Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 27, Sayı 1, 152-159, 2023

Süleyman Demirel University
Journal of Natural and Applied Sciences
Volume 27, Issue 1, 152-159, 2023

DOI: 10.19113/sdufenbed.1190150

Konut Fiyatlarının Tahmini için Polinomsal Regresyon ve Yapay Sinir Ağları Yöntemlerinin Uygulamalı Karşılaştırılması

Zeynep BARUT*[1], Turgay Tugay BİLGİN[2]

[1,2]Bursa Teknik Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği Bölümü, 16310, Bursa, Türkiye

(Alınış / Received: 16.10.2022, Kabul / Accepted: 16.01.2023, Online Yayınlanma / Published Online: 25.04.2023)

Anahtar Kelimeler
KNIME,
Makine Öğrenmesi,
Yapay Sinir Ağları,
Polinomsal Regresyon

Öz: Gayrimenkul sektörünün hızlı ekonomik büyümesi nedeniyle, konut fiyatlarının tahmini geleceğe yönelik planlamalar için önemlidir. Bu çalışmanın amacı makine öğrenmesi yöntemlerini kullanarak bir konutun potansiyel fiyatını tahmin etmektir. Makine öğrenmesi yöntemleri olarak Yapay sinir ağları ve Polinomsal regresyon kullanılarak bunların tahmin performansları karşılaştırılmıştır. Makine öğrenmesi yöntemlerinin uygulanabilmesi için KNIME veri analiz platformu kullanılmıştır. Yöntemlerin başarısını ölçmek için R Kare performans metriği kullanılmıştır. Uygulama sonuçları, Yapay sinir ağları yönteminin Polinomsal regresyon yöntemine göre ev fiyatlarını daha yüksek doğrulukla tahmin ettiğini göstermektedir. Yapılan çalışmanın ev değerlendirilmesi için kullanılan uygulamaların geliştirilmesine ve bu alanda yapılan bilimsel çalışmalara katkı sağlayacağı düşünülmektedir. Sonraki çalışmalarda farklı yöntemler veya ev özniteliklerinin bulunduğu veri setleri kullanılarak çalışmanın genişletilmesi hedeflenmektedir.

Applied Comparison of Polynomial Regression and Artificial Neural Networks Methods for Prediction of House Prices

Keywords
KNIME,
Machine Learning,
Neural Networks,
Polynomial Regression

Abstract: Due to the rapid economic growth of the real estate sector, the prediction of housing prices is important for future planning. The aim of this study is to predict the potential price of a house using machine learning methods. As machine learning methods, artificial neural networks and polynomial regression were used and their prediction performances were compared. KNIME data analysis platform was used to apply machine learning methods. R Squared performance metric was used to measure the success of the methods. The application results show that the artificial neural network method predicts house prices with higher accuracy than the Polynomial regression method. It is thought that the study will contribute to the development of applications used for home evaluation and to scientific studies in this field. In future studies, it is aimed to expand the study by using different methods or data sets with home attributes.

1. Giriş

Gayrimenkul sektörü ülkemizde ekonomik büyümeye katkı sağlayan ve gittikçe gücü artan bir yapıya sahiptir. Sektöre yönelik yüksek yatırımlar fiyatlar düzeyinde önemli artışlara neden olmaktadır. Alım satım işlemleri, ilgili fiyatlar yüksek olduğunda önemli ulusal etkilere neden olmaktadır. Konut fiyatları ülke ekonomisinin değişimi için önemli olduğundan, bu fiyatların tahmini geleceğe yönelik planlamalar için önemli bir veridir. Bu nedenle, teknolojinin gelişmesiyle birlikte bu tür tahminler için Makine öğrenmesi, Yapay sinir ağları ve Derin öğrenme yöntemleri sıklıkla kullanılmaya başlanmıştır. Değişkenler arasındaki ilişkileri bulmak için kullanılan yöntemler, çeşitli alanlarda kullanılmaktadır. En iyi sonucu verecek yöntemlerin kullanılan veri setine uygun olarak belirlenmesi gerekmektedir [1].

