A STUDY ON HOUSE PRICE PREDICTION USING MACHINE LEARNING METHODS: THE CASE OF MADRID

Conference Paper · December 2021

CITATION
1

READS
584

3 authors, including:

Ersan Okatan
Burdur Mehmet Akif Ersoy University
15 PUBLICATIONS **14** CITATIONS
SEE PROFILE

Ismail Kirbas
Burdur Mehmet Akif Ersoy University
132 PUBLICATIONS **710** CITATIONS
SEE PROFILE

MAKİNE ÖĞRENME YÖNTEMLERİ KULLANARAK KONUT FİYAT TAHMİNİ ÜZERİNE BİR ÇALIŞMA: MADRİD ÖRNEĞİ

Mehmet ORAL*, Ersan OKATAN, İsmail KIRBAŞ*****

**Burdur Mehmet Akif Ersoy Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, Burdur, Türkiye, oralmehmet@gmail.com*
***Burdur Mehmet Akif Ersoy Üniversitesi, Gölhisar Uygulamalı Bilimler Yüksekokulu, Bilgisayar Teknolojileri ve Bilişim Sistemleri Bölümü, Burdur, Türkiye, ersanokatan@mehmetakif.edu.tr*
**** Burdur Mehmet Akif Ersoy Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Burdur, Türkiye, ismailkirbas@mehmetakif.edu.tr*

Özet: Gayrimenkul sektöründe alım satımın, ilgili tutarların çok yüksek olması nedeniyle önemli küresel, ulusal ve kişisel etkileri vardır. Sektörde profesyonel olarak çalışan firmaların fiyat belirleyebilmesi, fırsatları yakalayabilmesi ve değerleme uzmanlarının doğru fiyatlandırma yapabilmesi sektörle ilgili tüm aktörlerin ayakta kalabilmesi açısından önemlidir. Son yıllarda birçok alanda makine öğrenmesi yöntemlerinden artan beklentiler ile bu alanda da beklentiler artmıştır. Bu çalışmada konu ile ilgili yapılacak çalışmalara yol göstermesi açısından konut fiyat tahmini için çeşitli makine öğrenmesi yöntemleri kullanılmış ve performans karşılaştırması yapılmıştır. Bir veri depolama platformundan alınan gerçek veriler farklı makine öğrenmesi algoritmalarında fiyat tahmini yapmak için kullanılmıştır. Konutlara ait 22 farklı özellik modellerin oluşturulması için kullanılmıştır. Farklı yöntemlerin denenmesi sonucu ortaya çıkan sonuçlardan en iyi performans değerine sahip yöntemler sırasıyla Bagged Trees Ensemble, Fine Tree, Exponential GPR, Wide Neural Network, Quadratic SVM olmuştur. Yapılan çalışmanın konut değerleme için kullanılan uygulamaların geliştirilmesine ve bu alanda yapılan bilimsel çalışmalara katkı sağlayacağı düşünülmektedir.

Anahtar Kelimeler: Ağaç Toplulukları, Konut Fiyat Tahmini, Makine Öğrenmesi, Yapay Sinir Ağları.

A STUDY ON HOUSE PRICE PREDICTION USING MACHINE LEARNING METHODS: THE CASE OF MADRID

ABSTRACT: Buying and selling in the real estate market has significant global, national and personal impacts because the amounts involved are so high. It is important for the companies working professionally in the sector to be able to determine prices, seize opportunities, and the valuation experts to make the right pricing for the survival of all actors in the sector. In recent years, with the increasing expectations from machine learning methods in many fields, expectations in this field have also increased. In this study, various machine-learning methods were used for house price prediction and performance comparison was made in order to guide the studies to be made on the subject. Real data from a data storage platform is used to make price predictions in different machine learning algorithms. As a result of trying different methods, the methods with the best performance values are Bagged Tress Ensemble, Fine Tree, Exponential GPR, Wide Neural Network, Quadratic SVM, respectively. It is thought that the study will contribute to the development of applications used for housing valuation and to scientific studies in this field.

