T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KONUT SATIŞ FİYATLARININ COĞRAFİ BİLGİ SİSTEMİ GİRDİLİ MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE TAHMİNLENMESİ: İSTANBUL ÖRNEĞİ

Duygu ATASEVER

YÜKSEK LİSANS TEZİ
İstatistik Anabilim Dalı
İstatistik Programı

Danışman
Prof. Dr. Ersoy ÖZ

Temmuz, 2024

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KONUT SATIŞ FİYATLARININ COĞRAFİ BİLGİ SİSTEMİ GİRDİLİ MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE TAHMİNLENMESİ: İSTANBUL ÖRNEĞİ

Duygu ATASEVER tarafından hazırlanan tez çalışması 17.07.2024 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı İstatistik Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Prof. Dr. Ersoy ÖZ
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Prof. Dr. Ersoy ÖZ, Danışman
Yıldız Teknik Üniversitesi _____

Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK, Üye
Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üye. Damla İLTER FAKHOURI, Üye
Mimar Sinan Güzel Sanatlar Üniversitesi _____

Danışmanım Prof. Dr. Ersoy ÖZ sorumluluğunda tarafımca hazırlanan Konut Satış Fiyatlarının Coğrafi Bilgi Sistemi Girdili Makine Öğrenmesi Yöntemleri İle Tahminlenmesi: İstanbul Örneği başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Duygu ATASEVER

İmza

Aileme
ve
dostlarıma

TEŞEKKÜR

Bu tez çalışmasının tamamlanmasında büyük bir katkısı olan danışmanım Prof. Dr. Ersoy Öz'e içtenlikle teşekkür etmek istiyorum. Bilimsel bilgisi, kapsamlı görüşleri, sürekli destekleri ve sabrı olmasaydı, bu çalışmayı tamamlamam mümkün olmazdı. Çalışmalarımda sürekli ilham verici oldu ve bilimsel araştırmaya olan tutkumu besledi. Bu yolculuğum boyunca bana inanmayı ve beni desteklemeyi asla bırakmayan aileme ve dostlarıma sonsuz teşekkürler. Anlayışları, motivasyonları ve inançları, bu tezi tamamlama sürecimde hayati bir rol oynadı. Bu tez, akademik ilerlememde ve kişisel gelişimimde önemli bir kilometre taşıdır ve bu yolculukta beni destekleyen herkese sonsuz teşekkürlerimi sunarım.

Duygu ATASEVER

İÇİNDEKİLER

KISALTMA LİSTESİ vii

ŞEKİL LİSTESİ viii

TABLO LİSTESİ ix

ÖZET x

ABSTRACT xii

1 GİRİŞ 1

1.1 Literatür Özeti . . . 2
1.2 Tezin Amacı . . . 7
1.3 Hipotez . . . 7

2 COĞRAFİ BİLGİ SİSTEMLERİ 9

2.1 CBS'de Kullanılan Analiz Türleri . . . 10
2.1.1 Coğrafi sorgulama . . . 10
2.1.2 Coğrafi Analiz . . . 11
2.1.3 Ağ Analizi . . . 12
2.1.4 Sayısal Arazi Analizi . . . 12
2.1.5 İstatistiksel Analiz . . . 13
2.1.6 Izgara (*Grid*) Analizi . . . 13
2.2 CBS'nin Uygulama Alanları . . . 14
2.3 CBS ve İstatistik . . . 16
2.4 Konut Fiyat Tahminlenmesinde CBS'nin Kullanımı . . . 16

3 MAKİNE ÖĞRENMESİ 18

3.1 Makine Öğrenmesinin Temelleri . . . 18
3.2 Makine Öğrenmesi Yöntemleri . . . 19
3.2.1 Denetimli Öğrenme . . . 19
3.2.2 Denetimsiz Öğrenme . . . 21
3.2.3 Yarı Denetimli Öğrenme . . . 22
3.2.4 Pekiştirmeli Öğrenme . . . 22

3.3 Makine Öğrenmesinin Uygulama Alanları 22
3.4 Konut Fiyat Tahminlemesinde Makine Öğreniminin Kullanımı . . . 24
3.5 Uygulamada Kullanılan Makine Öğrenmesi Algoritmaları 25
3.5.1 XGBoost Algoritması 26
3.5.2 LightGBM Algoritması 28
3.5.3 CatBoost Algoritması 29
3.6 Performans Değerlendirme Metrikleri 31
3.6.1 R Kare (R^2) 31
3.6.2 Ortalama Kare Hata 33
3.6.3 Kök Ortalama Kare Hata 34
3.6.4 Ortalama Mutlak Hata 35

4 UYGULAMA **37**
4.1 Veri Kümesi 38
4.2 Verilerin Analizi 40
4.3 Modellerin Uygulanması ve Sonuçlar 48

5 SONUÇ **57**

KAYNAKÇA **60**

TEZDEN ÜRETİLMİŞ YAYINLAR **64**

KISALTMA LİSTESİ

AHP	Analitik Hiyerarşi Süreci
AVM	Otomatik Değerleme Modelleri
CBS	Coğrafi Bilgi Sistemi
ÇKKV	Çok Kriterli Karar Verme
DTs	Karar Ağaçları
EKK	En Küçük Kareler
GBDT	Gradyan Artırıcı Karar Ağacı
GWR	Coğrafi Olarak Ağırlıklandırılmış Regresyon
HPM	Hedonik Fiyat Modeli
KNN	k En Yakın Komşu
LightGBM	Hafif Gradyan Artırma Makineleri
MRA	Çoklu Regresyon Analizi
MAE	Ortalama Mutlak Hata
MSE	Ortalama Kare Hata
NLP	Doğal Dil İşleme
PCA	Temel Bileşenler Analizi
RF	Rastgele Orman
RL	Pekiştirmeli Öğrenme
RMSE	Kök Ortalama Kare Hatası
SVM	Destek Vektör Makinaları
XGBoost	Aşırı Gradyan Artırma
YSA	Yapay Sinir Ağları

ŞEKİL LİSTESİ

Şekil 3.1 Makine öğrenmesi yöntemleri 19
Şekil 4.1 İlçelere göre ortalama konut fiyatları 40
Şekil 4.2 Asansör değişkeni dağılımı 41
Şekil 4.3 Otopark değişkeni dağılımı 41
Şekil 4.4 Kredi durumu değişkeni dağılımı 42
Şekil 4.5 Site içerisinde olma durumu dağılımı 42
Şekil 4.6 Kat sayısı değişkeni dağılımı 43
Şekil 4.7 Konutun bulunduğu kat değişkeni dağılımı 44
Şekil 4.8 Bina yaşı dağılımı 44
Şekil 4.9 Daire kategorisi değişkeninin dağılımı 45
Şekil 4.10 Daire kategorilerine göre ortalama fiyat dağılımı 45
Şekil 4.11 En yüksek ortalama fiyata sahip 10 ilçe 46
Şekil 4.12 En düşük ortalama fiyata sahip 10 ilçe 47
Şekil 4.13 Özelliklerin model üzerindeki etki dereceleri 55

TABLO LİSTESİ

Tablo 4.1 Veri kümesinde çalışmada kullanılacak konuta ait değişkenler . . 38
Tablo 4.2 Uzaklık bilgisinin bulunduğu değişkenler 38
Tablo 4.3 Veri kümesinden üretilen değişkenler 39
Tablo 4.4 Konutların fiyatlarına ilişkin istatistiksel özet 46
Tablo 4.5 Temel değişkenlerin kullanıldığı modelin sonuçları 49
Tablo 4.6 Hiperparametre optimizasyonuyla geliştirilmiş temel değişkenlerin kullanıldığı modelin sonuçları 50
Tablo 4.7 CBS değişkenlerinin eklendiği modelin sonuçları 52
Tablo 4.8 Hiperparametre optimizasyonuyla geliştirilmiş CBS modeli performans sonuçları 53

ÖZET

Konut Satış Fiyatlarının Coğrafi Bilgi Sistemi Girdili Makine Öğrenmesi Yöntemleri İle Tahminlenmesi: İstanbul Örneği

Duygu ATASEVER

İstatistik Anabilim Dalı
Yüksek Lisans Tezi

Danışman: Prof. Dr. Ersoy ÖZ

Konut piyasası, her konutun benzersiz özelliklere sahip olması nedeniyle çeşitlilik gösterir. Bu yüzden konut satış fiyatlarının tahmini, hem bireyler hem de sektör profesyonelleri için önemli bir konudur. Özellikle İstanbul gibi büyük ve dinamik bir şehirde, fiyatların nasıl şekillendiğini anlamak, doğru yatırım kararları alabilmek için kritik bir öneme sahiptir. Bu bağlamda, gelişmiş analitik yöntemler ve teknolojilerin kullanımı, konut satış fiyat tahminlerinin doğruluğunu artırmada büyük bir rol oynar. Coğrafi Bilgi Sistemleri (CBS) ve makine öğrenmesi tekniklerinin entegrasyonu, bu alanda yeni ve etkili bir yaklaşım sunmaktadır. Bu yöntemler, konut satış fiyatlarını etkileyen çeşitli faktörlerin, özellikle de konumsal faktörlerin, derinlemesine analiz edilmesini sağlar.

Araştırmanın temel amacı, İstanbul'un çeşitlilik gösteren konut piyasasında, satış fiyatlarının tahmin edilmesi sürecinde CBS ve makine öğrenmesi tekniklerinin etkin bir şekilde kullanılmasıdır. Makine öğrenmesi algoritmaları, konutların değerini etkileyen birçok faktörü analiz ederek, fiyat tahminlerinde bulunurken, CBS teknikleri konutların coğrafi konumlarını dikkate alarak, fiyatlar üzerindeki konumsal etkilerini gözler önüne serer. Bu yaklaşım, konut satış fiyatlarının daha kapsamlı bir şekilde anlaşılmasını ve daha doğru tahminler yapılmasını sağlamaktadır. Bu bakımdan çalışmada iki farklı model geliştirilmiştir. İlk modelde, konutun temel öznitelikleri (örneğin, metrekare, oda sayısı, bina yaşı) kullanılmıştır. İkinci modelde ise, CBS ile hesaplanmış uzaklık değişkenleri

(örneğin, park, toplu taşıma gibi önemli noktalara olan mesafeler) temel özelliklere eklenmiştir. XGBoost, LightGBM ve CatBoost algoritmaları ile gerçekleştirilen analizlerde, CBS ile zenginleştirilmiş modelin performansının daha yüksek olduğu gözlemlenmiştir. Özellikle LightGBM algoritması, CBS değişkenlerinin eklendiği modelde en yüksek performansı göstermiştir. XGBoost ve CatBoost algoritmaları da başarılı sonuçlar elde etmiştir. Araştırma sonuçları, konut satış fiyatlarının tahmininde CBS ve makine öğrenmesi tekniklerinin birlikte kullanımının önemini vurgulamaktadır.

Bu tez çalışması, konut piyasasının analiz edilmesi ve fiyat tahminlerinin yapılması konusunda yeni bir perspektif sunarak, sektördeki uygulamalara katkıda bulunmayı amaçlamaktadır. Çalışmada kullanılan yaklaşım, konut piyasasının daha kapsamlı bir şekilde anlaşılmasını ve daha doğru tahminler yapılmasını sağlamaktadır.

Anahtar Kelimeler: Konut değerleme, coğrafi bilgi sistemleri, makine öğrenmesi.

ABSTRACT

House Sales Price Prediction With Geographic Information System Inputed Machine Learning Methods: Istanbul Sample

Duygu ATASEVER

Department of Statistics
Master of Science Thesis

Supervisor: Prof. Dr. Ersoy ÖZ

The housing market is characterised by a high degree of diversity, with each residential property exhibiting distinctive characteristics. It is therefore of great importance to be able to estimate house sales prices, both for individuals and for industry professionals. In a large and dynamic city like Istanbul, understanding how prices are shaped is of critical importance in making the right investment decisions. In this context, the use of advanced analytical methods and technologies plays a significant role in improving the accuracy of home sales price forecasts. The integration of Geographic Information Systems (GIS) and machine learning techniques offers a new and effective approach in this field. These methods enable in-depth analysis of various factors affecting home sales prices, especially locational factors.

The principal objective of the research is to ascertain the most efficacious methodologies for utilising GIS and machine learning techniques in the estimation of sales prices within Istanbul's heterogeneous housing market. While machine learning algorithms make price predictions by analysing a multitude of factors that affect the value of houses, geographic information system (GIS) techniques take into account the geographical locations of houses and reveal the locational effects on prices. This approach offers a more comprehensive understanding of home sales prices and more accurate predictions. In this context, two distinct models were developed in the course of the study. In the initial model, the fundamental characteristics of the residence, including square footage, number of rooms, and

age, were utilized. In the second model, distance variables calculated with GIS (for example, distances to important points such as parks and public transportation) were incorporated into the basic characteristics. The results of the analyses conducted with the XGBoost, LightGBM, and CatBoost algorithms indicated that the performance of the GIS-enriched model was superior. In particular, the LightGBM algorithm demonstrated the most pronounced performance in the model where CBS variables were incorporated. Furthermore, the XGBoost and CatBoost algorithms yielded favourable outcomes. The findings of the research study underscore the significance of integrating geographic information systems (GIS) and machine learning techniques for the accurate prediction of house sale prices.

The objective of this thesis is to make a contribution to the practices in the sector by offering a new perspective on the analysis of the housing market and the formulation of price predictions. The approach used in the study provides a more comprehensive understanding of the housing market and more accurate forecasts.

Keywords: House sales price prediction, geographic information systems, machine learning.

1
GİRİŞ

Günümüzde, konut piyasalarındaki dalgalanmalar ve fiyat değişimleri, hem bireysel mülkiyet sahipleri hem de gayrimenkul sektörü paydaşları için önemli bir konu haline gelmiştir. Bu değişkenlik, konut satışlarını etkileyen bir dizi faktörün karmaşıklığından kaynaklanmaktadır. İstatistik ve makine öğrenmesi, bu karmaşıklığı anlamak ve gelecekteki konut satışlarını tahmin etmek için önemli araçlar sunmaktadır. Bu çalışma, konut satışlarının tahmin edilmesi sürecinde istatistik ve makine öğrenmesi tekniklerinin yanı sıra CBS tabanlı analizlerin önemini vurgulamaktadır.

Makine öğrenmesi, büyük ve karmaşık veri kümelerinin analiz edilerek içlerindeki örüntülerin ve ilişkilerin keşfedilmesinde kritik bir rol oynar. Makine öğrenmesi yöntemleri finans, yatırım, gayrimenkul değerleme gibi çeşitli ekonomik alanlarda yaygın olarak kullanılmaktadır [1]. Bu yöntemler, konut piyasasındaki dinamiklerin daha iyi anlaşılmasını sağlayarak, konut satış fiyatlarını etkileyen çeşitli faktörleri tanımlamak ve gelecekteki satışları daha doğru bir şekilde tahmin etmek amacıyla kullanılmaktadır.

CBS, konum verilerini haritalama ve analiz etme yeteneği ile öne çıkar. Konut satışlarındaki eğilimlerin ve kümelenmelerin tespit edilmesi, fiyat değişikliklerinin ve talep artışının coğrafi faktörlerle ilişkilendirilmesi gibi işlemler için etkin bir platform sunar. Ayrıca, sonuçların görsel olarak sunulması, karar alıcılara daha kapsamlı bir bakış açısı sağlar [2].

Bu tez çalışması, makine öğrenmesi modelleriyle yapılan konut satış fiyat tahminlerinin yanı sıra, CBS teknikleriyle oluşturulan tematik haritaların sunduğu coğrafi veri analizleri sayesinde konut piyasasındaki etkileşimleri daha derinlemesine anlamamıza olanak tanıyacaktır. Analizler, konut satış fiyatlarının coğrafi konumlarına göre nasıl değiştiğini, belirli bölgelerdeki talep artışını ve konut piyasasındaki diğer dinamikleri ortaya çıkaracaktır.

CBS tabanlı analizler, konut satış tahminlerine coğrafi bir boyut kazandırarak, konumun konut satış fiyatları üzerindeki etkisini detaylı bir şekilde incelemeyi amaçlar. Bu yöntem, konut satış fiyatlarının yalnızca ekonomik ve sosyal faktörlerle değil, aynı zamanda coğrafi konumla da nasıl şekillendiğini ortaya koyar. Böylece, konut piyasası analizlerinde daha kapsamlı ve derinlemesine bir anlayış geliştirmek mümkün olacaktır. Makine öğrenmesi modelleri ile CBS tabanlı analizlerin entegrasyonu, konut satışları üzerine yapılan tahminlerin doğruluğunu ve güvenilirliğini artırarak karar vericilere ve yatırımcılara değerli içgörüler sunmayı amaçlamaktadır.

Giriş bölümü, tezin konusunu, literatürdeki yerini ve çalışmanın amacını tanıtmaktadır. Araştırmanın önemini vurgulayarak, konunun seçimindeki motivasyon ve tezin genel yapısını ortaya koymaktadır.

İkinci bölümde, CBS'nin ne olduğu, analiz türleri, uygulama alanları, CBS ile istatistik arasındaki ilişki ve konut tahminlemesinde CBS'nin nasıl kullanıldığına dair detaylı bir inceleme sunulmaktadır. Bu bölüm, CBS'nin temel kavramlarını açıklar. Ayrıca, konut piyasası analizinde CBS'nin önemini vurgulamaktadır.

Üçüncü bölümde, makine öğrenmesinin temel prensipleri ve konut tahminlemesinde kullanılan makine öğrenmesi algoritmalarının detaylı bir analizi yapılır. Bu bölüm, algoritmaların teorik temellerini, uygulama yöntemlerini ve konut satış fiyat tahminleri üzerindeki etkilerini incelenmektedir.

Dördüncü bölüm, çalışmada ele alınan teorik bilgilerin pratiğe döküldüğü uygulama bölümüdür. Bu bölümde, CBS ve makine öğrenmesi tekniklerinin birleştirilmesiyle oluşturulan modeller üzerinden konut fiyat tahminlemesi yapılacaktır. Kullanılan veri seti, modelin geliştirilmesi, test edilmesi ve sonuçların analizi detaylı bir şekilde açıklanmaktadır.

Araştırmanın bulguları, CBS ve makine öğrenmesi tekniklerinin konut fiyat tahminlemesindeki etkinliği hakkında değerlendirmeler içeren beşinci bölüm ise sonuç bölümüdür. Çalışmanın sınırlılıkları, sonuçların uygulamadaki önemi ve gelecek araştırmalar için öneriler bu bölümde yer alacaktır. Bu bölüm, tezin genel değerlendirilmesini sunar ve araştırmanın katkılarını özetler niteliktedir.

1.1 Literatür Özeti

Çalışmanın bu bölümü hem ulusal hem de uluslararası alanda yapılmış çeşitli çalışmaları inceleyerek, konut fiyatları üzerindeki mekansal özelliklerin etkisini ve

gayrimenkul değerlemesi için kullanılan metodolojileri ele almaktadır. İncelenen çalışmalar, regresyon modeli ve coğrafi olarak ağırlıklandırılmış regresyon (GWR) gibi yaklaşımların yanı sıra, yapay sinir ağları (YSA), çoklu regresyon analizi (MRA) ve CBS gibi tekniklerin kullanımını kapsar. Bu çalışmalar kronolojik sıraya göre aşağıda kısaca özetlenmiştir.

Konut ve arazi fiyatları üzerindeki mekansal özelliklerin etkisini analiz etmeyi amaçlayan çalışma iki keşifsel yaklaşıma dayanmaktadır. Bu yaklaşımlar regresyon modelinin artıklarına dayalı bir analiz ve GWR tekniğidir. GWR modeli, regresyon katsayılarının konuma göre farklılık göstermesine izin vererek daha fazla mekansal deseni açığa çıkartır. Bu iki yöntemin karşılaştırılması, her birinin hedonik regresyon modellerini geliştirmede tamamlayıcı roller oynadığını göstermektedir. Genel regresyon modelinin artıklarının analizi, özellikle araziyle ilişkili değişkenlerin belirlenmesinde etkili olurken, GWR yöntemi alanla ilişkili mekansal özelliklerin etkisinin daha iyi anlaşılmasına yardımcı olmuştur. 2005 yılında gerçekleştirilen bu çalışma, mekansal özelliklerin konut ve arazi fiyatları üzerindeki etkisini anlamada hem küresel hem de yerel regresyon tekniklerinin kullanımının önemini vurgulamaktadır. Bu bulgular, hedonik regresyon modellerinin iyileştirilmesi için mekansal analiz tekniklerinin geliştirilmesine yönelik daha fazla araştırmaya yol gösterici olma niteliği taşımaktadır [3].

Konya/Selçuklu Bölgesi'nde gerçekleştirilen bir çalışma, Bosna Hersek Mahallesi'nde yer alan 300 konutu kapsayan bir veri seti üzerinden, YSA ve MRA metodolojilerini kullanarak gayrimenkul değerlemesi için bir model geliştirmeyi ve bu modellerin piyasa değerleriyle karşılaştırmasını amaçlamaktadır. Çalışmada konutların yaşları, kat sayıları, cepheleri, odalarının sayısı gibi konuta ait değişkenlerinin yanında CBS ile hesaplanan sosyal altyapılara olan mesafeleri gibi çeşitli kriterler dikkate alınmıştır. Elde edilen sonuçlar, YSA modelinin, MRA modeline göre gayrimenkul değerlerini daha başarılı bir şekilde tahmin edebildiğini ortaya koymaktadır. Çalışma, YSA ve CBS teknolojilerinin entegrasyonunun, gayrimenkul değerlemesi için güçlü ve etkili bir araç olduğunu göstermektedir. Araştırma sonuçları, YSA modelinin, gayrimenkul değerlemelerinde kullanılmak üzere diğer bölgelerdeki uygulamalar için bir temel oluşturabileceğini ve bu modelin, değerlemeye ihtiyaç duyulan farklı alanlarda kullanılabileceğini belirtmektedir [4].

Ankara ilindeki konut fiyatlarının ve bu fiyatların coğrafi dağılımının CBS yardımıyla nasıl incelenebileceği üzerine yapılan bu çalışmada, Ankara'nın çeşitli ilçelerinde satışa sunulan konutlara ait veriler kullanılarak coğrafi ve sözel verilerin entegrasyonu ve analizi ile konut fiyat tahminleri yapılması hedeflenmiştir.

Çalışma, çoklu regresyon analizi yoluyla konut fiyatlarını etkileyen faktörleri (örneğin, metrekare, bina yaşı, merkezi ısıtma varlığı, asansör, otopark gibi) belirlemiş ve bu faktörlere göre bir fiyat denklemi oluşturmuştur. Analiz sonuçlarına göre, bir dizi özelliğin, özellikle de ankastre mutfak gibi bazı özelliklerin, Ankara ilindeki konut fiyatları üzerinde belirgin bir etkisi olduğu saptanmıştır. Bu araştırma, konut fiyatlarının tahmin edilmesinde CBS'nin ve regresyon analizinin etkin bir şekilde kullanılabileceğini göstermekte ve bu alanda daha fazla araştırma yapılması için bir temel oluşturmaktadır. Elde edilen bulgular, konut alım-satım süreçlerinde karar verme mekanizmalarını destekleyecek şekilde, tematik haritaların oluşturulması ve analiz edilmesi ile kullanıcılara sunulmuştur [5].

