Gazi Üniversitesi
Fen Bilimleri Dergisi
PART C: TASARIM VE TEKNOLOJİ

Gazi University
Journal of Science
PART C: DESIGN AND TECHNOLOGY

GU J Sci, Part C, 13(1): 10-28 (2025)

Optuna Tabanlı Hiper Parametre Optimizasyonu ile Konut Fiyat Tahminlemede Makine Öğrenmesi Tekniklerinin Karşılaştırmalı Analizi

Vahid SİNAP[1*]

[1]*Ufuk Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Yönetim Bilişim Sistemleri Bölümü, Ankara, Türkiye*

Makale Bilgisi

Araştırma makalesi
Başvuru: 07/09/2024
Düzeltme: 09/12/2024
Kabul: 11/12/2024

Anahtar Kelimeler

Konut Fiyat Tahmini
Makine Öğrenmesi
Performans Karşılaştırması
Hiper Parametre Optimizasyonu
Optuna

Article Info

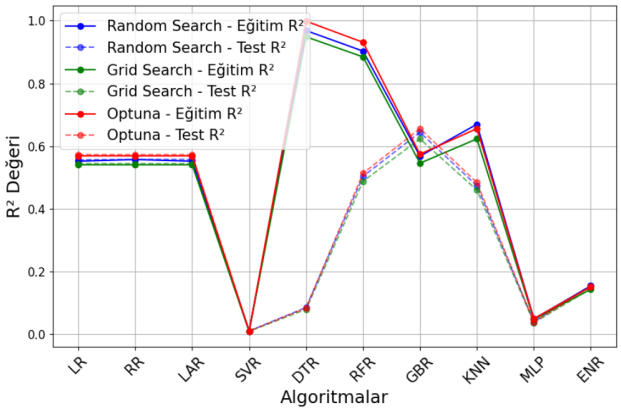
Research article
Received: 07/09/2024
Revision: 09/12/2024
Accepted: 11/12/2024

Keywords

House Price Prediction
Machine Learning
Performance Comparison
Hyperparameter Optimization
Optuna

Grafik Özet (Graphical/Tabular Abstract)

Bu çalışma, konut fiyatlarını tahmin etmek için 10 farklı regresyon algoritmasını ve çeşitli hiper parametre optimizasyon yöntemlerini karşılaştırmıştır. Optuna ile optimize edilen Gradyan Artırma Regresyonu modeli, yüksek R^2 (0.6558) ve düşük RMSE (4469.48) değerleriyle en başarılı model olmuştur. / This study compared 10 regression algorithms and various hyperparameter optimization methods for predicting housing prices. The Gradient Boosting Regression model optimized with Optuna emerged as the best, achieving a high R^2 (0.6558) and low RMSE (4469.48), demonstrating Optuna's precision and effectiveness in hyperparameter optimization.



Şekil A: *Modellerin doğruluk karşılaştırması /* ***Figure A:*** *Accuracy comparison of models*

Önemli noktalar (Highlights)

- *Optuna, regresyon modellerinde hiperparametre optimizasyonunda hassasiyet ve etkinlik avantajları sunmaktadır. / Optuna provides precision and efficiency advantages in hyperparameter optimization for regression models.*
- *Gradyan Artırma Regresyonu, RMSE ve R^2 metriklerine göre konut fiyatlarını tahmin etmede diğer modelleri geride bırakmıştır. / Gradient Boosting Regressor outperforms other models in predicting house prices based on RMSE and R^2 metrics.*
- *Makine öğrenmesi yöntemleri, konut fiyatı belirleyicilerindeki doğrusal olmayan etkileşimleri modellemede geleneksel yöntemlerden daha yüksek doğruluk oranlarına ulaşmıştır. / Machine learning methods achieve higher accuracy than traditional methods for modeling non-linear interactions in house price determinants.*

Amaç (Aim): *Optuna tabanlı hiperparametre optimizasyonu kullanarak konut fiyatı tahmini için en başarılı makine öğrenmesi algoritmasını belirlemek. / To identify the most successful machine learning algorithm for house price prediction using Optuna-based hyperparameter tuning.*

Özgünlük (Originality): *Bu çalışma, ekonomik sürdürülebilirlik boyutlarına odaklanarak üç hiperparametre ayarlama stratejisini dahil eden ve 10 farklı denetimli regresyon modelini karşılaştıran bir analiz sunmaktadır. / This study compares 10 different supervised regression models while incorporating three hyperparameter tuning strategies, highlighting the economic sustainability aspects of automatic forecasting systems.*

Bulgular (Results): *Optuna ile optimize edilen Gradyan Artırma Regresyonu, RMSE 4469.48 ve R^2 0.6558 ile test setinde en iyi sonuçları elde ederek diğer yöntemleri geride bırakmıştır. / Optuna-optimized Gradient Boosting Regressor achieved the best test results with an RMSE of 4469.48 and an R^2 of 0.6558, outperforming alternatives.*

Sonuç (Conclusion): *Optuna, en verimli hiperparametre ayarlama yöntemi olarak öne çıkmakta olup, makine öğrenmesi modelleri konut fiyatlarını tahmin etmede geleneksel yöntemlere kıyasla önemli avantajlar sunmaktadır. / Optuna stands out as the most efficient hyperparameter tuning method, and machine learning models offer significant advantages over traditional methods in predicting house prices.*

*Corresponding author, e-mail: vahidsinap@gmail.com
DOI: 10.29109/gujsc.1544987

GU J Sci, Part C, 13(1): 10-28 (2025)

Gazi Üniversitesi
Fen Bilimleri Dergisi
PART C: TASARIM VE TEKNOLOJİ

Gazi University
Journal of Science
PART C: DESIGN AND TECHNOLOGY

http://dergipark.gov.tr/gujsc

Optuna Tabanlı Hiper Parametre Optimizasyonu ile Konut Fiyat Tahminlemede Makine Öğrenmesi Tekniklerinin Karşılaştırmalı Analizi

Vahid SİNAP[1*]

[1]*Ufuk Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Yönetim Bilişim Sistemleri Bölümü, Ankara, Türkiye*

Makale Bilgisi

Araştırma makalesi
Başvuru: 07/09/2024
Düzeltme: 09/12/2024
Kabul: 11/12/2024

Anahtar Kelimeler

Konut Fiyat Tahmini
Makine Öğrenmesi
Performans Karşılaştırması
Hiper Parametre Optimizasyonu
Optuna

Öz

Konut fiyatlarının etkili bir şekilde tahmin edilmesi, ekonominin şekillenmesinde kritik bir rol oynamaktadır. Bu çalışmanın amacı, konut fiyatlarını tahminlemede en iyi performans gösteren makine öğrenmesi modelini belirlemektir. Bu amaçla, 10 farklı denetimli regresyon algoritması kullanılarak çeşitli modeller eğitilmiştir. Modellerin performansını optimize etmek amacıyla Grid Search, Random Search ve Optuna gibi hiper parametre ayarlama yöntemleri uygulanmıştır. Eğitim ve test setlerinde elde edilen metrik değerler, modellerin genel performansını değerlendirmek için kullanılmıştır. Araştırma sonuçları, hiper parametre ayarlama yöntemlerinin modellerin genel başarısını etkileyen kritik bir faktör olduğunu göstermiştir. Optuna ile optimize edilen Gradyan Artırma Regresyonu modeli, test veri setinde elde ettiği yüksek R^2 değeri (0.6558) ve düşük RMSE değeri (4469.48) ile konut fiyatlarını tahminlemede en başarılı model olarak belirlenmiştir. Optuna, hiper parametre optimizasyonunda sağladığı hassasiyet ve etkinlik ile diğer yöntemlere kıyasla belirgin bir üstünlük sunmuştur.

A Comparative Analysis of Machine Learning Techniques for House Price Prediction with Optuna-Based Hyperparameter Optimization

Article Info

Research article
Received: 07/09/2024
Revision: 09/12/2024
Accepted: 11/12/2024

Keywords

House Price Prediction
Machine Learning
Performance Comparison
Hyperparameter Optimization
Optuna

Abstract

Effectively predicting house prices plays a critical role in shaping the economy. This study aims to identify the best-performing machine learning model for predicting house prices. For this purpose, various models were trained using 10 different supervised regression algorithms. Hyperparameter tuning methods such as Grid Search, Random Search, and Optuna were applied to optimize the performance of these models. Metric values obtained from the training and test sets were used to evaluate the overall performance of the models. The research results indicate that hyperparameter tuning methods are a critical factor influencing the overall success of the models. The Gradient Boosting Regressor model optimized with Optuna was identified as the most successful model for predicting house prices, achieving a high R^2 score (0.6558) and a low RMSE value (4469.48) on the test dataset. Optuna demonstrated a significant advantage in hyperparameter optimization compared to other methods due to its precision and efficiency.

1. GİRİŞ (INTRODUCTION)

Konut fiyatlarının etkili bir şekilde tahmin edilmesi, ekonominin şekillenmesinde kritik bir rol oynamaktadır. Bu önemli konunun altında yatan sebeplerden biri, gayrimenkul sektöründeki dalgalanmaların önceden tahmin edilebilmesi ve bu sayede ekonomik istikrarın sürdürülebilir bir şekilde sağlanabilmesidir. Gayrimenkul piyasasındaki ani çalkantılar genellikle ekonomik dengesizliklere yol açmakta ve bu durum, konut talebi ile arzı dengelemede zorluklar yaşanmasına neden olabilmektedir [1]. Bununla birlikte, konut fiyatlarının doğru tahmin edilmesi aynı zamanda hükümetlerin gayrimenkul piyasasını daha iyi düzenleyebilmesi anlamına gelmektedir [2]. Bu düzenleme, ekonominin sürdürülebilir bir büyüme patikasında ilerlemesine katkı sağlamaktadır. Hükümetler, konut talebi ile arzını dengeleyerek, gayrimenkul sektöründeki dengesizlikleri önleyerek ekonomik istikrarı koruyabilmektedir [3]. Diğer bir etken, konut üreticilerinin zamanında ve bilinçli

*Corresponding author, e-mail: vahidsinap@gmail.com
DOI: 10.29109/gujsc.1544987

yatırım kararları almasını sağlamaktır. Konut fiyatlarının doğru bir şekilde tahmin edilmesi, konut projeleri geliştiren şirketlere piyasadaki değişimlere önceden uyum sağlama imkânı sunmaktadır [4]. Örneğin, bölgesel altyapı projeleri veya kentsel dönüşüm planları gibi faktörlerin etkisi önceden tahmin edilebilirse, şirketler projelerini buna göre planlayabilir ve rekabet avantajı elde edebilirler. Konut sadece bir bireyin temel ihtiyacını karşılamakla kalmayıp, aynı zamanda bir yatırım şekli olarak da önemlidir [5]. Bu nedenle, konut fiyatlarının doğru bir şekilde tahmin edilmesi, alıcılar ve satıcılar için büyük bir ilgi konusudur [6]. Gerçekçi fiyat tahminleri, alıcıların bütçelerine uygun konut seçeneklerini değerlendirmelerine ve satıcıların rekabetçi bir fiyat belirlemelerine yardımcı olabilmektedir.

Geleneksel olarak, konut fiyatı tahmininde çoğunlukla profesyonel değerleme uzmanlarına başvurulmaktadır. Bu uzmanlar, gayrimenkul değerlemesi yaparak, konutların piyasa değerini belirlemeye yönelik geleneksel yöntemleri uygulamaktadırlar [7]. Değerleme sürecinde konutun fiziksel özellikleri (örneğin, oda sayısı, konutun yaşına ilişkin bilgiler), coğrafi konumu, çevresel faktörler ve benzeri unsurlar dikkate alınmaktadır. Geleneksel değerleme yöntemleri arasında sıkça kullanılan modellerden biri hedonik regresyon analizidir. Bu model, konut fiyatını etkileyen çeşitli faktörleri inceleyerek, bu faktörlerin konutun değeri üzerindeki etkisini analiz etmeyi amaçlamaktadır [8]. Ancak, bu geleneksel yöntemlerin bazı sınırlamaları bulunmaktadır. Örneğin, hedonik regresyon modeli, bazı durumlarda model varsayımlarının ihlal edilmesine duyarlı olabilir ve doğrusal olmayan ilişkileri yeterince ele alamayabilir [9]. Ayrıca, hedonik fiyat modeli, model varsayımları, tahmin ve doğrusal olmayan sorunları çözme konusunda yetersiz kalabilmektedir [10]. Buna ek olarak, konut fiyatlarını tahmin etmek için geleneksel yöntemler genellikle bireysel öznelliklere dayanır ve bu durum, değerleme uzmanlarının önyargılı olma riskini artırabilir [11]. Bunun yanı sıra, bu yöntemler, geniş veri setlerini işlemekte zorlanabilir ve özellikle karmaşık, dinamik piyasa koşullarında doğru tahminler yapma konusunda sınırlamalara sahiptir [12]. Bu noktada, bağımsız bir üçüncü taraf kaynağı olarak hizmet edebilecek otomatik bir tahmin sistemi, daha az önyargılı bir yaklaşım sunabilir.

