

**Yüksek Lisans Tezi**

**Simay MİRGEN**

**Eskişehir 2019**

# SIRALI LOJİSTİK REGRESYON VE SINIFLANDIRMA AĞAÇLARININ PERFORMANS KARŞILAŞTIRMASI

**Simay MİRGEN**

**YÜKSEK LİSANS TEZİ**

**İstatistik Anabilim Dalı**

**Danışman: Doç. Dr. Betül KAN KILINÇ**

**Eskişehir**

**Eskişehir Teknik Üniversitesi**

**Lisansüstü Eğitim Enstitüsü**

**Temmuz 2019**

## JÜRİ VE ENSTİTÜ ONAYI

Simay Mirgen'in "Sıralı Lojistik Regresyon ve Sınıflandırma Ağaçlarının Performans Karşılaştırması" başlıklı tezi 31/07/2019 tarihinde aşağıdaki jüri tarafından değerlendirilerek "Eskişehir Teknik Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği'nin ilgili maddeleri uyarınca, İstatistik Anabilim dalında Yüksek Lisans Tezi olarak kabul edilmiştir.

| **Jüri Üyeleri** | **Ünvanı Adı Soyadı** | **İmza** |
|---|---|---|
| Üye (Tez Danışmanı) | : Doç. Dr. Betül KAN KILINÇ | ..... |
| Üye | : Prof. Dr. Berna YAZICI | ..... |
| Üye | : Prof. Dr. Özlem ALPU | ..... |

Prof. Dr. Murat TANIŞLI
Lisansüstü Eğitim Enstitü Müdürü

# ÖZET

SIRALI LOJİSTİK REGRESYON VE SINIFLANDIRMA AĞAÇLARININ PERFORMANS KARŞILAŞTIRMASI

Simay MİRGEN

İstatistik Anabilim Dalı
İstatistik Teorisi Bilim Dalı
Eskişehir Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, Temmuz,2019
Danışman: Doç. Dr. Betül KAN KILINÇ

İstatistiksel uygulamalarda, bağımlı ve bağımsız değişken arasındaki ilişkiyi tanımlamak için parametrik ve parametrik olmayan yöntemlere sıklıkla başvurulmaktadır. Bu çalışmada, bu yöntemlerden biri olan sınıflama ve regresyon ağaçları ile sıralı lojistik regresyonun sınıflama başarılarının karşılaştırılması amaçlanmıştır. Bu amaç doğrultusunda Eskişehir ilindeki satılık konutların birim fiyatını etkileyen faktörleri incelemek amacıyla bir uygulama yapılmıştır. Veri seti eğitim, doğrulama ve test olarak üçe ayrılmış, 5-katlı çapraz geçerlilik ile yöntemlerin performansları ortaya konmuştur. Elde edilen sonuçlar doğrultusunda, sınıflama ağaçları algoritmasıyla daha başarılı bir sınıflama yapıldığı saptanmıştır.

**Anahtar Sözcükler:** Sınıflandırma ağaçları, Sıralı lojistik regresyon, Çapraz geçerlilik

## ABSTRACT

PERFORMANCE COMPARISON BETWEEN ORDINAL LOGISTIC REGRESSION AND CLASSIFICATION TREE

Simay MİRGEN

Department of Statistics
Programme in Theory of Statistics
Eskişehir Technical University, Institute of Graduate Programs, July 2019
Supervisor: Assoc. Prof. Dr. Betül KAN KILINÇ

In statistical applications, parametrical and nonparametrical methods are often applied to identify the relationship between the dependent and independent variables. In this thesis, among these methods the comparison of the classification performance of the classification tree model and ordinal logistic regression model is purposed. For this aim, an application study is conducted to determine the factors effecting the unit price of real estate for sale in Eskisehir. The performance comparison of the methods is presented by **s**plitting dataset into three parts: training, validation and test using 5-fold cross validation. Based on the results, classification tree algorithm over performed the ordinal logistic regression model.



**Keywords:** Classification trees, Ordinal logistic regression, Cross validation

## TEŞEKKÜR

Yüksek lisansım süresince ve bu çalışmanın her aşamasında beni yönlendiren ve bilgilendiren, sonuçlanana kadar her türlü desteği ve anlayışı gösteren, kıymetli zamanını bana ayıran değerli hocam Doç. Dr. Betül KAN KILINÇ' a en içten dileklerimle teşekkür ederim. Ayrıca her zaman destekleyip güler yüzünü hiç eksik etmeyen değerli hocam Prof. Dr. Aladdin SHAMILOV' a teşekkürü borç bilirim.

Aynı zamanda iş yerim ile okulum arasında beni her yere yetiştiren babam Mehmet MİRGEN' e, her türlü yardımcım annem Sibel MİRGEN' e, desteğini esirgemeyen abim Öğ. Tğm. Çağatay MİRGEN' e, her zaman yardımda bulunan kuzenim Ebru Sarıışık Ünal' a çok teşekkür ederim.

Ayrıca iş yerimde izin konusunda yardımcı olan müdürüm Mutlu DEMİRCİ' ye, her zaman bana destek olup, gösterdiği sabır ve anlayış için yüksek lisansı bitirmemde büyük katkısı olan müdür yardımcım Şenay URSAVAŞ' a ve hep yanımda olduklarını hissettiren iş arkadaşlarıma bana her zaman inanan dostlarıma bütün samimiyetimle teşekkür ederim.

31/07/2019

## ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Bu tezin bana ait, özgün bir çalışma olduğunu; çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; bu çalışmanın Eskişehir Teknik Üniversitesi tarafından kullanılan "bilimsel intihal tespit programıyla tarandığını ve hiçbir şekilde "intihal içermediğini" beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

SİMAY MİRGEN

## İÇİNDEKİLER

## TABLOLAR DİZİNİ

**Sayfa**

## ŞEKİLLER DİZİNİ

**Sayfa**

## SİMGELER VE KISALTMALAR DİZİNİ

| | |
|---|---|
| CART | :Classification and Regression Tree (Sınıflandırma ve Regresyon Ağaçları) |
| GDM | : Genelleştirilmiş Doğrusal Modeller |
| OOM | : Orantısal Oran Modeli |
| SH | : Standart Hata |
| SLR | : Sıralı Lojistik Regresyon |

# 1. GİRİŞ

Günümüzde araştırmalara konu olan problemin çözümünde, problemi etkileyen birçok faktör vardır ve çözülecek problem için bu birçok faktörü incelemek gerekir. Bu nedenle karşılaşılan problemlerde birden fazla bağımsız değişkenin olduğu durumlar ortaya çıkmıştır. Bu gibi durumlarda verinin ortaya koyduğu değişkenliği, çok değişkenli istatistiksel analizlerin teknikler kullanarak açıklamak daha sağlıklı olabilir. Çok değişkenli istatistiklerin uygulanma amaçları arasında kümeleme ve sınıflama gibi kavramlar yer almaktadır. Verilerin sınıflandırılması için kümeleme analizi, diskriminant analizi, faktör analizi gibi yaygın teknikler kullanılabilir.

Sınıflama ve regresyon modelleri içinde, lojistik regresyon ve karar ağaçları yöntemleri sıklıkla tercih edilen yöntemler arasındadır. Klasik doğrusal regresyon modelindeki varsayımlar olmaksızın kategorik bağımlı değişkeni tahmin etmek için karar ağaçları kullanılabilirken, lojistik regresyon için bazı varsayımlar söz konusu olabilir.

Lojistik regresyon ile amaçlanan, en az değişkenle en iyi uyuma sahip olacak şekilde bağımlı ve bağımsız değişkenler arasındaki ilişkiyi tanımlayıp istatistiksel olarak anlamlı olarak kabul edilebilen bir model kurmaktır. Benzer bir ilişkiyi açıklamak için kullanılan karar ağaçları ise daha kolay anlaşılır ve daha rahat yorumlanabildiği için tercih edilir. Bu yöntemle, bağımsız değişkene sorulan soruların oluşturduğu cevaplar ağaç dalları şeklinde görülür ve bu sayede çok karmaşık veri seti, bağımlı değişkeni etkileyen değişkenler ve bu değişkenlerin modeldeki önemi görsel olarak incelenebilmektedir.

Lojistik regresyon modelinin kullanımı için yapılan ilk çalışmalar Berkson (1944) tarafından yapılmıştır. Model Finney (1972) tarafından biyolojik deneyler için probit analize alternatif olarak kullanılmıştır. Lojistik modele ilişkin çeşitli uygulamalar yapan Cox (1970) daha sonrasında Anderson (1979, 1983) tarafından gelişmeye devam etmiştir (Bircan, 2004). Lojistik modele uyum ile ilgili birçok çalışmalarda bulunan Aranda-Ordaz (1981) ve Johnson (1985) başlıca isimlerdendir. Daha sonra Lesaffre ve Albert (1989) ise çoklu grup lojistik modellerde etkin ve aykırı gözlemleri belirleme ölçütlerini incelemişlerdir.

Temel (2004), nöroloji bölümünde yapılan anket çalışması ile hastalığa sahip olup olmama durumuna etki eden değişkenleri sınıflandırma ağaçları analizi ile incelemiştir. Türkiye'deki hemşirelerin iş bırakma niyetini etkileyen faktörleri sıralı lojistik regresyon analizi ile Ayhan (2006) incelemiş ve iş bırakma niyetlerinin nelere bağlı olduğunu anlatmıştır. Sosyal güvenlik kurumu ilaç provizyon sistemi verileri üzerinde bir uygulama

yapan Kıran (2010) CART analizi ile lojistik regresyon analizini karşılaştırmıştır. Akın ve Şentürk (2012) bireylerin mutluluk düzeylerinin sıralı lojistik regresyon analizini yaparak mutluluk düzeyinin yıllara göre farklı çalışmalarda nasıl değiştiğini incelemişlerdir. Korkmaz (2018), kötü amaçlı yazılımların Network cihazında olup olmadığını, sınıflandırma ve regresyon ağaçları (CART) ve rastgele orman teknikleriyle inceleyerek sınıflandırma yapmıştır.

Bu çalışmada ise lojistik regresyonun bir çeşidi olan sıralı lojistik regresyon ile sınıflandırma ağaçları sonuçları değerlendirilerek, gerçek bir veri seti üzerinde uygulama yapılıp bu iki yöntemin sınıflama başarısının karşılaştırılması amaçlanmıştır.

Bu nedenle çalışmanın ikinci bölümünde, yöntemler hakkında genel bilgiler verilerek sıralı lojistik regresyon modeli ile sınıflandırma ve regresyon ağaçları tanıtılmış, karşılaştırma ölçülerinden olan karışıklık matrisi anlatılmıştır.

Çalışmanın üçüncü bölümü uygulama için düzenlenmiştir. Bu amaçla, satın alma ve satma işlerinde yaygın ve popüler bir platform olarak kullanılan bir internet sitesinden veriler toplanmıştır. Verinin tanıtılması ve sıralı lojistik regresyon algoritması ve sınıflandırma ağacı algoritması ile analizler ve sonuçlar gösterilmiştir. Her iki yöntem için kurulan modellerin sınıflandırma başarısı karşılaştırılmıştır.

Sonuç bölümde, her iki yöntemin performans başarısı için değerlendirmeler yapılmış ve yorumlanmıştır.

## 2. YÖNTEMLER

### 2.1. Genelleştirilmiş Doğrusal Modeller

Klasik doğrusal regresyon modelinde, normallik ve sabit varyanslılık sağlanmadığında veri dönüştürme teknikleri yerine genelleştirilmiş doğrusal model (GDM) yaklaşımı kullanılır (Montgomery, Peck ve Vining, 2013).

GDM, bağımlı değişkeni yanıt değişken, bağımsız değişkenleri ise açıklayıcı değişken olarak adlandırır (Karadağ, 2014).

GDM, yanıt değişkenin normal dağılım göstermediği durumlarda hem doğrusal hem de doğrusal olmayan regresyon modelleri için çözüm olarak geliştirilmiştir. Yanıt değişkeninin dağılımı; normal, Poisson, binom, üstel ve gamma dağılımlarını içeren üstel dağılım ailesidir (Montgomery, Peck ve Vining, 2013).

Üstel dağılım ailesinde; rastgele değişkenin olasılık veya olasılık yoğunluk fonksiyonu f(x;θ) şeklinde gösterilir. $\boldsymbol{\Theta}$ parametre kümesini göstermek üzere ($\boldsymbol{\Theta} \subset \mathbb{R}$) olasılık veya olasılık yoğunluk fonksiyonlarının ailesi,

$$\mathcal{C} = \{f(x;\theta): \theta \in \boldsymbol{\Theta}\} \tag{2.1}$$

ile gösterilir.

Her $f(x;\theta) \in \mathcal{C}$ için $f(x;\theta)$ fonksiyonu,

$$f(x;\theta) = h(x)c(\theta)\exp\left(\sum_{i=1}^{k} t_i(x)w_i(\theta)\right) \tag{2.2}$$

şeklinde yazılabiliyorsa, $\mathcal{C} = \{f(x;\theta): \theta \in \boldsymbol{\Theta}\}$ sınıfına bir üstel aile denir.

Burada, $h(x)$ ve i=1,2,3, … , k için $t_i(x)$' ler $x$ in bir fonksiyonu, $w_i(\theta)$' ler ve $c(\theta)$ ise sadece parametrenin bir fonksiyonudur. Ayrıca $h(x) > 0$ ve $c(\theta) > 0$ olmalıdır. Parametre tahmini ve istatistiki sonuç çıkarımlar açısından olasılık fonksiyonlarının üstel aileye ait olması önemli kolaylıklar sağlar (Akdi,2014).

GDM' lerin yapısı aşağıdaki gibi verilebilir:

$y_1, y_2, \ldots, y_n$ yanıt değişkeninin değerleri sırasıyla $\mu_1, \mu_2, \ldots, \mu_n$ ortalamalı birbirinden bağımsız gözlemler ve $y_i$ gözlemleri üstel aileden gelen bir dağılıma sahip olsun. $x_1, x_2, \ldots, x_k$ açıklayıcı değişkenlerini göstermek üzere,

$\eta_i = \mathbf{x'_i\beta} = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$ şeklinde model oluşturulur. Model bir bağlantı fonksiyonu (link function) yardımıyla bulunur.

$$\eta_i = g(\mu_i) = g(E(y_i)) \qquad i=1,2,\ldots,n \qquad \textbf{(2.3)}$$

Eşitlik (2.3)' te gösterilen bağlantı terimi, ortalama ve doğrusal tahmin edici arasındaki bağlayıcı fonksiyonu ifade eder. Bağlantı fonksiyonu türevlenebilen ve monoton bir fonksiyondur.

Yanıt değişkeninin beklenen değeri, $E(y_i) = g^{-1}(\eta_i) = g^{-1}(\mathbf{x_i'\beta})$' dir. Çoklu doğrusal regresyonda,

$$\mu_i = \eta_i = \mathbf{x_i'\beta} \qquad i=1,2,\ldots,n \qquad \textbf{(2.4)}$$

modeli özel bir durum göstermektedir. Bu durum $g(\mu_i) = \mu_i$ şeklindeki birim bağlantı fonksiyonudur (Nelder ve Wedderburn, 1972; Yarbaşı,2015).

Bağlantı fonksiyonu yanıt değişkeninin dağılımının taşıdığı özellikleri üstlenmektedir. Normal dağılımlı yanıt değişkeninin beklenen değeri ile doğrusal tahmin edici arasında bağ kuran g bağlantı fonksiyonu $g(\mu_i) = \mu_i$ şeklinde ise birim bağlantı fonksiyonunu oluşturur.

Poisson dağılımlı yanıt değişkeni için, $g(\mu_i) = \ln\mu_i$ logaritmik (log) bağlantı fonksiyonu kullanılır.

Binom dağılımlı yanıt değişkeni için $g(\mu_i) = \ln\frac{\mu_i}{1-\mu_i}$ lojit bağlantı fonksiyonu kullanılır. Dolayısıyla $g(\mu_i)$ bağlantı fonksiyonu yanıt değişkeninin dağılımına göre seçilir (Karadağ, 2014).

Sonuç olarak, GDM' ler bir bağlantı fonksiyonu kullanarak, doğrusal tahmin ediciye bağlı olarak yığın ortalaması hesaplayan klasik doğrusal modellerin genelleştirilmiş bir durumudur (Garrido ve Zhou, 2009).

### 2.1.1. Lojistik regresyon

GDM' lerde yanıt değişkeninin nitel olması halinde lojistik regresyon modeli kullanılır. Lojistik regresyon modeli yanıt değişkeni ile bir veya birden fazla açıklayıcı değişken arasındaki ilişkiyi inceler.

Doğrusal regresyon analizinde bağımlı değişkenin değeri tahmin edilmeye çalışılırken, lojistik regresyon analizinde yanıt değişkeninin alacağı değerlerden birinin gerçekleşme olasılığı tahmin edilir (Şensoy, 2009). Lojistik regresyonda, doğrusal

olmayan ilişki korunarak, ilişkinin formunu doğrusal hale getiren logaritmik dönüşümler yapılır.

Lojistik regresyon modeline, yanıt değişkeninin sahip olduğu kategoriye göre farklı isimler verilmektedir. Nitel yapıdaki yanıt değişkeni iki değer alıyorsa "İkili Lojistik Regresyon Modeli" (Binary Logistic Regression Model) kullanılır. Yanıt değişken en az üç kategoriden oluşuyor ve sınıflanabilir bir yapıda, nominal ve sıralama yapılamıyorsa "Çoklu Lojistik Regresyon Modeli" (Multinominal Logistic Regression Model) ve yanıt değişken en az üç kategoriden oluşuyorken sıralı (ordinal) bir yapıda ise "Sıralı Lojistik Regresyon Modeli" (Ordinal Logistic Regression Model) kullanılmaktadır.

İki sonuçlu yanıt değişkenli modellerde yanıt değişkeni sadece 0 ve 1 gibi iki olası değer alabilen yapıya sahiptir. Bu niteliksel yapıda olan yanıt değişkeni için başarı ve başarısızlık şeklinde verilmiş değerler tamamen keyfi olarak tanımlanabilir.

$$y_i = \boldsymbol{x'}_i\boldsymbol{\beta} + \varepsilon_i \qquad i=1,2,\ldots,n \qquad \textbf{(2.5)}$$

şeklinde modeli ele alınsın. Burada açıklayıcı değişken $\boldsymbol{x'}_i = [1, x_{i1}, x_{i2}, \ldots, x_{ik}]$, $\boldsymbol{\beta'} = [\beta_0, \beta_1, \beta_2, \ldots, \beta_k]$' dır ve yanıt değişkeni $y_i$' de 0 veya 1 değerini alır. Tablo 2.1' deki olasılık dağılımı ile yanıt değişkeni $y_i$'nin Bernoulli raslantı değişkeni olduğunu varsayılsın.

