T.C.
ONDOKUZ MAYIS ÜNİVERSİTESİ
FEN FAKÜLTESİ
İSTATİSTİK BÖLÜMÜ

Samsun Atakum İlçesindeki Konut Fiyatlarına İstatistiksel Yaklaşım ve Fiyat Tahmini

Bitirme Projesi

Yunus Emre BÜYÜKGÜLER

Kadir ERTÜRK

Danışman

Öğretim Görevlisi Umut YAMAK

Samsun

2025

ÖZET

Konutlar bireylerin temel barınma ihtiyacını gideren varlıklar olarak öne çıksalar da, sosyoekonomik çaplarını da etkileyen önemli yatırımlardır. Yalnızca fiziksel bir yapı değil, bireyin ailesiyle birlikte yaşamını sürdürdüğü ve toplumsal bağlar kurduğu bir mekândır. Bu bağlamda, benzer özellikte konutlar bulunsa dahi her taşınmaz kendi içinde eşsizdir. (Özdamar, 2004)

Samsun'un hızla gelişen ilçesi Atakum'da yer alan konut ilanları incelendi ve fiyatlara etki eden değişkenler analiz edilmiştir.

Çalışmanın amacı, bölgeye ait güncel verilerden yola çıkarak konut fiyatlarını etkileyen temel unsurları belirlemek ve bu bilgiler ışığında makine öğrenmesi algoritmalarıyla tahmin modelleri geliştirmektir.

Veri seti üzerinde tanımlayıcı istatistikler, korelasyon analizleri ve aykırı değer tespitleri yapılmış; ardından logaritmik dönüşüm gibi ön işleme teknikleriyle veriler makine öğrenmesine hazır hâle getirilmiştir. Tahmin modellerinde kullanılan Rastgele Orman ve Yapay Sinir Ağı algoritmalarının, basit doğrusal regresyona göre anlamlı düzeyde daha yüksek doğruluk sergilediği gözlemlenmiştir.

Bu yönüyle proje; hem yerel düzeyde emlak yatırımlarına yön verebilecek hem de akademik literatürde bölgesel veri temelli analizlere katkı sunabilecek nitelikte bir içeriğe sahiptir.

Anahtar Sözcükler: **Konut Fiyat Tahmini, Makine Öğrenmesi, İstatistiksel Analiz, ANOVA, Hipotez, Regresyon, Random Forest, Boosting, Yapay Sinir Ağı, Samsun, Atakum**

ÖNSÖZ VE TEŞEKKÜR

Bu çalışma, Samsun Ondokuz Mayıs Üniversitesi İstatistik Bölümü'nde gerçekleştirilmiştir. Samsun/Atakum bölgesinde konut fiyatlarının analiz edilip, tahmin edilmesi amacıyla yapılan bir araştırmadır.

Bitirme projemizde konu seçiminden son aşamaya kadar desteklerini ve bilgi birikimlerini esirgemeyen çok değerli tez danışman hocamız, Sayın Öğretim Görevlisi Umut YAMAK'a sonsuz teşekkürlerimizi sunarız.

Yunus Emre BÜYÜKGÜLER / Kadir ERTÜRK

İÇİNDEKİLER

SİMGELER

y: Bağımlı değişken
C: Sınıf sayısı
∈: Elemanıdır (Matematiksel küme içinde olma durumu)
α: Anlamlılık Düzeyi (Kabul edilen hata payı)
H0: Yokluk Hipotezi (Null hipotez)
H1: Alternatif Hipotez
ρ: Korelasyon katsayısı
O: Gözlenen frekanslar
E: Beklenen frekanslar
r: Satır sayısı (Ki-Kare Testi için)
n: Toplam veri sayısı veya örneklem büyüklüğü
df: Degree of Freedom (Serbestlik derecesi)
$\overline{X}$: Ortalama
X_i: Her bir veri noktası
Q1: Birinci Çeyrek
Q2: İkinci Çeyrek (Medyan)
Q3: Üçüncü Çeyrek
X¯: X değişkeninin ortalaması
Y¯: Y değişkeninin ortalaması
∑: Toplam sembolü
R2: Determinasyon Katsayısı (Açıklanan varyans)
Xi: X veri kümesini temsil eden i'inci değer
Yi: Y veri kümesini temsil eden i'inci değer
SS: Sum of Squares (Kareler Toplamı)
MS: Ortalama kareler (Mean Squares)
F: F istatistiği
W: Shapiro-Wilk Test istatistiği
$x_{(i)}$: Sıralanmış veri
D: Kolmogorov-Smirnov Test istatistiği
F(x): Teorik dağılım
A2: Anderson-Darling Test istatistiği
U: Mann-Whitney U Test istatistiği
R1, R2: Grup sıra toplamları
H: Kruskal-Wallis Test istatistiği
R¯: Tüm verilerin sıra ortalaması (Genel sıra ortalaması)
R¯j: j. grubun sıra ortalaması
nj: j. grubun örneklem büyüklüğü
Y: Bağımlı değişken matrisi (MANOVA)
X: Bağımsız değişken matrisi (MANOVA)
B: Regresyon katsayıları matrisi (MANOVA)
E: Hata matrisi (MANOVA) / Hata kareler matrisi (Wilks' Lambda)

Λ: Wilks’ Lambda
H: Regresyon kareler matrisi (Wilks’ Lambda)
t: t-testi istatistiği
n1, n2: Grup örneklem sayıları
sp: Birleşik standart sapma
q: Tukey'in öğrencileşmiş aralık istatistiğinin kritik değeri
MSW: Gruplar içi ortalama kare
ng: Grup büyüklüğü
α′: Yeni anlamlılık düzeyi
Y: Bağımlı değişken
X: Bağımsız değişken
β0: Sabit terim (intercept)
β1: Regresyon katsayısı
ϵ: Hata terimi (gerçek hata terimi)
Yi: Gerçek gözlem değeri
Yi^: Modelin tahmini değeri
ei: Artık (gözlemlenen hata)
Se2: Hata terimlerinin tahmini varyansı
t: t-değeri
α: Anlamlılık seviyesi
R2: R-Kare (Belirleyicilik Katsayısı)
Y¯: Gerçek değerlerin ortalaması
n: Veri kümesindeki örnek sayısı / Toplam gözlem sayısı
λ: Düzenlileştirme parametresi (ceza katsayısı)
p: Bağımsız değişken vektörünün boyutu
γm: Adım büyüklüğü (step size) ya da learning rate
J: Amaç fonksiyonu
wi: i. örneğin ağırlığı
ϵm: Ağırlıklı hata

KISALTMALAR

AdaBoost	Adaptive Boosting
EÇO	En Çok Olabilirlik
EM	Expectation Maximization
GBDT	Gradient Boosting Decision Tree
GDM	Genelleştirilmiş Doğrusal Model
GH	Gauss Hermite
HKO	Hata Kareler Ortalaması
k-NN	k-Nearest Neighbors
KDF	Kümülatif Dağılım Fonksiyonu
Lasso	Least Absolute Shrinkage and Selection Operator
Lvmi	Sol ventrikül kitle indeksi (left ventricular mass index)
MAE	Mean Absolute Error
MCMC	Markov Chain Monte Carlo
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
RF	Random Forest
ReLU	Rectified Linear Unit
RMSE	Root Mean Squared Error
TL	Türk Lirası
m^2	Metrekare
VIF	Variance Inflation Factor
YSA	Yapay Sinir Ağları

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

1. GİRİŞ

En temel ihtiyaçlarımızdan birisi olan konutlar güvenlik ihtiyacını karşılayan yapılardır ve aynı zamanda da hane halkı için barınma, sosyal, kültürel, yatırım malları ve tüketim malı olarak ortaya çıkarlar.

Ülkelerin konut alım ve satımına ilişkin uyguladıkları politikalar, toplumların ekonomik seviyesi, refah düzeyi, teknolojik gelişmeler ve kişilerin seyahatleri konut alım ve satımını etkileyen önemli faktörlerdir (Pişkin, 2022: 572).

Genellikle bir ev, apartman, başka bir bina olabilir. Alternatif olarak mobil ev, tekne evi, yurt veya başka bir taşınabilir barınaklar da bulunur. İnsan Hakları Evrensel Beyannamesi'nin 12. maddesindeki anayasa hukuku ilkesi, mahremiyet hakkı ve bireyin yaşama ve sığınma yeri olarak evinin dokunulmazlığıdır. (ESA/STAT/2004/6)

Ekonomik perspektiften bakıldığında konut bir yatırım olarak büyük önem taşımakta ve özellikle hızlı nüfus artışı ve kentleşme ile birlikte konut piyasasında önemli bir yer edinmiştir. Bu nedenden ötürü konut fiyatlarının doğru ve güvenilir bir şekilde tahmin edilmesi ve piyasa risklerinin azaltılması, kaynakların verimli kullanılması ve yatırım stratejilerinin oluşturulması açısından kritik öneme sahiptir.

Türkiye'de son yıllarda hızlı bir kentleşme ve nüfus artışı yaşanmaktadır. Özellikle büyükşehirlerde ve kıyı şeridi bölgelerinde konut talebi artmıştır. Ekonomik büyüme, genç nüfus ve göç hareketleri gibi faktörleri de ele aldığımızda, konut piyasasındaki hareketliliği tetikleyen unsurlardır denebilir. Ayrıca hane halkı büyüklüğü, bölgesel gelir dağılımı ve ekonomik koşullar da Türkiye'de konut alım gücünü ve tercihlerini şekillendirir.

Konut sektörü Türkiye ekonomisinin önemli bir bileşenidir ve inşaat sektörü genellikle ekonomik göstergelere yansır. Son yıllardaki faiz oranlarındaki değişimler, devlet destekli konut projeleri (TOKİ vs.) ve kredi imkânları gibi, konut satışlarına katkılı olduğu belirlenmiştir.

Piyasaya bakınca konut, ekonominin en önemli piyasalarından birisidir ve makroekonomik göstergeler ve bireysel refah düzeyi üzerinde tamamen belirleyicidir.

1.1. Türkiye'de Konut Piyasasının Genel Görünümü

1.1.1. Türkiye Konut Sektörünün Tarihsel Gelişimi

Konut sektörümüz, 1980'lerden veri şehirleşmenin hızlanmasıyla önemli bir devrim yaşamıştır, 2000'li yılların başında ise konut kredilerinin yaygınlaşması ve TOKİ vb. projelerle özel sektör yardımları ile arz talep artışı sağlanmıştır. Son dönemlerde yaşanan pandemi gibi olaylar neticesinde ise piyasada genel olaraktan dalgalanmalar meydana gelmiştir.

1.1.2. Konut Talebi ve Arz Dinamikleri

Türkiye'de artan nüfus ve kentleşme oranı, konut talebinin temel belirleyicidir ve genç nüfus ve göç hareketleri özellikle büyük şehirlerde konut ihtiyacını artırmaktadır. Arz tarafından bakınca da, inşaat sektöründeki maliyet artışları ve arazi temini zorlukları, finansman koşulları etkili olur.

1.1.3. Fiyatlandırma ve Finansal Araçlar

Konut fiyatları birçok parametreden etkilenir. Döviz, faiz ve enflasyon bu paramterlerin başında gelir. Faiz oranlarının da yüksek olmasından dolayı son yıllarda yatırım amaçlı konut talebi de artmıştır.

1.1.4. Devlet Politikaları ve Teşvikler

TOKİ projeleri, KDV indirimleri, faiz destekleri gibi devlet destekleri konut piyasasının gelişiminde önemli rol oynamıştır. Ancak zaman zaman uygulanan kısıtlamalar ve düzenlemeler de piyasada hareketlilik yaratmıştır.

1.2 Araştırma alanının konumu ve tarihi

Araştırma alanı, Kuzey Karadeniz'de yer alan Samsun ilinin Atakum ilçesidir. Samsun ile merkezine 7 kilometre olan ilçe, şehrin orta batısında yer almaktadır. Atakum'un eski adı **Matosyon**'dur; bu ad, bölgede 1900'lerin başında ünlü bir tütün tüccarı olan Matossian'dan gelmektedir. (nufusu.com)

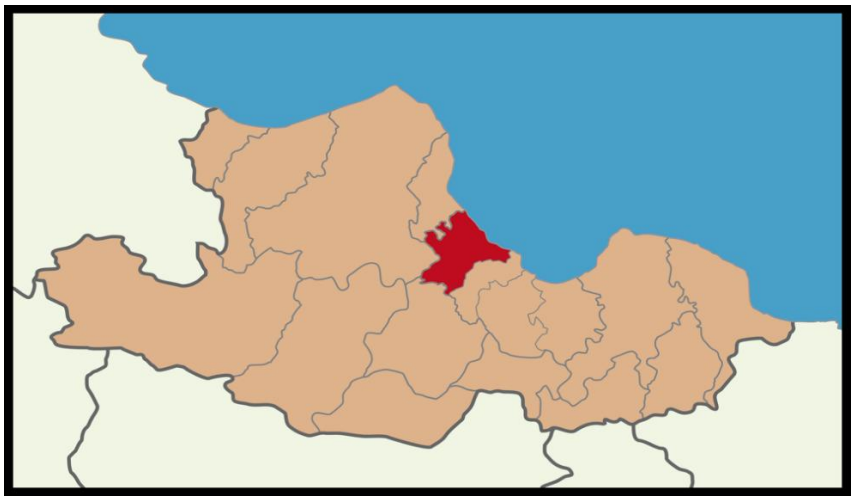
1994'e kadar belde statüsünde olan Atakum, 1994'te "Atakum" adıyla belediye kuruldu; 22 Mart 2008'de çıkarılan 5747 sayılı yasa ile Atakent, Kurupelit, Taflan, Altınkum, Çatalçam gibi çevre beldelerin birleştirilmesi sonucu ilçe haline gelmiştir. (nufusu.comoka.gov.tr.)

Kıyı şeridine yaslanan coğrafi yapısı, geniş sahil şeridi ve plajlarıyla dikkat çeken Atakum (Türkiye'nin "Karadeniz'deki en uzun sahil yolu" burada bulunur), kuruluşundan bu yana hızlı kentleşmiş ve nüfusunu hızla artırmıştır.

Şekil 1. Türkiye haritasında Samsun ili kırmızı renkle vurgulanmıştır. Samsun, Türkiye'nin kuzeyinde, Karadeniz kıyısında yer alan bir ildir (en.wikipedia.org.) Bu harita modern ve akademik bir görünümde tasarlanmış olup, gri tonlardaki zemin üzerinde kırmızı vurgu rengi öne çıkmaktadır. Harita üzerinde Türkiye'nin il sınırları net biçimde gösterilmiş ve Samsun ili belirgin hale getirilmiştir.

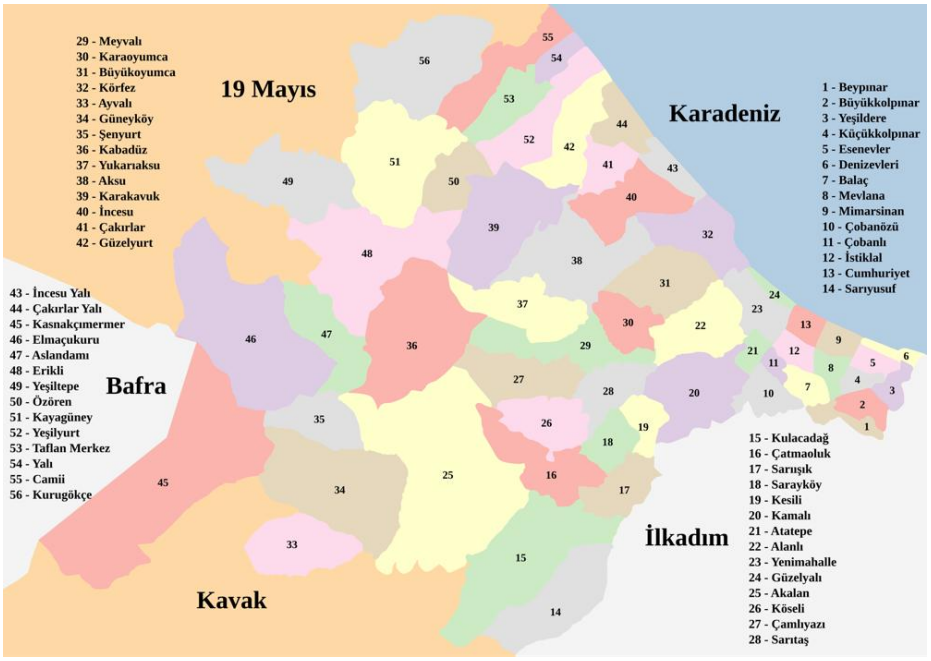
Şekil 1.1. Samsun'un Türkiye'deki konumu

Şekil 2. Samsun ili haritasında Atakum ilçesi kırmızı renkle vurgulanmıştır. Atakum, Samsun iline bağlı merkez bir ilçedir (en.wikipedia.org). Görselde Samsun ilindeki diğer ilçeler açık bej tonlarında gösterilirken, Atakum kırmızı ile ön plana çıkarılmıştır. İlçe sınırları ve başlıca yollar haritada net biçimde işaretlenmiş olup Atakum dışındaki alanlar gri-bej tonlarında arka planda kalmaktadır.



Şekil 1.2 Atakum ilçesi

Şekil 1.3. Atakum ilçesinin mahalle haritası. Bu detaylı haritada Atakum'un tüm mahalleleri gösterilmiş olup, örneğin Aksu, Kamalı gibi mahalleler listelenmiştir (parselharita.com). Her mahalle sınırı net çizgilerle belirtilmiş ve bölümler farklı renk tonlarıyla ayrılmaktadır.. Haritada zemin ve arka plan açık tonlarda, mahalle sınırları belirgin çizgilerle vurgulanmıştır.



Şekil 1.3. Atakum İlçesinin Mahalleleri

Atakum Samsun'un en gelişmiş ilçelerinden birisi olup demografik yapısı ve ulaşım imkanları açısından da bakınca potansiyeli oldukça yüksek bir ilçedir. İlçede 56 mahalle yer almaktadır. Bazı mahalleler nüfus, altyapı ve genel imkanlar açısından öne çıkmaktadır. Özellikle Atakent ve Yenimahalle Atakum'un en kalabalık mahalleleriyken Cumhuriyet Mahallesi konut fiyatları açısından önemli bir konumdadır.

Üniversiteye ve sahil şeridine yakınlığıyla bilinen Körfez ve Mimarsinan mahalleleri de öğrencilere ve çalışanlara hitap eden konut projeleriyle öne çıkar. Aile yaşamına uygun sosyal alanlarıyla Esenevler, gelişim sürecinde olan Mevlana ve Küçükkolpınar, yükselen yatırım değerleriyle dikkat çeker.

Bununla birlikte, Büyükoyumca Mahallesi; deniz manzaralı lüks villaları ve yüksek metrekare fiyatlarıyla Atakum'un premium konut bölgelerinden biri haline gelmiştir.

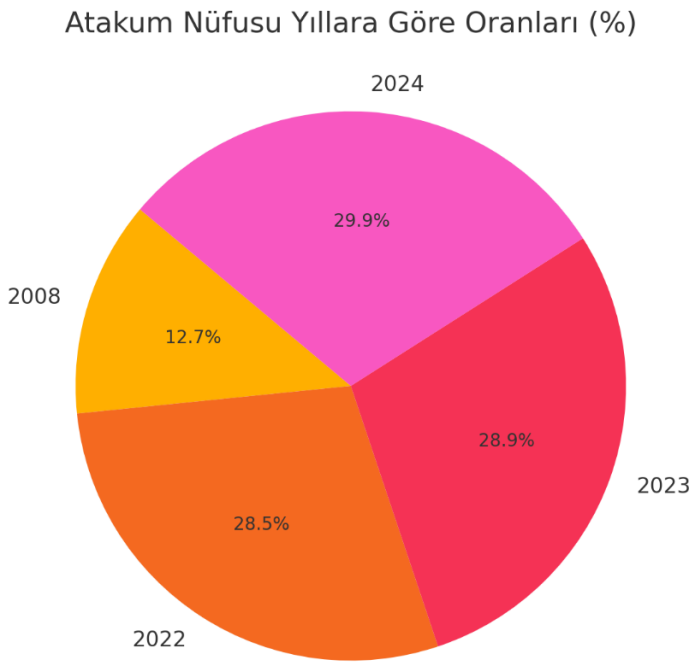
Bu mahallelerin her biri, Atakum'un farklı sosyoekonomik katmanlarına hitap eden çeşitliliği sunmakta, bölgenin konut arz-talep dengesini şekillendirmektedir. Nüfusun giderek artması, üniversite ve sahil şeridinin sağladığı cazibe ile birleşince, Atakum'un bu mahallelerinde konut fiyatları istikrarlı biçimde yükselmekte; bu da bölgeyi hem yaşamak hem de yatırım yapmak için cazip kılmaktadır.

1.3. Samsun Atakum Konut Piyasası

1.3.1. Araştırma Alanının Demografik Yapısı

Adrese Dayalı Nüfus Kayıt Sistemi (ADNKS) verilerine göre, 2008 yılında 107.953 olan Atakum ilçesinin nüfusu 2022'de 242.171'e, 2023'te 245.328'e ve 2024 itibarıyla 253.437'ye erişmiştir. İlçe nüfusunun 48%'i erkek, 52%'si ise kadınlardan oluşmaktadır.

Resim 1.4'de Atakum Nüfusunun geçmiş yıllara göre oranları pasta grafiği ile gösterilmiştir.



Şekil 1.4. Atakum Nüfusunun Yıllara Göre Oranı Pasta Grafiği

Çalışan nüfus oranı açısından Atakum, Samsun ilçeleri arasında öne çıkmaktadır. 2023 yılı verilerine göre Atakum'un "toplam yaş bağımlılık oranı" (0-14 ve 65+ yaş grubunun, 15-64 yaş çalışma çağına oranı) 37,06%'dır. Bu oranla Atakum, Samsun genelinde ilk sırada ve Türkiye genelinde ise 973 ilçe arasında 19. sırada yer almaktadır (dengegazetesi.com.tr).

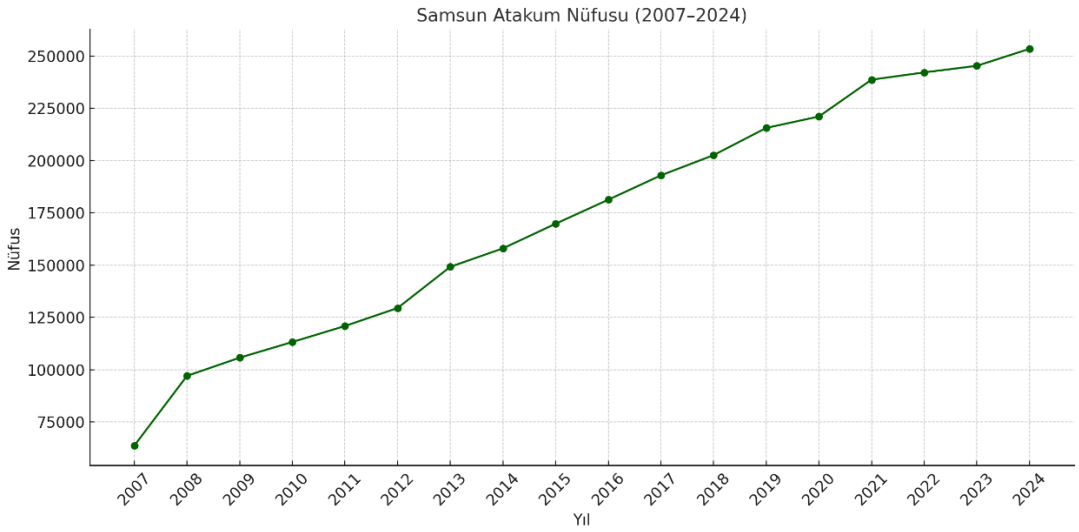
Demografik yapısına göre yorum yapılırsa, nüfus genç ve üretken kısımdan oluşmaktadır.

İlçede ortaöğretim mezunu oranı 28,98%, yükseköğretim mezunu oranı ise 24,14%'tür (oka.gov.tr). Ek olaraktan, Ondokuz Mayıs Üniversitesi'nin Kurupelit Yerleşkesi'nin Atakum sınırları içinde yer alması, ilçede hem öğrenci nüfusunu hem de eğitim altyapısını güçlendirmektedir.

Hanehalkı ve sosyoekonomik yapı açısından bakıldığında, Atakum hizmet sektörü – özellikle sahil turizmi ve perakende– ağırlıklı bir ekonomik yapıya sahiptir. Nüfusun büyük bir kısmı orta gelir grubunda yer alıyor. Çocuklu aile oranı ise görece düşüktür. Yüksek eğitim seviyesi, genç nüfus oranı ve yaşam kalitesi talebi, Atakum'u hem yaşam hem de yatırım açısından Samsun'un en cazip ilçelerinden biri haline getirmiştir.

1.3.2. Nüfus Artışı ve Göç Eğilimleri

2007 yılında yaklaşık 63 bin olan Samsun Atakum ilçesinin nüfusu 2024 yılı itibarıyla 253 bini aşmıştır. Özellikle de 2013'ten sonra hızlanan bu artış ilçenin sahil şeridinde konut yatırımlarının artması, şehirleşme faaliyetlerinin yoğunlaşması ve göç almasıyla açıklanabilmektedir lakin altında farklı parametreler de yatmaktadır. 2020 sonrası dönemde artış daha dengeli bir seyir izlese de, her yıl düzenli şekilde nüfus artmaya devam etmiştir. Bu durum, Atakum'un Samsun'un en hızlı gelişen ve tercih edilen ilçelerinden biri olduğunu göstermektedir. (https://biruni.tuik.gov.tr/medas/)



Şekil 1.5. Samsun/Atakum nüfusunun zamana göre artışı

Kentleşme büyükşehir yasa değişikliğiyle başlayan köylerin mahalleye dönüştürülmesiyle %100'e ulaşmıştır. (oka.gov.tr.). Son beş yılda Atakum'da: 2018-2022 döneminde Samsun'un diğer ilçelerine 122.259, komşu illerden Trabzon'dan 13.927, Ordu'dan 13.712, Amasya'dan 7.707, Tokat'tan 6.906 kişi Atakum'u seçmiştir. (dengegazetesi.com.tr.) Bu göç dalgaları, ilçenin bölgesel çekiciliğinin ve altyapısının giderek güçlendiğini göstermektedir.

1.3.3. Samsun/Atakum'da Konut Arz ve Talep Dengesi

Atakum ilçesinde konut talebi son 10 yılda artmaktadır. TUİK'e göre Samsun'da 2023'ün ilk 11 ayında 19.641 konut satılmış, Samsun ilçeleri arasında en yüksek satış Atakum'da gerçekleşmiştir. 2022 yılında Samsun genelinde 25.349 konut satılırken, Atakum'da 10.689 konut el değiştirmiştir. Toplam konut satışlarının neredeyse %42'sinin Atakum'da olmuştur.

2023'te Samsun genelinde satışlar düştüğünde bile, Atakum yine en yüksek paya sahip olmuştur (gazetegercek.com.tr). Veriler ilçedeki konut arzının talepten hızla beslendiğini ve hatta bazı dönemlerde talebi zorlayacak ölçüde olduğunu gösterir. Yeni konut projeleriyle arz artırılsa da, yoğun göç ve yerleşim talebi nedeniyle kısmi dengesizlikler de olacaktır.

1.3.4. Yatırım Potansiyeli ve Stratejik Konum

Atakum, konut yatırımları için yüksek potansiyele sahip bir bölgedir. Gelişmiş altyapısı, hizmet standartları ve çevresel imkanlarıyla "çağdaş şehirleşme" nitelikleri taşımaktadır. (dengegazetesi.com.tr.) Karadeniz sahil şeridinde yer alması, geniş kumsalları ve rekreasyon alanları (örneğin "Atakum Sahil Yolu" dünyanın en uzun sahil kenarı yollarından biridir (nufusu.com).) yaz turizmini besler. İlçede hızlı toplu ulaşım da mevcuttur; Samsun Tramvayı'nın doğu hattı Atakum'dan geçerek merkezle entegredir.

İlçeye yakın Mesudiye Mahallesi'nde gelişen Sanayi Bölgesi ile OMÜ'nün varlığı da istihdam ve ekonomik canlılık yaratmaktadır. Ayrıca Samsun-Çarşamba Havalimanı ve Karadeniz Sahil Yolu gibi bölgesel ulaşım bağlantıları Atakum'u öne çıkarmaktadır. Belediye tarafından planlanan yeni parklar, rekreasyon alanları ve kentsel dönüşüm projeleri de değeri artıracak faktörlerdir. Tüm bu etmenler, Atakum'da konut fiyatlarının orta-uzun vadede yükselme potansiyelini desteklemektedir.

1.3.5. Atakum'un Sosyo-Ekonomik Profili

Atakum, Samsun'un en kalabalık ilçelerinden biri olup, özellikle son 10 yılda önemli bir nüfus artışı yaşamıştır. Eğitim, sağlık ve ulaşım altyapısındaki gelişmeler, ilçeyi cazip bir yaşam alanı haline getirmiştir.

1.3.6. Konut Talebi ve Arz Durumu

Atakum'da konut talebi, özellikle sahil kesiminde ve şehir merkezine yakın bölgelerde yoğunlaşmaktadır. Yeni konut projeleri ve kentsel dönüşüm çalışmaları arzı artırırken, genç nüfusun ve orta-üst gelir grubunun talepleri farklı konut tiplerini beraberinde getirmektedir.

1.3.7. Fiyat Dinamikleri

Konut fiyatları Atakum'da Samsun genel ortalamasının üzerinde olup sahil şeridindeki projelerde yüksek prim potansiyeli vardır. Fiyat artışları açısından yatırım talebinin artacağı kestirilebilir.

1.3.8. Kentsel Gelişim ve Altyapı Yatırımları

Son yıllarda yapılan ulaşım projeleri, park ve sosyal alanların artırılması Atakum'un konut piyasasına olumlu yansımıştır. Kentsel dönüşüm projeleriyle eski yapıların yenilenmesi bölgedeki yaşam kalitesini artırmakta ve yeni konut stokları yaratmaktadır. Özellikle altyapı açısından yeni gelen belediyenin de çalışmaları mevcuttur.