Otomatik tahmin sistemleri, makine öğrenmesi veya istatistiksel yöntemleri kullanarak, belirli bir olayın veya durumun gelecekteki olası sonuçlarını tahmin etmeye odaklanan bilgisayar tabanlı sistemlerdir [13]. Bu sistemler, büyük miktarda veriyi analiz ederek desenleri tanımlamakta ve bu desenlere dayanarak gelecekteki olayları öngörmeye çalışmaktadır. Otomatik tahmin sistemleri, büyük veri setlerini analiz ederek desenleri belirlemek ve gelecekteki olayları tahmin etmek için makine öğrenmesi tekniklerini kullanır. Bu nedenle, makine öğrenmesi tabanlı tahmin sistemleri, otomatik tahmin sistemlerinin daha gelişmiş ve veri odaklı bir versiyonunu temsil etmektedir. Makine öğrenmesi tabanlı tahmin sistemleri, belirli bir algoritma tarafından öğrenilen modelleri kullanarak veri setlerinden öğrenme yeteneğine sahiptir [14]. Bu modeller, karmaşık ilişkileri ele alarak değişken koşullara uyum sağlayabilmektedir. Otomatik tahmin sistemleri, büyük veri setlerini analiz edebilme, hızlı öğrenme ve gerçek zamanlı adaptasyon gibi avantajlar sunmaktadır.

Makine öğrenmesi teknikleri, konut fiyat tahmininde geleneksel yöntemlere kıyasla önemli avantajlar sunmaktadır. Bu teknikler, esnek modelleme yetenekleri ile öne çıkmaktadır. Geleneksel değerleme yöntemleri genellikle lineer ilişkileri ele alabilirken, makine öğrenmesi modelleri karmaşık ve doğrusal olmayan ilişkileri daha etkili bir şekilde modelleyebilmektedir [15]. Bu, konut fiyatlarını etkileyen faktörler arasındaki daha ince ve karmaşık ilişkilerin anlaşılmasını sağlamaktadır. Şehir planlama süreçlerinde, altyapı projelerinin ve kentsel dönüşüm planlarının konut fiyatları üzerindeki etkisini doğru bir şekilde modelleyerek, kamu yatırımlarının etkinliğini artırmak mümkün hale gelebilir. Benzer şekilde, yatırım analizlerinde, piyasadaki değişkenlerin karmaşık etkileşimlerini değerlendirerek daha öngörülebilir risk ve getiri analizleri yapılabilir. Ayrıca, makine öğrenmesi modelleri büyük veri setleriyle daha etkili bir şekilde çalışabilmektedir [16]. Konut piyasasındaki çeşitli değişkenlerin karmaşıklığı göz önüne alındığında, geniş veri setlerini işleme yetenekleri, daha kapsamlı ve doğru tahminlere olanak tanımaktadır. Bu modeller, konut fiyatlarını etkileyen pek çok değişkeni aynı anda değerlendirebilmekte ve bu faktörler arasındaki etkileşimleri analiz edebilmektedir. Şehir planlamasında, geniş veri setlerinden elde edilen öngörüler, yerel yönetimlerin konut talebine uygun projeler geliştirmesine yardımcı olabilir. Yatırım analizleri açısından ise bu modeller, yatırımcıların pazar trendlerini daha iyi anlamasını ve daha bilinçli yatırım kararları almasını sağlayabilir. Makine öğrenmesi teknikleri, öğrenme yetenekleri sayesinde zaman içindeki değişen konut piyasası koşullarına uyum sağlayabilmektedir. Bu, modelin güncel ve dinamik verilere dayalı olarak sürekli

olarak iyileştirilebilmesine olanak tanımaktadır. Örneğin, makine öğrenmesi modelleri, ani piyasa değişikliklerine hızlı bir şekilde uyum sağlayarak, şehir planlama kararlarının ve yatırım stratejilerinin gerçek zamanlı olarak optimize edilmesine katkı sunabilir. Böylece hem kamu sektöründe hem de özel sektörde daha stratejik ve etkili kararlar alınabilir.

Makine öğrenmesi modellerinin faydalarının yanı sıra bazı zorlukları da bulunmaktadır. Bu modeller, genellikle karmaşık yapılara dayandığı için anlaşılması ve yorumlanması zor olabilmektedir [17]. Modelin içsel mekanizmalarının şeffaflık eksikliği, karar süreçlerinin bilinmezliğine yol açabilir ve bu da güvenilirlik sorunlarına neden olabilir [18]. Ayrıca, makine öğrenmesi modelleri, geniş veri setlerini etkili bir şekilde işleyebilme yeteneklerine rağmen, veri setlerindeki gürültü ve anlamsız ilişkilerle başa çıkma konusunda zorluklar yaşayabilir [19]. Yanlılık ve varyans arasındaki dengeyi bulma sürecinde modelin aşırı öğrenme veya yetersiz öğrenme eğiliminde olması, tahminlerin doğruluğunu etkileyebilir [20]. Bunlara ek olarak, makine öğrenmesi modelleri, belirli bir döneme veya bağlam içerisindeki geçmiş verilere aşırı bağımlı hale gelebilir [21]. Bu durum, modellerin gelecekteki beklenmeyen olaylara tepki verme yeteneklerini sınırlayabilir ve tahminlerin güvenilirliğini azaltabilir.

Makine öğrenmesi teknikleri konut fiyat tahmininde önemli avantajlar sunsa da kullanımlarıyla ilgili bu zorlukları anlamak ve ele almak önemlidir. Güvenilirlik ve genel model performansı açısından dengeli bir yaklaşım benimsemek, makine öğrenmesi tabanlı tahmin modellerinin etkin bir şekilde kullanılması açısından kritiktir. Bu bağlamda araştırmanın amacı, konut fiyatlarını tahminlemede en iyi performans gösteren makine öğrenmesi modelinin tespitini yapmaktır. Araştırmada, Doğrusal Regresyon (Linear Regression - LR), Ridge Regresyonu (Ridge Regression - RR), Lasso Regresyonu (Lasso Regression - LAR), Destek Vektör Regresyonu (Support Vector Regression - SVR), Karar Ağacı Regresyonu (Decision Tree Regression - DTR), Rastgele Orman Regresyonu (Random Forest Regression - RFR), Gradyan Artırma Regresyonu (Gradient Boosting Regression - GBR), K-En Yakın Komşu Regresyonu (K-Nearest Neighbors Regression - KNN), Çok Katmanlı Algılayıcı Regresyonu (Multilayer Perceptron Regression - MLP), ElasticNet Regresyonu (ElasticNet Regression - ENR) olmak üzere 10 denetimli regresyon algoritması kullanılmıştır. Makine öğrenmesi modellerinin oluşturulması sırasında veri ön işleme aşamaları ve modellerin performansını etkileyen durumlar ayrıntılı ele alınarak alanda yapılacak gelecekteki araştırmalara bir yol haritası çizilmesi hedeflenmektedir. Araştırmanın bir diğer önemli amacı, Grid Search, Random Search ve Optuna hiper parametre ayarlama yöntemlerinin kullanılmasıyla hiper parametrelerin optimize edilmesi ve bu optimizasyonların model performansları üzerindeki etkilerinin incelenmesidir. Buna göre, farklı makine öğrenmesi modellerinin çeşitli hiper parametre ayarları altında tahminleme yeteneklerinin, performans metriklerine dayalı olarak objektif bir şekilde değerlendirilmesi, araştırmanın ana odak noktalarını oluşturmaktadır. Ayrıca, otomatik tahmin sistemlerinin kullanımının, konut piyasasındaki dalgalanmaların önceden tahmin edilmesi ve bu bilgilerin ekonomik istikrarın sürdürülebilirliğine nasıl katkı sağladığının değerlendirilmesi amaçlanmaktadır. Belirlenen amaç ve hedeflere ulaşılması dahilinde, konut fiyat tahmininde en etkili ve güvenilir modelin belirlenmesine yönelik kapsamlı bir değerlendirme elde edilecektir. Araştırma bulgularının, konut sektöründeki paydaşlara, hükümetlere ve ekonomi uzmanlarına daha bilinçli kararlar almalarında rehberlik etme potansiyeline sahip olacağı ön görülmektedir.

2. İLGİLİ ARAŞTIRMALAR (RELATED WORKS)

Bu araştırmada, konut fiyatlarını daha etkili bir şekilde tahmin etmek için çeşitli makine öğrenmesi algoritmalarının analizi gerçekleştirilerek karşılaştırmaları yapılmıştır. Konut fiyatlandırmasındaki trendler, mevcut ekonomik durumu göstermekte ve doğrudan alıcılar ve satıcılarla ilgili bir konu olarak karşımıza çıkmaktadır. Bir evin gerçek fiyatı birçok faktöre bağlıdır. Bunlar arasında yatak odası sayısı, banyo sayısı ve konum gibi faktörler bulunmaktadır. Kırsal bölgelerde fiyatlandırma genellikle şehirlere göre daha düşüktür. Ev fiyatları, otoyola, alışveriş merkezine, süpermarkete, iş olanaklarına, iyi eğitim tesislerine gibi faktörlere yakınlıkla artmaktadır. Gayrimenkul şirketlerinin birkaç yıl öncesine kadar mülk fiyatını manuel olarak tahmin etmeye çalıştıkları bilinmektedir. Gayrimenkul satışı yapan şirketlerde genellikle herhangi bir gayrimenkul mülkünün fiyatını tahmin etmek için özel bir yönetim ekibi bulunmaktadır. Ancak, bu manuel tahminlerde, alıcılar ve satıcılar için önemli bir kayba neden olacak, ortalama %25 civarında hata oluşabilmektedir. Bu nedenle, ev fiyatlarının daha etkili ve tutarlı bir şekilde belirlenmesine yönelik birçok araştırma yapılmıştır.

Lu ve diğerleri [22] tarafından gerçekleştirilen araştırmada gelişmiş bir ev fiyat tahminleme sistemi önerilmiştir. Bu sistem, çeşitli özelliklere dayalı olarak iyi bir ev fiyat tahmini sunan etkili bir makine öğrenmesi modeli içermektedir. Bu sayede ev satın alacakların bütçelerine ve önceliklerine göre makul bir fiyatın belirlenmesi hedeflenmiştir. Modelin geliştirilmesinde hibrit regresyon tekniği kullanılmıştır. Ayrıca, çalışmada sınırlı veri seti ve veri özellikleri ile özellik mühendisliği (feature engineering) yöntemleri incelenmiştir. Çalışmada oluşturulan regresyon modeli ile veri setindeki özniteliklere bağlı olarak tahminlemeler gerçekleştirilmiş ve oluşturulan modellerin ev fiyatlarını tahminlemede önemli bir başarı elde ettiği vurgulanmıştır. Bu şekilde, ev piyasasındaki dalgalanmalara karşı daha dirençli ve kullanıcı dostu bir tahmin modeli oluşturulmuştur.

Durganjali ve Pujitha [23] sınıflandırma algoritmalarını kullanarak ürettikleri model ile ev yeniden satış fiyatı tahmini gerçekleştirmişlerdir. Araştırmada, evin yeniden satış fiyatının tahmininde Lineer Regresyon, Karar Ağacı, K-Means ve Rastgele Orman gibi farklı sınıflandırma algoritmaları kullanılmıştır. Ev fiyatını etkileyen birçok faktör bulunmaktadır. Bu faktörler arasında fiziksel özellikler, konum ve ekonomik koşullar yer almaktadır. Çeşitli performans metrikleri ile bu algoritmaların farklı veri setleri üzerindeki performansı değerlendirilmiştir. Sonuç olarak, Rastgele Orman'ın eğitim verilerine göre en iyi sonucu verdiği bulunmuştur.

Rahadi ve diğerleri [24] tarafından yürütülen bir araştırma, Jakarta, Endonezya'daki konut fiyatlarını, kavramsal model (conceptual model) ve anketler kullanarak analiz etmiştir. Araştırmanın temel amacı ev fiyatını etkileyen faktörleri sınıflandırmaktır. Araştırma sonuçlarına göre, her bir konutun fiyatını etkileyen özelliklerin farklılık gösterdiği belirlenmiştir. Bu faktörler arasında konum, yapısal özellikler ve çevresel koşullar öne çıkan başlıklar olmuştur.

Konut fiyat tahmini çalışmalarında, kullanılan makine öğrenmesi modelleri kadar veri setlerinde bulunan konut özellikleri de öneme sahiptir. Kurulan modellerin daha iyi performans verebilmesi için konut özelliklerinin doğru yorumlanması gerekmektedir. Literatürde konut özellikleri üzerine bazı değerli araştırmalar bulunmaktadır.

Konum, ev fiyatı belirlemede en önemli özellik olarak kabul edilmektedir [24-25]. Osmadi ve diğerleri [26] tarafından yapılan çalışmalarda, konut fiyatlarını etkileyen konum özelliklerinin önemi gözlemlenmiştir. Çalışmada mülkün konum özelliklerinin önemi üzerine bir gözlem yapılmıştır. Mülkün konumu, en yakın alışveriş merkezine olan mesafe veya tepeleri veya sahili gösteren konum özelliklerini içeren sabit bir konum özelliğine ayrılmıştır. Çalışma, alışveriş merkezine olan mesafe veya tepeleri veya sahili gösteren konum özellikleri gibi konumsal özelliklerin fiyatlandırma ile yakın ilişkiye sahip olduğunu göstermiştir.