**Tablo 2.1.** *Olasılık dağılımı (Montgomery, Peck ve Vining, 2013)*

| $y_i$ | Olasılık |
|---|---|
| 1 | P( $y_i$ =*1*)= $\pi_i$ |
| 0 | P( $y_i$ =*0*)= *1*- $\pi_i$ |

$E(\varepsilon_i) = 0$ olduğundan yanıt değişkeninin beklenen değeri,

$$E(y_i) = \boldsymbol{x'}_i\boldsymbol{\beta} = 1\,(\,\pi_i\,) + \,0(1 - \,\pi_i\,) = \pi_i \qquad \textbf{(2.6)}$$

olur. Buradan

$$E(y_i) = \boldsymbol{x'}_i\boldsymbol{\beta} = \,\pi_i \qquad \textbf{(2.7)}$$

olur. Bu Eşitlik (2.7), yanıt fonksiyonu ile verilen beklenen yanıtın sadece, yanıt değişkeninin 1 değerini alma olasılığı olduğunu ifade eder.

Eşitlik (2.5)' teki regresyon denkleminde bazı temel problemler vardır. Eğer yanıt değişkeni iki değer alıyorsa, $\varepsilon_i$ hata terimi sadece iki değer alabilir:

$$y_i = 1 \quad iken \;\; \varepsilon_i = 1 - \boldsymbol{x'}_i\boldsymbol{\beta} \tag{2.8}$$

$$y_i = 0 \quad iken \;\; \varepsilon_i = -\boldsymbol{x'}_i\boldsymbol{\beta} \tag{2.9}$$

Dolayısıyla, bu modeldeki hatalar normal olmayabilir.

Diğer problem ise hata varyansı aşağıdaki eşitlikten dolayı sabit değildir.

$$\begin{aligned} \sigma_{yi}^2 &= E\{\, y_i - \; E(y_i)\}^2 \\ &= (1-\pi_i\,)^2\,\pi_i \; + (0-\pi_i\,)^2\,(1-\pi_i\,) \\ &= \; \pi_i\,(1-\pi_i\,) \end{aligned} \tag{2.10}$$

Eşitlik (2.10)' da $E(y_i) = \boldsymbol{x'}_i\boldsymbol{\beta} = \; \pi_i$ olduğundan

$$\sigma_{yi}^2 = E(\, y_i)\,[1 - E(\, y_i)] \tag{2.11}$$

şeklinde yazılır. Buradan gözlemlerin varyansı ($\varepsilon_i = y_i - \pi_i$ ve $\pi_i$ bir sabit olduğundan hataların varyansı ile aynıdır) ortalamanın bir fonksiyonudur. Burada,

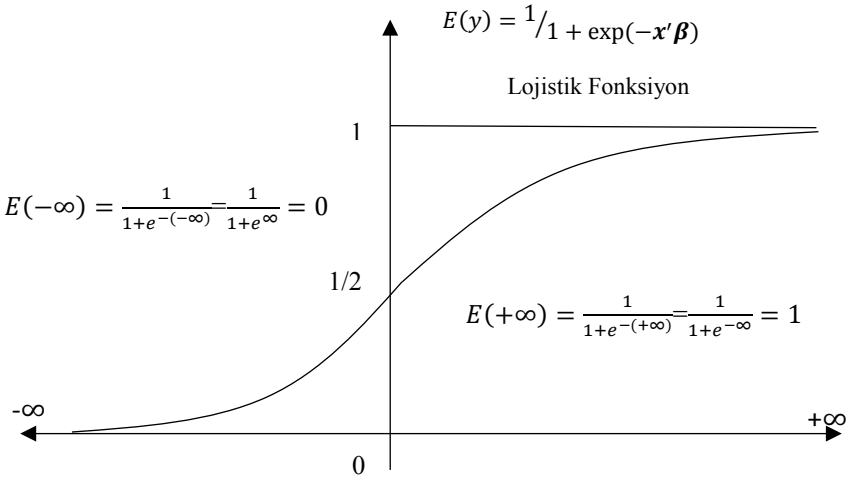
$$0 \le E(\, y_i) = \pi_i \; \le 1 \tag{2.12}$$

'dir. Eşitlik (2.12)'deki yanıt fonksiyonunda bir kısıt vardır. Bu kısıt başlangıçta kabul edilen Eşitlik (2.5)'deki gibi bir doğrusal yanıt fonksiyonunun seçilmesi ile problemlere sebep olabilir. Verilere yanıt değişkeninin tahmin değerlerinin [0,1] aralığı dışında olduğu bir model uydurulabilir (Montgomery, Peck ve Vining, 2013).

Lojistik yanıt fonksiyonu,

$$E(y) = \frac{\exp(\boldsymbol{x'}\boldsymbol{\beta})}{1+\exp(\boldsymbol{x'}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\boldsymbol{x'}\boldsymbol{\beta})} \tag{2.13}$$

biçiminde yazılır. Lojistik yanıt fonksiyonu, $x$'deki bir birim değişime karşılık $E(y)$'deki değişikliği gösterir.




**Şekil 2.1.** *Lojistik Yanıt Fonksiyonun Birikimli Dağılım Grafiği*

Şekil 2.1' de görüldüğü gibi lojistik dağılım fonksiyonunda, $y$, $(-\infty, +\infty)$ tanım aralığında değere sahiptir. Bu yüzden $E(y)$ fonksiyonunun değişim aralığı 0 ve 1 arasındadır. Lojistik regresyona ait olan eğrinin S şeklinde olması modelin tercih edilmesinin bir nedenidir. Eğri için, olasılık değerlerinin monoton artan ya da azalan S biçimli bir fonksiyonu kullanılır (Şensoy, 2009).

Lojistik yanıt fonksiyonu kolayca doğrusallaştırılabilir. Bunun için önerilen bir yaklaşım olarak, modelin yapısal kısmı, yanıt fonksiyonu ortalamasının bir fonksiyonu cinsinden tanımlanır. Doğrusal tahminci,

$$\eta = \boldsymbol{x'\beta} \tag{2.14}$$

olsun. Burada $\eta$,

$$\eta = \ln\left[\frac{P(y=1\,|x)}{P(y=0\,|x)}\right] = ln\frac{\pi}{1-\pi} \tag{2.15}$$

olarak tanımlanır. Bu dönüşüm ile doğrusallık sağlanır. Bu dönüşüme lojit dönüşüm denir ve $\pi/(1-\pi)$ oranına odds adı verilir (Montgomery, Peck ve Vining, 2013).

Odds; bir olayın olma olasılığını, olmama olasılığına bağlı olarak açıklar. İki odds değerinin birbirine oranlanması ile odds oranı (Odds Ratio-OR) hesaplanır. Odds oranı incelenen iki olayın gözlenme olasılıklarından birinin diğerine oranla kaç kat daha fazla veya kaç kat daha az olarak gerçekleşebileceğini gösterir (Güriş ve Astar, 2015).

Lojistik regresyon modelinin genel biçimi,

$$y_i = E(y_i) + \varepsilon_i \tag{2.16}$$

olarak yazılır. Burada $y_i$ gözlemleri, aşağıda verilen beklenen değer ile bağımsız Bernoulli rastlantı değişkenleridir:

$$E(y_i) = \pi_i = \frac{\exp(\boldsymbol{x_i'\beta})}{1 + \exp(\boldsymbol{x_i'\beta})} \tag{2.17}$$

$\boldsymbol{x_i'\beta}$' daki parametreleri tahmin etmek için en çok olabilirlik yöntemi kullanılabilir.

Gözlemlerin her biri Bernoulli dağılımı gösterir ve her bir gözlemin olasılık dağılımı,

$$f_i(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, \qquad i=1,2,\ldots,n \tag{2.18}$$

şeklinde olur ve her bir $y_i$ gözlemi 0 ve 1 değerini alır. Gözlemler bağımsız olduğu için olabilirlik fonksiyonu Eşitlik (2.19) ile verilebilir:

$$L(y_1, y_2, \ldots, y_n, \boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}\,. \tag{2.19}$$

Olabilirlik fonksiyonunun log-olabilirlik şeklinde düzenlenmiş hali;

$$\begin{aligned} \ln L(y_1, y_2, \ldots, y_n, \boldsymbol{\beta}) &= \ln \prod_{i=1}^{n} f_i(y_i) \\ &= \sum_{i=1}^{n} \left[ y_i \ln\left(\frac{\pi_i}{1 - \pi_i}\right)\right] + \sum_{i=1}^{n} \ln(1 - \pi_i) \end{aligned} \tag{2.20}$$

Eşitlik (2.20)'deki gibidir. Buna göre $1 - \pi_i = [1 + \exp(\boldsymbol{x_i'\beta})]^{-1}$ ve

$\eta_i = \ln[\pi_i/(1 - \pi_i)] = \boldsymbol{x_i'\beta}$ olduğu için log olabilirlik,

$$\ln L(\boldsymbol{y}, \boldsymbol{\beta}) = \sum_{i=1}^{n} y_i\, \boldsymbol{x_i'\beta} - \sum_{i=1}^{n} \ln[1 + \exp(\boldsymbol{x_i'\beta})] \tag{2.21}$$

Eşitlik (2.21)'deki gibi yazılabilir.

Lojistik regresyon modellerinde $x$ değişkenlerinin her düzeyinde birden fazla gözlem varsa $y_i$, $i$. gözlem için gözlenen 1'lerin sayısı ve $n_i$ her bir gözlemdeki denemelerin sayısı olsun. Bu durumda log olabilirlik Eşitlik (2.22)'deki gibi olur.

$$\begin{aligned} \ln L(\boldsymbol{y}, \boldsymbol{\beta}) &= \sum_{i=1}^{n} y_i \ln(\pi_i) + \sum_{i=1}^{n} n_i \ln(1-\pi_i) - \sum_{i=1}^{n} y_i \ln(1-\pi_i) \\ &= \sum_{i=1}^{n} y_i \ln(\pi_i) + \sum_{i=1}^{n} (n_i - y_i) \ln(1-\pi_i) \end{aligned} \tag{2.22}$$

$\widehat{\boldsymbol{\beta}}$ yukarıdaki Eşitlik (2.22)'deki algoritmanın verdiği model parametrelerinin son tahmini olsun. Eğer model varsayımları doğru ise asimtotik olarak,

$$E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \text{ ve } \mathrm{Var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1} \tag{2.23}$$

olduğu gösterilir. Burada **V**, her bir gözlemin tahmin edilen varyansını içeren $n \times n$ köşegen matrisidir. **V**' nin *i.* köşegen elemanı $V_{ii} = n_i \hat{\pi}_i (1 - \hat{\pi}_i)$ şeklindedir.

$\eta$ doğrusal tahmin edicinin tahmin değeri $\hat{\eta}_i = \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}$'dır ve lojistik regresyon modelinin uyum değeri,

$$\hat{y}_i = \hat{\pi}_i = \frac{\exp(\hat{\eta}_i)}{1 + \exp(\hat{\eta}_i)} = \frac{\exp(\boldsymbol{x}_i' \widehat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{x}_i' \widehat{\boldsymbol{\beta}})} = \frac{1}{1 + \exp(-\boldsymbol{x}_i' \widehat{\boldsymbol{\beta}})} \tag{2.24}$$

Eşitlik (2.24)'deki gibi yazılır (Montgomery, Peck ve Vining, 2013).

### 2.1.2. Sıralı lojistik regresyon modelleri

Lojistik regresyon modelinde, bağımlı değişken en az üç kategoriye sahip ve kategoriler doğal bir sıraya göre küçükten büyüğe doğru sıralanıyor ise sıralı lojistik regresyon modeli kullanılır. Örneğin, hastalığın şiddeti (en az şiddetliden en çok şiddetliye doğru), müşteri memnuniyeti (memnun değil, biraz memnun, çok memnun) gibi ölçeklerde ölçülebilir (Montgomery, Peck ve Vining, 2013).

Sıralı lojistik regresyon modeli, eşit aralıklı olmayan sıralı kategorilere sahip bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin belirlenmesi için en uygun tekniktir (Tıpiş, 2012).

Sıralı lojistik regresyon modellerinin elde edilmesi için bağlantı fonksiyonu kullanılmaktadır. Genellikle lojit, probit gibi bağlantı fonksiyonları kullanılır. Regresyon

katsayılarının değeri sıralı kategorik değişkenin kategorilerine bağlı olmadığı için bağlantı fonksiyonu kullanılarak tahmin edilen regresyon katsayıları her bir kesim noktasında (cut point) aynıdır (Chen ve Hughes,2004; Breslaw ve McIntosh, 1998).

Modelin oluşmasında önemli bir yaklaşım ise sıralı kategorik bağımlı değişkenin etkisi altında olduğu düşünülen gözlenmemiş sürekli, gizli (latent) bir bağımlı değişkenden tekrar düzenlenmesi varsayılır. İkinci olarak sıralı lojistik regresyon modelinin en belirgin özelliği olarak kabul edilen paralellik varsayımı önemlidir. Paralellik varsayımı gereği $\boldsymbol{\beta}$ parametresi farklı kategoriler ve farklı kesim noktaları için değişiklik göstermez ve dolayısıyla burada modeller arasındaki paralellik sınamasına gidilmesi gerekir.

#### *2.1.2.1. Sıralı lojistik regresyon modelinde gizli değişken kavramı*

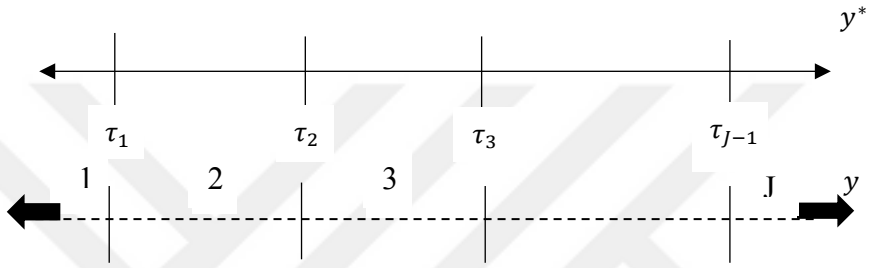
Gizli değişken kavramı; aralıklı ve kategorilere sahip gözlenebilir *y* bağımlı değişkenin altında sürekli $(-\infty, +\infty)$ ve gözlenemeyen gizli bir $y^*$ değişkeninin olduğu varsayımına dayanır (Long,1997; Breslaw ve McIntosh,1998). Genel olarak gizli değişken $y^*$, sıralı aralıklara bölünerek Eşitlik (2.25) ile gösterilebilir.

$$y_i = m \text{ olduğunda } m = 1, \dots, J \text{ için } \tau_{m-1} \leq y_i^* < \tau_m \qquad \textbf{(2.25)}$$

Burada *m*, bağımlı değişkenin kategorisini ifade eden sayılar, $\tau$'lar kesim noktaları olarak tanımlanır. En uç kategoriler 1 ve *J* için kesim noktaları $\tau_0 = -\infty$ ve $\tau_j = +\infty$ ile açık uçlu aralıklarla tanımlanır. Kesim noktaları bağımlı değişken ile gizli değişken arasındaki ilişkinin ifade edilmesini sağlar. Örneğin; anket uygulanan bireylere herhangi bir düşünceye katılıp katılmadıklarının sorusuna verilen cevap; (1) hiç katılmıyorum, (2) katılmıyorum, (3) kararsızım, (4) katılıyorum, (5)kesinlikle katılıyorum, şeklinde kategorilere ayrılmış olsun. Bu sıralı değişkenin sürekli gizli bir değişken ile ilişkili olduğu varsayılır. Bağımlı değişkenin *J* tane kategorisinin olduğu durumda, gözlenen y değişkeni ile gizli değişken $y^*$ arasındaki ilişki Eşitlik (2.26)'daki gibi gösterilir (Long, 1997; Liao, 1994).

$$y_i = \begin{cases} 1\ eğer\ , & -\infty = \tau_0 \le y_i^* < \tau_1 \\ 2\ eğer, & \tau_1 \le y_i^* < \tau_2 \\ 3\ eğer, & \tau_2 \le y_i^* < \tau_3 \\ & . \\ & . \\ & . \\ J\ eğer, & \tau_{J-1} \le y_i^* < \tau_J = +\infty \end{cases} \quad \textbf{(2.26)}$$

Eşitlik (2.26) 'da ardışık kategorileri ayıran $\tau$ kesim noktaları aşağıdaki Şekil 2.2.'deki gibi gösterilir.




**Şekil 2.2.** *Sıralı Bağımlı Değişkenin Kategorilerine Karşılık Gelen* $\boldsymbol{y^*}$ *Gizli Değişkenin Değerleri.*

Bu kesim noktaları $\beta_i$' lerle tahmin edilir. Burada $[\tau_1, \tau_{J-1}]$ noktaları $0 < \tau_1 < \tau_2 < \cdots < \tau_{J-1}$ şeklinde pozitif değer alabilir. Her zaman $\tau_0 = -\infty$ ve $\tau_J = +\infty$ olduğundan bağımlı değişkenin kategori sayısının 1 eksiği kadar kesim noktası bulunur (Liao,1994). Bu kesim noktaları vasıtasıyla $y$ bağımlı değişkene ait çıkarsamalarda bulunulur ve bu kesim noktaları aracılığıyla model kurulmaya çalışılır.

Gizli değişkenin bağımsız değişkenler ve $\varepsilon$ hata terimi yardımıyla gösterimi genel olarak Eşitlik (2.27)'deki gibi ifade edilir.

$$y^* = \sum_{k=1}^{K} \beta_k x_k + \varepsilon \quad \textbf{(2.27)}$$

$y^*$ değişkeninin dağılımını $\varepsilon$ hata terimi belirler. Burada $\varepsilon$ sıfır ortalama ile belirli bir dağılım gösteren ve dağılım fonksiyonu $F(\varepsilon)$ ile tanımlanan bir hata terimidir (Liao,1994).