Oral ve arkadaşlarının yayınladığı "Makine Öğrenme Yöntemleri Kullanarak Konut Fiyat Tahmini Üzerine Bir Çalışma: Madrid Örneği" [2] isimli çalışmasında, konut fiyatlarının tahmin edilmesi için çeşitli makine öğrenmesi yöntemleri kullanılarak performans değerleri karşılaştırılmıştır. Yapılan çalışma

*İlgili yazar: zebarut@gmail.com

sonucunda en iyi performans gösteren yöntemler Bagged Trees Ensemble, Fine Tree, Exponential Gaussian Process Regression, Wide Neural Network, Quadratic Support Vector Machine olarak bulunmuştur. Emrah ve arkadaşlarının yayınladığı "Yapay Zekâ ile Konut Fiyatlarının Tahmin Edilmesi" [3] isimli çalışmada, internette yayınlanan satılık konut verilerini toplayarak konut fiyat tahmini yapan bir model geliştirilmiştir. En başarılı fiyat tahmini Rastgele Orman yöntemi ile elde edilmiştir. Yapılan çalışma ile firmaların ve tüketicilerin konut fiyatlarını tahmin edebileceği bir sistem geliştirilmiştir. Fidan ve arkadaşlarının yayınladığı "Farklı Regresyon Analizi Yöntemleri Kullanılarak Ev Fiyatlarının Tahmini" [4] isimli çalışmada, konut fiyatlarının regresyon analizi yöntemleriyle tahmin edilmesi sağlanmıştır. Bu amaçla Doğrusal regresyon, Karar ağacı ve Rastgele Orman regresyon yöntemleri veriler üzerinde test edilmiştir. Yapılan çalışma ile Doğrusal regresyon yönteminin diğer yöntemlere göre daha doğru tahminler gerçekleştirdiği görülmüştür. Özgür ve arkadaşlarının yayınladığı "Konut Fiyat Tahmininde Yapay Sinir Ağları Yönteminin Kullanılması" [5] isimli çalışmada, konut fiyatlarının tahmin edilmesi için Yapay sinir ağları kullanılmıştır. Farklı fiziksel özellikler ve çeşitli parametreler ile Yapay sinir ağları modelleri oluşturulmuştur. Gizli katman nöron sayıları değiştirilerek modeller oluşturulmuş ve bu modellerin performansları karşılaştırılarak en uygun gizli katman nöron sayısı belirlenmiştir. Yapılan çalışma ile Yapay sinir ağlarının konut fiyatlarının tahmin edilmesinde iyi bir yöntem olduğu görülmüştür. Hadavandi ve arkadaşlarının yayınladığı "An Econometric Panel Data-Based Approach for Housing Price Forecasting in Iran" [6] isimli çalışmada, Tahran'ın 20 farklı bölgesindeki konut fiyatlarının tahmin edilmesi için bir model oluşturulmuştur. Modelleme için tek yönlü sabit etkiler ve tek yönlü rastgele etkiler yaklaşımları (panel veri yaklaşımları) uygulanmıştır. Sonuçlar, bu alanda yaygın olarak kullanılan en küçük kareler yaklaşımı ile karşılaştırılmıştır. Yapılan çalışma ile tek yönlü sabit etkiler yaklaşımının daha doğru tahminler sağladığı görülmüştür. Bu çalışma Tahran'daki konut piyasasını analiz etmek için panel veri yaklaşımını kullanan ilk çalışmadır. Phan'ın yayınladığı "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia" [7] isimli çalışmada, Avustralya'daki tarihi mülk işlemlerini analiz etmek için makine öğrenmesi teknikleri uygulanmıştır. Yapılan çalışma ile Melbourne şehrinin en pahalı ve en uygun bölgelerindeki ev fiyatları arasında yüksek tutarsızlık ortaya çıkmıştır. Ayrıca, ortalama hata karesi ölçümüne dayalı Stepwise ve Destek Vektör Makinesi kombinasyonunun rekabetçi bir yaklaşım olduğu görülmüştür. Aynı çalışma Avustralya genelindeki farklı konumlardan konut piyasasının işlemsel veri kümeleri için de uygulanmaktadır. Park ve arkadaşının yayınladığı "Using Machine Learning Algorithms for Housing Price Prediction: The Case of Fairfax County, Virginia Housing Data" [8] isimli çalışmada, konut fiyat tahmin modeli geliştirmek için makine öğrenmesi yöntemleri kullanılmıştır. C4.5, RIPPER, Naive Bayes ve AdaBoost makine öğrenmesi yöntemleri kullanılarak bir model geliştirilmiş ve doğruluk performansları karşılaştırılmıştır. Yapılan çalışma ile, doğruluğa dayalı RIPPER algoritmasının diğer modellerden tutarlı bir şekilde daha iyi performans gösterdiği görülmüştür. Teoh ve arkadaşlarının yayınladığı "Explainable Housing Price Prediction with Determinant Analysis" [9] isimli çalışmada, farklı veri setleri kullanılarak konut fiyatlarını tahmin etmek için regresyon tabanlı bir makine öğrenmesi modeli geliştirilmiştir. Konut fiyatlarını etkileyen önemli belirleyiciler Çok Terimli Lojistik Regresyon kullanılarak belirlenmiştir. Konut fiyatlarında büyük değişimlere neden olan özellikleri incelemek için Shapley Additive Explanations analizi kullanılarak kapsamlı bir çalışma yapılmıştır.