Ev fiyatını etkileyen diğer önemli bir özellik de fiziksel veya yapısal özellikler olarak belirtilmektedir [25,27]. Bir evin yapısal özellikleri yatak odası ve banyo sayısı, kat alanı, garaj veya veranda olup olmaması gibi özellikleri içermektedir. Potansiyel alıcıları çekmek ve evleri cazip kılmak için inşaat firmaları da söz konusu yapısal özelliklere dikkat ederek, potansiyel alıcıların isteklerini karşılamayı hedeflemektedirler. Ball [28] çalışmasında, yapısal özelliklerin ev alacaklar için ne satın alacaklarını belirlemede temel bir düşünce olacağını belirtmiştir. Rodriguez ve Sirmans [29] gerçekleştirdikleri çalışmada, yapısal özelliklerin ev fiyatlarının artmasında önemli bir etkiye sahip olduğunu vurgulamıştır.

Çevresel özellikler, ev fiyatını belirlemede dahil edilebilecek diğer bir faktördür. Chau ve Chin'e [27] göre evin içerisinde bulunduğu çevre halkının eğitim seviyesinin ve sosyal statüsünün evin değerini genellikle artırdığı belirtilmektedir. Owusu-Manu ve diğerleri [30] çalışmalarında, benzer yapısal özelliklere sahip iki ev arasında mahallenin maddi gelir ortalamasına göre ev fiyatlarında önemli bir artış olduğu tespit edilmiştir.

Bu çalışma, literatürdeki konut fiyat tahmin araştırmalarından birkaç önemli noktada farklılık göstermektedir. İlk olarak, daha önceki araştırmalar genellikle tek bir model veya sınırlı sayıda model üzerinde yoğunlaşırken, bu çalışmada konut fiyat tahmini için 10 farklı denetimli regresyon algoritmasının performans karşılaştırması yapılmıştır. Ayrıca, mevcut literatürün aksine, hiper parametre optimizasyonu için Grid Search, Random Search ve Optuna gibi üç farklı yöntem sistematik olarak değerlendirilmiş ve bu yöntemlerin model performansları üzerindeki etkileri incelenmiştir. Bu yaklaşım, sadece en iyi tahmin modelini belirlemekle kalmamış, aynı zamanda hiper parametre ayarlama stratejilerinin önemini de vurgulamıştır. Literatürde genellikle hiper parametre optimizasyonunun sınırlı şekilde ele alındığı görülmektedir. Bu kapsamlı değerlendirme, konut fiyat tahmin modellerinin daha doğru ve güvenilir sonuçlar üretebilmesi için önemli bir

rehber niteliğindedir. Buna ek olarak, araştırma, otomatik tahmin sistemlerinin konut piyasasındaki dalgalanmaları öngörme potansiyelini inceleyerek, ekonomik istikrarın sürdürülebilirliğine nasıl katkı sağlayabileceğini değerlendiren bir perspektif sunmaktadır. Bu yaklaşım, konut fiyat tahmini çalışmalarında ekonomik sürdürülebilirlik boyutunun nadiren ele alınan bir konu olması nedeniyle literatürde önemli bir boşluğu doldurmayı hedeflemektedir. Bu özgün yaklaşım, araştırmayı literatürdeki diğer çalışmalardan ayırmakta ve gelecekteki çalışmalara değerli bir temel oluşturmayı amaçlamaktadır.

3. MAKİNE ÖĞRENMESİ (MACHINE LEARNING)

Makine öğrenmesi, bilgisayar sistemlerinin veri setlerinden öğrenme yeteneği kazandığı bir yapay zekâ alt alanı olarak kabul edilmektedir. Bu alandaki temel kavramları anlamak adına 1959 yılında Arthur Samuel tarafından önemli bir tanım yapılmıştır. Samuel, makine öğrenmesini, "Bir programın bir görevi iyi yapmasını sağlayacak şekilde deneyimden öğrenmesi" olarak tanımlamıştır [31]. Bu, makine öğrenmesinin özünü, deneyimden öğrenme ve adaptasyon süreçleriyle ilişkilendiren önemli bir tanımdır. Makine öğrenmesinin temel amacını belirleyen bir başka tanım Tom Mitchell tarafından 1997 yılında sunulmuştur. Mitchell, makine öğrenmesini, "Bir bilgisayar programının belirli bir görevde performansını ölçmek amacıyla bir görevi öğrenmesi" olarak ifade etmiştir [32]. Bu, makine öğrenmesinin odak noktasının performans artışı ve öğrenme sürecinin ölçülmesi olduğunu vurgulayan bir tanımdır.

Makine öğrenmesi, genel olarak denetimli öğrenme, denetimsiz öğrenme ve pekiştirmeli öğrenme olmak üzere üç ana kategori altında incelenmektedir. Denetimli öğrenme, makine öğrenmesinin bir alt dalı olup, algoritmaya giriş verileri ile çıkış etiketleri arasındaki ilişkiyi öğrenme yeteneği sağlayan bir öğrenme türüdür. Bu süreçte, bir model, eğitim verileri üzerinden öğrenmekte ve daha sonra bu öğrenilen bilgileri yeni, önceden belirlenmemiş verilere uygulayarak çıkışları tahmin etmektedir. Bu yöntem, genellikle bir öğrenme problemi çerçevesinde kullanılmaktadır. Önceden etiketlenmiş bir veri seti kullanılarak model eğitilmektedir. Bu veri setinde her girişe karşılık gelen doğru çıkış etiketi bulunmaktadır. Model, bu giriş ve çıkışları kullanarak veri setindeki desenleri anlamaya çalışmaktadır. Denetimli öğrenme, sınıflandırma ve regresyon olmak üzere iki ana kategoriye ayrılmaktadır. Sınıflandırma, bir veri noktasını belirli bir kategoriye atama görevini üstlenirken, regresyon bir veri noktasının bir değeri tahmin etme görevini üstlenir. Regresyon analizleri nicel verilerle çalıştığından, özellikle sayısal değerlerle ilgilenen durumları kapsamaktadır [33]. Araştırmada da ele alınan durum olarak ev fiyatlarını belirleme süreci düşünüldüğünde ev fiyatları, bir dizi faktör tarafından etkilenebilecek karmaşık bir konsepttir. Bir regresyon modeli, bir evin metrekare büyüklüğü, oda sayısı, bulunduğu semt gibi özelliklere dayalı olarak bir evin fiyatını tahmin edebilir.

Eğitim aşamasında model bir veri seti kullanılır ve her ev için bilinen gerçek fiyatlar, modelin öğrenmesi için kullanılan çıkış etiketleri olarak kabul edilir [34]. Bu veri setindeki evlere ait özellikler (bağımsız değişkenler) ile fiyatlar (bağımlı değişken) arasındaki ilişki, model tarafından öğrenilir. Eğitilen model, daha sonra yeni bir evin özelliklerini kullanarak tahmin yapabilmektedir. Örneğin, bir regresyon modeli, 200 metrekare büyüklüğünde, 3 odalı ve şehir merkezine yakın bir semtte bulunan bir evin fiyatını tahmin edebilir. Model, bu özelliklere dayanarak benzer özelliklere sahip diğer evlerin fiyatlarından yola çıkarak tahminini gerçekleştirmektedir. Model, eğitim verileri üzerinden öğrenirken, belirli bir kayıp fonksiyonu kullanılır. Bu fonksiyon, modelin tahmin ettiği çıkış ile gerçek etiket arasındaki farkı ölçer. Eğitim süreci, bu kaybı minimize etmeye çalışarak modelin doğruluğunu artırır [35].

Denetimli öğrenmenin geniş bir uygulama yelpazesi bulunmaktadır. Görüntü tanıma, doğal dil işleme, tıbbi teşhis, finansal tahminler gibi birçok alanda başarıyla kullanılmaktadır. Bu yöntem, veri bilimi ve istatistiksel analizde yaygın olarak benimsenmiş bir araçtır [36]. Bu araştırma, konut fiyatlarını tahminleme bağlamında gerçekleştirilmiş olup, LR, RR, LAR, SVR, DTR, RFR, GBR, KNN, MLP ve ENR olmak üzere 10 denetimli regresyon algoritması kullanılmıştır. Bu algoritmalar, konut fiyatlarını belirleyen faktörleri ve bu faktörler arasındaki ilişkileri modelleme yetenekleri nedeniyle seçilmiştir. Denetimli regresyon algoritmalarının konut fiyatı tahminleme görevi için seçilmesiyle her bir algoritmanın konut piyasasındaki değişkenlikleri yakalama yeteneklerinin değerlendirilmesi amaçlanmaktadır.

3.1. Doğrusal Regresyon (Linear Regression)

LR, bir istatistiksel modelleme tekniğidir ve bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkiyi ifade etmek için kullanılmaktadır. LR, veri setindeki bu ilişkiyi temsil eden bir doğrusal fonksiyonun bulunmaya

çalışıldığı bir istatistiksel tekniktir [37]. Bu denklem Eşitlik 1'de verilmiştir.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_n X_n + \varepsilon \qquad (1)$$

Eşitlik 1'deki denklemin içerisinde Y bağımlı değişkeni, $X_1, X_2, \ldots, X_n$ bağımsız değişkenleri, β_0, $\beta_1, \ldots, \beta_n$ regresyon katsayıları ve ε hata terimi yer almaktadır. LR'nin amacı, regresyon katsayılarını gözlemlerle uyumlu hale getirilerek tahmin edilen değerlerle gerçek değerler arasındaki hatayı minimize etmektir. Bu optimizasyon genellikle en küçük kareler yöntemiyle gerçekleştirilmektedir. Algoritmanın işleyişi, veri setindeki gözlemler arasındaki doğrusal ilişkiyi temsil eden en iyi uyan doğruyu bulmayı içermektedir [38]. Bu, regresyon katsayılarının tahmin edilmesi ve modelin eğitilmesi sürecini içermektedir. Eğitim sonrasında, elde edilen model yeni bağımsız değişken değerleriyle kullanılarak bağımlı değişkenin tahminini yapabilir.

3.2. Ridge Regresyonu (Ridge Regression)

RR, doğrusal regresyonun bir genişlemesi olarak kabul edilmekte ve regresyon katsayılarının tahmin edilmesi sürecinde bir düzenleme (regülarizasyon) eklenmektedir. Temelde, en küçük kareler yöntemine benzer bir şekilde çalışılmakta, ancak regresyon katsayılarının aşırı uyum (overfitting) riskini azaltmak ve modelin genelleme yeteneğini artırmak için ek bir terim eklenmektedir. RR, L2 normu kullanılarak düzenleme uygular ve regresyon katsayılarının karelerinin toplamını sınırlayan bir terim eklenilerek gerçekleştirilir [39].

3.3. Lasso Regresyonu (Lasso Regression)

LAR, doğrusal regresyonun bir türevidir ve temel amacı regresyon katsayılarını tahmin ederken bir düzenleme yöntemi uygulamaktır. LAR, en küçük kareler yöntemini kullanır, ancak aynı zamanda regresyon katsayılarının mutlak değerlerini kontrol altında tutarak değişken seçimini gerçekleştirir. Lasso'nun belirgin özelliği, regresyon katsayılarını sıfıra yaklaştırma eğiliminde olmasıdır. Bu özellik, gereksiz veya düşük etkili değişkenlerin modelden çıkartılmasına olanak tanır. Lasso, bu seçici özelliği sayesinde, regresyon modelinin daha basit ve genelleştirilebilir olmasını sağlar. Algoritmanın çalışma prensibi, hedef değişken ile bağımsız değişkenler arasındaki ilişkiyi ifade eden bir denklem üzerinde düzenlemeli bir terim ekleyerek gerçekleşir. Bu düzenlemeli terim, regresyon katsayılarını kontrol altına alır ve aşırı uymayı önler. Lasso, düzenleme parametresi olarak adlandırılan bir katsayı kullanır, bu parametre arttıkça regresyon katsayıları sıfıra daha fazla yaklaşır ve değişken seçimi daha etkili hale gelir [40].

3.4. Destek Vektör Regresyonu (Support Vector Regression)

SVR, doğrusal ve doğrusal olmayan regresyon görevlerini gerçekleştirmek için kullanılan bir öğrenme algoritmasıdır. SVR, destek vektör makinelerinin bir regresyon uygulamasıdır ve özellikle aykırı değerlere dayanıklı bir regresyon modeli sağlar. SVR'nin temel amacı, veri noktalarının çoğunun bir hiperdüzlem (hyperplane) tarafından belirlenen bir bölge içinde bulunmasını sağlamaktır. Bu hiperdüzlem, regresyon modelinin doğrusal veya doğrusal olmayan ilişkileri ifade etmesi amaçlanan bir düzlem veya uzaydır. Algoritma, destek vektörler olarak adlandırılan ve regresyon modelini tanımlayan kritik veri noktalarını kullanır. Bu destek vektörler, hiperdüzlemle en iyi uyum sağlayacak şekilde seçilir. SVR, bu destek vektörleri arasındaki mesafeyi (margin) maksimize etmeye çalışırken, aynı zamanda aykırı değerlere karşı dirençli olacak şekilde tasarlanmıştır [41].