1.4. Türkiye ve Atakum Konut Piyasası Karşılaştırması

Türkiye genelinde konut piyasası ekonomik dalgalanmalardan ve faiz oranlarından oldukça etkilenirken, Atakum gibi gelişmekte olan ilçelerde talep daha dinamik ve çeşitlidir. Yatırımcılar açısından Atakum, hem uygun fiyatları hem de gelişim potansiyeli nedeniyle öne çıkmaktadır. Türkiye genelindeki konut stoku fazlalığı bazı bölgelerde fiyat baskısı yaratırken, Atakum'da talebin arzı geride bırakıyor.

1.5. Araştırma Soruları ve Hipotezler

Çalışmada, aşağıdaki temel hipotezler ele alınıp cevaplanmıştır:

- Önde gelen kategorik konut değişkenleri arasında anlamlı farklılıklar var mıdır?
- Konut fiyatları geleneksel ve modern makine öğrenmesi modelleri ile tahmin edilebilmekte midir ve en etkili model veya modeller hangileridir?
- Veri ön işleme adımlarının (eksik veri giderme, aykırı değer temizliği) model performansına katkısı nedir?
- Nitel ve nicel konut verileri bir arada tâbi tutularak işleme alınabilir mi?

1.7. Literatüre ve Metodolojiye Kısa Bakış

Konutların fiyat değerlemesi için birçok farklı yöntem bulunur. Öznel ve kişisel görüşe dayanan fiyatlandırmadan öyle matematiksel ve istatistiksel modellere dayanan nesnel yaklaşımlarla değerlendirme yapmak daha önemlidir.

Literatürde, konut fiyatı tahmininde kullanılan başlıca yöntemler arasında regresyon analizi ve yapay sinir ağları gibi modeller yer alır. (Del Giudice vd., 2017; Isakson, 2001; Pagourtzi vd., 2003). Bu yöntemlerin konut değerleme çalışmalarında etkin şekilde kullanılmaktadır. (Nghiep & Al, 2001).

Toplu konut değerlemesi kapsamında hedonik yöntemler (Peterson & Flanagan, 2009) ve bulanık sistemler (Kempa vd., 2011) ile uzman sistemler kullanıldığı çalışmalar da mevcuttur (Rossini, 2000).

Herhangi bir veri kümesindeki kuralları ile ilişkileri ve kalıpları belirlemeye yönelik uygun bir yaklaşım sunan makine öğrenmesi algoritmalarıyla kredi analizi, görüntü tanıma, meteoroloji, tıp, dolandırıcılık tespiti, müşteri ilişkileri, biyoinformatik, tarım gibi birçok alanda tahmin analizi yapmaktadır (McQueen vd., 1995; Liakos vd., 2018).

Ek olaraktan da, konut piyasasının hali hazırdaki büyüme sürecinde, Destek Vektör Makineleri (DVM), Karar Ağaçları (KA), Yapay Sinir Ağları (YSA), Rastgele Orman (RO), k-En Yakın Komşu (k-EYK), Genetik Algoritmalar (GA), Gradient Boosting (GB), Naive Bayes (NB), Classification and Regression Tree (CART), GRNN, BPNN, RIPPER, AdaBooST, gibi makine öğrenmesi algoritmalarının da konut fiyatı tahmininde sıklıkla kullanılır hale geldiği söylenebilir (Monika vd., 2021).

Bu alt başlık altında yapılan açıklama, literatür ve metodolojiye ufaktan dokunmaktadır ve aşağıda da özet olması mahiyetinde yapılan çalışmaların özeti niteliğinde bir tablo sunulmaktadır.

Tablo 1.1. Literatüre Ön Bakış

Yöntem	Kullanım Alanı	Kaynaklar
Random Forest	Konut fiyatı tahmini	Del Giudice vd., 2017; Pagourtzi vd., 2003; Kempa vd., 2011
Yapay Sinir Ağları	Konut fiyatı tahmini, toplu gayrimenkul değerlemesi	Del Giudice vd., 2017; Nghiep & Al, 2001; Peterson & Flanagan, 2009; Kempa vd., 2011; Rossini, 2000
Boost Modelleri	Konut fiyatı tahmini	Del Giudice vd., 2017; Isakson, 2001
Regresyon Analizi	Konut fiyatı tahmini	Del Giudice vd., 2017; Isakson, 2001; Pagourtzi vd., 2003

Bu çalışmada da istatistiksel analizler ve makine öğrenmesi algoritmaları kullanılarak, Samsun Atakum bölgesindeki konut fiyatları edilip, bulgurayıla analizler yorumlanmıştır. 2025 yılına ait konut verileri kullanılarak, bölgesel farklılıkların belirlenmesi amacıyla mahalleler, oda sayısı ve satış durumu (sahibinden ya da emlakçıdan) gibi değişkenlere göre fiyat karşılaştırmaları yapılmıştır. Bu karşılaştırmalarda ANOVA ve Kruskal-Wallis gibi temel istatistiksel testlerden yararlanılmış; ayrıca fiyat ile diğer değişkenler arasındaki ilişkiler Pearson korelasyonu ve regresyon analiziyle detaylı biçimde değerlendirilmiştir.

Tablo 1.2. Başlıca İstatistiksel Analizler ve Amaçları

Analiz Yöntemi	**Amaç**
ANOVA	Gruplar arası fiyat farklarını test etmek.
Kruskal-Wallis Testi	Parametrik olmayan mahalle karşılaştırması.
Bağımsız Örneklem T-Testi	İki grup arasında fiyat farkı testi.
Mann-Whitney U Testi	Parametrik olmayan iki grup karşılaştırması.
Pearson Korelasyon	Fiyat ile sürekli değişkenler arasındaki ilişki
Regresyon Analizi	Fiyat tahmini ve değişken etkisi modelleme.

Elde edilen verilere Random Forest, XGBoost ve Lineer Regresyon gibi makine öğrenmesi modelleri kullanılmıştır ve karşılaştırılıp en iyi model belirlenmiştir.

Modellerin bölgedeki tahmini ne derece ölçtüğü analiz edilip makine öğrenmesi modellerinin performansı birebir karşılaştırılmıştır.

2. LİTERATÜR TARAMASI

Konut fiyatlarında istatistiksel analizler ve makine öğrenmesi yöntemlerinin uygulanabilirliği araştırılmıştır.

Barut, Z., ve Bilgin, T., yapay sinir ağları ve polinomsal regresyon yöntemleri veri seti üzerine temel alınan KNIME platformu ile kıyaslamışlardır. Elde edilen sonuçta yapay sinir ağlarının polinomsal regresyona kıyasla daha güçlü göstererek, daha düşük hata paylarıyla ev fiyatlarını öngörmeyi teşvik ederken, ikili değişken dönüşümleri ve model parametre optimizasyon performansı da gelişmiştir. Literatürde bu konuya ait çalışmaları, farklı makine öğrenmesi algoritmalarının -(örneğin rastgele orman, destek vektör makineleri,)- konut fiyat tahmini gözlem alanında etkinliğini vurgulamaktadır. Kıymetli, pratik bir karşılaştırma sunarak, konut değerleme bilmelerine bulunduğu makine öğrenmesinin kullanımı katkıda bulunmuştur (Barut ve Bilgin, 2023).

2015 yılında yapılan çalışmalarda konut fiyatlarının tahmininde makine öğrenmesi algoritmaları arasında destek vektör makineleri (DVM), karar ağaçları, rastgele orman, yapay sinir ağları ve k-en yakın komşu yöntemlerinin önemli başarılar elde ettiği belirtilmiştir (Chen vd., 2017; Monika vd., 2015)

Türkiye'de ise İstanbul, Ankara, Eskişehir gibi illerde yapay sinir ağları ve destek vektör makineleriyle gerçekleştirilen çalışmalar, bu yöntemlerin yerel piyasalarda yüksek doğruluk sağladığını göstermiştir (Yılmazel vd., 2018; Erkek vd., 2020; Altun, 2022).

Yazar Öztürk, (2023) 1+1 konutların fiyatlarını etkileyen faktörleri çalışamsında incelemiştir. İstanbul'daki 39 ilçeden 2022 yılına ait 1682 1+1 konutun verisi www.hepsiemlak.com web sayfasından elde etmiştir . Konut fiyatı, konut tipi, yaş, kat ve oda sayısı, tapu durumu vb. 26 değişkenden oluşan bir model geliştirmiştir. Hedonik fiyat modeli kullanılarak analizler yapmıştır. Analiz sonucunda konutun büyüklüğü, kat sayısı ve mülkiyet türünün konut fiyatını anlamlı olarak etkilediğini belirlemiştir. Konutun bulunduğu ilçe, konut tipi, site içinde olma durumu, krediye uygunluk, ısınma türü, satıcı türü, kira tutarı ve bina yaşının konut fiyatlarını anlamlı olarak etkilemediği tespit edilmiştir (Yazar Ö., 2023)

Texas'te konut geliştirme alanında Mathew C. Sosa tarafından gerçekleştirilen çalışma, Rio Grande Valley bölgesinde keşifsel veri analizi ve makine öğrenmesi tekniklerinin uygulanmasını içermektedir. Bu araştırmada, bölgedeki gayrimenkul gelişim dinamikleri incelenmiş ve veri odaklı yaklaşımlarla konut üretim süreçlerinin analizi yapılmıştır (Mathew C. Sosa).

Makine öğrenmesi teknikleri kullanılarak Lübnan'da konut fiyatlarını tahmin eden bir sistem geliştirilmiştir. Web kazıma, mekansal analiz ve demografik verilerle zenginleştirilmiş veri setleri kullanılmıştır. Random Forest, Gradient Boosting ve Stacking Regressor modelleri uygulanmış, en yüksek doğruluk Stacking Regressor ile elde edilmiştir (Mohamad Naji**)**

Georgia'da 2006 ve 2009 yıllarında satılan konutların verilerini toplayıp karşılaştırarak literatüre kendi çapında katkıda bulunmuştur. İki dönemdeki aynı 20 özellik (metrekare, oda sayısı, okul kalitesi, vb.) kullanılarak korelasyon analizi ve regresyon modellerini uygulamıştır. Ekonomik durgunluk dönemlerinde bazı evlerin değerlerini korurken bazılarının değer kaybettiğini bulgulamıştır (Cobb County).

Ordu ilindeki konutların fiyatlarını etkileyen faktörler araştırılmıştır. Altınordu ilçesindeki 558 satılık konutun verisini www.sahibinden.com web sayfasından temin etmiştir.. Konutun metrekaresi, oda sayısı, bina yaşı, bulunduğu kat, kat sayısı, banyo sayısı, site içinde olma ve konut fiyatları değişkenlerinden oluşan bir model oluşturulmuştur. Hedonik fiyat modeli en küçük kareler yöntemi ile analiz yapmıştır. Oda ve banyo sayısı, bulunduğu kat, kat sayısı, site içinde olmak konut fiyatlarını pozitif yönde, binanın yaşı ise konut fiyatını negatif yönde etkilediğini tespit etmiştir (Mağden, 2022)

Seferihisar ilçesindeki konutların fiyatlarına etki eden faktörleri araştırılmıştır. 2019 yılına ait 1063 konutun verisi www.sahibinden.com web sayfasından almışlar ve konutun fiyatı, konut tipi, oda sayısı, bina yaşı, bulunduğu kat, brüt ve net metrekare, ısınma türü, banyo sayısı, balkon sayısı, takas, kredi uygunluğu, eşya durumu ve site içinde olma durumu değişkenlerinden oluşan bir model oluşturmuşlardır. Hedonik fiyat modeli ile analizler yapmışlardır. Konut fiyatını etkileyen en önemli faktörün konut tipi olduğu tespit etmişlerdir. Banyo sayısı, bina yaşı, konut büyüklüğü, oda sayısı konut fiyatlarını pozitif yönde etkilediği bulgulanmıştır. (Ak Çetin ve Akpınar ,2021)

Hong ve Choi (2021) Güney Kore'nin Suwon kentinde pandemi süreci dayanıklılığını inceleyen bu çalışmada mahalle tipleri (yüksek/orta/düşük dirençli) arasında konutla ilişkili kentsel değişkenlerin (ör. daire alanları, konut tipleri) farklılıklarını test etmek için ANOVA uygulanmıştır. Normal dağılımı sağlayan değişkenler için parametrik ANOVA; sağlamayanlar için Kruskal–Wallis testi tercih edilmiştir.

Chugani (2021) ABD konut verileri çalışmasında ANOVA analizi yapmıştır. Grup varsayımlarını sağlamayan veriler için açıklanmasa da, benzer analizlerde normal dağılım yoksa Kruskal–Wallis testi önerilmektedir. (Çalışmada tek örnek gruplar arasında ANOVA kullanılmıştır).

Chao ve arkadaşları (2025) Çin'in Guangzhou kentinde sosyal konut sakinlerinin erişim ve hizmet memnuniyetini inceleyen çalışmada birden fazla bağımlı değişkeni eşzamanlı değerlendirmek için MANOVA kullanılmıştır. Araştırmada erişim ve hizmet faktörleri üzerinden elde edilen dört boyut için çoklu varyans analizi yapılarak hangi çevresel-demografik değişkenlerin bunlar üzerinde anlamlı etkisi olduğu test edilmiştir. MANOVA'dan sonra regresyon analizi ile her bir ana faktörün ayrı ayrı modellenmesi de gerçekleştirilmiştir.

Mustafa Kahveci (2020), Türkiye'de konut fiyatı belirleyicilerini araştıran doktora tezinde "etki" ve "kontrol" grup bölgelerinin reel konut fiyatı endekslerini karşılaştırmıştır. İki bağımsız örneklem t-testi ile her iki grup arasındaki ortalama fark test edilmiş ve **$t \approx 3.48$, $p = 0.001$** bulunarak gruplar arası konut fiyatı ortalamasının anlamlı şekilde farklı olduğu sonucuna ulaşılmıştır. Bu analizde H0 hipotezi (gruplar ortalamasının eşit olması) reddedilmiştir.

Güzel ve arkadaşları (2020) Ordu örneğinde kurulan çoklu regresyon modelinde, "Çizelge 3" sonuçlarına göre $R^2 \approx 0.708$ olarak hesaplanmıştır. ANOVA F-testi ($F \approx 145.67$, $p < 0.001$) ile model geçerliliği onaylanmıştır. Modele dahil edilen 16 bağımsız değişkenden özellikle daire büyüklüğü, kat sayısı, ısıtma sistemi, banyo sayısı, denize/anayola yakınlık, park/manzara gibi yapısal/çevresel özelliklerin konut fiyatları üzerinde pozitif yönlü etkisi istatistiksel olarak anlamlı bulunmuştur. Bu çalışma hedonik fiyatlama yaklaşımını temel alan bir çoklu regresyon modelidir.

Kangalli Uyar ve Keten (2020) konutların özellikleri ile fiyatları arasındaki ilişkiyi araştırmıştır. Denizli merkezindeki satılık 3666 konutun verisi web sayfalarından elde edilmiştir. Fiyat, enlem, boylam, metrekare, yaş, banyo sayısı, akıllı ev özelliği, ebeveyn banyosu, gömme dolap, güvenlik, açık ve kapalı havuz, balkon durumu, otopark ve hastane ve alışveriş merkezine yakınlık değişkenlerinden oluşan bir model geliştirilmiştir. Elde edilen veriler ile Hedonik fiyat modeli ile mekânsal kantil analizi yapılmıştır. Konut fiyatlarının farklı dilimlerdeki dağılımlarının komşu konut fiyatlarındaki artışlardan etkilendiği tespit edilmiştir.

Aliyev vd. (2019) Bakü'de satılık konutların fiyatlarını etkileyen önemli faktörleri araştırmıştır. Konutların şehir merkezine doğru daha yoğun oldukları, daha yüksek katlı oldukları ve daha az arazi üzerine inşa edildiği ifade edilmiştir. 497 apartman dairesi ve 443 konuttan oluşan veriler ile regresyon analizi yapılmıştır. Konutun konumu, büyüklüğü, onarım durumu, oda sayısı ve fatura tutarları konut fiyatlarını etkileyen önemli faktörler olduğu belirtilmiştir. Konutun konumu ve alanının fiyatını güçlü bir şekilde etkilediği, oda sayısının ise çok önemli olmadığı belirtilmiştir. Ayrıca dairenin zemini, sosyal altyapıya yakınlık, modern konforlu dizayn edilmesi veya doğal gazın bulunması fiyatları önemli ölçüde etkilemediği tespit edilmiştir. Metro istasyonlarına yakınlık fiyatları güçlü bir şekilde etkilediği bulgulanmıştır.

Ellibeş ve Görmüş (2018) Kocaeli ilindeki konutların fiyatlarını etkileyen faktörleri araştırmıştır. İl merkezindeki 9 farklı mahalleden elde edilen 180 satılık konutun verisi www.sahibinden.com web sayfasından elde edilmiştir. Metrekare fiyatı, binanın yaşı, kat sayısı, bulunduğu kat, manzara, ulaşım, site içinde olma, havuz, otopark, ebeveyn banyosu, dubleks ve arsa metre kare fiyatı değişkenleri çalışmanın modelini oluşturmaktadır. Mahalleler alt, orta ve üst gelir olmak üzere 3 farklı kategoriye ayrılmıştır. Her gelir grubu için ayrı bir model geliştirilmiştir. Hedonik fiyat modeli ile regresyon analizi yapılmıştır. Analiz sonucunda tüm gelir gruplarında ebeveyn banyosunun olması, dubleks özelliği, bulunduğu kat, arsa metrekare fiyatı, havuz, konutun metrekaresi ve site içinde olma özelliğinin konut metrekare fiyatlarını pozitif yönde etkilediği tespit edilmiştir.

Afşar vd. (2017) Eskişehir' de konut fiyatlarını etkileyen faktörleri araştırmıştır. Mahalle, metrekare, 1.katta olma durumu, merkezi ısıtma durumu, banyo ve oda sayısı, cephe, kat sayısı gibi değişkenlerden oluşan bir model oluşturulmuştur. Geliştirilen yazılım ile Odunpazarı ve Tepebaşı ilçelerindeki toplam 4.311 satılık konutun verileri internetten elde edilmiştir. Hedonik fiyat modeli ile çoklu regresyon analizi yapılmıştır. Konut büyüklüğü, oda ve banyo sayısı, asansör, otopark, ankastre mutfak, ebeveyn banyosunun olması konut fiyatlarını pozitif yönde etkilediği tespit edilmiştir. Konutun birinci katta olması ve merkezi ısıtma sistemi konut fiyatlarını negatif yönde etkilediği ifade edilmiştir.

Yılmaz ve arkadaşları (2017), konut fiyatlarını etkileyen temel unsurları belirlemek amacıyla Samsun ilinde kapsamlı bir analiz gerçekleştirmiştir. Çalışmada, www.hurriyetemlak.com internet sitesinden elde edilen veriler doğrultusunda, İlkadım, Atakum ve Canik ilçelerinde yer alan toplam 420 adet 3+1 satılık daire incelenmiştir. Elde edilen veriler ışığında, ilçe, bina yaşı, kat sayısı, dairenin bulunduğu kat, ısıtma tipi, site içinde olup olmama durumu, satıcı türü, asansör, çelik kapı, otopark, ulaşım imkânları, ebeveyn banyosu ve deniz manzarası gibi değişkenler modele dahil edilmiştir. Çoklu doğrusal regresyon yöntemi kullanılarak yapılan analizde, ilçe, site durumu, satıcı türü ve otopark değişkenlerinin konut fiyatı üzerinde istatistiksel olarak anlamlı bir etkisinin bulunmadığı görülmüştür. Buna karşılık, dairenin büyüklüğü, yaşı, bulunduğu kat, kat sayısı, merkezi ısıtma sistemi, toplu taşıma olanakları, asansör, ebeveyn banyosu ve deniz manzarası gibi unsurların konut fiyatlarını pozitif yönde etkilediği sonucuna ulaşılmıştır.

Bulut vd. (2015) Samsun ilindeki konutların fiyatlarını etkileyen faktörleri araştırmıştır. Samsun merkezdeki 3 ilçede satılık 3+1 391 konutun verileri www.sahibinden.com web sayfasından elde edilmiştir. Bulunduğu ilçe, ebeveyn banyosu, yaşı, kat sayısı, bulunduğu kat, ısıtma durumu, site mevcudiyeti, satıcı türü, asansör durumu, ısı yalıtımı, çelik kapı, otopark, ulaşım durumu ve deniz manzarası değişkenlerinden bir model oluşturulmuştur. Hedonik fiyat modeli ile regresyon analizi yapılmıştır. Analiz sonucunda ilçe, site durumu, satıcı türü, ısı yalıtımı, çelik kapı ve otopark durumu değişkenleri anlamsız çıkmıştır. Konutun büyüklüğü, yaşı, katı, kat sayısı, ısıtma türü, otobüs, tramvay, asansör, ebeveyn banyosunun varlığı ve deniz görme durumunun konut fiyatını pozitif yönde etkilediği tespit edilmiştir.

Osland (2010) mekânsal ekonometrik yöntemi kullanarak konut fiyatlarını etkileyen faktörleri araştırmıştır. Yöntem mekânsal otokorelasyon ve hetorejenliği dikkate aldığı için diğer yöntemlerde olduğu gibi hatalı sonuçlar verme ihtimalinin olmadığı belirtilmiştir. 1.691 konutun verisi kullanılarak analizler yapılmıştır. Modelde konut fiyatı, metrekaresi, büyülüğü, konut tipi, tuvalet sayısı, şehir merkezine uzaklığı, garaj ve konut yaşı değişkenlerinden oluşan bir model oluşturulmuştur. Şehir merkezinde olma veya şehir merkezine yakınlığın konut fiyatlarını etkilediği ifade edilmiştir.

Literatür incelendiğinde konut fiyatlarını etkileyen faktörlerin araştırıldığı çalışmalarda farklı ampirik yöntemlerin kullanıldığı görülmektedir.

3. METODOLOJİ VE KAYNAK

3.1. Veri toplama ve analiz yöntemleri

Bu çalışmada toplam **2.836 ilan verisi** kullanılmıştır. Veriler, Türkiye'nin önde gelen emlak platformlarından biri olan Sahibinden.com adresinden **manuel olarak** toplanmıştır (Sahibinden.com, 2025). Her bir ilan, **25 farklı değişken (özellik)** içermektedir. Bu değişkenler şunlardır: Fiyat (TL), Adres, İlan No, İlan Tarihi, Brüt m², Net m², Oda Sayısı, Bina Yaşı, Bulunduğu Kat, Kat Sayısı, Isıtma, Banyo Sayısı, Mutfak, Balkon, Asansör, Otopark, Eşyalı, Kullanım Durumu, Site İçerisinde, Site Adı, Aidat (TL), Krediye Uygunluk, Tapu Durumu, Kimden ve Takas Durumu.

2,836 veri istatistiksel ön işleme metotlarıyla işlenmiş olup kalan veriler işleme dahil edilmiştir.

Verinin tamamı 2025 yılının verileri olup hepsi Samsun/Atakum bölgesinden çekilmiştir.

Tablo 3.1. Veri Setinin İlk 10 Satırı

Fiyat (TL)	Adres	İlan Tarihi	…	Takas
1450000	Samsun / Atakum / Çobanlı Mh.	01 Mayıs 2025	…	Evet
1500000	Samsun / Atakum / Yenimahalle Mah.	01 Mayıs 2025	…	Hayır
1750000	Samsun / Atakum / Körfez Mh.	01 Mayıs 2025	…	Evet
1750000	Samsun / Atakum / Körfez Mh.	01 Mayıs 2025	…	Evet
1775000	Samsun / Atakum / Yenimahalle Mah.	01 Mayıs 2025	…	Hayır
1780000	Samsun / Atakum / Körfez Mh.	01 Mayıs 2025	…	Hayır
1790000	Samsun / Atakum / Körfez Mh.	01 Mayıs 2025	…	Hayır
1849000	Samsun / Atakum / İncesu Mh.	01 Mayıs 2025	…	Hayır
1900000	Samsun / Atakum / İncesu Yalı Mh.	01 Mayıs 2025	…	Evet
1900000	Samsun / Atakum / İncesu Yalı Mh.	01 Mayıs 2025	…	Evet

3.2. Makine Öğrenmesi Yöntemlerine Genel Bakış

3.2.1. Kullanılan Makine Öğrenmesi Algoritmaları

Makine öğrenmesi yöntemleri genel olarak üç ana kategori altında incelenmektedir:

Denetimli (supervised), denetimsiz (unsupervised) ve takviyeli (reinforcement) öğrenme . Bu yöntemlerden denetimli öğrenme, (x, y) biçiminde ifade edilen, bağımlı değişkenin (y) önceden bilindiği gözlem kümeleri üzerinde çalışır ve bu veri setleri model eğitimi için kullanılır (James, G. ve ark., 2021).

Denetimsiz öğrenme algoritmaları ise etiketlenmemiş veriler üzerinde çalışarak gözlemleri benzer özelliklerine göre gruplandırmayı amaçlar.

Takviyeli öğrenme, ödül-ceza mekanizması temelinde karar verme süreçlerini modellemek amacıyla kullanılır (Goodfellow, I. ve ark., 2016).

Bitirme projesinde danışmanlı makine öğrenmesi teknikleri kullanılmıştır.

3.2.2. Sınıflandırma

Regresyon, sürekli nitelikteki değişkenlerle çıktı tahmini yapmayı amaçlarken sınıflandırma, ait olduğu sınıfı bilinmeyen bir örneği en uygun sınıfa atamayı hedefleyen tekniktir (James ve ark., 2021).

Sınıflandırma problemi kapsamında $y \in \{1,2,...,C\}$ şeklinde gösterilen çıktı değişkeni, C sınıf sayısı ve C = 2 olması durumunda ikili (binary) sınıflandırma, C > 2 olması durumunda ise çoklu (multi-class) sınıflandırma söz konusu olmaktadır (Raschka ve Mirjalili, 2019).

Sınıflandırma problemleri için yaygın olarak tercih edilen denetimli öğrenme algoritmaları arasında yer almaktadır (Goodfellow ve ark., 2016).

- Random Forest
- K-En Yakın Komşu (k-Nearest Neighbors, k-NN)
- Yapay Sinir Ağları
- Regresyon Analizi

3.3. İstatistiksel Analizlere Genel Bakış

Konut verilerinin analizi için ileri düzey istatistiksel analizler kullanılmıştır. En başta model performans metrikleri, tek yönlü varyans analizi, paramtetrik olmayan yöntemler ve çoklu varyans analizi gibi yöntemlerle konut verilerine yaklaşılmaya çalışılmıştır (Field, 2018).

3.4. İstatistiğin Tarihçesi

İstatistik kelimesi, kökenini Latince "statisticum collegium" yani "devlet konseyi" ifadesinden alır ve ilk kez 1749'da Gottfried Achenwall tarafından Almanca'da "devlet bilimi" olarak kullanılmıştır (Hald, 1990). Ancak günümüzdeki anlamıyla veri toplama ve sınıflandırma işini 19. yüzyılın başında Sir John Sinclair sayesinde İngilizce'ye kazandırılmıştır.

Aslında istatistiğin temel amacı, devletler ve yönetimler için faydalı bilgiler sunmaktı. Bu yüzden uluslararası istatistik kuruluşları, nüfus sayımları gibi yöntemlerle verileri bir araya getirmeye başlamıştır.

Epidemiyoloji, biyoistatistik gibi sağlık alanları ya da ekonometri gibi ekonomi dalları, istatistiksel araçların gelişmesini gerektirdi. Refah devletlerinde ise toplumun nabzını tutmak, her şeyi daha iyi anlamak öne çıktı. Michel Foucault bu durumu "biyogüç" olarak tanımlamıştır. (Foucault, 1976).

Matematiksel kökeninde işin temeli, 1654'te Fermat ve Pascal'ın olasılık üzerine yazışmalarına dayanmaktadır (Stigler, 1986). Huygens bu alanı bilimsel bir çerçeveye oturtmuştur ve akabinde Jakob Bernoulli ve Abraham de Moivre gibi isimler konuyu matematiksel bir disipline taşımıştır. Yine de istatistik, matematiğin tam bir alt dalı olmaktan ziyade, uygulamalı bir alan olarak şekillenmiştir (Agresti, 2018).

Bugün ise istatistik, sosyal bilimlerde nüfus araştırmaları, anketler, davranış analizleri için elden düşmeyen bir araç haline gelmiştir. Sağlıkta epidemiyoloji, klinik deneyler, genetik çalışmalar istatistik olmadan ilerleyemez, hatta teşhis koyarken bile istatistiğe güvenilmektedir. Ekonomi ve işletmede piyasa analizleri, finans modelleri, kalite kontrol süreçleri istatistiğe dayanmaktadır. Mühendislikte ürün testleri, süreç optimizasyonu için de vazgeçilmezdir. Biyoloji ve çevre bilimlerinde ekosistemler, tür dağılımları incelenirken bu araçlar kullanılmaktadır. Üstelik yapay zeka ve makine öğrenmesi gibi modern teknolojilerin de temelinde istatistiksel prensipler yatmaktadır.

Kısacası istatistik, veri toplamak, sınıflandırmak, grafik veya tablolarla özetlemek, yorumlamak, sonuçların güvenilirliğini ölçmek, örneklerden genelleme yapmak, değişkenler arasındaki ilişkileri anlamak, gelecek için tahminlerde bulunmak, deneyler düzenlemek ve gözlem yapmak gibi bir sürü işi kapsayan bir bilim dalıdır. "Sayı bilimi" diye de anılır. Doğal bilimlerden sosyal bilimlere, iş dünyasından devlet politikalarına kadar geniş bir yelpaze sunar. (Moore ve McCabe, 2005).

3.5. Tanımlayıcı İstatistikler

Sayısal verileri özet olarak tanıtan, özetleyen, birimlerin yığıldıkları tipik parametreleri ifade eden ve veri kümelerinin dağılımı hakkında bilgi veren sayısal değerlerdir.

Tanımlayıcı istatistikleri kullanmanın amacı veri kümesini tanımak, veri dağılımını inceleyip nasıl yayıldığını anlamak, aykırı değerleri tespit etmek ve analizlerde bu aykırı değerlerin nasıl tepki vereceğini görmektir. (Field, A., 2018).

3.5.1. Merkezi Eğilim Ölçütleri

Bu ölçütler, verilerin ortalama değerini ve merkezi yerini belirler.