Keywords: Ensemble Trees, House Price Prediction, Machine Learning, Neural Networks

1.GİRİŞ

Dünyada gayrimenkul sektör büyüklüğü 2020 yılında COVID-19'un küresel etkilerine rağmen 10,5 trilyon dolar hacme ulaşmıştır [1]. Böylesine büyük bir hacimde yapılan işlemler hem ülke, hem şehir hem de bölge ekonomisi için çok önemlidir. Bunun yanında insanların hayatları boyunca oluşturdukları birikimleri değerlendirmek için kullandıkları gayrimenkul alımları kişisel ölçüde çok büyük önem taşımaktadır[2]. İsteklere uygun bir konutu doğru fiyata alabilmek için büyük çaba sarf etmek gerekir. Bununla birlikte emlak sektöründeki firmaların aldıkları konutu uygun fiyatla alabilmesi, doğru fiyatla satabilmesi firmanın geleceği adına da büyük önem taşımaktadır. Büyük rakamlarla yapılan işlemler nedeniyle firmalar yanlış fiyatlandırma sonucu kısa sürede iflas edebilir. Sadece alım satım yapan firmalar değil aynı zamanda inşaat sektöründeki firmalarında konutları doğru fiyata satabilmesi çok önemlidir[3]. İnşaat sürecinin uzun sürmesi nedeniyle fiyatlar süreç içerisinde sürekli değişebilmekte bazen ülke, bazen küresel çapta oluşan maliyet artışları nedeniyle, bazen de konutun bulunduğu bölgelerdeki ani talep artışları nedeniyle fiyatlar hızla değişmektedir. Satış için doğru fiyat belirlenmemesi düşük fiyatla satarak zarara ya da yüksek fiyat nedeniyle satışların gecikmesine neden olmaktadır. Fiyatlar bazı bölgelerde hızla değişebilmekte, talep alanları günler içerisinde farklılaşmakta, yerel politik gelişmelerle değişmekte ve alıcı ya da satıcıların değişimi takip edemeyip zararlarına yol açabilmektedir. Alım ve satım sırasında ortaya çıkan anlaşmazlıkların daha az olması için de fiyat tahmin eden uygulamaların kullanılması yararlı olacaktır[3].

Fiyat belirlenmesi için kullanılan yöntemlerden biri Konut Fiyat Endeksi'ni kullanmaktır. Bu endeks ev alım işlemlerinin banka kredisi ile yapılması durumunda hesaplanmakta ve ortalama olarak o bölgedeki fiyatı temsil etmektedir. Böylelikle genel olarak fiyatlardaki düşüş ve yükselişler izlenebilmektedir. Ancak ortalama olarak yapılan bu hesap ile bir evin detaylı özelliklerine göre fiyatlandırılması mümkün değildir [2]. Makine öğrenmesi yöntemleri kullanılarak konutların fiyatlarının belirlenmesi hem yeni iş modelleri oluşturulması hem de alıcı – satıcıların daha kolay karar verebilmesini sağlar. Ancak oluşturulacak modellerin kullanılabilmesi için etkin bir model oluşturulması çok önemlidir. Bu amaçla konut fiyat tahmini ile ilgili birçok çalışma yapılmıştır. Yapılan bir çalışmada Random Forest(RF), Light GBM ve XGBoost yöntemleri ve bunların çeşitli melez varyasyonları konut fiyatları tahmini için incelenmiştir, RF en iyi performansa sahip yöntem olarak bulunmuş ancak işlem süresi daha fazla olmuştur [4]. Başka bir çalışmada RF, XGBoost ve Adaboost yöntemleri kullanılmış daha sonra lineer regresyon ile birleştirilerek performans artışı sağlanmıştır [5], ayrıca konut özelliklerinden tahmin sonucunda etkili olanlar belirlenmiştir. Lineer, Ridge, Lasso, Support Vector ve XGBoost regresyonlarının karşılaştırıldığı bir çalışmada XGBoost ile en iyi sonuç elde edilmiştir [6]. Başka bir çalışmada konut fiyat tahmini için farklı yöntemler karşılaştırılmış ve Stepwise – SVM en iyi yöntem olarak bulunmuştur [7]. Fan ve arkadaşları çalışmalarında konut fiyat tahmini için özellik seçimi ve kayıp veri tamamlama yöntemleri kullanılmıştır [8]. Jamil ve arkadaşları yaptıkları çalışmada yeşil binaların özelliklerini kullanarak fiyat tahmini yapmış, 5 farklı yöntemde en iyi performans sonucunu Karar Ağaçları yöntemi ile elde etmişlerdir [9]. Varma ve arkadaşları çalışmalarında bazı makine öğrenmesi yöntemlerini yapay sinir ağları yöntemi ile birlikte kullanarak daha iyi sonuçlar elde etmişlerdir [10]. Lu ve arkadaşları Lasso ve