### *2.1.2.2. Gizli değişkende dağılımsal varsayımlar*

Eşitlik (2.27)'de gösterilen modeli tahmin etmek için, sıradan en küçük kareler yöntemi bağımlı değişken gözlemlenemeyen olduğundan, en çok olabilirlik yöntemi kullanılabilir. Bu yönteminin kullanılabilmesi için, hata terimlerinin dağılımına ilişkin varsayıma ihtiyaç vardır. Çoğunlukla, hatalar normal olduğundan probit model veya lojistik hatalar olduğunda lojit model düşünülür (Long, 1997). Benzer şekilde, hata terimlerinin dağılımı için çok fazla kullanılmayan başka dağılımlar da mevcuttur (McCullagh, 1980).

Sıralı lojistik regresyon modellerinin oluşturulmasında McCullagh (1980) tarafından geliştirilen probit ve lojit modellerin dağılımsal varsayımları uygulamada ve yorumlamada kolaylık sağladığı için çok sık kullanılır Örneğin; sıralı probit model için, $\varepsilon$ hata terimi 0 ortalama ve 1 varyansla normal dağılım gösterir (Long, 1997). Buna ilişkin olasılık yoğunluk fonksiyonu ve birikimli dağılım fonksiyonları sırasıyla Eşitlik (2.28) ve (2.29)'da gösterilmiştir:

$$\lambda(\varepsilon) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\varepsilon^2}{2}\right) \tag{2.28}$$

$$\Lambda(\varepsilon) = \int_{-\infty}^{\varepsilon} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt. \tag{2.29}$$

Sıralı lojit modelde, $\varepsilon$ hata terimi 0 ortalama ve $\frac{\pi^2}{3}$ varyans ile lojistik dağılıma sahiptir.

$$f(\varepsilon) = \frac{\exp(\varepsilon)}{[1 + \exp(\varepsilon)]^2} \tag{2.30}$$

$$F(\varepsilon) = \frac{\exp(\varepsilon)}{1 + \exp(\varepsilon)} \tag{2.31}$$

Burada olasılık yoğunluk fonksiyonu ve birikimli dağılım fonksiyonları sırasıyla Eşitlik (2.30) ve (2.31)'de gösterilmiştir.

### *2.1.2.3. Gözlemlenmiş kategorilerin olasılıklarının elde edilmesi*

Hata terimlerinin dağılımı belirlendikten sonra verilen $\boldsymbol{x}$ değişkenine bağlı olarak $y$ değişkeninin gözlemlenmiş değerlerinin olasılıkları hesaplanır. Eşitlik (2.26)'da gözlenen bağımlı değişkenin birinci kategoride meydana gelme olasılığı $P(y = 1)$ için Eşitlik (2.32)'deki gibi gösterilir (Long, 1997).

$$P(y = 1) = P(\tau_0 \leq y^* \leq \tau_1) \tag{2.32}$$

Eşitlik (2.32)' de $y^* = \boldsymbol{x\beta} + \varepsilon$' ın değeri yerine yazılırsa;

$$P(y = 1) = P\left(\tau_0 \leq \sum_{k=1}^{K} \beta_k x_k + \varepsilon \leq \tau_1\right)$$

$$= P\left(\tau_0 - \sum_{k=1}^{K} \beta_k x_k \leq \varepsilon \leq \tau_1 - \sum_{k=1}^{K} \beta_k x_k\right)$$

$$P(y = 1) = P\left(\tau_1 - \sum_{k=1}^{K} \beta_k x_k\right) - P\left(\tau_0 - \sum_{k=1}^{K} \beta_k x_k\right)$$

$$P(y = 1) = F\left(\tau_1 - \sum_{k=1}^{K} \beta_k x_k\right) - F\left(\tau_0 - \sum_{k=1}^{K} \beta_k x_k\right)$$

Birinci kategori için birikimli olasılık değeri hesaplanırsa,

($F(\tau_0 - \sum_{k=1}^{K} \beta_k x_k) = F(-\infty - \sum_{k=1}^{K} \beta_k x_k) = 0$ şeklinde yazılır.)

$$P(y = 1) = P\left(\tau_1 - \sum_{k=1}^{K} \beta_k x_k\right) \tag{2.33}$$

Eşitlik (2.33) elde edilir.

Diğer kategorilere ait olasılıklar;

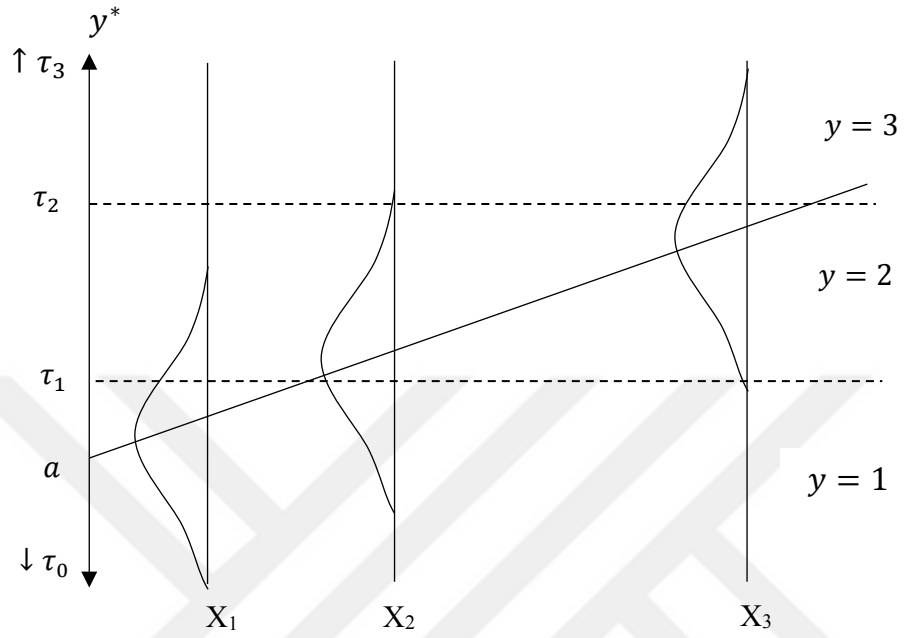
$$P(y = 2) = F\left(\tau_2 - \sum_{k=1}^{K} \beta_k x_k\right) - F\left(\tau_1 - \sum_{k=1}^{K} \beta_k x_k\right)$$

$$P(y = 3) = F\left(\tau_3 - \sum_{k=1}^{K} \beta_k x_k\right) - F\left(\tau_2 - \sum_{k=1}^{K} \beta_k x_k\right)$$

$$\vdots$$

$$P(y = J) = F\left(\tau_J - \sum_{k=1}^{K} \beta_k x_k\right) - F\left(\tau_{J-1} - \sum_{k=1}^{K} \beta_k x_k\right) \tag{2.34}$$

Eşitlik (2.34) ile verilir ve $F(\varepsilon)$ hata teriminin dağılım fonksiyonu tanımlanır. Gizli değişken $y^*$ için hata terimi $\varepsilon$, dağılımı lojistik dağılımı gösterir (Liao, 1994).




**Şekil 2.3.** *Sıralı Lojistik Regresyon Modeli için x Değerleri verildiğinde, Gözlenen 4 Kategorili y için Gizli Değişken* $\boldsymbol{y^*}$ *Değişkeninin Dağılım Grafiği.*

Gözlenen bağımlı değişkenin son kategoride meydana gelme olasılığı $P(y = J)$ hesaplanırken $F\left(\tau_J - \sum_{k=1}^{K} \beta_k x_k\right) = 1$ olur ve Eşitlik (2.34)'teki ifade, Eşitlik (2.35) gibi yazılır (Liao, 1994; Long, 1997).

$$P(y = J) = 1 - F\left(\tau_{J-1} - \sum_{k=1}^{K} \beta_k x_k\right) \tag{2.35}$$

Şekil 2.3.' de gözlenen 4 kategorili bağımlı değişken için gizli değişkenin dağılımı gösterilmiştir. Bu dağılımın kategori olasılıkları Eşitlik (2.36)' daki gibi elde edilir.

$$P(y_i = 1|\boldsymbol{x_i}) = P(\tau_0 \leq y^* \leq \tau_1|\boldsymbol{x_i}) \tag{2.36}$$

Eşitlik (2.36)' da $y^* = \boldsymbol{x_i\beta} + \varepsilon_i$ yazılırsa, aşağıdaki Eşitlik(2.37) elde edilir.

$$P(y_i = 1|\boldsymbol{x}_i) = P(\tau_0 \leq \boldsymbol{x_i\beta} + \varepsilon_i \leq \tau_1|\boldsymbol{x_i}) \tag{2.37}$$

Eşitlik (2.37)'de $\varepsilon_i$ hata terimlerini yalnız bırakmak için eşitliğin her iki tarafına negatif olarak $\boldsymbol{x_i\beta}$ eklenirse Eşitlik (2.38) elde edilir.

$$P(y_i = 1|\boldsymbol{x_i}) = P(\tau_0 - \boldsymbol{x_i\beta} \leq \varepsilon_i \leq \tau_1 - \boldsymbol{x_i\beta}|\boldsymbol{x_i})$$
$$P(y_i = 1|\boldsymbol{x_i}) = P(\varepsilon_i \leq \tau_1 - \boldsymbol{x_i\beta}|\boldsymbol{x}_i) - P(\varepsilon_i \leq \tau_0 - \boldsymbol{x_i\beta}|\boldsymbol{x_i})$$
$$= F(\tau_1 - \boldsymbol{x_i\beta}) - F(\tau_0 - \boldsymbol{x_i\beta})$$
$$(F(\tau_0 - \boldsymbol{x_i\beta}) = F(-\infty - \boldsymbol{x_i\beta}) = 0)$$
$$P(y_i = 1|\boldsymbol{x}_i) = F(\tau_1 - \boldsymbol{x_i\beta}) \quad \textbf{(2.38)}$$

Diğer 3 kategori için de birikimli olasılık dağılım fonksiyonu aşağıdaki Eşitlik (2.39), (2.40) ve (2.41)' deki gibi elde edilir.

$$P(y_i = 2|\boldsymbol{x_i}) = F(\tau_2 - \boldsymbol{x_i\beta}) - F(\tau_1 - \boldsymbol{x_i\beta}) \quad \textbf{(2.39)}$$

$$P(y_i = 3|\boldsymbol{x_i}) = F(\tau_3 - \boldsymbol{x_i\beta}) - F(\tau_2 - \boldsymbol{x_i\beta}) \quad \textbf{(2.40)}$$

$$P(y_i = 4|\boldsymbol{x_i}) = F(\tau_4 - \boldsymbol{x_i\beta}) - F(\tau_3 - \boldsymbol{x_i\beta}) \quad \textbf{(2.41)}$$

Eşitlik (2.41)'de gözlenen bağımlı değişkenin son kategoride meydana gelme olasılığı $P(y = 4)$ hesaplanırken $F(\tau_4 - \boldsymbol{x_i\beta}) = +\infty = 1$ olduğu için ifade,

$P(y_i = 4|\boldsymbol{x}_i) = 1 - F(\tau_3 - \boldsymbol{x_i\beta})$ gibi yazılabilir.

Sıralı probit modelde, $\varepsilon$ hata teriminin dağılımı normal dağılım gösterdiği için gözlenen bağımlı değişkenin son kategoride meydana gelme olasılığı Eşitlik (2.42)' deki gibi gösterilir (Liao, 1994).

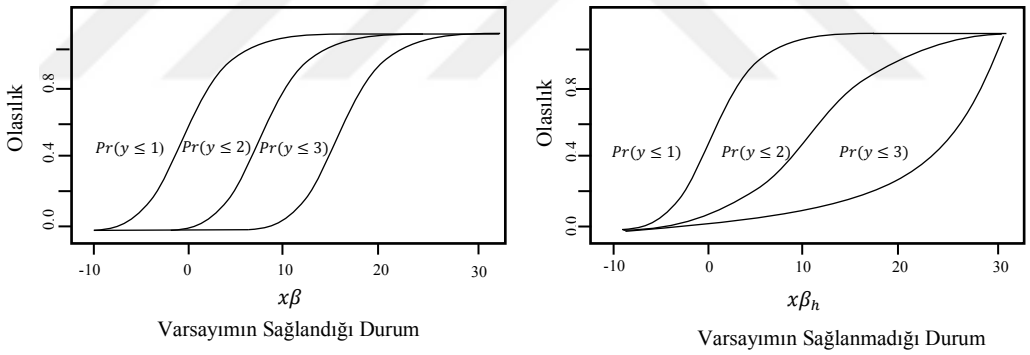
$$P(y = 1) = \phi\left(-\sum_{k=1}^{K}\beta_k x_k\right)$$
$$P(y = 2) = \phi\left(\tau_2 - \sum_{k=1}^{K}\beta_k x_k\right) - \phi\left(\tau_1 - \sum_{k=1}^{K}\beta_k x_k\right)$$
$$P(y = 3) = \phi\left(\tau_3 - \sum_{k=1}^{K}\beta_k x_k\right) - \phi\left(\tau_2 - \sum_{k=1}^{K}\beta_k x_k\right)$$
$$\vdots$$
$$P(y = J) = 1 - \phi\left(\tau_{J-1} - \sum_{k=1}^{K}\beta_k x_k\right) \quad \textbf{(2.42)}$$

Sıralı lojit ve sıralı probit modelleri arasındaki farklılık hata terimlerinin dağılımından kaynaklanmaktadır.

### *2.1.2.4. Paralel doğrular varsayımı*

Paralel doğrular varsayımı, bağımlı değişken açıklanırken belirlenen regresyon katsayılarının, sıralı bağımlı değişkenin tüm kategorilerinde eşit olduğunu varsayar (Ayhan, 2006). Bir başka deyişle, bağımsız değişkenler ile bağımlı değişken arasındaki ilişki bağımlı değişkenin kategorilerine göre değişiklik göstermez. Sıralı lojistik regresyonda, *J* tane kategoriye sahip bağımlı değişken için *J*-1 tane lojit karşılaştırma ile varsayım sağlandığında $\tau_{J-1}$ tane kesim noktası ve yalnızca bir adet $\beta$ parametresi bulunur (Kleinbaum ve Klein, 2010).

Bu varsayım ile bağımlı değişkenin kategorilerinin birbirine paralel olduğu söylenebilir. Varsayımın sağlanmadığı durumda kategorilerin paralelliği bozulmuş olur (Fullerton ve Xu, 2012). Bu durum Şekil 2.4.'te gösterilmiştir.




**Şekil 2.4.** *Kategorilere Ait Olasılıklar için Varsayımın Sağlandığı ve Sağlanmadığı Durumlara Göre Olasılık Dağılım Grafikleri*

Paralel doğrular varsayımını daha ayrıntılı açıklamak gerekirse, *h*' ye eşit ya da daha düşük kategoriye sahip modelin birikimli olasılık modelini ele alalım.

$$P(y \leq h|\boldsymbol{x}) = F(\tau_h - \boldsymbol{x\beta}) \quad \textbf{(2.43)}$$

Eşitlik (2.43)'de verilen birikimli dağılım fonksiyonundaki $\tau_h - \boldsymbol{x\beta}$ değer birikimli olasılık değeridir. $\boldsymbol{\beta}$ tüm kategoriler için aynı olduğundan Eşitlik (2.43) farklı kesim noktalarıyla iki kategorili modelde gösterilecek olursa,

$$\tau_h - \boldsymbol{x\beta} = (\tau_h - \beta_0) - \sum_{k=1}^{K} \beta_k \, x_k \tag{2.44}$$

Eşitlik (2.44) şeklinde ifade edilir.

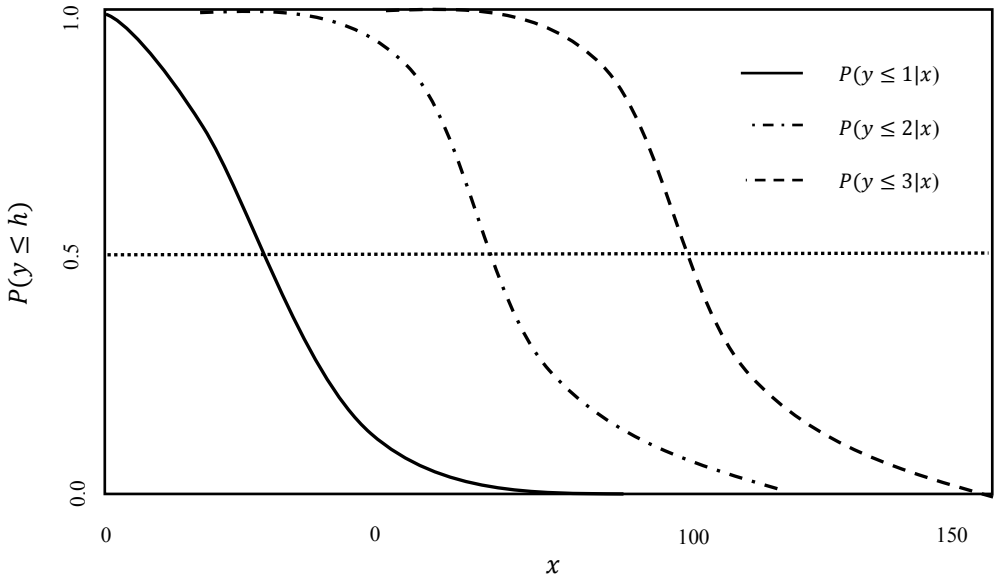
$y \leq 1$ için model, $\tau_1 - \beta_0$ kesim noktasıyla Eşitlik (2.45)

$$P(y \leq 1|\boldsymbol{x}) = F\left( (\tau_1 - \beta_0) - \sum_{k=1}^{K} \beta_k \, x_k \right) \tag{2.45}$$

ve $y \leq 2$ için model, Eşitlik (2.46)'daki gibi elde edilir.

$$P(y \leq 2|\boldsymbol{x}) = F\left( (\tau_2 - \beta_0) - \sum_{k=1}^{K} \beta_k \, x_k \right) \tag{2.46}$$

Eşitlik (2.45) ve Eşitlik (2.46)'da görüldüğü gibi kesim noktaları değişiklik gösterir ama katsayılar aynıdır. Kesim noktalarındaki bu değişiklik olasılık eğrisinde sağa veya sola doğru hafif bir değişikliğe sebep olur ancak eğrinin şeklini değiştirmez (Long,1997,s.140).