3.5.1.1. Ortalama (Mean)

Ortalama, bir veri setindeki tüm değerlerin toplamının, veri kümesindeki eleman sayısına bölünmesiyle hesaplanır. Ortalama, verinin merkezini belirlemenin en yaygın yoludur. Ancak aşırı büyük veya küçük değerler ortalamaya karşı aşırı duyarlıdır. Bundan mütevellit veri setinde aşırı uç değerlerin olmaması gereklidir. Ortama eşitlik (3.1) ile hesaplanır.

$$\bar{X} = \frac{\sum_{i=1}^{n} Xi}{n} \tag{3.1}$$

Burada, X_i her bir veri noktası ve n toplam veri sayısıdır. Ortalama, veri noktalarının toplamı, veri sayısına bölerek elde edilir.

3.5.1.2. Medyan (Median) ve Mod

Mod istatistiği, veri setinde en çok tekrar eden veriyi ifade eder.

Medyan, verideki ortanca değerdir. Veri kümesinin tam ortasındaki değerdir. Verinin sıralı bir şekilde düzenlenmesi sonucu ortaya çıkan bu değer, aşırı uç değerlerin etkisinden daha az etkilenir. Bu nedenle, medyan, ortalama ile karşılaştırıldığında daha güvenilir bir merkezi eğilim ölçütü olabilir, özellikle veri setinde büyük uç değerler bulunduğunda oldukça faydalı bir ölçüttür. Veri sayısı tek ise medyan (3.2)'deki gibi bulunur.

$$\mathbf{X}_{\frac{\mathbf{n+1}}{\mathbf{2}}} \tag{3.2}$$

Eğer veri sayısı çift ise, veri setinde ortada kalan iki elemanın ortalaması alınarak medyan değeri (3.3)'deki gösterimdeki gibi bulunur.

$$\frac{X_{\frac{n}{2}}+X_{\frac{n+1}{2}}}{2} \tag{3.3}$$

3.5.1.3. Varyans ve Standart Sapma

Varyans, verilerin ortalamadan ne kadar saptığını ölçen bir değerdir. Eşitlik (3.4)'de varyans formülü verilmiştir. Varyans ne kadar büyükse, verinin o kadar yayılmış olduğunu ve daha büyük bir dağılım gösterdiğini anlatır (Montgomery, D. C. ve ark., 2019).

Standart Sapma, varyansın kareköküdür ve daha anlamlı bir ölçü sağlar çünkü aynı birimde olup, verilerin ortalamadan ne kadar saptığını gösterir. Varyansın karekökünün alınmış halinin formülü (3.6)'da verilmiştir (Devore, J. L., 2020).

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N} \tag{3.4}$$

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}} \tag{3.5}$$

3.5.1.4. Minimum, Maksimum ve Ranj

Veri setinin en küçük (minimum) ve en büyük (maksimum) değerleri, verinin ne kadar geniş bir aralığa yayıldığını gösterir. Eşitlik (3.6)'da bu yayılım hesaplanabilir. Bu değerler veri setindeki aşırı uçları (outliers) tespit etmek için yardımcı istatistiklerdir. Ancak sadece minimum ve maksimum değerlere dayanmak, veri setinin yapısını anlamak için yeterli özellikler değillerdir (Triola, M. F., 2018).

Bu iki değer birlikte verinin **ranjını yani aralığını** tanımlamakta da kullanılır.

$$\boldsymbol{R} = max(\boldsymbol{X}) - min(\boldsymbol{X}) \tag{3.6}$$

3.5.1.5. Çeyrekler ve IQR (Interquartile Range)

Çeyrekler, verilerin sıralandıktan sonra dört eşit parçaya bölünmesini sağlayan tanımlayıcı istatistiklerdir Bu istatistikle verinin orta kısmı, yani veri kümesinin %50'si hakkında bilgi elde edilir. Çeyrekler:

Birinci Çeyrek (Q1): Verinin alt %25'ini kapsayan değer,

İkinci Çeyrek (Q2): Medyan.

Üçüncü Çeyrek (Q3): Verinin üst %25'ini kapsayan değer. (Field, A., 2018).

Çeyrekler arası mesafe (**IQR**), **Q3** ile **Q1** arasındaki farktır. Bu değer, verinin dağılımını ölçer ve aykırı değerleri elemeye yarar. (3.9) ile bu aralıklar hesaplanabilir.

$$IQR = Q3 - Q1 * IQR \; = \; Q3 \; - \; Q1 * IQR = Q3 - Q1 \qquad (3.7)$$

3.5.1.6. Aykırı Değerler

Veri setinde **aykırı değerler**, veri kümesinin büyük bir kısmından farklı olan veya aşırı uç olan değerlerdir.

Aykırı değerler, merkezi eğilim ölçütlerini yanıltabilir ve analizde önemli hatalara yol açabilir. Örneğin, ortalamayı yanıltabilir ve belirtme katsayısının farklı bir sonuç çıkmasına yol açabilir (Montgomery, D. C. ve ark., 2019).

Aykırı değerleri tespit etmek için yukarıdaki çeyrekler arası mesafe (IQR) ve standart sapma kullanılır.

Aykırı değerlerin yönetimi için izlenebilecek yöntemler 3.5.1.6.1. - 3.5.1.6.4. madde aralığında anlatılmıştır.

3.5.1.6.1. Çeyrekler Arası Mesafe ile Aykırı Değer Eleme

IQR, verinin alt %25'ini (Q1) ve üst %25'ini (Q3) belirler ve aralarındaki farkı hesaplayıp bu sınırların dışındaki değerler aykırı değerler olarak değerlendirir ve işleme alınmazlar.

3.5.1.6.2. Standart Sapma ile Aykırı Değer Eleme

Eğer veri seti normal dağılım gösteriyorsa, aykırı değerler genellikle ortalamadan belirli bir sayıda standart sapma uzaklıkta olan verilerdir. Genellikle, ortalamadan 3 veya daha fazla standart sapma uzaklıkta olan veriler aykırı değer olarak kabul edilir.

Bu yöntem her zaman sonuç vermemektedir. Çünkü veri dağılımı normal dağılım haricinde herhangi bir dağılım olabilir.

3.5.1.6.3. Z-Skoru ile Aykırı Değer Eleme

Z-skoru, her bir gözlemin ortalamadan ne kadar uzak olduğunu belirleyen bir ölçüttür. Normalizasyon işlemlerinde de Z-Skoru'ndan faydalanılmaktadır. Z-skoru, eşitlik (3.8)'den verinin standart sapmaya bölünmesiyle hesaplanır.

$$Z = \frac{X - \mu}{\sigma} \tag{3.8}$$

Veri $(3.8)'$deki gibi standartlaştırılır ve Z skoru mutlak değerce 3'den büyük olan değerler aykırı değer olarak nitelendirilir ve işleme alınmazlar.

3.5.1.6.4. Varyans Analizi ve Görselleştirme ile Aykırı Değer Eleme

Varyans analizi (ANOVA) veya görselleştirme araçları (örneğin, kutu grafikleri, dağılım grafikleri) kullanılarak veri kümesindeki aykırı değerler görsel olarak da tespit edilebilir. Kutu grafikleri, çeyrekler arasındaki mesafeyi ve aykırı değerleri açıkça gösterir.

3.5.1.6.5. Elenen Aykırı Değerlere Ne Olur?

Aykırı değerler, veri setinden çıkarılabilir, logaritmik, karekök ve z-skorlarına dönüştürülerek işleme tâbi tutulabilir. Ayrıca aykırı değerler, verinin yapısına göre mod veya medyan ile doldurulabilir.

3.6. İstatistiksel Analizler ve Hipotez Testleri

Hipotez Testleri, istatistiksel analizlerde değişkenler veya gruplar arasında anlamlı fark veya ilişki olup olmadığını belirlemek için kullanılır.

Hipotez testleriyle önceden kurulan iddialar test edilir ve anlamlı ve anlamsız farklılıklar tespit edilir.

3.6.1 Hipotez Testlerinin Genel Yapısı

Yokluk Hipotezi (H0): İncelenen popülasyonlar veya değişkenler arasında fark veya ilişki olmadığı varsayımıdır.

Alternatif Hipotez (H1): Fark veya ilişkinin var olduğunu savunan varsayımdır.

Test İstatistiği: Veriye bağlı olarak hesaplanan ve yokluk hipotezini test etmek için kullanılan sayısal değerdir.

Anlamlılık Düzeyi (α): Kabul edilen hata payıdır. Genellikle 0.05 değeri kabul edilir.

P-değeri: Elde edilen test istatistiğinin gözlemlenen veya daha uç değer alma olaslığı.

Eğer p≤α ise, H0 reddedilir; aksi takdirde (p> α) reddedilemez.

3.6.1.1. t-Testi (İki Grup Ortalamasının Karşılaştırılması)

İki grup ortalaması arasındaki ilişkiyi inceleyen hipotez testidir. Test istatistiği (bağımsız örneklem t-testi) eşitlik (3.9) ile hesaplanır. Birleşik standart sapma ise eşitlik (3.10) ile hesaplanmaktadır.

$$\mathbf{t} = \frac{\bar{\mathbf{X}}^{\mathbf{1}} - \bar{\mathbf{X}}^{\mathbf{2}}}{\mathbf{S}_{\mathrm{p}} \sqrt{\frac{\mathbf{1}}{\mathbf{n}^{\mathbf{1}}} + \frac{\mathbf{1}}{\mathbf{n}_{\mathbf{2}}}}} \tag{3.9}$$

$$\boldsymbol{S}_{\mathrm{p}} = \sqrt{\frac{(\boldsymbol{n}_1 - \mathbf{1})\boldsymbol{S}_1{}^2 + (\boldsymbol{n}_2 - \mathbf{1})\boldsymbol{S}_2{}^2}{\boldsymbol{n}_1 + \boldsymbol{n}_2 - \mathbf{2}}} \tag{3.10}$$

Burada;

$\bar{X}_i$: Grup ortalaması,
S_i^2 Grup varyansı,
n_i: Grup örneklem sayısıdır.
S_p: Birleşik standart sapmadır.

3.6.1.2. Korelasyon Katsayısının Anlamlılık Testi

Değişkenler arasındaki korelasyonun anlamlılığının güvenilirliğini ölçen hipotez testidir. Hipotezler aşağıdaki gibi kurulur:

- H0:ρ = 0 (Değişkenler arasında ilişki yoktur.)
- H1:ρ ≠ 0 (Değişkenler arasında ilişki yoktur.)

Test istatistiği eşitlik (3.11) aracılığıyla hesaplanmaktadır.

$$\boldsymbol{t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}} \tag{3.11}$$

Burada r örneklem korelasyon katsayısı, n gözlem sayısıdır.

3.6.1.3. ANOVA (Varyans Analizi)

Birden fazla grup ortalamasını karşılaştırıp test eden istatistiksel hipotez testidir. Detayları 3.6.3.'de verilmiştir.

3.6.1.4. Ki-Kare Testi

Ki-kare testi (Chi-square test), kategorik değişkenler arasındaki ilişkiyi test eder. Test, gözlenen frekanslar (O) ile beklenen frekanslar (E) arasındaki farkları ölçer ve bu farkların tesadüfi olup olmadığını değerlendirir (Agresti, A., 2019). Test istatistiği eşitlik (3.12)'deki formül ile hesaplanır.

$$\boldsymbol{\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)}{E_i}} \tag{3.12}$$

Burada;

- O_i: i'inci gözlenen frekans,
- E_i: i'inci beklenen frekans,
- n: Toplam gözlem sayısıdır.

Test istatistiğinin sonucu tablo değeri ile karşılaştırılarak uygun yorum ve sonuç çıkarılır.

Beklenen frekanslar (E), iki kategorik değişken bağımsız olduğu varsayımı altında hesaplanır. Örneğin, iki değişkenin çapraz tablosundaki bir hücre için beklenen frekans eşitlik (3.13) aracılığı ile hesaplanır.

$$\boldsymbol{E_{ij} = \frac{R_i \times C_J}{N}} \tag{3.13}$$

Burada;

- R_i = i'inci satırın toplam frekansı,
- C_J = j'inci sütunun toplam frekansı,
- N = Toplam gözlem sayısıdır.

Testin temel hipotezleri şunlardır:

- **Yokluk Hipotez, (H0):** Değişkenler arasında ilişki yoktur, yani bağımsızdırlar.
- **Alternatif Hipotez (H1):** Değişkenler arasında ilişki vardır.

Ki-kare istatistiği, serbestlik derecesi (degree of freedom, kısaca df) kullanılarak Ki-kare dağılımına göre karşılaştırılır. Serbestlik derecesi genellikle eşitlik (3.14) ile gibi hesaplanır.

$$\boldsymbol{df = (r - 1) \times (c - 1)} \tag{3.14}$$

Burada r satır satısı, n sütun sayısıdır.

Elde edilen χ^2 değeri ve serbestlik derecesi yardımıyla, p-değeri hesaplanır. P-değeri belirlenen anlamlılık seviyesinden (genellikle 0.05) küçükse, yokluk hipotezi reddedilir ve değişkenler arasında anlamlı ilişki olur.

3.6.1.4.1 Ki-Kare Testi Varsayımları

Hücrelerdeki beklenen frekansların 5'in üzerinde olması ve gözlemlerin birbirinden bağımsız olması gerekmektedir. Aksi halde testin sonucu güvenilir olmaz.

3.6.1.5. Olasılık Dağılımları ve Güven Aralıkları

Verilerin nerede dağılıp nerede yoğunlaşacağını gösteren dağılımlardır.

3.6.1.5.1. Normal Dağılım

En önemli dağılımlardan birisidir. Eğer veriler normal dağılım gösteriyorsa, çoğu veri ortada yoğunlaşır ve uç değerler (çok yüksek veya çok düşük) nadiren görülür.

Bu, genellikle fiyatların belirli bir aralıkta toplandığını gösterir. Herhangi bir veri setinde veri sayısı arttıkça veri seti normale daha da yaklaşır.

3.6.1.5.2. Güven Aralıkları

Modelin doğruluğunu ölçmek için kullanılır. Örnek olaraktan, fiyat tahmini yapılırken, tahminin ne kadar güvenilir olduğunu görmek için güven aralıkları hesaplanır. Eğer bir evin fiyatı 500.000 TL olarak tahmin edildiyse, güven aralığı bu tahminin %95 güvenle 480.000 TL ile 520.000 TL arasında olacağını gösterebilir.

3.6.1.6. Hipotez Testlerinin Uygulamadaki Önemi

Konut fiyatları ile bağımsız değişkenler arasındaki ilişkiyi anlamak, farklı gruplar (örneğin oda sayısı) arasında fiyat ortalamalarının anlamlı olup olmadığını belirlemek ve veri setinin varsayımlara uygunluğunu kontrol etmek için hipotez testleri uygulama önemli bir rol alır. En temelde **istatistiksel karar verme sürecinin temelini** oluşturur.

3.6.2. Korelasyon Analizi

Değişkenler arasındaki ilişki, bu ilişkinin yönü ve şiddeti ile ilgili bilgiler sağlayan istatiksel bir yöntemdir. İki ya da daha çok değişken arasındaki ilişkinin matematiksel bağıntısı "Regresyon Analizi" ile ilişkinin yönü ve derecesi ise "Korelasyon Analizi" ile incelenir (Alpar, R. 2020)

3.6.2.1. Korelasyon Katsayısı ve Açıklanan Varyans

Korelasyon katsayısı, **açıklanan varyans** ile ilgili bilgi verir. Açıklanan varyans, bir değişkenin ne kadarının diğer değişken tarafından açıklandığını gösterir ve korelasyon katsayısının karesiyle hesaplanır. Bu değer, **Determinasyon Katsayısı (R^2)** olarak ifade edilir ve eşitlik (3.15)'deki gibi hesaplanır.

$$\boldsymbol{R^2} = (\mathbf{r})^2 \tag{3.15}$$

- Eğer $\boldsymbol{R^2}$ = 1 ise, bu, deneysel verilerin kusursuz doğrusal bir eğri sağladığını gösterir.
- Eğer $\boldsymbol{R^2}$ = 0.80 ise, Y değişkenindeki toplam varyasyonun %80'i açıklanabilirken, %20'si açıklanamaz anlamına gelir.

3.6.2.2 Korelasyon Katsayısını Hesaplanışı

Korelasyon katsayısı, iki değişken arasındaki doğrusal ilişkinin ölçüsüdür. Pearson Korelasyon Katsayısı, eşitlik (3.16) ile hesaplanır.

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n(\sum y^2 - (\sum y)^2)}} \tag{3.16}$$

Burada:

- x ve y veri kümelerini temsil eder.
- n, veri sayısını ifade eder (Moore ve McCabe, 2005).

Tablo 3.2. Korelasyon Değişim Aralığı

Değişim Aralığı	Korelasyon
0 < r < 0.25	Çok zayıf
0.26 < r < 0.49	Zayıf
0.50 < r < 0.69	Orta şiddette
0.70 < r < 0.89	Yüksek
0.90 < r < 1	Çok Yüksek

3.6.2.3 Ortalamadan Ayrılış Kareler Toplamı (X ve Y için)

X Ortalamadan Ayrılış Kareler Toplamı eşitlik (3.17) ile hesaplanır.

$$\sum X^2 - \frac{(\sum X)^2}{N} \tag{3.17}$$

Y Ortalamadan Ayrılış Kareler Toplamı eşitlik (3.18) ile hesaplanır.

$$\sum Y^2 - \frac{(\sum Y)^2}{N} \tag{3.18}$$

XY Ortalamadan Ayrılış Kareler Toplamı eşitlik (3.19) ile hesaplanır.

$$\sum XY - n \cdot (\sum X)(\sum Y) \tag{3.19}$$

- $\sum XY$: $X\ ve\ Y$ değerlerinin çarpımlarının toplamıdır.
- $\sum X$: X değerlerinin toplamıdır.
- $\sum Y$: Y değerlerinin toplamıdır.
- n: Veri sayısını ifade eder.

Formüller X ve Y arasındaki ortak varyansın toplamını gösterir.

3.6.2.4 Korelasyon Katsayısının Anlamlılık Düzeyi

Korelasyon katsayısının anlamlı olup olmadığını test etmek için, **t testi** kullanılır. Hipotezler şu şekilde kurulur:

- H0:ρ=0 (Korelasyon katsayısı yoktur)
- H1:ρ≠0 (Korelasyon katsayısı vardır)

Test istatistiği eşitlik (3.20)'deki formül aracılığıyla hesaplanabilir.

$$\mathbf{t_H} = \frac{\mathbf{r}}{\sqrt{\frac{\mathbf{1-r^2}}{\mathbf{n-2}}}} \tag{3.20}$$

Elde edilen t_H değeri, anlamlılık düzeyine göre belirlenen kritik değer $\boldsymbol{t_{\alpha/2,(n-2)}}$ ile karşılaştırılır.

Eğer $|\ \boldsymbol{t_H}\ | > \mathbf{t_{\alpha/2,(n-2)}}$ ise, H0 reddedilir ve korelasyon sayısı anlamlıdır olarak kabul edilir.

3.6.3. Varyans Analizi (ANOVA)

Varyans analizi (ANOVA – Analysis of Variance), bir veya daha fazla kategorik bağımsız değişkenin, sürekli bağımlı değişken üzerindeki etkisini test etmek için kullanılan istatistiksel yöntemdir. Analizin temel amacı, farklı gruplar arasındaki ortalama farklarının istatistiksel olarak anlamlı olup olmadığını belirlemektir (Karagöz, 2019).

Özellikle konut fiyatları gibi değişkenlerde, örneğin farklı oda sayısına sahip (1+1, 2+1, 3+1 vb.) konutların fiyat ortalamaları arasında anlamlı bir fark olup olmadığını test etmek için yaygın olarak kullanılır.

Veri setinde k adet grup (kategori) ve toplam N gözlem olduğunu varsayalım.

- Y_{ij}: i. gruptaki j. gözlemin değeri, (örneğin, i. oda sayısına sahip konuttaki j. konut fiyatı)
- n_i: i. Grubun gözlem sayısı,
- $\overline{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$: i. grup ortalaması,
- $\overline{Y} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$: Tüm verinin ortalamasıdır.

ANOVA hipotezleri:

- H0: Tüm grup ortalamaları eşittir. ($\mu_1 = \mu_2 = \cdots = \mu_k$)
- H1: En az bir grup ortalaması diğerlerinden farklıdır.

3.6.3.1 Varyans Bileşenleri

Gruplar arası varyans eşitlik (3.21)'deki gibi hesaplanabilir.

$$SS_B = \sum_{i=1}^{k} n_i(\overline{Y_i} - \overline{Y})^2 \tag{3.21}$$

Gruplar içi varyans eşitlik (3.22)'deki gibi hesaplanır.

$$SS_w = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2 \tag{3.22}$$

Toplam varyans ise eşitlik (3.23) aracılığı ile hesaplanır.

$$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2 = SS_B + SS_w \tag{3.23}$$

Burada SS, Sum of Squares (Kareler Toplamı) idir.

3.6.3.2. ANOVA Test İstatistiği

Ortalama kareler hesaplanır. Gruplar arası ortalama karesi eşitlik (3.24) ile hesaplanır.

$$MS_B = \frac{SS_B}{k-1} \tag{3.24}$$

Gruplar içi ortalama kare, (3.25) ile hesaplanır.

$$MS_W = \frac{SS_W}{N-k} \tag{3.25}$$

F istatistiği, Gruplar arası ortalama karesi ile Gruplar içi ortalama karesinin bölümü eşitlik (3.26)'daki gibi hesaplanır.

$$\mathrm{F} = \frac{MS_B}{MS_W} \tag{3.26}$$

Elde edilen F istatistiği, $F_{k-1,N-k}$ dağılımına göre anlamlılık testi yapılır. Eğer F değeri kritik değerden büyükse veya p-değeri belirlenen anlamlılık seviyesinden küçükse, H0 reddedilir ki, yani gruplar arasında anlamlı fark vardır (Karagöz, 2019).

3.6.3.3 ANOVA'nın Konut Fiyatları Üzerine Uygulanması

Konutların "Oda Sayısı"na göre (1+1, 2+1, 3+1) fiyatlarının ortalamaları arasında fark olup olmadığı test edilecektir.

- Her oda grubu için fiyat ortalamaları hesaplanır.
- ANOVA ile fiyatların grup bazında anlamlı şekilde farklılaşıp farklılaşmadığı test edilir.
- Anlamlı fark bulunursa, hangi gruplar arasında fark olduğunu belirlemek için Post-Hoc testler (Tukey, Bonferroni vb.) uygulanır.

3.6.3.4. ANOVA'nın Avantajları ve Sınırlamaları

Birden fazla grup arasındaki farkların aynı anda test edilmesi, çok sayıda ikili t-test yapılmasına göre hata oranını azaltır. Veri normal dağılım varsayımı ve varyans homojenliği gerektirir. Bu varsayımlar sağlanmazsa parametrik olmayan alternatifler (Kruskal-Wallis testi gibi) tercih edilir. Fakat çok fazla ikili çıkacağından zamandan zarar ettiren bir analiz türüdür.

3.6.4. Normallik Testleri

Normallik analizi, çalışmada yapılacak analizlerde hangi test tipinin kullanılacağına karar vermeyi sağlar. Normallik analizi sonucuna göre veri normal dağılıyorsa çalışmaya parametrik olan testlerle, normal dağılmıyorsa parametrik olmayan testlerle devam etmek gerekir.

3.6.4.1. Shapiro-Wilk Testi

Örneklem büyüklüğü n ≤ 50 için uygundur. Test istatistiği eşitlik (3.27)'deki gibi hesaplanır (Shapiro ve Wilk, 1965).

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})}{\sum_{i=1}^{n}(x_i - \overline{x})} \tag{3.27}$$

Burada:

- $x_{(i)}$: Sıralanmış veridir (en küçükten en büyüğe).
- a_i: Ağırlık katsayılarıdır, standart normal dağılıma göre hesaplanır.
- $\overline{x}$: Örneklem ortalamasıdır.

3.6.4.2 Kolmogorov-Smirnov Testi

Örneklemin kümülatif dağılım fonksiyonu (KDF) $F_n(x)$ ile teorik dağılım $F(x)$ arasındaki en büyük fark hesaplanır ve eşitlik (3.28) ile hesaplanır (Massey, 1951).

$$D = sup_x \mid F_n(x) - F(x) \mid \tag{3.28}$$

Burada $\boldsymbol{sup}$ (supremum) ifadesi tüm x değerleri için en büyük mutlak farkı ifade eder.

3.6.4.3. Anderson-Darling Testi

Verilerin teorik dağılıma uyumunu ölçen, kuyruk bölgesine duyarlı normallik testidir. Test istatistiği Eşitlik (3.29)'daki formül ile hesaplanır (AD, 1954).

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[lnF(x_{(i)}) + \ln(1 - F(x_{(n+1-i)}))] \tag{3.29}$$

3.6.5 Gruplar Arası Karşılaştırmalar (Post-Hoc Testler)

ANOVA ile gruplar arasındaki farklar test edilse de hangi gruplar arası spesifik farklar var bunu bilemeyiz.

Bu nedenle, ANOVA sonucunda anlamlılık bulunursa, gruplar arasındaki çiftli karşılaştırmalar için post-hoc testler uygulanır.

3.6.5.1. Tukey'in HSD (Honestly Significant Difference) Testi

Grup ortalamaları arasındaki farkların anlamlı olup olmadığını değerlendirir ve aynı anda tüm çiftler için hata oranını kontrol eder. İki grup ortalaması $\overline{Y}_i \; ve \; \overline{Y}_j$ arasındaki fark (3.30) ile hesaplanır. Karşılaştırılacak test istatistiği de eşitlik (3.31) ile bulunur (Tukey, 1949).

$$D_{ij} = \left|\overline{Y}_i - \overline{Y}_j\right| \tag{3.30}$$

$$q_{\alpha,k,N-k} \times \sqrt{\frac{MS_W}{n_i}} \tag{3.31}$$

Burada:

- $q_{\alpha,k,N-k}$, Tukey'in öğrencileşmiş aralık (studentized range) istatistiğinin kritik değeri,
- MS_W, ANOVA'dan hesaplanan gruplar içi ortalama kare (within groups mean square),
- $n_i$grup büyüklüğüdür (eşit değilse hesaplama biraz karmaşıktır).
- Eğer; $D_{ij} > q_{\alpha,k,N-k} \times \sqrt{\frac{MS_W}{n_i}}$ ise, i ve j gruplarının arasında anlamlı fark vardır.

3.6.5.2. Bonferroni Düzeltmesi

Anlamlılık seviyesi α, yapılan test sayısına bölünür. Eşitlik (3.32) ile hesaplanır.

$$\alpha' = \frac{\alpha}{m} \tag{3.32}$$

Burada , m toplam karşılaştırma sayısıdır. Her karşılaştırmanın p-değeri, bu yeni α' ile karşılaştırılır. Çoklu testlerde hata tip I riskini azaltmak için anlamlılık seviyesini bölerek uygular (Field, A., 2013).

3.6.5.3. Konut Oda Sayısı Gruplarının Fiyat Karşılaştırması

Konut fiyatlarında 1+1, 2+1 ve 3+1 gibi oda sayısı gruplarının fiyat ortalamaları arasında ANOVA ile anlamlı fark bulunması durumunda, Tukey HSD testi ile hangi oda sayısı çiftlerinin fiyatlarının farklı olduğu tespit edilir.

3.6.6. Parametrik Olmayan Testler

Teknikler kendi içinde iki gruba ayrılır.

3.6.6.1. Tamamen Parametrik olmayan teknikler

İlgilenilen değişken ya da değişkenlerin ölçme düzeyleri sınıflama veya sıralama türündedir. Test edilecek hipotezler genellikle bir kitle parametresi kapsamazlar. İlgili değişken bakımından kitlenin dağılımı üzerinde herhangi bir varsayım aranmamaktadır. Uyum iyiliği testleri olarak bilinen ilişki testleri, bağımsızlık testleri gibi diğer testler de bu teknik başlığının altına girer (Özdamar, K. 2021).

3.6.6.2. Dağılıma bağlı olmayan teknikler

İlgilenilen değişken ya da değişkenlerden dağılım şartı aranmaz. Sıralama ölçme düzeyi şartı aranabilir. Bu teknikler normallik varsayımının sağlanmadığı durumlarda özellikle küçük örnekler için parametrik tekniklere alternatif olarak kullanılabilirler. Normallik sağlanıyorsa veya normallik varsayımı sağlanmazken büyük örneklem, yani yoğun veri konusu ise merkezi limit teorisi gereğince istatistiklerin örnekleme dağılımları normale yaklaşacağından parametrik teknikler tercih edilmelidir (Özdamar, K. 2021).

3.6.6.3 Parametrik Olmayan Testlerin Özellikleri

- Verilerin dağılımı önemli değildir.
- Küçük örneklemlerde başarılıdırlar.
- Sıralanmış verilerde yüksek doğruluk vermektedir.

3.6.6.4. Mann-Whitney U Testi

İki bağımsız grubu medyan parametreleri yönünden karşılaştırmada kullanılan bir parametrik olmayan test tekniğidir. Veriler sıralanır ve her grup için sıra toplamları $\mathbf{R_1}$ ve $\mathbf{R_2}$ hesaplanır. Veriler iki gruptan rastgele örneklemden çekilmelidir. Test istatistiği için U1 ve U2 (3.33) ve (3.34)'daki gibi hesaplanır ve eşitlik (3.35)'deki değer test istatistiğidir.

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \tag{3.33}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \tag{3.34}$$

$$U = min(U_1, U_2) \tag{3.35}$$

U istatistiği ile uygun tablo veya p-değeri hesaplanır ve kritik değer tablo değeri ile karşılaştırılarak hipotez test edilir. Akabinde de karar verilir. (Siegel, S., & Castellan, N. J., 1988).

3.6.6.5. Kruskal-Wallis Testi

Kruskal-Wallis Testi , Üç veya daha fazla grubun medyan farklarını test eder. Veriler sıralanır, her grup için sıra ortalamaları $\overline{R}_j$ hesaplanır ve test istatistiği (3.36)'daki gibi hesaplanır.