Gradient Boosting metotları ile oluşturulan hibrit bir yöntemle en iyi sonucu elde etmişlerdir. Hibrit modelde %65 Lasso ile %35 Gradient Boosting kombinasyonu kullanılmıştır [11].

Bu çalışmada konut fiyat tahmini için makine öğrenme yöntemleri incelenmiştir. Bunun için Madrid'de çeşitli dijital ortamlarda satış için ilan edilmiş bazı konutlara ait veriler, bir veri depolama platformundan alınmıştır. Madrid Gayrimenkul Piyasası adındaki bu veri kümesi gerekli bazı düzenlemeler yapılarak farklı makine öğrenme yöntemleri ile fiyat tahminleri gerçekleştirilmiştir. Veri kümesindeki verilerin bir kısmı eğitim bir kısmı test için kullanılmış, test için kullanılan verilerin gerçek değerleri ile tahmin değerleri arasındaki farktan yararlanılarak performans sonuçları elde edilmiştir. Performans sonuçları karşılaştırılarak konut fiyat tahmini ile ilgili yöntemlerin analiz edilmesi hedeflenmiştir.

Bu çalışma 4 bölümden oluşmaktadır. 1. Bölüm 'de konut fiyat tahmini konusunda bilgiler verilmiş, 2. Bölüm' de veri kümesi ve düzenlenmesi ve kullanılan makine öğrenmesi yöntemleri hakkında bilgiler verilmiştir. 3. Bölüm 'de elde edilen bulgular verilmiş, son bölümde de ise bulgular ile ilgili değerlendirmeler yapılmıştır.

2.YÖNTEM

Bu çalışmada bir veri kaynağından alınan emlak fiyatları ve özellikleri kullanılmış, veriler eğitim ve test için ayrılıp kullanılan algoritmalarla oluşturulan model test edilmiştir.

2.1. Veri Kümesi ve Analizi

Kullanılan veriler Kaggle veri depolama platformundan alınan Madrid Gayrimenkul Piyasası veri kümesidir. 2021 yılında Avrupa'nın en kalabalık 5. Şehri olan İspanya'nın başkenti Madrid ülkenin siyasi, ekonomik ve kültür merkezidir. Ekonomik üretimi, yüksek yaşam standardı ve pazar büyüklüğüne sahip Madrid şehrinin metropolit nüfusu yaklaşık 6,5 milyondur. Şehir, ilçe adı verilen 21 idari bölüme ayrılmıştır. Bu veri kümesinde Madrid şehrindeki bu ilçelerde bulunan ve satışa çıkarılan konutların fiyatları, özellikleri ile birlikte verilmiştir [12]. Veri kümesinde 21742 kayıt bulunmaktadır. Her kayıtta 58 farklı özellik verilmiştir. Bu özelliklere oda sayısı, metrekare, banyo sayısı, cadde adı, cadde numarası, kiralama fiyatı, konut tipi, yapım yılı ve asansör, park alanı, yüzme havuzu gibi seçeneklerin bulunma durumu örnek olarak verilebilir. Veriler makine öğrenmesi algoritmalarında kullanılabilmesi için bazı ön işlemlerden geçirilmiştir. Bu işlemler şunlardır;

- Tamamı veya büyük çoğunluğu eksik özelliklerin tamamen silinmesi (örneğin has_public_parking, portal, are_pets_allowed)
- Tamamı tek bir değerden oluşan özelliklerin silinmesi (örneğin is_rent_price_known)
- Anlamsız değer bulunan kayıtların silinmesi (örneğin negatif fiyat değeri içeren kayıtlar)
- Birbirinin aynısı veya benzeri olan özelliklerden sadece birer tanesinin kullanılması diğerlerinin silinmesi (örneğin sq_mt_built saklanırken, sq_mt_allotment ve sq_mt_useful silindi)