Yapılan bir diğer çalışmada, İstanbul konut piyasası'nda hedonik fiyatlama yaklaşımı kullanılarak konut fiyatlarının tahmin edilmesi üzerine odaklanmaktadır. Çalışmanın amacı, mekansal bağımlılığı dikkate alan hedonik konut fiyatlama modeli ile İstanbul'daki konut piyasasının talep yönü hakkında daha fazla bilgi edinmektir. Ekim-Aralık 2013 döneminde, İstanbul'da satışa sunulan 2797 apartman dairesinin özelliklerinin incelenmesiyle, bu özelliklerin konut fiyatları üzerindeki etkisi ortaya konmuştur. Araştırma, olağan en küçük kareler (EKK) modeli ile başlamış ve mekânsal bağımlılığı dikkate alan mekânsal modellerle devam etmiştir. Bulgular, mekânsal Durbin modeli'nin konut fiyatları ve özellikleri arasındaki ilişkiyi daha iyi açıkladığını ve EKK regresyon modeline göre katsayı tahminlerinin güvenilirliği açısından daha etkin sonuçlar verdiğini göstermiştir. Özellikle İstanbul'da apartman dairesi türündeki konutların fiyatlarını en fazla arttıran özellik boğaz manzarası olduğu; fiyatı en fazla düşüren özelliğin ise konutun bodrum katta olması olduğu tespit edilmiştir. Çalışma, İstanbul konut piyasası için kapsamlı bir analiz sunarak, konut talep edenlerin tercihleri ve profilleri hakkında detaylı bilgiler sağlamaktadır. Bu bilgiler, konut politikalarının etkin bir şekilde uygulanması ve konut piyasasındaki fiyatlandırma stratejilerinin geliştirilmesi açısından önem taşımaktadır [6].

D.A.J. Oud Hollanda'nın Alkmaar belediyesinde gerçekleştirdiği çalışma, CBS kullanılarak gayrimenkul değerlemesinin nasıl geliştirilebileceğini inceler. Bu çalışma, geleneksel otomatik değerleme modellerine (AVM) dahil edilmeyen mekansal özelliklerin, özellikle gayrimenkullerin manzaralarının, objektif olarak nasıl değerlendirilebileceğini ve mekansal regresyon teknikleri ile birleştirilerek gayrimenkul değer tahminlerinin doğruluğunun nasıl artırılabileceğini gösterir. Mekansal ve geleneksel regresyon modellerini karşılaştıran araştırma, mekansal değişkenlerin entegrasyonunun, özellikle CBS teknolojileri kullanılarak elde edilen manzara verilerinin, gayrimenkul değer tahminlerini önemli ölçüde iyileştirdiğini

bulmuştur. Sonuçlar, mekansal analizin gayrimenkul değerlemesi uygulamalarında daha fazla kullanılması gerektiğini ve CBS'nin gayrimenkul değerlemesindeki potansiyelini vurgulamaktadır. Bu çalışma, gayrimenkul değerlemesine modern teknolojik yaklaşımların entegrasyonunun önemini ve etkisini ortaya koymaktadır [7].

Bir diğer çalışmada, Türkiye'de ilçe bazında gayrimenkul satışlarını mekansal ve zamansal yöntemlerle inceleyerek, gayrimenkul piyasasının mevcut durumunu ve zaman içindeki önemli değişiklikleri ortaya koymaktır. Çalışma, 2004 ile 2017 yılları arasındaki 14 yıllık gayrimenkul satış verilerini kullanarak, bu verileri CBS ile analiz etmiştir. Bulgular, büyük şehirlerin ve çevre ilçelerinin gayrimenkul piyasalarının canlı olduğunu, doğu ve kuzey bölgelerinde ise gayrimenkul satışlarının daha durağan olduğunu göstermektedir. Ayrıca çalışmada yıllık ortalama gayrimenkul satış artış oranının, yıllık ortalama nüfus artış oranından yaklaşık yedi kat daha yüksek olduğu bulunmuştur. Çalışma, gayrimenkul piyasasının mekansal-zamansal analizine yeni bir bakış açısı sunarak, yatırım planlaması için önemli bilgiler sağlar. Bu tür bir mekansal-zamansal çalışma, gayrimenkul piyasasının performansını ve ekonomik durumu 14 yıllık verilerle sunarak, ülke genelinde yatırım planlaması için öne çıkan bölgeleri vurgular [8].

Niğde kenti için gerçekleştirilen çalışmada taşınmazların değerlendirmesi için çok kriterli karar verme (ÇKKV) yöntemlerinden biri olan analitik hiyerarşi süreci (AHP) ve CBS kullanılarak bir değerleme modeli geliştirilmiştir. Bu model sayesinde, taşınmazların piyasa değerleri ile AHP ve CBS yardımıyla belirlenen değerlerin karşılaştırılması yapılmıştır. Taşınmaz değerlemesinde beş ana kriter (yapı özellikleri, konum, altyapı, ulaşım, nüfus) ve bu kriterlere bağlı 38 alt kriter dikkate alınmıştır. Bu kriterler, taşınmaz değerlemesinde önem dereceleri uzman görüşleriyle belirlenerek ağırlıklandırılmıştır. Sonuç olarak, Niğde kentinde seçilen 30 dairenin AHP ve CBS yöntemleri kullanılarak yapılan değerlemesi, taşınmazların piyasa değerleri ile uyumlu olduğunu göstermiştir. Bu yaklaşım, taşınmaz değerlemesinde objektif ve güvenilir bir yöntem olarak önerilmektedir [9].

2020 yılında Çin'deki konut fiyatlarının tahmin modelini oluşturmak için CBS teknolojisi ve taşıma erişilebilirliğini kullanarak yapılan bir çalışmada, yaya, otobüs ve metro erişilebilirliği olmak üzere üç kentsel toplu taşıma endeksi oluşturulmuştur. Çalışma rastgele orman (*random forest*, RF), gradyan artırıcı karar ağacı (*Gradient Boosting Decision Tree*, GBDT), hafif gradyan artırma makineleri (*Light Gradient Boosting Machine*, LightGBM) algoritmalarını kullanarak konut fiyatlarını tahmin etmek için bir model geliştirir. Araştırmanın bulguları, taşıma erişilebilirliğinin, farklı kentsel alanlardaki toplu taşımanın kolaylığını temsil

ettiğini ve kentsel toplu taşımanın erişilebilirliğinin kentsel konut fiyatları üzerinde önemli bir etkisi olduğunu gösterir. Model, %89,2 tahmin doğruluğu ve 1761,84 kök ortalama kare hatası (RMSE) ile yüksek bir tahmin başarısı sergiler. Bu çalışma, taşıma erişilebilirliğinin kentsel konut ekonomisi üzerindeki rolüne ve toplu taşıma gelişiminin konut ekonomisinde nasıl önemli bir faktör olduğuna dair anlayışı derinleştirmektedir [10].

Bir başka çalışmada, CBS ve makine öğrenmesi yöntemlerinin entegrasyonuyla konut amaçlı taşınmazların toplu değerlemesine odaklanmaktadır. Türkiye ve Birleşik Krallık'taki konutların toplu değerlemesi için geliştirilen hibrit bir yaklaşım sunulmaktadır. Açık lisanslı coğrafi veri kaynakları kullanılarak yapılan yakınlık, yüzey ve görünürlük analizleri sonucunda bir taşınmaz değer haritası oluşturulmuştur. XGBoost, CatBoost, LightGBM ve RF gibi algoritlar kullanılarak fiyat tahmin modelleri geliştirilmiştir. Konumsal kriterler, değerleme verilerine öznitelik zenginleştirme amacıyla dahil edilmiş ve tahmin doğruluğu önemli derecede artırılmıştır. RF modeli, diğer regresyon yöntemlerine göre daha yüksek bir doğruluğa ulaşmıştır. Bu çalışma, taşınmaz değerlemesi alanında CBS ve makine öğrenmesi yöntemlerinin potansiyelini vurgulamakta ve önerilen hibrit yaklaşımın diğer bölgelerde de uygulanabileceğini öngörmektedir [1].

Gerçekleştirilen başka bir çalışmada, İstanbul'daki konut fiyatlarını tahmin etmek için hedonik fiyat modeli (HPM) ve YSA yöntemlerini karşılaştırarak, Eylül 2015 ile Nisan 2019 arasında satılan 900 konut üzerinden bir analiz sunar. Araştırma, YSA modelinin HPM'ye göre konut fiyatlarını daha başarılı bir şekilde tahmin ettiğini ve düşük ortalama mutlak hata (MAE) değerleriyle daha doğru sonuçlar sağladığını bulmuştur. Sonuç olarak, konut fiyat tahminlerinde YSA yönteminin daha etkili olduğu gösterilmiştir [11].

Son olarak, coğrafi verilerin işlenmesi ve RF algoritması kullanılarak konutlar için bir değer tahmin modeli geliştirilmiştir. Konutların ekonomik değerlemesi için 120'den fazla kriter belirlenmiş, bu kriterler üzerinden coğrafi ve yapısal veriler analiz edilmiştir. Model geliştirmede Pearson korelasyon analizi ve kutu grafik (*boxplot*) yaklaşımı ile aykırı veri analizi yapılmıştır. RF yöntemi ile geliştirilen model, çoklu doğrusal regresyon analizi ile karşılaştırılarak test edilmiş ve RF yönteminin daha etkin sonuçlar verdiği tespit edilmiştir. Bu bulgular, makine öğrenmesi tekniklerinin ve coğrafi bilgi sistemlerinin taşınmaz değerleme süreçlerinde önemli faydalar sağlayabileceğini göstermektedir [12].

1.2 Tezin Amacı

Çalışmanın amacı, İstanbul'un çeşitlilik gösteren konut piyasasında, satış fiyatlarının tahmin edilmesi sürecinde CBS ve makine öğrenmesi tekniklerinin etkin bir şekilde entegrasyonunu sağlamaktır. Araştırma kapsamında, konutların stratejik öneme sahip noktalara (örneğin; ulaşım hatları, sahil/deniz, park/yeşil alanlar ve alışveriş merkezleri) olan fiziksel uzaklıkları CBS kullanılarak hesaplanacak ve bu uzaklık verileri, makine öğrenmesi modelinde bağımsız değişkenler olarak kullanılacaktır. Bu entegre yaklaşım, konut fiyatlarının daha doğru ve güvenilir bir şekilde tahmin edilmesine yardımcı olmayı hedeflemektedir. Ayrıca, model tarafından elde edilen tahmin sonuçları ile birlikte, İstanbul'un farklı bölgelerindeki konut fiyat dağılımları detaylı tematik haritalar aracılığıyla görselleştirilecektir. Bu görselleştirme, piyasa analizlerinin ve karar verme süreçlerinin zenginleştirilmesine katkıda bulunacak, coğrafi dağılım ve piyasa dinamiklerine dair derinlemesine anlayış sağlayacaktır. Tez, konut piyasasıyla ilgilenen tüm paydaşların (örneğin; yatırımcılar, geliştiriciler, şehir planlamacıları ve politika yapıcılar) daha bilinçli ve veriye dayalı kararlar almasına olanak tanıyacak metodolojik bir çerçeve sunmayı amaçlamaktadır. Bu bağlamda, CBS ve makine öğrenimi teknolojilerinin birleşimi, İstanbul gibi canlı bir şehirde konut piyasasının daha etkin şekilde kavranması ve yönetilmesi açısından hayati bir rol oynayacaktır.

1.3 Hipotez

Bu çalışmanın hipotezi, İstanbul'daki konut fiyatlarının, konum ve çevresel faktörler gibi coğrafi özelliklerle yakından ilişkili olduğu ve bu özelliklerin makine öğrenmesi modeline entegre edilmesiyle fiyat tahminlerinin önemli ölçüde iyileştirilebileceğidir. CBS aracılığıyla elde edilen uzaklık ve konum verilerinin, konut piyasası analizlerinde kullanılan geleneksel yöntemlere kıyasla daha kapsamlı ve detaylı bilgiler sunarak, konut fiyat tahminlerinin doğruluğunu artıracağı öngörülmektedir. Bu entegre yaklaşımın, piyasadaki coğrafi eğilimleri ve desenleri daha net bir şekilde ortaya çıkaracağı ve böylece konut fiyatlarını etkileyen dinamikleri derinlemesine anlama imkanı sunacağı varsayılmaktadır. Ayrıca, araştırma, makine öğrenmesi ve CBS'nin birleştirilmesinin, konut piyasasının daha iyi anlaşılmasına ve şehir içindeki konut fiyat dağılımının daha detaylı haritalandırılmasına olanak tanıyacağını ortaya koymaktadır. Bu yaklaşımın, İstanbul'daki konut piyasası için stratejik planlama, yatırım kararları ve politika geliştirme süreçlerinde kullanılabilecek değerli ve yenilikçi içgörüler sağlayacağı öngörülmektedir. Sonuç olarak, bu çalışma, konut piyasası analizlerinde coğrafi bilginin entegre edilmesinin, sektöre yönelik daha bilinçli,

verimli ve etkili stratejilerin geliştirilmesine nasıl katkıda bulunabileceğini ortaya koymayı amaçlamaktadır.

2
COĞRAFİ BİLGİ SİSTEMLERİ

CBS, konum bazlı verilerin toplanması, saklanması, işlenmesi, analiz edilmesi ve ilgili kullanıcılara aktarılmasını mümkün kılan bilgisayar tabanlı bir teknolojidir. CBS, arazi özellikleri, yol ağları, nüfus dağılımı ve demografik bilgiler gibi çeşitli kategorileri içerir ve bu bilgileri katmanlar şeklinde düzenler. Bu katmanlar, verilerin analizini ve görselleştirmesini kolaylaştırır [13]. Coğrafi bilgi sistemlerindeki veriler coğrafi bir boyuta sahiptir ve konumsal verilerle birleştirilerek daha kapsamlı analizler yapılmasına olanak tanır. Özetle CBS, mekansal ve sözel verileri bütünleşik bir şekilde sunar. Bu, verilerin sistemli bir biçimde toplanmasını, yönetilmesini ve analiz edilmesini kolaylaştırır [14]. CBS, kullanıcıların coğrafi veriler üzerinde sorgulamalar yapmasına, istatistiksel analizler gerçekleştirmesine, haritalar oluşturmasına ve coğrafi bilgiyi diğer veri türleriyle entegre etmesine olanak tanır.

CBS'nin temel bileşenleri;

- Donanım: Bilgisayarlar ve coğrafi veri toplama cihazları gibi fiziksel ekipmanları içerir.
- Yazılım: Verilerin işlenmesi ve analiz edilmesi için gerekli araçları sağlar.
- Veri: CBS'nin en kritik bileşenidir ve coğrafi konumlarla ilişkili çeşitli bilgileri içerir.
- İnsan: Sistemin etkin bir şekilde kullanılmasını sağlayan eğitimli personeli ifade eder.
- Yöntemler: Verilerin toplanması, saklanması, işlenmesi ve analiz edilmesine yönelik prosedürler ile teknikleri kapsar.

şeklindedir [15].

CBS'de kullanılan veriler konumsaldır ve iki farklı veri tipi vardır. Bunlar; vektör ve raster veri modelleridir. Vektör veriler, coğrafi verilerin konum ile ilgili koordinat değerleri ve bilgilerin saklandığı veri türüdür. Belli koordinatlara sahip olan bu veriler, nokta, çizgi ve alan olmak üzere üç farklı şekilde kategorize edilir [16]. Raster veriler, pikseller (ya da hücreler) kullanılarak ifade edilen bir veri türüdür ve vektör verilere göre farklı bir yapıda bulunur. Bu veriler, komşu hücrelerin bir araya gelmesiyle oluşur ve genellikle vektör verilerin dönüştürülmesiyle elde edilir. Raster veriler, konumsal analizlerin gerçekleştirilmesi açısından vektör verilere kıyasla daha basit ve kullanışlıdır. Ek olarak, verinin depolanması için de uygun bir formata sahiptir. Bu veriler hava fotoğrafları, uydu görüntüleri, taranmış haritalar gibi verilerdir. Hem raster veri hem de vektör veri modelleri, kullanılacak olan özel uygulamalara bağlı olarak değişiklik gösterebilir [17].

CBS, şehir planlama, çevre yönetimi, tarım, afet yönetimi, ulaşım ve sağlık hizmetleri gibi geniş bir yelpazede uygulama alanlarına sahiptir. Örneğin, bir şehirdeki trafik akışını analiz etmek, sel riski yüksek alanları belirlemek, tarım arazilerinin verimliliğini artırmak veya halk sağlığı sorunlarını izlemek için CBS teknolojileri kullanılabilir.

2.1 CBS'de Kullanılan Analiz Türleri

CBS'de temel analizler aşağıdaki şekilde gruplandırılabilir:

- Coğrafi Sorgulama
- Coğrafi Analiz
- Ağ Analizi
- Sayısal Arazi Analizi
- İstatistiksel Analiz
- Izgara (*Grid*) Analizi

2.1.1 Coğrafi sorgulama

CBS, konumla ilişkilendirilmiş grafiksel ve metinsel bilgileri bir araya getirerek, bu bilgiler arasında bağlantılar kurar. Bu sistemlerde, hem grafik (harita, şekil vb.) hem de grafik olmayan (metin, sayısal veri vb.) bilgiler birbiriyle ilişkilendirilerek, çeşitli sorgulamalar yapılabilir. Bu tür sorgulamalar, coğrafi sorgulama olarak adlandırılır ve kullanıcıların grafik bilgiler üzerinden metinsel bilgilere, metinsel

bilgiler üzerinden grafik bilgilere veya metinsel bilgiler üzerinden yine metinsel bilgilere ulaşmasını sağlar [18].

Grafiksel bilgilere dayalı metinsel bilgi sorgulaması, veri tabanındaki bir nesnenin harita, şekil gibi grafiksel verileri üzerinden, o nesneyle ilişkili adres gibi metinsel bilgilere erişim sağlamayı ifade eder. Örneğin, bir binanın harita üzerindeki konumunu kullanarak, o binanın adresine ve orada yaşayan kişilerin listesine ulaşılabilir.

Metinsel bilgilere dayalı grafiksel bilgi sorgulaması ise kullanıcıların belirli metinsel kriterlere göre sorgulama yaparak, bu sorgulamaya uyan nesnelerin grafiksel verilerini (haritalar, şekiller vb.) elde etmelerini sağlar. Bu, örneğin, belirli bir alana sahip arazilerin veya belirli bir sokakta yer alan gayrimenkul ofislerinin haritasını çıkarmak gibi işlemleri içerir.

Metinsel bilgilerden metinsel bilgilere yapılan sorgulama ise daha çok sözel veri üzerinden gerçekleşir. Örneğin, bir adadaki parseller hakkında bilgi almak ve bu parsellerin sahip veya hissedarları hakkında detaylı bilgilere erişmek gibi işlemler bu kategoriye girer. Bu tür sorgulamalar, coğrafi bilgileri derinlemesine analiz etmeyi ve çeşitli bilgiler arasındaki ilişkileri ortaya çıkarmayı mümkün kılar.

2.1.2 Coğrafi Analiz

Coğrafi analiz, CBS'de kritik bir role sahiptir. Bu analizin temel amacı, belirli bir konumdaki varlıkların, diğer konumda bulunan varlıklarla olan ilişkilerini keşfetmek ve çeşitli analizler yapmaktır. Coğrafi analiz üç ana işlemi kapsar: coğrafi birleştirme, yakınlık analizi ve sınır işlemleri [18].

Coğrafi birleştirme, CBS'de yaygın olarak kullanılan bir analiz türüdür. Bu işlem, yeni yerleşim alanlarının, otoyol veya doğal gaz hatları için en uygun güzergâhların belirlenmesi gibi çeşitli planlama faaliyetlerinde temel bir araçtır. CBS'de grafik veriler (nokta, çizgi ve poligon olarak) saklanabilir ve bu veriler, alan tabanlı birleştirme işlemleriyle bir araya getirilebilir. Nokta, çizgi veya alanlar arasında birleştirme yapılarak, çeşitli şeffaf harita katmanlarının birleştirilmesiyle detaylı bilgi haritaları oluşturulabilir. Bu birleşimler, birleştirilen katmanlardaki varlıkların detaylı özelliklerini içerebilir.

Yakınlık analizi, bir konumdan belirli bir mesafe içindeki tüm coğrafi detayların tanımlanmasını gerektirebilir. Bu işlem, bir nokta, çizgi veya alan etrafında tampon bölgeler (*buffer*) oluşturarak, bu bölgeler içinde yer alan varlıkların tespit edilmesi ve analiz edilmesini sağlar. Yakınlık analizi, coğrafi detayların konumlarına göre

değerlendirilmesinde önemli bir yöntemdir.

Coğrafi sınırlar üzerinde gerçekleştirilebilecek işlemler çeşitlidir ve bunlar arasında coğrafi bölme, silme, güncelleştirme, birleştirme ve sınırların kaldırılması yer alır. Bu işlemler, coğrafi alanların nasıl yönetileceğini ve çeşitli coğrafi verilerin nasıl entegre edileceğini belirler.

Bu üç ana coğrafi analiz işlemi, coğrafi bilgi sistemlerinin verimli kullanımı için temel yöntemlerdir ve bölgesel planlama, kaynak yönetimi ve çevresel değerlendirme gibi birçok alanda uygulanabilirler.

2.1.3 Ağ Analizi

Ağ analizleri, fiziksel ağların (yollar, su yolları, elektrik iletim hatları vb.) performansını ve etkileşimlerini değerlendirmek için kullanılır. Bu analizler sayesinde, en kısa yol hesaplamaları, erişilebilirlik analizleri, rota optimizasyonu ve ağ üzerindeki akış analizleri gibi çeşitli sorulara cevaplar üretilebilir. Ağ analizleri, veriye dayalı karar verme süreçlerini destekleyerek, daha etkili planlama, tasarım ve yönetim sağlar. Bu analizler, özellikle büyük ve karmaşık ağlar söz konusu olduğunda, stratejik planlama ve operasyonel verimliliğin artırılmasında kritik bir rol oynar [19].

2.1.4 Sayısal Arazi Analizi

Sayısal arazi analizi, bilgisayar programları kullanarak arazinin şeklini ve özelliklerini incelememizi sağlar. Bu analizle arazi üzerinde hangi bölgelerin daha yüksek veya daha alçak olduğunu, hangi yönlerin güneşi daha fazla aldığını veya hangi alanların gözden uzak kaldığını gibi birçok bilgiyi elde edebiliriz [20]. Bu işlemler, arazinin nasıl kullanılacağına dair karar verirken bize yardımcı olur.

- Eğim Hesabı: Arazinin ne kadar eğimli olduğunu ölçeriz. Bu, örneğin bir yamaçta suyun nereye akacağını veya hangi alanların erozyon riski taşıdığını belirlememize yardımcı olabilir. Ayrıca, çok dik alanlarda inşaat yapmanın uygun olup olmadığına karar vermemize de yardımcı olur.
- Bakı Hesabı: Bir noktadan bakıldığında neyin görülebileceğini tahmin ederiz. Bu, örneğin bir evin pencerelerinden manzaranın nasıl olacağını veya bir yerin gün içinde ne kadar güneş aldığını anlamamıza yardımcı olur.
- Kesit Çıkarma: İki nokta arasındaki arazinin yükseklik profilini çıkarırız. Bu, yolu veya tüneli nereye yapmamız gerektiğini planlamak için önemlidir.