2. Materyal ve Metot

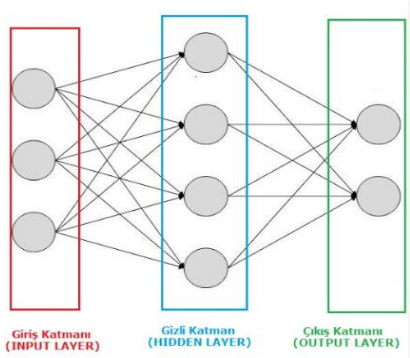
2.1. KNIME Platformu

KNIME, düğümler arasındaki ilişkiler sayesinde verinin işlenerek raporlanmasını sağlayan bir veri analiz platformudur. Aynı isme sahip bir firma tarafından açık kaynak olarak geliştirilmiştir. Mevcut sabit disk alanıyla sınırlı olan büyük veri süreçlerinde kullanıma uygundur. Yazılımcılar tarafından geliştirilen ek özellikleri sisteme eklemeyi sağlayan uzantı mekanizmasına sahiptir. Genel olarak iş zekası sürecindeki veri analizi uygulamalarında kullanılmaktadır. Makine öğrenmesi ve veri madenciliğine yönelik çeşitli bileşenlere sahiptir ve bu bileşenler uygulama içerisinde düğüm olarak ifade edilir. Düğümler aracılığıyla kod yazmadan işlemler gerçekleştirilir. İlişkilendirilen düğümler akış sırasına göre çalıştırılır. Ayrıca her düğüme ait çıktı tek tek görüntülenebilmektedir. Çok fazla sayıda düğüm ile fonksiyonel içerik sağlar. Görsel arayüzü sayesinde veri akışının tasarlanmasını ve çeşitli algoritmaların çalıştırılmasını sağlar. Veri analitiği ile ilgili araştırmalarda çok yaygın bir şekilde kullanılmaktadır. Güçlü fonksiyonları ve açık sistem yapısı ile giderek yaygınlaşmaktadır [10,11].

2.2. Yapay Sinir Ağları

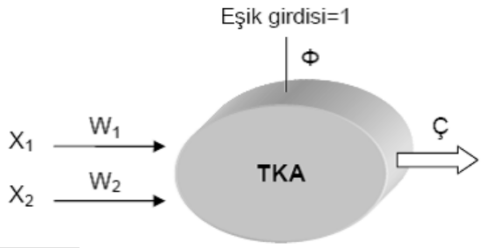
Yapay sinir ağları, öğrenme ile yeni bilgiler türetebilme ve oluşturabilme gibi işlemleri gerçekleştirebilmek amacıyla geliştirilen bilgisayar sistemleridir. Yapay sinir ağları insan beyni örnek alınarak, öğrenme sürecinin matematiksel olarak modellenmesiyle oluşturulmuştur. Beyindeki biyolojik sinir ağlarının yapısını ve yeteneklerini taklit etmektedir. Giriş katmanı, gizli katmanlar ve çıkış katmanı olmak üzere üç katmadan oluşmaktadır. Şekil 1'de Yapay sinir ağlarının katman yapısı gösterilmiştir. Bilgiler ağa girdi katmanından iletilir, ara katmanlarda işlenir ve çıktı katmanına