- Verilerin derlendiği ilanlarda havuz, teras gibi ek özellikler sadece mevcut olduğunda belirtildiği için, girilmeyen değerlerin eklenmesi (örneğin sadece "true" değerleri bulunuyorsa kalan değerler "false" olarak belirlendi)
- Kategorik değişkenler yerine kukla değişkenler eklenmesi (örneğin energy_certificate değişkeni)
- Açık adres, ilan başlığı vb. gerekli olmayan özelliklerin silinmesi
- Bölge ortalama metrekare fiyatı ilgili özelliklerden alınarak oluşturulması

Gerekli düzenlemeler yapıldıktan sonra 8956 kayıt ve 22 özellikten oluşan veri kümesinde bulunan özellikler Tablo 1 de gösterilmiştir.

Tablo 1: Konut özellikleri

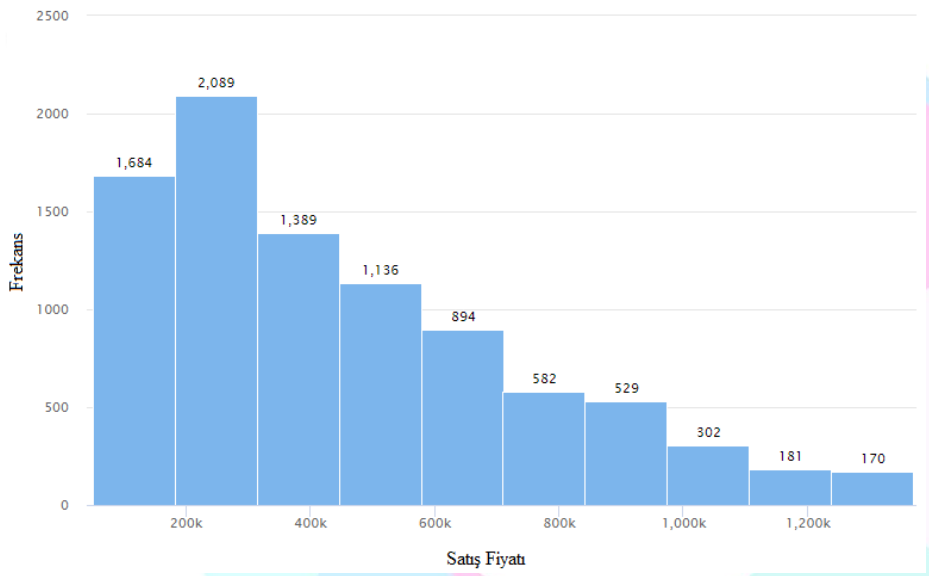
Özellik Adı	Açıklama	Özellik Adı	Açıklama
sq_mt_built	Metrekare alanı	has_fitted_wardrobes	Gömme gardırop var mı?
n_rooms	Oda sayısı	has_lift	Asansör var mı?
n_bathrooms	Banyo sayısı	has_garden	Bahçe var mı?
floor	Bulunduğu kat	has_pool	Havuz var mı?
average_sm_price	Ortalama metrekare fiyatı	has_terrace	Teras var mı?
rent_price	Kiralama fiyatı	has_balcony	Balkon var mı?
buy_price	Satış fiyatı	has_storage_room	Depo var mı?
house_type	Konut tipi	has_green_zones	Yeşil alan var mı?
is_renewal_needed	Yenileme gereklimi	energy_certificate	Enerji sertifikası
is_new_development	Yeni yapı mı?	has_parking	Park alanı var mı?
has_ac	Klima var mı?	is_orientation	Cephe bilgisi

Verilerin, modelleme için kullanılmasından önce analiz çalışması yapılmıştır. Konut özellikleri için korelasyon matrisi incelendiğinde en yüksek korelasyona sahip özellikler Tablo 2'de verilmiştir. Buna göre satış fiyatında en çok kiralama fiyatı, konut metrekaresi, banyo sayısının etkili olduğu söylenebilir. Diğer korelasyon değerleri incelendiğinde beklenen değerlerin ortaya çıktığı görülmüştür.