- Görünürlük Analizi: Belli bir noktadan bakıldığında hangi alanların görülebildiğini ve hangilerinin görülemeyeceğini belirleriz. Bu, örneğin bir gözetleme kulesi yerleştirmek veya bir bölgenin mahremiyetini değerlendirmek için kullanılabilir.

Bu araçlar, arazinin nasıl kullanılacağına karar verirken önemli bilgiler sağlar. Örneğin, hangi alanların güneşten en iyi şekilde yararlanabileceğini belirlemek için güneş ışığının düştüğü yerler incelenebilir.

2.1.5 İstatistiksel Analiz

CBS, mekansal veriler üzerinde kapsamlı analizler yapma ve bu verileri yönetme yeteneği sunan güçlü teknolojik araçlardır. CBS kullanılarak gerçekleştirilen istatistiksel analizler, toplam değer belirleme, ortalama hesaplama, en yüksek ve en düşük değerlerin tespiti gibi temel işlemlerin ötesine geçerek, daha detaylı ve derinlemesine analizler sunar [18]. Yüzde hesaplamaları, standart sapma ve varyans gibi ölçümler, veri setlerinin dağılımı ve eğilimleri hakkında daha fazla bilgi sağlar.

CBS ile yapılan çalışmalarda, mekansal interpolasyon gibi ileri düzey teknikler kullanılarak, ölçülen noktalar arasındaki değerler tahmin edilebilir ve bu sayede veri setleri arasında daha geniş bir anlayış geliştirilebilir. Ayrıca, mekansal otokorelasyon analizleri, belirli özelliklerin coğrafi dağılımındaki benzerlikleri veya farklılıkları belirleyerek, mekansal desenlerin ve ilişkilerin anlaşılmasına olanak tanır.

Bu analizlerin sonuçları, kullanıcıların verileri kolayca anlamasını ve yorumlamasını sağlayacak şekilde metin, tablo ve grafiklerle sunulabilir. Özellikle, coğrafi veri görselleştirme, CBS'nin en güçlü yönlerinden biridir. Renkli haritalar, ısı haritaları, çizgi grafikler ve 3D modeller gibi çeşitli görselleştirme teknikleri, mekansal verilerin ve analiz sonuçlarının anlaşılmasını kolaylaştırır. Interaktif haritalar ve görseller, kullanıcılara veriler üzerinde doğrudan etkileşimde bulunma ve belirli alanlarda detaylı incelemeler yapma imkanı sunar.

2.1.6 Izgara (*Grid*) Analizi

CBS içerisinde kullanılan raster tabanlı verilerle gerçekleştirilen analiz işlemleri, temel olarak üç farklı ızgara analiz türünü kapsar: Optimum Koridor Belirleme, Modelleme ve Simülasyon, ve Komşuluk Analizi [18].

- Optimum Koridor Belirleme analizi, iki bölge veya nokta arasında, arazi

eğimi, toprak türü, maliyet gibi faktörlerin incelenerek en uygun geçiş yolunun tespit edilmesi sürecidir. Bu tür bir analiz, örneğin, iki şehir arasında yapılacak yeni bir karayolu için en uygun güzergahın belirlenmesinde kullanılabilir.

- Modelleme ve Simülasyon, belirli coğrafi olayların potansiyel etkilerinin analizi ve bu etkilerin simülasyonu ile ilgilidir. Bu yaklaşım, özellikle doğal afetlerin etkilerinin önceden tahmin edilmesi gibi durumlar için değerlidir. Örneğin, bir tsunami sonrası etkilenebilecek alanların modellemesi, risk yönetimi ve acil müdahale planlamasında kritik öneme sahiptir [20].
- Komşuluk Analizi ise, belirli bir konum veya özelliğe yakın olan diğer özelliklerin belirlenmesiyle ilgilidir. Bu analiz türü, örneğin, bir kıyı şeridine komşu olabilecek rekreasyon alanlarının tespitinde kullanılabilir.

Bu üç ana ızgara analiz yöntemi, CBS'de raster verilerin kullanımının çeşitliliğini ve gücünü göstermektedir. Her biri, coğrafi verileri farklı amaçlar için analiz etmek ve değerlendirmek üzere özelleşmiş teknikler sunar. Bu yöntemler, çevresel planlama, kentsel gelişim, afet yönetimi ve daha pek çok alanda önemli kararlar alınmasına yardımcı olur.

2.2 CBS'nin Uygulama Alanları

CBS, sahip olduğu üstün analiz yetenekleri sayesinde taşınmazlara yönelik olarak, doğal kaynak yönetimi ve doğal kaynakların korunması, ticaret ve işletme planlama, kentsel büyümenin kontrol altına alınması ve sosyo-ekonomik analizler gibi alanlarda sıklıkla kullanılmaktadır [21, 22]. CBS, çok çeşitli alanlarda uygulama bulur ve bu alanlardaki karar verme süreçlerini destekler. Bu alanlardan bazılarının aşağıdaki gibi olduğu söylenebilir [15].

1. Şehir Bilgi Sistemleri: Şehirlerdeki farklı verilerin, çeşitli amaçlar ve kurumlar tarafından toplanıp, bir sistem altında düzenlenerek kolaylıkla erişilebilir ve yönetilebilir hale getirilmesi ile oluşturulan sistemdir. Bu sistem sayesinde, kaçak yapıların tespiti, ruhsatsız işletmelerin kontrolü, emlak ve çevre temizlik vergilerinin toplanması gibi işlemler daha etkili bir şekilde yönetilerek, gelirlerin artırılması hedeflenmektedir.
2. Çevre ve Peyzaj Yönetimi: CBS, çevre kirliliği analizi, çevresel veri tabanının oluşturulması ve gelişmelerin izlenmesi gibi işlemlerde aktif olarak kullanılmaktadır. CBS'nin kullanımı, yeşil alanların, milli parkların, oyun

ve dinlenme alanlarının, ormanların, tarihi alanların ve turistik yerlerin yanı sıra çevrelerindeki yerleşim bölgelerinin planlanmasında, kirlilik tespit ve önlemlerinde büyük katkı sağlar.

3. Doğal Kaynak Yönetimi: CBS, petrol, doğalgaz ve diğer doğal kaynakların araştırılmasında, maden kaynaklarının değerlendirilmesinde etkili bir araç olarak kullanılmaktadır.

4. Turizm Sektörü Yönetimi: CBS teknolojisi, turizm alanlarının planlanması, korunması ve bu bölgelerdeki işletmelerin denetlenmesi işlemlerini kolaylaştırmaktadır.

5. Ulaşım Planlaması: Yeni yol projelerinin uygun lokasyon ve güzergahlarının tespit edilmesi, ulaşım sorunları olan bölgelerde çözümler üretilmesi gibi alanlarda CBS vazgeçilmez bir araçtır.

6. Tarım ve Gıda Güvenliği: CBS, tarımsal üretim tahminleri, hassas tarım uygulamaları, biyolojik çeşitliliğin korunması, hastalık ve zararlı analizleri, tarımsal sigortacılık, kuraklık etkilerinin izlenmesi ve sulama planlaması gibi konularda önemli katkılar sunar.

Coğrafi bilgi sistemlerine gayrimenkul alanında da sıklıkla başvurulmaktadır. Bu alandaki uygulamalar ise aşağıdaki şekilde özetlenebilir [23].

- Konut ararken belirli bir alan etrafında tampon bölge oluşturarak sorgulama yapmak,
- Müşteri profilleri, satış potansiyellerini belirlenerek uygun alanların/konutların tespit edilmesi,
- Bir konutun değerini anlamak için belirli noktalara (Okullar, parklar, toplu taşıma durakları) olan yakınlık veya manzara gibi faktörleri dikkate almak için kolaylık,
- Bir mülkün parsel şeklini analiz ederek her bir parselin alanını ve boyutlarını hesaplama,
- Doğru konut değerleme yaklaşımlarını belirlemek,
- İpotekli veya kredi ile satılmış konutları tematik haritada göstermek,
- Kullanılabilir ticari alanların seçimini kolaylaştırmak,

- Karşılaştırmalı gayrimenkul analizi için mülklerin metrekare başına değerini hesaplamak,
- Ticari gayrimenkulleri yaş, büyüklük, konum ve çevresel faktörler gibi özelliklere göre değerlendirerek mülk değerlerinin piyasa fiyatı ile uyumlu olmasını sağlamak,
- Aktif ulaşım, bisiklet yolları ve yürüyüş yolları gibi sağlıklı yaşam faktörlerini mahalle bazında tespit etmek

2.3 CBS ve İstatistik

CBS ve istatistik, mekânsal verilerin analizinde kritik öneme sahip disiplinlerdir. CBS, coğrafi verilerin toplanması, depolanması, işlenmesi, analiz edilmesi ve görselleştirilmesiyle ilgilenirken, istatistik bu veriler üzerinden anlamlı sonuçlar çıkarılmasını sağlayan matematiksel yöntemler sunar. Bu iki alanın entegrasyonu, mekânsal veri analizinde derinlemesine anlayış ve kapsamlı çözümler sunar [24].

- Mekânsal İstatistik ve Desen Analizi: Mekânsal istatistik, coğrafi konum bilgisini içeren veriler üzerinde yapılan istatistiksel analizlerdir. CBS ortamında mekânsal istatistik kullanılarak, verilerdeki mekânsal desenler, ilişkiler ve anomaliler detaylı bir şekilde incelenebilir [25]. Desen analizi, belirli bir coğrafi alandaki olayların veya olguların dağılımını anlamak için kritik öneme sahiptir. Örneğin, suç desenlerinin veya hastalık salgınlarının coğrafi dağılımı bu yöntemlerle analiz edilebilir.
- Tahmin Modelleri ve Uygulama Alanları: İstatistiksel modeller ve tahmin yöntemleri, CBS veri setleri üzerinde uygulandığında, gelecekteki eğilimleri ve olası senaryoları tahmin etme konusunda güçlü araçlar sunar. Çevresel modeller, hava durumu tahminleri, nüfus artışı projeksiyonları ve suç oranı tahminleri, bu entegrasyonun somut örnekleridir [26].

CBS ve istatistiğin birleşimi, çevresel bilimler, kamu sağlığı, şehir ve bölge planlaması, suç analizi gibi daha pek çok alanda değerli içgörüler sağlar. Özellikle çevresel araştırmalarda, bu entegrasyon su kaynaklarının yönetimi, kirlilik izleme ve habitat değişikliklerinin izlenmesi gibi konularda önemli katkılar sağlar [27] .

2.4 Konut Fiyat Tahminlenmesinde CBS'nin Kullanımı

CBS, konut satış trendlerinin analizi ve tahmin edilmesinde yenilikçi bir bakış açısı sunarak, konut sektöründe değerli bilgiler sağlamaktadır. Konuma dayalı bir sistem

olan CBS, mekana ilişkin coğrafi verileri toplama, kullanma, analiz etme, depolama işlemlerini destekleyen donanım ve yazılım altyapısına sahiptir.Bu özellikleri sayesinde, konut değerlemesi için doğru, tarafsız ve güvenilir kaynaklardan elde edilen verilerin yönetimi kolaylaşır. Sistemsel bir veri tabanında depolama imkanı sunarak değerleme uygulamaları için büyük bir önem taşır [28].

CBS teknolojisi, konut satış fiyatları, kiralık gayrimenkuller ve pazar dinamikleri hakkında derinlemesine bilgiler sağlar. Coğrafi konum, altyapı kalitesi, ulaşım bağlantıları ve çevresel faktörler gibi verilerin analizi yapılır. Bu analizler, konut satış fiyatları ve pazar trendleri konusunda detaylı bir anlayış sunar. Yatırımcılar ve geliştiriciler bu bilgileri kullanarak daha bilinçli kararlar alabilirler. Örneğin, CBS bir bölgenin konut satış fiyatları üzerindeki potansiyel etkilerini değerlendirerek yatırımcılara ve geliştiricilere önemli fırsatlar sunabilir.

Aynı zamanda, CBS gayrimenkullerle ilgili birçok işlemde analiz aracı olarak kullanılabilir. Yatırımcılar, yatırım yapmak istedikleri alanlardaki satılık ya da yapım aşamasındaki tüm arsa ve binaları görüntüleyebilirler. Alıcının kriterlerine bağlı olarak gerçekleştirilen sorgulamalarla, potansiyel olarak uygun arsalar veya binalar belirlenebilir. Bu bilgiler, arsa ve binaların şehir merkezine, parklara, postanelere, hastanelere ve diğer önemli yerlere olan uzaklığını hemen gösterir. CBS, gayrimenkul satış ve kira fiyatlarının coğrafi dağılımını analiz ederek, nerelerde değerlerin düşük veya yüksek olduğunu belirler. Ticari amaçla alınacak veya kiralanacak iş yerleri için yapılan sorgulamalarla, müşterilere alternatif yerler sunulabilir, böylece emlak piyasasına ilişkin kapsamlı ve anlayışlı bir görünüm sağlanır [18].

CBS teknolojisinin kullanımı, mekânsal verilerin konut satış tahminleri üzerindeki etkisinin daha iyi anlaşılmasını sağlar. Bu teknoloji, emlak piyasası analizlerinde ve tahminlerinde, konum temelli derinlemesine analizler yaparak, yatırımcıların ve geliştiricilerin piyasa dinamiklerini daha iyi anlamalarına olanak tanır. Bu yenilikçi yaklaşım, konut sektöründe karar verme süreçlerini önemli ölçüde iyileştirir ve pazarın geleceğine dair daha kesin tahminler yapılmasına yardımcı olur.

3
MAKİNE ÖĞRENMESİ

Makine öğrenmesi, büyük veri kümelerinden bilgi çıkarmayı kolaylaştıran algoritmaların kullanımı üzerine kurulu bir bilim alanıdır [29]. Bu alan, bilgisayarların açıkça programlanmadan belirli görevleri yerine getirebilmesini sağlamak için algoritmaların geliştirilmesi ve değerlendirilmesi üzerine odaklanır [30]. Makine öğrenmesi, mevcut verileri analiz ederek örüntü tanıma, sınıflandırma ve tahmin yapma süreçlerinde kullanılacak etkili özelliklerin seçilmesine olanak tanıyan modellere dayalı bir yaklaşımdır [31]. Bu tanımlar ışığında, makine öğrenimi, veriye dayalı modelleme kullanarak algoritmik çıktılar üreten bir veri analizi metodudur. Arthur Samuel'in klasik tanımıyla, "Makine öğrenmesi, bilgisayarların açıkça programlanmadan öğrenme yeteneği kazanmasıdır" [32].

3.1 Makine Öğrenmesinin Temelleri

Makine öğrenmesi, matematiksel ve istatistiksel modellerin uygulanmasıyla verilerden anlamlı bilgilere ulaşmayı amaçlar. Bu süreç, büyük veri kümelerinin analiz edilmesi ve bu verilerden öğrenilen bilgilerin yeni veri kümelerine uygulanmasıyla gerçekleşir. Öğrenme süreci genellikle üç aşamadan oluşur:

1. Model Eğitimi: Bu aşamada, algoritma belirli bir veri kümesi üzerinde eğitilir. Eğitim sürecinde, algoritma veri kümesindeki kalıpları ve ilişkileri öğrenir. Örneğin, konut fiyatlarını tahmin etmek için kullanılan bir model, tarihsel satış verilerini analiz eder. Bu aşamada model, geçmiş verilere dayalı olarak konut fiyatlarını tahmin etmeyi öğrenir. Verilerdeki bağımlılıkları ve örüntüleri keşfeder ve bu bilgiyi yeni verilere uygulamak üzere saklar.

2. Model Doğrulama: Eğitilen modelin performansını değerlendirmek için kullanılır. Bu aşamada, modelin doğruluğu ve genelleme yeteneği test edilir. Bu, genellikle çapraz doğrulama teknikleri kullanılarak yapılır. Çapraz doğrulama, veri setini birkaç alt sete bölerek her alt set için

modelin performansını ölçer ve bu şekilde modelin genelleme yeteneğini değerlendirmiş oluruz.

3. Model Testi: Eğitilen modelin performansını değerlendirmek için kullanılır. Bu aşamada, modelin doğruluğu ve genelleme yeteneği test edilir. Bu, genellikle çapraz doğrulama teknikleri kullanılarak yapılır. Çapraz doğrulama, veri setini birkaç alt sete bölerek her alt set için modelin performansını ölçer ve bu şekilde modelin genelleme yeteneğini değerlendirmiş oluruz.

3.2 Makine Öğrenmesi Yöntemleri

Makine öğrenmesi yöntemleri öğrenme şekline göre; denetimli, denetimsiz, yarı denetimli ve pekiştirmeli olarak 4 sınıfa ayrılmaktadır.



Şekil 3.1 Makine öğrenmesi yöntemleri

3.2.1 Denetimli Öğrenme

Denetimli öğrenme, makine öğrenmesinin temel yöntemlerinden biridir ve modelin, giriş (özellikler) ile bunlara karşılık gelen çıktılar (etiketler) arasındaki ilişkiyi öğrenmesi amacıyla tasarlanmıştır. Bu yöntemde, her veri noktası bir etiketle ilişkilendirilir, böylece model, verilen girişlere dayanarak çıktıları tahmin edebilmek için bu ilişkiyi öğrenir. Örneğin, bir evin özelliklerini (boyutu, konumu, odaların sayısı gibi) girdi olarak alıp, evin fiyatını tahmin etmeye çalışan bir regresyon modeli denetimli öğrenme ile eğitilebilir.

Sınıflandırma ve regresyon, denetimli öğrenmenin iki ana dalıdır.

- Sınıflandırma, makine öğrenimi alanında, gözlemlenen verilerin kategorilere ayrılması işlemidir. Bu süreçte, model önceden tanımlanmış kategorilere dayanarak yeni verilerin hangi kategoriye ait olduğunu belirler [33]. Konut fiyatlarıyla ilgili bir örnekte, bir şehirdeki farklı mahallelerdeki evlerin özelliklerine (örneğin, büyüklük, oda sayısı, mahallenin sosyoekonomik durumu gibi bağımsız X değişkenleri) dayanarak, bu evleri "yüksek fiyatlı" veya "düşük fiyatlı" (bağımlı Y değişkeni) olarak sınıflandırmak bu işlemin bir uygulaması olacaktır. Bu süreç, evlerin özelliklerine göre fiyat aralıklarını öğrenmek için geçmiş verilerden yararlanır ve böylece model, verilen özelliklere sahip yeni evlerin fiyat kategorisini tahmin edebilir.
- Regresyon, veriler arasındaki sayısal ilişkileri analiz eden bir modelleme yöntemidir [34]. Bu yaklaşım, bir dizi bağımsız X değişkeni kullanarak, bir bağımlı Y değişkeninin değerini tahmin etmeye çalışır. Sınıflandırmadan temel farkı, kategorik değil, sürekli sonuçlar üretmesidir. Örneğin, bir şehrin çeşitli mahallelerindeki evlerin özelliklerini (bağımsız X değişkenleri - X1: Alanı, X2: Yapım Yılı, X3: Oda Sayısı, X4: Konumu, Xn: Yakın Altyapı Olanakları gibi) dikkate alarak, bu evlerin fiyatlarını (bağımlı Y değişkeni) tahmin etmek, regresyon analizinin bir uygulamasıdır. Bu yöntem, özellikle konut fiyatlarını etkileyen faktörlerin hangi derecede önemli olduğunu belirlemeye ve belirli özelliklere sahip bir evin piyasa değerini öngörmeye yardımcı olur.

Bu öğrenme türünün başarısı büyük ölçüde doğru ve kapsamlı etiketlenmiş veri kümelerine bağlıdır. Yüksek kaliteli veri kümeleri, modelin gerçek dünya koşullarını daha iyi genellemesine ve dolayısıyla daha doğru tahminler yapmasına olanak tanır.

Literatürde sıklıkla kullanılan makine öğrenme algoritmalarından bazıları aşağıda sıralanmıştır.

- k En Yakın Komşu (*K-Nearest Neighbors*, KNN)
- Destek Vektör Makinaları (*Support Vector Machine*, SVM)
- Yapay Sinir Ağları (*Artificial Neural Network*, ANN)
- Rastgele Orman (*Random Forest*)
- Doğrusal Regresyon (*Linear Regression*)
- Lojistik Regresyon (*Logistic Regression*)

- Karar Ağaçları (*Decision Trees*, DTs)
- Naive Bayes

3.2.2 Denetimsiz Öğrenme

Denetimsiz öğrenme, makine öğrenimi modellerinin etiketlenmemiş veri kümeleri üzerinde gizli düzenleri ve ilişkileri ortaya çıkarmak üzere çalıştığı bir süreçtir. Bu yaklaşım, veri kümeleri içerisindeki doğal yapıları keşfetmek için modellerin kullanılmasını içerir, böylece sistemler etiket veya sınıflandırma gerektirmeyen veriler arasında bağlantılar kurabilir. Bu süreçte, algoritmalar veri kümesini analiz eder ve herhangi bir dış müdahale olmaksızın veriler arasındaki ilişkileri belirlemeye çalışır [35].

Denetimsiz öğrenmede kullanılan başlıca metotlar şunlardır:

- Kümeleme Yöntemi
- Temel Bileşenler Analizi (PCA)

Kümeleme, denetimsiz öğrenmede verilerin benzerliklerine göre gruplandığı temel bir yöntemdir. Bu teknik, pazar segmentasyonundan görüntü analizine kadar çeşitli uygulamalarda geniş bir kullanım alanına sahiptir. Yaygın olarak K-Ortalamalar kümeleme ve Hiyerarşik kümeleme yöntemleri kullanılmaktadır. K-Ortalamalar kümeleme, en basit ve en popüler kümeleme algoritmalarından biridir. Verileri, özellik benzerliğine göre K adet farklı kümeye ayırır. K-Ortalamalar, kümelerin içindeki kareler toplamını en aza indirmeyi hedefler, bu da kümeleri mümkün olduğunca kompakt hale getirir. Ancak, K-Ortalamalar kümelerin dışbükey olduğunu varsayar ve uzatılmış veya düzensiz şekilli kümelerle başa çıkmakta zorlanabilir. Ayrıca, seçilen başlangıç merkezlerine bağlı olarak yerel minimumda sıkışabilir ve K değerinin doğru seçimi sonuçları önemli ölçüde etkiler. Hiyerarşik kümeleme ise veri noktalarını temsil eden ağaç benzeri bir yapı (dendrogram) oluşturur. Kümeler arasındaki yakınlık, Öklidyen mesafe, Manhattan mesafesi veya Mahalanobis mesafesi gibi çeşitli metriklerle ölçülebilir. Hiyerarşik kümeleme, küme sayısını önceden belirlemeyi gerektirmez ve sonuçlar dendrogramlar aracılığıyla daha anlaşılabilir olur. Ancak, büyük veri kümeleri için hesaplama yoğun olabilir, çünkü zaman karmaşıklığı dörtgen oranda artar.

Temel Bileşenler Analizi (PCA), bir veri kümesindeki değişkenlerin sayısını azaltırken orijinal varyansın çoğunu korumayı amaçlayan bir boyut indirgeme tekniğidir. PCA, orijinal değişkenleri, ana bileşenler adı verilen ve yakalanan

varyansa göre sıralanan yeni, ilişkisiz değişkenler setine dönüştürür. PCA, yüksek boyutlu verileri görselleştirmek ve gürültüyü azaltmak için kullanışlıdır.