gönderilirler. Bilgi işleme, ağa gelen bilgilerin ağın ağırlık değerleri kullanılarak çıktıya dönüştürülmesidir. Ağırlıkların doğru değerleri verildiğinde, girdiler için doğru çıktılar üretilmektedir. Birçok nöron ve gizli katmandan oluşan sinir ağlarına çok katmanlı sinir ağları ve tek bir katmandan oluşan sinir ağlarına tek katmanlı sinir ağları denir [12].



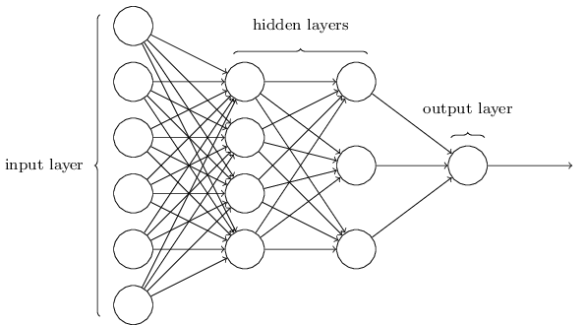
Şekil 1. Yapay sinir ağlarının katman yapısı [12].

Tek katmanlı Yapay sinir ağları, girdi ve çıktı katmanlarından oluşmaktadır. Doğrusal olmayan problemlerde kullanılmamaktadır. Birden fazla girdi değerlerine sahip olabilmektedir. Girdi değerleri ile ağırlık değerleri çarpılarak çıktı değeri hesaplanır. Eşik değeri çıktının 0 olmasını engellediği için çıktı her zaman 1 değerini almaktadır. Tek katmanlı Yapay sinir ağı modelinin yapısı Şekil 2'de verilmiştir [13].



Şekil 2. Tek katmanlı Yapay sinir ağı modeli [13].

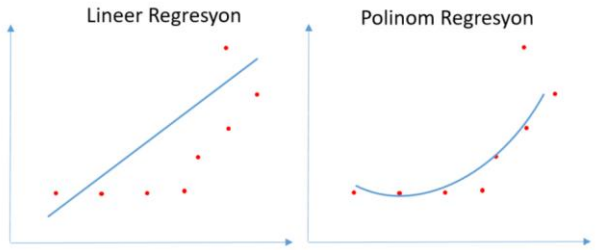
Çok katmanlı Yapay sinir ağları doğrusal olmayan problemlerin çözümünde kullanılmaktadır. Birden fazla girdi ve gizli katmana sahip yapılardır. Gizli katmanların sayısı problemin akışına göre değiştirilmektedir. Gizli katman problemin yapısına göre farklı fonksiyonlar ile işlenip çıktı katmanına aktarılmasını sağlamaktadır. Çok katmanlı Yapay Sinir ağı modelinin yapısı Şekil 3'de verilmiştir [13].



Şekil 3. Çok katmanlı Yapay sinir ağı modeli [13].