**Şekil 2.5.** *Birikimli Olasılık Fonksiyonu*

Şekil 2.5.' te 4 kategorinin birikimli olasılık değerleri verilmektedir. Sıralı lojistik regresyonda, dört kategorili bağımlı değişken için 3 tane birikimli olasılık eğrisi elde edilir. Olasılık değeri olarak 0.5' in, x ekseninde noktalarla ifade edildiğini görebiliriz. Bu noktadaki 3 olasılık eğrisinin eğim katsayısı Eşitlik (2.47) ile bulunabilir (Long, 1997).

$$\frac{\partial P(y \leq 1|x)}{\partial x} = \frac{\partial P(y \leq 2|x)}{\partial x} = \frac{\partial P(y \leq 3|x)}{\partial x} \quad \textbf{(2.47)}$$

Eşitlik (2.47)' de verilen ifadeye dikkat edilirse, bağımlı değişkenin bütün kategorilerinde eğim katsayısı eşittir. Bu durumda eğriler paraleldir.

### *2.1.2.5. Paralel doğrular varsayımının test edilmesi*

Sıralı lojistik regresyonda bilgilerin doğru ve güvenilir olması için paralel doğrular varsayımının sağlanması şarttır. Bu varsayım sağlanmadığı takdirde sonuçlar anlamsız ve yanlış olur. Bu yüzden modelin varsayımını test etmemiz gerekir.

Paralel doğrular varsayımının uygunluğunu doğrulamak için Olabilirlik Oran Testi, Wald Ki-Kare testi gibi testler kullanılmaktadır (Long, 1997; Agresti, 2019). Eğer varsayım sağlanmazsa alternatif modeller kullanılabilir (O'Connell, 2006).

Varsayımı test ederken, bağımlı değişken açıklanırken belirlenen regresyon katsayılarının her bir kategoride birbirine eşit olup olmadığına bakılır. Hipotezler aşağıdaki gibi kurulur (Ayhan, 2006).

$H_0$: $\boldsymbol{\beta_1 = \beta_2 \cdots = \beta_{J-1} = \beta}$

$H_1$: En az bir $\boldsymbol{\beta}$ farklıdır.

Bir başka değişle, "Modeldeki regresyon katsayıları, bağımlı değişkenin bütün kategorilerinde aynıdır" hipotezine karşılık "Modeldeki regresyon katsayıları, bağımlı değişkenin bütün kategorilerinde farklıdır" hipotezi test edilir.

#### *Olabilirlik oran testi*

Olabilirlik oran testi, kısıtlı bir model tahmin eder ve kısıtların kaldırılmasıyla log olabilirlik fonksiyonundaki değişimi gösterir. Bu test, sabit parametre dışındaki diğer parametrelerin anlamlılığını beraber test eder. Bu özellik ile regresyon modellerindeki F testinin karşılığıdır ve ki-kare dağılımı kullanılır (Güriş ve Astar, 2015).

Log olabilirlik fonksiyonu ile gözlenen ve tahmin edilen değerlerin karşılaştırması Eşitlik (2.48)'de gösterilmektedir.

$$D = 2lnL(Kısıtsız\ model) - 2lnL(Kısıtlı\ model) \quad \textbf{(2.48)}$$

Eşitlik (2.48)'de kısıtlı model; sadece kesim noktasının bulunduğu modeli ifade ederken, kısıtsız model; modelde bulunan kesim noktası, tüm eğim ve diğer parametreleri içeren modeli ifade eder. $lnL$ ifadesine olabilirlik oranı (Likelihood ratio) adı verilir. $D$ istatistiği doğrusal regresyondaki hata kareler toplamına karşılık gelir ve uyum iyiliğine karar vermede yardımcı olur. Bu eşitliğe göre kısıtsız model ve sabit terimli model (kısıtlı) için ayrı ayrı $D$ istatistiği bulunduktan sonra, $G^2$ istatistiği ile bağımsız değişkenin istatistiksel anlamlılığı test edilmektedir (Hosmer ve Lemeshow, 2000).

$$G^2 = D(Kısıtlı\ Model) - D(Kısıtsız\ Model) \quad \textbf{(2.49)}$$

$G^2$ asimtotik olarak ki-kare dağılımına sahiptir. Burada $\chi^2$dağılımı için serbestlik derecesi $k(J-2)$'dir. "$k$" bağımsız değişken sayısı, "$J$" ise kategori sayısını ifade eder. Eğer test istatistiği tablo değerinden büyükse, $H_0 = \beta_1 = \beta_2 \cdots = \beta_{J-1} = 0$ hipotezi reddedilir ve en az bir eğim katsayısının diğerinden farklı olduğu sonucuna varılmış olur. Bu yüzden paralel doğrular varsayımı sağlanmadığı ortaya çıkar (Long,1997).

***Wald Ki-Kare testi***

Paralel doğrular varsayımının uygunluğunu kontrol etmek için kullanılan diğer test Wald ki-kare test istatistiğidir. Bu test ile regresyon katsayılarının paralelliği genel olarak test edilip hangi değişken veya değişkenler nedeniyle paralel doğrular varsayımının bozulduğu gösterilir (Brant, 1990; Long, 1997; Kleinbaum ve Klein, 2010).

Wald ki-kare testi, regresyon katsayılarının standart hata tahminine oranlanması sonucunda bulunur. Wald ki-kare testi en çok olabilirlik tahmincilerinin yaklaşık standart normal dağılım gösterdiğini varsaymaktadır (Agresti, 2019).

Wald test istatistiği, $H_0$ doğru iken 1 serbestlik derecesi ile ki-kare dağılımına uygunluk gösterir ve Eşitlik (2.50) ile verilmiştir.

$$\mathrm{W} = z^2 = \left[\frac{(\hat{\beta}_1 - \beta^*)}{SH(\hat{\beta}_1)}\right]^2 \quad \textbf{(2.50)}$$

Eşitlik (2.50)' de görüldüğü gibi W değeri standart normal dağılmış bir ($z^2$) istatistiğin karesidir. Bir başka deyişle; Wald testi, $\beta_1$ eğim parametresinin en çok olabilirlik tahmininin bu parametrenin standart hata tahminine oranlanmasıyla bulunur. Bu oran $H_0: \beta_1 = 0$ hipotezi altında standart normal dağılıma yaklaşır. Hipotezler daha karmaşık hale getirilebilir, $H_0: \beta_1 = \beta_2 = \cdots = \beta_{J-1} = 0$ (Long, 1997, s.92).

Olabilirlik oran testi ve Wald ki-kare testi yapısal olarak aynı sonuçları vermesine rağmen, küçük örneklem araştırmalarında farklı sonuçlar ortaya çıkmaktadır. Testlerden hangisinin kullanılması hakkında bir kesinlik yoktur, asimtotik olarak aynı oldukları söylenebilir. Ancak literatürde Wald ki-kare testinin biraz daha güçlü olduğu düşünülmektedir (Long,1997, s.97).

#### *2.1.2.6. Orantısal oran modeli*

Sıralı lojistik regresyonda sıralı kategoriye sahip bağımlı değişkenin kategorilerinin arasında paralellik varsayımı sağlanıyorsa model kurmak için genellikle orantısal oran modeli ( Proportional Odds Model) kullanılır (Brant, 1990).

Sıralı lojistik regresyon için McCullagh (1980) tarafından tanımlanan orantısal oran modeli (OOM), modelin birikimli olasılıklarının dağılımına dayanır. Birikimli olasılık, belirlenen kategori ve daha düşük kategorilerde gerçekleşme olasılığını birlikte gösterir (O'Connell, 2006).

Birikimli olasılıklar kullanılarak Eşitlik (2.51) gösterilir (Kleinbaum ve Ananth, 1997).

$$P(y \leq y_j|\boldsymbol{x}) = \left[\frac{\exp(\tau_j - \boldsymbol{x}'\boldsymbol{\beta})}{1 + \exp(\tau_j - \boldsymbol{x}'\boldsymbol{\beta})}\right] \quad j = 1,2,\dots,J-1 \tag{2.51}$$

Eşitlik (2.51)'de $\tau_j$ bilinmeyen kesim parametrelerini ve $J-1$ tane tahminciye karşılık gelen $\tau_1 \leq \tau_2 \leq \cdots \leq \tau_{J-1}$ ; $\beta$'lar ise $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)'$ $\boldsymbol{x}$'e karşılık gelen bilinmeyen regresyon katsayılarının bir vektörüdür. Odds oranlarının logaritması alınıp model doğrusal bir formda Eşitlik (2.52)'deki gibi yazılabilir (Kleinbaum ve Ananth, 1997).

$$P(y \leq y_j|\boldsymbol{x}) = \left[\frac{\exp(\tau_j - \boldsymbol{x}'\boldsymbol{\beta})}{1 + \exp(\tau_j - \boldsymbol{x}'\boldsymbol{\beta})}\right]$$

$$1 - P(y \leq y_j|\boldsymbol{x}) = \frac{1}{1 + \exp(\tau_j - \boldsymbol{x}'\boldsymbol{\beta})}$$

$$odds = \frac{[P(Y \leq y_j|\boldsymbol{x})]}{[P(Y > y_j|\boldsymbol{x})]}$$

$$= \frac{\left[\frac{\exp(\tau_j - \boldsymbol{x}'\boldsymbol{\beta})}{1 + \exp(\tau_j - \boldsymbol{x}'\boldsymbol{\beta})}\right]}{\frac{1}{1 + \exp(\tau_j - \boldsymbol{x}'\boldsymbol{\beta})}}$$

$$= \exp(\tau_j - \boldsymbol{x}'\boldsymbol{\beta}) \qquad j = 1,2, \dots, J-1$$

$$log\frac{[P(Y \leq y_j|\boldsymbol{x})]}{[P(Y > y_j|\boldsymbol{x})]} = log\exp(\tau_j - \boldsymbol{x}'\boldsymbol{\beta})$$

$$= (\tau_j - \boldsymbol{x}'\boldsymbol{\beta}) \quad j = 1,2, \dots, J-1 \qquad \textbf{(2.52)}$$

Orantısal oran modeli için her bir birikimli lojitin kendisine ait kesim noktası vardır. Tüm birikimli lojitler için bağımsız değişkenlere ait $\boldsymbol{\beta}$'lar birbirine eşittir. Dolayısıyla önceden test edilmesi gereken paralel doğrular varsayımını sağlamış olur.

### 2.1.3. Model katsayılarının açıklanması

Modelleme lojit bağlantı fonksiyonu yardımıyla yapılır. Lojit modellerde katsayıları yorumlamak için odds oranlarından faydalanılır. Gizli değişken ve diğer bütün değişkenler sabitken $exp(\hat{\beta}_k)$ değeri odds oranlarını verir. Değişkenler için bir referans kategorisi seçilir. Bu referans kategorileri kullanılarak yorumlamalar yapılır.

## 2.2. Sınıflandırma ve Regresyon Ağaçları (CART)

İstatistiksel verilerin görsel olarak incelenmesi, değişkenler arasındaki ilişkinin belirlenip tahminler yapılabilmesi için karar ağaç modelleri kullanılmaktadır. Ağaç modellerinde, bağımsız değişkene ait sorulardan alınan cevaplarla ortaya çıkan yollar görsel olarak ağaç dallarını oluşturmakta ve bağımlı değişkeni etkileyen bağımsız değişken ya da değişkenleri göstermektedir (Temel, 2004).

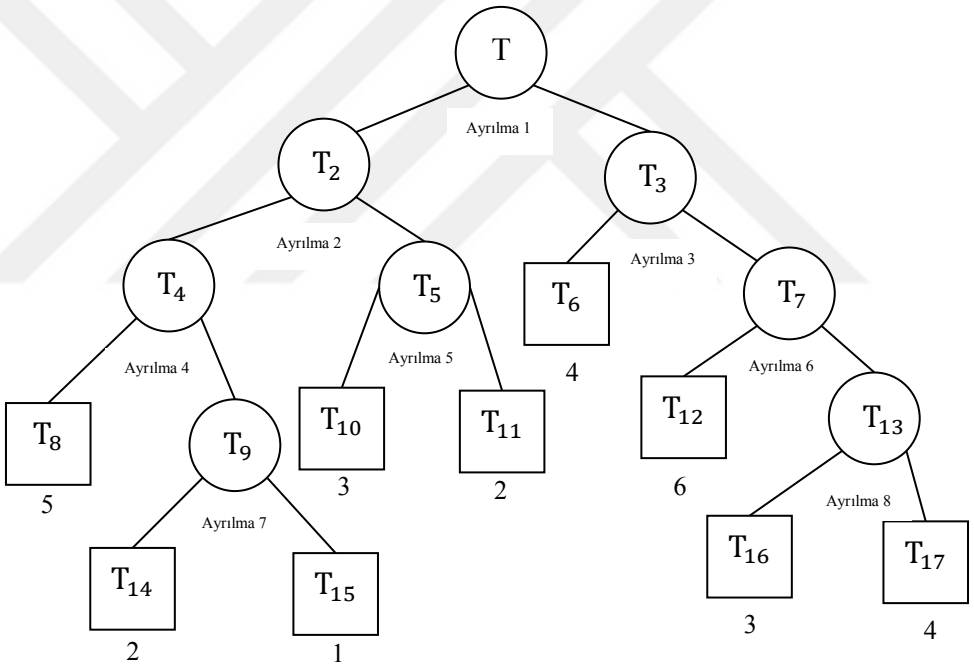
Veri seti için hiçbir varsayıma ihtiyaç duymaması nedeniyle tercih edilen sınıflama ve regresyon ağaçları (Classification and Regression Tree, CART) hem kategorik hem de

sürekli bağımlı değişkenlerin alacağı değerleri analiz ve tahmin etmek için kullanılan parametrik olmayan istatistiksel bir yöntemdir (Ceyhan, 2014).

1984 yılında Breiman ve arkadaşları tarafından geliştirilen CART analizinde incelenen bağımlı değişken kategorik ise sınıflama ağaçları (Classification Tree), sürekli ise regresyon ağaçları (Regression Tree) şeklinde isimlendirilir (Fu, 2004).

### 2.2.1. Ağaç yapılı sınıflandırıcılar

Ağaç yapılı sınıflandırıcılar, daha doğru bir ifade ile ikili ağaç yapılı sınıflandırıcılar, bir T kümesini kendisi ile başlayarak, T'nin alt kümelerinin tekrar eden bir şekilde iki alt kümesine ayrılmasıyla oluşturulur. Bu süreç varsayımsal altı sınıflı bir ağaç için Şekil 2.6' da gösterilmiştir (Breiman vd., 1984).




**Şekil 2.6.** *Örnek Sınıflandırma Ağaç Yapısı*

Şekil 2.6.' da $T_2$ ve $T_3$ ayrıktır ve $T = T_2 \cup T_3$ şeklinde gösterilir. Aynı şekilde $T_4$ ve $T_5$ ayrıktır ve $T_2 = T_4 \cup T_5$ gösterilir. $T_3 = T_6 \cup T_7$ şeklinde devam eder. Ayrılmayan alt kümeler $T_6$, $T_8$, $T_{10}$, $T_{11}$, $T_{12}$, $T_{14}$, $T_{15}$, $T_{16}$ ve $T_{17}$ terminal alt kümeler olarak isimlendirilir. Şekilde görüldüğü gibi terminal alt kümeler dikdörtgen terminal olmayanlar ise daire ile gösterilir.

T' nin ayrılmalarından elde edilen terminal alt kümeler bir sınıf etiketi ile belirtilmiştir. Aynı sınıf etiketine sahip iki veya daha fazla terminal altkümesi olabilir. Sınıflandırıcıya karşılık gelen etiket, aynı sınıfa karşılık gelen bütün terminal alt kümelerinin bir araya getirilmesiyle elde edilir. Bu $A_1 = \mathrm{T}_{15}$, $A_2 = \mathrm{T}_{11} \cup \mathrm{T}_{14}$, $A_3 = \mathrm{T}_{10} \cup \mathrm{T}_{16}$, $A_4 = \mathrm{T}_6 \cup \mathrm{T}_{17}$, $A_5 = \mathrm{T}_8$, $A_6 = \mathrm{T}_{12}$ şekilde gösterilir. Ayrılmalar $\boldsymbol{x} = (x_1, x_2, \dots)$ ölçüm vektöründeki koşullar tarafından şekillendirilir.

Örneğin, T' nin $\mathrm{T}_2$ ve $\mathrm{T}_3$' e ayrılması Eşitlik (2.53)'teki; $\mathrm{T}_3$' ün $\mathrm{T}_6$ ve $\mathrm{T}_7$' ye ayrılması Eşitlik (2.54)'deki şekilde olur.

$$\mathrm{T}_2 = \{\boldsymbol{x};\ x_4 \leq 7\}, \qquad \mathrm{T}_3 = \{\boldsymbol{x};\ x_4 > 7\} \tag{2.53}$$

$$\mathrm{T}_6 = \{\boldsymbol{x} \in \mathrm{T}_3;\ x_3 + x_5 \leq -2\}, \quad \mathrm{T}_7 = \{\boldsymbol{x} \in \mathrm{T}_3;\ x_3 + x_5 > -2\} \tag{2.54}$$

Ağaç sınıflandırıcı, ölçüm uzayı $\boldsymbol{x}$' in (bağımsız değişkenler matrisi) ve sınıfları (bağımlı değişken vektörü) şu şekilde tahmin eder: İlk ayrılmanın tanımından, $\boldsymbol{x}$' in $\mathrm{T}_2$ veya $\mathrm{T}_3$' e gidip gitmeyeceği belirlenir. Örneğin, Eşitlik (2.1)'de $x_4 \leq 7$ ise $\boldsymbol{x}$, $\mathrm{T}_2$ kümesine, eğer $x_4 > 7$ ise $\mathrm{T}_3$ kümesine gider. Eğer $\boldsymbol{x}$, $\mathrm{T}_3$'e giderse Eşitlik (2.2)'deki gibi $\boldsymbol{x}$'in $\mathrm{T}_6$ veya $\mathrm{T}_7$'ye gidip gitmediği belirlenir. Bir terminal alt kümeye ulaşan $\boldsymbol{x}$' in öngörülen sınıfı, gittiği terminal alt kümesine tanımlanan sınıf etiketi ile verilir (Breiman vd., 1984).

Teorik olarak, T' nin alt kümesine *t* düğümü, T' nin kendisine (gözlem değerlerinin tümünü içine alan karar verme noktasına) $t_1$ kök düğümü (root node) denir. Her zaman verinin maksimum homojenliğe sahip iki bölüme ayrılması gerekir. Ayrılan bölümlere çocuk düğümü (child node) adı verilir (Timofeev, 2004). Karar vermenin daha gerçekleşmediği çocuk düğümleri bir sonraki aşamada aile düğümü olur ve tekrardan ikiye ayrılır. Amaç, çocuk düğümlerini ayırma kriterlerine göre tekrarlı ikili ayırmalarla terminal düğümlere ulaştırmaktır. Terminal düğümlerinden sonra ayrılma gerçekleşmez ve ağaçtaki en homojen düğüme ulaşılmış olur.