3.2.3 Yarı Denetimli Öğrenme

Yarı denetimli öğrenme (*semi-supervised learning*) algoritmaları, denetimli ve denetimsiz öğrenme metodlarının avantajlarını bir araya getiren bir yaklaşımdır. Bu teknik, hem etiketli hem de etiketlenmemiş verileri kullanarak makine öğrenimi ve veri madenciliği modellerini geliştirir. Etiketli verilerin sınırlı olduğu durumlarda, yarı denetimli öğrenme, etiketlenmemiş verileri de dikkate alarak modellerin daha geniş bir veri kümesinden öğrenmesini ve genellemesini sağlar. Bu yaklaşım, özellikle etiketleme maliyetinin yüksek olduğu veya yeterli etiketli veriye ulaşmanın zor olduğu durumlar için tercih edilir. Bu yöntem, etiketli verilerle ön eğitim yapıp daha sonra bu bilgiyi etiketlenmemiş veriler üzerinde uygulayarak modelin performansını artırır [36].

3.2.4 Pekiştirmeli Öğrenme

Pekiştirmeli öğrenme (*Reinforcement Learning*, RL), karar alma bilimini ifade eder. Bir ortam içerisinde maksimum kazanım sağlamak amacıyla en uygun eylemleri öğrenme sürecini kapsar. Bu en uygun eylemler, çevrelerini keşfederek belirli bir amaca ulaşmalarını sağlayacak davranışları öğrenen çocuklar gibi, çevre ile olan etkileşimleri ve bu etkileşimlere verilen tepkiler üzerinden öğrenilir.

RL, denetimli öğrenmeden birkaç önemli yönden farklılık gösterir; en belirgin fark, girdi-çıktı çiftlerinin sunulmamasıdır. Bunun yerine, bir eylem seçtikten sonra aracı, hemen alınan ödül ve sonraki durum hakkında bilgilendirilir ancak uzun vadeli çıkarları açısından hangi eylemin en iyisi olacağı söylenmez. Bu, ajanın optimal şekilde hareket etmek için aktif bir şekilde deneyim toplamasını gerektirir. RL'de çevrimiçi performans önemlidir, çünkü sistem genellikle öğrenme ile eşzamanlı olarak değerlendirilir [37].

3.3 Makine Öğrenmesinin Uygulama Alanları

Makine öğrenmesi, çeşitli endüstrilerde ve alanlarda geniş bir uygulama yelpazesine sahiptir. Bu uygulama alanları, makine öğrenmesinin ne kadar geniş ve etkili bir şekilde kullanılabileceğini göstermektedir. Bazı önemli uygulama alanları aşağıda listelenmiştir:

- Finans: Makine öğrenmesi, kredi skorlama, dolandırıcılık tespiti, portföy

optimizasyonu ve algoritmik ticaret gibi finansal hizmetlerde kullanılır. Örneğin, bankalar, müşteri kredi riskini değerlendirmek için makine öğrenmesi modelleri kullanır. Ayrıca, finansal piyasaların analizinde ve tahmininde de önemli rol oynar. Makine öğrenmesi, algoritmik ticarette, yatırım stratejilerini optimize etmek ve piyasadaki anormallikleri tespit etmek için kullanılır. Dolandırıcılık tespiti konusunda ise, büyük veri analitiği ve derin öğrenme yöntemleri sayesinde finansal kurumlar anormal işlemleri hızlıca tespit edebilirler [38].

- Sağlık: Sağlık sektöründe, makine öğrenmesi, hastalık teşhisi, genetik analizler ve hasta izleme gibi uygulamalarda kullanılır. Örneğin, kanser teşhisinde, tıbbi görüntüleme verilerini analiz ederek anormallikleri tespit eden derin öğrenme algoritmaları kullanılır. Bu tür uygulamalar, hastalıkların erken teşhis edilmesine ve tedavi süreçlerinin iyileştirilmesine katkı sağlar. Ayrıca, genetik verilerin analizinde de makine öğrenmesi, hastalık risklerini tahmin etmek ve kişiselleştirilmiş tedavi planları oluşturmak için kullanılır [39].

- Pazarlama: Müşteri segmentasyonu, tahminsel analizler ve kişiselleştirilmiş öneriler gibi pazarlama stratejilerinde makine öğrenmesi kullanılır. Örneğin, e-ticaret siteleri, kullanıcıların önceki alışveriş davranışlarına dayanarak kişiselleştirilmiş ürün önerileri sunar. Makine öğrenmesi, müşteri davranışlarını analiz ederek daha hedeflenmiş ve etkili pazarlama kampanyaları oluşturulmasına olanak tanır. Ayrıca, müşteri memnuniyetini artırmak ve müşteri kaybını azaltmak için de makine öğrenmesi yöntemleri kullanılır [40].

- Doğal Dil İşleme (NLP): Metin analizi, dil çevirisi, sesli komutlar ve sohbet botları gibi uygulamalarda doğal dil işleme teknikleri kullanılır. Örneğin, dil modelleri, büyük metin verilerini analiz ederek belirli bir konudaki bilgiyi özetleyebilir ve kullanıcılara hızlı bilgi sağlar. Sohbet botları, müşteri hizmetlerinde yaygın olarak kullanılmakta olup, kullanıcılarla doğal ve etkili iletişim kurarak müşteri deneyimini iyileştirir [41].

- Görüntü Tanıma: Makine öğrenmesi, yüz tanıma, nesne tanıma ve görüntü sınıflandırma gibi uygulamalarda yaygın olarak kullanılır. Örneğin, sosyal medya platformları, kullanıcıların fotoğraflarını etiketlemek için yüz tanıma teknolojisini kullanır. Güvenlik alanında, kameralardan alınan görüntüleri analiz ederek anormallikleri tespit eden sistemler geliştirilmiştir. Bu tür uygulamalar, güvenlik ihlallerini önlemek ve suçların hızlı bir şekilde tespit edilmesine yardımcı olur [42].

- Otonom Araçlar: Otonom araçların navigasyonu ve çevre algısı, makine öğrenmesi algoritmaları tarafından desteklenir. Bu araçlar, çevrelerindeki nesneleri tanımlayarak güvenli bir şekilde hareket edebilirler. Sensörlerden alınan verileri işleyerek, trafik durumunu analiz edebilir ve en güvenli güzergahı belirleyebilirler. Bu tür teknolojiler, trafik kazalarını azaltmak ve ulaşım verimliliğini artırmak için büyük bir potansiyele sahiptir [43].

3.4 Konut Fiyat Tahminlemesinde Makine Öğreniminin Kullanımı

Konut fiyatlarının doğru bir şekilde tahmin edilmesi, gayrimenkul sektörü için kritik bir öneme sahiptir. Geleneksel yöntemler, konut fiyatlarını belirleyen karmaşık faktörleri ve dinamikleri tam anlamıyla yansıtmakta yetersiz kalabilir. Bu noktada, makine öğrenimi teknikleri, gelişmiş analiz yetenekleri ve yüksek doğruluk oranları ile önemli avantajlar sunmaktadır.

Makine öğrenimi, geniş veri kümeleri üzerinden öğrenme yaparak geleceğe yönelik tahminlerde bulunabilen algoritmalar bütünüdür. Konut fiyat tahminlemesinde, makine öğrenimi çeşitli aşamalarda kullanılabilir:

1. Veri Toplama ve Ön İşleme: Makine öğrenimi modelleri, konut fiyatlarını etkileyen çok sayıda değişkeni (konum, büyüklük, oda sayısı, yaş, mahalle vb.) kullanarak tahmin yapar. Bu nedenle, verilerin doğru bir şekilde toplanması ve eksik veri hatalarının giderilmesi gereklidir [44].

2. Özellik Mühendisliği: Modellerin performansını artırmak için verilerden anlamlı özellikler çıkarma sürecidir. Örneğin, bir konutun yakınındaki okulların kalitesi veya toplu taşıma erişimi gibi özellikler, fiyatı etkileyen önemli faktörler olabilir.

3. Model Seçimi ve Eğitim: Makine öğrenimi tekniklerinden hangisinin kullanılacağına karar verilmesi ve bu modelin veri kümesi üzerinde eğitilmesi sürecidir [45].

4. Model Değerlendirme: Modelin performansını değerlendirmek için doğruluk, hata oranı ve diğer performans metrikleri kullanılır. En iyi performansı gösteren model, tahminlerde kullanılır.

5. Tahmin ve Yorumlama: Eğitilen model, yeni veri üzerinde tahmin yapar. Bu tahminler, yatırım kararları, fiyatlandırma stratejileri ve piyasa analizleri için kullanılabilir [46].

Makine öğreniminin konut fiyat tahminlemesinde kullanılmasının birçok faydası bulunmaktadır:

1. Yüksek Doğruluk: Makine öğrenimi algoritmaları, verilerdeki karmaşık ilişkileri ve desenleri tanımlamada oldukça başarılıdır. Bu sayede, konut fiyatlarının daha doğru bir şekilde tahmin edilmesini sağlar [47].
2. Özelleştirilebilirlik: Modeller, spesifik veri kümelerine ve problemlere göre özelleştirilebilir. Bu esneklik, farklı bölgeler ve pazarlarda etkili tahminler yapılmasına olanak tanır [45].
3. Hız ve Verimlilik: Gelişmiş makine öğrenimi algoritmaları, büyük veri kümeleri üzerinde hızlı ve verimli bir şekilde çalışabilir. Bu, geniş çaplı analizlerin ve gerçek zamanlı tahminlerin yapılmasına imkan tanır [48].
4. Gelişmiş Analiz Yeteneği: Makine öğrenimi teknikleri, veri kümesindeki önemli faktörleri ve trendleri ortaya çıkararak daha derinlemesine analiz yapmayı mümkün kılar. Bu, piyasa dinamiklerini daha iyi anlamayı ve stratejik kararlar almayı kolaylaştırır.
5. Uyarlanabilirlik: Makine öğrenimi modelleri, zamanla değişen piyasa koşullarına ve yeni verilere kolayca uyum sağlayabilir. Bu, modellerin sürekli olarak güncellenmesini ve performanslarının korunmasını sağlar [45].

Makine öğrenimi teknikleri, konut fiyat tahminlemesinde sundukları yüksek doğruluk, esneklik ve verimlilik ile geleneksel yöntemlerin ötesinde avantajlar sunmaktadır. Bu tekniklerin kullanımı, gayrimenkul sektöründe daha bilinçli ve stratejik kararlar alınmasına katkı sağlar.

3.5 Uygulamada Kullanılan Makine Öğrenmesi Algoritmaları

Bu çalışmada, konut satış fiyatlarını tahmin etmek amacıyla denetimli öğrenme yöntemlerinden XGBoost, LightGBM ve CatBoost olmak üzere üç farklı makine öğrenmesi algoritması kullanılmış ve performansları ayrıntılı bir şekilde tartışılmıştır. Bu üç algoritmanın seçilme nedeni, performans ve verimlilik açısından üstün özellikler sunmalarıdır.

XGBoost, yüksek performansı ve verimli bellek yönetimi ile büyük veri kümelerinde hızlı çalışmasıyla bilinir. Ayrıca, eksik verilerle çalışabilme yeteneği sayesinde herhangi bir veri tamamlama (*imputation*) yöntemine gerek kalmadan veri analizini mümkün kılar. LightGBM, çok büyük veri kümeleriyle başa çıkma

yeteneği sayesinde tercih edilir ve histogram tabanlı algoritması olmasından dolayı çok daha hızlı çalışır. CatBoost ise kategorik verilerle doğrudan çalışabilme kapasitesi ve bu verileri ön işleme gereksinimini azaltması ile öne çıkar.

Boosting algoritmalarının avantajları arasında genel performans ve kullanım kolaylığı öne çıkar. XGBoost, aşırı uyum (*overfitting*) sorunlarını azaltma yeteneği ile bilinirken, LightGBM, büyük veri kümelerinde hızlı ve bellek kullanımını optimize eden bir performans sergiler. CatBoost ise kategorik verilerle doğrudan çalışarak eğitim sürecinde veri dengesizliği ve önyargıyı azaltır.

Paralel işleme ve dağıtık hesaplama destekleri sayesinde XGBoost ve LightGBM, büyük veri kümeleri üzerinde hızlı model eğitimi sağlarlar. CatBoost ise model açıklanabilirliğini artırmak için çeşitli araçlar sunar. Bu üç algoritmanın yüksek performans, esneklik ve çeşitli veri türleri ile uyumlu çalışma yetenekleri, konut satış fiyatlarını tahmin etme gibi karmaşık bir problem için ideal seçimler olmasını sağlar.

3.5.1 XGBoost Algoritması

Aşırı Gradyan Artırma (*Extreme Gradient Boosting*) yani XGBoost, gradyan artırma algoritmalarının verimli ve ölçeklenebilir bir uygulamasıdır. Tianqi Chen ve Carlos Guestrin tarafından 2016 yılında geliştirilen bu algoritma, makine öğrenimi ve veri bilimi alanlarında geniş bir kullanım alanı bulmuştur. XGBoost, yüksek performansı, hızı ve esnekliği ile bilinir ve çeşitli veri bilim yarışmalarında ve endüstriyel uygulamalarda tercih edilen bir araç haline gelmiştir [47].

XGBoost'un temel prensibi, bir dizi karar ağacını ardışık olarak eğiterek hataları minimize etmektir. Her ağaç, önceki ağaçların yaptığı hataları düzeltmek üzere eğitilir. Bu süreç, başlangıçta bir tahmin ile başlar ve ardından gerçek değerler ile tahminler arasındaki farklar hesaplanarak devam eder. Yeni bir karar ağacı, bu farkları tahmin etmek için eğitilir ve bu ağacın tahminleri önceki tahminlere eklenir. Bu adımlar, belirli bir ağaç sayısına ulaşılana veya hatalar minimize edilene kadar tekrarlanır.

XGBoost, performansını artırmak için birkaç yenilikçi teknik kullanır. Düzenleme (*regularization*) teknikleri, modelin karmaşıklığını cezalandırarak aşırı uyumu önler ve modelin genelleme yeteneğini artırır. Ağırlıklı çeyrek çizelgesi (*weighted quantile sketch*), veri ağırlıklarını etkin bir şekilde kullanarak bölme bulma işlemlerinde hassasiyet sağlar. Seyrek verilerle çalışma (*sparsity-aware split finding*) tekniği ise eksik değerler gibi seyrek verilerle etkin bir şekilde başa çıkar [47].

XGBoost'un avantajları arasında yüksek performans, ölçeklenebilirlik ve esneklik bulunur. Düzenleme teknikleri sayesinde çeşitli veri kümelerinde güvenilir sonuçlar verir ve aşırı uyumu önler. Hız ve ölçeklenebilirlik açısından optimize edilmiş olan XGBoost, büyük veri kümelerini işleyebilir ve disk üstü hesaplama (*out-of-core computing*) gibi teknikler sayesinde belleğe sığmayan verilerle bile çalışabilir. Ayrıca, regresyon, sınıflandırma ve sıralama gibi çeşitli amaç fonksiyonlarını destekler, bu da onu geniş bir yelpazede makine öğrenimi görevleri için uygun hale getirir. Paralel ve dağıtık hesaplama yetenekleri, XGBoost'un birden fazla çekirdekte ve dağıtık küme yapılarında paralel olarak çalışabilmesini sağlar, bu da ölçeklenebilirliğini ve performansını artırır [47] [49].

XGBoost'un performansında hiperparametre ayarlarının büyük bir rolü vardır. Bu hiperparametrelerin dikkatlice ayarlanması, model performansını önemli ölçüde artırabilir. Önemli hiperparametreler şunlardır:

- Öğrenme hızı ($learning\ rate$)
- Ağaç sayısı ($n_{estimators}$)
- Maksimum derinlik (max_depth)
- Yaprak düğüm ağırlığı (min_child_weight)
- Gama (γ)
- Alt örnekleme oranı ($subsample$)
- Ağaç başına örnekleme oranı ($colsample_bytree$)

Öğrenme hızı, her ağacın katkısının ne kadar olacağını belirlerken, ağaç sayısı ($n_{estimators}$) toplam ağaç sayısını belirler. Maksimum derinlik, her ağacın derinliğini kontrol eder ve modelin karmaşıklığını düzenler. Yaprak düğüm ağırlığı (min_child_weight), bir yaprak düğümde bulunması gereken minimum toplam ağırlığı belirlerken, gama (γ) bir düğümün bölünmesinin gerektirdiği minimum kayıp azaltma miktarını ifade eder. Alt örnekleme oranı ($subsample$), her ağacın eğitiminde kullanılacak veri oranını belirler ve ağaç başına örnekleme oranı ($colsample_bytree$) her ağacın oluşturulmasında kullanılacak özelliklerin oranını kontrol eder. Bu hiperparametrelerin optimize edilmesi, modelin aşırı uyum yapmasını engelleyebilir ve genelleme yeteneğini artırabilir [47] [49].

XGBoost, sağlık, finans ve perakende gibi birçok farklı alanda uygulanmaktadır. Sağlık alanında, hastalık sonuçlarının tahmin edilmesi ve hasta tekrar yatışlarının

öngörülmesi gibi öngörücü modelleme uygulamalarında kullanılır. Finans sektöründe, kredi skorlama, sahtekarlık tespiti ve algoritmik ticaret gibi alanlarda uygulanır. Perakende sektöründe ise satış tahmini, envanter yönetimi ve müşteri segmentasyonu gibi alanlarda kullanılmaktadır [47] [49].

3.5.2 LightGBM Algoritması

LightGBM, Microsoft tarafından geliştirilmiş olan ve gradyen arttırmanın verimli ve ölçeklenebilir bir uygulamasıdır. Guolin Ke ve ekibi tarafından 2017 yılında tanıtılan bu algoritma, makine öğrenimi ve veri bilimi alanlarında geniş bir kullanım alanı bulmuştur. LightGBM, hızlı eğitim süresi, düşük bellek kullanımı ve büyük veri kümeleri ile başa çıkabilme yeteneği ile bilinir. Bu algoritma, genellikle çeşitli veri bilim yarışmalarında ve endüstriyel uygulamalarda tercih edilen bir araç haline gelmiştir [48].

LightGBM'nin temel prensibi, gradyen arttırma karar ağaçları (GBDT) oluştururken, eğitim süresini ve bellek kullanımını optimize etmektir. Bu algoritma, veri kümelerini histogram tabanlı yöntemler kullanarak böler. Bu sayede büyük veri kümelerinde bile hızlı ve verimli bir eğitim süreci sağlar. LightGBM, aynı zamanda paralel ve dağıtık hesaplama yetenekleri ile büyük veri kümelerini işlemek için idealdir [48].

LightGBM, performansını artırmak için çeşitli yenilikçi teknikler kullanır. Bu tekniklerden biri, yaprak temelli büyüme (*leaf-wise growth*) yöntemidir. Bu yöntem, her iterasyonda en büyük kayıp azalmasını sağlayan yaprağı seçerek ağacın daha derin dallarının oluşturulmasını sağlar. Diğer bir teknik ise histogram tabanlı karar kurallarıdır. Bu yaklaşım, bellek kullanımını optimize ederek ve hesaplama süresini azaltarak, özelliklerin diskretize edilmesini ve histogramların kullanılmasını içerir [48].

LightGBM'nin avantajları arasında yüksek hız, düşük bellek kullanımı ve büyük veri kümeleri ile başa çıkabilme yeteneği bulunur. Ayrıca, modelin doğruluğunu artıran ve aşırı uyumu önleyen çeşitli düzenleme teknikleri de sunar. LightGBM, regresyon, sınıflandırma ve sıralama gibi çeşitli amaç fonksiyonlarını destekler, bu da onu geniş bir yelpazede makine öğrenimi görevleri için uygun hale getirir. Paralel ve dağıtık hesaplama yetenekleri, LightGBM'nin birden fazla çekirdekte ve dağıtık küme yapılarında paralel olarak çalışabilmesini sağlar, bu da ölçeklenebilirliğini ve performansını artırır [48] [50] [45] .

LightGBM'nin başarısında hiperparametre ayarlarının büyük bir rolü vardır. Bu hiperparametrelerin dikkatlice ayarlanması, model performansını önemli ölçüde

artırabilir. Önemli hiperparametreler şunlardır:

- Öğrenme hızı ($learning\ rate$)
- Ağaç sayısı ($n_{estimators}$)
- Maksimum derinlik (max_depth)
- Minimum yaprak ağırlığı (min_child_weight)
- Özellik fraksiyonu ($feature_fraction$)
- Torbalama fraksiyonu ($bagging_fraction$)
- L2 düzenleme terimi (λ)

Öğrenme hızı, her ağacın katkısının ne kadar olacağını belirlerken, ağaç sayısı ($n_{estimators}$) toplam ağaç sayısını belirler. Maksimum derinlik (max_depth), her ağacın derinliğini kontrol eder ve modelin karmaşıklığını düzenler. Minimum yaprak ağırlığı (min_child_weight), bir yaprak düğümde bulunması gereken minimum toplam ağırlığı belirlerken, özellik fraksiyonu ($feature_fraction$) her iterasyonda kullanılacak özelliklerin oranını, torbalama fraksiyonu ($bagging_fraction$) ise her iterasyonda kullanılacak veri örneklerinin oranını belirler. L2 düzenleme terimi (λ), modelin karmaşıklığını kontrol eden bir parametredir. Bu hiperparametrelerin optimize edilmesi, modelin aşırı uyum yapmasını engelleyebilir ve genelleme yeteneğini artırabilir [48] [50] [45].

LightGBM, sağlık, finans ve perakende gibi birçok farklı alanda uygulanmaktadır. Sağlık alanında, hastalık sonuçlarının tahmin edilmesi ve hasta tekrar yatışlarının öngörülmesi gibi öngörücü modelleme uygulamalarında kullanılır. Finans sektöründe, kredi skorlama, sahtekarlık tespiti ve algoritmik ticaret gibi alanlarda uygulanır. Perakende sektöründe ise satış tahmini, envanter yönetimi ve müşteri segmentasyonu gibi alanlarda kullanılmaktadır [48] [50] [45] .

3.5.3 CatBoost Algoritması

CatBoost, Yandex tarafından geliştirilen ve gradyen arttırma algoritmasının özel bir versiyonudur. Bu algoritma, özellikle kategorik verilerle çalışırken üstün performans gösterir. 2017 yılında Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush ve Andrey Gulin tarafından tanıtılmıştır. CatBoost, makine öğrenimi ve veri bilimi alanlarında geniş bir kullanım alanı

bulmuştur. Genellikle yüksek performansı, hızı ve kategorik verilerle çalışma yeteneği ile bilinir [45] [51].

CatBoost'un temel prensibi, gradyen arttırma algoritmasını kullanarak bir dizi karar ağacını ardışık olarak eğitmek ve model hatalarını minimize etmektir. CatBoost'un en önemli özelliklerinden biri, kategorik verileri doğrudan işleyebilme yeteneğidir. Diğer algoritmalar genellikle kategorik verileri işleyebilmek için one-hot encoding gibi ön işleme adımları gerektirirken, CatBoost bu tür ön işleme gereksinimlerini ortadan kaldırır. Bu sayede, modelin hem eğitim süresi kısalır hem de doğruluğu artar [45] [51] .