Tablo 2:Korelasyon Değerleri

Attributes	sq_mt_built	n_rooms	n_bathrooms	average_sm_price	rent_price	buy_price
sq_mt_built	1	0.709	0.809	0.235	0.670	0.802
n_rooms	0.709	1	0.635	0.069	0.470	0.529
n_bathrooms	0.809	0.635	1	0.262	0.643	0.740
average_sm_price	0.235	0.069	0.262	1	0.650	0.621
rent_price	0.670	0.470	0.643	0.650	1	0.835
buy_price	0.802	0.529	0.740	0.621	0.835	1

Konut fiyatı için dağılım grafiği Şekil 1'de verilmiştir.



Şekil 1: Fiyat dağılım grafiği

Makine öğrenmesi yöntemlerinin uygulanması için MATLAB programı kullanılmıştır. Fiyat değeri bağımlı değişken olarak belirlenmiş, diğer özellikler bağımsız değişkenler olarak kullanılmıştır.

2.2 Model Seçimi ve Performans Kriterleri

- Belirlilik katsayısı (R^2):

Tahmin edilen değerler ile gerçek değerler arasındaki farkın ne kadar çok olduğuna bağlı olarak ortaya çıkan modelin performans ölçümünde kullanılabilecek bir değerdir. Elde edilen sonuçlar gerçek değere ne kadar yakınsa o kadar iyi uyum sağlandığı kabul edilir. R^2 İfadesinde sonuç her zaman pozitiftir ve [0,1] aralığında bir değer alır. Sonucun 1'e yakın olması tahmin edilen değerin gerçek değerleri açıklamada yüksek doğruluğa sahip olduğunu ifade eder. Aşağıda Denklem 1'de ifadesi verilmiştir.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \mu)^2}$$

(1)

- Mean Squared Error (MSE) – Ortalama Kare Hata:

Ortalama Kare Hatası (MSE) tüm değerler için hataların karesini toplayıp ortalamasının alınması ile bulunmaktadır. Denklem 2'de gösterilmiştir. MSE değeri her zaman pozitif çıkacaktır. Verilerin farklı skaladaki değerlerle modele uygulanması durumunda sonuçlar değerlerin büyüklüğüne göre çıkacağı için hatalı karşılaştırmalara neden olabilir. Çalışmamızda değerler aynı skalada olduğu için bu değerlendirme kriteri kullanılabilir.

$$MSE = \frac{1}{n}\sum_{j=1}^{n} e_j^2 \quad (2)$$

Kare nedeniyle, büyük hataların MSE üzerinde küçük hatalardan daha fazla etkisi vardır. Bu nedenle MAE, kareden faydalanmadığı için aykırı değerlere karşı daha sağlamdır.
Tahmin değerleri ile gerçek değerler arasındaki uzaklığın bulunmasında sıkılıkla kullanılan, hatanın büyüklüğünü ölçen RMSE bir nevi tahmin hatalarının standart sapmasıdır. RMSE ifadesi Denklem 3'te gösterilmiştir. Mutlak değerin kullanılması istenilmeyen durumlarda RMSE kullanılması tercih edilir.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n} e_j^2} \quad (3)$$

$$RMSE = \sqrt{SME} \quad (4)$$

Verilerde aykırı değer varsa ve bunları yok saymak istiyorsanız, MAE daha iyi bir seçenektir ancak bunları kayıp fonksiyonunuzda hesaba katmak istiyorsanız MSE yerine RMSE seçilmesi daha iyidir.