SVR'nin çalışma prensibi, bir çekirdek fonksiyonu kullanarak girdi verilerini yüksek boyutlu uzaya taşımasıdır. Bu sayede, doğrusal olmayan ilişkileri ele alabilir ve daha karmaşık veri yapılarını modelleyebilir. Çekirdek fonksiyonları, veri noktalarının orijinal uzayda lineer olmayan ilişkilerini ifade etmektedir [42].

3.5. Karar Ağacı Regresyonu (Decision Tree Regression)

DTR, veri kümesini kullanarak bir regresyon modeli oluşturan bir öğrenme algoritmasıdır. Bu algoritma, veriyi bölme ve sınıflandırma işlemlerini gerçekleştiren bir ağaç yapısı kullanır. Algoritmanın çalışma prensibi, veri kümesini özelliklere göre bölme ve bu bölmelerde hedef değişkenin ortalamasını tahmin etme şeklindedir. DTR, bu bölme işlemlerini gerçekleştirirken, her bir bölme noktasının belirli bir özellik değeri ve eşik değeri ile belirlendiği bir yapı oluşturur [43]. Algoritmanın temel amacı, veri kümesini en iyi şekilde açıklamak ve hedef değişkenin değerini doğru bir şekilde tahmin etmektir. Eşitlik 2'de DTR'nin formülüne yer verilmiştir.

$$f(x) = \sum_{m=1}^{M} c_m \cdot I(x \in R_m) \quad (2)$$

Bu formül incelendiğinde $f(x)$, tahmin edilen hedef değişkenin değerini; M, ağaçtaki terminal düğüm sayısını; R_m, m numaralı terminal düğümdeki bölgeyi; c_m, m numaralı terminal düğümdeki tahmin edilen değeri ifade etmektedir. $I(x \in R_m)$, x girdi verisi R_m bölgesinde ise 1, değilse 0 değerini almaktadır.

3.6. Rastgele Orman Regresyonu (Random Forest Regression)

RFR, bir makine öğrenmesi algoritması olarak geniş bir kullanım alanına sahiptir. Bu algoritma, bir dizi karar ağacının bir araya getirilmesiyle oluşturulan bir topluluk öğrenmesi (ensemble) modelidir. Regresyon problemlerinde her bir karar ağacı belirli bir hedef değişkenin tahminini gerçekleştirir. RFR'nin temel işleyişini anlamak için, her bir ağacın oluşturulma sürecine odaklanmak gerekmektedir. İlk olarak, eğitim veri setinden rastgele örneklemler alınır. Bu örneklemler, her bir ağacın eğitiminde kullanılacak alt veri setlerini oluşturur. Daha sonra, her bir ağaç, bu alt veri setleri üzerinde bağımsız olarak eğitilir. Eğitim sırasında, her bir karar ağacı belirli bir özellik alt kümesi üzerinde düğüm bölünmeleri yapar. Bu bölünmeler, her düğümde en iyi bölünmeyi seçmek için belirli bir ölçü kullanılarak gerçekleştirilir. Bu sayede her ağaç, veri setindeki desenleri farklı yollarla öğrenir. Son olarak, her bir ağacın tahminleri bir araya getirilir ve genel tahmin değeri elde edilir. Bu birleştirme süreci, regresyon problemlerinde genellikle ağaç tahminlerinin aritmetik ortalamasını içerir. Bu yöntem, her bir ağacın bağımsız olarak öğrenmesini ve genelleme yeteneklerini artırarak modelin performansını artırmaktadır [44].

3.7. Gradyan Artırma Regresyonu (Gradient Boosting Regression)

GBR, bir hata fonksiyonunu minimize etmek amacıyla artan gradyan adımları kullanarak bir modelin eğitildiği bir regresyon tekniğidir. Başlangıçta belirlenen bir model üzerinden gerçek ve tahmin edilen değerler arasındaki hataların karesinin toplamını içeren bir hata fonksiyonu hesaplanmaktadır. Bu hata fonksiyonunun gradyanı alınarak, parametrelerin değiştirilme yönü belirlenmekte ve parametreler, gradyanın tersine doğru küçük adımlarla güncellenmektedir. Yeni parametrelerle elde edilen modelin performansı değerlendirilmekte ve bu süreç, hata fonksiyonunun minimize edildiği bir noktaya ulaşana kadar tekrarlanmaktadır. Matematiksel formülde, her bir parametrenin, öğrenme oranı ile hatanın türevinden çıkartılmasıyla güncellendiği gradyan iniş (gradient descend) adımı ifade edilmektedir. Bu iteratif süreç, modelin eğitim verilerine uyum sağlaması ve optimize olması için kullanılmaktadır [45]. GBR, Eşitlik 3'te yer alan formülle temsil edilmektedir.

$$\theta_j := \theta_j - \alpha \frac{\alpha}{\alpha \theta_j} J(\theta_0, \theta_1, \dots, \theta_n) \quad (3)$$

Formülde θ_j, j. parametreyi; α, öğrenme oranını; $J(\theta_0, \theta_1, \dots, \theta_n)$, hata fonksiyonunu temsil etmektedir. Bu formül, her bir parametrenin, öğrenme oranı ile hatanın türevinden çıkartılmasıyla güncellendiği gradyan iniş adımını temsil eder. Bu adımlar, hata fonksiyonunu minimize edecek parametre değerlerini bulmak üzere tekrarlanır.

3.8. K-En Yakın Komşu Regresyonu (K-Nearest Neighbors Regression)

KNN Regresyonu, örnekler arasındaki benzerlik temelinde çalışarak bir tahmin yapma yöntemini benimsemektedir. İlk aşamada, bir veri setindeki her bir örnek, sahip olduğu özelliklere göre uzayda bir nokta olarak temsil edilir. Tahmin yapılacak yeni bir örnek geldiğinde, bu örneğin uzaydaki konumu, mevcut veri setindeki diğer örneklerle karşılaştırılır. Özellik uzayındaki benzer örnekler belirlenerek bu örneklerin çıkış değerleri tahminde kullanılmaktadır. Bu belirleme işlemi Öklidyen uzaklık metriği kullanılarak gerçekleştirilir. Yani, iki örnek arasındaki uzaklık, özelliklerine göre hesaplanır ve bu uzaklıklar kullanılarak en yakın k komşu belirlenir [46]. Matematiksel formülle ifade edildiğinde, N adet örneğin bulunduğu veri setinde, i -inci örneğin çıkış değeri y_i ve j-inci örneğin çıkış değeri y_j olmak üzere, örneğin tahmini şu şekilde hesaplanır (Eşitlik 4):

$$\hat{Y} = \frac{1}{k} \sum_{i \in N_k} y_i \quad (4)$$

Formülde, $\hat{Y}$ yeni örneğin tahmin edilen çıkış değerini, k ise belirlenen komşu sayısını temsil eder. N_k, yeni örneğe en yakın k komşuyu ifade eden bir kümedir.

3.9. Çok Katmanlı Algılayıcı Regresyonu (Multilayer Perceptron Regression)

MLP Regresyonu, bir yapay sinir ağı modelidir ve regresyon problemlerini çözmek için kullanılmaktadır. Bu algoritma, giriş katmanı, bir veya daha fazla gizli katman ve bir çıkış katmanından oluşan bir yapıya sahiptir. Giriş katmanında bulunan nöronlar, özelliklerle ilişkilendirilmiş veri setinin her bir öğesini temsil

eder. Gizli katmanlardaki nöronlar, öğrenilecek karmaşık ilişkileri modellemek için kullanılır. Çıkış katmanındaki nöronlar ise regresyon sonuçlarını üretir. MLP Regresyonunun, geniş bir uygulama yelpazesi bulunmaktadır [47]. Özellikle, ev fiyat tahmininden finansal analize kadar birçok alanda başarıyla kullanılmaktadır. Eğitim süreci boyunca ağırlıkların ve yanlılıkların (bias) optimize edilmesi, modelin veri setine daha iyi uymasını sağlamakta ve doğru tahminler yapmasına olanak tanımaktadır.

3.10. ElasticNet Regresyonu (ElasticNet Regression)

ENR, bir regresyon yöntemi olup, hem L1 (Lasso) hem de L2 (Ridge) ceza terimlerini içeren bir lineer regresyon türüdür. Bu algoritma, öznitelik seçimi yapabilme özelliğini sağlayan Lasso regresyonunun avantajları ile çoklu korelasyonlu özniteliklerle başa çıkabilme yeteneğini temin eden Ridge regresyonunun avantajlarını birleştirmektedir. ElasticNet, ağırlıkları güncellemek için hatanın gradyanı ile iki ceza terimini de kullanır. ENR, belirli bir hiper parametre olan alpha tarafından kontrol edilen bir karışım oranı kullanarak L1 ve L2 terimlerini birleştirmektedir [48]. Eğitim süreci, modelin genelleştirilmiş bir performans elde etmesi amacıyla hatanın minimize edilmesini amaçlamaktadır. ElasticNet regresyonunun matematiksel formülü, belirli bir hiper parametre olan alpha, cezalandırma terimleri ve model parametrelerini içerir. Eğitim sırasında bu parametreler, veri setine uygun bir şekilde güncellenir. Bu sayede, ElasticNet, veri setlerindeki karmaşıklıkları ele alarak etkili bir regresyon modeli oluşturabilir.

4. YÖNTEM (METHOD)

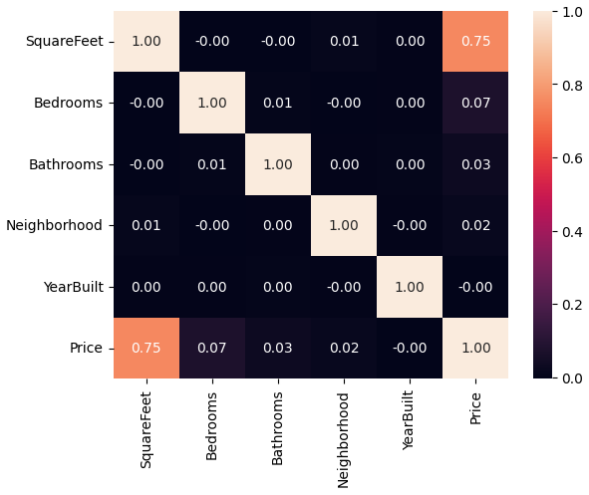
Bu bölümde, araştırmada kullanılan veri setinin özelliklerine, algoritmaların karşılaştırılmasında kullanılan performans ölçütlerine ve veri hazırlama sürecine ilişkin bilgilere yer verilmiştir.

4.1. Veri Seti (Dataset)

Araştırmada kullanılan veri seti "Housing Prices Dataset (Konut Fiyatları Veri Seti)" olarak adlandırılmaktadır ve geniş bir veri kümesini içermektedir. Bu veri seti, ev fiyatlarını tahminlemek, konut piyasasındaki çeşitli faktörleri anlamak ve bu faktörlerin ev fiyatları üzerindeki etkilerini değerlendirmek amacıyla kullanılan popüler bir açık kaynaklı veri setidir. Kaggle platformundan da erişilebilen bu veri seti, çeşitli makine öğrenimi modellerini eğitmek ve ev fiyatlarının belirlenmesinde etkili olan faktörleri değerlendirmek için bir kaynak olarak kullanılmaktadır. Veri seti 49.841 kayıt ve evlerin özelliklerini belirten altı öznitelik içermektedir. Veri setindeki özniteliklere ve açıklamalarına aşağıda yer verilmiştir:

- SquareFeet: Mülkün toplam alanı (m^2).
- Bedrooms: Mülkteki yatak odalarının sayısı.
- Bathrooms: Mülkteki banyoların sayısı.
- Neighborhood: Mülkün bulunduğu mahalle.
- YearBuilt: Mülkün inşa edildiği yıl.
- Price: Mülkün fiyatı.

Şekil 1'de veri setine ait ısı haritası verilmiştir. Isı haritası incelendiğinde, mahalle (Neighborhood) ve inşa yılı (YearBuilt) öznitelikleri arasında belirgin bir ilişki bulunmamaktadır. Ancak, metrekare (SquareFeet) özniteliğinin, hedef özniteliği olan fiyat (price) üzerinde güçlü bir etkisi olduğu gözlemlenmektedir. Yapılan ısı haritası incelemesi, ev fiyatlarını belirlemede SquareFeet özniteliğinin önemli bir faktör olduğunu ortaya koymaktadır.