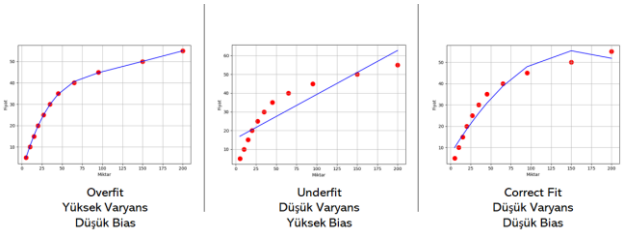
2.3. Polinomsal Regresyon

Doğrusal bir regresyonda girdi ve çıktı arasında doğrusal bir ilişki vardır. Polinomsal regresyonda ise aralarındaki ilişki eğri biçimindedir. Polinomsal regresyon bu eğrinin fonksiyonunu vermektedir. Polinomsal fonksiyonunun hangi dereceyi alacağı önemlidir. Basit ve çoklu lineer regresyonda bağımlı ve bağımsız değişkenler arasındaki ilişki doğrusaldır. Gerçek hayattaki doğrusal olmayan ilişkileri modellemek için Polinomsal regresyon kullanılır. Şekil 4'te Doğrusal ve Polinomsal regresyonun gösterimi verilmiştir [14].



Şekil 4. Doğrusal ve Polinomsal regresyonun grafiksel gösterimi [14].

Bias dengelenmesi, modelin verileri sığdırmadaki basit varsayımlarından kaynaklanan hatadır. Bias yüksek olduğunda, modelin verilerdeki desenleri yakalayamadığı anlamına gelir ve düşük öğrenme oluşur. Varyans dengelenmesi, verileri sığdırmaya çalışan karmaşık modelden kaynaklanan hatadır. Yüksek varyans, modelin veri noktalarının çoğundan geçtiği ve aşırı öğrenmeye neden olduğu anlamına gelir. Şekil 5'te Bias ve Varyans dengelenmesi grafik ile açıklanmıştır [15].



Şekil 5. Bias ve Varyans dengelenmesinin grafik ile açıklanması [15].

Polinomsal regresyon geniş bir eğrilik alanına uymaktadır. Bağımlı ve bağımsız değişken arasındaki ilişkiye en iyi yaklaşımı verir. Ancak veride aykırı değer bulunması, doğrusal olmayan bir analizin sonuçlarını önemli şekilde etkiler. Sınırları belirli olan veri setleri için uygun bir algoritmadır. Ancak veri setinin sınırlarının dışından gelen yeni veriler için yüksek hataya sahip tahminlerde bulunabilir [15].

2.4. Model Performansını Değerlendirme

Makine öğrenmesi ile ilgili çalışmalarda model performansını değerlendirmek için kullanılabilecek çeşitli metrikler vardır. R Kare, modeldeki bağımsız değişkenlere göre bağımlı değişkenin varyasyon oranını ölçer. Regresyon modelinin çok fazla bağımsız değişkeni varsa test verilerinde istenilen başarıyı göstermez. Bu durumlarda Düzeltilmiş R Kare kullanılır. Modele eklenen ek bağımsız değişkenleri cezalandırır ve aşırı uyum sorununu çözer. MSE (Mean Squared Error), gerçek değerler ile tahmin edilen değerler arasındaki hataların karesinin ortalaması ile tanımlanır. Root Mean Squared Error olan RMSE ise, MSE'nin karekökü'dür. Eğer büyük değerlerdeki yanlış hataların cezalandırılması istenirse MSE, problemde açıklanabilirlik önemliyse RMSE kullanılır. MAE (Mean Absolute Error), gerçek değerler ile tahmin edilen değerler arasındaki hataların mutlak değerlerinin ortalaması ile tanımlanır. MAPE (Mean Absolute Percentage Error), hataların mutlak değerlerinin gerçek değerlere oranının ortancası ile tanımlanır. Bu metriklerin amacı en küçük değere sahip sonuca ulaşmaktır. Bölüm 2.4.1 ve 2.4.5 arasında kullanılan metriklerin matematiksel formülleri ve açıklamaları verilmiştir. Formüllerde bulunan ifadelerde ti, i.birimin tahmin değerini, gi, i.birimin gerçek değerini, ej, tahmindeki hata değerini ifade etmektedir [16,17].