3.BULGULAR

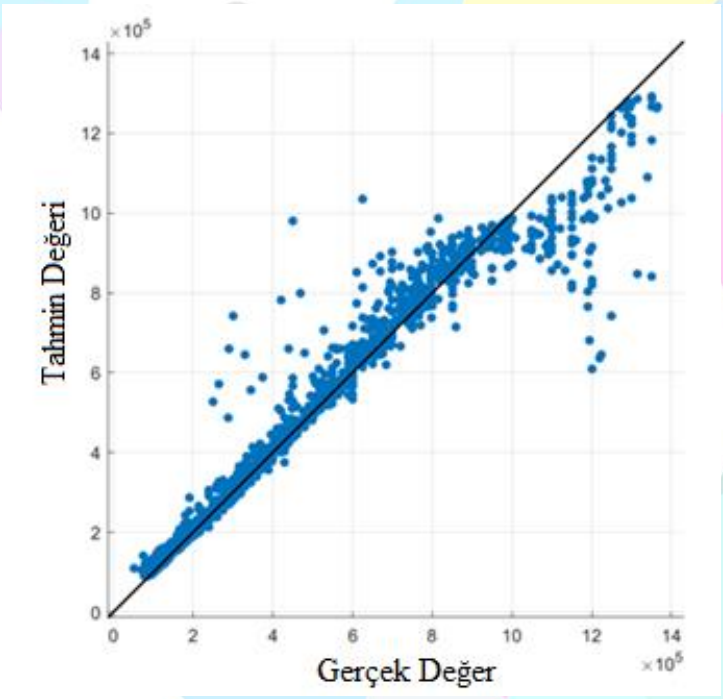
Veri kümesi kullanılarak makine öğrenme metotları ile konutlara ait fiyat değerleri tahmin edilmeye çalışılmıştır. Veriler rastgele şekilde %75 eğitim için, %25 test için kullanılmıştır. Matlab programında bulunan çok sayıda makine öğrenme algoritmaları ile test edilmiştir. Bunlardan en iyi R^2 değerine ulaşan 5 metot ve sonuçları Tablo 3'te verilmiştir.

Tablo 3: Performans sonuçları

Model	Quadratic SVM	Wide Neural Network	Exponential GPR	Fine Tree	Bagged Trees Ensemble
RMSE	76130	**68605**	**66661**	**66464**	**61526**

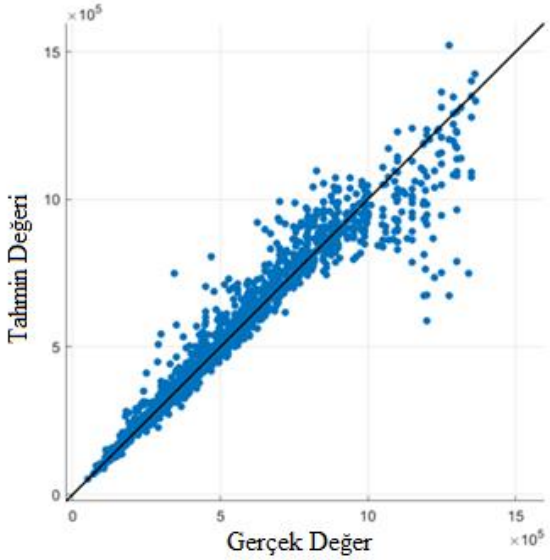
R^2	0.93	**0.95**	**0.95**	0.95	**0.96**
MSE	5.7958e+09	**4.7066e+09**	**4.4436e+09**	**4.4175e+09**	**3.7855e+09**
MAE	33579	**24886**	**31340**	**12206**	**25305**
Tahmin hızı: ~ (obs/sec)	10000	**5300**	**5000**	**160000**	**10000**
Eğitim süresi: (sec)	50.512	**215.88**	**128.87**	**1.1873**	**12.209**

En iyi sonuç Bagged Trees Ensemble algoritması ile elde edilmiştir. Bagged Trees Ensemble algoritması ile yapılan tahmin sonucu oluşan grafik Şekil 2'de verilmiştir.