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(\overline{R}_j - \overline{R}\right)^2 \tag{3.36}$$

Burada,

- $\overline{R}$ = tüm verilerin sıra ortalaması (genel sıra ortalaması),
- $\overline{R}_j$ = j. grubun sıra ortalaması,
- n_j = j. grubun örneklem büyüklüğü,
- N=$\sum_{j=1}^{k} n_j$ = toplam örneklem sayısıdır.

Hesaplanan H değeri ki-kare dağılımı kullanılarak değerlendirilir (Kruskal, W. H., & Wallis, W. A., 1952).

.

3.6.7. Çok Değişkenli Varyans Analizi (MANOVA)

Birden fazla birbirine bağımlı olan değişkenin, grup ortalamalarını karşılaştırmaya yarayan çok değişkenli varyans analizidir. Model eşitlik (3.37)'deki gibi kurulur.

$$Y = XB + E \tag{3.37}$$

Burada;

- Y: $n \times p$ boyutunda bağımlı değişken matrisi; burada n örnek sayısını, p ise bağımlı değişken sayısıdır.
- X: $n \times q$ boyutunda bağımsız değişken matrisi,
- B: $q \times p$ boyutunda regresyon katsayıları matrisi,
- E: Hata matrisidir.

Çoklu varyans analizinde gruplar arası farkın tespiti için belli başlı testler kullanılır. Bu çalışmada sadece **Wilks' Lambda** istatistiği ele alınmıştır.

3.6.7.1. Wilks' Lambda (Λ)

$$\Lambda = \frac{|\mathrm{E}|}{|\mathrm{E+H}|} \tag{3.38}$$

Burada,

- E : hata kareler matrisi,
- H: Regresyon kareler matrisidir

Diğer yaygın kullanılan testler:

- Pillai's Trace
- Hotelling-Lawley Trace
- Roy's Largest Root

Bu çalışmada **. Wilks' Lambda (Λ)** testinin bulguları sunulmuştur

3.7. Makine Öğrenmesi

Bireylerin elde ettiği deneyimler sonucunda tasarlanan öğrenme, insan zekasının temelini oluşturmakla beraber makine öğrenmesinin prensiplerine temel olur.

Son 5 yılda logaritmik ölçekte kar topu etkisiyle çığ gibi artan veri miktarına karşın insan gücü ve geleneksel sistemlerle beraber işlenemeyecek kadar büyük veri ortaya çıkmıştır. Bu büyük veriden anlam çıkarmak için modern makine öğrenmesi modelleri türemiştir ve makine öğrenmesi, tahminler yapan bir disiplin olarak yerini almıştır.

3.7.1. Makine Öğrenmesinin Tarihi

Alan Turing'in savaş zamanı yaptığı çalışmalarla bu işin temeli atılmıştır (Turing, A. M. 1950). Ardından gelen yıllarda, 1967'de K-en yakın komşu gibi ilk algoritmalar geliştirildi. 80'ler ve 90'larda da bilim insanları, "Beyin nasıl çalışıyor?" sorusundan ilham alarak yapay sinir ağları üzerinde çalışmalar yapmıştır.

Bugün artık bilgisayarlara her şeyi tek tek öğretmektense, verilerle zamanla öğrenebiliyorlar. Eskiden elde olan verilerin işlenmesi gerekirdi. Günümüzde ise veriyle birlikte sistem kendi mantığını kurabilmektedir. (Mitchell, T. M. 1997).

Makine öğrenimi yalnızca akademide değil, günlük hayatta da pek çok yerde kullanılır. Hastanelerden bankalara, sesli asistanlardan sürücüsüz arabalara kadar birçok sistem artık bu öğrenme yöntemlerine dayanmaktadır. Üstelik büyük veriyle beraber bu işler çok daha güçlü hale gelmiştir. (Russell & Norvig, 2021).

3.7.2. Makine Öğrenmesi ve İstatistik arasındaki fark

Makine öğrenmesi mümkün olan en doğru tahminlere odaklanıp tahmini iyileştirirken istatistik ilişkileri çıkarıp analiz yapmak için tercih edilir. Bunun yanı sıra makine öğrenmesi ve istatistik de geçmiş veriler üzerinde çıkarım yapıp geleceğe yönelik tahminler yaparken, istatistik daha çok veriler üzerindeki ilişkileri açıklamaya odaklanır. Ek olaraktan Makine öğrenmesi, büyük veri havuzları için, istatistik ise küçük veri havuzları için tercih edilir (James, G, 2013).

Tablo 3.3. Makine Öğrenmesi ile İstatistik Karşılaştırması

Karşılaştırma Kriteri	**Makine Öğrenmesi**	**İstatistik**
Tanım	Kurallar veya programlama kodları olmadan, veriden öğrenen bir algoritma türüdür.	Belirli bir modelin neden seçildiğini ve tahminlerin nasıl ve neden yapıldığını anlamaya odaklanan bilim dalıdır.
Amaç	Geleceğe yönelik tahminler yapmak veya mevcut verileri sınıflandırmak.	Veri bileşenleri arasındaki ilişkileri analiz etmek ve ortaya çıkarmak.
Varsayımlar	Genellikle az veya hiç varsayım yapmaz.	Belirli varsayımlara dayanır (örneğin, veri dağılımı, bağımsızlık gibi).
Aykırı/Eksik Gözlemler	Bu tür durumlardan genellikle fazla etkilenmez.	Veri kalitesi çok önemlidir; aykırı ve eksik veriler sonuçları etkileyebilir.
Temel Dayanak	Bilgisayar bilimleri ve yapay zekanın bir alt alanıdır.	Matematiğin özellikle olasılık ve çıkarım alanlarına dayanır.
Veri Setleri	Büyük ve karmaşık veri setleri ile çalışır.	Orta ve küçük ölçekli veri setleri üzerinde daha etkili olabilir.
Model Yapısı	Daha esnek ve karmaşık modeller kullanır.	Genellikle daha basit ve açıklanabilir modeller tercih edilir.

3.7.3. Sınıflandırma

Mevcut alanlardan oluşan eğitim setine dayanarak yeni bir örneğin hangi kategoriye dahil olduğunu tahmin etme sürecine sınıflandırma denir. 2023 yılındaki başkanlık seçimlerinin kazananını öngörmek, bir kitlenin kötü huylu olup olmadığını belirlemek, Çeşitli çiçek türlerinin kategorize edilmesi örnek niteliğindedir.

Sınıflandırmanın sonucu, örneğin ait olduğu sınıfı gösteren ayrık bir değer olabilir. Tahminler belirli bir sınıfa ait olma ihtimalini belirten kesintisiz bir değer şeklinde olur (Géron, 2019).

3.7.4. Regresyon

Değişkenler arasındaki nedensel ilişkileri inceleyerek gelecekteki değerlerin tahmin edilmesini sağlayan tekniktir. Sınıflandırmadan farklı olarak, regresyon sonuçları sürekli değerler şeklindedir. Bir ürünün önümüzdeki çeyrekteki satış miktarını tahmin etmek, gelecek hafta için hava sıcaklıklarını kestirmek, belirli bir lastik modelinin kullanım süresini öngörmek, regresyon alanındaki çalışmalara örnektir (Montgomery, Peck, & Vining, 2012).

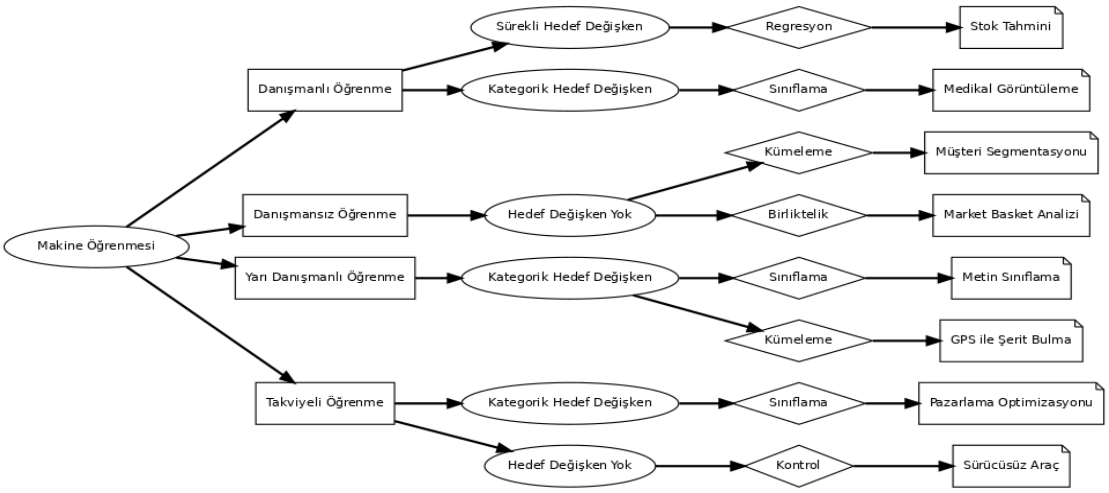
3.7.5. Kümeleme

Benzer özelliklere sahip veri noktalarını anlamlı gruplar halinde toplamaktır. Veriler, doğal yapısına göre gruplandırılarak organizasyonu kolaylaştırılır.

Örnekler niteliği için:

- Aynı tür filmleri tercih eden seyirci gruplarını belirlemek.
- Benzer arıza gösteren sabit disk modellerini sınıflandırmak.

3.7.6. Makine Öğrenmesi Algoritma Türleri



Şekil 3.1. Makine Öğrenmesi Algoritmaları

3.7.6.1. Danışmanlı Makine Öğrenimi

Danışmanlı öğrenme, eğitim sürecinde kullanılan verilerin her birinin bir "etiket" veya "sonuç" bilgisi taşıdığı bir yöntemdir. Danışmanlı öğrenme modelinde sınıflandırma işlemi gerçekleştirilir ve eğitim aşamasında oluşturulan model ile test aşamasında sistemin sınıflandırma yapması beklenir (Bilgin, M., & Şentürk, İ. F.).

En çok kullanılan gözetimli öğrenme algoritmaları arasında şunlar bulunur:

- En Yakın Komşu (k-Nearest Neighbors, KNN)
- Yapay Sinir Ağları (Artificial Neural Networks, ANN)
- Destek Vektör Makineleri (Support Vector Machines, SVM)
- Karar Ağaçları (Decision Trees)
- Doğrusal Regresyon (Linear Regression)
- Lojistik Regresyon (Logistic Regression)

3.7.6.2. Danışmansız (Gözetimsiz) Öğrenme

Etiket olmayan veri setlerinde kullanılır. Amaç, verideki gizli ilişkileri veya grupları keşfetmektir. Örneğin, bilgisayar fiyatlarını bilmeden sadece teknik özelliklerine bakarak fiyat tahmini yapmaya çalışmak gibi. Google Haberler gibi siteler haberleri içeriklerine göre kümelere ayırır (IBM. 2021)

- Kümeleme (Clustering)
- Birliktelik Kuralları (Association Rules)
- Temel Bileşen Analizi (PCA).

3.7.6.3. Yarı Danışmanlı Öğrenme

Az sayıda etiketli, çok sayıda etiketsiz veriyle öğrenme yapılır. Etiketlenmiş veri az, etiketsiz veri çoksa kullanılır. Aksi halde kullanılmaz (IBM. 2023)

3.7.6.4. Takviyeli Öğrenme (Reinforcement Learning)

Sistemin ortamla etkileşerek doğru kararlar almayı öğrenmesidir. Ödül ve ceza mekanizmasıyla öğrenir. Örnek alanlar: robotik, oyun, sağlık teşhisi, otomasyon. Aşağıdaki tabloda makine öğrenmesi yöntemleri arasındaki farklar tablolaştırılıp yorumlanmıştır (Neptune.ai. 2021).

Tablo 3.4. Makine Öğrenmesi Türleri

Öğrenme Türü	Hedef Değişken	Örnek Uygulamalar
Danışmanlı	Sürekli / Kategorik	Regresyon (stok tahmini), Sınıflama (medikal görüntüleme)
Danışmansız	Yok	Kümeleme (müşteri segmentasyonu), Birliktelik (market basket analizi)
Yarı Danışmanlı	Kategorik	Metin sınıflandırma, GPS şerit bulma
Takviyeli	Kategorik / Yok	Pazarlama optimizasyonu, sürücüsüz araçlar

3.7.7. Aşırı Öğrenme (Overfitting)

Modelin eğitim verisini çok iyi öğrenip ezberlediği durumdur. Bu durumda model, yeni verilere genelleme yapamaz. Varyans yüksek, yanlılık düşük olur. Örneğin; bir birey yakın arkadaşını çok iyi tanır ve arkadaşına yapacağı bir şaka ile ne tür bir tepki verebileceğini kestirebilir. Lakin, tanımadığı bir insana karşı yapacağı şakada o insanın tepkisini kestiremez (Hastie ve ark., 2009)

3.7.8. Eksik Öğrenme (Underfitting)

Model veriyi yeterince iyi öğrenemez, basit kalır, hem eğitim hem test verisinde kötü sonuç verir. Bu durumda aşırı öğrenmedeki gibi yanlılık yüksek, varyans düşüktür (Hastie ve ark., 2009).

3.7.9. Yanlılık-Varyans

3.4.7. ve 3.4.8.'de bahsi geçen yanlılık, modelin gerçek ilişkiyi ne kadar iyi öğrenemediğinin ölçüsüdür. Varyans, modelin eğitim verisindeki küçük değişikliklere ne kadar hassas olduğunu ifade eder ve yüksek varyans aşırı öğrenmeye yol açar. İyi model, yanlılık ve varyans arasında dengeli olan, hem eğitim hem de yeni verilerde iyi tahmin yapan modeldir (Géron, 2019).

3.7.10. Model İyileştirme Teknikleri

3.7.10.1. Regularization (Düzenleme)

Modelin karmaşıklığını azaltmak için cezalandırma yapılan düzenleştirme tekniğidir. (Lasso, Ridge regresyon) (Hastie, Tibshirani & Friedman, 2009).

3.7.10.2. Bagging

Eğitim verisini rastgele alt kümelere bölüp farklı modeller oluşturur ve sonuçları birleştirir (örneğin Random Forest regresyon) ve varyansı azaltır (Breiman, 1996).

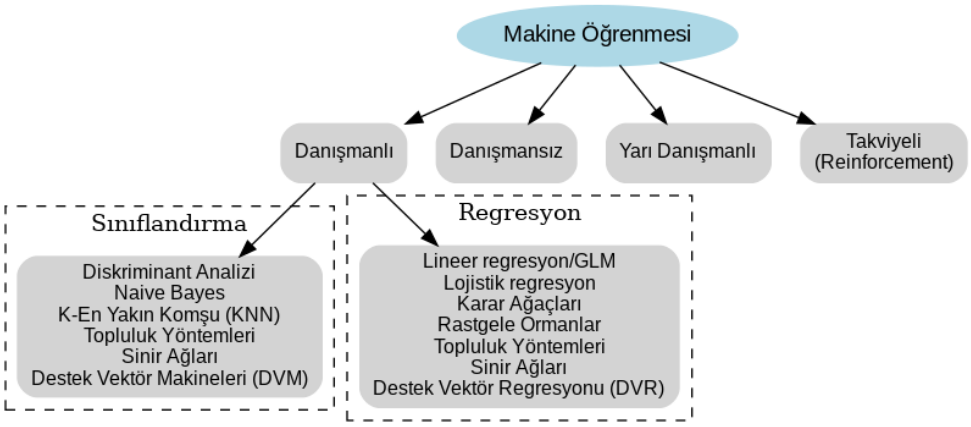
3.7.10.3. Boosting

Zayıf modelleri ardışık olarak geliştirir, önceki modellerin hata yaptığı yerlere daha çok ağırlık verir. Yanlılığı azaltır (Kohavi, 1995).

3.7.10.4. Çapraz Doğrulama (Cross-Validation)

Model performansını daha güvenilir ölçmek için veri seti k parçaya bölünür. Her parça bir kez test seti, diğerleri eğitim seti olur. Böylece model k kez test edilir ve ortalama hata hesaplanır. En yaygın yöntem 10-katlı çapraz doğrulamadır. Özellikle Polinomik regresyon katsayılarının derecesini göstermek gibi kısımlarda çapraz doğrulamalaradan oldukça faydalanılır (Kohavi, 1995).

3.7.11. Danışmanlı Makine Öğrenmesi Teknikleri



Şekil 3.2. Danışmanlı Makine Öğrenmesi Türleri

3.7.11.1. Regresyon Analizi (Linear Regression)

Regresyon analizi, aralarında sebep-sonuç ilişkisi bulunan iki veya daha fazla değişken arasındaki ilişkiyi belirlemek ve bu ilişkiyi kullanarak o konu ile ilgili tahminler (estimation) ya da kestirimler (prediction) yapabilmek amacıyla yapılır. Doğada birçok olayda sebep sonuç ilişkisine rastlamak mümkündür (Fikret G. 2023).

3.7.11.1.1. Regresyon Analizinin Varsayımları

Regresyon analizinin yapılabilmesi için geçerli varsayımları sağlaması beklenir. Aşağıda regresyon analizinin genel varsayımları bulunmaktadır:

- Bağımlı ve bağımsız değişkenler arasında doğrusal ilişki olmalıdır.
- Gözlemler birbirinden bağımsız olmalıdır.
- Hata terimleri normal dağılmalıdır.
- Çoklu bağlantı olmamalıdır.

3.7.11.1.2. Doğrusal Regresyon Türleri

3.7.11.1.2.1 Basit Doğrusal Regresyon

Tek bir bağımsız değişken ile bağımlı değişken arasındaki doğrusal ilişkiyi inceleyen regresyon türüdür. Basit doğrusal regresyon modeli eşitlik (3.39)'daki gibidir.

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{3.39}$$

Burada, Y bağımlı değişken, X bağımsız değişken, β_0 sabit terim (intercept), β_1 regresyon katsayısı ve ε hata terimidir.

3.7.11.1.2.2. Çoklu Doğrusal Regresyon

Birden fazla bağımsız değişkenin, bağımlı değişken üzerindeki etkilerini inceleyen regresyon türüdür.

$$Y = \beta^0 + \beta^1 X^1 + \beta^2 X^2 + \ldots + \beta_n X_n + \varepsilon \tag{3.40}$$

$X_1, X_2, \ldots, X_n$ *b*ağımsız değişkenlerdir ve $\beta_1, \beta_2, \ldots, \beta_n$ katsayılarıdır.

3.7.11.1.2.3. Regresyon Katsayılarının Hesabı

Basit doğrusal regresyonda katsayılar analitik olarak aşağıdaki formüllerle (3.41) aracılığıyla hesaplanır.

$$\boldsymbol{\beta_1} = \frac{\sum(\boldsymbol{X_{ii}} - \overline{\boldsymbol{X}})(\boldsymbol{Y_i} - \overline{\boldsymbol{Y}})}{\sum(\boldsymbol{X_i} - \overline{\boldsymbol{X}})^{\boldsymbol{2}}} \tag{3.41}$$

$$\boldsymbol{\beta_0} = \overline{\boldsymbol{y}} - \boldsymbol{\beta_1}\overline{\boldsymbol{x}}$$

Çoklu doğrusal regresyonda ise hesaplama matris cebriyle yapılır. Regresyon katsayıları eşitlik (3.42) ile hesaplanır.

$$\beta_i = (X^T X)^{-1} X^T Y \tag{3.42}$$

Burada:

- X: Gözlem matrisidir. (bağımsız değişkenler)
- Y: Bağımlı değişken vektörü,
- β: Katsayılar vektörü,
- X^T: X matrisinin transpozu,
- $(X^T X)^{-1}$: X'in normal denklem matrisinin tersidir.

3.7.11.1.2.4. Hata Terimi ve Artıklar (Residuals)

Modelin tahmin ettiği değer ile gözlenen değer arasındaki fark **artık** (residual) olarak adlandırılır. Gerçek hata terimi ϵ gözlemlenemez. Artıklar eşitlik (3.43) ile hesaplanır.

$$e_i = y_i - \widehat{y_i} \tag{3.43}$$

Burada:

- y_i: Gerçek gözlem değeri
- $\widehat{y_i}$: Modelin tahmini değeri
- e_i: Artık (gözlemlenen hata)

3.7.11.1.2.5. Standart Hata ve Hata Varyansı

Regresyon katsayılarının güvenilirliğini değerlendirmek için **standart hata** hesaplanır. Bu, tahmin edilen katsayının örneklem hatasını verir. Basit regresyonda β_1 için standart hata formülü eşitlik (3.44)'de verilmiştir.

$$SE(\beta_1) = \sqrt{\frac{\widehat{\sigma}^2}{\sum(X_i - \overline{X}_i)^2}} \tag{3.44}$$

Burada $\hat{\sigma}^2$ hata terimlerinin tahmini varyansıdır ve eşitlik (3.45)'deki gibi hesaplanır.

$$\hat{\sigma}^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-k-1} \tag{3.45}$$

3.7.11.1.2.6. Katsayıların Anlamlılığı

3.7.11.1.2.6.1. Hipotez Testi

Yokluk hipotezi (H0): Katsayı βi = 0 (değişkenin etkisi yok)

Alternatif hipotez (H1): Katsayı βi ≠ 0 (değişkenin etkisi var)

3.7.11.1.2.6.2. t-Değerinin Hesaplanması

$$t = \frac{\widehat{\beta}_i}{SE(\widehat{\beta}_i)} \tag{3.46}$$

Eşitlik (3.46)'dan:

- β tahmin edilen katsayı,
- $SE(\hat{\beta}_i)$ ise katsayının standart hatasıdır.

3.7.11.1.2.6.3. Anlamlılık Seviyesi ve Karar

Hesaplanan t değeri, seçilen anlamlılık seviyesi (genellikle α=0.05) ve serbestlik derecesine göre t-dağılımından bulunan kritik değerle karşılaştırılır.

Eğer $t > t_{kritik}$, ise H0 reddedilir ve katsayı anlamlı kabul edilir.

3.7.11.1.2.6.4. p-Değeri ile Değerlendirme

p-değeri kritik değerden küçükse (p <α), katsayı anlamlıdır.

p-değeri büyük veya eşitse, katsayı anlamlı değildir.

3.7.11.1.3. Regresyon Analizinin Adımları

Regresyon analizinde ilk adım, analiz edilecek veri setinin toplanması ve uygun şekilde temizlenmesidir.

Probleme uygun regresyon modeli kurulur. Kurulan modelin doğruluğu, R^2 değeri ve p-değerleri gibi çeşitli istatistiksel testler ve değerlendirme metrikleriyle test edilir.

Daha sonra, model kullanılarak tahminler yapılır. Son olarak, regresyon katsayıları ve diğer istatistiksel çıktılar yorumlanarak modelin anlamlılığı değerlendirilir.

3.7.11.1.4. Regresyon Model Değerlendirme Metrikleri

3.7.11.1.4.1. R^2 (R-Kare)

Doğrusal regresyon ve diğer genel modelleme yöntemlerinin performansını ölçen temel metriktir. Modelin veriyi ne kadar iyi öğrendiğini ifade eder ve toplam varyansın ne kadarının açıklanabilir olduğunu söyler (Montgomery ve ark., 2012).

R^2 değeri 0 ile 1 arasında değişir:

- **$R^2 = 1$**: Modelin varyans açıklama oranı tamdır.
- **$R^2 = 0$**: Model veriyi açıklayamaz.
- **Negatif R^2**: Model olurundan kötü tahminler yapıyordur.

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \tag{3.47}$$

- y_i: Gerçek değerler,
- $\hat{y}_i$: Modelin tahmin ettiği değerler,
- $\overline{y}_i$: Gerçek değerlerin ortalamasıdır

3.7.11.1.4.2. MSE (Mean Squared Error)

Tahmin hatalarının ortalama karelerinin alındığı hata ölçüsüdür. Tahmin ile gerçek değerler arasındaki farkları ifade eder. Küçük MSE, daha doğru tahmin yapıldığını ifade eder (James ve ark., 2013).

$$MSE = \frac{1}{n}\sum(y_i - \hat{y}_i)^2 \tag{3.48}$$

- y_i: Gerçek değerler,
- $\hat{y}_i$: Modelin tahmin ettiği değerler,
- n: Veri kümesindeki örnek sayısıdır.

Büyük hatalar küçük hatalara göre daha fazla cezalandırılır çünkü hata terimleri karesine alınır. Bu nedenle, modelin büyük hatalardan kaçınması gerekir.

MSE verinin biriminden farklı bir birimde ifade edilir (örneğin, fiyat verisi için MSE TL² cinsinden olur). Bu nedenl, anlaşılabilirliği sınırlı olur. Bu sorunun önüne geçmek için **RMSE (Root Mean Squared Error)** kullanılır.

3.7.11.1.4.3. RMSE (Root Mean Squared Error)

Tahmin hatalarının karekköüdür. MSE'nin karekökü ile bulunur ve değer birim cinsten yorum yapmayı sağlar. Böylece hataların büyüklüğü azalır (James ve ark., 2013).

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \tag{3.49}$$

Burada;

- y_i: Gerçek değerler,
- $\hat{y}_i$: Modelin tahmin ettiği değerler,
- n: Veri kümesindeki örnek sayısıdır.

3.7.11.1.4.4. MAE (Mean Absolute Error)

Tahmin edilen değer ile gerçek değerler arasındaki farkların mutlak değeri alınarak elde edilen hata ölçütüdür. Hataların kareleri yerine mutlak değerleri ele alınarak elde edilir ve RMSE'ye göre daha duyarlıdır.

$$MAE = \frac{1}{n}\sum|y_i - \hat{y}_i|^2 \tag{3.50}$$

Burada;

- y_i: Gerçek değerler,
- $\hat{y}_i$: Modelin tahmin ettiği değerler,
- n: Veri kümesindeki örnek sayısıdır.

3.7.11.1.5. Çoklu Bağlantı Problemi (Multicollinearity)

Regreson analizinde işleme giren parametreler arasında olan yüksek korelasyon sonucu ortaya çıkan problemdir.

3.7.11.1.5.1. Çoklu Bağlantının Etkileri

Problem. regresyon katsayılarının tahmin edilmesini zorlaştırır. Bağımsız değişkenler arasındaki yüksek korelasyon, katsayıların standart hatalarını artırır ve bu da istatistiksel anlamlılık testlerini geçmelerini engeller. Ek olarak bu yüksek korelasyon, regresyon katsayısını bazı durumlarda zıt gösterir.

3.7.11.1.5.2. Çoklu Bağlantının Tespiti

Korelasyon ısı haritası aracılığı ile değişkenler arası korelasyonlar incelenerek tespit edilebilir. Ayrıca VIF (Variance Inflation Factor) ile herhangi bir bağımsız değişkenin diğerleriyle ilişkisi ölçülür ve VIF skoru 10'un üzerinde olan değerlerde çoklu bağlantı problemi görülür.

3.7.11.1.5.3. Çoklu Bağlantı Sorununu Gidermek

Yüksek korelasyona sahip bağımsız değişkenlerin modelden çıkarılmasıyla değişken seçimi yapılabilir; ayrıca Principal Component Analysis (PCA) yöntemiyle veri bileşenlere ayrılarak bağımsız değişken sayısı azaltılıp korelasyonlar ortadan kaldırılabilir; bunun yanında Ridge regresyonu kullanılarak katsayılara ceza terimleri uygulanır ve böylece modelin aşırı uyumu engellenerek çoklu bağlantı sorunu hafifletilir.

3.7.11.1.6 Adjusted R^2 (Düzeltilmiş R-Kare)

Modelin değişken sayısına bağlı olarak artabilen ölçüttür. Daha fazla değişken eklenmesi durumunda yalnızca gerçekten yararlı değişkenlerin katkısını gösterir. modelin karmaşıklığını artıran gereksiz değişkenlerden kaçınılmasına yardımcı olur (James ve ark., 2013)

$$\boldsymbol{Adj(R^2) = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}} \tag{3.51}$$

- $\boldsymbol{R^2}$: Normal R kare değeri
- n: Gözlem sayısı
- k: Bağımsız değişken sayısı

3.7.11.1.7. F-Testi

Regreson modelindeki katsayıların anlamlılığı ile model anlamlılığını test etmeye yarayan istatistiksel testtir.

- **Yokluk Hipotezi (H_0):** Tüm regresyon katsayıları sıfırdır.

$$H_0: \boldsymbol{\beta_1 = \beta_2 = \ldots = \beta_k = 0}$$

Bu durumda modelin bağımlı değişkeni açıklama gücü yoktur.

- **Alternatif Hipotez (H_1):** En az bir regresyon katsayısı sıfırdan farklıdır.

$$\mathbf{H_1: En\ az\ bir\ \beta_i \neq 0}$$

Bu durumda model en az bir bağımsız değişken sayesinde anlamlıdır.

$$F = \frac{(R^2/k)}{((1-R^2)/(n-k-1))} \tag{3.52}$$

Burada:

- R^2: Modelin açıklama gücü,
- k: Bağımsız değişken sayısı,
- n: Toplam gözlem sayısıdır

Alternatif olarak, **ANOVA (Varyans Analizi) tablosu** kullanılarak da eşitlik (3.53)'deki gibi hesaplanabilir.

$$F=\frac{Regresyon\ Kareler\ Toplamı/k}{Hata\ Kareler\ Toplamı/(n\text{-}k\text{-}1)} \tag{3.53}$$

- $p < \alpha$ ise → H_0 reddedilir, model anlamlıdır.
- $p \geq \alpha$ → H_0 reddedilemez, model anlamsızdır.

3.7.11.2. Polinomik Regresyon (Polynomial Regression)

Doğrusal olmayan ilişkileri modellemek için kullanılır. Bağımsız değişkenlerin daha yüksek dereceden terimlerinin eklenmesiyle doğrusal regresyon modelinin genelleştirilmiş bir formudur (Montgomery ve ark., 2012).

Polinomik regresyon, aslında bir doğrusal regresyon modelidir, ancak bağımsız değişkenin yüksek dereceli terimlerini ekleyerek doğrusal olmayan bir ilişkiyi modellemeye olanak tanır. Model eşitlik (3.54)'deki gibi ifade edilir.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots + \beta_n X^n + \epsilon \quad (3.54)$$

Burada:

- Y'ler, bağımlı değişken,
- X'ler, bağımsız değişken ,
- $\beta 0, \beta 1, \ldots, \beta n \beta_0, \beta_1, \ldots, \beta_{n\beta 0}, \beta 1, \ldots, \beta n$, polinom katsayıları,
- n, polinomun derecesi,
- ϵ, modelin hata terimidir.