2.4.1. MSE (Mean Squared Error)

Hata değerlerinin büyüklükleri benzer olduğu durumlarda tercih edilmektedir. Sonuç pozitif değerli çıkmaktadır. Sonucun 0'a yakın olması modelin performansının iyi olduğunu göstermektedir. Denklem 1'de fonksiyonun hesaplanma şekli gösterilmektedir [17].

$$HKO = \frac{1}{n}\sum_{i=1}^{n}(t_i - g_i)^2 = \frac{1}{n}\sum_{i=1}^{n} {e_j}^2 \qquad (1)$$

2.4.2. RMSE (Root Mean Squared Error)

Tahmin hatalarındaki standart sapmadır. Tahmin hatalarının yayılma durumunu ölçmektedir. Model hatasız olduğunda sonuç 0 çıkmaktadır. Denklem 2'de bu metrik için kullanılan denklem gösterilmiştir [17].

$$HKOK = \sqrt{\frac{1}{n}\sqrt{\sum_{i=1}^{n}(t_i - g_i)^2}} = \sqrt{\frac{1}{n}\sqrt{\sum_{i=1}^{n} {e_j}^2}} \qquad (2)$$

2.4.3. MAPE (Mean Absolute Percentage Error)

Hata değerlerinin birimlerinde farklılık olduğunda kullanılan bir fonksiyondur. Yüzde olarak ifade edilir. Fonksiyonun %10'un altında çıkması yüksek doğruluğa sahip model, %10 ile %20 arasında çıkması doğru tahmin modeli olduğunu göstermektedir. Denklem 3'te fonksiyonun hesaplanma şekli gösterilmektedir [17].

$$OMHY = 100X\frac{\sum_{i=1}^{n}(t_i - g_i)^2/t_i}{n} \qquad (3)$$

2.4.4. MAE (Mean Absolute Error)

Denklem 4'te yer alan denkleme göre hesaplanan bu hata fonksiyonu, tahmin ve gerçek değerler arasındaki farkın ölçüsüdür. Genellikle regresyon ve zaman serisi problemlerinde kullanılmaktadır [17].

$$OMH = \frac{\sum_{i=1}^{n}|t_i - g_i|}{n} = \frac{\sum_{i=1}^{n} e_j}{n} \qquad (4)$$

2.4.5. R Kare

Gerçek değerler ile tahmin edilen değerler arasındaki ilişkiyi ifade eder. Eksi sonsuz ile 1 arasında değer alır. Sonuç 1'e ne kadar yakınsa model o kadar hassas ve uyum iyiliği uygun demektir [18].

$$R\ Kare = \frac{\sum(x - \overline{x})(y - \overline{y})}{\sum(x - \bar{x})^2 \sum(y - \overline{y})^2} \qquad (5)$$

2.5. Veri Seti

Bu çalışmada ev satışlarında kullanılan çeşitli bilgiler kullanılarak bir evin potansiyel fiyatını tahmin etmeye yönelik bir uygulama yapılmıştır. Geliştirilen bu uygulama ile bir tür karar destek modeli oluşturulmuştur. Bunun için ev fiyatlarının ve çeşitli özniteliklerin olduğu King County House veri seti kullanılmıştır. Veri seti Kaggle web sitesinde "House Sales in King County, USA" başlığıyla indirilebilir durumdadır [19]. Veri seti içerisinde 17384 adet veri ve 21 tane öznitelik bulunmaktadır. Tablo 1'de veri setinin bir kısmı gösterilmiştir. Çalışmada kullanılan veri seti, Mayıs 2014 ve Mayıs 2015 döneminde Washington's King County'deki evlerin fiyatı, boyutu, konumu gibi diğer çeşitli özellikleri hakkında bilgiler içerir. Kullanılan değişkenlerin açıklamaları Tablo 2'de ve değişkenlerin veri türleri Tablo 3'de verilmiştir [20,21].