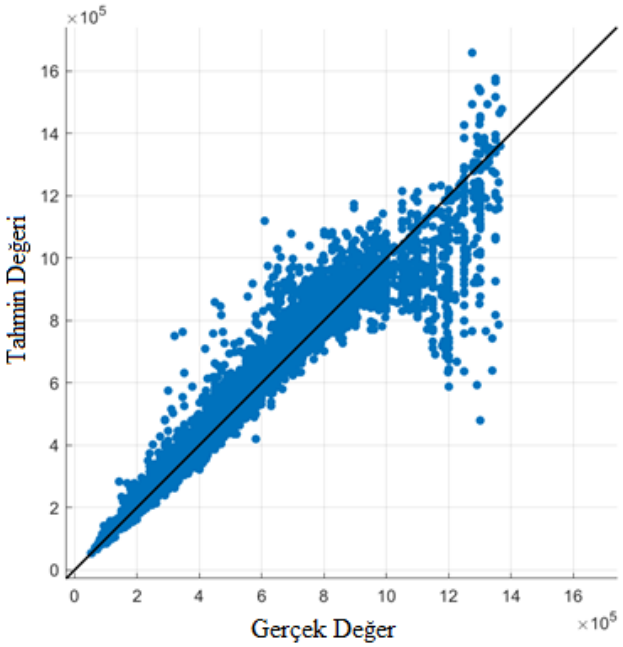
Şekil 2: Bagged Trees Ensemble algoritması ile elde edilen tahmin grafiği

Fine Tree algoritması ile elde edilen sonuç Şekil 3'te verilmiştir.



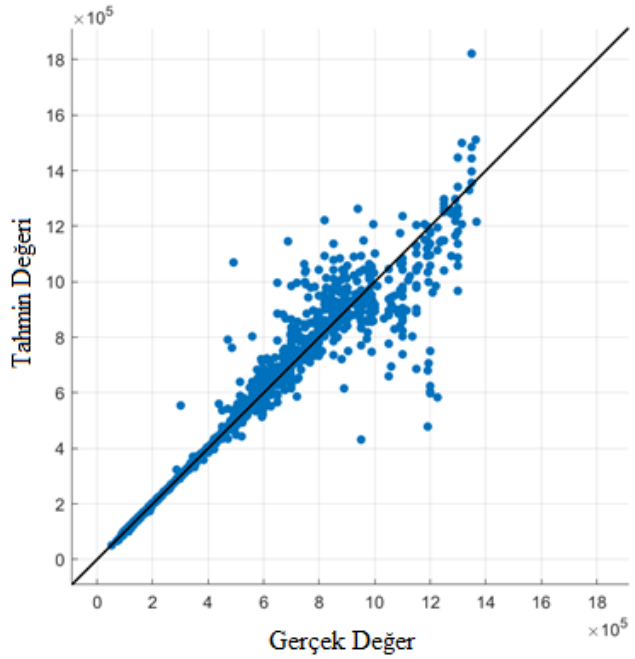
Şekil 3: Fine Tree algoritması ile elde edilen tahmin grafiği

Exponential GPR algoritması ile elde edilen sonuç Şekil 4'te verilmiştir.



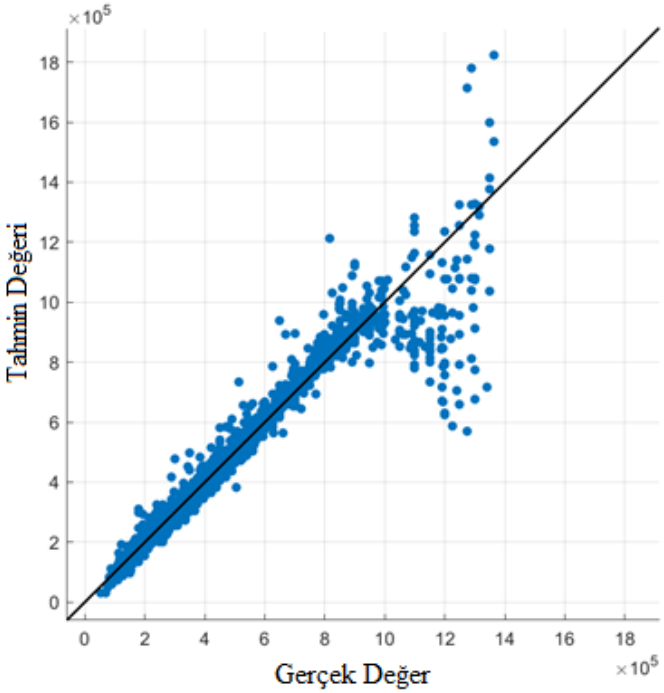
Şekil 4: Exponential GPR algoritması ile elde edilen tahmin grafiği

Wide Neural Network algoritması ile elde edilen sonuç Şekil 5'te verilmiştir.