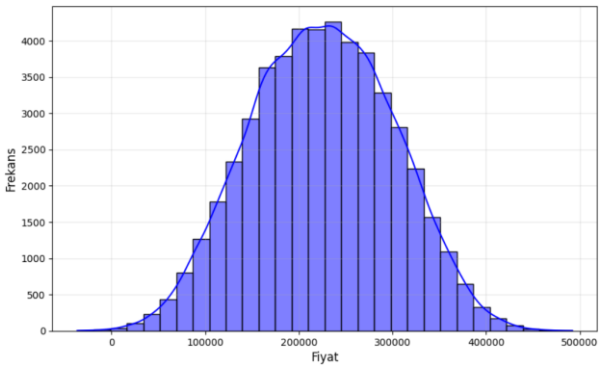


Şekil 1. Isı haritası (Heatmap)

4.2. Verilerin Hazırlanması (Data Preparation)

Araştırma sürecinde, ev fiyat tahminleme modellerini oluşturmaya başlamadan önce veri setinin ön işleme aşamasında bazı önemli işlemler gerçekleştirilmiştir. İlk olarak, ev fiyatları negatif değerlere sahip olan kayıtlar, bu durumun model performansını olumsuz etkilememesi amacıyla veri setinden çıkarılmıştır. Bu olgu, veri setindeki hatalı veya yanlış etiketlenmiş verilerin temizlenmesini sağlamaktadır. Ardından, veri setinde tekrarlanan kayıtların silinmesi işlemi gerçekleştirilmiştir. Bu adım, veri setindeki tutarsızlık ve gereksiz karmaşıklığı azaltarak modelin genel performansını artırmayı hedeflemektedir. Coğrafi konumunun ev

fiyatlarındaki etkisini doğru bir şekilde yakalamak amacıyla kategorik bir değişken olan mahalle (Neighborhood) özniteliğinin kodlanması gerçekleştirilmiştir. Evin bulunduğu mahalle, ev fiyatlarını etkileyen önemli bir faktördür ve bu bilgiyi modelin daha etkili bir şekilde öğrenmesi için sayısallaştırmak önemlidir. Son olarak, modelin daha güvenilir ve genelleme yeteneği yüksek tahminler yapmasına yardımcı olmak amacıyla ev fiyatlarının yuvarlanması işlemi gerçekleştirilmiştir. Bu adım, fiyatları daha anlamlı ve modelin anlayabileceği bir formda temsil etmektedir. Bu ön işleme adımları, modelin eğitim verilerini daha sağlam ve güvenilir hale getirerek, ev fiyatlarını daha doğru bir şekilde tahmin etmesine olanak sağlamaktadır.



Şekil 2. Fiyat dağılımı (Price distribution)

Veri setindeki fiyat (Price) dağılımı, konut fiyatlarının istatistiksel özelliklerini anlamak amacıyla incelenmiştir. Bu bağlamda, fiyatların dağılımını görselleştirmek için histogram ve Çekirdek Yoğunluğu Tahmini (Kernel Density Estimate - KDE) kullanılmıştır. Fiyat dağılımı, Şekil 2'deki grafikte yer aldığı gibi normal dağılım özelliklerini göstermektedir. Grafik, fiyatların çoğunluğunun belirli bir aralıkta yoğunlaştığını ortaya koymaktadır. Normal dağılımı simüle eden eğri, verilerin simetrik ve çan şeklinde dağıldığını ve bu dağılımın ortalama etrafında yoğunlaştığını göstermektedir. Bu dağılımın normal dağılıma yakın olması, konut fiyatlarının çoğunlukla merkezî bir değere yakın olduğunu, ancak bazı istisnaların (örneğin, lüks konutlar veya kriz sonrası fiyatlar gibi) bu dağılımın uçlarında yer aldığını göstermektedir. Normal dağılımın bu şekilde sergilenmesi, istatistiksel analizler için verinin uygun olduğunu ve ileri düzey tahmin modelleri geliştirmek için uygun bir temel oluşturduğunu ifade etmektedir.

4.3. Performans Metrikleri (Performance Metrics)

Regresyon modellerinin etkin bir şekilde değerlendirilmesi ve oluşturulan modeller arasında performans karşılaştırması yapılmasını sağlamak amacıyla çeşitli performans değerlendirme metriklerinin kullanılmasını gerektirmektedir. Bu bağlamda, araştırmada tercih edilen metrikler arasında Ortalama Mutlak Hata (Mean Absolute Error – MAE), Ortalama Kare Hata (Mean Squared Error – MSE), Kök Ortalama Kare Hata (Root Mean Squared Error – RMSE) ve belirleme katsayısı (coefficient of determination - R^2) skoru bulunmaktadır.

MAE, modelin tahminlerinin gerçek değerlerden ortalama sapmasını ölçen bir metrik olarak kabul edilmektedir. Düşük bir MAE değeri, modelin daha keskin ve doğru tahminler gerçekleştirdiğinin göstergesi olarak kabul edilir. Bu metrik, her bir tahmin hatasının mutlak değeri alınarak bunların toplandığı ve ardından ortalama değerinin hesaplandığı bir formülle ifade edilir. Bu sayede, modelin ne kadar yanıltıcı olmadığı ve tahminlerin gerçek değerlere ne kadar yakın olduğu değerlendirilmiş olur [49]. MAE'nin formülü Eşitlik 5'te gösterilmektedir.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \check{y}_i| \qquad (5)$$

MSE, hataların karelerinin ortalamasını temsil eder ve bu nedenle büyük hataların, küçük hatalara kıyasla daha fazla ağırlığa sahip olduğu durumları vurgular. Bu metrik, her bir tahmin hatasının karesinin alınması, bunların toplanması ve ardından ortalama değerinin hesaplanmasıyla elde edilmektedir. Bu hesaplama ile modelin tahminlerinin gerçek değerlere ne kadar yakın veya uzak olduğunu değerlendirmek mümkün olmaktadır. MSE, regresyon modelinin performansını ölçerken hataların büyüklüğünün önemli olduğu durumlar için kullanışlı bir değerlendirme kriteridir [50]. MSE'nin formülü Eşitlik 6'da verilmiştir.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \check{y}_i)^2 \qquad (6)$$

RMSE, regresyon modelinin tahminlerinin gerçek değerlere ne kadar yakın veya uzak olduğunu değerlendirmek için kullanılan bir performans metriğidir. RMSE, MSE'nin karekökü olarak geçmektedir ve bu sayede hataların orijinal biriminde ifade edilmesine olanak tanır. Büyük hataların model performansını daha fazla etkilediği durumları vurgulamaktadır. Her bir tahmin hatasının karesinin alınmasıyla bunların toplanması sağlanır. Ardından ortalama değeri hesaplanarak karekök alınır. RMSE, regresyon modellerinin hata düzeyini daha açıklayıcı bir şekilde ifade etmek için

yaygın olarak tercih edilen bir metriktir [51]. RMSE'nin formülü Eşitlik 7'de yer almaktadır.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \check{y}_i)^2} \qquad (7)$$

R^2 skoru, bağımlı değişkenin varyansının bağımsız değişkenler tarafından ne kadarının açıklandığını ölçen bir metriktir. Bir modelin veriyi ne kadar iyi açıkladığını belirlemek için kullanılmaktadır. R^2 skoru, 0 ile 1 arasında bir değer alır, 1'e ne kadar yakınsa, modelin veriyi o kadar iyi açıkladığı anlamına gelir. R^2 skoru, regresyon modelinin toplam varyansın yüzde kaçını açıkladığını ifade eder. Modelin açıklama gücüne dair bir ölçüdür ve ne kadar yüksekse, modelin başarısı o kadar yüksek kabul edilir [52]. Formülü Eşitlik 8'de yer almaktadır.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \check{y}_i)^2}{\sum_{i=1}^{n}(y_i - \grave{y}_i)^2} \qquad (8)$$

5. DENEYSEL ÇALIŞMA VE BULGULAR (EXPERIMENTAL STUDY AND FINDINGS)

Bu araştırmada, konut fiyatlarını tahminleme amacıyla 10 farklı denetimli regresyon algoritması kullanılmıştır. Modeller oluşturulurken, veri seti %80 eğitim ve %20 test olmak üzere ikiye ayrılmıştır. Kullanılan bütün algoritmalarda rastgele durum (random state) 42 olarak ayarlanmıştır. Modeller için en iyi hiper parametre ayarları, Grid Search, Random Search ve Optuna yöntemleri kullanılarak belirlenmiştir ve bu belirlenen hiper parametrelere göre performans karşılaştırmaları gerçekleştirilmiştir. Grid Search, belirtilen hiper parametre aralıkları içinden farklı kombinasyonlar deneyerek en iyi performansı veren hiper parametreleri seçmektedir. Grid Search kullanarak bir RF modelinin hiper parametrelerinin ayarlanması durumunda, Grid Search belirlenen hiper parametre aralıkları içinden her bir kombinasyonu deneyerek birçok farklı RF modeli oluşturmaktadır. Daha sonra bu modelleri belirli bir metrik kullanarak değerlendirmekte ve en iyi performansı veren hiper parametre kombinasyonunu seçmektedir [53]. Random Search, hiper parametrelerin rastgele seçilen kombinasyonlarını deneyerek modelin performansını optimize etmeyi amaçlamaktadır. Bu yöntem, Grid Search'in aksine, belirli bir hiper parametre aralığında rastgele örneklemeler yaparak hiper parametre uzayını daha geniş bir şekilde keşfetmekte ve daha hızlı sonuçlar elde edilmesine olanak tanımaktadır. Optuna ise hiper parametre optimizasyonunda daha gelişmiş bir yöntem sunmaktadır. Optuna, bir optimizasyon algoritması kullanarak hiper parametre aralığından en iyi sonuçları aramak üzere deneyler yapmaktadır. Bu yöntemde hiper parametrelerin etkilerinin daha hızlı ve etkili bir şekilde değerlendirilmesi için Bayes optimizasyonu gibi teknikler kullanılmaktadır. Araştırmada, modellerin performanslarını daha güvenilir bir şekilde değerlendirmek amacıyla beş katmanlı çapraz doğrulama yöntemi uygulanmıştır. Modellere ait en iyi sonuçlar Optuna yöntemi ile belirlenen hiper parametreler ile elde edilmiştir. Buna göre algoritmaların hiper parametre değerleri Tablo 1'de verilmiştir.

Araştırmada kullanılan her bir algoritmanın performansını değerlendirmek, regresyon modelinin ne kadar iyi tahmin yaptığını ölçmek amacıyla MAE, MSE, RMSE ve R^2 metrikleri kullanılmıştır. Tablo 2'de Random Search yöntemi ile, Tablo 3'te Grid Search yöntemi ile ve Tablo 4'te ise Optuna yöntemi ile ayarlanmış hiper parametrelerle elde edilmiş modellerin tahminleme yeteneklerine dair performans ölçümleri sunulmuştur.

Tablo 1. Algoritmaların hiper parametre ayarları (Hyperparameter settings of algorithms)

Algoritma	Parametreler	Değer
LR	fit_intercept, normalize, solver, alpha	True, False, 'lbfgs', 0.0001
RR	alpha, fit_intercept, normalize	1.0, True, False
LAR	alpha, fit_intercept, normalize	1.0, True, False
SVR	C, epsilon, kernel, degree	Belirlenmiş, Belirlenmiş, 'rbf', -
RFR	n_estimators, max_features, min_samples_split, min_samples_leaf	100, "auto", 2, 1
GBR	n_estimators, learning_rate, max_depth, subsample, min_samples_split	100, 0.1, None, 1.0, 2
DTR	max_depth, min_samples_split, min_samples_leaf, criterion	5, 2, 1, 'squared_error'
KNN	n_neighbors, weights, algorithm	5, 'uniform', 'auto'
MLP	hidden_layer_sizes, activation, learning_rate, solver, alpha	(100, 50), 'relu', 'constant', 'adam', 0.0001

ENR	alpha, l1_ratio, fit_intercept, normalize	0.5, 0.5, True, False

Tablo 2. Random Search ile ayarlanmış hiper parametrelerle modellerin performansı (Performance of models with hyperparameters tuned with Random Search)

Algoritma	MAE		MSE		RMSE		R^2	
	Eğitim	Test	Eğitim	Test	Eğitim	Test	Eğitim	Test
LR	4114.88	4078.94	2577.32	2522.21	5102.68	5068.68	0.5519	0.5564
RR	4076.20	4112.98	2542.17	2519.52	5056.96	5067.03	0.5572	0.5574
LAR	4114.88	4112.88	2577.32	2519.61	5102.68	5067.05	0.5519	0.5575
SVR	6700.79	6634.06	5922.49	5833.70	7909.95	7860.58	0.0101	0.0101
DTR	1972.01	6316.91	8314.29	5428.43	2974.52	7466.34	0.9686	0.0850
RFR	1611.58	4368.61	4086.93	2870.13	2026.32	5453.65	0.9031	0.5034
GBR	4037.82	3787.24	2518.91	2118.19	5120.94	4691.60	0.5680	0.6439
KNN	3662.91	4470.33	2037.68	3041.92	4511.15	5615.21	0.6697	0.4735
MLP	6076.87	6136.66	5350.86	5446.10	7511.67	7611.89	0.0485	0.0392
ENR	5865.44	5799.30	5079.17	4999.94	7246.43	7218.47	0.1541	0.1555

Tablo 3. Grid Search ile ayarlanmış hiper parametrelerle modellerin performansı (Performance of models with hyperparameters tuned with Grid Search)