CatBoost, performansını artırmak için birkaç yenilikçi teknik kullanır. Bunlardan biri, "ordered boosting" tekniğidir. Bu teknik, veri kümesinde meydana gelen önyargıyı (bias) ortadan kaldırmak için verileri rastgele sıralar ve ardından her bir veri noktası için doğru tahminler yapar. Diğer bir yenilikçi teknik ise "oblivious trees" adı verilen ve karar ağaçlarının daha verimli bir şekilde oluşturulmasını sağlayan yöntemdir. Bu teknik, her düğümde aynı bölünme kuralını kullanarak, ağaç yapısının hafızada daha az yer kaplamasını ve hesaplama süresinin azalmasını sağlar [45] [51] .

CatBoost'un avantajları arasında yüksek performans, düşük bellek kullanımı ve kategorik verilerle doğrudan çalışma yeteneği bulunur. Düzenleme teknikleri sayesinde çeşitli veri kümelerinde güvenilir sonuçlar verir ve aşırı uyumu önler. Ayrıca, regresyon, sınıflandırma ve sıralama gibi çeşitli amaç fonksiyonlarını destekler, bu da onu geniş bir yelpazede makine öğrenimi görevleri için uygun hale getirir. CatBoost'un paralel ve dağıtık hesaplama yetenekleri vardır. Bu yetenekler, algoritmanın birden fazla çekirdekte ve dağıtık küme yapılarında paralel olarak çalışabilmesini sağlar. Bu da ölçeklenebilirliği ve performansı artırır [45] [51] [50].

CatBoost'un başarısında hiperparametre ayarlarının büyük bir rolü vardır. Bu hiperparametrelerin dikkatlice ayarlanması, model performansını önemli ölçüde artırabilir. Önemli hiperparametreler şunlardır:

- Öğrenme hızı ($learning\ rate$)
- Ağaç derinliği ($depth$)
- L2 düzenleme terimi ($l2_leaf_reg$)
- Iterasyon sayısı ($iterations$)
- Torbalama sıcaklığı ($bagging_temperature$)

- Rastgelelik kuvveti ($random_strength$)

Öğrenme hızı, her ağacın katkısının ne kadar olacağını belirlerken, ağaç derinliği her ağacın maksimum derinliğini belirler. L2 düzenleme terimi, modelin karmaşıklığını kontrol eden bir parametredir. Iterasyon sayısı, toplam ağaç sayısını belirlerken, torbalama sıcaklığı modelin aşırı uyum yapmasını engellemek için veri örneklemeyi kontrol eder ve rastgelelik kuvveti ise modelin aşırı uyum yapmasını engellemek için rastgeleliği kontrol eder. Bu hiperparametrelerin optimize edilmesi, modelin aşırı uyum yapmasını engelleyebilir ve genelleme yeteneğini artırabilir [45] [51].

CatBoost, sağlık, finans ve perakende gibi birçok farklı alanda uygulanmaktadır. Sağlık alanında, hastalık sonuçlarının tahmin edilmesi ve hasta tekrar yatışlarının öngörülmesi gibi öngörücü modelleme uygulamalarında kullanılır. Finans sektöründe, kredi skorlama, sahtekarlık tespiti ve algoritmik ticaret gibi alanlarda uygulanır. Perakende sektöründe ise satış tahmini, envanter yönetimi ve müşteri segmentasyonu gibi alanlarda kullanılmaktadır [45] [51] [50].

3.6 Performans Değerlendirme Metrikleri

3.6.1 R Kare (R^2)

R-kare (R^2), bir regresyon modelinin performansını değerlendirmede kullanılan önemli bir metriktir. Bu metrik, bağımlı değişkendeki toplam varyansın ne kadarının bağımsız değişkenler tarafından açıklandığını gösterir ve 0 ile 1 arasında bir değer alır. Yüksek bir R^2 değeri, modelin veriyi iyi açıkladığını gösterirken, düşük bir R^2 değeri modelin veriyi yeterince açıklayamadığını gösterebilir.

R^2, aşağıda verilen 3.1 formülü ile hesaplanır:

$$R^2 = 1 - \frac{SS_{resudial}}{SS_{total}} \tag{3.1}$$

$SS_{resudial}$ (*Residual Sum of Squares*) gerçek değerler ile tahmin edilen değerler arasındaki farkların karelerinin toplamıdır ve aşağıda verilen 3.2 formülü ile hesaplanır:

$$SS_{resudial} = \sum (y_i - \hat{y}_i)^2 \tag{3.2}$$

SS_{total} (*Total Sum of Squares*) gerçek değerler ile bu değerlerin ortalaması

arasındaki farkların karelerinin toplamıdır ve aşağıda verilen 3.3 formülü ile hesaplanır:

$$SS_{total} = \sum (y_i - \bar{y})^2 \tag{3.3}$$

Bu formüllerdeki değerler:

- y_i gerçek değerleri
- $\hat{y}_i$ tahmin edilen değerleri
- $\bar{y}$ ise gerçek değerlerin ortalamasını

temsil eder.

R² değeri, yüzde olarak ifade edilerek modelin veriyi ne kadar iyi açıkladığını kolayca gösterir. Bu, modelin performansını anlamayı ve yorumlamayı basit hale getirir. Farklı regresyon modellerinin performanslarını karşılaştırmak için oldukça kullanışlıdır, çünkü R² değeri, hangi modelin veriyi daha iyi açıkladığını objektif bir şekilde belirlenmesine yardımcı olur. Ayrıca, modelin bağımlı değişkendeki varyansın ne kadarını açıkladığını belirlemek, modelin açıklayıcılık gücünü değerlendirmekte önemli bir rol oynar.

R²'nin kullanımı, model performansını değerlendirmede güçlü bir araçtır. Finans, ekonomi, mühendislik ve pazarlama gibi çeşitli alanlarda yaygın olarak kullanılır. Örneğin, finans sektöründe R², varlık fiyatlama modellerinin performansını değerlendirmek için kullanılabilir. Pazarlamada ise reklam kampanyalarının etkinliğini ölçmek için kullanılabilir.

Ancak, R²'nin sınırlamaları da bulunmaktadır. Yüksek bir R² değeri, modelin veriyi aşırı öğrenmiş olabileceği riskini gösterir. Bu durumda model, eğitim verisindeki gürültüyü de öğrenmiş olabilir ve yeni verilerde iyi performans göstermeyebilir. Ayrıca, bağımsız değişkenlerin sayısını artırmak her zaman R² değerini artırır, ancak bu durum modelin genel performansını iyileştirmez. Eklenen değişkenler anlamlı değilse, modelin tahmin gücü artmayabilir. R² değeri genellikle doğrusal modeller için uygundur ve doğrusal olmayan modellerde yanıltıcı sonuçlar verebilir. Doğrusal olmayan ilişkilerin bulunduğu durumlarda, R²'nin kullanımı sınırlı olabilir.

Bu nedenle, bazı durumlarda alternatif metriklerin kullanılması daha uygun olabilir. Örneğin, kök ortalama kare hata (RMSE), modelin tahmin hatalarının

büyüklüğünü hesaba kattığı için büyük hataları daha fazla cezalandırır ve bu sayede benzer performans gösteren modeller arasında daha iyi ayrım yapabilir. RMSE, modelin doğruluğunu değerlendirmede önemli bir metrik olup, doğrusal olmayan modellerde R²'ye göre daha uygun sonuçlar verebilir.

Sonuç olarak, R², regresyon modellerinin performansını değerlendirmede güçlü bir araçtır, ancak modelin karmaşıklığı ve doğrusal olmayan ilişkiler gibi faktörleri göz önünde bulundurarak kullanılmalıdır. Doğrusal olmayan modellerde veya overfitting riskinin yüksek olduğu durumlarda, RMSE gibi alternatif metrikler daha uygun olabilir [52].

3.6.2 Ortalama Kare Hata

Ortalama kare hata (*Mean Squared Error*, MSE), makine öğrenmesi ve istatistikte yaygın olarak kullanılan bir performans değerlendirme metriğidir. MSE, modelin tahmin ettiği değerler ile gerçek değerler arasındaki farkların karesinin ortalamasını alır. Bu metrik, tahmin hatalarının büyüklüğünü ifade eder ve 0'a yakın değerler modelin yüksek doğruluğunu gösterir.

MSE, aşağıda verilen 3.4 formülü ile hesaplanır:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.4}$$

Formüldeki değerler:

- n veri kümesindeki örnek sayısını
- y_i gerçek değerleri
- $\hat{y}_i$ tahmin edilen değerleri
- $\sum_{i=1}^{n}$ i indisi 1'den n'ye kadar olan toplamı
- $(y_i - \hat{y}_i)^2$ her bir gözlem için tahmin hatasının karesini

ifade etmektedir.

MSE, tahmin hatalarının büyüklüğünü net bir şekilde gösterir. Büyük hatalar daha fazla cezalandırıldığı için modelin hatalarını belirlemede etkilidir. Hesaplaması ve yorumlanması kolaydır; farkların kareleri alınarak tüm hatalar pozitif hale getirilir

ve ortalamaları alınır. Ayrıca, kareli hatalar, özellikle büyük hataların etkisini artırdığı için, modelin ciddi hatalarını daha net bir şekilde gösterir.

MSE, büyük hatalara karşı oldukça hassastır. Büyük hatalar, kare alındığında daha da büyüyerek MSE değerini artırır. Ayrıca, MSE veri kümesinin birimlerine bağımlıdır, bu da farklı veri kümelerinin MSE değerlerini karşılaştırmayı zorlaştırır. MSE'nin tek başına anlamlılığı düşük olabilir; bu nedenle modelin genel performansını değerlendirmek için diğer metriklerle birlikte kullanılması gerekebilir.

MSE, regresyon modellerinde yaygın olarak kullanılır ve finansal tahminler, mühendislik uygulamaları ve pazarlama analizlerinde modelin doğruluğunu değerlendirmek için kullanılır. Model performansını iyileştirmek ve en uygun modeli seçmek için önemli bir araçtır.

3.6.3 Kök Ortalama Kare Hata

RMSE, makine öğrenmesi ve istatistikte yaygın olarak kullanılan bir performans değerlendirme metriğidir. RMSE, modelin tahmin ettiği değerler ile gerçek değerler arasındaki farkların karesinin ortalamasının karekökünü alır. Bu metrik, tahmin hatalarının büyüklüğünü ifade eder ve 0'a yakın değerler modelin yüksek doğruluğunu gösterir.

RMSE, aşağıda verilen 3.5 formülü ile hesaplanır:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.5}$$

Formüldeki değerler:

- n veri kümesindeki örnek sayısını
- y_i gerçek değerleri
- $\hat{y}_i$ tahmin edilen değerleri
- $\sum_{i=1}^{n}$ i indisi 1'den n'ye kadar olan toplamı
- $(y_i - \hat{y}_i)^2$ her bir gözlem için tahmin hatasının karesini

ifade etmektedir.

RMSE, tahmin hatalarının büyüklüğünü doğrudan ifade eder ve büyük hatalar daha fazla cezalandırılır. Bu, modelin ciddi hatalarını belirlemede etkilidir. RMSE, aynı zamanda verilerin birimlerine geri dönmesini sağlar, böylece tahmin hatalarının anlamlı bir ölçüsünü verir. Hesaplanması ve yorumlanması kolaydır, çünkü hataların karekökü alınarak pozitif hale getirilir ve ortalamaları alınır.

RMSE, büyük hatalara karşı oldukça hassastır. Büyük hatalar, kare alındığında daha da büyüyerek RMSE değerini artırır, bu da bazı durumlarda yanıltıcı olabilir. Ayrıca, RMSE'nin birim bağımlılığı, farklı veri kümeleri arasında doğrudan karşılaştırmayı zorlaştırabilir. Bu nedenle, RMSE'nin tek başına kullanılması modelin genel performansını değerlendirmede yeterli olmayabilir; diğer metriklerle birlikte kullanılması gerekebilir.

RMSE, regresyon modellerinde yaygın olarak kullanılır. Aynı zamanda finansal tahminler, mühendislik uygulamaları ve pazarlama analizlerinde modelin doğruluğunu değerlendirmek için kullanılır. Model performansını iyileştirmek ve en uygun modeli seçmek için önemli bir araçtır. Ancak, özellikle büyük hataların daha az önemli olduğu durumlarda alternatif metriklerle birlikte kullanılması daha uygun olabilir. Örneğin, RMSE, doğrusal modellerin yanı sıra zaman serisi tahminleri ve çeşitli makine öğrenimi uygulamalarında da yaygın olarak kullanılır.

Sonuç olarak, RMSE, tahmin hatalarının büyüklüğünü anlamada güçlü bir araçtır, ancak büyük hatalara duyarlılığı nedeniyle dikkatli kullanılması ve diğer performans metrikleriyle birlikte değerlendirilmesi önemlidir.

3.6.4 Ortalama Mutlak Hata

Ortalama mutlak hata (MAE), makine öğrenmesi ve istatistikte yaygın olarak kullanılan bir performans değerlendirme metriğidir. MAE, modelin tahmin ettiği değerler ile gerçek değerler arasındaki farkların mutlak değerlerinin ortalamasını alır. Bu metrik, tahmin hatalarının büyüklüğünü ifade eder ve 0'a yakın değerler modelin yüksek doğruluğunu gösterir.

MAE, aşağıda verilen 3.6 formülü ile hesaplanır:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3.6}$$

Formüldeki değerler:

- n veri kümesindeki örnek sayısını
- y_i gerçek değerleri
- $\hat{y}_i$ tahmin edilen değerleri
- $\sum_{i=1}^{n}$ i indisi 1'den n'ye kadar olan toplamı
- $|y_i - \hat{y}_i|$ her bir gözlem için tahmin hatasının mutlak değerini

ifade etmektedir.

MAE, tahmin hatalarının büyüklüğünü doğrudan ifade eder ve hesaplaması oldukça basittir. Hataların mutlak değerleri alınarak pozitif hale getirilir ve ortalamaları alınır. MAE, her bir hatayı eşit derecede değerlendirir ve büyük hataları aşırı cezalandırmaz. Bu nedenle, modelin genel performansını değerlendirmede dengeli bir yaklaşım sunar.

MAE, bazı durumlarda büyük hataları yeterince vurgulamayabilir. Büyük hatalar, mutlak değerleri alındığında daha az belirgin hale gelir, bu da bazı durumlarda modelin performansını tam olarak yansıtmayabilir. Ayrıca, MAE'nin birim bağımlılığı, farklı veri kümeleri arasında doğrudan karşılaştırmayı zorlaştırabilir. Bu nedenle, MAE'nin tek başına kullanılması modelin genel performansını değerlendirmede yeterli olmayabilir; diğer metriklerle birlikte kullanılması gerekebilir.

MAE, regresyon modellerinde yaygın olarak kullanılır ve finansal tahminler, mühendislik uygulamaları ve pazarlama analizlerinde modelin doğruluğunu değerlendirmek için kullanılır. Model performansını iyileştirmek ve en uygun modeli seçmek için önemli bir araçtır. MAE, tahmin hatalarının büyüklüğünü anlamada dengeli bir metrik olduğu için, farklı alanlarda yaygın olarak tercih edilir. Özellikle, büyük hataların daha az önemli olduğu durumlarda veya hataların mutlak değerlerinin daha anlamlı olduğu senaryolarda kullanımı uygundur. Sonuç olarak, MAE, tahmin hatalarının büyüklüğünü anlamada etkili bir araçtır, ancak büyük hataların vurgulanmasında yetersiz kalabileceği için dikkatli kullanılması ve diğer performans metrikleriyle birlikte değerlendirilmesi önemlidir.

4
UYGULAMA

Bu bölümde çalışmanın veri toplama, veri ön işleme ve veri setinin modellenmesi aşamaları ele alınacaktır. Ardından, gerçekleştirilen analizler ve bu analizlerin sonuçları, modellerin doğruluk oranları ve başarıları ile birlikte değerlendirilecektir. Modellerin performansları karşılaştırılarak, en iyi sonuçları veren model ve elde edilen bulgular detaylı bir şekilde incelenecektir.

CBS analizleri için QGIS yazılımı kullanılmıştır. QGIS, veri görüntüleme, düzenleme ve çözümleme yetenekleri sağlayan çoklu platform destekli özgür ve açık kaynaklı bir CBS yazılımıdır. Bu yazılım, konutların şehir içindeki stratejik noktalara olan uzaklıklarını ölçmek ve bu verileri analize entegre etmek için tercih edilmiştir. Ayrıca, konutların coğrafi dağılımını ve çevresel etkileşimlerini görselleştirmek için tematik haritalar oluşturulmuştur. QGIS, kullanıcı dostu arayüzü ve geniş eklenti seçenekleri ile veri analiz sürecini daha verimli ve esnek hale getirmektedir.

Veri ön işleme, özellik mühendisliği ve model optimizasyonu adımları Python ortamında gerçekleştirilmiştir. Python, veri bilimi, web geliştirme, otomasyon ve yapay zeka gibi çeşitli alanlarda yaygın olarak kullanılan güçlü ve esnek bir programlama dilidir. Python'da veri analizi ve makine öğrenmesi uygulamaları için popüler olan Scikit-Learn kütüphanesi kullanılmıştır. Bu kütüphane, veri ön işleme aşamasında eksik verilerin doldurulması, kategorik verilerin kodlanması ve verilerin normalize edilmesi gibi işlemler için kapsamlı araçlar sunmaktadır.

Modelleme sürecinde, konut fiyatlarını tahmin etmek için XGBoost, LightGBM ve CatBoost makine öğrenmesi algoritmaları uygulanmıştır. XGBoost, yüksek doğruluk oranı ve hız ile bilinen bir gradyan artırma algoritmasıdır. LightGBM, büyük veri kümeleri üzerinde hızlı ve verimli bir şekilde çalışan hafif bir gradyan artırma algoritmasıdır. CatBoost ise, kategorik verilerin otomatik olarak işlenmesi ve güçlü performansı ile öne çıkan bir algoritmadır. Bu algoritmaların her biri, model performansını optimize etmek için hiperparametre ayarlamaları ile

birlikte kullanılmıştır. Analiz sonuçları, her bir algoritmanın model performansı açısından nasıl bir fark yarattığını belirlemek için ayrıntılı olarak incelenmiş ve karşılaştırılmıştır.

4.1 Veri Kümesi

Bu tez çalışmasında, İstanbul iline ait konut satış verileri kullanılmıştır. Veri kümesi, bir gayrimenkül firması tarafından sağlanmış olup, 09.11.2023 - 24.01.2024 tarihleri arasındaki konut satışlarını içermektedir. Veri kümesi toplamda 45.693 örnek ve 17 öznitelik içermektedir. Tablo 4.1, orijinal veri kümesindeki konutlara ait değişkenleri, türlerini ve eksik gözlem sayısını göstermektedir.

Tablo 4.1 Veri kümesinde çalışmada kullanılacak konuta ait değişkenler

Değişken Adı	Türü	Eksik Gözlem
Konutun Büyüklüğü (m² (Net))	Numerik	0
Oda Sayısı	Numerik	0
Salon Sayısı	Numerik	0
Bina Yaşı	Numerik	140
Bulunduğu Kat	Kategorik	3988
Kat Sayısı	Kategorik	0
Isıtma	Kategorik	136
Banyo Sayısı	Numerik	0
Asansör	Kategorik	13584
Otopark	Kategorik	9450
Site İçerisinde	Kategorik	437
Krediye Uygun	Kategorik	470
İlçe	Kategorik	0
Mahalle	Kategorik	0
Fiyat	Numerik	0
Enlem	Numerik	0
Boylam	Numerik	0

Tablo 4.2'de CBS teknikleriyle hesaplanan uzaklık değişkenlerine ait değişken adı, türü ve eksik gözlem bilgisi bulunmaktadır.

Tablo 4.2 Uzaklık bilgisinin bulunduğu değişkenler

Değişken Adı	Türü	Eksik Gözlem
Ulaşım Noktalarına Uzaklık	Numerik	0
Sahile/Denize Uzaklık	Numerik	0
Park/Yeşil Alan Uzaklık	Numerik	0
Alışveriş Merkezine Uzaklık	Numerik	0

Çalışmada, mevcut veri kümesinden üretilen ek değişkenler de bulunmaktadır. Bu değişkenler, konut satışlarını daha iyi analiz edebilmek için veri ön işleme ve özellik

mühendisliği adımlarıyla oluşturulmuştur. Üretilen değişkenler ve özellikleri Tablo 4.3'te verilmiştir.

Tablo 4.3 Veri kümesinden üretilen değişkenler

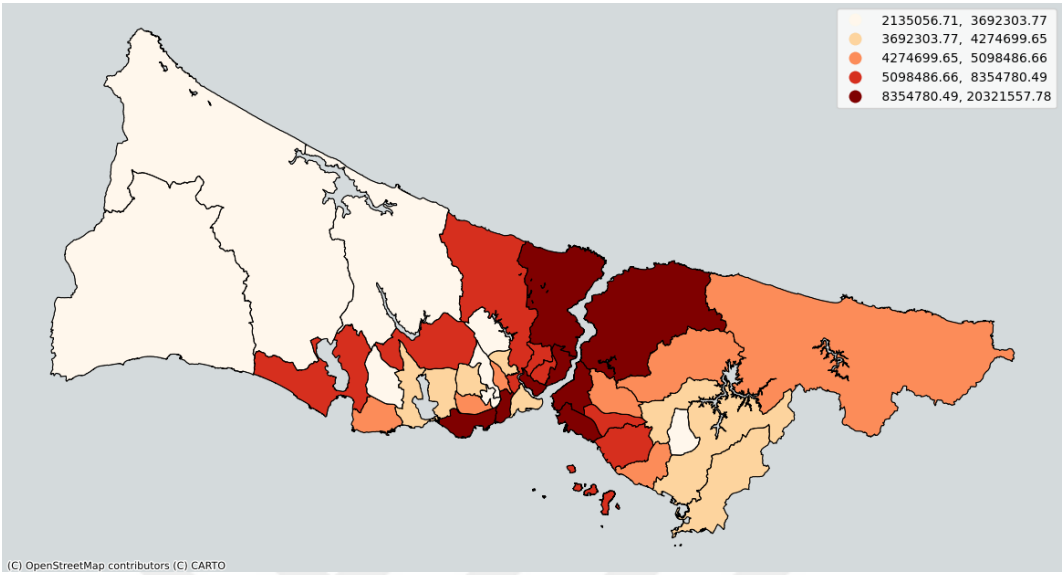
Değişken Adı	Türü	Eksik Gözlem
Daire Kategorisi	Kategorik	0
Yıl	Numerik	0
Ay	Numerik	0
Gün	Numerik	0
Haftanın Günü	Numerik	0
Hafta Gününün İsmi	Kategorik	0
Hafta Sonu	Kategorik	0

Oluşturulan değişkenlerin açıklamaları şu şekildedir:

- Konutların metrekare büyüklüğü, oda sayısı ve salon sayısı temel özelliklerini göz önünde bulundurarak bir kategorize etme fonksiyonu geliştirildi. Bu fonksiyonun sonucunda "Daire Kategorisi" değişkeni oluşturulmuştur. Net metrekare alanı 90 m² veya daha az olan, oda sayısı 2 veya daha az olan ve salon sayısı 1 veya daha az olan konutlar "küçük ve ekonomik", net metrekare alanı 150 m² veya daha az olan, oda sayısı 3.5 veya daha az olan ve salon sayısı 2 veya daha az olan konutlar "orta düzey konforlu" ve daha büyük ve daha fazla odaya sahip konutlar "geniş ve lüks" olarak kategorize edildi. Bu değişken, farklı konut türlerinin fiyat üzerindeki etkisini analiz etmeyi sağlar.
- Satış tarihinden çıkarılan "Yıl, Ay, Gün" değişkenleri, konut satışlarının yıl, ay ve gün bazında analiz edilmesine olanak tanır.
- Satışların haftanın hangi gününde gerçekleştiğini gösteren "Haftanın Günü" değişkeni, haftalık döngülerin etkisini incelemek için kullanılır. Haftanın günlerinin isimlerini içeren "Hafta Gününün İsmi" değişkeni ise kategorik analizlerde kullanılabilir.
- Satışın hafta sonu mu, yoksa hafta içi mi gerçekleştiğini belirten "Hafta Sonu" değişkeni, haftalık eğilimlerin tespitinde faydalıdır.