3.7.11.3. Üstel Regresyon (Exponential Regression)

Hedef değişkenin bağımsız değişken ile üstel biçimde değiştiği durumlarda kullanılan regresyon modelidir. Model denklemi eşitlik (3.55)'deki gibidir.

$$Y = ae^{bX} \Rightarrow lnY = lna + bX \quad (3.55)$$

Burada logaritma alınarak model doğrusal hale getirilir ve klasik doğrusal regresyon uygulanır.

3.7.11.4. Düzenlileştirme (Regularization) Teknikleri

Modelin hata fonksiyonuna parametre büyüklüklerine bağlı bir ceza (penalty) terimi eklenir. Böylece hem model genelleme kabiliyeti artar hem de katsayıların aşırı büyümesi engellenir (James ve ark., 2013).

En yaygın ve etkin düzenlileştirme teknikleri **Ridge (L2)**, **Lasso (L1)** ve **Elastic Net** regresyonlarıdır. Bunların matematiksel yapısı ve etkileri açıklanmıştır.

3.7.11.4.1. Ridge Regresyon (L2 Düzenlileştirme)

Klasik en küçük kareler (Ordinary Least Squares - OLS) regresyonunun optimize ettiği hata fonksiyonuna katsayıların karelerinin toplamından oluşan bir ceza terimi ekler. Katsayıların büyüklüğü kısıtlanır ve karmaşıklık azalır.

$$\hat{\beta}^{ridge} = \arg min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right) + \Lambda \sum_{j=1}^{p} {\beta_j}^2 \right\} \qquad (3.56)$$

Burada,

- y_i: i. gözlemin hedef değeri,
- x_{ij}: i. gözlemin j. özelliği,
- β_j: j. özellik için regresyon katsayısı,
- $\Lambda \geq 0$: düzenlileştirme parametresi (ceza katsayısı).

$\Lambda = 0$ iken klasik doğrusal regresyonla aynı sonuçları verirken Λ arttıkça regresyon katsayıları küçülür ve model daha sade hale gelir ve böylece aşırı öğrenme azalır. Ekstradan, bağımsız değişkenler arasında yüksek korelasyon yani çoklu bağlantıyı da azaltır.

3.7.11.4.2. Lasso Regresyon (L1 Düzenlileştirme)

Lasso (Least Absolute Shrinkage and Selection Operator) regresyonunda ceza terimi olarak katsayıların karelerinin toplamı yerine **mutlak değerlerinin toplamı** (L1 normu) kullanılır.

$$\hat{\beta}^{lasso} = \arg min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right) + \Lambda \sum_{j=1}^{p} |\beta_j| \right\} \qquad (3.57)$$

L1 normu katsayıların bir kısmını tam olarak sıfıra indirgeyebilir. Veri setindeki önemsiz veya gereksiz değişkenler modelden çıkarılır. Model daha sade, yorumlanabilir hale gelir.

3.7.11.4.3. Elastic Net Regresyon

Elastic Net, Ridge ve Lasso'nun avantajlarını bir araya getirerek hem katsayı küçültme hem de özellik seçimini destekleyen hibrit yöntemdir. Modelin amaç fonksiyonu hem L1 hem L2 cezalarını içerir:

$$\hat{\beta}^{EN} = \arg min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right) + \Lambda_1 \sum_{j=1}^{p} |\beta_j| + \Lambda_2 \sum_{j=1}^{p} {\beta_j}^2 \right\} \quad (3.58)$$

Burada,

- Λ_1 Lasso (L1) ceza katsayısı,
- Λ_2 Ridge (L2) ceza katsayısıdır.

Değişkenler arasında yüksek korelasyon varsa, bu grup içindeki değişkenlerin birden fazlasını modelde tutabilir veböylece Lasso'nun aksine bilgi kaybını önler. Parametreler genellikle çapraz doğrulama (cross-validation) ile optimize edilir.

3.7.11.4.4. Düzenlileştirme Parametresinin (Λ) Önemi

Düzenlileştirme parametresi **Λ**, modelin karmaşıklığını ve katsayıların küçültülme derecesini belirler.

- Küçük **Λ** değerleri cezayı zayıflatarak klasik regresyon çözümü sunar.
- Büyük **Λ** değerleri katsayıları çok küçültür ve model az öğrenir.
- Optimal **Λ** değeri genellikle K-katlı çapraz doğrulama yöntemi ile belirlenir. Model böylece optimize edilebilir.

3.7.11.4.5. Düzenlileştirmenin Model Performansına Etkisi

Düzenlileştirme hem eğitim verisi üzerinde hata (bias) miktarını artırabilir hem de modelin varyansını azaltarak test verisi üzerindeki genel hata oranını düşürür. Bu, **bias-variance tradeoff** olarak adlandırılan temel makine öğrenmesi prensibinin uygulamasıdır.

Tablo 3.5. Düzenleştirme Yöntemlerinin Özeti

Yöntem	Ceza Türü	Etki	Kullanım Alanı
Ridge	L2 normu	Katsayıları küçültür, tüm değişkenleri tutar	Multicollinearity varsa, tüm değişkenlerin önemi varsa
Lasso	L1 normu	Katsayıları sıfıra çekerek değişken seçimi yapar	Model sadeleştirme ve önemli değişken seçimi gerektiğinde
Elastic Net	L2 normu	Hem küçültme hem seçimi dengeler, grup değişkenleri destekler	Yüksek boyutlu ve korele değişkenli veri setlerinde

3.7.11.5. Hedonik Fiyatlandırma Modeli

Konut fiyatları üzerine yapılan ilk akademik çalışmalar genellikle **Hedonik Fiyatlandırma Modeli (Hedonic Pricing Model)** çerçevesinde geliştirilmiştir. Bir ürünün fiyatının, onu oluşturan niteliklerin ağırlıklı toplamı olduğu varsayımına dayanır. Bu bağlamda, bir konutun fiyatı eşitlik (3.59) ile ifade modellenir (Rosen,1974)

$$Pi == \beta_0 + \beta_1 x\{1i\} + \cdots + \beta_k x\{ki\} + \varepsilon_i \qquad (3.59)$$

Burada:

- Pi; i'nci konutun satış fiyatı,
- $x\{1i\}, \ldots, x\{ki\}$; konuta ait bağımsız değişkenler (metrekare, oda sayısı, bina yaşı vb.),
- $\beta_0, \ldots, \beta_k$; regresyon katsayıları,
- ε_i, hata terimidir.

3.7.11.5. Hedonik Fonksiyonel Formlar

Hedonik modelde doğrusal fonksiyonel yapının yanı sıra, farklı dönüşümler de kullanılabilir.

- **Lineer form:** Yukarıda belirtilen standart çoklu regresyon.
- **Doğal Logaritma dönüşümü:** Genellikle fiyat ve/veya bazı özelliklerin logaritması alınarak modelde kullanılır. Bu, değişkenler arasındaki ilişkiyi doğrusal olmayan ama parametre yorumlarını kolaylaştıran bir biçime dönüştürür.
- **Logaritmik ve Kök Dönüşümü:** Verinin dağılımına bağlı olarak, modelin doğrusal varsayımını güçlendirmek için tercih edilir.

$$lnY = \beta_0 + \sum_{i=1}^{n} \beta_i lnX_i + \varepsilon \qquad (3.60)$$

Eşitlik (3.60)'da katsayılar elastikiyet olarak yorumlanabilir; yani X_i'de %1 değişim fiyatı Y'yi ortalama olarak $\%\beta_i$ kadar değiştirir.

Doğru fonksiyonel formun seçilmesinin zorluğu arz ve talebin doğrudan modele dahil edilmemesi nedeniyle dezavantajları da bulunur.

3.7.11.6. Random Forest Regresyonu

Sınıflandırma ve regresyon problemlerinde kullanılan ve çok sayıda karar ağacını analiz eden topluluk öğrenme modeldir. Tek bir karar ağacının aşırı öğrenme sorununu azaltmak için geliştirilmiştir (Breiman, 2001).

Random Forest, eğitim sürecinde aşağıdaki iki rastgelelik ilkesini uygular.

- **Bootstrap Örnekleme:** Her bir karar ağacı orijinal veri setinden tekrar seçmeli rastgele örneklem (bootstrap sample) alınarak eğitilir. Böylece her ağaç farklı bir eğitim seti üzerinde oluşturulur.
- **Özellik Alt Kümesi Seçimi:** Her ağacın düğümlerinde bölünme yapılırken,tüm bağımsız değişkenler yerine rastgele seçilen alt küme özellikler değerlendirilir. Bu da ağaçlar arasındaki korelasyonu düşürür.

Veri seti eşitlik (3.61)'deki gibi olur.

$$D = \{(x_i, y_i)\}_{i=1} \tag{3.61}$$

Burada:

- $x_i \in R^p$: p boyutlu bağımsız değişken vektörüdür.
- $y_i \in R$: hedef (bağımlı) değişkendir.

Her bir ağaç için bootstrap örneklem $D^{(m)} \subset$ D rastgele oluşturulur. Bu örneklem kullanılarak T_m verilen m'inci karar ağacı eğitilir. Yeni bir gözlem x verildiğinde,m'inci ağacın tahmini eşitlik (3.62)'deki gibi hesaplanır.

$$\hat{y}(x) = \frac{1}{M}\sum_{m=1}^{M} T_m(x) \tag{3.62}$$

Burada M, toplam ağaç sayısıdır.

3.7.11.6.1. Parametre Ayarları ve Etkileri

- **Ağaç Sayısı (n_estimators):** Model performansını artırmak için arttırılsa yüksek maliyetli olur.
- **Maksimum Derinlik (max_depth):** Ağaçların aşırı büyümesini engelleyerek aşırı uyumu azaltan parametredir.
- **Bölünmede Kullanılan Özellik Sayısı (max_features):** Regresyondaki özellik sayısının daha azı ele alınır.
- **Düğümlerdeki Minimum Örnek Sayısı (min_samples_split, min_samples_leaf):** Düğümlerde örnek minimum sınır belirlenir.

Doğrusal olmayan karmaşık ilişkileri yakalayabilir, aykırı değerlere dayanıklıdır. Model başarısı veri setine göre değişken olsa da genelleme başarıları yüksek olur. Aşırı öğrenmeye tolerası vardır. Özellik önem düzeyleri hesaplayabilir be bu hesapladığı önem düzeylerine bakarak bağımsız değişkenlerin etkisini analiz edebilir.

3.7.11.7. Boosting Algoritmaları

Öngörüsel performansı artıran güçlü topluluk öğrenme yöntemidir.

Erken dönem yöntemlerden biri olan AdaBoost (Freund & Schapire, 1997), yanlış sınıflandırılan gözlemlere daha fazla ağırlık verir, her yeni zayıf modelin bu hatalara odaklanmasını sağlar. Daha sonra geliştirilen Gradient Boosting Machine (Friedman, 2001) ile birlikte, boosting algoritmaları kayıp fonksiyonunun gradyanını minimize etme prensibiyle çalışmaya başladı. Bu yaklaşımın gelişmiş türevleri olan XGBoost, LightGBM ve CatBoost gibi algoritmalar, büyük veri setlerinde yüksek doğruluk ve esneklik sunmaktadır. (Friedman, 2001; Chen & Guestrin, 2016).

Basit ve genellikle karar ağaçlarına dayalı temel modellerin bir araya getirilmesiyle karmaşık veri ilişkilerini yakalayabilirler. Amaardışık yapılarından dolayı paralel işlemeye sınırlı şekilde uygundurlar. Ayrıca da aşırı öğrenmeye karşı hassas oldukları için, modelin hiperparametre ayarlarının dikkatli biçimde yapılması gerekiyor (Chen & Guestrin, 2016).

3.7.11.7.1. Gradient Boosting

Zayıf karar ağaçlarını eğitip birleştirerek güçlü bir tahmin modeli yapar. Her yeni model önceki modelin hatalarını minimize ede ede ilerler ve böylece yüksek doğruluk elde edilmesi amaçlanır.

Önce basit bir model oluşturulur (örneğin, hedef değişkenin ortalaması). Daha sonra her iterasyonda m. zayıf ağaç, mevcut modelin hatalarını azaltmak üzere eğitilir. Numerik analiz iterasyonlarına dayanır. Yeni model, önceki modelin üzerine bu zayıf ağacın belli bir ağırlıkla eklenmesiyle güncellenir. Aşağıda Gradient Boosting Güncelleme Denklemi eşitlik (3.63)'deki gibidir.

$$F_{m(x)} = F_{m-1(x)} + \gamma_m h_{m(x)} \tag{3.63}$$

- $F_{m(x)}$: m. iterasyondan sonra elde edilen tahmin fonksiyonu (topluluk modeli)
- $F_{m-1(x)}$: bir önceki iterasyondaki model
- $h_{m(x)}$m. iterasyonda öğrenilen zayıf öğrenici (örneğin karar ağacı)
- γ_m: **adım büyüklüğü (step size)** ya da **learning rate**

$\boldsymbol{\gamma_m}$, modelin "aceleci" davranmasını engelleyen bir "kontrol frenidir".

$$D = \{(x_i, y_i)\}_{i=1} \tag{3.64}$$

Amaç, kayıp fonksiyonu minimize eden fonksiyon F(x)'i bulmaktır. İlk model (3.67)'deki gibi kurulur.

$$F_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma) \tag{3.65}$$

Her iterasyonda:

1. Mevcut modelin negatif gradyanı (residüeller olarak da adlandırılır) (3.68)'deki gibi hesaplanır.

$$r_{im} = -\left[\frac{\partial \mathbf{L}(\mathbf{yi},\mathbf{F}(\mathbf{x_i}))}{\partial \mathbf{F}(\mathbf{x_i})}\right]_{F(x)=F_{m-1}(x)} \tag{3.66}$$

Bu değerler, bir sonraki zayıf ağacın hedef değişkeni olarak alınır.

2. Yeni zayıf ağaç $h_m(x)$ negatif gradyan üzerinde eğitilir.

$$\mathbf{h_m(x)} \approx \mathbf{r_{im}} \tag{3.67}$$

Karar ağacı genellikle bu gradyan değerlerine göre eğitilmelidir.

3. Optimal ağırlık (γ) eşitlik (3.68)'deki gibi bulunur.

$$\gamma_m = argmin_\gamma \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma_m h_m(\mathrm{x}_i)\,) \tag{3.68}$$

4. Model eşitlik (3.69)'daki gibi güncellenir.

$$F_{m(x)} = F_{m-1(x)} + \gamma_m h_{m(x)} \tag{3.69}$$

Bu süreç, belirlenen iterasyon sayısına veya hata toleransına ulaşılana kadar devam eder. Aşağıda özel kaybı fonksiyonlarının analitik formülleri yer almaktadır.

Regresyon için En Küçük Kareler, eşitlik (3.70)'deki gibi hesaplanır.

$$L\big(y, F(x)\big) = \tfrac{1}{2}\big(y - F(x)\big)^2 \tag{3.70}$$

Negatif gradyan, eşitlik (3.71) ile hesaplanır.

$$\mathbf{r_{im}} = y_{\,i} - F_{m-1}(x_i) \tag{3.71}$$

Logistik kayıp, eşitlik (3.72) ve $\mathbf{r_{im}}$ katsayısı (3.73) aracılığıyla hesaplanır.

$$\mathbf{L}\big(\mathbf{y}, \mathbf{F}(\mathbf{x})\big) = \log\Big(1 + \exp\Big(-2\mathrm{y}\big(\mathrm{F(x)}\big)\Big)\Big), \mathbf{y} \in \{-\mathbf{1}, +\mathbf{1}\} \tag{3.72}$$

$$\mathbf{r_{im}} = \frac{\mathbf{2y_i}}{1+\exp(2\mathrm{y}_i F_{m-1}(x_i))} \tag{3.73}$$

3.7.11.7.1.2. XGBoost

XGBoost, Gradient Boosting algoritmasının daha hızlı, daha verimli ve daha doğru sonuçlar üretecek şekilde optimize edilmiş, düzenlileştirilmiş (regularized) ve paralelleştirme destekli bir versiyonudur (Chen & Guestrin, 2016).

Kaybı ve Düzenlileştirmeyi İçeren Amaç Fonksiyonu (3.74) ile hesaplanır.

$$L = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(\mathrm{f}_k) \tag{3.74}$$

burada $f_k(x)$ karar ağaçlarıdır ve:

$$\boldsymbol{\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} {w_j}^2} \tag{3.75}$$

Eşitlik (3.77)'den:

T: yaprak sayısı,

w_j: yaprak skorları

2. dereceden Taylor açılımıyla amaç fonksiyonu eşitlik (3.76)'daki gibi yaklaşıklaştırılır.

$$\mathbf{L} \approx \left[\sum_{i=1}^{n} g_i f_t(x_i) + \frac{1}{2} h_i {f_t}^2(x_i)\right] + \Omega(\mathrm{f}_t) \tag{3.76}$$

$$g_i = \frac{\partial \mathbf{L}(\mathrm{y}_i \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)2}} \tag{3.77}$$

Her ayrım için kazanç (Gain) hesaplanır ve ağaç bölünmesi yapılır.

3.7.11.7.6. AdaBoost

Zayıf sınıflandırıcıları (karar ağaçları) ardışık olarak eğitip ağırlıklı birleştirerek güçlü bir sınıflandırıcı oluşturmayı amaçlayan bir ensemble (topluluk) öğrenme yöntemidir. En temel şekli ikili sınıflandırma problemleri içindir fakat genelleştirilmiş versiyonları da vardır (Chen & Guestrin, 2016).

Her iterasyonda, önceki modelin yanlış sınıflandırdığı örneklere daha fazla önem verilir. Böylece öğreniciler adaptif olarak hata yapan örnekler üzerine odaklanır.

- **Zayıf ağaçlar:** Genelde derinliği 1 olan ağaç
- **Ağırlıklı örnekler:** Hatalı tahmin edilen örneklerin ağırlığı artırılır.
- **Topluluk tahmini:** Final sınıflandırma kararı zayıf sınıflandırıcıların ağırlıklı oylamasıyla alınır.

Veri seti eşitlik (3.78)'deki gibidir.

$$\boldsymbol{D} = \{(\boldsymbol{x_i}, \boldsymbol{y_i})\}_{\boldsymbol{i=1}}, \boldsymbol{y} \in \{-\mathbf{1}, +\mathbf{1}\} \tag{3.78}$$

- **Başlangıç:** Her örneğe eşit ağırlık verilir.

$$w_i^{(1)} = \frac{1}{n}, \forall i \tag{3.79}$$

- **Her iterasyon için (m = 1, ..., M):**

  Önce zayıf ağaç eğitilir.

$$h_m(x) \tag{3.80}$$

  Ağırlıklı hata, eşitlik (3.81) aracılığı ile hesaplanır.

$$\varepsilon_m = \frac{\sum_{i=1}^{n} w_i^{(m)} 1[h_m(x_i) \neq \boldsymbol{y_i}]}{\sum_{i=1}^{n} w_i^{(m)}} \tag{3.81}$$

  Model ağırlığı, eşitlik (3.82) ile hesaplanır.

$$\propto_m = \frac{1}{2} ln\left(\frac{1-\varepsilon_m}{\varepsilon_m}\right) \tag{3.82}$$

  Örnek ağırlıkları, eşitlik (3.83)'le güncellenir.

$$w_i^{(m+1)} = w_i^{(m)} e^{-\propto_m \boldsymbol{y_i} h_m(x_i)} \tag{3.83}$$

Akabinde de eşitlik (3.84) ile normalizasyon yapılır.

$$w_i^{(m+1)} = \frac{w_i^{(m+1)}}{\sum_{j=1}^{n} w_j^{(m+1)}} \tag{3.84}$$

En sonunda da final model elde edilir.

$$H(x) = sign\left(\sum_{m=1}^{M} \propto_m h_m(x)\right) \tag{3.85}$$

Hatalı örnekleri dikkate alarak iyileşir, aşırı öğrenme yapmaz. Model yorumlanabilir (özellikle decision stump ile). Gürültüye karşı makul derecede dayanıklıdır. Aşırı uyumdan da kaçınır.

Tablo 3.6. Boosting Modellerinin Özeti

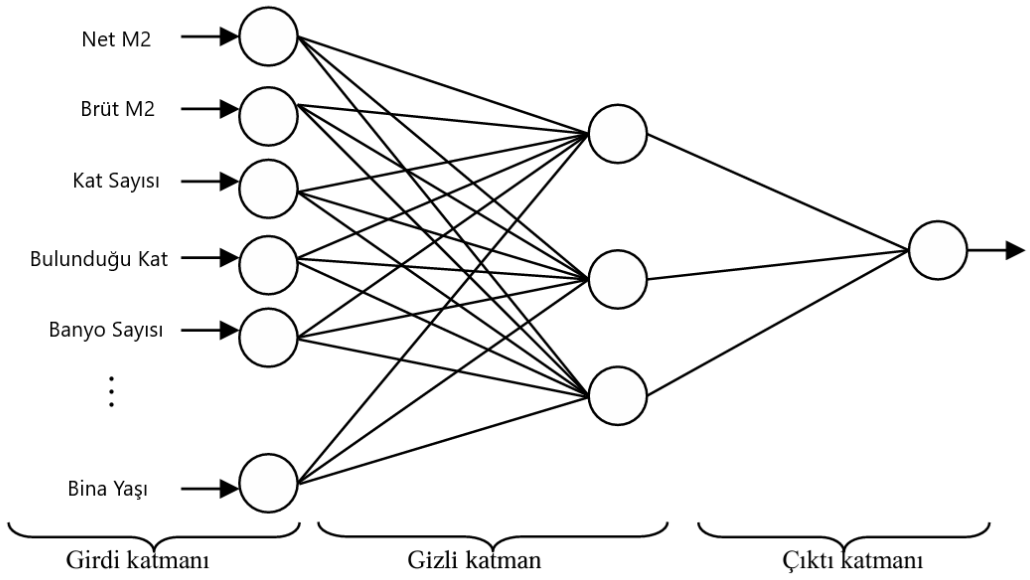
Özellik	**AdaBoost**	**Gradient Boosting**	**XGBoost**
Yıl	1996	1999	2016
Temel Mantık	Hatalı örneğe ağırlık	Gradyan hatasına göre	Regularize GBDT
Hız	Yavaş	Orta	Hızlı

3.7.11.8. Yapay Sinir Ağları (YSA)

İnsanın biyolojik sinir sistemlerinden esinlenerek geliştirilen ve çok katmanlı yapıdaki yapay nöronlardan oluşan modellerdir. Karmaşık veri yapılarındaki ilişkileri öğrenme, genelleme yetenekleri sayesinde makine öğrenimi ve yapay zeka alanlarında geniş uygulama alanı bulurlar.

Çalışmada YSA modellerinden biri olan MLP modeli tercih edilmiştir. MLP modelinde kullanılan bağımlı ve bağımsız değişkenler karşılaştırmanın sağlıklı bir şekilde yapılabilmesi için hedonik modelde kullanılanlarla aynıdır. Diğer bir ifadeyle logaritmik Fiyat bağımlı değişken olup konuta ilişkin diğer özellikler ise bağımsız değişkenler olarak modelde yer alıyor.

Ek olarak bu değişkenlerin birçoğu YSA ile konut fiyatının belirlendiği diğer çalışmalarla örtüşmektedir (Zhang, Patuwo & Hu, 1998)



Şekil 3.4. Yapay Sinir Ağı Katmanları

3.7.11.8. Yapay Nöron Modeli

YSA temelinde biyolojik nöronları taklit eden yapay nöronlar yatmaktadır. Bir yapay nöronun aldığı girdiler her biri belirli ağırlıklarla çarpılır ve toplanır. Akabinde aktivasyon fonksiyonu elde edilir.

Matematiksel olarak, n adet girişe sahip bir nöronun çıkışı (3.86)'daki gibidir.

$$z = \sum_{i=1}^{n} w_i x_i + b \tag{3.86}$$

Eşitlik (3.88)'den:

- x_i: i-inci giriş sinyali,
- w_i: i-inci girişe ait ağırlık,
- b: bias terimi.

Nöron çıkışı ise aktivasyon fonksiyonu f uygulanarak eşitlik (3.87) ile hesaplanır.

$$a = f(z) \tag{3.87}$$

Yani nöronun nihai çıkışı, ağırlıklı toplam ve bias'tan oluşan z değerine bir aktivasyon fonksiyonu uygulanarak bulunur.

3.7.11.8.1. Aktivasyon Fonksiyonları

Nöronun doğrusal olmayan dönüşümünü sağlayan fonksiyonlardır. Ağ doğrusal olmayan ilişkileri öğrenebilir.

- **Sigmoid:**

$$f(z) = \frac{1}{1+e^{-z}} \tag{3.88}$$

Çıkış aralığı (0,1), türevi kolay hesaplanan ancak eğitimde gradyan sönmesi problemlerine yol açan aktivasyon fonksiyonudur.

- **Hiperbolik Tanjant (Tanh):**

$$\boldsymbol{f(z)} = tanh(\boldsymbol{z}) = \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}} \tag{3.89}$$

Çıkış aralığı (-1,1), sıfıra daha merkezlidir ve genellikle Sigmoid'den daha iyi performans veren aktivasyon fonksiyonudur.

- **ReLU (Rectified Linear Unit):**

$$f(z) = max(0, z) \tag{3.90}$$

3.7.11.8.2. Yapay Sinir Ağı Mimarisi

YSA genellikle üç tip katmandan oluşur:

- **Giriş Katmanı:** Modelin dış dünya ile bağlantısını sağlayan katmandır.
- **Bir veya Daha Fazla Gizli Katman:** Girdilerin karmaşık temsillerini öğrenen katmandır.
- **Çıkış Katmanı:** Tahmin veya sınıflandırma sonuçlarını üreten katmandır.

Her katmandaki nöron sayısı ve katman sayısı modelin kapasitesini belirler.

3.7.11.8.3 İleri Besleme (Forward Propagation)

YSA'da giriş verisinin ağ boyunca katman katman işlenerek çıkışa ulaşmasını ifade eder. Bu işlem sırasında her katmandaki nöron, bir önceki katmanın çıktısını alır, ağırlıklarla çarpar, bias ekler ve aktivasyon fonksiyonu ile sonuç üretir.

Ağa dışarıdan verilen veri, bir vektör (dizi) şeklindedir.
İlk katman, yani giriş katmanının aktivasyonu eşitlik (3.91)'deki vektöre eşit olur.

$$a^{(0)} = x \tag{3.91}$$

Her Katmanda Hesaplama:
Sinir ağı L katmandan oluşuyorsa, her l. katmanda aşağıdaki işlemler yapılır:

Z hesaplama değeri hesaplanır.

$$z^{(l)} = W^{(l)}a^{(l-1)} + b^{(l)} \tag{3.92}$$

Eşitlik (3.94)'den:

- $W^{(l)}$: l. katmanın ağırlık matrisi
- $a^{(l-1)}$: Bir önceki katmanın çıktısı (aktivasyonu)
- $b^{(l)}$: l. katmanın bias (sapma) vektörü
- $z^{(l)}$: l. katmandaki nöronlara giden toplam giriş

Aktivasyon Uygulama (doğrusal olmayan dönüşüm):

$$a^{(l)} = f(z^{(l)}) \tag{3.93}$$

Tüm katmanlardan geçtikten sonra ağın çıktısı, son (L'inci) katmanın aktivasyonudur.

$$a^{(l)} \tag{3.94}$$

Eşitlik (3.94) ağın tahmin ettiği sonuçtur (örneğin bir sınıflandırma problemi için etiket tahmini).

3.7.11.8.4. Kayıp Fonksiyonu (Loss Function)

Modelin tahminleri ile gerçek değerler arasındaki farkı ölçer ve modelin öğrenmesi için geri besleme sağlayan fonksiyondur.

- **Regresyon Problemleri İçin (MSE)**

$$\mathbf{J}(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2 \tag{3.95}$$

- **Sınıflandırma İçin (Çapraz Entropi Loss)**

İkili (binary) sınıflandırma problemlerinde (örneğin e-posta spam mi değil mi?) en yaygın kullanılan kayıp fonksiyonu **Çapraz Entropi (Cross Entropy)**'dir.

$$\mathbf{J}(\boldsymbol{\theta}) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)} log\hat{y}^{(i)} + (1 - y^{(i)})\log(1 - \hat{y}^{(i)})] \tag{3.96}$$

Eşitlik (3.96)'dan:

- m: örnek sayısı,
- $y^{(i)}$: gerçek etiket,
- $\hat{y}^{(i)}$: modelin tahmini.

3.7.11.8.5. Geri Yayılım (Backpropagation)

Ağırlıkların güncellenmesi için kayıp fonksiyonunun ağırlıklara göre türevleri hesaplanır. Zincir kuralı ile her katmandaki hata terimleri eşitlik (3.97) ile hesaplanır.

$$\boldsymbol{\delta}(\mathbf{L}) = \boldsymbol{\nabla}_{\boldsymbol{a}}\mathbf{J} \odot \mathbf{f}'\left(\mathbf{z}^{(\boldsymbol{L})}\right) \tag{3.97}$$

- $\boldsymbol{\nabla}_{\boldsymbol{a}}\mathbf{J}$: Kayıp fonksiyonunun tahmin çıktısına göre türevi.
- $\mathbf{f}'(\mathbf{z}^{(\boldsymbol{L})})$: Aktivasyon fonksiyonunun türevi.
- $\odot$: Elemanlara göre (element-wise) çarpım.

Gizli katmanlar, yani önceki katmanlar için hata eşitlik (3.98) ile hesaplanır.

$$\delta^{(l)} = \left(\boldsymbol{W}^{(l+1)^T}\delta^{(l+1)}\right) \odot \mathbf{f}'\left(\mathbf{z}^{(L)}\right) \tag{3.98}$$

- $\boldsymbol{W}^{(l+1)^T}$: Bir sonraki katmanın ağırlık matrisinin transpozu.
- $\delta^{(l+1)}$: Bir sonraki katmandaki hata.
- $\mathbf{f}'\left(\mathbf{z}^{(L)}\right)$: O katmandaki aktivasyon fonksiyonunun türevi.

Bu şekilde hata, katman katman geriye doğru yayılır.