### 2.2.2. Sınıflandırma ağacının oluşturulması

Bir ağacın yapısını oluşturmak için üç önemli husus vardır. Bunlar;

1. Ayrılmaların seçimi
2. Terminal düğümünün ne zaman oluşacağının veya devam edeceğinin kararı
3. Her terminal düğümün bir sınıfa atanması

şeklinde ifade edilir (Breiman vd., 1984).

Ağaç modellerinin iki temel amacı vardır. Bunlardan birincisi sınıflama ve ikincisi tahmindir. Bu sayede ağaç modelleri, verinin tanımlanması ve gelecekteki verilerin tahmin edilmesi için pek çok araştırmacı tarafından tercih edilir.

N adet deney vasıtasıyla ölçümler bir başlangıç veri seti (Learning Sample) *L*'de toplanır. Varsayalım ki başlangıç veri seti *L*, J sınıflı bir problem için ele alınsın. $N_j$, *j* sınıfındaki durumların sayısı olsun. Genellikle önsel olasılıklar olarak $\pi(j)$ alınır ve $N_j/N$ ile hesaplanır. Önsel olasılıklar ya veriden tahmin edilir ($N_j/N$) ya da araştırmacı tarafından sağlanır.

Bir *t* düğümde, $\boldsymbol{x}_n \epsilon\, t$ olmak üzere *L*' deki toplam durum sayısı $N(t)$ ve *t*' deki j sınıf durum sayısı $N_j(t)$ ile göstersin. *L*' deki *j* sınıf durumlarının *t'* de bulunma oranı $N_j(t)/N_j$ ile bulunur. Önsel olasılıklar verildiğinde $\pi(j)$, *j* sınıf durumunun olasılığı olarak gösterilecektir. Bu yüzden bir durumun hem *j* sınıfında hem de *t* düğümünde bulunma olasılığı (resubstitution estimate) Eşitlik (2.55)'teki gibi gösterilir.

$$p(j,t) = \pi(j)\ N_j(t)/N_j \tag{2.55}$$

Herhangi bir durumun *t* düğümünde olma olasılığının $p(t)$ tahmini Eşitlik (2.56)'daki gibi gösterilir.

$$p(t) = \sum_j p(j,t) \tag{2.56}$$

Bir durumun *t* düğümünde olduğu bilindiğinde *j* sınıfında olma olasılığı ise Eşitlik (2.57)'de gösterilir.

$$p(j|t) = p(j,t)/p(t) \tag{2.57}$$

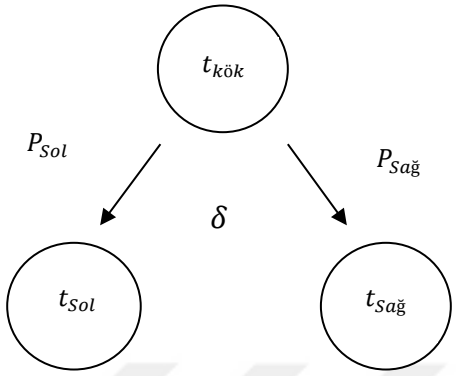
Burada,

$$\sum_j p(j|t) = 1 \tag{2.58}$$

Eşitlik (2.58)'deki gibi gösterilir.

Özel olarak $\pi(j) = N_j/N$ olarak alındığında, $p(j|t) = N_j(t)/N(t)$ olur.

Herhangi bir $t$ düğümü için, varsayalım ki, bu düğümü $t_{Sağ}$ ve $t_{Sol}$ olarak $P_{Sağ}$ oranında $t_{Sağ}$ düğümüne, $P_{Sol}$ oranında $t_{Sol}$ düğümüne ayıran bir $\delta$ aday ayrılması olsun. Bu durum Şekil 2.7.' deki gibi gösterilir.




**Şekil 2.7.** *CART için Bölme Algoritması*

Bu durumda bu ayırmanın iyiliği (goodness of the split) Eşitlik (2.59) ile tanımlanır.

$$\Delta \mathrm{i}(\delta, t) = \mathrm{i}(t) - P_{Sol}\mathrm{i}(t_{Sol}) - P_{Sağ}\mathrm{i}(t_{Sağ}) \quad \textbf{(2.59)}$$

Burada;

$\mathrm{i}(t)$ : $t$ düğümünün safsızlık (impurity) ölçüsü

$P_{Sol}$ : $t.$ düğümünden sol çocuk düğümüne atanan gözlemlerin oranı

$P_{Sağ}$ : $t.$ düğümünden sağ çocuk düğümüne atanan gözlemlerin oranı

$\mathrm{i}(t_{Sol})$ : Sol çocuk düğümünün safsızlık ölçüsü

$\mathrm{i}(t_{Sağ})$ : Sağ çocuk düğümünün safsızlık ölçüsü

olarak tanımlanır (Temel, 2004).

Ağaç oluşturmak için kullanılan dört temel yöntem vardır;

1. $\boldsymbol{x}$, A'nın elemanı mıdır? ($\boldsymbol{x} \in$ A?) şeklinde oluşturulan ikili sorular kümesi (A $\subset$ T)
2. Herhangi bir t düğümünün herhangi bir ayırması $\delta$ için $\phi(\delta, t)$ ayırma kriteri iyiliğinin hesaplanabilmesi,
3. Ayrılmayı durdurma kuralı,
4. Her bir terminal düğümünü bir sınıfa atama kuralı (Breiman, 1984).

### 2.2.3. Standart soru seti

Bir veri seti standart bir yapıya sahipse, $\varphi$ sorular seti de standartlaştırılabilir. Ölçüm matrisinin $\boldsymbol{x} = (x_1, x_2, \dots, x_M)$ yapıda olduğunu varsayalım. Burada *M* sonlu bir sayı ve $x_1, x_2, \dots, x_M$ değişkenleri kategorik tipte sıralı değişkenler grubudur. Standartlaştırılmış sorular seti $\varphi$ aşağıdaki gibi tanımlanır:

1. Her ayrılma sadece tek bir değişkenin değerine bağlıdır.
2. Sıralı yapıdaki her $x_m$ değişkeni için, $\varphi$ soru setindeki tüm soruların formu $\{x_m \leq \mathrm{c?}\}$ şeklindedir ve $c \in (-\infty, \infty)$' dır.
3. Eğer $x_m$, $\{b_1, b_2, \dots, b_L\}$ değerlerini alan kategorik bir değişken ise $\varphi$ soru seti $\{x_m \in S?\}$ şeklindeki sorulardan oluşur ve *S*, $\{b_1, b_2, \dots, b_L\}$' in tüm alt kümelerini belirtir.

Tüm *M* değişkenleri için 2 ve 3 aşamalarındaki görülen ayrılmalar standartlaştırılmış kümeyi oluşturmaktadır (Breiman, 1984).

### 2.2.4. Ayrılma ve ayrılmanın durma kuralı

Ayrılma kriterinin iyiliği başlangıçtaki bir safsızlık fonksiyonundan üretilmektedir. Bu fonksiyon Tanım 1 ile verilir.

*Tanım1: Bir safsızlık fonksiyonu $\phi$, bütün J-katlı sayıların kümesi $(p_1, p_2, \dots, p_J)$ üzerinde tanımlanmış, $p_j \geq 0$ ve $j = 1, \dots, J$ için $\sum_j p_j = 1$ ve aşağıdaki şartları sağlayan bir fonksiyondur.*

1. *$\phi$ sadece $(\frac{1}{j}, \frac{1}{j}, \dots, \frac{1}{j})$ noktasında maksimumdur.*
2. *$\phi$ minimuma sadece bu (1, 0, … , 0), (0, 1, 0, … , 0), (0, 0, …, 0,1) noktalarda erişir.*
3. *$\phi$, $p_1, p_2, \dots, p_j$' nin simetrik bir fonksiyonudur.*

*Tanım2: $\phi$ safsızlık fonksiyonu verildiğinde herhangi bir düğüm t' nin i(t) safsızlık ölçüsü Eşitlik (2.60)' daki gibi tanımlanır;*

$$\mathrm{i}(t) = \phi\big(p(1|t), p(2|t), \dots, p(J|t)\big) \qquad \textbf{(2.60)}$$

Herhangi bir t düğümü için, varsayalım ki, bu düğümü $t_{Sağ}$ ve $t_{Sol}$ olarak $P_{Sağ}$ oranında $t_{Sağ}$ düğümüne, $P_{Sol}$ oranında $t_{Sol}$ düğümüne ayıran bir $\delta$ aday ayrılması için,

safsızlıktaki azalmayı daha önce tanımlanan Eşitlik (2.59) ile gösterilmiştir. O zaman $\phi(\delta, t)$' nin ayrılma iyiliği $\Delta i(\delta, t)$ olur.

Yapılan ayrılmalardan sonra terminal düğümlerin bir kümesine ulaştığımızı varsayalım. Kullanılan ayrılmış kümeler, kullanıldıkları sıra ile birlikte ikili ağaç olarak tanımlanan *T* olsun. Geçerli düğümler kümesi terminal $\tilde{T}$ ile gösterilsin, $I(t) = i(t)p(t)$ ve ağaç safsızlığı Eşitlik (2.61) ile tanımlansın.

$$I(T) = \sum_{t \epsilon \tilde{T}} I(t) = \sum_{t \epsilon \tilde{T}} i(t)p(t) \tag{2.61}$$

$\Delta i(\delta, t)$' yı maksimum yapan ayrılmaları seçmenin, tüm ağaç safsızlığını minimum yapan $I(T)$ değerini seçmeye eşdeğer olduğu kolaylıkla görülmektedir. Herhangi bir $t \epsilon \tilde{T}$ düğümünü $t_{Sol}$ ve $t_{Sağ}$ olarak bölen bir $\delta$ ayrılması alınsın. Yeni ağaç olan $T'$ için safsızlık Eşitlik (2.62)'de gösterilir.

$$I(T') = \sum_{\tilde{T}-\{t\}} I(t) + I(t_{Sol}) + I\left(t_{Sağ}\right) \tag{2.62}$$

Safsızlıktaki azalma ise Eşitlik (2.63)' te gösterilir.

$$I(T) - I(T') = I(t) - I(t_{sol}) - I\left(t_{sağ}\right) \tag{2.63}$$

Bu sadece *t* düğümüne ve onu ayıran $\delta$' ya bağlıdır. Bu nedenle ağaç safsızlığındaki azalma ile *t* maksimizasyonu eşdeğerdir. Bu ifade Eşitlik (2.64)'teki gibi gösterilir.

$$\Delta I(\delta, t) = I(t) - I(t_{sol}) - I\left(t_{sağ}\right) \tag{2.64}$$

$t_{Sol}$ ve $t_{Sağ}$' a giden t düğüm kümesinin sırasıyla $P_{Sol}$ ve $P_{Sağ}$ oranları ise Eşitlik (2.65a)' da, oranların toplamı Eşitlik (2.65b)' de gösterilir ve Eşitlik (2.66)'daki gibi yazılır.

$$P_{Sol} = {p(t_{Sol})}\Big/{p(t)}\,, \qquad P_{Sağ} = {p(t_{Sağ})}\Big/{p(t)} \tag{2.65a}$$

$$P_{Sol} + P_{Sağ} = 1 \tag{2.65b}$$

$$\Delta I(\delta, t) = \left[i(t) - P_{Sol} i(t_{sol}) - P_{Sağ} i\left(t_{sağ}\right)\right] p(t) = \Delta i(\delta, t)\ \mathrm{p(t)} \tag{2.66}$$

$\Delta I(\delta, t)$, $\Delta i(\delta, t)$'den $p(t)$ faktörü ile farklılık gösterdiği için, aynı $\delta^*$ ayrılması iki ifadeyi de maksimum yapan ayrılmadır. Bu yüzden ayrılmış seçim prosedürü, genel ağaç safsızlığını en aza indirmek için tekrarlanan bir girişim olarak düşünülebilir. Ayrılmayı durdurmak için $\beta > 0$ şeklinde bir kesim noktası seçilir ve eğer $max_{\delta \epsilon S} \Delta I(\delta, t) < \beta$ ise bu t düğümü bir terminal düğüm olarak belirlenir (Breiman, 1984).

Sınıflama ağaçlarında çeşitli safsızlık ölçüleri (Gini, Twoing, Chi-square, G-square) kullanılabilir. Safsızlık ölçüleri literatürde en iyi ayırma yöntemleri olarak bilinirler. Genellikle, Gini çeşitlilik indeksi (Gini Diversity Index) ve Twoing kuralı tercih edilir (Temel, 2004).

### 2.2.5. Gini çeşitlilik indeksi

Sınıflandırma ağaçlarında herhangi bir düğümde en iyi ayırmayı bulmak için genellikle çok farklı kriterler vardır. Bunlardan bir ayırma yöntemi olarak aşağıdaki Eşitlik (2.67)'deki ayırma yöntemi kullanılır.

$$i(t) = -\sum_j p(j|t) \log[p(j|t)] \tag{2.67}$$

Bir diğeri düğümdeki safsızlığın ölçüsünü veren Gini çeşitlilik indeksidir (Gini Diversity Index). Bir t düğümü ile verilen tahmini sınıf olasılıkları $p(j|t)$, $(j = 1, \dots, J)$ olan safsızlık ölçüsü önceki bölümde Eşitlik (2.60)' taki gibi tanımlanmıştır. Gini çeşitlilik indeksi için Eşitlik (2.68) kullanılır.

$$i(t) = \sum_{j \neq i} p(j|t)\, p(i|t) \tag{2.68}$$

Gini indeksi Eşitlik (2.69a)' daki gibi de yazılabilir.

$$i(t) = \left( \sum_j p(j|t) \right)^2 - \sum_j p^2(j|t) \tag{2.69a}$$

$\sum_j p(j|t) = 1$ olduğundan;

$$= 1 - \sum_j p^2(j|t) \tag{2.69b}$$

Eşitlik (2.69b) olarak gösterilebilir.

İki sınıf kriteri için indeks Eşitlik (2.70a) kullanılarak Eşitlik (2.70b) elde edilir.

$$\mathrm{i}(t) = 1 - \sum_{j=1}^{2} p^2(j|t) = 1 - [p^2(1|t) + p^2(2|t)] \qquad \textbf{(2.70a)}$$

$p(1|t) + p(2|t) = 1$ eşitliğinin her iki tarafının karesi alınırsa;

$p^2(1|t) + p^2(2|t) = 1 - 2[p(1|t).p(2|t)]$ olur. Eşitlik (2.70a)' da yerine yazılırsa;

$$\mathrm{i}(t) = 1 - \left[1 - 2[p(1|t).p(2|t)]\right]$$
$$\mathrm{i}(t) = 2p(1|t)p(2|t) \qquad \textbf{(2.70b)}$$

Eşitlik (2.70b) şeklinde olur.

Son olarak $p_1, \dots, p_J$' nin $\phi(p_1, \dots, p_J)$ fonksiyonu olarak kabul edilen Gini indeksinin negatif olmayan katsayılarla ikinci dereceden bir polinom olduğunu unutmamalıyız. Bu nedenle $r + s = 1,\ r \geq 0,\ s \geq 0$,

$$\phi(rp_1 + sp_1', rp_2 + sp_2', \dots, rp_J + sp_J' \geq r\phi(p_1, \dots, p_J) + s\phi(p_1', \dots, p_J')$$

bu $\delta$'nın herhangi bir ayrılması için $\Delta\mathrm{i}(\delta, t) \geq 0$ sağlar.
Sadece $p(j|t_R) = p(j|t_L) = p(j|t),\quad j = 1, \dots, J$ gerçekleşirse $\Delta\mathrm{i}(\delta, t) = 0$ olur (Breiman, 1984).

## 2.3.Karışıklık Matrisi (Confusion Matrix)

Sınıflandırma başarısını ölçen yardımcı ölçülerin tanımları için Tablo2.2' de verilen karışıklık matrisinden (hata matrisi) yararlanılabilir.

**Tablo 2.2.** *Üç Sınıf için Karışıklık Matrisi*

| | | **Gerçek Sınıf** | | | |
|---|---|---|---|---|---|
| | | **A** | **B** | **C** | **Toplam** |
| **Tahmin Sınıfı** | **A** | $TP_{AA}$ | $E_{AB}$ | $E_{AC}$ | D= $TP_{AA}$+ $E_{AB}$+ $E_{AC}$ |
| | **B** | $E_{BA}$ | $TP_{BB}$ | $E_{BC}$ | F= $E_{BA}$ +$TP_{BB}$ + $E_{BC}$ |
| | **C** | $E_{CA}$ | $E_{CB}$ | $TP_{CC}$ | G= $E_{CA}$+ $E_{CB}$+ $TP_{CC}$ |
| | **Top.** | H= $TP_{AA}$+ $E_{BA}$+ $E_{CA}$ | I= $E_{AB}$+ $TP_{BB}$+ $E_{CB}$ | J= $E_{AC}$+ $E_{BC}$+ $TP_{CC}$ | T=D+F+G=H+I+J |

Tablo 2.2.'deki kısaltmalar aşağıdaki gibi tanımlanabilir.

TP: True positive (Doğru pozitif) Gerçekte pozitif durumun tahmin sonucunda pozitif çıkmasıdır. ($TP_{AA}$, $TP_{BB}$, $TP_{CC}$)

FP: False positive (Yanlış pozitif) Gerçekte negatif durumun tahmin sonucunda pozitif çıkmasıdır. A için, $FP_A=E_{AB}+E_{AC}$; B için, $FP_B=E_{BA}+E_{BC}$; C için, $FP_C=E_{CA}+E_{CB}$ şeklinde gösterilir.

TN: True negative (Doğru negatif) Gerçekte negatif sonucun tahmin sonucunda negatif çıkmasıdır.

A için, $TN_A= TP_{BB}+ E_{BC}+ E_{CB} +TP_{CC}$;

B için, $TN_B= TP_{AA}+ E_{AC}+ E_{CA} +TP_{CC}$;

C için, $TN_C= TP_{AA}+ E_{AB}+ E_{BA} +TP_{BB}$

şeklinde ifade edilir.

FN: False negative (Yanlış negatif) Gerçekte pozitif sonucun tahmin sonucunda negatif çıkmasıdır. A sınıfı için, $FN_A= E_{BA}+ E_{CA}$; B için, $FN_B =E_{AB}+ E_{CB}$; C için, $FN_B=E_{AC}+ E_{BC}$ şeklinde gösterilir.