Algoritma	MAE		MSE		RMSE		R^2	
	Eğitim	Test	Eğitim	Test	Eğitim	Test	Eğitim	Test
LR	4196.88	4144.52	2627.36	2592.31	5252.36	5217.10	0.5404	0.5442
RR	4196.89	4172.85	2627.36	2592.32	5252.36	5217.09	0.5404	0.5442
LAR	4196.88	4172.92	2627.36	2592.32	5252.36	5217.10	0.5404	0.5442
SVR	6500.32	6489.20	6035.33	6008.24	7960.58	7842.70	0.0093	0.0094
DTR	1908.91	6127.82	8051.91	5562.17	2907.66	7642.16	0.9487	0.0792
RFR	1657.90	4450.44	4187.77	2953.38	2096.93	5568.70	0.8846	0.4877
GBR	4169.55	3748.67	2594.03	2097.51	5219.95	4692.95	0.5455	0.6230
KNN	3748.51	4581.38	2097.51	3131.61	4692.95	5723.78	0.6230	0.4597
MLP	6259.20	6275.98	5793.25	5835.11	7799.31	7827.44	0.0470	0.0365
ENR	5985.47	5969.61	5173.94	5142.80	7370.64	7348.43	0.1435	0.1449

Tablo 4. Optuna ile ayarlanmış hiper parametrelerle modellerin performansı (Performance of models with hyperparameters tuned with Optuna)

Algoritma	MAE		MSE		RMSE		R^2	
	Eğitim	Test	Eğitim	Test	Eğitim	Test	Eğitim	Test
LR	3997.03	3947.16	2502.25	2468.87	5002.25	4968.67	0.5689	0.5728
RR	3997.04	3974.14	2502.25	2468.76	5002.25	4968.66	0.5689	0.5728
LAR	3997.03	3974.16	2502.25	2468.76	5002.25	4968.67	0.5689	0.5728
SVR	6190.78	6170.67	5747.93	5722.14	7581.51	7564.48	0.0098	0.0099
DTR	1818.01	5836.02	7668.49	5297.30	2769.20	7278.25	0.9986	0.0834
RFR	1578.95	4237.56	3988.35	2812.74	1997.08	5303.53	0.9312	0.5133
GBR	3971.95	3570.16	2471.46	1997.63	4971.38	4469.48	0.5742	0.6558
KNN	3570.01	4363.22	1997.63	2982.49	4469.48	5451.22	0.6558	0.4839
MLP	5961.14	5977.12	5517.38	5557.25	7427.91	7454.70	0.0495	0.0384
ENR	5700.45	5675.82	4927.56	4897.90	7019.66	6998.50	0.1511	0.1525

Tablo 2, Tablo 3 ve Tablo 4'te modellerin hiper parametre ayarlama yöntemlerine göre hem eğitim hem de test veri setlerinde aldığı skorlar verilmiştir. Modellerin genel performansının değerlendirilmesi açısından test seti daha kritik kabul edilmektedir. Modelin gerçek dünya verileriyle nasıl başa çıkabildiğini değerlendirmek, modelin genelleme yeteneğini daha iyi yansıtmaktadır. Ancak, eğitim setindeki performans değerleri, modelin veriyi ne kadar iyi öğrendiğini anlamak için önemlidir. İdeal durumda, model eğitim ve test setlerinde benzer performans göstermelidir. Eğer model eğitim setinde çok iyi performans sergilerken, test setinde kötü performans gösteriyorsa, bu durum aşırı uyuma işaret edebilir. Bu nedenle, modelin hem

eğitim hem de test setlerinde iyi performans sergilemesi hedeflenir.

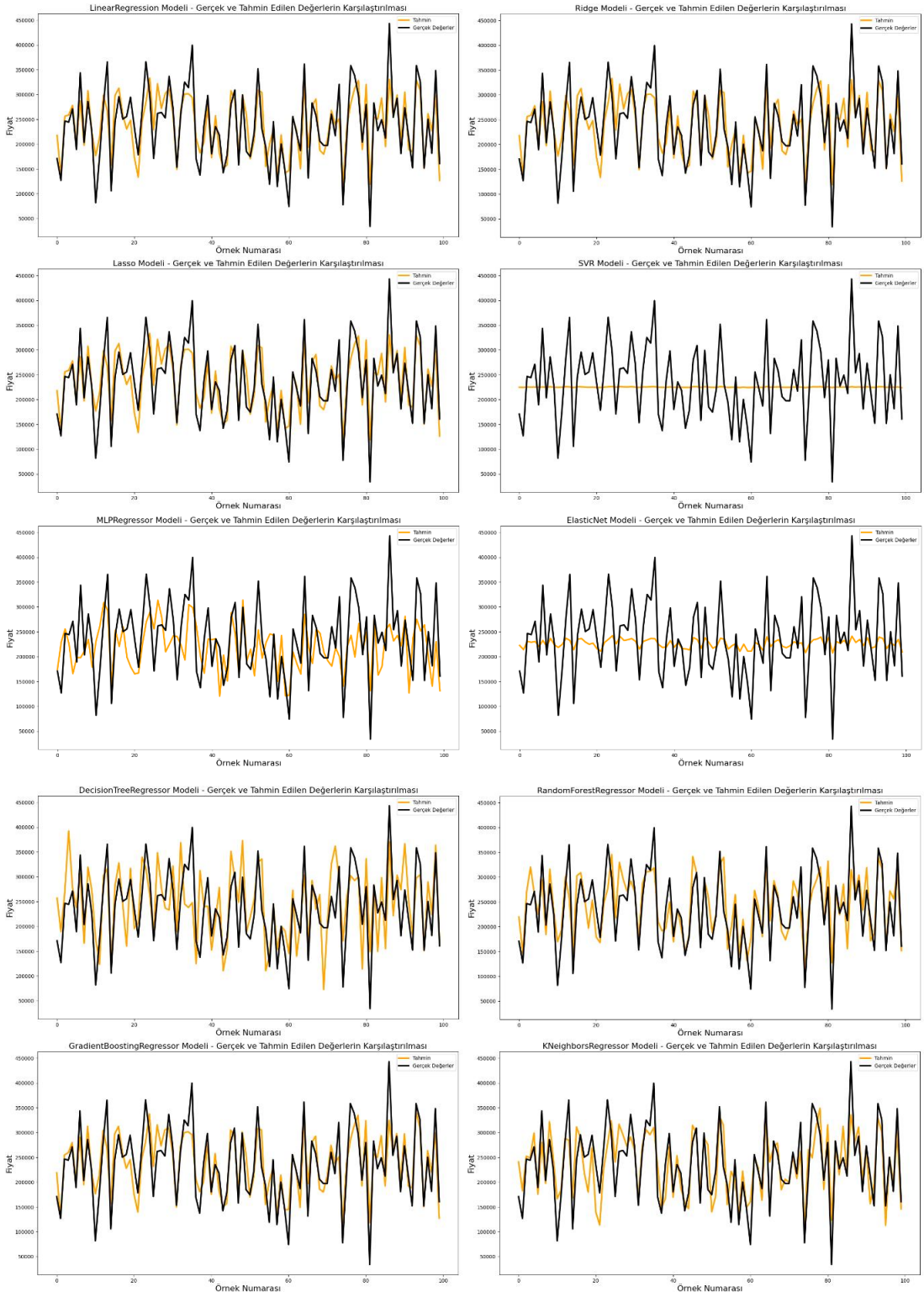
Araştırma bulgularına göre Optuna ile hiper parametreleri ayarlanmış GBR modeli, diğer modeller arasında test veri seti üzerinde alınan en yüksek R^2 değeri ile konut fiyatlarını belirlemede en iyi model olarak belirlenmiştir (RMSE = 4469.48, R^2 = 0.6558). LR, RR ve LAR modelleri eğitim ve test veri setlerinde benzer performans sergilemektedir. Bahsedilen modeller açısından R^2 değerleri yaklaşık olarak %57'dir. Bu değer, hedef değişkenin varyansının %57'sinin açıklandığını göstermektedir. SVR modeli eğitim ve test veri setlerinde düşük performans göstermektedir. R^2 değeri yaklaşık olarak 0'a yakındır, bu da modelin verilere uygun olmadığı şeklinde yorumlanabilir. Buna ek olarak MAE (6170.67) ve RMSE değerlerinin (7564.48) yüksek bulunmasıyla da daha büyük bir tahmin hatasının olduğuna işaret etmektedir. DTR, eğitim setinde olağanüstü performans göstermektedir (RMSE = 2769.20, R^2 = 0.9986). Ancak test setinde performans çok ciddi ölçüde düşmektedir (RMSE = 7278.25, R^2 = 0.0834). Bu durum modelin potansiyel bir aşırı uyum durumuna girdiğini göstermektedir. RFR, eğitim setinde DTR'ye benzer şekilde yüksek bir performans göstermiştir (RMSE = 1997.08, R^2 = 0.9312). Eğitim seti üzerindeki yüksek R^2 değeri, modelin iyi bir genelleme performansı gösterdiğine işaret etmektedir. Ancak, test setindeki performansa bakıldığında DTR kadar olmasa da modelin performansının önemli ölçüde düştüğü görülmektedir (RMSE = 5303.53, R^2 = 0.5133). Bunların yanı sıra, bulgular, Optuna'nın genel olarak en yüksek performansı sağladığını ve diğer yöntemlere kıyasla belirgin bir üstünlük sunduğunu göstermektedir. Optuna'nın hiper parametre optimizasyonundaki hassasiyeti ve etkinliği, algoritmaların performansını artırmada en etkili yaklaşım olarak öne çıkmaktadır. Buna karşılık, Random Search en düşük performansı gösterirken, Grid Search daha iyi sonuçlar elde etmiştir ancak, Optuna'nın sağladığı yüksek verimliliği yakalayamamıştır. Grid Search, sistematik bir tarama ile Random Search'tan üstün performans sergilese de Optuna'nın sunduğu optimizasyon avantajlarını sağlayamamıştır.

Tablo 5. Optuna ile hiper parametreleri ayarlanmış modellerin doğrulama veri setinde performansı
(Performance of Optuna hyper-parameterized models on validation dataset)

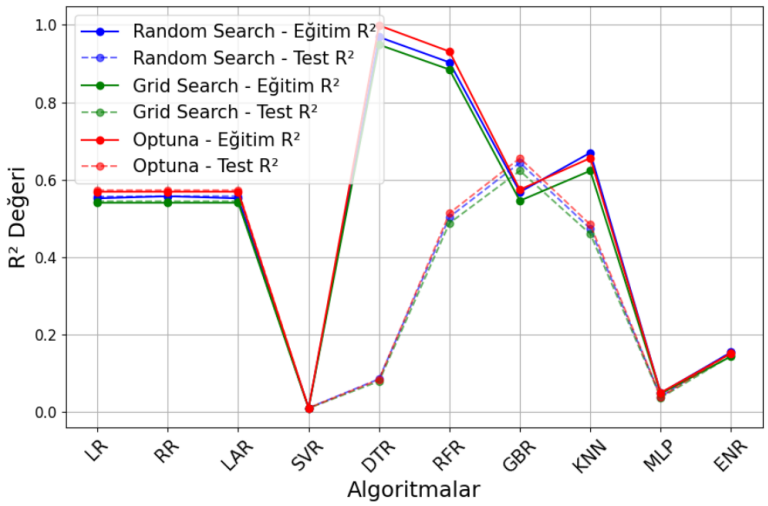
Algoritma	MAE		MSE		RMSE		R^2	
	Doğrulama	Test	Doğrulama	Test	Doğrulama	Test	Doğrulama	Test
LR	3983.21	3965.16	2481.19	2456.35	4981.55	4956.16	0.5774	0.5682
RR	3983.21	3965.08	2481.19	2456.32	4981.63	4956.13	0.5774	0.5682
LAR	3983.21	3965.11	2481.19	2456.34	4981.56	4956.15	0.5774	0.5682
SVR	6211.48	6129.86	5812.46	5631.82	7623.95	7504.55	0.0099	0.0099
DTR	5835.94	5827.88	5340.62	5298.47	7307.95	7279.05	0.0903	0.0685
RFR	4273.26	4208.20	2864.71	2773.79	5352.30	5266.68	0.5120	0.5124
GBR	3991.73	3969.73	2490.29	2465.43	4990.28	4965.31	0.5758	0.5666
KNN	4399.74	4326.71	3029.14	2935.85	5503.76	5418.35	0.4840	0.4839
MLP	6174.98	6037.88	5885.91	5704.85	7671.97	7553.05	0.0025	0.0029
ENR	5714.81	5636.84	4975.00	4820.79	7053.37	6943.19	0.1525	0.1525

Veri seti, modellerin genelleme yeteneğini daha ayrıntılı değerlendirebilmek amacıyla eğitim (%70), doğrulama (%15) ve test (%15) olarak yeniden bölünmüştür. Daha önceki ölçümlerde en yüksek performans değerleri Optuna ile hiper parametre optimizasyonu gerçekleştirildiğinde elde edildiği için modeller bu ayarlamalarla yeniden test edilmiştir. Doğrulama veri seti üzerindeki modellerin performans sonuçları Tablo 5'te sunulmuştur. Tablo 5'teki değerlendirmeler, Tablo 4'te yer alan test sonuçlarıyla kıyaslandığında bazı modellerin performansında önemli değişimler olduğu gözlemlenmiştir. GBR modeli, doğrulama veri setinde en iyi performansı sergilemiştir (R^2 = 0.5758, RMSE = 4990.28). Test seti sonuçları ile karşılaştırıldığında, doğrulama veri setindeki RMSE değerinin %11.3 oranında daha yüksek olduğu görülmektedir. Ancak, bu durum modelin genelleme başarısını hala koruduğunu ve veri bölme yönteminden kaynaklanan sapmanın kabul edilebilir düzeyde olduğunu göstermektedir. RFR modeli, doğrulama veri setinde R^2 = 0.5120 ve RMSE = 5352.3 skorlarıyla eğitim ve test veri setlerine kıyasla daha düşük performans göstermiştir. Bu düşüş, modelin aşırı uyuma yatkın olabileceğine işaret ederken, RFR'nin genel anlamda güçlü bir tahmin aracı olmaya devam ettiğini göstermektedir. Özellikle test veri setindeki RMSE değeriyle kıyaslandığında, doğrulama veri setindeki RMSE'nin %0.9 oranında daha yüksek olduğu gözlemlenmiştir. Genel olarak, GBR modeli en iyi performansı sergilerken, doğrulama seti

sonuçları Tablo 4'teki test seti sonuçlarına göre bazı modellerde küçük sapmalar olduğunu, ancak genel sıralamanın büyük ölçüde değişmediğini ortaya koymuştur.