Öznitelik seçim sürecinde, literatürde konut seçiminde etkili olan kriterler dikkate alınmıştır. Literatür taraması sonucunda, konut seçiminde etkili olan ana kriterler ve alt kriterler belirlenmiştir. Bu kriterler; konuta ilişkin fiziki özellikler, çevresel özellikler, yapısal özellikler ve finansal özellikler olarak tanımlanmıştır. Alt kriterler arasında oda sayısı, konutun bulunduğu kat, cephe, manzara, denize

yakınlık, alışveriş merkezi ve markete yakınlık, ulaşım kolaylığı, oyun alanı veya yeşil alan, bina yaşı, ısıtma sistemi, otopark, asansör, fiyat, krediye uygunluk ve ödeme koşulları yer almaktadır [53].



Şekil 4.1 İlçelere göre ortalama konut fiyatları

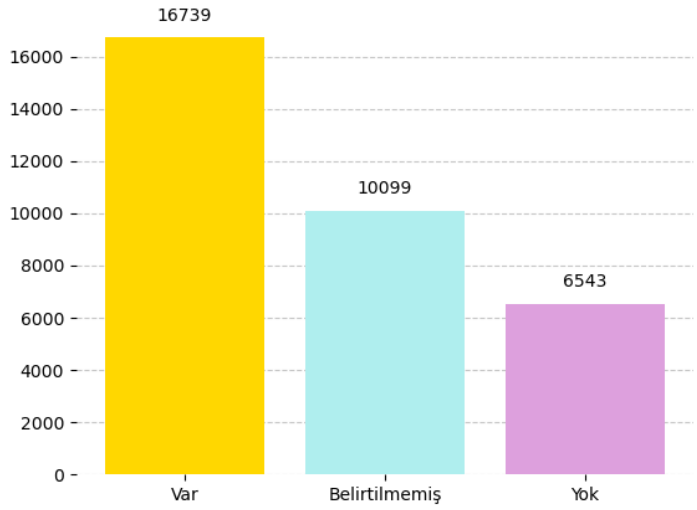
Şekil 4.1'deki harita İstanbul'un farklı bölgelerinde konut fiyatlarının nasıl dağıldığını görsel olarak ifade etmekte ve merkezi, gelişmiş bölgelerde fiyatların daha yüksek olduğunu, daha uzak bölgelerde ise fiyatların daha düşük olduğunu göstermektedir. İstanbul'un merkezi ve daha gelişmiş bölgelerindeki konut fiyatlarının daha yüksek olduğunu görebiliriz. En koyu renkle gösterilen ilçeler, en yüksek konut fiyatlarına sahip bölgeler olup, haritada bunlar genellikle İstanbul'un Avrupa yakasında yer almaktadır. Özellikle Beşiktaş, Sarıyer, Kadıköy gibi merkezi ilçeler en yüksek fiyat kategorisinde yer alıyor. Açık renklerle gösterilen ilçeler ise daha düşük ortalama konut fiyatlarına sahip bölgeler olup, genellikle İstanbul'un daha uzak ve gelişmekte olan bölgelerinde yer almaktadır. Bu bölgeler arasında Silivri, Çatalca gibi ilçeler öne çıkmaktadır.

4.2 Verilerin Analizi

Veri analizi aşamasında, ilk olarak verilerin ön işlenmesi gerçekleştirilmiştir. Bu aşamada eksik gözlemlerin incelenmesi, kategorik verilerin sınıflandırılması, aykırı gözlemlerin temizlenmesi ve her mahalle için yeterli verinin mevcut olma durumu incelenmiştir. İşlemlerin sonunda ham veri kümesinde 33381 adet örnek kalmıştır.

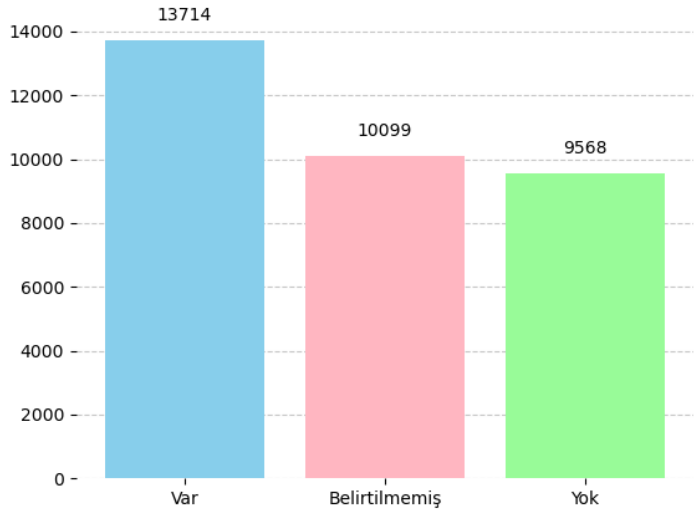
İlk olarak otopark, asansör, krediye uygunluk ve site içerisinde sütunlarındaki eksik değerler 'Belirtilmemiş' olarak doldurulmuştur.

Şekil 4.2'deki dağılım, üç farklı asansör durumu kategorisini içermektedir: "Var", "Belirtilmemiş" ve "Yok". Bu kategoriler, asansör durumu bilgisi mevcut olan konutların sayısını, belirtilmemiş olanları ve asansörü bulunmayanları temsil eder. Asansörü olan konutlar 19.842 adet, asansör durumu belirtilmemiş konutlar 12.135 adet, asansörü olmayan konutlar ise 7.548 adettir.



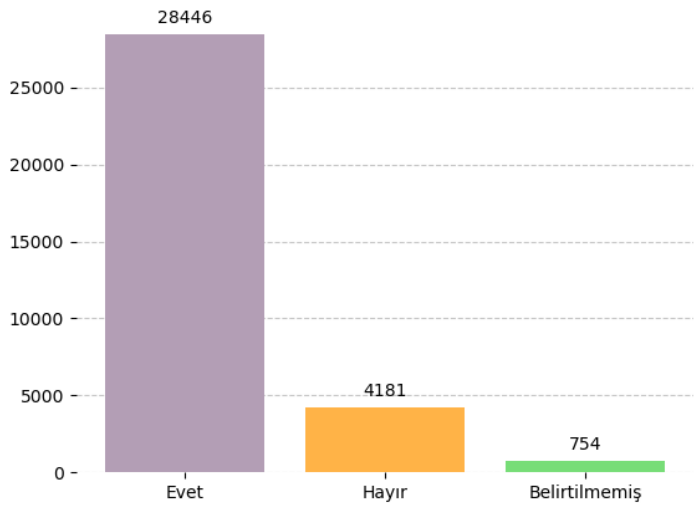
Şekil 4.2 Asansör değişkeni dağılımı

Şekil 4.3'de, İstanbul ilinde konut satışlarının otopark durumu açısından dağılımı gösterilmektedir. Otoparkı olan konutlar, 16.642 adetle en yüksek sayıya sahiptir. Otopark durumu belirtilmemiş konutlar, 12.135 adetle ikinci sırada yer almaktadır. Otoparkı olmayan konutlar ise 10.748 adetle en düşük sayıya sahiptir.



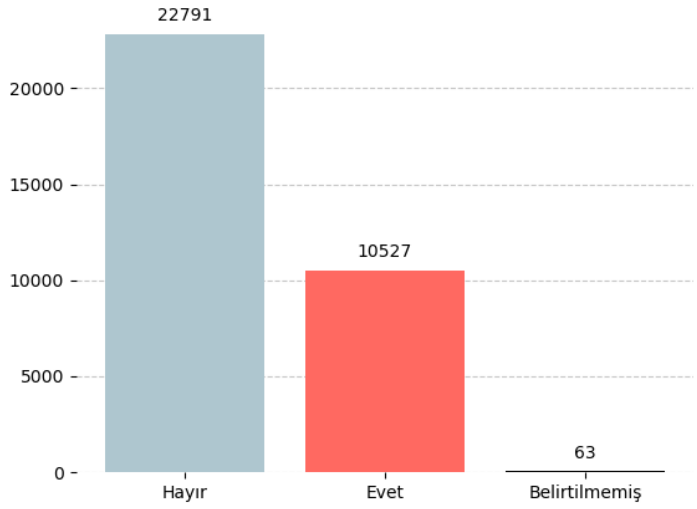
Şekil 4.3 Otopark değişkeni dağılımı

Şekil 4.4'de, İstanbul ilinde konut satışlarının krediye uygunluk durumu açısından dağılımı gösterilmektedir. Krediye uygun konutlar, 33.414 adetle en yüksek sayıya sahiptir. Krediye uygun olmayan konutlar, 5.206 adetle ikinci sırada yer almaktadır. Kredi durumu belirtilmemiş konutlar ise 905 adetle en düşük sayıya sahiptir.



Şekil 4.4 Kredi durumu değişkeni dağılımı

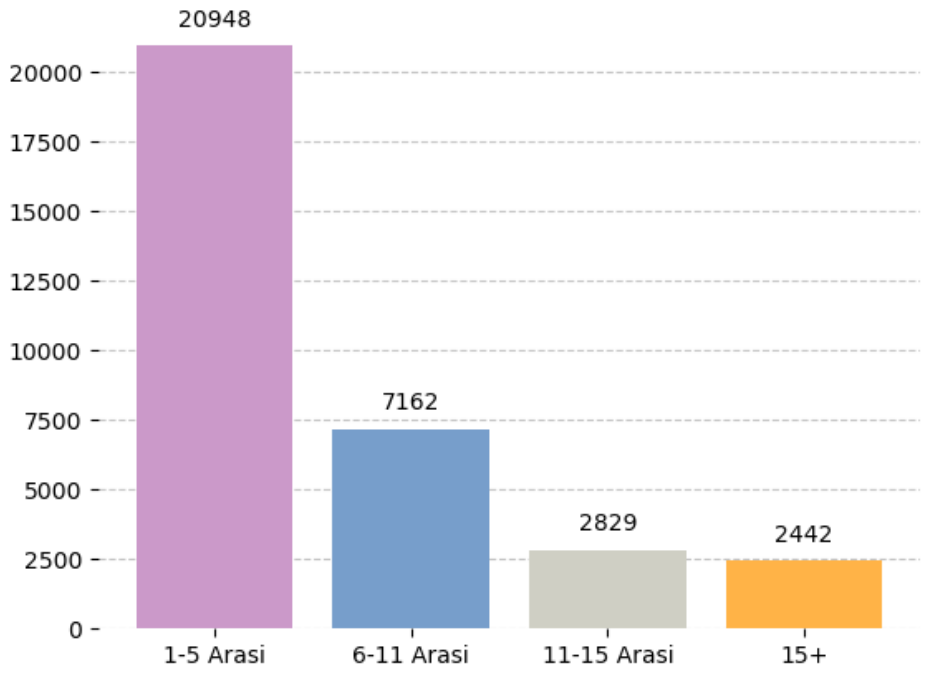
Şekil 4.5'de, İstanbul ilindeki konutların site içerisinde olup olmama durumuna göre nasıl sınıflandırıldığını ve her kategoride kaç adet konut bulunduğunu ortaya koymaktadır. Site içerisinde olmayan konutlar 26.084 adetle en yüksek sayıya sahiptir. Site içerisinde olan konutlar 13.319 adetle ikinci sırada yer almaktadır. Site durumu belirtilmemiş konutlar ise 122 adetle en düşük sayıya sahiptir.



Şekil 4.5 Site içerisinde olma durumu dağılımı

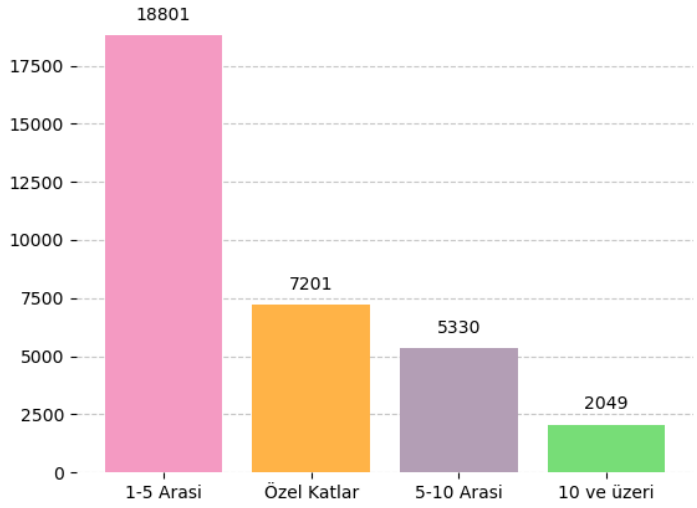
Bunun yanı sıra verilerin kategorik özellikleri belirli kurallara göre düzenlenmiştir. 'Kat Sayısı' sütunundaki veriler belirli aralıklara göre yeniden sınıflandırılmıştır. Örneğin, 1-5 arası kat sayısına sahip binalar "1-5 Arası", 6-10 arası kat sayısına sahip binalar "6-10 Arası" olarak gruplandırılmıştır. Aynı şekilde konutun bulunduğu kat bilgisinin olduğu "Bulunduğu Kat" değişkeninin kategorileri de düzenlenmiştir.

Şekil 4.6'teki grafikte, İstanbul ilindeki konutların kat sayısına göre nasıl sınıflandırıldığını ve her kategoride kaç adet konut bulunduğunu göstermektedir. 1-5 kat arası olan konutlar, 24.669 adetle en yüksek sayıya sahiptir. 6-11 kat arası olan konutlar, 8.382 adetle ikinci sırada yer almaktadır. 11-15 kat arası olan konutlar, 3.279 adetle üçüncü sırada yer alırken, 15 ve üzeri katlı konutlar 3.195 adetle en düşük sayıya sahiptir.



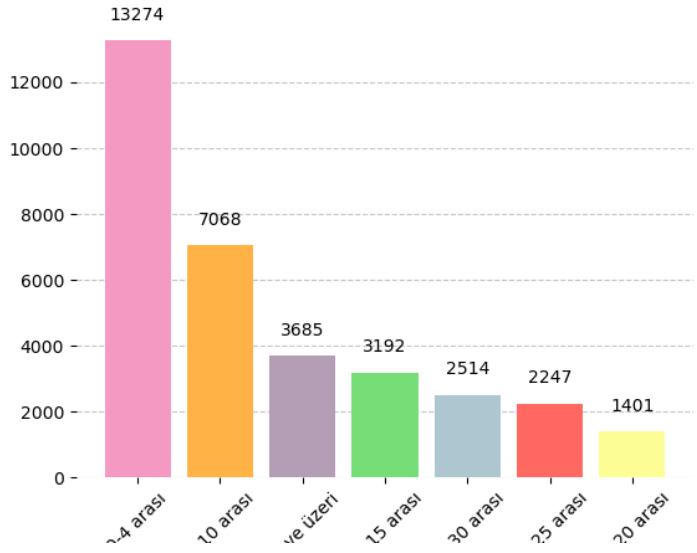
Şekil 4.6 Kat sayısı değişkeni dağılımı

Şekil 4.7, konutların bulunduğu kata göre nasıl sınıflandırıldığını ve her kategoride kaç adet konut bulunduğunu göstermektedir. 1-5 kat arasında bulunan konutlar, 21.892 adetle en yüksek sayıya sahiptir. Özel katlarda (çatı katı) bulunan konutlar, 8.554 adetle ikinci sıradadır. 5-10 kat arasında bulunan konutlar, 6.501 adetle üçüncü sırada yer almaktadır. 10 ve üzeri katlarda bulunan konutlar ise 2.563 adetle en düşük sayıya sahiptir.



Şekil 4.7 Konutun bulunduğu kat değişkeni dağılımı

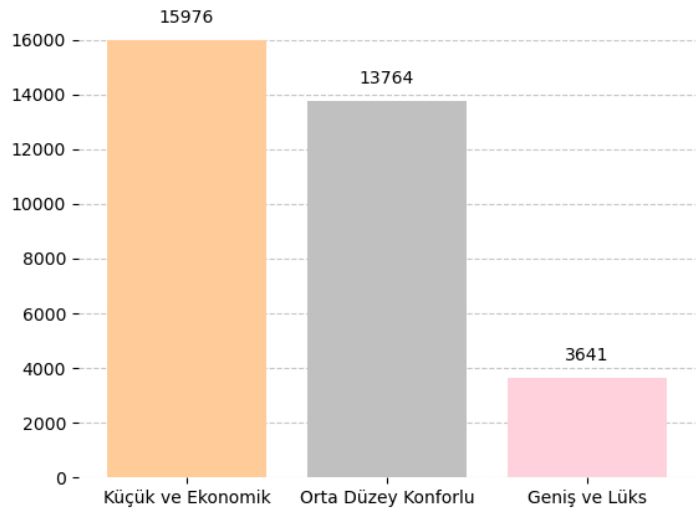
Şekil 4.8'de bina yaşlarına göre daire sayılarının dağılımı gösterilmektedir. 0-4 yaş arası binalar 15636 adetle en yaygın olanıdır, bu da yeni binaların pazarda önemli bir yer tuttuğunu göstermektedir. 5-10 yaş arası binalar 8202 adetle ikinci sırada yer almaktadır. 21 yaş ve üzeri binalar 4574 adetle pazarda belirli bir yer tutarken, 11-15 yaş arası binalar 3745 adet, 16-20 yaş arası binalar 1741 adet, 25 yaş binalar 2670 adet ve 30 yaş binalar 2942 adet daire ile temsil edilmektedir.



Şekil 4.8 Bina yaşı dağılımı

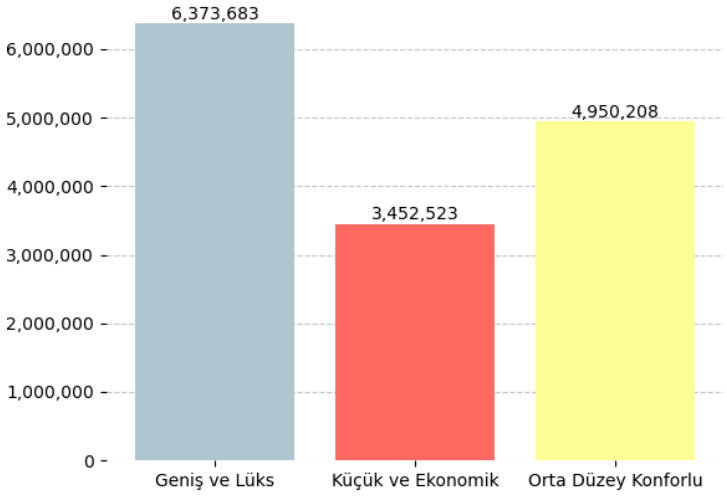
Şekil 4.9'de, daire kategorilerinin sayısal dağılımı gösterilmektedir. Küçük ve Ekonomik daireler 17321 adetle en yaygın olanıdır. Orta Düzey Konforlu daireler 15962 adetle ikinci sırada yer alırken, Geniş ve Lüks daireler 6227 adetle en az

sayıda bulunmaktadır.



Şekil 4.9 Daire kategorisi değişkeninin dağılımı

Şekil 4.10'da üç farklı daire kategorisinin ortalama fiyatları karşılaştırılmaktadır. Geniş ve lüks daireler, ortalama fiyatı 6.373.683 TL ile en yüksek ortalama fiyata sahip kategori olarak öne çıkmaktadır. Bu kategoriyi 4.950.208 TL ortalama fiyat ile orta düzey konforlu daireler takip etmektedir. En düşük ortalama fiyata sahip kategori ise 3.452.523 TL ile küçük ve ekonomik dairelerdir.



Şekil 4.10 Daire kategorilerine göre ortalama fiyat dağılımı

Fiyat değişkeninde aykırı değerler, çeyrekler arası aralık (IQR) yöntemi kullanılarak temizlenmiştir. Bu yöntemde, verilerin alt çeyrek ve üst çeyrek değerleri

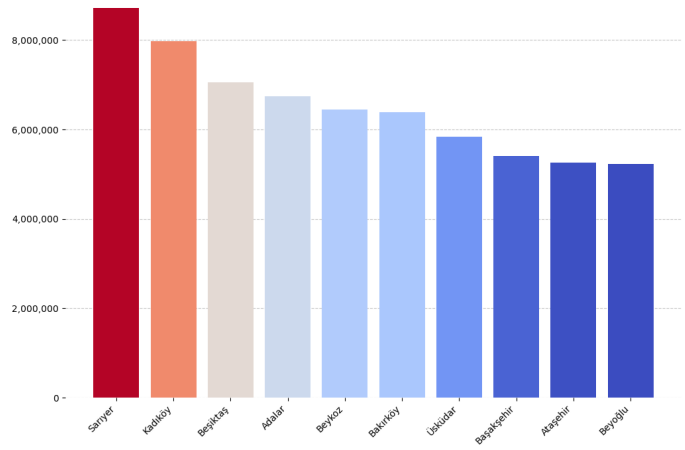
hesaplanmış, bu değerlerin 1.5 katı kadar dışındaki veriler aykırı değer olarak kabul edilip çıkarılmıştır.

Tablo 4.4'teki özet, satılık konut fiyatlarının geniş bir aralıkta dağıldığını göstermektedir. Ortalama fiyatın (4.371.970,22) medyan fiyattan (3.650.000) yüksek olması, bazı yüksek fiyatlı ilanların ortalamayı yukarı çektiğini göstermektedir. Standart sapmanın makul bir seviyede olması, fiyatların büyük ölçüde tutarlı olduğunu ve aşırı uç değerlerin etkisinin sınırlı olduğunu gösterir. En düşük ve en yüksek fiyatlar arasındaki geniş aralık, piyasanın çok çeşitli fiyat segmentlerine sahip olduğunu ve hem uygun fiyatlı hem de lüks konutların bulunduğunu ortaya koymaktadır. Bu çeşitlilik, modelleme ve analiz süreçlerinde dikkate alınması gereken önemli bir faktördür.

Tablo 4.4 Konutların fiyatlarına ilişkin istatistiksel özet

Ölçüm	Değer (TL)
Ortalama	4.371.970,22
Minimum	1.190.000
%25 (1. çeyrek)	2.650.000
%50 (medyan)	3.650.000
%75 (3. çeyrek)	5.500.000
Maksimum	12.500.000
standart Sapma	2.397.722,12

Şekil 4.11'da İstanbul'un farklı ilçelerine göre ortalama konut fiyatlarının dağılımı gösterilmektedir. En yüksek ortalama fiyata sahip 10 ilçe sırasıyla belirtilmiştir ve bu ilçeler arasında belirgin fiyat farkları gözlemlenmektedir.