Doğruluk (Accuracy): Doğru sınıflandırılmış sayısının toplama bölünmesi ile bulunur. Formül olarak $TP_{AA} + TP_{BB} + TP_{CC} / T$ gösterilir.

Duyarlılık (Sensitivity): Doğru pozitif oran şeklinde yorumlanır. Pozitif kararın doğru olma olasılığı olarak düşünülebilir. Bir gözlemin gerçekte A sınıfında iken tahmin sonucunda A sınıfında olması oranı $TP_{AA} / H$, gerçekte B sınıfında iken tahmin sonucunda B sınıfında olması oranı $TP_{BB} / I$, gerçekte C sınıfında iken tahmin sonucunda C sınıfında olması oranı $TP_{CC} / J$ şeklinde hesaplanır.

Belirleyicilik-Seçicilik-Özgüllük (Specificity): Negatif doğru oran şeklinde yorumlanır. Negatif kararın doğru olma olasılığı olarak da düşünülebilir. Gerçekte A sınıfında değil iken tahmin sonucunda A olmaması oranı ( B veya C olması) $TN_A / FP_A + TN_A$ şeklinde hesaplanabilir. Gerçekte B sınıfında değil iken tahmin sonucunda B sınıfında olmaması oranı ( A veya C olması) $TN_B / FP_B + TN_B$ şeklinde hesaplanabilir. Aynı şekilde C için de $TN_C / FP_C + TN_C$ şeklinde belirleyicilik hesaplanır.

Duyarlılık ve belirleyicilik değerlerinin artmasıyla, geçerliliğin arttığı kabul edilir.

Pozitif tahmin değeri-PPV(Positive predictive value) : Tahmin sonucuna göre belirlenen pozitifler içerisindeki doğru pozitiflerin oranıdır. A sınıfı için, $TP_{AA}/D$; B sınıfı için, $TP_{BB}/F$; C sınıfı için, $TP_{CC}/G$ şeklinde hesaplanır.

Negatif tahmin değeri- NPV ( Negative predictive value) : Tahmin sonucuna göre belirlenen negatifler içerisindeki doğru negatiflerin oranıdır. A sınıfı için, $TN_A$/ F+G; B sınıfı için, $TN_B$/ D+G; C sınıfı için, $TN_C$/ D+F şeklinde hesaplanır.

Yaygınlık (Prevalence) : Pozitif sınıfın gerçek sıklığını verir. A sınıfı için, H/T; B sınıfı için I/T; C sınıfı için J/T olarak hesaplanır.

Cohen's Kappa: Diğer ölçülerden biri de Kappa istatistiğidir. Aşağıda verilen Eşitlik (2.71) ile hesaplanır.

$$k = \frac{P_0 - P_e}{1 - P_e} = 1 - \frac{1 - P_0}{1 - P_e} \quad \textbf{(2.71)}$$

$P_0 = \frac{TP_{AA}+TP_{BB}+TP_{CC}}{T}$ ve bu uyumun şansa bağlı ortaya çıkma olasılığı

$P_e = \left(\frac{H}{T} \times \frac{D}{T}\right) + \left(\frac{I}{T} \times \frac{F}{T}\right) + \left(\frac{J}{T} \times \frac{G}{T}\right)$ şeklinde hesaplanır (Altman ve Bland, 1994).

Kappa değerlerini Landis ve Koch Tablo 2.3.'deki gibi yorumlamıştır (Landis ve Koch,1977).

**Tablo 2.3.** *Kappa Değerlerinin Yorumu*

| $\kappa$ değeri | Yorum |
|---|---|
| $< 0$ | Şansa bağlı olan uyumdan daha kötü uyumlu |
| $0{,}01 - 0{,}20$ | Önemsiz derecede uyumlu |
| $0{,}21 - 0{,}40$ | Zayıf derecede uyumlu |
| $0{,}41 - 0{,}60$ | Orta derecede uyumlu |
| $0{,}61 - 0{,}80$ | Iyi derecede uyumlu |
| $0{,}81 - 1{,}00$ | Çok iyi derecede uyumlu |

Bu değer (-1, +1) arasında değer almaktadır. +1 değeri pozitif anlamda uyumluluk, 0 değeri uyumsuzluk ve -1 değeri negatif anlamda uyumsuzluk olarak yorumlanabilir.

## 3.UYGULAMA

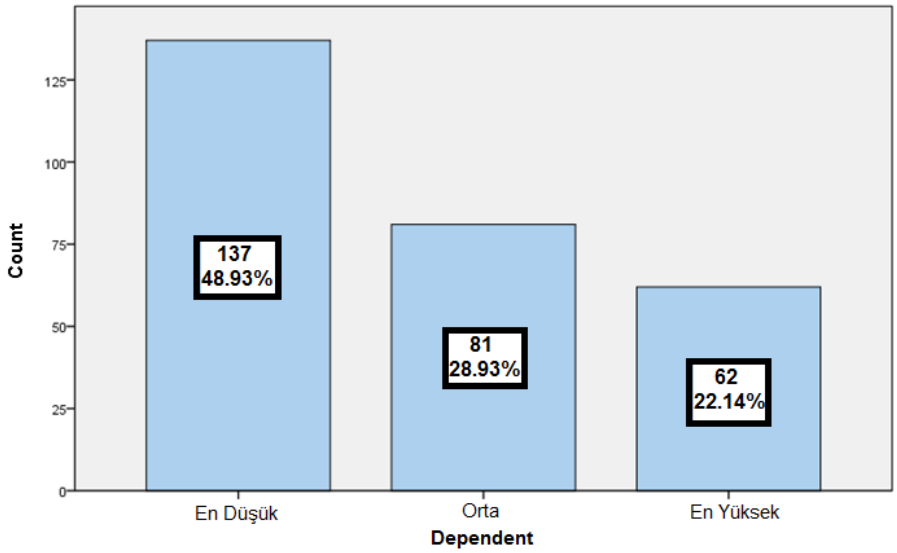
Bu çalışmadaki veri, yaygın olarak kullanılan bir internet sitesinde yayınlanan satılık konut ilanları kullanılarak elde edilmiştir. Veri, 2018 yılının ekim ayından 2019 yılının mayıs ayına kadar olan aylarda Eskişehir ili için internet sitesinde yayınlanan 280 adet satılık konutların belirli özelliklerini içermektedir. Bu özellikler; konutların fiyatı (price), metrekaresi (land), yaşı (age), bahçe durumu (type), oda sayısı (bedroom), banyo sayısı (bathroom), garaj durumu (garage), çevresinde sosyal ortamların (hastane, okul, alışveriş merkezi vb.) durumu (amenities), balkon durumu (balcony), dairenin kaçıncı katta olduğu (floor), asansör durumu (elevator) bulunduğu ilçenin *Odunpazarı* veya *Tepebaşı'nda* olması durumu (district) şeklinde seçilmiştir.

Bağımlı değişken olarak konut fiyatlarının metrekarelerine bölünmesiyle bulunan oran (konut birim fiyatı, TL/$m^2$) seçilmiştir. Bağımlı değişken konut birim fiyatı (TL/$m^2$) sıralı yapıda 3 kategoriye ayrılmıştır. Kategorileri belirlemek için 11 Mart 2019 tarihinde TC Merkez Bankası'nın EVDS internet sitesinde yayınlanan 2018 yılına ait Türkiye'deki konut birim fiyatları (TL/$m^2$) kullanılmıştır. EVDS' nin internet sitesinde Konut birim fiyatı, en düşük 2.118,52 (TL/$m^2$) ve en yüksek 2.314,83 (TL/$m^2$) olarak verilen değerler sıralı kategorilerin oluşturulması için kullanılmıştır.

Konut birim fiyatının 2.118,52 (TL/$m^2$)'den küçük olduğu değerler birinci kategori (En düşük); 2.118,52 (TL/$m^2$)'den büyük ve 2.314,83 (TL/$m^2$)'den küçük olduğu değerler ikinci kategori (Orta); 2.314,83 (TL/$m^2$)'den büyük olduğu değerler üçüncü kategori (En Yüksek) olarak tanımlanmıştır (EVDS, Verinin Merkezi).

Bu çalışmada, Eskişehir ilindeki konut birim fiyatları üzerinde etkili olan potansiyel değişkenler seçilerek, sıralı lojistik regresyon ve sınıflandırma ağaçları yöntemleri ile konut birim fiyatı için sınıflandırma modelleri oluşturularak, en iyi performans gösteren yöntemi seçmek amaçlanmıştır. Bu yöntemlerden sıralı lojistik regresyon, sağlaması gereken varsayımlar nedeniyle değişken seçiminde öncelikli etkiye sahip olmuştur. Buna göre, tüm değişkenlerin modelde yer aldığı sıralı lojistik regresyon analizi için paralellik testi sonucunda ki-kare test istatistiği 10 serbestlik derecesiyle 30,678 olarak elde edilmiştir. Ki-kare istatistiğine karşılık gelen p olasılığı (=0,001), alfa=0,05 anlamlılık düzeyinde reddedilmiştir. Bir sonraki aşamada, paralellik varsayımını sağlamayan değişkenler göz ardı edilerek konutun bahçe durumu, asansör durumu, bulunduğu ilçe bilgisi değişkenleri ile çalışılmaya karar verilmiştir. Uygulama için SPSS 24.0 ve R-Studio 1.0.153 (Mac OS X 10_12_6) paket programları kullanılmıştır.

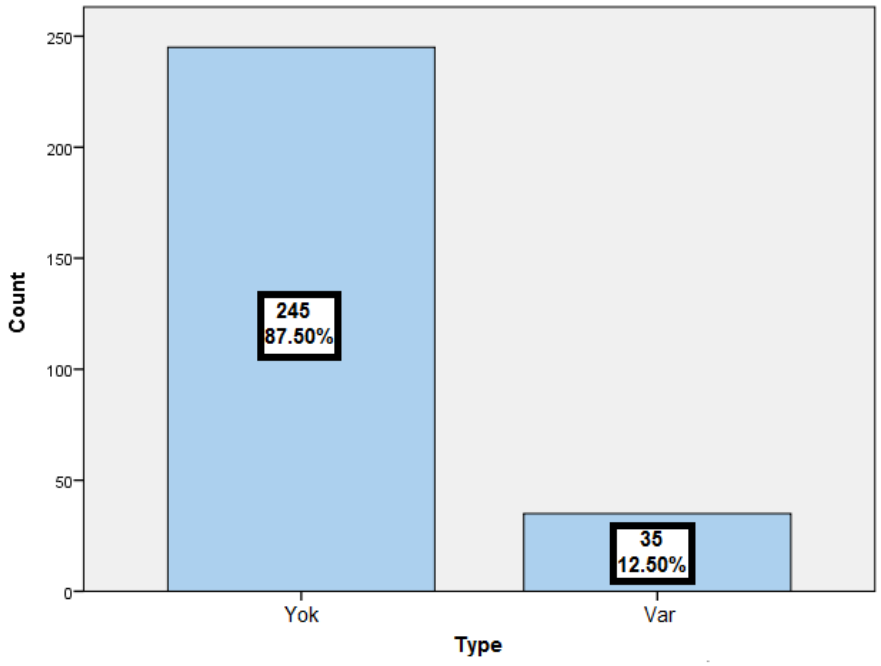
Bağımlı değişken olarak belirlenen konut birim fiyatı için Şekil 3.1.' de gözlemlerin oranı ve sayısı gösterilmiştir. En düşük kategorisi %48,93 oranında 137 adet gözlemden, Orta kategorisi %28,93 oranında 81 adet gözlemden ve En yüksek kategorisi de %22,14 oranında 62 adet gözlemden oluşmaktadır.




**Şekil 3.1.** *Bağımlı Değişken için Sütun Grafiği*

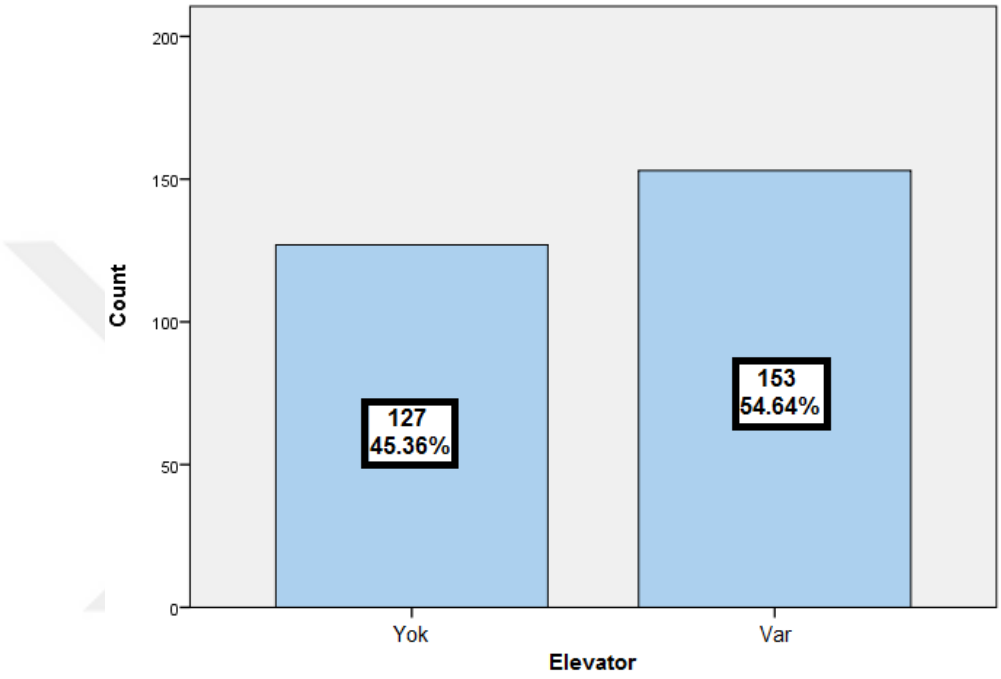
Bağımlı değişken olan konut birim fiyatını etkilediği düşünülen bağımsız değişkenler; bahçe durumu, asansör, ilçe bilgisi olarak tanımlanmıştır. Bu üç değişken paralellik varsayımını sağladığı için seçilmiştir. Bu değişkenler için varsayımın sonuçları ileride gösterilecektir.

Bahçe değişkeni, seçilen konut bahçeliyse *1=var*; bahçeli değilse *0=yok* şeklinde 2 kategoriden oluşmaktadır. Şekil 3.2.'de gösterildiği gibi konutun bahçesi *yok* kategorisinde %87,50 oranında 245 adet gözlemden, konutun bahçesi *var* kategorisinde ise %12,50 oranında 35 adet gözlemden oluşmaktadır.



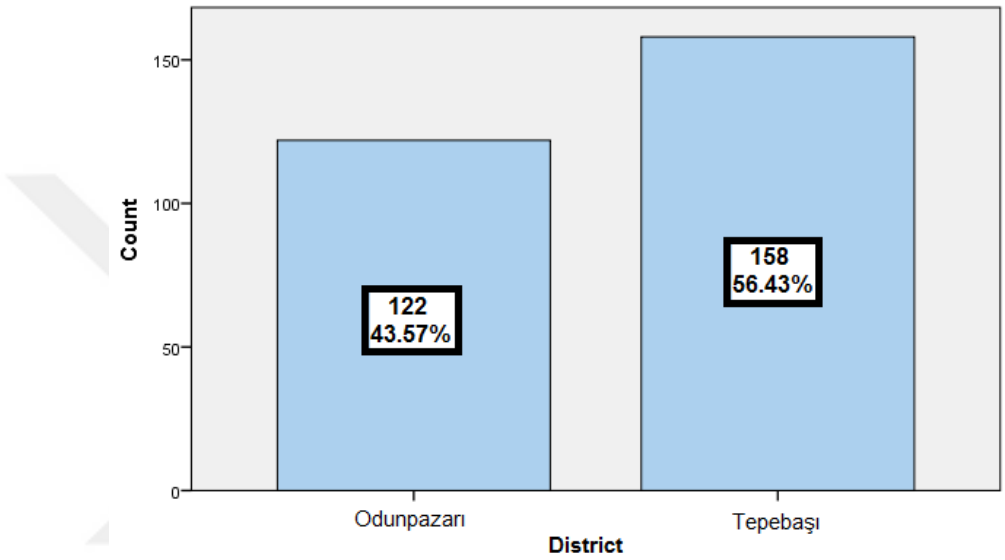

**Şekil 3.2.** *Bahçe Durumu Bağımsız Değişkeni için Sütun Grafiği*

Asansör değişkeni, seçilen konutun asansörü varsa *1=var* asansörü yoksa *0=yok* şeklinde 2 kategoriden oluşmaktadır. Şekil 3.3.'de gösterildiği gibi konutun asansörü *yok* kategorisinde %45,36 oranında 127 adet gözlemden, asansörü *var* kategorisinde ise %54,64 oranında 153 adet gözlemden oluşmaktadır.




**Şekil 3.3.** *Asansör Durumu Bağımsız Değişkeni için Sütun Grafiği*

İlçe bilgisi değişkeni, seçilen konutun bulunduğu ilçenin *Odunpazarı* veya *Tepebaşı* ilçesinde olması durumuna göre 2 kategoriden oluşmaktadır. Şekil 3.4.'te gösterildiği gibi konutun bulunduğu ilçe *Odunpazarı* ise %43,57 oranında 122 adet gözlemden, konutun bulunduğu semt *Tepebaşı* ilçesinde ise %56,43 oranında 158 adet gözlemden oluşmaktadır.




**Şekil 3.4.** *İlçe Bilgisi Bağımsız Değişkeni için Sütun Grafiği*

Uygulamada, birim konut fiyatları üzerinden kategorik hale dönüştürülen bağımlı değişkenin, açıklayıcı değişkenlerle parametrik ve parametrik olmayan modellerinin tahmin performanslarının karşılaştırılması yapılacaktır. Bu amaçla parametrik yöntemlerden lojistik regresyonun özel bir şekli olan sıralı lojistik regresyon kullanılmıştır. Öte yandan bağımlı değişkenin sürekli değişken olmadığı durumlarda kullanılan sınıflandırma ağaçları ile parametrik olmayan bir sınıflandırma yapılacaktır. Her iki yöntemin de performans başarılarının ortaya konması amacıyla 5-katlı çapraz geçerlilik sınaması ile modellerin geçerliliği sınanacaktır. Elde edilen sonuçlar kullanılarak en iyi modele ilişkin istatistiksel çıkarımlar yapılacaktır.

### 3.1.Sıralı Lojistik Regresyon için Uygulama

Sıralı lojistik regresyon yapabilmek için gerekli varsayımlar aşağıda sıralanmıştır.