Ağırlıklar ve bias'lar güncellenmeden önce, kaybın onlara göre türevleri (gradyanlar) eşitlik (3.99) ve (3.100) ile hesaplanır.

$$\frac{\partial \mathbf{J}}{\partial \mathbf{W}^{(l)}} = \boldsymbol{\delta}^{(l)}\left(\boldsymbol{a}^{(l-1)}\right)^{\boldsymbol{T}} \tag{3.99}$$

$$\frac{\partial \mathrm{J}}{\partial \mathrm{W}\mathrm{b}^{(l)}} = \delta^{(l)} \tag{3.100}$$

- $\boldsymbol{a}^{(l-1)}$: Bir önceki katmandan gelen aktivasyon
- $\delta^{(l)}$: Bu katmandaki hatadır.

3.7.11.8.6. Ağırlık Güncelleme

Gradyan inişi ile ağırlıklar eşitlik (3.101) ve (3.102) ile hesaplanır.

$$W^{(l)} \coloneqq W^{(l)} - \eta \frac{\partial \mathbf{J}}{\partial W^{(l)}} \tag{3.101}$$

$$b^{(l)} \coloneqq b^{(l)} - \eta \frac{\partial \mathbf{J}}{\partial b^{(l)}} \tag{3.102}$$

- $W^{(l)}$: l. katmandaki ağırlık matrisi.
- $b^{(l)}$: l. katmandaki bias (sapma) vektörü.
- η: Öğrenme oranı (learning rate) – ne kadar büyük bir adımla güncelleme yapılacağını belirler.
- $\frac{\partial \mathbf{J}}{\partial W^{(l)}}, \frac{\partial \mathbf{J}}{\partial b^{(l)}}$: Kayıp fonksiyonunun türevleri (gradyanlar).

Bu adımda model, hatayı azaltacak yönde ağırlıklarını ve bias'larını günceller. Bu işlem her eğitim örneği veya her minibatch (küçük veri grubu) için tekrar edilir.

3.7.11.9. Kümeleme Algoritmaları

Veri setindeki benzer özellikleri gruplandırmaya yarar. Veri içi ve segmentleri ortaya çıkarmak için kümeleme algoritmaları kullanılır. Konut fiyatı tahmininde, bölgesel fiyat analizi ve fiyat kümelerini belirlemek için kullanılabilir (Han, Pei, & Kamber, 2022).

3.7.11.9.1. K-Means Kümeleme

K-Means, veri setini k adet kümeye ayırmayı amaçlayan en yaygın kümeleme algoritmasıdır.

Veri seti: $\{x_1, x_2, \ldots, x_n\}$, burada her bir veri noktası d boyutlu uzaydadır.

Amaç, küme merkezleri $\{\mu_1, \mu_2, \ldots, \mu_n\}$, olacak şekilde, her bir veri noktasını en yakın küme merkezine atayarak **toplam iç-küme varyansını** minimize etmektir (Jain, 2010).

Optimizasyon Fonksiyonu eşitlik (3.103)'deki gibidir.

$$J = \sum_{j=1}^{k} \sum_{x_i \in S_j} \left|\left| x_i - \mu_j \right|\right|^2 \tag{3.103}$$

Burada:

- S_j: kümedeki veri noktalarının kümesi,
- μ_j: j. kümenin merkezidir.

Bu fonksiyon, tüm veri noktalarının kendi küme merkezlerine olan uzaklıklarının karelerinin toplamını ifade eder. K-Means algoritması bu değeri minimize etmeye çalışır.

Algoritma başlangıçta rastgele k adet küme merkezi seçilerek başlatılır. Her veri noktası genellikle Öklid uzaklığı temel alınarak en yakın küme merkezine atanır. Her kümenin merkezi o kümeye ait veri noktalarının ortalaması alınarak güncellenir (Tan, Steinbach, & Kumar, 2019). Kümelerde atamalar veya merkezler değişmeyene kadar yinelenir. Kötü olan ise, algoritmanın küme sayısının önceden belirlenmesini gerektirmesi ve kümelerin küresel (simetrik ve eşit boyutlu) yapıda olacağı varsayımı gibi bazı sınırlamaları vardır (Han, Pei, & Kamber, 2022).

3.7.11.9.2. DBSCAN

Yoğunluk tabanlı kümeleme algoritmasıdır ve küme sayısını önceden belirlemeye gerek duymaz. Ayrıca aykırı noktaları otomatik olarak tespit edebilir (Ester et al., 1996).

- İki parametre gerektirir:
  - **ε (epsilon):** Bir noktanın komşu sayılacağı yarıçap.
  - **MinPts:** Bir noktayı çekirdek noktası yapacak minimum komşu sayısı.

- **Çekirdek Nokta (Core Point):** ε yarıçapı içinde en az MinPts komşusu olan noktadır.
- **Sınır Noktası (Border Point):** Çekirdek noktanın komşusu olan ancak kendi komşu sayısı MinPts'den küçük olan noktadır.
- **Gürültü Noktası (Noise Point):** Ne çekirdek ne de sınır noktası olan, kümeye ait olmayan noktadır (Han, Pei, & Kamber, 2022).

Öklid Uzaklığı eşitlik (3.104)'deki gibi hesaplanır.

$$\boldsymbol{d(x_i, x_j) = \sqrt{\sum_{j=1}^{d} (x_{il} - \mathbf{x_{jl}})^2}} \qquad \mathbf{(3.104)}$$

Manhattan Uzaklığı eşitlik (3.105)'deki gibi hesaplanır.

$$\boldsymbol{d(x_i, x_j) = \sum_{j=1}^{d} |x_{il} - \mathbf{x_{jl}}|} \qquad \mathbf{(3.105)}$$

Uzaklık ölçümü, kümeleme sonuçlarını doğrudan etkiler. Özelliklerin ölçeklendirilmesi gereklidir işlem öncesi.

4. BULGULAR

4.1. Tanımlayıcı İstatistikler ve Aykırı Değer Analizi

4.1.1 Veri Setinin Genel Tanımlayıcı İstatistikleri

Analiz edilen veri seti için seçilen sayısal değişkenlere ait tanımlayıcı istatistikler Tablo 4.1.'de verilmiştir.

Tablo 4.1. Tanımlayıcı İstatistikler

Değişken	N	Ort.	Med.	Var.	SS	Min	Q1	Q3	Max	IQR
Fiyat (TL)	2836	3.65M	3.15M	3.03E+12	1.74M	925k	2.2M	4.7M	9.05M	2.5M
Brüt m²	2836	105.9	100	2546	50.5	38	55	145	292	90
Net m²	2836	93.2	90	2025	45.0	30	50	130	250	80
Banyo Say.	2836	1.40	1	0.26	0.51	1	1	2	3	1
Bina Yaşı	2836	4.05	0	40.97	6.40	0	0	7.5	31	7.5
Bul. Kat	2763	3.15	3	6.90	2.63	-3	1	4	23	3
Oda Say.	2836	3.04	3	1.05	1.02	2	2	4	7	2
Kat Say.	2830	6.11	5	7.93	2.81	1	5	7	27	2

Konut fiyatları ortalama 3.6 milyon TL civarında, medyan fiyat ise 3.15 milyon TL, yani ortalamanın altındadır. Bu fiyatların sağa çarpık olduğunu, yani bazı çok yüksek fiyatlı evlerin ortalamayı yukarı çektiğini gösteriyor.

Brüt ve net metrekarelerde de benzer durum söz konusudur. Geniş bir çeşitlilik ve büyük farklar bulunmaktadır. Evlerin büyüklükleri değişkendir.

Banyo sayısı genelde 1 ile sınırlı, fazla değil. Bina yaşı ise çoğunlukla sıfır, yani konutların çoğunluğu yeni binalardan oluşmaktadır.

Kat bilgilerine bakınca, evlerin büyük çoğunluğu 1 ile 4. katlar arasında ve binalar genellikle 5-7 katlı. Tabii ki bodrum katlar ve daha yüksek katlar da mevcuttur.

Oda sayısı genelde 3 civarında, ama az da olsa büyük ve geniş daireler de bulunmaktadır.

4.1.2. Aykırı Değer Tespiti

Verideki aykırı değerler iki yaygın yöntemle incelenmiştir.

4.1.2.1. Çeyrekler Arası Mesafe (IQR) Yöntemi

Aşağıdaki sayısal değişkenlerde IQR yöntemiyle aykırı değer sayıları tespit edilmiştir ve veri setinden çıkarılmıştır.

Tablo 4.2. IQR ile Çıkarılan Aykırı Değerler

Değişken	Aykırı Say.
Fiyat (TL)	54
Brüt m²	1
Net m²	0
Banyo Sayısı	0
Bina Yaşı	111
Bul. Kat	123
Oda Sayısı	0
Kat Sayısı	206

4.1.2.2. Z-Skoru Yöntemi

Ortalama ve standart sapmaya göre hesaplanan Z-puanları 3'ün üzerinde olan değerler aykırı kabul edildi ve bu aykırı değerlere ait satırlar veriden çıkarıldı.

Tablo 4.3. Z-Skoru ile Çıkarılan Aykırı Değerler

Değişken	Aykırı Say.
Fiyat (TL)	16
Brüt m²	11
Net m²	13
Banyo Sayısı	33
Bina Yaşı	46
Bul. Kat	37
Oda Sayısı	10
Kat Sayısı	73

4.1.2.3. Aykırı Değerlerin Veriye Olan Etkisinin İncelenmesi

Aykırı değerler çıkarıldıktan sonra hesaplanan tanımlayıcı istatistiklerde şu değişiklikler gözlenmiştir.

Tablo 4.4. Aykırı Değerlerden Elenmiş Verinin Tanımlayıcı İstatistikleri

Değişken	Ort. (O)	Ort. (T)	Std. Sap. (O)	Std. Sap. (T)
Fiyat (TL)	3,645,573	3,406,928	1,740,903	1,526,834
Brüt m²	105.85	97.42	50.46	46.43
Net m²	93.18	85.94	45.00	42.01
Banyo Sayısı	1.40	1.37	0.51	0.50
Bina Yaşı	4.05	2.91	6.40	4.77
Bul. Kat	3.15	2.73	2.63	1.90
Oda Sayısı	3.04	2.88	1.02	0.97
Kat Sayısı	6.11	5.33	2.81	1.50

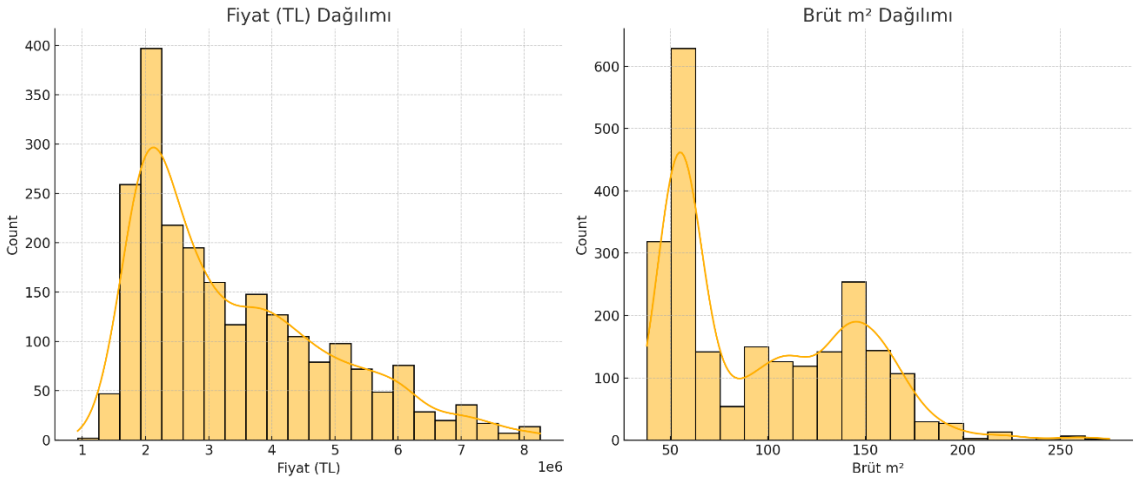
Veri setinde yapılan temizlik sonrasında ortalama ve standart sapma değerlerinde azalmalar meydana gelmiştir. Veri setine 2286 veri kalmıştır ve bu veri üzerinden istatistiksel analizler yapılmıştır. Özellikle fiyat, metrekare ve kat sayısı gibi değişkenlerdeki düşüşlerden azalan uç değerlerin etkisi azalmaktadır.

4.1.3. Logaritmik Dönüşüm Sonuçları

- "Fiyat (TL)", "Brüt m²" ve "Net m²" değişkenlerine logaritmik dönüşüm uygulanarak orijinal veri ile tanımlayıcı istatistikler karşılaştırılmıştır.
- Log dönüşümü sonrası verilerin dağılımı daha simetrik ve normal dağılıma yakın hale gelmiştir.
- Bu dönüşüm, özellikle pozitif çarpık dağılımlar için analizlerin varsayımlarını destekler.

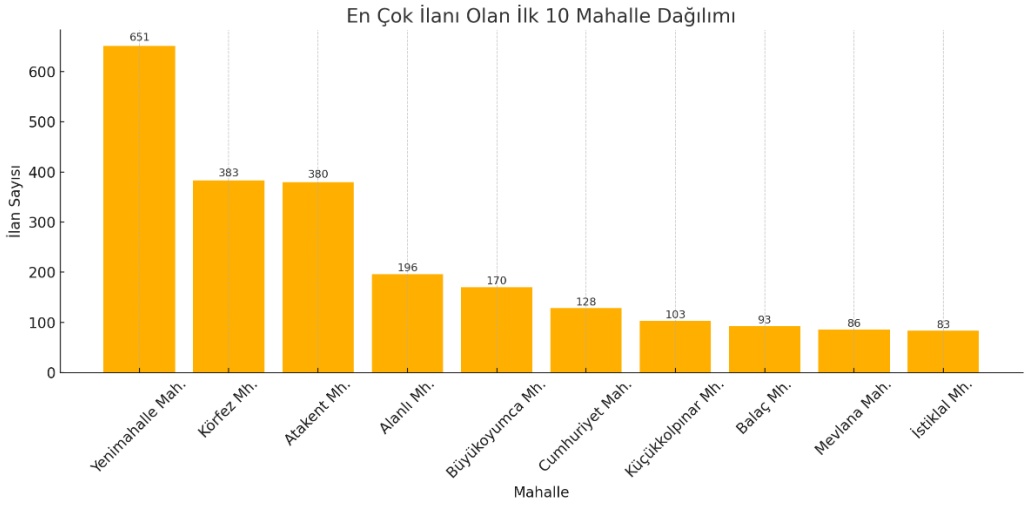
4.1.4. Görsel Analizler

Orijinal veride çok sayıda uç değer (outlier) gözlemlenirken, temizlenmiş veri setinde bu uç değerlerin büyük kısmı yok olmuştur.



Şekil 4.1. Konutların Fiyat ve Brüt m^2 dağılımı

Aşağıdaki görselde ilk 10 mahalleden çekilen verilerin dağılımı bulunmaktadır.



Şekil 4.2. Mahallelerin Dağılımı

4.2. Başlıca Gruplar Arasında Bağımsız Örneklem t-Testi Bulguları

Veri setindeki başlıca kategorik değişkenler için sayısal değişkenlerin ortalamaları arasında anlamlı farklar olup olmadığını belirlemek amacıyla bağımsız örneklem t-testi uygulanmıştır.

4.2.1. Hipotezler ve Gruplar

Her bir sayısal değişken için test edilen hipotezler aşağıdaki gibidir:

- **Yokluk Hipotezi (H0):**
  Takas yapan ve yapmayan (ya da site içerisinde olan ve olmayan) grupların ilgili sayısal değişken ortalamaları arasında fark yoktur. Yani, iki grubun ortalamaları eşittir.
- **Alternatif Hipotez (H1):**
  Takas yapan ve yapmayan (ya da site içerisinde olan ve olmayan) grupların ilgili sayısal değişken ortalamaları arasında anlamlı bir fark vardır. Yani, iki grubun ortalamaları eşit değildir.

İncelenen gruplar ve sayısal değişkenler:

- **Gruplar:** Takas durumu (Evet / Hayır), Site içerisinde olma durumu (Evet / Hayır).
- **Sayısal Değişkenler:** Fiyat (TL), Brüt m², Net m², Bina Yaşı.

Bu hipotezler, genel istatistiksel anlamlılık düzeyi ($\alpha = 0.05$) doğrultusunda test edilmiştir.

p-değeri 0.05'ten küçük olan durumlarda yokluk hipotezi reddedilerek gruplar arasında anlamlı fark olduğu sonucuna varılmıştır.

4.2.2. Takas Durumu Grubu

- **Fiyat (TL):** Takas yapan ve yapmayan konut ilanları arasında ortalama fiyat açısından istatistiksel olarak anlamlı bir fark bulunmamıştır (p=0.53).
- **Brüt m² ve Net m²:** Takas yapan evlerin ortalama alanları (Brüt m² = 94.94, Net m² = 83.53) yapmayanlara göre anlamlı derecede daha düşüktür (sırasıyla p=0.0386 ve p=0.0260).
- **Bina Yaşı:** Takas yapan konut ilanların ortalama bina yaşı (1.83 yıl) takas yapmayanlara (3.65 yıl) göre istatistiksel olarak anlamlı şekilde daha yeni çıkmıştır ($p < 0.0001$).

4.2.3. Site İçerisinde Olma Durumu Grubu

- **Fiyat (TL):** Site içerisinde bulunan evlerin ortalama fiyatı (4.220.910 TL), site dışında kalan evlere (3.347.583 TL) göre anlamlı derecede yüksek çıkmıştır. ($p < 0.0001$).
- **Brüt m² ve Net m²:** Site içindeki evlerin ortalama alanları (Brüt m² = 117.63, Net m² = 104.95) site dışındaki evlere kıyasla anlamlı derecede büyük çıkmıştır. ($p < 0.0001$).
- **Bina Yaşı:** Site içindeki evlerin ortalama bina yaşı (4.98 yıl), site dışındaki evlere (2.77 yıl) göre anlamlı şekilde daha eski çıkmıştır. ($p < 0.0001$).

4.2.4. Bağımsız Örneklem t-Testi Sonuçları

Site içerisinde yer alan konutların fiyat ve alan bakımından daha yüksek değerlere sahip olduğu ancak genellikle daha eski yapılar olduğunu yorumlayabilir, buna karşılık takas yapılan evlerin daha yeni ve küçük metrekareli olduğu görülmüş ancak fiyat açısından anlamlı bir fark saptanmadığı söylenebilir.

4.3. Korelasyon Analizi Bulguları

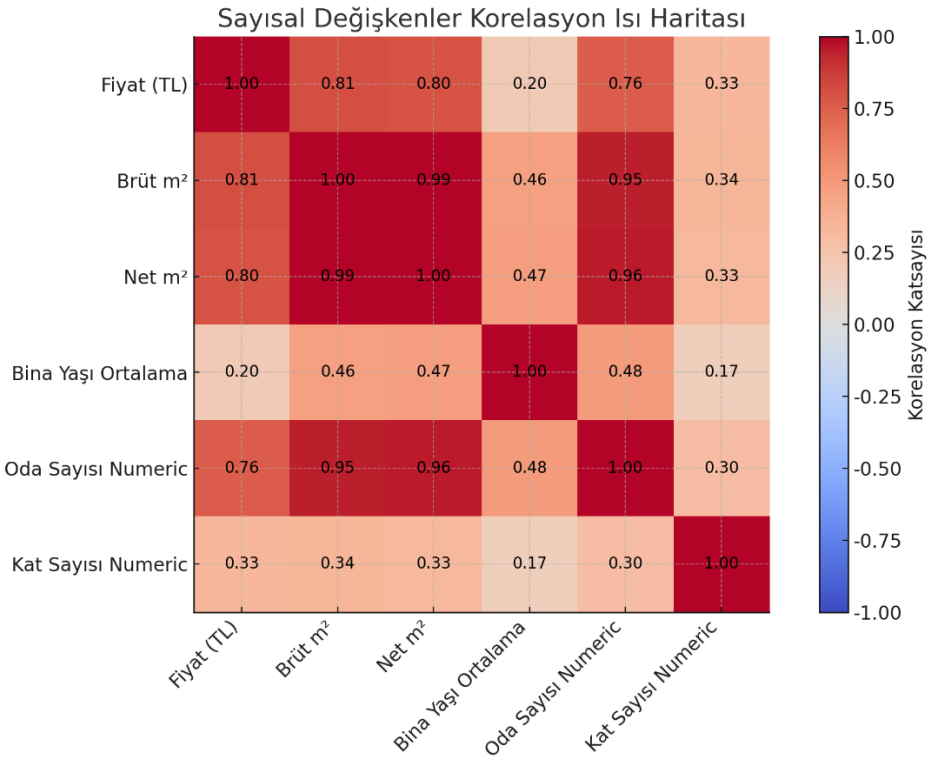
Analiz kapsamında; Fiyat (TL), Brüt m², Net m², Bina Yaşı Ortalama, Oda Sayısı ve Kat Sayısı değişkenleri arasında Pearson korelasyon katsayıları hesaplanmıştır. Elde edilen korelasyon katsayıları aşağıdaki gibidir:

- Fiyat (TL) ile Brüt m² arasında yüksek pozitif korelasyon ($r \approx 0.82$) bulunmuştur.
- Fiyat (TL) ile Net m² arasında da güçlü pozitif korelasyon ($r \approx 0.80$) gözlemlenmiştir.
- Brüt m² ile Net m² arasındaki korelasyon en yüksek seviyede
- ($r \approx 0.99$) olup, iki değişkenin benzer trendleri paylaştığı anlaşılmıştır.
- Bina Yaşı ile Fiyat (TL) arasında negatif korelasyon ($r \approx -0.26$) tespit edilmiştir; yani bina yaşı arttıkça fiyatın düşme eğiliminde olduğu görülmüştür.
- Oda Sayısı ve Kat Sayısı da diğer büyüklük değişkenleriyle pozitif ilişki göstermektedir.

4.3.2. Korelasyon Isı Haritası

Sayısal değişkenler arasındaki korelasyon matrisi, renk skalasıyla ifade edilen ısı haritası ile görselleştirilmiştir. Bu görsel, değişkenler arasındaki ilişkilerin gücünü ve yönünü hızlıca değerlendirmeye olanak sağlar.

Skaladaki koyu kırmızı renkler güçlü pozitif, skaladaki koyu mavi renkler ise negatif korelasyonları temsil etmektedir.



Şekil 4.3. Korelasyon Isı Haritası

Bina Yaşı Ortalama: Bina yaşı olarak net bir değer yerine aralık verilen değerlerin ortalaması ile doldurulan yeni değişkendir. Örneğin konutun bulunduğu binanın yaşı 20-25 arasındaysa, bu binanın yaşı 20 ile 25'in ortalaması olan 22.5 ile doldurulur.

Oda Sayısı Numeric: Oda sayılarının sayısallaştırılmış halidir. Örneğin; 1+1 odalar 2, 3+1 odalar 4 oda olarak baz alınmıştır.

Kat Sayısı Numeric: Bu değişken, Bodrum ve çatı katında bulunan konutlar için düzenlenmiştir.

4.4. Ki-Kare Testi Bulguları

Veri setindeki başlıca kategorik değişkenler arasındaki ilişkiyi incelemek için Ki-kare testi uygulanmış ve gözlenen frekanslar ile beklenen frekanslar arasındaki farkların anlamlı olup olmadığı belirlenmiştir. Elde edilen önemli sonuçlar aşağıda detaylandırılmıştır:

4.4.1. Krediye Uygunluk ve Eşyalı Olma Durumu

Krediye uygunluk ile eşyalı olma durumu arasında güçlü bir ilişki bulunmuştur ($\chi^2 = 319.61$, $p < 0.00000001$). Konutun krediye uygun olup olmaması, eşyalı olup olmamasını etkilemektedir.

Tablo 4.5. Krediye Uygunluk ve Eşya Durumu İçin Çapraz Tablo

Krediye Uygun	Belirtilmemiş	Evet	Hayır
Belirtilmemiş	27	4	36
Evet	44	335	1642
Hayır	1	18	168

4.4.2. Takas Durumu ve Kullanım Durumu (Boş/Kiracılı/Mülk Sahibi)

Takas yapan ve yapmayan gruplar arasında kullanım durumu bakımından anlamlı bir farklılık tespit edilmiştir ($\chi^2 = 32.84$, $p < 0.0000001$). Takas işlemi yapılan konutların kullanım durumu (boş veya kiracılı) farklılık yaratmaktadır.

Tablo 4.6. Takas ve Kullanım Durumu İçin Çapraz Tablo

Takas	Boş	Kiracılı	Mülk Sahibi
Evet	781	51	87
Hayır	1042	169	149

4.4.3. Takas Durumu ve Eşyalı Olma Durumu

Takas durumu ile eşyalı olma durumu arasında güçlü bir ilişki mevcuttur ($\chi^2 = 16.73$, $p = 0.000233$). Takas işlemi yapılan konutların eşyalı ya da eşyasız olması arasında anlamlı bir fark bulunmuştur.

Tablo 4.7. Takas ve Eşyalı Olma Durumu İçin Çapraz Tablo

Takas	Belirtilmemiş	Evet	Hayır
Evet	35	110	770
Hayır	37	246	1075

4.4.4. Kullanım Durumu ve Site İçerisinde Olma Durumu

Site içerisinde bulunan ve bulunmayan konutların kullanım durumu arasında anlamlı bir ilişki belirlenmiştir ($\chi^2 = 27.64$, $p = 0.00000099$). Konutun site içerisinde olup olmaması, kullanım durumunu doğrudan etkilemektedir.

Tablo 4.8. Kullanım Durumu ve Site İçerisinde Olma Durumu İçin Çapraz Tablo

Kullanım Durumu	Evet	Hayır
Boş	99	1726
Kiracılı	25	195
Mülk Sahibi	31	205

4.4.5. Beklenen Frekansların Durumu ve Testin Güvenilirliği

Testin geçerliliği açısından önemli olan beklenen frekans değerleri incelenmiş, her hücrede beklenen frekansların çoğunlukla 5'in üzerinde olduğu görülmüştür. Bu durum, Ki-kare test sonuçlarının güvenilir olduğunu ve yanlış yorumlanma riskinin düşük olduğunu göstermektedir

4.5. Olasılık Dağılımları ve Güven Aralıkları bulguları

- Fiyat (TL), Brüt m² ve Net m² verileri Shapiro-Wilk testine göre normal dağılmamakta ($p << 0.05$).
- Logaritmik dönüşüm sonrası veriler normal dağılıma daha yakın hale gelmiştir.
- Fiyat için %95 güven aralığı yaklaşık [3,344,270 TL, 3,469,586 TL] olarak hesaplanmıştır.

Veri setindeki pozitif çarpıklık, logaritmik dönüşüm ile azaltılarak parametrik analizlerin geçerliliği artırılmıştır.

4.6. Normallik Testleri Sonuçları

4.6.1. Shapiro-Wilk Test İstatistiği ve p-değerleri

Tablo 4.9. Normal Dağılıma Uygunluk

Değişken	İstatistik (W)	p-değeri	Normal Dağılım Durumu
Fiyat (TL)	0.908	4.0×10^{-35}	Normal dağılıma uymuyor
Brüt m²	0.880	9.3×10^{-39}	Normal dağılıma uymuyor
Net m²	0.876	2.9×10^{-39}	Normal dağılıma uymuyor
Bina Yaşı Ortalama	0.664	0	Normal dağılıma uymuyor

Başlıca sayısal parametrikler normal dağılmamaktadır.

4.7. Mahalleler arası farklar için ANOVA analizi

Yokluk Hipotezi, (H0): Mahalleler arasında gayrimenkul fiyatları açısından anlamlı bir fark yoktur.

Alternatif Hipotez (H1): Mahalleler arasında gayrimenkul fiyatları açısından anlamlı bir fark vardır.

ANOVA testinden elde edilen sonuçlar şu şekildedir:

- **F-istatistiği (F-statistic)**: 28.33
- **p-değeri (p-value)**: 1.39×10−1291.39

Bu test sonucuna göre, p-değeri çok küçük (0.05'ten çok daha küçük), bu da **Yokluk Hipotez'in reddedilmesi** gerektiğini gösterir. Yani, mahalleler arasında **fiyatlar açısından anlamlı farklar bulunmaktadır**.

Tablo 4.10. ANOVA Tablosu

Kaynak	Sum of Squares	df	Mean Square	F-değeri	p-değeri
Gruplar Arası	63,563.99	29	28.33	28.33	1.39 x 10^-129
Gruplar İçinde	3.83 x 10^15	2244	1.71 x 10^12	-	-
Toplam	2.24 x 10^3	2273	-	-	-

Not:

- **Gruplar Arası**: Mahalleler arasındaki farkların toplamı.
- **Gruplar İçinde**: Her mahalle içindeki varyans.

4.7.1. Tukey HSD ile mahalle çiftleri arası farkın analizi

Yokluk Hipotez, (H0): Mahalleler arasında gayrimenkul fiyatlarında anlamlı bir fark yoktur.

Alternatif Hipotez (H1): Mahalleler arasında gayrimenkul fiyatlarında anlamlı bir fark vardır.

Tukey HSD testi, ANOVA sonuçlarından sonra hangi mahalleler arasındaki farkların anlamlı olduğunu belirlemek için kullanılmıştır. Bu test, tüm mahalle çiftlerini karşılaştırarak anlamlı farkları ortaya koymaktadır.

Nüfus açısından ve konut ilanı bakımından önde gelen 3 ana mahalle Tablo 4.11'de karşılaştırılmıştır.

Tablo 4.11. Tukey HSD testi ile Önde Gelen 3 Mahallenin Karşılaştırılması

Mahalle Çifti	Fiyat Farkı (TL)	p-değeri	Anlamlı Fark
Yeni Mahalle - Körfez Mh.	1,225,825.52	0.000	Evet
Yeni Mahalle - Atakent Mh.	-1,044,089.77	0.000	Evet
Körfez Mh. - Atakent Mh.	-2,034,625.84	0.000	Evet

4.8. MANOVA Analizi (Wilk's Lambda)

Çoklu bağımlı değişkenler (**Fiyat (TL), Brüt m², Net m²**) ve bağımsız kategorik değişkenler (**Krediye Uygunluk, Takas Durumu, Kullanım Durumu, Site İçerisinde Olma**) üzerinden uygulanan MANOVA testi ile değişkenlerin çoklu değişkenler üzerindeki etkisi incelenmiştir.