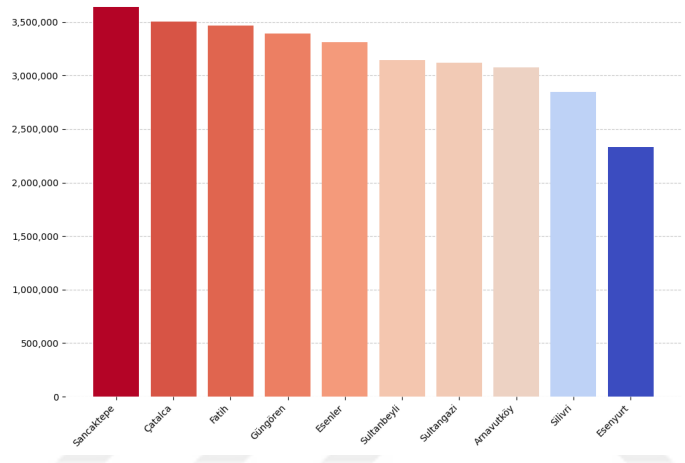


Şekil 4.11 En yüksek ortalama fiyata sahip 10 ilçe

En yüksek ortalama fiyata sahip ilçe Sarıyer olup, burada ortalama konut fiyatı yaklaşık 8.000.000 TL civarındadır. Sarıyer'i sırasıyla Kadıköy, Beşiktaş, Adalar

ve Beykoz izlemektedir. Bu ilçelerin ortalama konut fiyatları 6.000.000 TL ile 7.000.000 TL arasında değişmektedir. Bakırköy, Üsküdar, Başakşehir, Ataşehir ve Beyoğlu ilçeleri ise nispeten daha düşük ortalama fiyatlara sahiptir, bu fiyatlar 4.000.000 TL ile 5.000.000 TL arasında seyretmektedir.

Şekil 4.12'de İstanbul'un farklı ilçelerine göre en düşük ortalama konut fiyatlarına sahip 10 ilçe gösterilmektedir. Bu ilçeler arasında belirgin fiyat farkları gözlemlenmektedir ve grafikte düşük fiyatlar için açık renkler, yüksek fiyatlar için koyu renkler kullanılmıştır.



Şekil 4.12 En düşük ortalama fiyata sahip 10 ilçe

En düşük ortalama fiyata sahip ilçe Esenyurt olup, burada ortalama konut fiyatı yaklaşık 2.000.000 TL civarındadır. Esenyurt'u sırasıyla Silivri, Arnavutköy, Sultangazi, Sultangazi, Sultanbeyli, Esenler, Güngören, Fatih, Çatalca ve Sancaktepe izlemektedir. Bu ilçelerin ortalama konut fiyatları 2.000.000 TL ile 3.500.000 TL arasında değişmektedir. Sancaktepe en yüksek fiyatlı ilçe olarak öne çıkarken, Esenyurt en düşük fiyatlı ilçe olarak dikkat çekmektedir.

Bazı mahallelerde yeterli veri bulunmaması, modelin performansını olumsuz etkileyebilir. Bu nedenle, belirli bir minimum veri miktarı eşiği belirleyerek veri kümesinin filtrelenmesi işlemi gerçekleştirildi. Bu adımda, mahalle bazlı en az veri sayısının kaç olması gerektiği incelendi ve bu eşik değerine göre veri kümesinin yüzde kaçının korunabileceği hesaplandı. Farklı minimum veri miktarı eşikleri belirlenerek, bu eşiklerde veri kümesinin yüzde kaçının kullanılabileceği değerlendirildi. Böylece, minimum veri miktarı eşiklerine göre veri kümesinin ne kadarının korunduğu anlaşılabilir hale geldi. Hesaplamalara göre her mahallede en az 20 veri bulunması durumunda veri kümesinin %95'ı açıklanabilmektedir. Buna bağlı olarak, belirlenen minimum veri miktarına göre yeterli veriye sahip olmayan

mahalleler veri kümesinden çıkarıldı. Bu adım, modelin her mahalle için yeterli veriyle eğitilmesini sağlayarak, daha doğru tahminler yapılmasını ve modelin genel performansının artırılmasını amaçlamaktadır.

4.3 Modellerin Uygulanması ve Sonuçlar

Çalışmada, iki ayrı veri kümesinin konut satış fiyatlarının tahmini için gelişmiş regresyon modelleri kullanılmıştır. İlk veri kümesinde yalnızca temel özellikler (emlak tipi, bina yaşı, bulunduğu kat, kat sayısı, ısıtma türü, banyo sayısı, asansör, otopark, site içerisinde olup olmadığı, krediye uygunluk, ilçe, mahalle, daire kategorisi ve hafta günü) kullanılmıştır. İkinci veri kümesinde ise ek olarak CBS ile hesaplanan uzaklık değişkenleri eklenmiştir. Bu iki veri kümesiyle yapılan modellemelerin karşılaştırılması, CBS ile hesaplanan değişkenlerin konut fiyat tahminlerindeki etkisini ortaya koymak amacıyla yapılmıştır.

Veri kümelerinin hazırlanmasında ilk adım olarak gereksiz sütunlar ('Enlem', 'Boylam') veri kümesinden çıkarılmış ve kategorik değişkenler encode edilmiştir. Kategorik değişkenlerin encode edilmesi işlemi sonrasında veri kümesi eğitim, doğrulama ve test setlerine ayrılmıştır. Eğitim ve doğrulama kümeleri %80 ve %20 oranında ayrılmış, doğrulama seti ise eşit oranda ikiye bölünerek test seti elde edilmiştir.

Modellerin kurulumu ve eğitimi için iki aşama gerçekleştirilmiştir:

İlk aşamada, temel veri seti kullanılarak XGBoost, LightGBM ve CatBoost modelleri eğitilmiştir. Bu üç model için temel pipeline oluşturulmuş ve hiperparametre ayarlamaları yapılmadan modellerin performansları karşılaştırılmıştır. Modellerin her biri için:

- XGBoost modeline 100 ağaç ($n_estimators = 100$) ve 0.1 öğrenme oranı ($learning_rate = 0.1$) parametreleri atanmıştır.
- LightGBM modeline 100 ağaç ($n_estimators = 100$) ve 0.1 öğrenme oranı ($learning_rate = 0.1$) parametreleri atanmıştır.
- CatBoost modeline 100 ağaç ($n_estimators = 100$) ve 0.1 öğrenme oranı ($learning_rate = 0.1$) parametreleri atanmıştır.

parametreleri atanmıştır. Tablo 4.5'deki sonuçlar, temel değişkenler kullanılarak oluşturulan XGBoost, LightGBM ve CatBoost modellerinin performanslarını karşılaştırmaktadır.

Tablo 4.5 Temel değişkenlerin kullanıldığı modelin sonuçları

Model	R^2	RMSE	MSE	MAE
XGBoost	0.7502	1200000.0000	1440000000000.0000	850000.0000
LightGBM	0.7734	1150000.0000	1322500000000.0000	800000.0000
CatBoost	0.7056	1300000.0000	1690000000000.0000	940000.0000

XGBoost modeli, bağımsız değişkenlerin bağımlı değişkeni açıklama gücünü %75.02 oranında göstermektedir (R^2: 0.7502). Modelin tahminlerinde ortalama sapma yaklaşık 1.2 milyon TL olup, bu değer RMSE olarak belirtilmiştir (RMSE: 1.200.000). Hata karelerinin ortalaması ise 1.440.000.000.000 TL olarak hesaplanmıştır (MSE). Ortalama mutlak hata değeri (MAE) ise 850.000 TL'dir. Bu sonuçlar, XGBoost modelinin oldukça iyi bir performans sergilediğini göstermektedir.

LightGBM modeli ise %77.34 açıklama gücüne sahiptir (R^2: 0.7734). Modelin tahminlerinde ortalama sapma yaklaşık 1.15 milyon TL'dir (RMSE: 1.150.000) ve hata karelerinin ortalaması 1.322.500.000.000 TL olarak hesaplanmıştır (MSE). Ortalama mutlak hata değeri ise 800.000 TL'dir. LightGBM modeli, XGBoost modelinden daha iyi bir performans sergilemektedir.

CatBoost modeli ise %70.00 açıklama gücüne sahiptir (R^2: 0.7056). Modelin tahminlerinde ortalama sapma yaklaşık 1.3 milyon TL olup (RMSE: 1.300.000), hata karelerinin ortalaması 1.690.000.000.000 TL'dir (MSE). Ortalama mutlak hata değeri (MAE) ise 940.000 TL'dir. CatBoost modeli, diğer iki modele kıyasla daha düşük bir performans sergilemektedir.

Genel olarak, temel değişkenlerle yapılan modellemelerde LightGBM modeli en iyi performansı sergilemektedir. En yüksek R^2 (0.7734) değerine sahip olan LightGBM modeli, aynı zamanda en düşük RMSE (1.150.000) ve MAE (800.000) değerlerine sahiptir. XGBoost modeli de iyi bir performans göstermektedir, ancak LightGBM modeline kıyasla biraz daha düşük sonuçlar vermektedir. CatBoost modeli ise diğer iki modele göre daha düşük performans sergilemektedir. Bu değerlendirmeler, temel değişkenlerle yapılan modellemelerde LightGBM modelinin en iyi performansı sergilediğini, XGBoost modelinin ise iyi bir alternatif olduğunu, ancak CatBoost modelinin diğer iki modele göre daha düşük performans gösterdiğini ortaya koymaktadır.

XGBoost, LightGBM ve CatBoost modellerinin hiperparametre optimizasyonu sonrasında elde edilen en iyi hiperparametreler şu şekildedir:

- XGBoost modelinde, $colsample_bytree$ ve subsample parametrelerinin 0.8 olarak ayarlanması, modelin her ağaç için daha az veri ve özellik kullanmasını sağlayarak overfitting riskini azaltmıştır. Öğrenme oranının 0.1 olarak belirlenmesi, modelin daha hızlı ve dengeli bir şekilde öğrenmesine olanak tanımıştır. Maksimum derinlik 10 ve ağaç sayısı 300 olarak belirlenmiş, bu da modelin daha karmaşık ilişkileri öğrenmesini sağlamıştır.

- LightGBM modelinde, $n_estimators$ parametresinin 200 olarak belirlenmesi, modelin yeterli sayıda ağaç oluşturmasını sağlamıştır. Maksimum derinliğin 7 olarak ayarlanması, modelin daha derin ağaçlar oluşturmasını engelleyerek overfitting'in önlenmesine yardımcı olmuştur. Öğrenme oranının 0.2 olarak belirlenmesi, modelin daha hızlı öğrenmesini sağlamıştır.

- CatBoost modelinde, derinlik (depth) parametresinin 10 olarak belirlenmesi, modelin daha karmaşık ilişkileri öğrenmesini sağlamıştır. Iterasyon sayısının 189 olarak ayarlanması, modelin yeterli sayıda öğrenme döngüsünden geçmesini sağlamıştır. Öğrenme oranının yaklaşık 0.198 olarak belirlenmesi, modelin hızlı ancak dengeli bir şekilde öğrenmesini sağlamıştır.

Bu hiperparametre optimizasyonları sırasında, her modelin hiperparametrelerinin performansını değerlendirmek için k-fold cross-validation kullanılmıştır. K-fold cross-validation, veri setini k alt parçaya bölerek modelin her bir parçada eğitilip test edilmesini sağlar. Bu çalışma için k=5 olarak belirlenmiş ve her model, 5 farklı veri seti kombinasyonunda eğitilerek ve test edilerek ortalama performans metrikleri hesaplanmıştır. Tabloda 4.6'daki sonuçlar, hiperparametre optimizasyonuyla geliştirilmiş temel değişkenlerle oluşturulan XGBoost, LightGBM ve CatBoost modellerinin performanslarını karşılaştırmaktadır.

Tablo 4.6 Hiperparametre optimizasyonuyla geliştirilmiş temel değişkenlerin kullanıldığı modelin sonuçları

Model	R^2	RMSE	MSE	MAE
XGBoost	0.8100	1050000.0000	1102500000000.0000	780000.0000
LightGBM	0.8394	1000000.0000	1000000000000.0000	740000.0000
CatBoost	0.7663	1150000.0000	1322500000000.0000	860000.0000

XGBoost modeli, bağımsız değişkenlerin bağımlı değişkeni açıklama gücünü %81.00 oranında göstermektedir (R^2: 0.8100). Modelin tahminlerinde ortalama sapma yaklaşık 1.05 milyon TL olup, bu değer RMSE olarak belirtilmiştir (RMSE: 1.050.000). Hata karelerinin ortalaması ise 1.102.500.000.000 TL olarak hesaplanmıştır (MSE). Ortalama mutlak hata değeri (MAE) ise 780.000

TL'dir. Bu sonuçlar, XGBoost modelinin oldukça iyi bir performans sergilediğini göstermektedir.

LightGBM modeli ise %83.94 açıklama gücüne sahiptir (R^2: 0.8394). Modelin tahminlerinde ortalama sapma yaklaşık 1 milyon TL'dir (RMSE: 1.000.000) ve hata karelerinin ortalaması 1.000.000.000.000 TL olarak hesaplanmıştır (MSE). Ortalama mutlak hata değeri ise 740.000 TL'dir. LightGBM modeli, XGBoost modelinden daha iyi bir performans sergilemektedir.

CatBoost modeli ise %76.63 açıklama gücüne sahiptir (R^2: 0.7663). Modelin tahminlerinde ortalama sapma yaklaşık 1.15 milyon TL olup (RMSE: 1.150.000), hata karelerinin ortalaması 1.322.500.000.000 TL'dir (MSE). Ortalama mutlak hata değeri (MAE) ise 860.000 TL'dir. CatBoost modeli, diğer iki modele kıyasla daha düşük bir performans sergilemektedir.

Genel olarak, hiperparametre optimizasyonu ile geliştirilmiş temel değişkenlerle yapılan modellemelerde LightGBM modeli en iyi performansı sergilemektedir. En yüksek R^2 (0.8394) değerine sahip olan LightGBM modeli, aynı zamanda en düşük RMSE (1.000.000) ve MAE (740.000) değerlerine sahiptir. XGBoost modeli de iyi bir performans göstermekte olup, LightGBM modeline yakın sonuçlar vermektedir. CatBoost modeli ise diğer iki modele kıyasla daha düşük performans sergilemektedir. Bu değerlendirmeler, hiperparametre optimizasyonu yapılmış temel değişkenlerle yapılan modellemelerde LightGBM modelinin en iyi performansı sergilediğini, XGBoost modelinin ise iyi bir alternatif olduğunu, ancak CatBoost modelinin diğer iki modele göre daha düşük performans gösterdiğini ortaya koymaktadır.

İkinci aşamada, ilk veri setine ek olarak CBS ile hesaplanan uzaklık değişkenleri eklenmiş ve modeller yeniden eğitilmiştir. Bu aşamada da XGBoost, LightGBM ve CatBoost modelleri kullanılmıştır. Modellerin her biri için oluşturulan parametre yapısı:

- XGBoost modeline 100 ağaç ($n_estimators = 100$) ve 0.1 öğrenme oranı ($learning_rate = 0.1$) parametreleri atanmıştır.
- LightGBM modeline 100 ağaç ($n_estimators = 100$) ve 0.1 öğrenme oranı ($learning_rate = 0.1$) parametreleri atanmıştır.
- CatBoost modeline 100 ağaç ($n_estimators = 100$) ve 0.1 öğrenme oranı ($learning_rate = 0.1$) parametreleri atanmıştır.

şeklinde açıklanabilir. Tabloda 4.7'deki sonuçlar, CBS değişkenlerinin

eklenmesiyle oluşturulan XGBoost, LightGBM ve CatBoost modellerinin performanslarını karşılaştırmaktadır.

Tablo 4.7 CBS değişkenlerinin eklendiği modelin sonuçları

Model	R^2	RMSE	MSE	MAE
XGBoost	0.7875	1130000.0000	1276900000000.0000	820000.0000
LightGBM	0.8180	1050000.0000	1102500000000.0000	750000.0000
CatBoost	0.7275	1250000.0000	1562500000000.0000	900000.0000

XGBoost modeli, bağımsız değişkenlerin bağımlı değişkeni açıklama gücünü %78.75 oranında göstermektedir (R^2: 0.7875). Modelin tahminlerinde ortalama sapma yaklaşık 1.13 milyon TL olup, bu değer RMSE olarak belirtilmiştir (RMSE: 1.130.000). Hata karelerinin ortalaması ise 1.276.900.000.000 TL olarak hesaplanmıştır (MSE). Ortalama mutlak hata değeri (MAE) ise 820.000 TL'dir. Bu sonuçlar, XGBoost modelinin iyi bir performans sergilediğini göstermektedir.

LightGBM modeli ise %81.80 açıklama gücüne sahiptir (R^2: 0.8180). Modelin tahminlerinde ortalama sapma yaklaşık 1.05 milyon TL'dir (RMSE: 1.050.000) ve hata karelerinin ortalaması 1.102.500.000.000 TL olarak hesaplanmıştır (MSE). Ortalama mutlak hata değeri ise 750.000 TL'dir. LightGBM modeli, XGBoost modelinden daha iyi bir performans sergilemektedir.

CatBoost modeli ise %72.75 açıklama gücüne sahiptir (R^2: 0.7275). Modelin tahminlerinde ortalama sapma yaklaşık 1.25 milyon TL olup (RMSE: 1.250.000), hata karelerinin ortalaması 1.562.500.000.000 TL'dir (MSE). Ortalama mutlak hata değeri (MAE) ise 900.000 TL'dir. CatBoost modeli, diğer iki modele kıyasla daha düşük bir performans sergilemektedir.

Genel olarak, CBS değişkenlerinin eklenmesiyle yapılan modellemelerde LightGBM modeli en iyi performansı sergilemektedir. En yüksek R^2 (0.8180) değerine sahip olan LightGBM modeli, aynı zamanda en düşük RMSE (1.050.000) ve MAE (750.000) değerlerine sahiptir. XGBoost modeli de iyi bir performans göstermekte olup, LightGBM modeline yakın sonuçlar vermektedir. CatBoost modeli ise diğer iki modele kıyasla daha düşük performans sergilemektedir. Bu değerlendirmeler, CBS değişkenlerinin eklenmesiyle yapılan modellemelerde LightGBM modelinin en iyi performansı sergilediğini, XGBoost modelinin ise iyi bir alternatif olduğunu, ancak CatBoost modelinin diğer iki modele göre daha düşük performans gösterdiğini ortaya koymaktadır.

Hiperparametre optimizasyonu sırasında K-fold cross-validation kullanılmıştır ve k = 5 olarak belirlenmiştir. XGBoost, LightGBM ve CatBoost modellerinin

hiperparametre optimizasyonu sonrasında elde edilen en iyi hiperparametreler şu şekildedir:

- XGBoost modelinde, $colsample_bytree$ ve $subsample$ parametrelerinin 0.8 olarak ayarlanması, modelin her ağaç için daha az veri ve özellik kullanmasını sağlayarak aşırı öğrenme riskini azaltmıştır. Öğrenme oranının 0.05 olarak belirlenmesi, modelin daha yavaş ancak daha kararlı bir şekilde öğrenmesine olanak tanımıştır. Maksimum derinlik 10 ve ağaç sayısı 300 olarak belirlenmiş, bu da modelin daha karmaşık ilişkileri öğrenmesini sağlamıştır.
- LightGBM modelinde, öğrenme oranının 0.2 olarak belirlenmesi, modelin hızlı bir şekilde öğrenmesini sağlamıştır. Ancak bu durum, kısmen daha yüksek bir RMSE değeri ile sonuçlanmıştır. Maksimum derinliğin 7 olarak sınırlandırılması, modelin daha derin ağaçlar oluşturmasını engelleyerek overfitting'in önlenmesine yardımcı olmuştur.
- CatBoost modelinde, derinlik (depth = 9) ve iterasyon sayısı (iterations = 263) optimize edilerek, modelin daha karmaşık ilişkileri öğrenmesi sağlanmıştır. Öğrenme oranı (learning_rate: 0.18) ise, yeterince yüksek tutularak modelin öğrenme süreci hızlandırılmıştır.

Tabloda 4.8'deki sonuçlar, CBS değişkenlerinin eklenmesi ve hiperparametre optimizasyonuyla geliştirilmiş XGBoost, LightGBM ve CatBoost modellerinin performanslarını karşılaştırmaktadır.

Tablo 4.8 Hiperparametre optimizasyonuyla geliştirilmiş CBS modeli performans sonuçları

Model	R^2	RMSE	MSE	MAE
XGBoost	0.8449	990000.3452	980100000000.5249	750000.6838
LightGBM	0.8738	900000.3929	810000000000.0620	680000.7216
CatBoost	0.7811	1100000.9524	1210000000000.0776	830000.2222

XGBoost modeli, bağımsız değişkenlerin bağımlı değişkeni açıklama gücünü %84.49 oranında göstermektedir (R^2: 0.8449). Modelin tahminlerinde ortalama sapma yaklaşık 990,000 TL olup, bu değer RMSE olarak belirtilmiştir (RMSE: 990.000). Hata karelerinin ortalaması ise 980.100.000.000 TL olarak hesaplanmıştır (MSE). Ortalama mutlak hata değeri (MAE) ise 750.000 TL'dir. Bu sonuçlar, XGBoost modelinin oldukça iyi bir performans sergilediğini göstermektedir.

LightGBM modeli ise %87.38 açıklama gücüne sahiptir (R^2: 0.8738). Modelin tahminlerinde ortalama sapma yaklaşık 900,000 TL'dir (RMSE: 900.000) ve hata karelerinin ortalaması 810.000.000.000 TL olarak hesaplanmıştır (MSE). Ortalama mutlak hata değeri ise 680.000 TL'dir. LightGBM modeli, XGBoost modelinden daha iyi bir performans sergilemektedir.

CatBoost modeli ise %78.11 açıklama gücüne sahiptir (R^2: 0.7811). Modelin tahminlerinde ortalama sapma yaklaşık 1,100,000 TL olup (RMSE: 1.100.000), hata karelerinin ortalaması 1.210.000.000.000 TL'dir (MSE). Ortalama mutlak hata değeri (MAE) ise 830.000 TL'dir. CatBoost modeli, diğer iki modele kıyasla daha düşük bir performans sergilemektedir.

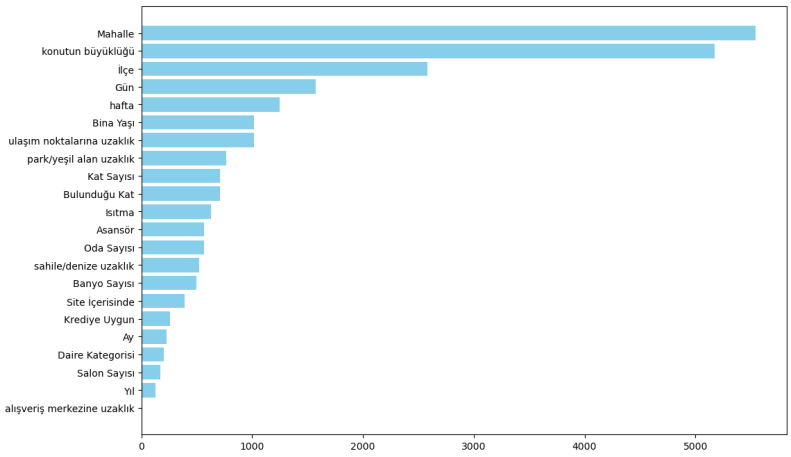
Genel olarak, CBS değişkenlerinin eklenmesi ve hiperparametre optimizasyonu yapılması ile elde edilen modeller arasında LightGBM modeli en iyi performansı sergilemektedir. En yüksek R^2 (0.8738) değerine sahip olan LightGBM modeli, aynı zamanda en düşük RMSE (900.000) ve MAE (680.000) değerlerine sahiptir. XGBoost modeli de iyi bir performans göstermekte olup, LightGBM modeline yakın sonuçlar vermektedir. CatBoost modeli ise diğer iki modele kıyasla daha düşük performans sergilemektedir. Bu değerlendirmeler, CBS değişkenleri eklenmiş ve hiperparametre optimizasyonu yapılmış modellerde LightGBM modelinin en iyi performansı sergilediğini, XGBoost modelinin de iyi bir alternatif olduğunu, ancak CatBoost modelinin diğer iki modele göre daha düşük performans gösterdiğini ortaya koymaktadır.