*Varsayım 1:* Bağımlı değişken sıralı düzeyde ölçülmelidir. Sıralı değişken için örnek olarak likert maddeleri içeren kategoriler seçilebilir. Likert tipi ölçek için kesinlikle katılıyorum, katılıyorum, kararsızım, katılmıyorum ve kesinlikle katılmıyorum şeklinde 5 kategorili olabilir.

*Varsayım 2:* Sürekli, sıralı ve kategorik (ikili değişkenleri içeren) olan bir veya daha fazla bağımsız değişkenler vardır. Ancak sıralı bağımsız değişkenler kategorik olarak değerlendirilmelidir. Bu kriteri sağlayan sürekli değişken örnekleri yaş (yıl olarak ölçülen), gelir (TL cinsinden ölçülen), sınav performansı (0 ile 100 arasında puan olarak), ağırlık (kg cinsinden) v.b. olarak verilebilir.

*Varsayım 3:* Çoklu bağlantı yoktur. Çoklu bağlantı sorunu, birbiriyle büyük ölçüde ilişkili iki veya daha fazla bağımsız değişken olduğunda oluşur. Bu, sıralı lojistik regresyonun hesaplanmasında hangi değişkenin bağımlı değişken olduğunu anlama sorunlarına yol açar. Sıralı lojistik regresyonda çoklu bağlantının olup olmadığının belirlenmesi önemli bir adımdır. Bu varsayım için test yapmak, kategorik değişkenler için kukla değişkenler (dummy variable) oluşturulmasını gerektirir.

*Varsayım 4:* Sıralı lojistik regresyon modelinin temel bir varsayımı olan orantısal oran modeli her bağımsız değişken, sıralı bağımlı değişkenin her kümülatif (birikimli) bölünmesinde aynı etkiye sahiptir.

Sıralı lojistik regresyonda, bağımlı değişkenin kategorileri arasında bir sıra olması sebebiyle, sıradan lojistik regresyondan farklı bir çözümleme yapılır. Sıralı lojistik regresyonda, bağımlı değişkenin her bir kategorisinin (sınıf) veya o kategoriye kadar olanların olasılık değerleri elde edilmeye çalışılır. Bu yönüyle sıradan lojistik regresyondan farklılık göstereceğinden, özellikle tahminci (predictor) değişkenlerin katsayılarına ilişkin yorumlarda dikkat edilmesi gerekmektedir. SPSS' de verilerin sınıflayıcı, sıralayıcı ve sürekli olarak ölçek tipleri belirlendikten sonra analiz için uygun şartlar sağlanmış olur.

Sıralı lojistik regresyonda, bağımlı değişkenin hangi değerinin referans değer olarak alındığı sorusu önemlidir. Bu anlamda SPSS, son kategoriyi varsayılan olarak referans değeri kabul etmektedir. Öte yandan R programı, SPSS' deki gibi referans değerini son değer olarak varsayılan kabul etmemektedir. Bu yüzden aynı sonuçları almak

için, her bir tahminci değişken ve bağımlı değişken için yeniden referans belirleme yapılmalıdır.

Tablo 3.1.' de, her iki model için (modelde sadece $\beta_0$ sabit terim varken ve kısıtsız model), hesaplanan -2log olabilirlik (-2logL) değeri, ki-kare değerleri, serbestlik derecesi, olasılık değeri bulunmaktadır. Bu iki modelin -2logL değerleri arasındaki farkın anlamlılığı ki-kare testi ile sınanır. Bir başka deyişle bu test ile -2logL istatistiği, bağımlı değişkenin açıklanmayan varyansının anlamlılığını sınamada kullanılır. Buradaki log olabilirlik değeri, bağımlı değişkenin tahminciler tarafından tahmin edilme olasılığını gösterirken, -2logL değeri yaklaşık olarak ki-kare dağılımına uygunluğu gösterir. Burada, -2logL istatistiği bu sonuçlara göre p=0,000 değeriyle, alfa=0,05 anlamlılık düzeyinde $H_0$ hipotezi reddedilir (p<0,05 olduğu için). Dolayısıyla, tahmincilerin olduğu kısıtsız model, istatistiksel olarak anlamlı bir katkı sağlamış sayılır.

**Tablo 3.1**. *Bahçe, Asansör ve İlçe Değişkenleri ile Kurulan Model Uyumu*

| Model | -2logL değeri | ki-kare değeri | Serbestlik derecesi | p |
|---|---|---|---|---|
| Kısıtlı Model | 91,798 | | | |
| Genel | 62,357 | 29,441 | 3 | 0,000 |

Tablo 3.2.' de uyum iyiliği testi sonucunda iki adet ki-kare değeri ve bunlara ilişkin olasılık değerleri verilmiştir. Uyum iyiliği testinde $H_0$ hipotezinde modelin veriyle uyumlu olduğu hipotezi öne sürülürken, alternatif hipotezde modelin veriyle uyumlu olmadığı iddia edilir. Tablo 3.2.' de görüldüğü gibi Pearson ki-kare değerine karşılık gelen p değeri (0,419) anlamlılık düzeyi olarak kabul edilen alfa=0,05' dan büyük olduğundan model veri ile uyumludur sonucuna ulaşılır.

**Tablo 3.2.** *Bahçe, Asansör ve İlçe Değişkenleri ile Kurulan Modelin Uyum İyiliği Testi*

| Model | ki-kare değeri | Serbestlik derecesi | p |
|---|---|---|---|
| Pearson | 11,300 | 11 | 0,419 |
| Sapma | 13,536 | 11 | 0,260 |

Modelin uyum iyiliği $R^2$ değerleri (Pseudo R-Square) aracılığıyla da incelenmiştir. $R^2$ değerleri ile bağımlı değişkenin yüzde kaçının bağımsız değişkenler tarafından açıklandığı gösterilmektedir. Ancak $R^2$ değerleri lojistik regresyon için iyi bir ölçüt olmadığından bu analizler için düşük çıkmaktadır. Tablo 3.3.' te görüldüğü gibi Cox ve Snell $R^2$ değeri 0,100 Nagelkerke $R^2$ değeri 0,114 ve McFadden $R^2$ değeri 0,050 olarak elde edilmiştir.

**Tablo 3.3**. *Bahçe, Asansör ve İlçe Değişkenleri ile Kurulan SLR Modeline ait $R^2$ değerleri*

| | |
|---|---|
| Cox and Snell | 0,100 |
| Nagelkerke | 0,114 |
| McFadden | 0,050 |

Olabilirlik oran testi ile anlamlı çıkan modelimizde (p=0,000<0,05) katsayıları ve odds oranlarını yorumlamak için paralellik varsayımının test edilmesi gerekir. Paralellik varsayımı için parametrelerin kategorilerin her birinde eşitliği test edilerek paralel doğrular varsayımının sağlanıp sağlanmadığına bakılır.

Tablo 3.4.'te olabilirlik oran testi ile incelendiğinde p=0,080 değeri alfa=0,05 anlamlılık düzeyinden büyük olduğu için $H_0$ hipotezi kabul edilir ve paralellik varsayımı kabul edilmiş olur. Aksi durumlarda açıklayıcı değişken sayısı özellikle büyük olduğunda (Brant, 1990) , örnek hacmi büyük olduğunda (Allison, 1999; Clogg ve Shihadhe, 1994) veya modelde bir sürekli açıklayıcı değişken olduğunda (Allison, 1999) Orantısal oran modeli varsayımı reddedilebilir.

**Tablo 3.4.** *Bahçe, Asansör ve İlçe Değişkenleri ile Kurulan SLR Modeline İlişkin Paralellik Testi Sonuçları*

| Model | -2LogL değeri | Ki-kare değeri | Serbestlik derecesi | p değeri |
|---|---|---|---|---|
| Kısıtsız | 62,357 | | | |
| Genel | 55,604 | 6,753 | 3 | 0,080 |

Paralellik varsayımının hangi değişkenler için sağlanıp sağlanmadığını Wald testi ile de incelememiz mümkündür.

Paralellik varsayımı sağlanan sıralı lojistik regresyonda orantısal oran modeli veriye uygulanırsa Tablo3.5.' teki sonuçlar elde edilir.

**Tablo 3.5**. *Bahçe, Asansör ve İlçe Değişkenleri ile Kurulan SLR Modeline İlişkin OOM Katsayısı, Standart Hata, Wald istatistiği, Odds Oranı Tahminleri ve p Değerleri*

| **Konut Birim Fiyatı** | **Değişken** | **Katsayı** $(\widehat{\boldsymbol{\beta}})$ | **Standart Hata** | **Wald Testi** | **Odds Oranı** $(e^{\widehat{\beta}})$ | **p değeri** | **95% Güven Aralığı** | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | **Alt Sınır** | **Üst Sınır** |
| Karşılaştırma1 | Eşik 1 | -1,011 | 0,371 | 7,405 | 0,363 | 0,007 | -1,738 | -0,283 |
| Karşılaştırma2 | Eşik 2 | 0,405 | 0,367 | 1,220 | 1,499 | 0,269 | -0,314 | 1,124 |
| | Bahçe= yok | -0,780 | 0,344 | 0,344 | 0,458 | 0,023 | -1,454 | -0,106 |
| | Bahçe= var | 0[a] | . | . | . | . | . | . |
| | Asansör= yok | -1,009 | 0,237 | 18,155 | 0,364 | 0,000 | -1,474 | -0,545 |
| | Asansör= var | 0[a] | . | . | . | . | . | . |
| | İlçe= *Odunpazarı* | 0,476 | 0,234 | 4,143 | 1,609 | 0,042 | 0,018 | 0,935 |
| | İlçe= *Tepebaşı* | 0[a] | . | . | . | . | . | . |

a: Referans değeri olarak alınmıştır.

Paralellik varsayımı sağlandığına göre orantısal oran modeli için Tablo 3.5.'teki değişkenler yorumlanırsa;

Karşılaştırma 1 (Düşük Kategoriye Karşı Orta ve Yüksek Kategori): Burada her bir bağımsız değişkendeki kategorilerin değişimi, orta ve yüksek kategorisinin, düşük kategorisine karşı odds değerini nasıl değiştirdiği aşağıdaki gibi yorumlanır:

Bahçesi olan bir konutun (type=1), konut birim fiyatı bakımından düşük kategoride olması yerine daha yüksek (orta ve yüksek) kategoride olma olasılığı, bahçesi olmayan bir konuta (type=0) göre 0,458 kat daha fazladır.

Asansörü olan bir konutun (elevator=1), konut birim fiyatı bakımından düşük kategoride olması yerine daha yüksek (orta ve yüksek) kategoride olma olasılığı, asansörü olmayan bir konuta (elevator=0) göre 0,364 kat daha fazladır.

*Tepebaşı* ilçesinde bulunan bir konutun (district=*Tepebaşı*) konut birim fiyatı bakımından düşük kategoride olması yerine daha yüksek (orta ve yüksek) kategoride olma olasılığı *Odunpazarı* ilçesinde bulunan bir konuta (district=*Odunpazarı*) göre 1,609 kat daha fazladır.

Karşılaştırma 2 (Düşük ve Orta Kategoriye Karşı Yüksek Kategori): Burada her bir bağımsız değişkendeki kategorilerin değişimi, yüksek kategorisinin, düşük ve orta kategorilerine karşı odds değerini nasıl değiştirdiği yorumlanır. Bu karşılaştırma için de odds oranları aynı olduğundan benzer şekilde yorum yapılabilir. Ancak bağımlı değişkenin bu kategorisi istatistiksel olarak alfa=0,05 anlamlılık düzeyinde önemli bulunmadığından yorumlama yapılmayacaktır.

### 3.2. Sınıflandırma Ağacı için Uygulama

Bu bölümde, popüler bir yöntem olan sınıflama ağaçları olarak isimlendirilen yönteme dayalı analiz sonuçları verilecektir. Bu yöntemde sıralı lojistik regresyondaki gibi varsayımlar bulunmamaktadır.

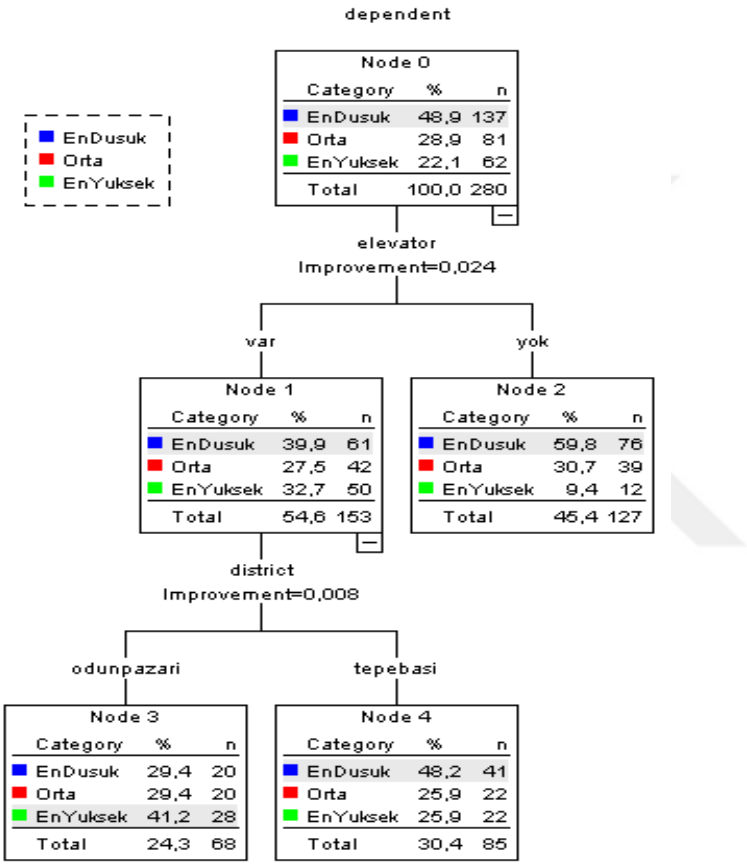
Konut birim fiyatı üzerinde etkili olduğu düşünülen asansör durumu, bahçe durumu, ilçe bilgisi değişkenlerinin sınıflandırma ağacı algoritması ile modellemesi gerçekleşmiştir. Bu algoritma ile elde edilen sınıflandırma ağacı Şekil 3.5.' te verilmiştir.

Şekil 3.5. incelendiğinde konut birim fiyatı olarak belirlenen bağımlı değişken için 280 tane konutun 137 tanesi En Düşük, 81 tanesi Orta, 62 tanesi En Yüksek kategoride yer almaktadır. Konut birim fiyatını belirleyen değişkenler olarak asansör durumu (elevator) ve ilçe bilgisi (district) yer almaktadır. Tablo 3.6.' de gösterildiği gibi bu belirleyici değişkenler konut birim fiyatını 3 ayrı sınıfa ayırmaktadır. Burada her bir terminal düğüm sınıf olarak adlandırılmıştır.

**Tablo 3.6**. *Konut Birim Fiyatlarının Belirleyicileri ve Sınıfları*

| Sınıf | Düğüm | Asansör | İlçe |
|---|---|---|---|
| 1 | 2 | Yok | |
| 2 | 3 | Var | *Odunpazarı* |
| 3 | 4 | Yok | *Tepebaşı* |

Sınıf 1; 127 adet konuttan oluşup konutların %45,4'ünü oluşturmaktadır. Sınıf 2; 68 adet konuttan oluşup konutların %24,3 ünü oluşturmaktadır. Sınıf 2'nin temel özelliği asansörü olup *Odunpazarı* ilçesinde bulunan konutları içermesidir. Sınıf 3 ise 85 adet konuttan oluşup konutların %30,4'ünü oluşturmaktadır. Sınıf 3'ün temel özelliği asansörü olup *Tepebaşı* ilçesinde bulunan konutları içermesidir.




**Şekil 3.5**. *Bahçe, Asansör ve İlçe Değişkenleri ile Kurulan Sınıflama Modeline ait Sınıflandırma Ağacı*

Konut birim fiyatı için ilk sınıflamanın asansör durumuna göre olduğu Şekil3.5.'te görülmektedir. Asansörü olan konutlar %54,6 oranıyla Düğüm 1'e (Node 1), asansörü olmayan konutlar %45,4 oranıyla Düğüm 2'e (Node 2) ayrılmıştır.

Düğüm 2'de asansörü olmayan konutların konut birim fiyatına göre en düşük kategorideki oranı %59,8 orta kategorideki oranı %30,7 en yüksek kategorideki oranı %9,4 olarak görülmektedir. Konut birim fiyatını konutların asansörünün olmaması büyük ölçüde etkilemektedir. Asansörü olmayan konutların, konut birim fiyatının en düşük

kategorisindeki oranı %59,8 olup, bu diğer kategoriler arasındaki en yüksek orandır. Bu durum asansörü olmayan konutların birim fiyatının düşük olduğu şeklinde yorumlanabilir.

Düğüm 1'de asansörü olan konutların konut birim fiyatına göre en düşük kategorideki oranı %39,9 orta kategorideki oranı %27,5 en yüksek kategorideki oranı %32,7 olarak görülmektedir. Asansörü olan konutların yaklaşık konut birim fiyatlarındaki oranlarının birbirine yakın olduğu görülmektedir.

Asansörü olup *Odunpazarı* ilçesinde bulunan konutlar %24,3 oranıyla Düğüm 3'e (Node 3), asansörü olup *Tepebaşı* ilçesinde bulunan konutlar %30,4 oranıyla Düğüm 4'e (Node 4) ayrılmıştır.

Düğüm 3'te asansörü olup, *Odunpazarı* ilçesinde bulunan konutların konut birim fiyatına göre 'En Düşük' ve 'Orta' kategorideki oranları aynı olarak gerçekleşirken (%29,4), 'En Yüksek' kategorideki oranı bunlara göre daha fazladır (%41,2). Bu durum asansörü olup *Odunpazarı* ilçesinde bulunan konutların birim fiyatının yüksek olduğu şeklinde yorumlanabilir.

Düğüm 4'te asansörü olup, *Tepebaşı* ilçesinde bulunan konutların konut birim fiyatına göre 'En Düşük' kategorideki oranı %48,2 olup, 'Orta' ve 'Yüksek' kategorideki oranları %25,9 olarak gerçekleşmiştir. Bu durum ise asansörü olup *Tepebaşı* ilçesinde bulunan konutların birim fiyatının düşük olduğu şeklinde yorumlanabilir.