Şekil 5: Wide Neural Network algoritması ile elde edilen tahmin grafiği

Quadratic SVM algoritması ile elde edilen sonuç Şekil 6'da verilmiştir.



Şekil 6: Quadratic SVM algoritması ile elde edilen tahmin grafiği

Kullanılan makine öğrenme yöntemleri ile elde edilen tahmin sonuçları grafiklerine bakıldığında küçük fiyat değerlerine sahip konutların değerlerinin daha iyi tahmin edildiği görülmektedir.

4.SONUÇ

Bu çalışmada konut özellikleri ve fiyat değerleri içeren bir veri kümesi kullanılarak farklı makine öğrenme yöntemleri karşılaştırılmıştır. Elde edilen sonuçlara göre en iyi performans Bagged Trees Ensemble ile elde edilmiştir. Diğer performans sonuçlarının da yakın değerlerde olduğu görülmüştür.

Yapılan çalışmanın konut fiyat tahmini ile ilgili çalışmalarda yol gösterici olacağı düşünülmektedir. Ülkemizde bu konuda yapılan çalışma sayısı az olduğu için sonraki çalışmalarda Türkiye'de belirli bölgelerdeki konut fiyat tahmini üzerinde çalışmalar yapılabilir. Sonraki yapılacak çalışmalarda özellik seçimi ile ilgili detaylı araştırmalar yapılmasının uygun olacağı düşünülmektedir.

5.KAYNAKÇA

[1] "Real Estate Market Size Report 20/21". [Çevrimiçi]. Erişim adresi: https://www.msci.com/documents/10199/a4535e8e-3b0d-f34d-4a0b-dc73058f7469

[2] (2021) Housing Price Prediction via Improved Machine Learning Techniques - ScienceDirect.

[3] Karlik, B. (2017) DEVELOPMENT OF OPTIMUM ANN STRUCTURES FOR HOUSE PRICE ESTIMATION.

[4] Truong, Q., Nguyen, M., Dang, H., and Mei, B. (2020) Housing price prediction via improved machine learning techniques. *Procedia Computer Science*. 174 433–442.

[5] Srirutchataboon, G., Prasertthum, S., Chuangsuwanich, E., Pratanwanich, P.N., and Ratanamahatana, C. (2021) Stacking Ensemble Learning for Housing Price Prediction: a Case Study in Thailand. in: 2021 13th Int. Conf. Knowl. Smart Technol. KST, IEEE, Bangsaen, Chonburi, Thailandpp. 73–77.

[6] Manasa, J., Gupta, R., and Narahari, N.S. (2020) Machine learning based predicting house prices using regression techniques. in: 2020 2nd Int. Conf. Innov. Mech. Ind. Appl. ICIMIA, IEEE, pp. 624–630.

[7] Phan, T.D. (2018) Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. in: 2018 Int. Conf. Mach. Learn. Data Eng. ICMLDE, pp. 35–42.

[8] Fan, C., Cui, Z., and Zhong, X. (2018) House prices prediction with machine learning algorithms. in: Proc. 2018 10th Int. Conf. Mach. Learn. Comput., pp. 6–10.

[9] Jamil, S., Mohd, T., Masrom, S., and Ab Rahim, N. (2020) Machine Learning Price Prediction on Green Building Prices. in: 2020 IEEE Symp. Ind. Electron. Appl. ISIEA, IEEE, TBD, Malaysiapp. 1–6.

[10] Varma, A., Sarma, A., Doshi, S., and Nair, R. (2018) House price prediction using machine learning and neural networks. in: 2018 Second Int. Conf. Inven. Commun. Comput. Technol. ICICCT, IEEE, pp. 1936–1939.

[11] Lu, S., Li, Z., Qin, Z., Yang, X., and Goh, R.S.M. (2017) A hybrid regression technique for house prices prediction. in: 2017 IEEE Int. Conf. Ind. Eng. Eng. Manag. IEEM, IEEE, Singaporepp. 319–323.

[12] (2021) "Madrid real estate market". https://kaggle.com/mirbektoktogaraev/madrid-real-estate-market (erişim Kas. 18, 2021).