Performans değerlendirmesi sonucunda, CBS ile hesaplanan değişkenlerin eklenmesi, modellerin performansında belirgin bir iyileşme sağlamıştır. Temel veri seti ile yapılan tahminlerde, XGBoost modelinin R^2 skoru 0.7502 iken, CBS ile hesaplanan değişkenlerin eklenmesi sonrasında bu skor 0.7875'e yükselmiştir. Benzer şekilde, LightGBM modelinin R^2 skoru 0.7734'ten 0.8180'e yükselmiş ve CatBoost modelinin R^2 skoru 0.7056'dan 0.7275'e yükselmiştir. Hiperparametre optimizasyonu sonrasında ise, XGBoost modelinin R^2 skoru 0.8449'a, LightGBM modelinin R^2 skoru 0.8738'e ve CatBoost modelinin R^2 skoru 0.7811'e yükselmiştir.

Bu sonuçlar, coğrafi faktörlerin konut fiyatlarının tahmininde önemli bir rol oynadığını ve bu tür verilerin modellerde kullanılması gerektiğini göstermektedir. CBS ile hesaplanan uzaklık değişkenlerinin konut fiyat tahmin modellerine eklenmesi, modellerin doğruluğunu artırmıştır.

Çalışmada kullanılan ve en iyi sonucu veren LightGBM modelinin en önemli özelliklerini belirlemek amacıyla özellik önemi (*feature importance*) analizi

gerçekleştirilmiştir. Bu analiz, modelin hangi değişkenlere daha fazla ağırlık verdiğini ve konut fiyatlarını tahmin etmede hangi faktörlerin daha kritik olduğunu ortaya koymaktadır. Şekil 4.13'te, LightGBM modelinin konut fiyatlarını tahmin etmede en önemli özelliklerinin sıralaması verilmiştir.



Şekil 4.13 Özelliklerin model üzerindeki etki dereceleri

En önemli özellik "Mahalle" olup, bu değişken konutun bulunduğu mahallenin konut fiyatlarını büyük ölçüde etkilediğini göstermektedir. Mahallenin yanı sıra, "konutun büyüklüğü" de oldukça önemli bir faktördür. Konutun büyüklüğü, brüt ve net alan gibi ölçümlerle belirlenir ve bu, konutun piyasa değerinde belirleyici bir rol oynar.

Üçüncü sırada "İlçe" yer almaktadır. Bu, konutun hangi ilçede bulunduğunun da fiyat üzerinde önemli bir etkisi olduğunu gösterir. "Gün" ve "hafta" gibi zamanla ilgili değişkenler de önemlidir, bu da piyasadaki mevsimsel veya haftalık değişimlerin konut fiyatlarına yansıdığını ortaya koyar.

Diğer önemli özellikler arasında "Bina Yaşı", "Ulaşım noktalarına uzaklık", "Park/yeşil alan uzaklık", "Kat Sayısı" ve "Bulunduğu Kat" bulunmaktadır. Bu değişkenler, konutun yaşı, merkezi yerlere ve yeşil alanlara olan uzaklığı, bina kat sayısı ve konutun bulunduğu kat gibi faktörlerin fiyatı önemli ölçüde etkilediğini göstermektedir.

Daha az önemli özellikler ise "Isıtma", "Asansör", "Oda Sayısı", "Sahile/denize uzaklık", "Banyo Sayısı", "Site İçerisinde", "Krediye Uygun", "Ay", "Daire Kategorisi", "Salon Sayısı", "Yıl" ve "alışveriş merkezine uzaklık" olarak sıralanmıştır. Bu özellikler, konut fiyatlarını etkileyen faktörler arasında yer alsa da, ilk sıralardaki değişkenlere kıyasla daha az etkiye sahiptirler.

Genel olarak, bu grafik, mahalle, konut büyüklüğü ve ilçe gibi lokasyon ve

büyüklükle ilgili faktörlerin konut fiyat tahminlerinde en kritik değişkenler olduğunu göstermektedir. Bu bulgular, modelin doğruluğunu artırmak için özellikle bu özelliklerin dikkate alınmasının önemini vurgulamaktadır.

5
SONUÇ

Konut fiyatlarının tahmin edilmesi, gayrimenkul piyasasında yatırım kararları, şehir planlaması ve ekonomik analizler açısından büyük önem taşımaktadır. Konut tahminlemesi, konutların fiyatlarının belirlenmesinde çeşitli faktörlerin analiz edilmesini gerektirir. Bu faktörler arasında konutun fiziksel özellikleri, ekonomik koşullar, sosyal çevre ve coğrafi konum gibi unsurlar bulunmaktadır. Özellikle büyük ve dinamik şehirlerde, konut fiyatlarının doğru bir şekilde tahmin edilmesi, yatırımcılar ve planlamacılar için kritik bir öneme sahiptir. Bu bağlamda, CBS ve makine öğrenmesi tekniklerinin entegrasyonu, konut fiyat tahminlerinin doğruluğunu artırmak için yenilikçi ve etkili bir yaklaşım sunmaktadır.

Bu tez çalışmasında, İstanbul ili özelinde konut fiyatlarının tahmin edilmesi ve bu fiyatların CBS kullanılarak analiz edilmesi hedeflenmiştir. Çalışmanın temel amacı, konut piyasasının dinamiklerini daha iyi anlamak ve konut fiyatlarını daha doğru bir şekilde tahmin edebilmek için CBS ve makine öğrenmesi algoritmalarının entegrasyonunu araştırmaktır. Elde edilen bulgular ve sonuçlar, coğrafi faktörlerin konut fiyatlarının belirlenmesinde büyük bir öneme sahip olduğunu ortaya koymuştur. Özellikle konum, çevresel özellikler ve erişilebilirlik gibi coğrafi değişkenler, konut fiyatlarının belirlenmesinde önemli rol oynamaktadır. Bu bulgu, konut piyasası analizlerinde coğrafi verilerin kullanımının gerekliliğini vurgulamaktadır.

Çalışmada kullanılan XGBoost, LightGBM ve CatBoost gibi makine öğrenmesi algoritmaları ile konut fiyat tahmin modelleri geliştirilmiştir. Bu modeller, CBS verilerinin entegrasyonu ile önemli ölçüde iyileştirilmiştir. Özellikle, LightGBM modeli CBS verileri ile desteklendiğinde en yüksek performansı göstermiştir. CBS ile hesaplanan mesafe değişkenlerinin makine öğrenmesi modellerine eklenmesi, modellerin doğruluğunu ve performansını artırmıştır. Örneğin, LightGBM modelinin R^2 skoru, CBS değişkenlerinin eklenmesi ile 0.8394'den 0.8738'ye yükselmiştir. Bu durum, coğrafi bilgi sistemleri ve makine öğrenmesi tekniklerinin birlikte kullanılmasının konut fiyat tahminlerindeki önemini vurgulamaktadır.

Bu entegrasyon, konut fiyatlarının daha doğru tahmin edilmesine ve piyasa analizlerinin daha güvenilir sonuçlar üretmesine olanak tanımaktadır.

İstanbul'daki konut fiyatlarının dağılımını gösteren detaylı tematik haritalar oluşturulmuş ve bu haritalar, CBS ve makine öğrenmesi tekniklerinin bir arada kullanımının pratik yararlarını ortaya koymuştur. Tematik haritalar, fiyat tahminleri ve pazar analizleri için değerli bir kaynak oluşturmaktadır. Bu görselleştirmeler, konut fiyatlarının coğrafi dağılımını görsel olarak sunarak, piyasa analizlerinin ve karar verme süreçlerinin zenginleştirilmesine katkıda bulunmaktadır. Haritalar, özellikle yatırımcılar ve şehir planlamacıları için önemli bilgiler sunmakta ve bu kişilerin daha bilinçli kararlar almasına yardımcı olmaktadır.

Çalışmanın sınırlılıkları değerlendirilmiş ve gelecekteki araştırmalar için çeşitli önerilerde bulunulmuştur. CBS ve makine öğrenmesi tekniklerinin entegrasyonunun farklı coğrafi bölgelerde ve farklı veri setleri ile test edilmesi, konut piyasası analizlerinin doğruluğunu ve genellenebilirliğini artırabilir. Ayrıca, daha kapsamlı ve çeşitli veri kaynaklarının kullanımı ile modellerin performansının daha da iyileştirilebileceği öngörülmektedir. Bu bağlamda, CBS verilerinin toplanması ve işlenmesi süreçlerinde daha fazla özen gösterilmesi, verilerin doğruluğunu ve güvenilirliğini artıracaktır. Ayrıca, makine öğrenmesi algoritmalarının sürekli olarak güncellenmesi ve iyileştirilmesi, modellerin performansını ve tahmin doğruluğunu daha da artırabilir.

Bu çalışma, konut piyasası analizlerinde CBS ve makine öğrenmesi tekniklerinin entegrasyonunun önemini vurgulamakta ve bu alandaki uygulamalara katkı sağlamaktadır. Elde edilen bulgular, yatırımcılar, şehir planlamacıları ve politika yapıcılar için önemli içgörüler sunmakta ve konut fiyatlarının daha doğru tahmin edilmesine yardımcı olmaktadır. Bu entegre yaklaşım, konut piyasasının daha doğru analiz edilmesi ve gelecekteki fiyat hareketlerinin daha isabetli tahmin edilmesi açısından önemli bir metodolojik çerçeve sunmaktadır. Ayrıca, çalışmanın bulguları, konut piyasası analizlerinde kullanılan geleneksel yöntemlerin CBS ve makine öğrenmesi teknikleri ile desteklenmesi gerektiğini göstermektedir. Bu yöntemlerin entegrasyonu, konut fiyat tahminlerinin doğruluğunu artırmakla kalmayıp, aynı zamanda piyasadaki coğrafi eğilimleri ve desenleri daha net bir şekilde ortaya koymaktadır.

Bu çalışmanın literatüre sağladığı en büyük katkı, CBS ve makine öğrenmesi tekniklerinin entegrasyonunun konut fiyat tahminlerindeki etkinliğini göstermesidir. Örneğin, Tuna ve diğerlerinin (2015) çalışmasında, Ankara ilinde konut fiyatlarının lineer regresyon ve CBS kullanılarak tahmin edilmesi

üzerine yoğunlaşılmıştır [5]. Fakat bu çalışmada, makine öğrenmesi algoritmaları (XGBoost, LightGBM ve CatBoost) kullanılmış ve bu algoritmaların CBS verileri ile entegrasyonu araştırılmıştır.

Bu çalışmanın mevcut literatürden bir diğer farkı ise, İstanbul iline odaklanarak coğrafi değişkenlerin konut fiyatları üzerindeki etkisinin detaylı bir şekilde analiz edilmesidir. Özsoy ve Şahin (2009), İstanbul'daki konut fiyatlarını etkileyen başlıca parametreleri incelemiş ve sınıflandırma ve regresyon ağacı (CART) yaklaşımını kullanmışlardır [54]. Ancak, bu çalışmada daha kapsamlı bir veri seti kullanılarak, çeşitli makine öğrenmesi algoritmaları ile daha yüksek doğrulukta tahminler elde edilmiştir.

Gelecekteki araştırmalar için, CBS ve makine öğrenmesi tekniklerinin entegrasyonunun farklı coğrafi bölgelerde ve veri setlerinde test edilmesi önerilmektedir. Ayrıca, daha çeşitli ve kapsamlı veri kaynaklarının kullanılması, modellerin performansını artırabilir. Özellikle, sosyal ve ekonomik faktörlerin de dahil edilmesi, konut fiyat tahminlerinin doğruluğunu daha da artırabilir. Makine öğrenmesi algoritmalarının sürekli güncellenmesi ve iyileştirilmesi de, tahminlerin doğruluğunu artıracak önemli bir faktördür.

Sonuç olarak, bu çalışma, CBS ve makine öğrenmesi tekniklerinin entegrasyonunun konut fiyat tahminlerindeki etkinliğini ve bu alandaki uygulamalara katkı sağlayabileceğini ortaya koymaktadır. Bu entegre yaklaşım, konut piyasasının daha doğru analiz edilmesi ve gelecekteki fiyat hareketlerinin daha isabetli tahmin edilmesi açısından önemli bir metodolojik çerçeve sunmaktadır. Bu bağlamda, araştırma, sektördeki paydaşların daha bilinçli ve veriye dayalı kararlar almasına yardımcı olmayı hedeflemektedir. Ayrıca, bu çalışma, gelecekteki araştırmalar için önemli bir temel oluşturarak, CBS ve makine öğrenmesi tekniklerinin entegrasyonunun daha geniş bir coğrafi alanda ve farklı veri setleri ile test edilmesi gerektiğini vurgulamaktadır. Bu şekilde, konut piyasası analizlerinin doğruluğu ve genellenebilirliği artırılabilir ve sektörde daha etkili stratejiler geliştirilmesine olanak tanınabilir.

KAYNAKÇA

[1] M. O. Mete, T. Yomralıoğlu, "Cbs ve makine öğrenmesi yöntemleri ile konut amaçlı taşınmazların toplu değerlemesi," in *TUFUAB -MERSIN 2022*, 2022.

[2] M. Yalçın, "Türkiye konut satış İstatistiklerinin cbs tabanlı analizleri," *Afyon Kocatepe Üniversitesi Fen Ve Mühendislik Bilimleri Dergisi*, vol. 20, no. 3, pp. 448–455, 2020.

[3] X. Gao, Y. Asami, "Influence of spatial features on land and housing prices," *Tsinghua Science and technology*, vol. 10, no. 3, pp. 344–353, 2005.

[4] S. Yalpir, S. S. Durduran, F. B. Unel, M. Yolcu, "Creating a valuation map ın gıs through artificial neural network methodology: A case study.," *Acta Montanistica Slovaca*, vol. 19, no. 2, 2014.

[5] M. F. Tuna, T. Türk, O. Kitapçı, "Lineer regresyon ve coğrafi bilgi sistemleri yardımıyla ev fiyatlarının tahmin edilmesi: Ankara örneği," *TMMOB HKMO*, vol. 15, 2015.

[6] S. G. Kangallı Uyar, N. Yayla, "Konut fiyatlarının hedonik fiyatlama yaklaşımına göre mekânsal ekonometrik modeller ile tahmini: İstanbul konut piyasası örneği," 2016.

[7] D. Oud, "Gıs based property valuation," Ph.D. dissertation, MS Thesis, Dept. of Geographical sciences-geo-information applications and . . ., 2017.

[8] S. Erdoğan, A. Memduhoğlu, "A spatio-temporal exploratory analysis of real estate sales in turkey using gıs," *Journal of European Real Estate Research*, vol. 12, no. 2, pp. 207–226, 2019.

[9] A. Bozdağ, E. Ertunç, "Cbs ve ahp yöntemi yardımıyla niğde kenti örneğinde taşınmaz değerleme," *Geomatik*, vol. 5, no. 3, pp. 228–240, 2020.

[10] C. Xue, Y. Ju, S. Li, Q. Zhou, Q. Liu, "Research on accurate house price analysis by using gıs technology and transport accessibility: A case study of xi'an, china," *Symmetry*, vol. 12, no. 8, p. 1329, 2020.

[11] S. E. Taşabat, M. Ersen, "House price prediction: A case study for ıstanbul," in *Industry 4.0 and the Digital Transformation of International Business*, Springer, 2023, pp. 233–250.

[12] S. Şişman, A. Ç. Aydınoğlu, "Toplu taşınmaz değerlemesi için coğrafi verinin İşlenmesi ve rastgele orman makine öğrenmesi tekniğiyle tahmin modelinin geliştirilmesi," *Harita Dergisi*, vol. 169, pp. 13–27, 2023.

[13] G. Adana Karaağaç, "Coğrafi bilgi sistemleri yardımıyla gayrimenkul değer artışı haritalarının oluşturulması: Eskişehir ili tepebaşı ilçesi örneği," M.S. thesis, Sosyal Bilimler Enstitüsü, 2018.

[14] G. J. Meaden, J. Aguilar-Manjarrez, "Advances in geographic information systems and remote sensing for fisheries and aquaculture," *FAO fisheries and aquaculture technical paper*, no. 552, p. I, 2013.

[15] S. Timur, "Coğrafi bilgi sistemleri destekli taşınmaz değer haritalarının oluşturulması: İstanbul ili, şişli ilçesi örneği," Ph.D. dissertation, Fen Bilimleri Enstitüsü, 2009.

[16] İ. Balcı, H. Çoban, E. Mehmet, "Coğrafi bilgi sistemi," *Turkish Journal of Forestry*, vol. 1, no. 1, pp. 115–132, 2009.

[17] G. Töreyen, İ. Özdemir, T. Kurt, "Arcgıs 10 desktop uygulama dokümanları," *Coğrafi Bilgi Sistemleri Mühendislik ve Eğitim Ltd. Şti*, pp. 1–10, 2010.

[18] V. Tecim, *Coğrafi bilgi sistemleri: Harita tabanlı bilgi yönetimi*. Vahap Tecim, 2008.

[19] A. Cabuk *et al.*, *Coğrafi Bilgi Sistemlerinde Analiz ve Yorumlama*. Anadolu Üniversitesi, 2010, ısbn: 978-975-06-0791-2.

[20] U. Çağatay, A. BayrakdaroĞlu, "Konut finansman sisteminde coğrafi bilgi sistemlerinin kullanılabilirliği," *Uluslararası Alanya İşletme Fakültesi Dergisi*, vol. 5, no. 1, pp. 101–112, 2013.

[21] M. P. Ralphs, P. Wyatt, *GIS in land and property management*. Taylor & Francis, 2003.

[22] R. L. Church, A. T. Murray, *Business site selection, location analysis, and GIS*. John Wiley & Sons Hoboken, NJ, 2009.

[23] *1000 gıs applications uses – how gıs ıs changing the world*. [Online]. Available: `https://gisgeography.com/gis-applications-uses/` (visited on 03/13/2024).

[24] P. A. Longley, M. F. Goodchild, D. J. Maguire, D. W. Rhind, *Geographic information science and systems*. John Wiley & Sons, 2015.

[25] S. Banerjee, B. P. Carlin, A. E. Gelfand, *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC, 2003.

[26] M. Goodchild, "Twenty years of progress: Gıscience in 2010," *Journal of spatial information science*, no. 1, pp. 3–20, 2010.

[27] C. Brannstrom, J. M. Vadjunec, *Land change science, political ecology, and sustainability*. Routeledge, 2013.

[28] M. Işıklı, "Coğrafi bilgi sistemleri ile taşınmaz değerlemesi," *Yapı Bilgi Modelleme*, vol. 1, no. 1, pp. 21–26, 2019.

[29] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.

[30] R. Behl, I. Kashyap, "Machine learning classifiers," in *Big Data, IoT, and Machine Learning*, CRC Press, 2020, pp. 3–36.

[31] A. L. Tarca, V. J. Carey, X.-w. Chen, R. Romero, S. Drăghici, "Machine learning and its applications to biology," *PLoS computational biology*, vol. 3, no. 6, e116, 2007.

[32] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of research and development*, vol. 3, no. 3, pp. 210–229, 1959.

[33] W.-M. Lee, *Python machine learning*. John Wiley & Sons, 2019.

[34] L. D. Schroeder, D. L. Sjoquist, P. E. Stephan, *Understanding regression analysis: An introductory guide*. Sage Publications, 2016, vol. 57.

[35] M. Özşahin, K. B. Yılmaz, S. Akcan, "Makine öğrenmesi algoritmaları ile ev kirası tahmini," *Araştırma ve Değerlendirmeler II*, p. 23, 2022.

[36] X. Zhu, A. B. Goldberg, *Introduction to semi-supervised learning*. Springer Nature, 2022.

[37] L. P. Kaelbling, M. L. Littman, A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[38] S. Bahoo, M. Cucculelli, X. Goga, J. Mondolo, "Artificial intelligence in finance: A comprehensive review through bibliometric and content analysis," *SN Business & Economics*, vol. 4, no. 2, p. 23, 2024.

[39] V. K. Verma, S. Verma, "Machine learning applications in healthcare sector: An overview," *Materials Today: Proceedings*, vol. 57, pp. 2144–2147, 2022.

[40] E. W. Ngai, Y. Wu, "Machine learning in marketing: A literature review, conceptual framework, and research agenda," *Journal of Business Research*, vol. 145, pp. 35–48, 2022.

[41] K. Chowdhary, K. Chowdhary, "Natural language processing," *Fundamentals of artificial intelligence*, pp. 603–649, 2020.

[42] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[43] C. Chen, A. Seff, A. Kornhauser, J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2722–2730.

[44] J. Han, J. Pei, H. Tong, *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[45] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, "Catboost: Unbiased boosting with categorical features," *arXiv preprint arXiv:1706.09516*, 2018.

[46] V. Limsombunchao, "House price prediction: Hedonic price model vs. artificial neural network," 2004.

[47] T. Chen, C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[48] G. Ke *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017, pp. 3146–3154.

[49] M. Zhu, Y. Zhang, Y. Gong, K. Xing, X. Yan, J. Song, "Ensemble methodology: Innovations in credit default prediction using lightgbm, xgboost, and localensemble," *arXiv preprint arXiv:2402.17979*, 2024.

[50] J. Hancock, T. M. Khoshgoftaar, "Catboost for big data: An ınterdisciplinary review," *Research Square*, 2020. doı: 10.21203/rs.3.rs-54646/v2.

[51] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, "Catboost: Gradient boosting with categorical features support," *arXiv preprint arXiv:1810.11363*, 2018.

[52] *R squared: Understanding the coefficient of determination*. [Online]. Available: `https://arize.com/blog-course/r-squared-understanding-the-coefficient-of-determination/`.

[53] H. Akbal, "Konut seçimini etkileyen faktörlerin incelenmesi: Balıkesir-akçay örneği," *Pamukkale Üniversitesi İşletme Araştırmaları Dergisi*, vol. 9, no. 2, pp. 332–341, 2022.

[54] O. Özsoy, H. Şahin, "Housing price determinants in ıstanbul, turkey: An application of the classification and regression tree model," *International Journal of Housing Markets and Analysis*, vol. 2, no. 2, pp. 167–178, 2009.

TEZDEN ÜRETİLMİŞ YAYINLAR

Konferans Bildirisi

1. Atasever D, Öz E., "Konut Satış Tahminlerinin Coğrafi Bilgi Sistemi Tabanlı Analizi: İstanbul İli Örneği", V. International Applied Statistic Congress 2024, İstanbul, Türkiye