### 3.2. Yöntemlerin Modelleme Başarılarının Karşılaştırılması

Her iki yöntemle elde edilen modellerin sınıflandırma başarılarını ölçmek için veri seti, eğitim (%50), doğrulama (%25) ve test (%25) olmak üzere 3' e ayrılmıştır. Eğitim ve doğrulama veri setleri, model oluşturmada kullanılmıştır. Aşırı uyum (overfitting) sorununu önlemek için 5-katlı çapraz doğrulama kullanılmıştır. Buna göre, doğrulama veri setinde en iyi performans gösteren yöntem seçilerek, daha önce kullanılmamış ve test olarak ayrılan veri setinde performans karşılaştırması yeniden yapılmıştır. Bu durumda daha başarılı model, sınıflandırma ağaçları ile elde edilmiştir.

Tablo 3.7.' de eğitim, doğrulama ve test verisi üzerinden sıralı lojistik regresyon ve sınıflandırma ağaçları yöntemleriyle elde edilen performans sonuçları yer almaktadır.

**Tablo 3.7**. *Eğitim, Doğrulama ve Test Verisi için Elde Edilen Performans Ölçüleri*

| Model | Eğitim Doğruluk (%) | Eğitim Yanlış Sınıflandırma (%) | Doğrulama Doğruluk (%) | Doğrulama Yanlış Sınıflandırma (%) | Test Doğruluk (%) | Test Yanlış Sınıflandırma (%) |
|---|---|---|---|---|---|---|
| SLR | 71,8 | 28,2 | 62,7 | 37,3 | 64,8 | 35,2 |
| CART | 73,5 | 26,5 | 72,9 | 27,1 | 74,1 | 25,9 |

Kurulan modelde elde edilen sonuçlar için doğruluk oranı yüksek çıkan algoritma tercih edilir.

Eğitim veri setinde karşılaştırılan sıralı lojistik regresyon ile CART oranlarına Tablo3.7'den bakıldığında, CART algoritması % 73,5 doğruluk oranında, sıralı lojistik regresyona göre daha başarılı bir model olarak görülmektedir.

Doğrulama veri seti iki yöntem arasındaki seçim için kullanılır. CART algoritması % 72,9 doğruluk oranında, sıralı lojistik regresyonunun % 62,7 doğruluk oranına göre daha başarılı olduğu görülmektedir. Test verisi (hiç kullanılmayan veri üzerinden elde edilen sonuçlar) içinde sonuçlar benzerlik göstererek doğrulama verisindeki sonuçları desteklemektedir.

Analiz sonuçlarından elde edilen üç kategori için karışıklık matrisi Tablo 3.8.'de hem sıralı lojistik regresyon analizi hem de sınıflandırma ağacı analizi için gösterilmiştir.

**Tablo 3.8**. *Sıralı Lojistik Regresyon ve Sınıflandırma Ağacı Analizleri için Üç Kategorili Karışıklık Matrisi*

| | | **Sıralı Lojistik Regresyon** | | | | **Sınıflandırma Ağacı** | | |
|---|---|---|---|---|---|---|---|---|
| | | Reference (Gerçek Sınıf) | | | | Reference (Gerçek Sınıf) | | |
| Prediction (Tahmin Sınıfı) | Kategoriler | En Düşük | Orta | En Yüksek | | En Düşük | Orta | En Yüksek |
| | En Düşük | 44 | 16 | 4 | En Düşük | 43 | 10 | 5 |
| | Orta | 0 | 0 | 0 | Orta | 9 | 12 | 0 |
| | En Yüksek | 9 | 9 | 26 | En Yüksek | 1 | 3 | 25 |

Sıralı lojistik regresyon analizi ve sınıflandırma ağacı analizi için Tablo 3.9.' da verilen ölçüler ile bağımlı değişken konut birim fiyatlarının, veri için gerçekte hangi

kategoriye sahip olduğunun ve tahmin sonucunda hangi kategoriye denk geldiğinin oranları verilmiştir.

**Tablo 3.9**. *Sıralı Lojistik Regresyon Analizi ve Sınıflandırma Ağacı Analizi Sonucu Elde Edilen Yardımcı Ölçüler*

| | **Sıralı Lojistik Regresyon Analizi** | | | | **Sınıflandırma Ağacı Analizi** | | |
|---|---|---|---|---|---|---|---|
| Sınıf / Yardımcı Ölçüler | En Düşük | Orta | En Yüksek | | En Düşük | Orta | En Yüksek |
| Duyarlılık (Sensitivity) | 0,8302 | 0,000 | 0,8667 | | 0,8113 | 0,4800 | 0,8333 |
| Belirleyicilik (Specificity) | 0,6364 | 1,000 | 0,7692 | | 0,7273 | 0,8916 | 0,9487 |
| PPV-Positive Predicted Value | 0,6875 | NAN | 0,5909 | | 0,7414 | 0,5714 | 0,8621 |
| NPV-Negative Predicted Value | 0,7955 | 0,7685 | 0,9375 | | 0,800 | 0,8506 | 0,9367 |
| Yaygınlık (Prevalence) | 0,4907 | 0,2315 | 0,2778 | | 0,4907 | 0,2315 | 0,2778 |
| | | | | | | | |
| Kappa | 0,41 | | | | 0,58 | | |
| Doğruluk (Accuracy) | 0,64 | | | | 0,74 | | |
| 95%Güven aralığı(CI) | (0,5504; 0,7376) | | | | (0,6475; 0,8203) | | |

Sıralı lojistik regresyon analizinde yapılan tahminler için gerçekte ‘En Düşük’ kategoride olan gözlemlerin aynı kategoride tahmin edilmesi oranı 0,8302 olup başarılı bir sonuç elde edilmiştir. Benzer şekilde gerçekte ‘En Yüksek’ kategoride olan gözlemlerin aynı kategoride tahmin edilmesi oranı 0,8667 olup başarılıdır. Ancak gerçekte ‘Orta’ kategoride olan gözlemlerin aynı kategoride tahmin edilme oranı 0,000 olarak elde edilmiştir. Sınıflandırma ağacı analizinde ise gerçekte ‘En Düşük’ ve ‘En Yüksek’ kategorilerde olan gözlemlerin aynı kategoride tahmin edilmesi oranları sırasıyla 0,8113 ve 0,8333 olup başarılı bir sonuç elde edilmiştir. Ancak sıralı lojistik regresyonda ‘Orta’ kategorinin duyarlılık oranı 0,00 iken sınıflandırma ağacında ‘Orta’ kategorinin

duyarlılık oranı 0,48 çıkmıştır. Buradan Sınıflandırma ağacı analizi duyarlılık açısından daha yüksek olduğu görülmektedir.

Belirleyicilik (Specificity) oranları negatif kararın doğru olma olasılığı anlamına gelir. Sıralı lojistik regresyon analizinde, gerçekte 'En Düşük' kategorisinde olmayan gözlemlerin diğer kategorilerde ('Orta ve En yüksek') tahmin edilmesi oranı 0,6364; gerçekte 'Orta' kategorisinde olmayan gözlemlerin diğer kategorilerde tahmin edilmesi oranı 1,000 ve gerçekte 'En yüksek' kategorisinde olmayan gözlemlerin diğer kategorilerde tahmin edilmesi oranı 0,7692 olarak elde edilmiştir. Buradan gerçekte negatif olan kararın negatif olarak seçilme olasılıklarının yüksek olduğu görülmektedir. Sınıflandırma ağaçları analizinde belirleyicilik oranları daha yüksek çıkmıştır.

Pozitif tahmin değerleri, sınıflandırma sonucunda pozitif sınıfta olanların gerçekte o sınıfta bulunma olasılığını verir. Sıralı lojistik regresyon analizinde 'Orta' kategorideki duyarlılık oranı sıfır olduğu için pozitif tahmin oranı tanımsız olmuştur.

Negatif tahmin değerleri, sınıflandırma sonucunda negatif sınıfta bulunanların ne kadarının gerçekte negatif sınıfta bulunması oranıdır. Tablo 3.9'daki değerler için negatif tahmin değerlerinin yüksek oranda çıktığı görülmektedir.

Gerçekte pozitif olan sınıfın toplamdaki sıklığını belirlemek için yaygınlık değerlerine bakılır. Yaygınlık değerleri karşılaştırılan iki analizde de aynıdır. 'En Düşük' kategorideki yaygınlık oranı 0,4907; 'Orta' kategorideki oranı 0,2315; 'En Yüksek' kategorideki oranı 0,2778' dir.

Kappa istatistiği, sıralı lojistik regresyon analizinde 0,41 ve sınıflandırma ağacı analizinde 0,58 olarak elde edilmiş ve bu durum kategorilerin arasında orta derecede uyumluluk olduğu şeklinde yorumlanır (Landis ve Koch, 1977).

Kategorilerin doğru sınıflandırma oranı olarak elde edilen doğruluk oranları ise sıralı lojistik regresyon analizinde 0,648; sınıflandırma ağacı analizinde 0,74 olarak elde edilmiştir.

# 4. SONUÇ

Bu çalışmada, Eskişehir ilindeki konut birim fiyatları üzerinde etkili olduğu düşünülen değişkenler bahçe durumu, asansör durumu, bulunduğu ilçe bilgisi ile uygun modeller kurulmuştur. Bu modeller ile sıralı lojistik regresyon ve sınıflandırma ağaçları yöntemleri kullanılarak konut birim fiyatı için sınıflandırma yapılmıştır. Modelin veri ile uyumlu olduğu uyum iyiliği testiyle gösterilmiştir. Paralellik varsayımı test edilerek, katsayıları yorumlamak için odds oranlarından yararlanılmıştır.

Sıralı lojistik regresyon algoritması için seçilen değişkenlerin tümü istatistiksel olarak anlamlı bulunurken, sınıflandırma ağacı algoritması için konutun asansör durumu ve bulunduğu ilçe bilgisi istatistiksel olarak anlamlı bulunmuştur.

Modellerin sınıflandırma başarılarını ölçmek için veri seti üçe ayrılmıştır. Söz konusu yöntemlerle yapılan analizlerde tüm verinin %50'si modeli oluşturmak için eğitim verisi, sınıflama kurallarının doğruluğunu test etmek amacıyla %25 doğrulama ve %25 test verisi olarak ayrılmıştır. Doğrulama veri setinde en iyi performans gösteren yöntem seçilerek, daha önce kullanılmamış ve test olarak ayrılan %25'lik veri setinde performans karşılaştırması yeniden yapılmıştır. Test veri setinde sıralı lojistik regresyon algoritması ile kurulan modelin sınıflandırma başarısı, sınıflandırma ağacı algoritması ile kurulan modelin sınıflandırma başarısına göre daha düşük hesaplanmıştır. Buna göre daha başarılı model, sınıflandırma ağaçları ile elde edilmiştir.

Mevcut veriler ile bu yöntemlerin başarılarına ilişkin elde edilen sonuçlar, aynı özelliğe sahip veriler için yapılacak farklı uygulama alanlarında gelecekteki çalışmalara yol gösterici olacaktır.

# KAYNAKÇA


Agresti, A. (2019). *An Introduction to Categorical Data Analysis*, Thrid Edition, Wiley Series in Probability and Statistics, printed in USA.

Akdi, Y. (2014). *Matematiksel İstatistiğe Giriş* (4. Baskı) Ankara: Gazi Kitabevi.

Akın, H.B. ve Şentürk E. (2012). *Bireylerin Mutluluk Düzeylerinin Ordinal Lojistik Regresyon Analizi ile İncelenmesi*, 10 (37). https://dergipark.org.tr/download/article-file/165694 (Erişim Tarihi:14.03.2019)

Altman, D. G. ve Bland, J.M. (1994). *Statistics Notes: Diagnostic tests 1: sensitivity and specificity,* British Medical Journal. Vol. **308**.

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

Antonio, K. ve Beirlant, J. (2007). *Actuarial Statistics with Generalized Linear Mixed Models*. Insurance: Mathematics and Economics, 40(1), 58-76.

Atılgan, E. (2011). *Karayollarında Meydana Gelen Trafik Kazalarının Karar Ağaçları ve Birliktelik Analizi ile İncelenmesi*. Yüksek Lisans Tezi. Ankara : Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı.

Ayhan, S. (2006). *Sıralı Lojistik Regresyon Analiziyle Türkiye'deki Hemşirelerin İş Bırakma Niyetini Etkileyen Faktörlerin Belirlenmesi.* Yüksek Lisans Tezi. Eskişehir: Osmangazi Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı.

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

Brant, R. (1990). *Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression.* Biometrics, (Vol. **46**) 1171-1178.

Breiman, L., Friedman, J.H., Olshen, R. A. ve Stone, C. J. (1984). *Classification and Regression Trees*. Florida, USA, CRC Press.

Ceyhan, G. (2014). *Üniversite Öğrencilerinin Yansıtıcı Düşünme Düzeyleri ve Araştırmaya Yönelik Kaygılarının Çeşitli Değişkenler Açısından CART Analizi ile İncelenmesi.* Yüksek Lisans Tezi. Van: Yüzüncü Yıl Üniversitesi, Eğitim Bilimleri Enstitüsü,Eğitim Bilimleri Anabilim Dalı.

Fu, C. Y. (2004). *Combining Loglinear Model with Classification and Regression Tree (CART): An Application to Birth Data*. Computational Statistics & Data Analysis, 45(4), 865-874.

Fullerton, A. S. ve Xu, J. (2012). The Proportional Odds with Partial Proportionality Constraints Model for Ordinal Response Variables. Social Science Research, 41(1), 182-198.

Garrido, J. ve Zhou, J. (2009). *Full Credibility with Generalized Linear and Mixed Models*. ASTIN Bulletin: The Journal of the IAA, 39(1), 61-80.

Güriş, S. ve Astar, M. (2015). *Bilimsel Araştırmalarda SPSS ile İstatistik* (2. Baskı) İstanbul: Der Kitabevi.

Hosmer Jr, D. W., Lemeshow, S. ve Sturdivant, R. X. (2013). *Applied Logistic Regression* (Vol. **398**). John Wiley ve Sons. New York.

Hosmer, D. W. ve Lemeshow, S. (2000). *Applied Logistic Regression,* John Wiley, New York.

Karadağ, Ö.G. (2014). *Genelleştirilmiş Doğrusal Modeller için Sınırlı Dalgalanmalı Kredibilite Yaklaşımı.* Yüksek Lisans Tezi. Ankara : Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, Aktüerya Bilimleri Anabilim Dalı.

Kıran, Z. B. (2010). *Lojistik Regresyon ve CART Analizi Teknikleriyle Sosyal Güvenlik Kurumu İlaç Provizyon Sistemi Verileri Üzerinde Bir Uygulama.* Yüksek Lisans Tezi. Gazi Üniversitesi, Fen Bilimleri Enstitüsü.

Kleinbaum, D. G. ve Klein M. (2010). *Logistic Regression: A Self-Learning Text*. Third Edition.

Korkmaz, D. (2018). *Sınıflandırma ve Regresyon Ağaçları ile Rastgele Orman Algoritması Kullanarak Botnet Tespiti: Van Yüzüncü Yıl Üniversitesi Örneği.* Yüksek Lisans Tezi. Van Yüzüncü Yıl Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı.

Landis JR., Koch GG. (1977). *The measurement of observer agreement for categorical data*. Biometrics. 33, 159-74.

Lawson, C. ve Montgomery D.C. (2006). *Logistic Regression Analysis of Customer Satisfaction Data*. Quality and Reliability Engineering İnternational, 22(8), 971-984.

Liao, T. F. (1994). *Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models, Quantitative Applications in the Social Sciences* (No. 101). Sage Publications.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables* (Vol. **7**). Advanced Quantitative Techniques in The Social Sciences. Sage Publications.

McCullagh, P. (1980). *Regression Models for Ordinal Data.* Journal of the Royal Statistical Society. Series B (Methodological), 42(2), 109-127.

Montgomery, D.C. , Peck, E.A. , and Vining G.G. (2013). *Doğrusal Regresyon Analizine Giriş*.( Çev:M.A.Erar).Ankara:Nobel.

Nelder, A.J. ve Wedderburn, R. W.M. (1972). *Generalized Linear Models*. *Journal of the Royal Statistical Society*: Series A (General), 135(3), 370-384.

O'Connell, A. A. (2006). *Logistic Regression Models for Ordinal Response Variables*, Quatitative Applications in the Social Sciences (Vol. **146)**. Sage Publications.

Özdamar, K. (2004). *Paket Programlar ile İstatistiksel Veri Analizi*, Eskişehir: Kaan Kitabevi.

Servet, O. (2018). *Dünya Ülkelerinde Mutluluk Eşitsizliği-Mutluluğu Belirleyen Faktörlerin Analizi.* Doktora Tezi. Gaziantep: Sosyal Bilimler Enstitüsü, İktisat Ana Bilim Dalı.

Şensoy, E.Z. (2009). *Nonlineer Lojistik Regresyon ve Uygulaması.* Yüksek Lisans Tezi. İstanbul: Marmara Üniversitesi, Fen Bilimleri Enstitüsü, Matematik Anabilim Dalı.

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)

Temel, G.O. (2004). *Sınıflama ve Regresyon Ağaçları*. Yüksek Lisans Tezi. Mersin: Mersin Üniversitesi,Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalı.

Tıpiş, Ö. (2012). *Özel ve Kamu Bankalarında Psikolojik Şiddet Boyutlarının İstatistiksel Tekniklerle Analizi.* Yüksek Lisans Tezi. Edirne: Trakya Üniversitesi, Sosyal Bilimler Enstitüsü, Ekonometri Anabilim Dalı.

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

Yarbaşı, İ.Y. (2015). *Genelleştirilmiş Lineer Modeller: Akademik Danışmanlık Hizmetlerinden Memnuniyet Üzerine bir Uyguluma.* Yüksek Lisans Tezi. Erzurum: Atatürk Üniversitesi, Sosyal Bilimler Enstitüsü, Ekonometri Anabilim Dalı.

# ÖZGEÇMİŞ


Adı Soyadı : Simay MİRGEN

Yabancı Dil : İngilizce

Doğum Yeri ve Yılı : Eskişehir / 1991

E-Posta : simaymirgen91@gmail.com


Eğitim ve Mesleki Geçmişi:

- 2017, Memur, T. Vakıflar Bankası T.A.O.
- 2016-2019, Eskişehir Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, İstatistik Anabilim Dalı.
- 2011-2019, Anadolu Üniversitesi, İşletme Fakültesi, İşletme Bölümü.
- 2009-2014, Gazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü.
- 2005-2009, H. Ahmet Kanatlı Anadolu Lisesi.