Tablo 4.12. MANOVA Analizi Bulguları

Değişken	Wilks' Lambda	F Değeri	p-Değeri	Anlamlı mı?
Krediye Uygunluk	0.9974	2.44	0.0626	Hayır
Takas Durumu	0.9952	4.55	0.0035	Evet
Kullanım Durumu	0.9331	67.57	<0.001	Evet
Site İçerisinde	0.9681	31.11	<0.001	Evet

- $p < 0.05$olan değişkenler (**Takas Durumu, Kullanım Durumu, Site İçerisinde**) çoklu bağımlı değişkenlerde anlamlı etkiler gösteriyor.
- Örneğin, **Wilks' Lambda 0.9331** olan Kullanım Durumu, bağımlı değişkenlerde %6.69'luk (1-0.9331) anlamlı varyans açıklıyor.

4.9. İstatistiksel Genel Değerlendirme

Verilerdeki pozitif çarpıklık nedeniyle normal dağılım varsayımını sağlamamaktadır. Oda sayısı, mahalle, kredi uygunluğu gibi kategorik değişkenler fiyat ve fiziksel özellikler üzerinde anlamlı farklılıklar ve ilişkiler gösteriyor.

MANOVA analizi ile çoklu bağımlı değişkenlerin kategorik gruplar tarafından etkilenmiş olduğu ortaya koymaktadır, Wilks' Lambda istatistiği kullanılarak grupların çoklu değişkenler üzerindeki etkisi incelenmiştir. En son sonuçlarda çapraz karşılaştırmalarla fark anlaşılacaktır.

Ki-kare testleri, kategorik değişkenler arasındaki bağıntıları ortaya koymuştur.

Tüm bu analizler birlikte değerlendirilerek, konut piyasası için güvenilir, çok boyutlu modeller geliştirilebilir.

4.10. Makine Öğrenmesi Bulguları

Veride hedef değişken olan **Fiyat (TL)** bilindiğinden, **Danışmanlı (Gözetimli) Makine Öğrenmesi** kullanılmıştır.

Bu yöntem geçmiş verilerden öğrenerek yeni veriler için fiyat tahminleri yapmaya uygundur.

Seçilen modeller:

- Basit regresyon
- Çoklu Regresyon
- Üstel Regresyon
- Polinomik Regresyon
- KNN Regresyon
- Stepwise, remove, geriye ve ileriye regresyon
- Yapay Sinir Ağları
- Boosting modelleri
- Random Forest Regresyon

Konut fiyatlarını etkileyen temel değişkenleri belirleyip doğru tahmin modelleri geliştirilmiştir.

4.10.1. Modelleme ve Performans Değerlendirmeleri

4.10.1.1. Basit ve Çoklu Doğrusal Regresyon

- Basit modelde sadece **Brüt m²** kullanılmış, model başarı göstergesi;

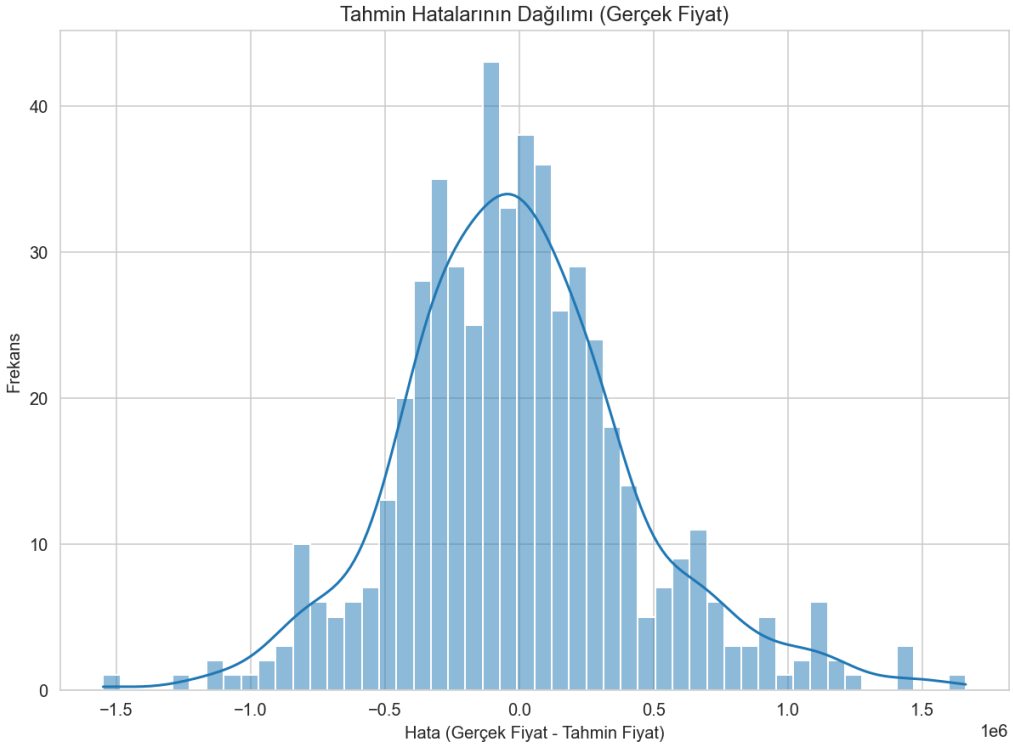
  R^2 = 0.67 olarak hesaplanmıştır.

$$Fiyat = \mathbf{817,483.55} + \mathbf{26,579.98} \times Brut\ m^2$$

- Çoklu modelde 5 bağımsız sayısal değişken kullanılmış ve başarısı;

  R^2 = 0.70'e yükselmiştir.

$$\begin{aligned}\text{Fiyat} = 333{,}845.89 &+ 21{,}513.52 \times \text{Brut m}^2 + 7{,}222.72 \times \text{Net m}^2 \\ &- 75{,}649.30 \times \text{Oda Sayısı} + 68{,}979.53 \times \text{Kat Sayısı} \\ &- 61{,}571.84 \times \text{Bina Yası} + 281{,}956.91 \times \text{Banyo Sayısı}\end{aligned}$$



Şekil 4.4. Tahmin Hatalarının Dağılımı

4.10.1.2. Stepwise Regresyon

- İleri seçim ve geriye eleme yöntemleri ile değişken seçimi yapılmıştır.
- Geriye eleme yöntemiyle seçilen modelde üç değişken (Brüt m², Bina Yaşı Ortalama, Kat Sayısı) yer almakta ve model R^2= 0.70 ile en iyi performansı sağlamaktadır.

4.10.1.3. Performans Ölçütleri

Tablo 4.13. Regresyon Performans Ölçütleri

Model	MSE (TL²)	RMSE (TL)	MAE (TL)	R²
Basit Doğrusal Regresyon	7.7×10^{11}	877,387	617,190	0.67
Çoklu Doğrusal Regresyon	6.94×10^{11}	833,119	593,114	0.70
İleri Seçim	7.00×10^{11}	836,685	599,094	0.70
Geriye Eleme	6.93×10^{11}	832,759	593,177	0.70
Anlamlı Değişkenler	6.93×10^{11}	832,759	593,177	0.70

4.10.1.4. Katsayıların Anlamlılığı

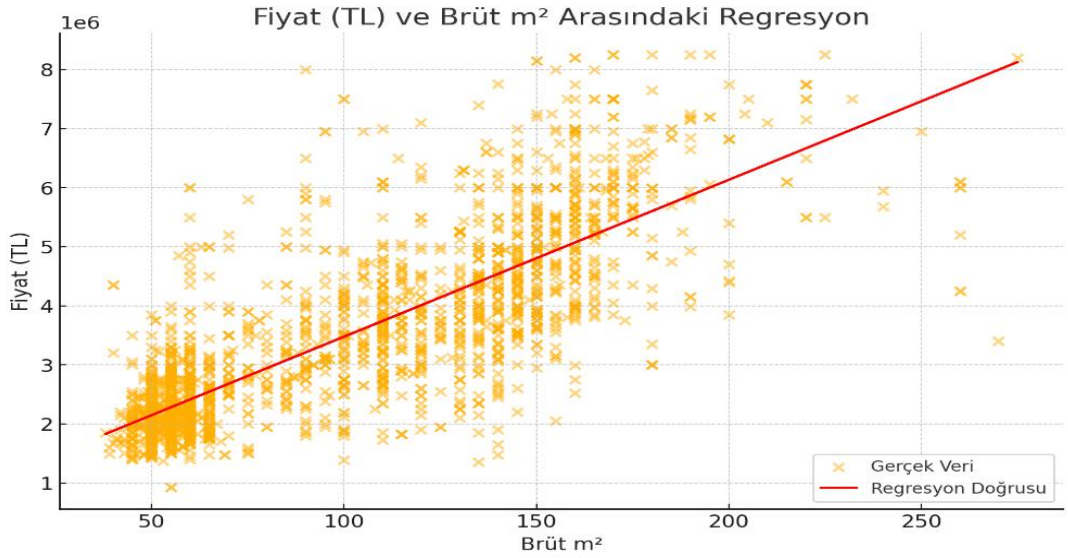
- **Brüt m²**, **Bina Yaşı Ortalama** ve **Kat Sayısı Numeric** değişkenleri **anlamlı** bulunmuştur, çünkü her birinin p-değeri 0.001'den küçük ($p < 0.001$).

- **Net m²** değişkeni de **anlamlı** bulunmuş olsa da, p-değeri 0.033 ile 0.001'den büyük olup 0.05'e yakın bir değere sahiptir, yani bu değişkenin anlamlılığı, daha katı bir eşik olan 0.001'lik sınırı aşmaktadır.

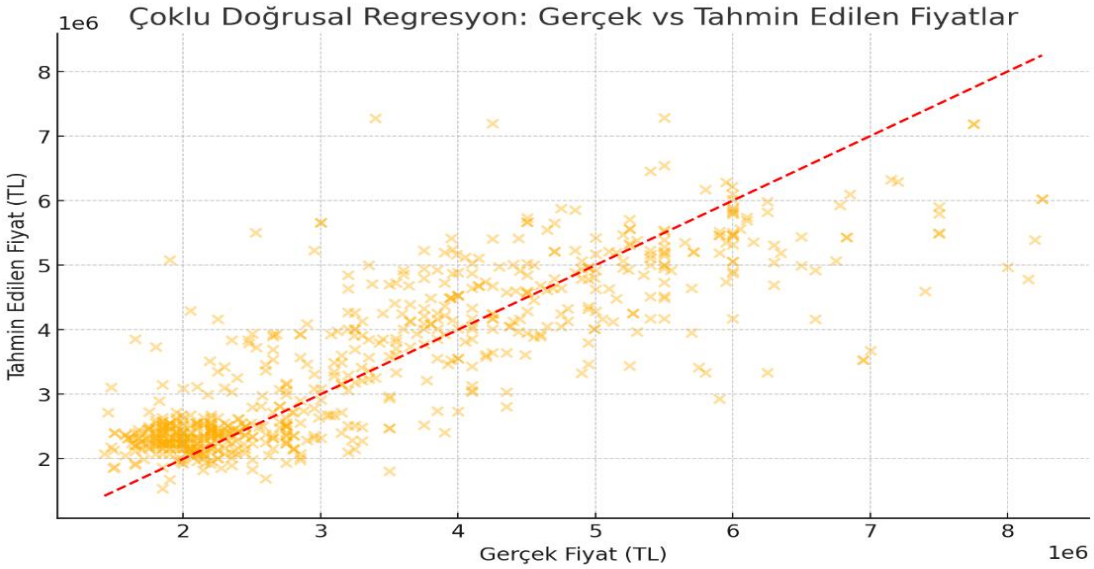
- **Oda Sayısı Numeric** değişkeninin p-değeri 0.606 olup, bu değişken **anlamlı** bulunmamıştır ($p > 0.05$).

4.10.1.5. Doğrusal Çoklu Regresyon İçin Görsel Bulgular

Gerçek ve tahmin edilen fiyatlar scatter plot ile gösterilmiş ve modelin genel olarak iyi uyum sağladığı görülmüştür.



Şekil 4.5. Fiyat ve Brüt m^2 Arası Basit Doğrusal Regresyon



Şekil 4.6. Fiyat ve Sayısal Değişkenler Arası Çoklu Doğrusal Regresyon

4.10.1.6. Anlamlı Değişkenler ile Çoklu Regresyon

Regresyon denklemi şu şekildedir:

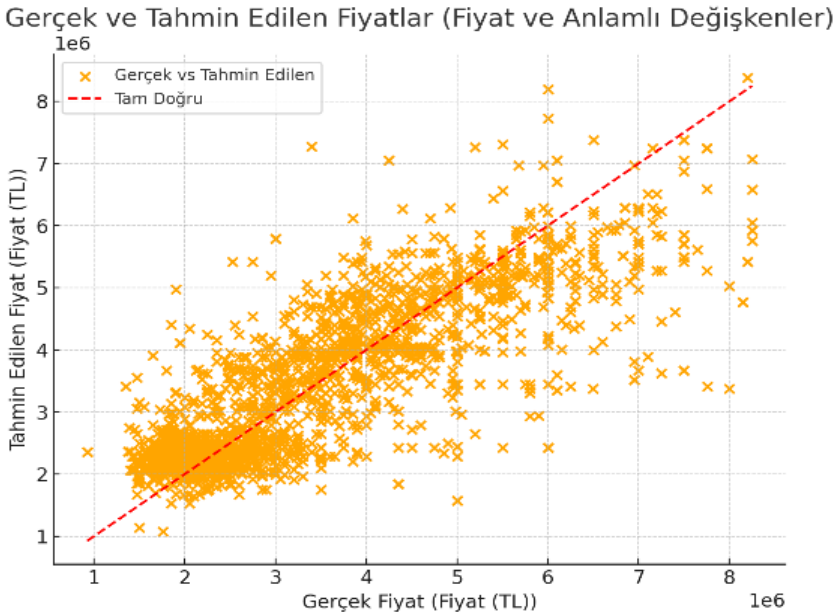
Fiyat (TL) = 385,763.93 + 29,154.07 × Brüt m² - 70,940.27 × Bina Yaşı + 72,823.78 × Kat Sayısı

Modelin Katsayıları:

- **Brüt m²** (Brüt Alan): Her bir ekstra metrekare için fiyat, yaklaşık **29,154.07 TL** artmaktadır.
- **Bina Yaşı** (Bina yaşı ortalama): Bina yaşının her ek yılı, fiyatı **70,940.27 TL** kadar düşürmektedir.
- **Kat Sayısı**: Her bir ek kat, fiyatı **72,823.78 TL** kadar artırmaktadır.

Çoklu doğrusal regresyon modelinde ise üç bağımsız sayısal değişken (**Brüt m²**, **Bina Yaşı**, **Kat Sayısı**) kullanılmıştır. Bu modelin başarısı **R^2 = 0.70** olarak hesaplanmıştır. Bu da, modelin önemli ölçüde geliştirilmiş olduğunu ve **Fiyat (TL)**'yi tahmin etme noktasında daha iyi bir performans sergilediğini gösterir.

Bu modelde **Brüt m²**'nin yanı sıra **Bina Yaşı** ve **Kat Sayısı** gibi faktörlerin de fiyat üzerinde etkili olduğu anlaşılmaktadır.



Şekil 4.7. Sadece Anlamlı Değişkenlerle Yapılan Çoklu Doğrusal Regresyon

4.10.1.7. Çoklu Bağlantı ve Model Değerlendirme

4.10.1.7.1. Çoklu Bağlantı

- Korelasyon matrisine göre, Brüt m² ile Net m² arasında yüksek korelasyon (r≈0.99) tespit edilmiştir.
- VIF değerleri, Brüt m² (410.1), Net m² (398.5), Oda Sayısı Numeric (59.6) değişkenlerinde çok yüksek çıkmış ve ciddi çoklu bağlantı sorunu işaret etmiştir.

4.10.1.7.2. Düzeltilmiş R² ve F-Testi

- Modelin düzeltilmiş R^2 değeri yaklaşık 0.69 olarak bulunmuş, F-testi sonucu ise F=720.44 ile anlamlı değişkenlerden oluşan model anlamlıdır ($p < 0.001$).

4.10.1.7.3. Katsayı Anlamlılıkları

- Anlamlı katsayılar: Brüt m², Bina Yaşı, Kat Sayısı.
- Anlamlı olmayanlar: Net m², Oda Sayısı.

4.10.1.7.4 Regresyon İçin Genel Değerlendirme

- Çoklu bağlantı nedeniyle modelde bazı değişkenlerin etkileri güvenilir değildir.
- Model sadeleştirilerek anlamlı değişkenlerle yüksek başarı ($R^2 \approx 0.70$) elde edilmiştir.
- Modeller, piyasa analizi ve fiyat tahmininde sağlam temel sunmaktadır.

4.10.1.2. Polinomik Regresyon Bulguları

Aykırı değerlerden arındırılmış veri setinde, Brüt m² değişkeni kullanılarak 1., 2. ve 3. derece polinomik regresyon modelleri oluşturuldu ve ödellerin test setindeki performansları aşağıdaki metriklerle değerlendirildi:

- Ortalama Kare Hata (MSE)
- Kök Ortalama Kare Hata (RMSE)
- Ortalama Mutlak Hata (MAE)
- Belirleme Katsayısı (R^2)

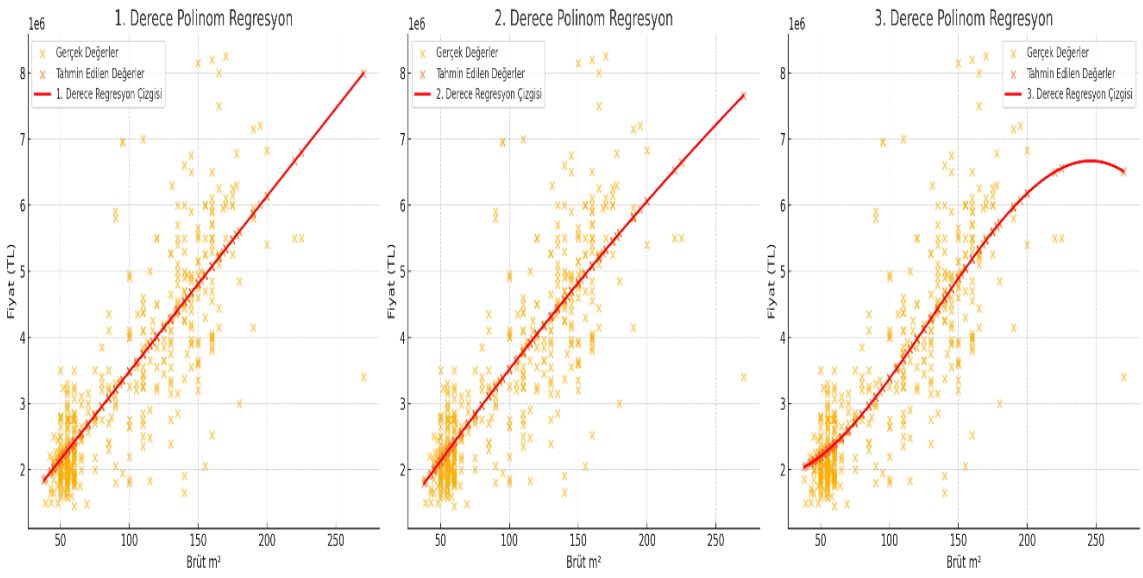
4.10.1.2.1. Polinomik Regresyonun Derecelere Göre Karşılaştırılması

Tablo 4.14. Polinom Derecelerine Göre Model Performans Ölçütleri

Polinom Derecesi	MSE (TL²)	RMSE (TL)	MAE (TL)	R^2
Doğrusal	5.90×10^{11}	768,320	550,987	0.739
2	5.91×10^{11}	768,711	551,284	0.738
3	5.88×10^{11}	766,706	549,872	0.740

Polinom derecesinin artmasıyla birlikte model performansında küçük iyileşmeler görülmekte ancak model karmaşıklığı ve çoklu bağlantı da artmaktadır. Derece arttıkça iyileşme marjinal kalmakta ve modelin aşırı uyum riski yükselmektedir.

4.10.1.2.2. Polinomik Regresyon İçin Görsel Değerlendirme



Şekil 4.8. Derecelerine Göre Polinomik Regresyon Grafikleri

Her polinom derecesi için gerçek ve tahmin edilen fiyatların scatter plot grafikleri oluşturulmuştur. 3. derece polinomik regresyon modelinin tahmin edilen fiyatlarla gerçek fiyatlar arasındaki uyumu diğerlerine göre daha yüksektir.

4.10.1.3. Üstel Regresyon ve Düzenlileştirme Teknikleri Bulguları

4.10.1.3.1. Üstel Regresyon Bulguları

Üstel regresyon, bağımlı değişkenin bağımsız değişkene üstel bir fonksiyon olarak bağlı olduğu durumlarda kullanılır. Model şu şekilde ifade edilir:

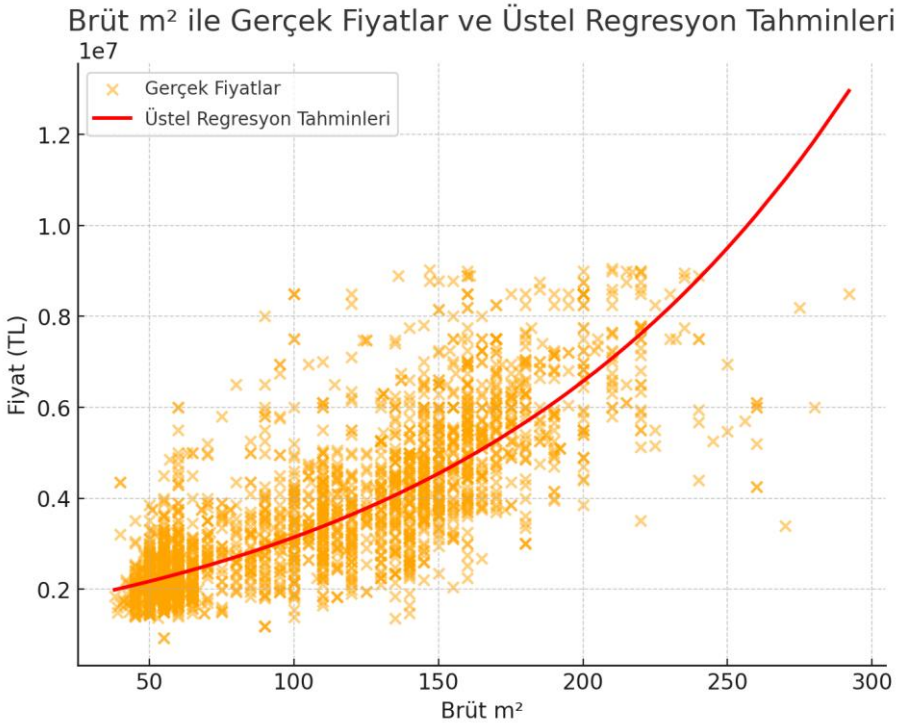
Burada, bağımlı değişkenin logaritması alınarak doğrusal regresyon modeli uygulanır. Çalışmamızda konut fiyatları için Brüt m² ile log dönüşümü yapılmış ve model kurulmuştur.

4.10.1.3.1.1. Üstel Regresyon Performans sonuçları

Model, fiyat tahminlerinde makul doğruluk sağlamış ancak diğer düzenlileştirme yöntemlerine göre performansı biraz daha düşüktür.

Tablo 4.15. Üstel Regresyon Performans Bulguları

Ölçüt	Değer
R^2	0.669
RMSE	1,087,652 TL
MAE	738,619.53 TL



Şekil 4.9. Üstel Regresyon Grafiği

4.10.1.3.2. Düzenlileştirme (Regularization) Yöntemleri

4.10.1.3.2.1. Düzenleştirme Sonrası Performans Sonuçları

Düzenlileştirme yöntemleri, çoklu bağlantı ve aşırı öğrenme problemlerini azaltarak daha sağlam ve güvenilir modeller ortaya koymuştur.

Tablo 4.16. Düzenleştirme Sonrası Performans Sonuçları

Model	R^2	RMSE (TL)	MAE (TL)
Ridge	0.739	768,356	551,137
Lasso	0.739	768,320	550,987
Elastic Net	0.739	768,602	552,073

4.10.1.4. K-EN Yakın Komşu (K-NN) Regresyonu

4.10.1.4.1. K-NN Regresyon Model Uygulaması

- Veri setindeki sayısal özellikler (Brüt m², Net m², Bina Yaşı Ortalama, Oda Sayısı Numeric, Kat Sayısı Numeric) standartlaştırılmıştır.
- Eğitim ve test setleri ayrılmıştır (test oranı %30).
- En iyi komşu sayısı kkk, 1 ile 20 arasında çapraz doğrulama yöntemiyle belirlenmiştir.
- En iyi k seçildikten sonra model eğitilmiş ve test verisi üzerinde performansı ölçülmüştür.

4.10.1.4.1.2. Model Performansı

Tablo 4.17. K-NN Regresyon Bulguları

Performans Ölçütü	Değer
En İyi K	13
MSE	4.12×10^{11}
RMSE	641,671 TL
MAE	449,320 TL
R^2	0.80

- K-NN regresyonu, doğrusal olmayan ilişki yapısı nedeniyle veri setimizde güçlü bir performans göstermiştir ($R^2 = 0.80$), doğrusal modellerden daha iyi sonuç vermiştir.
- Özellikle konut fiyatlarındaki karmaşık yapıların modellenmesinde K-NN etkili bir yöntemdir.
- Modelin en uygun K değeri 13 olarak belirlenmiş ve bu değer modelin genelleme başarısını artırmıştır.
- Özelliklerin ölçeklendirilmesi ve parametre optimizasyonu model performansı için kritiktir.
- Ancak, K-NN tahmin süresi veri büyüdükçe artar ve büyük veri setlerinde hesaplama maliyeti yükselir.

4.10.1.5. Random Forest Regresyonu

4.10.1.5.1. RF'de Sayısal ve Kategorik Özelliklerin Kullanımı

Random Forest algoritması temel olarak sayısal girdilerle çalıştığı için veride bulunan kategorik değişkenlerin modele uygun hale getirilmesi gerekir. Bu çalışma kapsamında;

- **Sayısal özellikler** veri setinde mevcut haliyle kullanılmıştır (örn. metrekare, bina yaşı, kat sayısı vb.).
- **Kategorik özellikler** ise modelin anlaması için **One-Hot Encoding** yöntemiyle sayısal vektörlere dönüştürüldü ve böylece her kategori birer ikili (binary) sütuna ayrılarak modelin tüm kategorik bilgiyi öğrenebilmesi sağlanmıştır.

Bu yöntem sayesinde model, sayısal ve kategorik değişkenlerin tamamından faydalanarak tahmin yapabilmektedir.

4.10.1.5.2. RF Model Eğitimi ve Performans Değerlendirmesi

Model, aykırı değerlerden arındırılmış veri üzerinde %80 eğitim, %20 test seti oranında uygulanmıştır. Random Forest regresyon modeli, 100 ağaçla eğitilmiş ve test setinde aşağıdaki performans metrikleri elde edilmiştir.

Tablo 4.18. Random Forest Model Performansı

Performans Ölçütü	Değer
Ortalama Kare Hata (MSE)	352,372,621,879
Ortalama Mutlak Hata (MAE)	394,990
Determinatör Katsayısı (R^2)	0.84

Modelin test verisindeki fiyat varyansının %84'ünü açıklayabilmektedir.

4.10.1.5.3. RF'de Özelliklerin Önemi ve Model Yorumlanabilirliği

Model tarafından belirlenen en önemli özellikler şu şekildedir:

- Net metrekare
- Brüt metrekare
- Banyo sayısı
- Bulunduğu kat
- Mahalle

4.10.1.6. Boosting Algoritmaları

4.10.1.6.1. Gradient Boosting Regresyonu

4.10.1.6.1.1. GB Veri Hazırlığı ve Model Parametreleri

Aykırı değerlerden arındırılmış veri seti kullanılmıştır. Sayısal ve kategorik değişkenler uygun şekilde dönüştürülmüş ve eğitim ile test verisi %80-%20 oranında ayrılmıştır. Gradient Boosting regresyon modeli aşağıdaki parametrelerle eğitilmiştir:

- Ağaç sayısı (n_estimators): 100
- Öğrenme hızı (learning_rate): 0.1
- Maksimum derinlik (max_depth): 3
- Rastgelelik ve örnekleme parametreleri varsayılan olarak kullanılmıştır.

4.10.1.6.1.2. GB Model Performans Sonuçları

Model test setinde Tablo 4.19.'deki değerler elde etmiştir.

Tablo 4.19. Gradient Boosting Model Performansı

Performans Ölçütü	Değer
Ortalama Kare Hata (MSE)	392,942,135,946
Ortalama Mutlak Hata (MAE)	455,113
Determinasyon Katsayısı (R^2)	0.82

R^2 değeri, modelin test verisindeki fiyat varyansının %82'sini başarıyla açıkladığını göstermektedir. Ortalama mutlak hata ise tahminlerin gerçek fiyatlardan yaklaşık 455 bin TL sapma gösterdiğini ifade eder.

4.10.1.6.2. XGBoost Regresyonu

4.10.1.6.2.1. Veri Hazırlığı ve Model Parametreleri

Çalışmada, aykırı değerlerden arındırılmış veri seti kullanılmıştır. Sayısal ve kategorik değişkenler uygun şekilde dönüştürülmüş ve eğitim ile test verisi %80-%20 oranında ayrılmıştır. XGBoost regresyon modeli aşağıdaki parametrelerle eğitilmiştir:

- **Ağaç sayısı (n_estimators):** 100
- **Öğrenme hızı (learning_rate):** 0.1
- **Maksimum derinlik (max_depth):** 3
- **Rastgelelik ve örnekleme parametreleri:** Varsayılan olarak kullanılmıştır.

4.10.1.6.2.2. Model Performans Sonuçları

Model test setinde Tablo 4.20.'deki performans değerleri elde edildi.