Şekil 3. Gerçek fiyat değerleri ile tahmin edilen değerlerin karşılaştırması (Comparison of actual price values and predicted values)



Şekil 4. Modellerin doğruluk karşılaştırması (Accuracy comparison of models)

Şekil 3'te gerçek satış değerleri ile Optuna ile hiper parametreleri ayarlanmış algoritmalar tarafından tahmin edilen değerlerin karşılaştırması bulunmaktadır. Şekil 3 incelendiğinde RFR ve GBR modellerinin, gerçek değerlere en yakın tahminleri ürettiği gözlemlenmiştir. GBR modelinin tahmin performansı, diğer modellere kıyasla daha kararlı ve doğruluğu yüksek bulunmuştur. Buna karşın SVR modelinde tahmin edilen değerlerin gerçek değerlerden önemli ölçüde sapma gösterdiği dikkat çekmektedir.

Hiper parametre optimizasyon yöntemlerinin modeller üzerindeki etkisi ise Şekil 4'te R^2 değerleri üzerinden değerlendirilmiştir. Random Search, Grid Search ve Optuna yöntemleri, modellerin performansını artırmada farklı derecelerde etkili olmuştur. GBR ve RFR hem eğitim hem de test veri setlerinde yüksek R^2 değerleri elde etmiştir. Bu durum, söz konusu modellerin hem eğitim setinde öğrenme kapasitesinin hem de test setinde genelleme yeteneğinin güçlü olduğunu göstermektedir. Hiper parametre optimizasyon yöntemleri arasında Optuna, modellerin performansını en çok artıran yöntem olarak öne çıkmıştır. Bunlara ek olarak SVR modeli, hiper parametre optimizasyon yöntemlerinden bağımsız olarak düşük R^2 değerleri göstermiştir.

6. SONUÇ VE TARTIŞMA (CONCLUSION AND DISCUSSION)

Bu araştırmada, konut fiyat tahminlemesinde en başarılı makine öğrenmesi algoritmasını belirleyebilmek amacıyla 10 farklı denetimli regresyon algoritmasının performans karşılaştırmaları gerçekleştirilmiştir. Eğitim ve test setlerinde elde edilen metrik değerler, modellerin genel performansını anlamak adına detaylı bir değerlendirmeye tabi tutulmuştur. Ayrıca, algoritmaların performansını optimize etmek için Grid Search, Random Search ve Optuna gibi hiper parametre ayarlama yöntemleri kullanılmıştır. Bu yöntemler, her bir modelin hiper parametrelerini en uygun şekilde ayarlayarak en iyi performansı elde etmek için kapsamlı bir şekilde uygulanmıştır. Hiper parametre ayarlama süreçleri, modellerin genel başarılarını etkileyen önemli bir faktör olarak değerlendirilmiştir.

Optuna ile hiper parametreleri ayarlanmış GBR modeli, test veri seti üzerinde elde ettiği yüksek R^2 değeri (0.6558) ve düşük RMSE değeri (4469.48) ile konut fiyatlarını tahminleme konusunda en başarılı model olarak belirlenmiştir. Optuna'nın hiper parametre optimizasyonundaki sağladığı hassasiyet ve etkinlik, diğer yöntemlerle karşılaştırıldığında belirgin bir üstünlük sunmuştur. Grid Search ve Random Search yöntemlerinin sınırlamaları göz önüne alındığında, Optuna'nın daha ileri düzeyde performans iyileştirmeleri sağladığı gözlemlenmiştir. Grid Search, hiper parametre kombinasyonlarını sistematik bir şekilde tarayarak geniş bir arama alanı sunmasına rağmen, belirli bir kombinasyonun optimal olduğunu garantileyememektedir. Random Search ise hiper parametreleri rastgele kombinasyonlarla denediğinden, hiper parametrelerin en iyi kombinasyonlarına ulaşmada sınırlı başarı göstermiştir. Buna karşın, Optuna'nın Bayesiyen optimizasyon yaklaşımı, hiper parametre alanında daha verimli ve hedefe yönelik bir arama yaparak daha iyi sonuçlar elde edilmesini sağlamıştır. Optuna, önceki deneyimlere dayalı olarak her yeni hiper parametre kombinasyonunu seçerken, performans fonksiyonunun olasılıksal modelini güncellemekte ve böylece daha verimli bir arama süreci sağlamaktadır [54]. Optuna'nın TPE (Tree-structured Parzen Estimator) gibi yöntemleri, hiper parametrelerin dağılımlarını modelleyerek, performansın yüksek olduğu bölgeleri

hedeflemekte ve daha az performans gösteren bölgelerden kaçınmaktadır [55]. Bu sayede, hiper parametre alanında daha hızlı ve etkili bir keşif süreci gerçekleştirilmekte ve model performansını önemli ölçüde artırılmaktadır. Optuna'nın diğer yöntemlere kıyasla üstünlüğü, arama alanını dinamik olarak daraltarak daha hedefe yönelik bir şekilde hiper parametre seçimi yapmasından kaynaklanmaktadır. Literatürde Optuna'nın özellikle derin öğrenme modelleri gibi yüksek boyutlu hiper parametre alanlarına sahip problemler için daha etkili olduğu belirtilmektedir. Çünkü TPE algoritması, her iterasyonda performansı yüksek bölgeleri daha iyi modelleyerek, performansı düşük olasılık alanlarını hızla elemine edebilmektedir [56]. Ayrıca, Optuna'nın "prune" özelliği, düşük performans gösteren denemeleri erken durdurma imkânı sağlayarak zaman ve kaynak israfını önlemektedir. Bu özellik, özellikle yüksek sayıda hiper parametre kombinasyonu gerektiren senaryolarda, diğer yöntemlere göre önemli bir avantaj sağlamaktadır [57]. Literatürde, Optuna'nın hiper parametre optimizasyon sürecinde öğrenme oranı, ağaç derinliği veya minibatch boyutu gibi parametrelerin hassas bir şekilde ayarlanmasında etkili olduğu ve bu sayede daha iyi genelleme performansı elde edildiği gösterilmiştir [58]. Bunun yanı sıra, TPE, karmaşık hiper parametre etkileşimlerini daha iyi modelleyerek, klasik Bayesiyen optimizasyon yöntemlerinden daha iyi sonuçlar vermektedir [59]. Araştırmada kullanılan modeller açısından değerlendirildiğinde ise GBR diğer modellere kıyasla daha iyi genelleme yeteneği göstermiş ve tahminlerde daha düşük hata payına sahip olmuştur. Bu durum, GBR'nin konut fiyatlarındaki karmaşıklığı daha iyi öğrenme yeteneğine sahip olduğunu göstermektedir.

SVR modeli, düşük R^2 değeri (0.0099) ve yüksek RMSE değeri (7564.48) ile test setinde zayıf performans sergilemiştir. Bu durum, SVR'nin konut fiyatlarını doğru bir şekilde modelleyemediğini ve genelleme yeteneğinin düşük olduğunu göstermektedir. SVR'nin düşük performans göstermesi, özellikle veri setindeki karmaşıklığın bu modelin öğrenme yeteneğini sınırladığını düşündürmektedir. SVR'nin başarılı bir şekilde uygulanabilmesi için parametre ayarının ve kernel fonksiyonunun doğru seçilmesi önemlidir.

DTR ve RFR modelleri ise eğitim setinde yüksek performans göstermelerine rağmen, test setinde performanslarında önemli düşüşler yaşamışlardır. Bu durum, bu modellerin aşırı uyuma yatkın olduğunu ve genelleme yeteneklerinin sınırlı olduğunu göstermektedir. Aşırı uyma, bir modelin eğitim setine aşırı derecede uyum sağlayarak, bu veri kümesindeki gürültü ve rastgele varyasyonları öğrenme eğiliminde olduğu durumu ifade etmektedir. Bu durum, modelin eğitim verilerine mükemmel bir şekilde uymasıyla sonuçlanırken, gerçek dünya verileri üzerinde beklenmedik ve kötü tahminlere yol açmaktadır.

DTR ve RFR gibi karar ağacı tabanlı modeller, veri setindeki karmaşıklığı öğrenme yetenekleri nedeniyle aşırı uyma eğilimindedir. Özellikle eğitim setindeki yüksek performans, modelin veri setindeki detayları ezberleme kapasitesine işaret eder. Ancak, test setinde performansın düşük olması, modelin bu ezberlenen bilgileri genelleme yeteneğinin zayıf olduğunu gösterir. Karmaşık yapıların kontrol edilmesi, bu modellerin aşırı uyuma eğilimini azaltmaya yönelik önemli bir stratejidir. Bu durumu kontrol edebilmek için literatürde birkaç strateji geçmektedir. Modelin karmaşıklığını düzenleyen hiper parametrelerin (max_depth, n_estimators gibi) ayarlanması, aşırı uyum eğilimini azaltmaktadır. Ayrıca, veri setindeki gürültüyü azaltmak ve gereksiz karmaşıklığı önlemek için özellik mühendisliği (feature engineering) tekniklerinin uygulanması önerilmektedir.

Diğer doğrusal modeller olan LR, RR ve LAR modelleri benzer performans göstermişlerdir. Eğitim ve test setlerindeki R^2 değerleri yaklaşık olarak %57 olduğundan, bu modellerin hedef değişkenin varyansının önemli bir kısmını açıkladığı söylenebilir. Ancak, GBR modeline kıyasla tahminlerde daha yüksek hata payına sahiptirler. Doğrusal regresyon modelleri hedef değişkenin doğrusal bir kombinasyonunu modellediği için yeterince esnek olmayabilir. Ancak, bu modellerin genel olarak anlaşılır ve yorumlanabilir olması, bazı uygulamalarda tercih edilmelerine neden olmaktadır [34].

6. GELECEK ARAŞTIRMALAR VE UYGULAMALAR İÇİN ÖNERİLER (RECOMMENDATIONS FOR FUTURE RESEARCH AND APPLICATIONS)

Araştırma sonuçları, konut fiyatlarını tahminlemede en yüksek performansa sahip modelin GBR olduğunu göstermektedir. Ancak, model seçimi sürecinde dikkate alınması gereken önemli faktörler bulunmaktadır. Bu faktörler arasında güvenilirlik ve genel performansın dengeli bir şekilde ele alınması önemli bir rol oynamaktadır. Güvenilirlik, modelin güvenilir sonuçlar üretme kabiliyetini içermektedir. Genel performans ise modelin hem eğitim hem de test veri setlerinde başarılı bir şekilde çalışma yeteneğini ifade etmektedir. Özellikle hiper parametre ayarlama yöntemleri bu süreçte önemli

bir rol üstlenmektedir. Hiper parametrelerin uygun şekilde ayarlanması, modelin hem eğitim hem de test veri setlerinde daha iyi genelleme yapabilmesine ve tahmin performansını maksimize etmesine olanak tanımaktadır. Bu çalışmada Optuna, hiper parametrelerin optimize edilmesinde daha verimli ve hedefe yönelik bir arama yaparak GBR modelinin performansını önemli ölçüde artırmıştır.