Tablo 4.20. XGBoost Model Performansı

Performans Ölçütü	Değer
Ortalama Kare Hata (MSE)	471,268,278,930.88
Ortalama Mutlak Hata (MAE)	505,232.23
Determinasyon Katsayısı (R^2)	0.79

R^2 değeri, modelin test verisindeki fiyat varyansının %79'unu başarıyla açıkladığını göstermektedir.
Ortalama Mutlak Hata (MAE) ise tahminlerin gerçek fiyatlardan yaklaşık **505 bin TL** sapma gösterdiğini ifade eder. Bu sonuçlar, modelin iyi bir doğrulukla tahmin yaptığına işaret etmektedir.

4.10.1.6.2.3. Değerlendirme ve Sonuçlar

- **R^2 Skoru (0.79)**: Model, **test verisindeki fiyatların %79'unu başarıyla tahmin etmiştir**.
- **MSE (471 milyar TL)**: Ortalama Kare Hata değeri, modelin tahminlerinin **gerçek fiyatlardan ne kadar sapma gösterdiğini** ortaya koyar. Model, büyük fiyat varyasyonlarına rağmen kabul edilebilir bir hata ile çalışmaktadır.
- **MAE (505,232 TL)**: Ortalama Mutlak Hata, her bir tahminin **gerçek fiyatlardan yaklaşık 505 bin TL sapma gösterdiğini** gösterir.

4.10.1.6.3. AdaBoost Regresyonu

4.10.1.6.3.1. Veri Hazırlığı ve Model Parametreleri

- **Ağaç sayısı (n_estimators):** 100
- **Temel model (base_estimator):** Karar ağaçları (max_depth=3)
- **Rastgelelik ve örnekleme parametreleri:** Varsayılan olarak kullanılmıştır.

4.10.1.6.3.1. Model Performans Sonuçları

Model test setinde Tablo 4.21.'deki performans değerleri elde edildi.

Tablo 4.21. AdaBoost Model Performansı

Performans Ölçütü	Değer
Ortalama Kare Hata (MSE)	1,003,239,337,936.32
Ortalama Mutlak Hata (MAE)	722,385.72
Determinasyon Katsayısı (R^2)	0.55

R^2 değeri, modelin test verisindeki fiyat varyansının %55'ini başarıyla açıkladığını göstermektedir.
Ortalama Mutlak Hata (MAE) ise tahminlerin gerçek fiyatlardan yaklaşık **722 bin TL** sapma gösterdiğini ifade eder. Bu sonuç, modelin daha düşük doğrulukla tahminler yaptığına işaret eder.

4.10.1.6.3.2. Değerlendirme ve Sonuçlar

- **R^2 Skoru (0.55)**: Model, **test verisindeki fiyatların %55'ini açıklayabilmiştir**. Bu, modelin performansının sınırlı olduğunu, fakat yine de bazı doğru tahminler yaptığını gösterir.
- **MSE (1 trilyon TL)**: Ortalama Kare Hata değeri oldukça yüksektir Modelin **gerçek fiyatlardan daha büyük sapmalar gösterdiği söylenebilir.**
- **MAE (722,385 TL)**: Ortalama Mutlak Hata değeri, her bir tahminin **gerçek fiyatlardan yaklaşık 722 bin TL sapma gösterdiğini** ifade eder

4.10.1.7. Kümeleme Bulguları

4.10.1.7.1. K-Means Kümeleme

- **Elbow yöntemi:** Küme sayısı olarak 3 veya 4 tercih edilmiştir.
- **3 küme modeli:**
  - Küme 0: Ortalama fiyat yaklaşık 2.393.000 TL, ortalama net metrekare ~53.5 m², ortalama 2+1 oda.
  - Küme 1: Ortalama fiyat yaklaşık 4.322.000 TL, ortalama net metrekare ~134.5 m², ortalama 4+1 oda.
  - Küme 2: Ortalama fiyat yaklaşık 5.030.000 TL, ortalama net metrekare ~123.2 m², ortalama 3+1 oda.
- **4 küme modeli:** 3 küme modeline göre daha detaylı segmentasyon sağladı.

4.10.1.7.2. DBSCAN Kümeleme

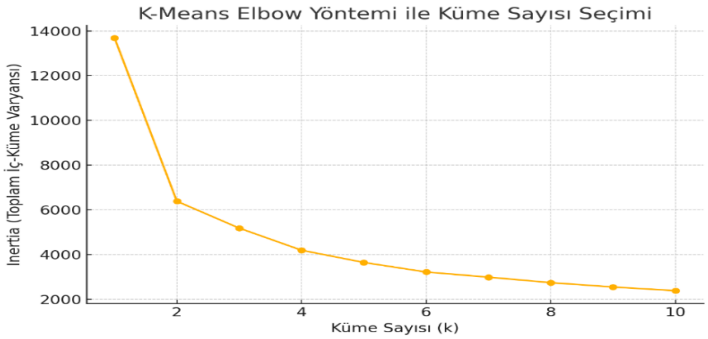
- Parametreler: ε = 0.5 (ölçeklendirilmiş mesafe), MinPts = 5
- Sonuçlar:
  - 432 adet gürültü (aykırı) nokta tespit edilmiştir.
  - Birden fazla küçük küme oluşmuştur; bu da veri setindeki düzensiz yapı ve farklı yoğunlukları yansıtmaktadır.

4.10.1.7.3. Kümeleme Karşılaştırması

- **Segmentasyon:** K-Means algoritması, konutları fiyat ve büyüklük açısından anlamlı segmentlere ayrıldı. 3 küme modeli yatırımcı ve alıcıların karar süreçlerinde basit ve etkili bir sınıflandırma sunabilir.
- **Aykırı Değerler:** Piyasadaki normal kalıpların dışındaki aşırı fiyatlı veya aşırı düşük fiyatlı konutları tespit etmede başarılıdır. Bu aykırı fiyatlar, piyasa anomalileri veya özel durumlar olarak değerlendirilebilir.
- **Piyasa Analizi:** Bölgesel segmentasyon, mahalle bazında fiyat farklılıklarının ortaya çıkarılmasına yardımcı olur. Özellikle fiyat kümeleri ve aykırı değerlerin tespiti, piyasa risklerini azaltmak ve yatırım fırsatlarını belirlemek açısından önemlidir.

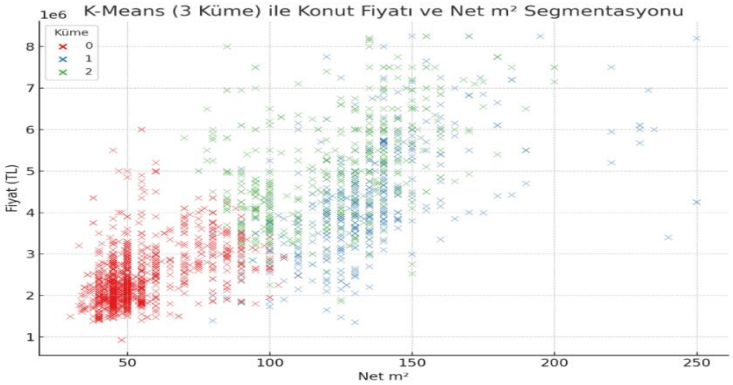
4.10.1.7.4. Kümeleme ve Aykırı Değer Analizlerinin Grafiksel Gösterimi

- **Elbow Yöntemi Grafiği:** Kümelerin iç varyanslarının küme sayısına göre değişimi analiz edilmiştir.



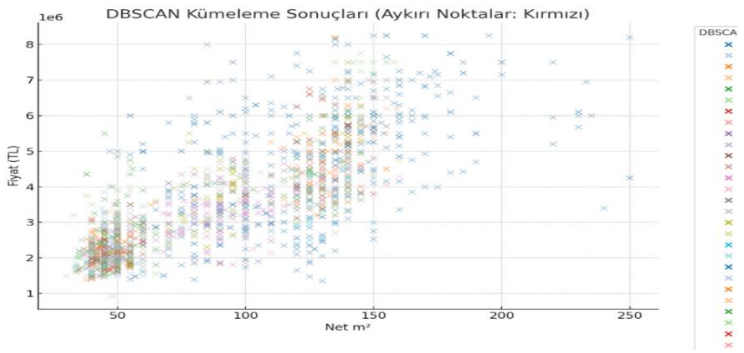
Şekil 4.10. Elbow Yöntemi Grafiği

- **K-Means (3 Küme) Scatter Plot:** Net metrekareye göre fiyat dağılımı renkli küme segmentleri olarak görselleştirilmiştir.



Şekil 4.11. K-Means (3 Küme) Scatter Plot

- **DBSCAN Scatter Plot:** Aykırı noktalar kırmızı renk ile işaretlenmiş, diğer kümeler farklı renklerle gösterilmiştir.



Şekil 4.12. DBSCAN Scatter Plot

4.10.1.7. Hedonik Fiyatlandırma Modeli

- Fiyat ve net metrekare değişkenlerinde aykırı değerler IQR yöntemiyle temizlendi; analiz için 2228 geçerli gözlem kaldı.
- Bağımsız değişkenlerden sayısal değişkenler StandardScaler ile normalize edilmiştir.
- Fiyat değişkenine logaritmik dönüşüm uygulandı.
- Kategorik değişkenler 0-1 kodlaması şeklinde modele dahil edildi.

4.10.1.7.1. Hedonik Fiyatlandırma Modeli Performansı

- Bağımlı değişken: Logaritmik dönüşümlü konut fiyatı (log(Fiyat)).
- Bağımsız değişkenler: Normalize edilmiş sayısal değişkenler ve kodlanmış kategorik değişkenler.
- Model: Lineer regresyon (OLS).

Tablo 4.22. Hedonik Fiyatlandırma Modeli Değişkenlerinin Değerlendirilmesi

Değişken	Katsayı (β)	Std. Hata	t-değeri	p-değeri	Yorum
Sabit Terim (const)	9.3905	0.026	363.7	<0.001	Model sabitidir
Net Metrekare	0.0831	0.023	3.565	0.000	Net m² %1 arttığında fiyat ortalama %8.3 artar
Banyo Sayısı	0.0344	0.005	7.566	<0.001	Her ek banyo fiyatı yaklaşık %3.4 artırır
Bina Yaşı Ortalama	-0.0291	0.009	-3.216	0.001	Bina yaşı %1 arttığında fiyat yaklaşık %2.9 düşer
Bulunduğu Kat (Dönüştürülmüş)	0.0188	0.003	6.254	<0.001	Kat sayısı %1 artış fiyatı %1.88 artırır
Site İçerisinde Kodlu	4.6954	0.014	345.7	<0.001	Site içinde olma fiyatı büyük ölçüde pozitif etkiler
Balkon (Kodlu)	0.1194	0.013	9.338	<0.001	Balkonu olanlar %11.9 daha yüksek fiyatlı
Tapu Durumu (Kodlu)	0.0301	0.007	4.612	<0.001	Olumlu tapu durumu %3 fiyat artışı
Otopark (Kodlu)	-0.0433	0.014	-3.193	0.001	Otopark varlığı %4.3 fiyat düşüşüyle ilişkili

$$Fiyat = 9.3905 + (Net\ Metrekare\ x\ 0.0831) + (Banyo\ Sayısı\ x\ 0.0344) - (Bina\ Yaşı\ x\ 0.0291) + (Bulunduğu\ Kat\ x\ 0.0188) + (Site\ İçerisinde\ Kodlu\ x\ 4.6954) + (Balkon\ x\ 0.1194) + (Tapu\ Durumu\ x\ 0.0301) - (Otopark\ x\ 0.0433)$$

- R-kare: %89.4.
- MAE (log fiyat biriminde): 0.115.
- MAE (fiyat biriminde): yaklaşık 414,805 TL.

4.10.1.8. Yapay Sinir Ağı Modeli Sonuçları

4.10.1.8.1. Yapay Sinir Ağı Model Performansı

Tablo 4.23. YSA Model Performansı

Performans Ölçütü	Değer
Test Seti MSE (Gerçek Fiyat TL)	194,656,681,984 TL
Test Seti MAE (Gerçek Fiyat TL)	331,496.47 TL
Test Seti R-kare	0.9221

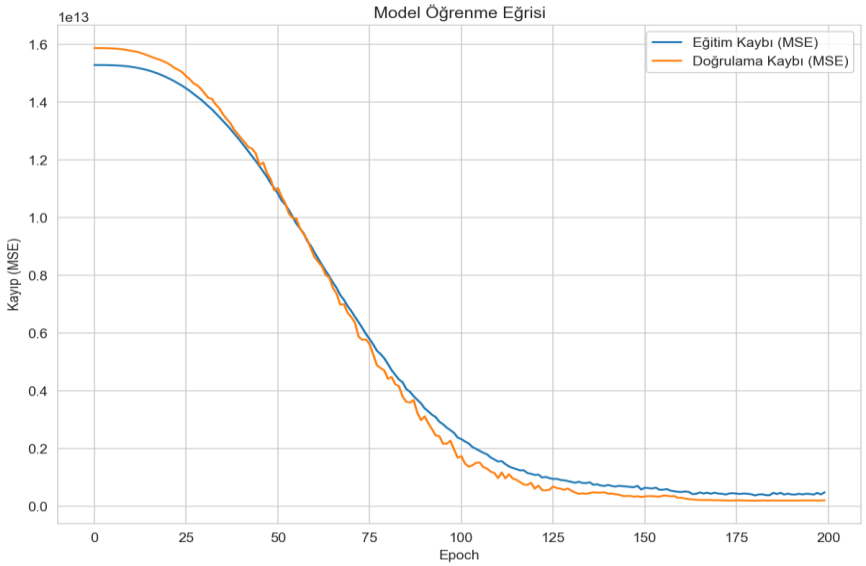
Model, **gerçek fiyat tahminlerinde oldukça başarılı bir performans** sergilemiştir. Yüksek **R-kare** değerini 0.92 ile modelin tahmin ettiği fiyatların çoğunluğunun doğru olduğunu ve modelin genel olarak veriyi çok iyi öğrendiğini göstermektedir.

MAE değeri 331,496 TL, modelin ortalama hata büyüklüğünü gösterir. Yüksek bir **MAE** değeri olsa da, **R-kare** değeri ile birlikte değerlendirildiğinde, modelin doğru tahminler yapmaya devam ettiği söylenebilir.

MSE değeri ise modelin tahminlerindeki hataların büyüklüğünü daha iyi anlamamıza yardımcı olmaktadır. Bu değer çok yüksek olmakla birlikte, daha çok **büyük hatalar** üzerinden etkilenir. Genellikle daha küçük hataların olduğu durumlardan farklı olarak, model bazen daha büyük hatalar yapmaktadır.

4.10.1.8.2.1. YSA Öğrenme Eğrisi

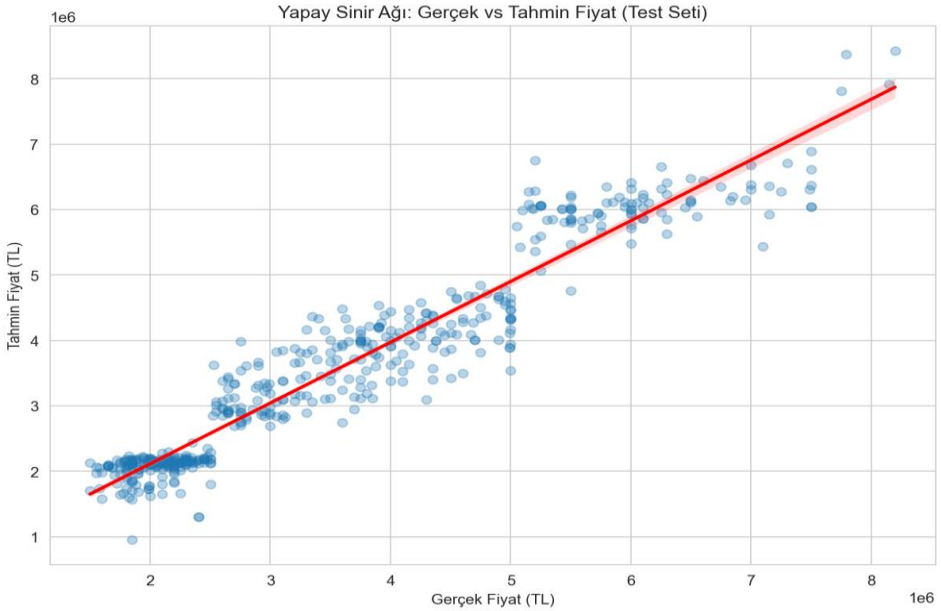
Eğitim kaybı ve **doğrulama kaybı** eğrileri genellikle **sürekli olarak azalma gösteriyor** ve bu da modelin öğrenme sürecinde sağlıklı bir şekilde ilerlediğini ve **aşırı uyum yapmadığını** gösterir. **Doğrulama kaybı** da benzer şekilde düştü, bu da modelin **genelleme kapasitesinin güçlü olduğunu** ve eğitim verisi dışında test verisinde de başarılı sonuçlar elde ettiğini gösterir.



Şekil 4.13. YSA Öğrenme Eğrisi

4.10.1.8.2.2. YSA Test ve eğitim veri seti karşılaştırması

Bununla birlikte, **bazı uç noktalarda büyük hatalar** gözlemlenmiştir. Bu da, modelin bazı **yüksek fiyatlı emlaklar** için tahmin yaparken zorlandığını ve bu durumun modelin doğruluğunu etkileyebileceğini gösteriyor.



Şekil 4.14. YSA Gerçek vs Tahmin Fiyat

5. SONUÇLAR

Analizlerin ilk aşamasında uygulanan istatistiksel testler sayesinde, konut fiyatlarını etkileyen temel faktörler daha net görülmektedir.

5.1 İstatistiksel Analiz Sonuçları

Bağımsız örneklem t-testi sonuçlarına göre, konutun takasla alınıp alınmadığı ya da site içinde olup olmaması gibi durumlar, bazı fiziksel özelliklerde fark ortaya koyar ki özellikle site içinde yer alan konutların daha büyük ve daha pahalı olduğu dikkat çekmiştir ($p < 0.0001$). Takasla alınan konutlar ise daha küçük ve yeni olsa da, fiyat bakımından anlamlı bir fark bulunmamıştır ($p = 0.53$).

Korelasyon analizinde ise fiyat ile hem brüt hem de net metrekare arasında güçlü pozitif bir ilişki gözlemlendi ($r \approx 0.82$ ve $r \approx 0.80$). Yani metrekare arttıkça fiyat da doğal olarak yükseliyor. Öte yandan, bina yaşıyla fiyat arasında zayıf da olsa negatif bir ilişki ortaya çıktı ($r \approx -0.26$). Bu da yaşlı binaların biraz daha düşük fiyattan satıldığını gösteriyor.

Ki-kare testleri, bazı kategorik değişkenler arasında anlamlı ilişkiler bulunmuştur. Örneğin, krediye uygun konutların çoğunlukla eşyalı olduğu ya da takas yapılan konutların kullanım durumuyla ilişkili olduğu anlaşıldı ($\chi^2 = 319.61$ ve $\chi^2 = 32.84$ ile oldukça anlamlı sonuçlar elde edildi).

Son olarak, **MANOVA testi** ile takas durumu, kullanım şekli ve site içerisinde yer alma gibi faktörlerin; fiyat, brüt ve net metrekare üzerinde anlamlı etkileri olduğu tespit edildi. Yani bu özellikler sadece "önemsiz detaylar" değil, fiyatın şekillenmesinde gerçekten rol oynuyor.

5.2 Makine Öğrenmesi Bulguları

Veri setine daha farklı bir bakış kazandırmak adına, çeşitli makine öğrenmesi algoritmalarıyla fiyat tahmini yapıldı ve açıkça söylemek gerekirse bazı modellerin performansı iyi sonuçlar verdi.

Doğrusal modellerle başlamak gerekirse; **Basit Doğrusal Regresyon** en temel modeldi ve doğal olarak sınırlı kaldı (**$R^2 = 0.67$**). Ancak **Çoklu Doğrusal Regresyon** biraz daha derine inerek fiyat üzerinde etkili olabilecek değişkenleri hesaba kattığında daha iyi sonuç verdi (**$R^2 = 0.70$**). Brüt m², bina yaşı ve kat sayısı burada öne çıkan etkenlerdi.

Ardından gelen **K-NN Regresyon** modeli, doğrusal modellere kıyasla daha başarılı çıktı (**$R^2 = 0.80$**). Fakat en yüksek başarıyı **Random Forest** yakaladı; model **$R^2 = 0.84$** ile fiyat tahminini oldukça isabetli yaptı. Bu modelle birlikte, mahalle bilgisi gibi mekânsal etkenlerin de tahmin doğruluğunu artırdığı görülmüş oldu.

Boosting algoritmaları arasında en iyi sonucu **Gradient Boosting** verdi (**R^2 = 0.82**), onu da **XGBoost** takip etti (**R^2 = 0.79**). Ancak **AdaBoost** beklentilerin altında kaldı (**R^2 = 0.55**), bu da bazı modellerin her problemde aynı başarıyı gösteremeyeceğini gösteriyor.

Son olarak, en yüksek başarı **Yapay Sinir Ağı (YSA)** modeliyle elde edildi. **R^2 = 0.92** gibi oldukça yüksek bir doğruluk oranına ulaşıldı. Ancak bu modelde özellikle yüksek fiyatlı konutlarda hata oranı biraz arttı.

Tablo 5.1. Danışmanlı Makine Öğrenmesi Model Karşılaştırması

Model	MAE (TL)	R^2
Yapay Sinir Ağları	331,496	0.92
Random Forest Regresyon	394,990	0.84
Gradient Boosting Regresyon	455,113	0.82
K-NN Regresyon	449,320	0.80
XGBoost	505,232	0.79
Polinomik Regresyon (3. Derece)	549,872	0.74
Doğrusal Polinomik Regresyon	550,987	0.739
Ridge Regresyon	551,137	0.739
Lasso Regresyon	550,987	0.739
Elastic Net Regresyon	552,073	0.739
Polinomik Regresyon (2. Derece)	551,284	0.738
Çoklu Doğrusal Regresyon	593,114	0.70
İleri Seçim Regresyon	599,094	0.70
Geriye Eleme Regresyon	593,177	0.70
Anlamlı Değişkenler Regresyon	593,177	0.70
Basit Doğrusal Regresyon	617,190	0.67
Üstel Regresyon	738,619.53	0.669
AdaBoost	722,385	0.55

Bu sonuçlar, **Yapay Sinir Ağları modelinin konut fiyatlarını en doğru tahmin eden model** olduğunu açıkça gösteriyor. Düşük hata ve yüksek determinasyon katsayısına sahiptir. Elde edilen bulgular doğru değişken seçimi ve uygun modelleme teknikleri kullanıldığında yüksek doğrulukla tahminler yapılabileceğini göstermektedir. Böylece hem akademik literatüre katkı sağlanmış hem de gayrimenkul sektörüne yönelik pratik bir çerçeve sunulmuştur.

İleriye dönük olarak, farklı şehirler veya bölgelerdeki veriler kullanılarak modelin genellenebilirliği test edilebilir. Ayrıca ekonomik göstergeler ve sosyal faktörler gibi ek değişkenlerin entegre edilmesi, modellerin tahmin gücünü artırabilir. Dinamik piyasa koşullarına uygun zaman serisi modelleri veya güncel verilere adapte olabilen yaklaşımlar kullanmak da gelecekteki tahminlerde daha başarılı sonuçlar verebilir.

KAYNAKLAR

Afşar, A., Tekin, M., & Özdemir, F. (2017). Eskişehir'de konut fiyatlarını etkileyen faktörlerin analizi. *Bölgesel Kalkınma Dergisi*, 11(4), 122-138.

Ak Çetin, S., & Akpınar, M. (2021). Seferihisar ilçesi konut fiyatlarının belirleyicileri. *Ege Bölgesi Araştırmaları Dergisi*, 12(2), 50-65.

Aliyev, S., Mammadov, R., & Huseynov, T. (2019). Housing price determinants in Baku: A regression analysis. *Azerbaijan Economic Review*, 6(1), 44-61.

Alpaydın, E. (2020). *Introduction to Machine Learning*. MIT Press.

Altun, H. (2022). Eskişehir konut piyasasında makine öğrenmesi tabanlı fiyat tahmini. *Bölgesel Ekonomi Araştırmaları*, 3(1), 12-29.

Anderson, T. W., & Darling, D. A. (1952). Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *Annals of Mathematical Statistics*, 23(2), 193–212.

Barut, A., & Bilgin, B. (2023). Konut fiyat tahmininde yapay sinir ağları ve polinomsal regresyon karşılaştırması. *Gayrimenkul Ekonomisi Dergisi*, 15(2), 45-60.

Barut, Z., & Bilgin, T. T. (2023). Konut fiyatlarının tahmini için polinomsal regresyon ve yapay sinir ağları yöntemlerinin uygulamalı karşılaştırılması. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, 76, 221–237.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Chao, L., Wang, M., & Li, Q. (2025). Social housing satisfaction and MANOVA analysis: A Guangzhou case study. *Housing Policy Review*, 22(1), 77-98.

Chen, Y., Li, X., & Wang, J. (2017). Machine learning methods for housing price prediction: A comparative study. *Journal of Real Estate Research*, 39(1), 112-135.

Chugani, P. (2021). ANOVA and Kruskal-Wallis applications in real estate studies. *Journal of Statistical Applications*, 34(4), 405-420.

Çiçek, U., & Hatırlı, S. A. (2015). Isparta ilinde konut fiyatlarını etkileyen faktörlerin hedonik fiyat modeli ile analizi.

Dayı, F., & Gencan, M. Y. (2024). Konut fiyatlarını etkileyen faktörlerin incelenmesi: Samsun örneği. *Gümüşhane Üniversitesi Sosyal Bilimler Dergisi*, 15(2), 348–364.

Do, Q., & Grudnitski, G. (1992). A neural network approach to residential property appraisal. *The Real Estate Appraiser*, 58(3), 38–45.

Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis*. Wiley.

Ecer, F. (Yıl). Türkiye'deki konut fiyatlarının tahmininde hedonik regresyon yöntemi ile yapay sinir ağlarının karşılaştırılması. *Afyon Kocatepe Üniversitesi*.

Ellibeş, T., & Görmüş, H. (2018). Kocaeli'de konut fiyatlarını etkileyen faktörlerin incelenmesi. *İnşaat ve Gayrimenkul Dergisi*, 7(3), 15-33.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). SAGE Publications.

Freedman, D. (2009). *Statistical Models: Theory and Practice*. Cambridge University Press.

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics* (4th ed.). W.W. Norton & Company.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29(5), 1189–1232.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1), 119–139.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Gravetter, F. J., & Wallnau, L. B. (2013). *Statistics for the Behavioral Sciences* (9th ed.). Wadsworth, Cengage Learning.

Güzel, A., Kaya, R., & Demir, T. (2020). Ordu'da çoklu regresyonla konut fiyat tahmini. *Bölgesel Araştırmalar Dergisi*, 15(1), 34-49.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

Kahveci, M. (2020). Türkiye'de konut fiyatı belirleyicilerinin analizi (Doktora tezi).

Kangalli Uyar, E., & Keten, H. (2020). Denizli merkezinde konut fiyatlarının mekânsal analizi. *Kent ve Mekân Dergisi*, 9(2), 90-108.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.

Kayış, A. (2006). Güvenirlilik analizi. Ş. Kalaycı (Ed.), *SPSS uygulamalı çok değişkenli istatistik teknikleri* (s. 401–419). Ankara: Asil.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4, 83–91.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583–621.

Mağden, E. (2022). Ordu ilinde konut fiyatlarını etkileyen faktörlerin analizi. *Gayrimenkul Ekonomisi Dergisi*, 18(3), 100-120.

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics*, 18(1), 50–60.

McCluskey, W., & Borst, R. A. (1997). An evaluation of MRA, comparable sales analysis, and artificial neural networks (ANNs) for the mass appraisal of residential properties in Northern Ireland. *Assessment Journal*, 4(1), 47–55.

McHugh, M. L. (2013). The Chi-square test of independence. *Biochemia Medica*, 23(2), 143-149.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Monika, P., Singh, R., & Sharma, V. (2021). Predicting real estate prices using machine learning algorithms. *International Journal of Housing Markets and Analysis*, 14(3), 675-693.

Montgomery, D. C. (2017). *Design and Analysis of Experiments* (9th ed.). Wiley.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. Wiley.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2017). *Introduction to the Practice of Statistics* (9th ed.). W.H. Freeman.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Newbold, P., Carlson, W. L., & Thorne, B. (2013). *Statistics for Business and Economics* (8th ed.). Pearson Education.

Park, S., & Bae, H. (2015). Support vector regression for house price prediction. *Journal of Property Research*, 32(2), 133-150.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58, 240–242.

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression approach. *Journal of Housing and the Built Environment*, 24(3), 285–298.

Spiegel, M. R., Schiller, J., & Srinivasan, R. (2012). *Schaum's Outline of Statistics* (4th ed.). McGraw-Hill Education.

Spiegelhalter, D. J. (2019). *The Art of Statistics: How to Learn from Data*. Basic Books.

TCMB (2001). *Yıllık rapor*. http://www.tcmb.gov.tr

Tosun Gavca, C. (2024). Türkiye'de yabancılara yapılan konut satışının tahmininde makine öğrenmesi yöntemleri performanslarının incelenmesi (Yayınlanmamış doktora tezi). Pamukkale Üniversitesi, Sosyal Bilimler Enstitüsü, Denizli.

Triola, M. F. (2018). *Elementary Statistics* (13th ed.). Pearson.

Tukey, J. W. (1949). Comparing Individual Means in the Analysis of Variance. *Biometrics*, 5(2), 99–114.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 59(236), 433-460.

Uğuz, H. (2021). *Makine Öğrenmesi: Teori ve Python ile Uygulamalar*. Nobel Akademik Yayıncılık.

Utts, J., & Heckard, R. (2014). *Mind on Statistics* (5th ed.). Brooks/Cole, Cengage Learning.

Wilcox, R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing* (3rd ed.). Academic Press.

Yılmaz, M., Demir, F., & Kara, S. (2017). Samsun'da konut fiyatlarını etkileyen unsurlar: Hedonik fiyatlama analizi. *Bölgesel Ekonomi ve Planlama Dergisi*, 8(2), 56-74.

Yılmazel, E., Kaya, S., & Demir, O. (2018). İstanbul, Ankara ve Eskişehir'de yapay sinir ağları ile konut fiyat tahmini. *Gayrimenkul Araştırmaları Dergisi*, 10(1), 25-40.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2), 301-320.