Gelecekteki araştırmalarda, hiper parametre optimizasyonu sürecinin ötesine geçilerek, farklı makine öğrenmesi yöntemlerinin performanslarının kıyaslanması ve değerlendirilmesi önerilmektedir. Model performansını artırmak için "genetik algoritma" veya "benzetilmiş tavlama (simulated annealing)" gibi evrimsel ve doğadan ilham alan arama yöntemlerinin kullanımı araştırılabilir. Bu yöntemler, Grid Search veya Random Search farklı olarak, daha karmaşık ve çok boyutlu hiper parametre alanlarında daha etkin arama yapabilme potansiyeline sahiptir. Ayrıca, modellerin genelleme yeteneklerini geliştirmek için "topluluk (ensembling)" tekniklerinin (örneğin, bagging, boosting ve stacking) daha kapsamlı bir şekilde incelenmesi önerilmektedir. Bu yaklaşımlar, tek bir modelin sınırlamalarını aşmak ve daha dengeli ve güvenilir tahminler elde etmek için farklı modellerin birlikte kullanımını mümkün kılmaktadır. Örneğin, konut fiyat tahminleri için GBR modelinin yanı sıra, XGBoost, LightGBM veya CatBoost gibi diğer gelişmiş yöntemlerle yapılan topluluk modellerinin performansı değerlendirilebilir. Bununla birlikte, konut fiyat tahmin modellerinin geliştirilmesinde sadece model performansını artırmakla sınırlı kalmamak, aynı zamanda model açıklanabilirliğini de ön planda tutmak önerilmektedir. "SHAP değerleri" veya "LIME" gibi açıklanabilirlik tekniklerinin kullanımı, model tahminlerinin nasıl ve neden yapıldığına dair daha fazla içgörü sağlayabilir ve bu da modelin kullanıcılar ve paydaşlar tarafından daha kolay kabul görmesine olanak tanıyabilir. Bu tür tekniklerin, konut fiyatlarını etkileyen temel faktörlerin daha iyi anlaşılmasına katkı sağlayacağı öngörülmektedir. Bunlara ek olarak, coğrafi konum, konut özellikleri ve piyasa koşulları gibi farklı değişkenlerin modeller üzerindeki etkisinin daha derinlemesine incelenmesi için "coğrafi bilgi sistemleri (Geographic Information System - GIS)" ile entegre edilen modellerin geliştirilmesi, konut piyasasının dinamiklerini daha kapsamlı ve doğru bir şekilde analiz etmeye yardımcı olabilir. Bu yaklaşım, bölgesel farklılıkların ve yerel piyasa trendlerinin daha iyi anlaşılmasına olanak tanımaktadır.

Otomatik tahmin sistemlerinin kullanımı ve bu bilgilerin ekonomik istikrara olan katkısı alanda değerlendirilmesi gereken önemli bir konu başlığı olarak öne çıkmaktadır. Bu tür sistemlerin konut piyasasındaki dalgalanmaları önceden tahmin edebilme potansiyeli, ekonomik planlamada ve karar almalarda önemli bir rol oynayabilir. Bu konuda yapılacak daha fazla araştırma, otomatik tahmin sistemlerinin gerçek dünya uygulamalarında nasıl kullanılabileceği konusunda daha kapsamlı bilgiler sunabilir. Bu çalışma, konut fiyat tahminleme konusunda makine öğrenmesi tekniklerinin önemini ve kullanılabilirliğini göstermesiyle gelecekteki araştırmalar için bir yol haritası çizmektedir.

ETİK STANDARTLARIN BEYANI
(DECLARATION OF ETHICAL STANDARDS)

Bu makalenin yazarı, çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan etmektedir.

The author of this article declares that the materials and methods they use in their work do not require ethical committee approval and/or legal-specific permission.

YAZARLARIN KATKILARI (AUTHORS' CONTRIBUTIONS)

***Vahid SİNAP*:** Deneyleri yapmış, sonuçlarını analiz etmiş ve makalenin yazım işlemini gerçekleştirmiştir.

He conducted the experiments, analyzed the results, and performed the writing process.

ÇIKAR ÇATIŞMASI (CONFLICT OF INTEREST)

Bu çalışmada herhangi bir çıkar çatışması yoktur.

There is no conflict of interest in this study.

KAYNAKLAR (REFERENCES)

[1] Sornette D, Woodard R. Financial bubbles, real estate bubbles, derivative bubbles, and the financial and economic crisis. In: Econophysics approaches to large-scale business data and financial crisis. Springer Japan; 2010: 101-148.
[2] Tse RY. An application of the ARIMA model to real-estate prices in Hong Kong. Journal of Property Finance. 1997; 8(2): 152-163.
[3] Cervero R. Jobs-housing balancing and regional mobility. Journal of the American Planning Association. 1989; 55(2): 136-150.
[4] Ghysels E, Plazzi A, Valkanov R, Torous W. Forecasting real estate prices. In: Handbook of Economic Forecasting. Vol. 2. 2013: 509-580.

[5] Hutchison NE. Housing as an investment? A comparison of returns from housing with other types of investment. Journal of Property Finance. 1994; 5(2): 47-61.
[6] Pai PF, Wang WC. Using machine learning models and actual transaction data for predicting real estate prices. Applied Sciences. 2020; 10(17): 5832.
[7] Peter NJ, Okagbue HI, Obasi EC, Akinola AO. Review on the application of artificial neural networks in real estate valuation. International Journal. 2020; 9(3): 5-11.
[8] Herath SK, Maier G. The hedonic price method in real estate and housing market research. A review of the literature. Institute for Regional Development and Environment. 2010: 1-21. Vienna, Austria: University of Economics and Business.
[9] Taylor LO. Theoretical foundations and empirical developments in hedonic modeling. In: Hedonic methods in housing markets: Pricing environmental amenities and segregation. 2008: 15-37. New York, NY: Springer New York.
[10] Landajo M, Bilbao C, Bilbao A. Nonparametric neural network modeling of hedonic prices in the housing market. Empirical Economics. 2012; 42: 987-1009.
[11] Northcraft GB, Neale MA. Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. Organizational Behavior and Human Decision Processes. 1987; 39(1): 84-97.
[12] Patil P. A comparative study of different time series forecasting methods for predicting traffic flow and congestion levels in urban networks. International Journal of Information and Cybersecurity. 2022; 6(1): 1-20.
[13] Venkatachalam AR, Sohl JE. An intelligent model selection and forecasting system. Journal of Forecasting. 1999; 18(3): 167-180.
[14] Bojer CS. Understanding machine learning-based forecasting methods: A decomposition framework and research opportunities. International Journal of Forecasting. 2022; 38(4): 1555-1561.
[15] Almeida JS. Predictive non-linear modeling of complex data by artificial neural networks. Current Opinion in Biotechnology. 2002; 13(1): 72-76.
[16] Vanschoren J, Van Rijn JN, Bischl B, Torgo L. OpenML: Networked science in machine learning. ACM SIGKDD Explorations Newsletter. 2014; 15(2): 49-60.
[17] L'heureux A, Grolinger K, Elyamany HF, Capretz MA. Machine learning with big data: Challenges and approaches. IEEE Access. 2017; 5: 7776-7797.
[18] Felzmann H, Fosch-Villaronga E, Lutz C, Tamò-Larrieux A. Towards transparency by design for artificial intelligence. Science and Engineering Ethics. 2020; 26(6): 3333-3361.
[19] Maharana K, Mondal S, Nemade B. A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings. 2022; 3(1): 91-99.
[20] Bejani MM, Ghatee M. A systematic review on overfitting control in shallow and deep neural networks. Artificial Intelligence Review. 2021; 1-48.
[21] Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. Electronic Markets. 2021; 31(3): 685-695.
[22] Lu S, Li Z, Qin Z, Yang X, Goh RSM. A hybrid regression technique for house prices prediction. In: 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). 2017: 319-323. IEEE.
[23] Durganjali P, Pujitha MV. House resale price prediction using classification algorithms. In: 2019 International Conference on Smart Structures and Systems (ICSSS). 2019: 1-4. IEEE.
[24] Rahadi RA, Wiryono SK, Koesrindartoto DP, Syamwil IB. Factors affecting housing products price in Jakarta metropolitan region. International Journal of Property Sciences. 2016; 6(1).
[25] Alfiyatin AN, Febrita RE, Taufiq H, Mahmudy WF. Modeling house price prediction using regression analysis and particle swarm optimization case study: Malang, East Java, Indonesia. International Journal of Advanced Computer Science and Applications. 2017; 8(10).
[26] Osmadi A, Kamal EM, Hassan H, Fattah HA. Exploring the elements of housing price in Malaysia. Asian Social Science. 2015; 11(24): 26.
[27] Chau KW, Chin TL. A critical review of literature on the hedonic price model. International Journal for Housing Science and Its Applications. 2003; 27(2): 145-165.
[28] Ball MJ. Recent empirical work on the determinants of relative house prices. Urban Studies. 1973; 10(2): 213-233.
[29] Rodriguez M, Sirmans C. Managing corporate real estate: evidence from the capital markets. Journal of Real Estate Literature. 1996; 4(1): 13-33.
[30] Owusu-Manu DG, Edwards DJ, Donkor-Hyiaman KA, Asiedu RO, Hosseini MR, Obiri-

Yeboah E. Housing attributes and relative house prices in Ghana. International Journal of Building Pathology and Adaptation. 2019; 37(5): 733-746.

[31] Samuel AL. Some studies in machine learning using the game of checkers. IBM Journal of Research and Development. 1959; 3(3): 210-229.

[32] Mitchell TM. Machine learning. McGraw Hill. 1997.

[33] Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. 2009: 1-758. New York: Springer.

[34] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Vol. 112. 2013: 18. New York: Springer.

[35] Xu C, Lu C, Liang X, Gao J, Zheng W, Wang T, Yan S. Multi-loss regularized deep neural network. IEEE Transactions on Circuits and Systems for Video Technology. 2015; 26(12): 2273-2283.

[36] Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. Emerging Artificial Intelligence Applications in Computer Engineering. 2007; 160(1): 3-24.

[37] Speelman D. Logistic regression. Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy. 2014; 43: 487-533.

[38] Menard S. Coefficients of determination for multiple logistic regression analysis. The American Statistician. 2000; 54(1): 17-24.

[39] McDonald GC. Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics. 2009; 1(1): 93-100.

[40] Ranstam J, Cook JA. LASSO regression. Journal of British Surgery. 2018; 105(10): 1348-1348.

[41] Zhang F, O'Donnell LJ. Support vector regression. In: Machine learning. 2020: 123-140. Academic Press.

[42] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press. 2000.

[43] Xu M, Watanachaturaporn P, Varshney PK, Arora MK. Decision tree regression for soft classification of remote sensing data. Remote Sensing of Environment. 2005; 97(3): 322-336.

[44] Li Y, Zou C, Berecibar M, Nanini-Maury E, Chan JCW, Van den Bossche P, et al. Random forest regression for online capacity estimation of lithium-ion batteries. Applied Energy. 2018; 232: 197-210.

[45] Fu MC, Qu H. Regression models augmented with direct stochastic gradient estimators. INFORMS Journal on Computing. 2014; 26(3): 484-499.

[46] Zhang L, Liu Q, Yang W, Wei N, Dong D. An improved k-nearest neighbor model for short-term traffic flow prediction. Procedia-Social and Behavioral Sciences. 2013; 96: 653-662.

[47] Aitkin M, Foxall R. Statistical modelling of artificial neural networks using the multi-layer perceptron. Statistics and Computing. 2003; 13: 227-239.

[48] Liu W, Dou Z, Wang W, Liu Y, Zou H, Zhang B, Hou S. Short-term load forecasting based on elastic net improved GMDH and difference degree weighting optimization. Applied Sciences. 2018; 8(9): 1603.

[49] Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geoscientific Model Development. 2014; 7(3): 1247-1250.

[50] Prasad NN, Rao JN. The estimation of the mean squared error of small-area estimators. Journal of the American Statistical Association. 1990; 85(409): 163-171.

[51] Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research. 2005; 30(1): 79-82.

[52] Onyutha C. A hydrological model skill score and revised R-squared. Hydrology Research. 2022; 53(1): 51-64.

[53] Ranjan GSK, Verma AK, Radhika S. K-nearest neighbors and grid search cv based real time fault monitoring system for industries. In: 2019 IEEE 5th International Conference for Convergence in Technology (I2CT). 2019: 1-5. IEEE.

[54] Hanifi S, Cammarono A, Zare-Behtash H. Advanced hyperparameter optimization of deep learning models for wind power prediction. Renewable Energy. 2024; 221: 119700.

[55] Nguyen HP, Liu J, Zio E. A long-term prediction approach based on long short-term memory neural networks with automatic parameter optimization by Tree-structured Parzen Estimator and applied to time-series data of NPP steam generators. Applied Soft Computing. 2020; 89: 106116.

[56] Wang L, Xie S, Li T, Fonseca R, Tian Y. Sample-efficient neural architecture search by learning actions for Monte Carlo Tree Search. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021; 44(9): 5503-5515.

[57] Srinivas P, Katarya R. hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using

XGBoost. Biomedical Signal Processing and Control. 2022; 73: 103456.

[58] Bian K, Priyadarshi R. Machine learning optimization techniques: A survey, classification, challenges, and future research issues. Archives of Computational Methods in Engineering. 2024; 1-25.

[59] Zulfiqar M, Gamage KA, Kamran M, Rasheed MB. Hyperparameter optimization of Bayesian neural network using Bayesian optimization and intelligent feature engineering for load forecasting. Sensors. 2022; 22(12): 4446.