Tablo 1. Çalışma için kullanılan veri seti

id	date	price	bedrooms
7129300520	20141013T000000	221900	3
6414100192	20141209T000000	538000	3
5631500400	20150225T000000	180000	2
2487200875	20141209T000000	604000	4
1954400510	20150218T000000	510000	3
7237550310	20140512T000000	1225000	4
1321400060	20140627T000000	257500	3
2008000270	20150115T000000	291850	3

Tablo 2. Veri setindeki alanlar ve açıklamaları

Veri Setindeki Alanlar	Açıklamaları
id	Ev için tanımlanan benzersiz numara
date	Evin satılma tarihi
price	Ev için hedeflenen fiyat tahmini
bedrooms	Evdeki yatak odası sayısı
bathrooms	Evdeki banyo sayısı
sqft_living	Evin metrekaresi
sqft_lot	Arsanın metrekaresi
floors	Evdeki katlar
waterfront	Deniz kıyısına bakan ev
view	Evden görülen manzara
condition	Evin genel durumu
grade	King County derecelendirme sistemine göre konut birimine verilen genel not
sqft_above	Bodrum dışında evin metrekaresi
sqft_basement	Bodrumun metrekaresi
yr_built	Evin inşa yılı
yr_renovated	Evin yenilendiği yıl
zipcode	Posta kodu
lat	Enlem koordinatı
long	Boylam koordinatı
sqft_living15	En yakın 15 komşu için iç konut yaşam alanı metrekaresi
sqft_lot15	En yakın 15 komşunun arsalarının metrekaresi

Tablo 3. Veri setindeki alanlar ve veri türleri

Veri Setindeki Alanlar	Veri Türleri
id	int64
date	object
price	float64
bedrooms	int64
bathrooms	float64
sqft_living	int64
sqft_lot	int64
floors	float64
waterfront	float64
view	float64
condition	int64
grade	int64
sqft_above	int64
sqft_basement	object
yr_built	int64
yr_renovated	float64
zipcode	int64
lat	float64
long	float64
sqft_living15	int64
sqft_lot15	int64

3. Bulgular

Öncelikle kullanılacak veri KNIME ortamına aktarılmıştır. "File Reader" düğümü ile veri okunduktan sonra modele verilecek değişkenler "Column Filter" düğümü ile seçilir. Evin satış tarihi gibi ilk etapta kullanılmayacak değişkenler filtre yardımıyla çıkartılır. Çalışmada kullanılacak olan Yapay sinir ağları ve Polinomsal regresyon yöntemleri için verinin normalize edilmesi gerekir. "Normalizer" düğümü yardımıyla veri seti normalleştirilir. Bunu takiben, kullanılacak veri seti, eğitim ve test için "Partitioning" düğümü yardımıyla ikiye ayrılır. Eğitim kümesi, modelin hedef değişkendeki bilgiyi ne kadar iyi açıkladığı konusunda bilgi verirken, test kümesi ise daha önce görülmemiş gözlemler verildiğinde modelin ne kadar iyi performans göstereceğini anlatır. Modelleme aşamasında Yapay sinir ağları ile öğrenme ve tahmin düğümleri modele eklenir. Performanslarının ölçülmesi için "Score" düğümü eklenir. Fiyat tahmini yapılacağı için sürekli bir çıktı vardır, bu yüzden Numeric Scorer düğümü kullanılır. Parametre optimizasyonundan gelen en iyi parametre değerleri verilerek RProp MLP (Multi Layer Perceptron) Learner ile ev fiyatları tahmin edilir. Tablo 4'te verildiği gibi test seti için R Kare değerine bakıldığında yaklaşık %72'lik bir performans değeri elde edilmiştir. Hatalar düşüktür. Tablo 5'te verildiği gibi eğitim seti için R Kare değerine bakıldığında yaklaşık %74'lük bir performans değeri elde edilmiştir. Overfitting (Aşırı öğrenme) olmaması için modelin test ve eğitim kümesindeki hatalar arasındaki fark düşük olmalıdır. Burada eğitim ve test performansı birbirine yakın olduğu için aşırı öğrenme olmamıştır.

Tablo 4. Yapay sinir ağları ile test performansının ölçülmesi

Metrikler	Sonuçlar
R^2	0,727
Mean Absolute Error	0,017
Mean Squared Error	0,001
Root Mean Squared Error	0,026
Mean Signed Difference	0
Mean Absolute Percentage Error	0,375
Adjusted R^2	0,727

Tablo 5. Yapay sinir ağları ile eğitim performansının ölçülmesi

Metrikler	Sonuçlar
R^2	0,745
Mean Absolute Error	0,016
Mean Squared Error	0,001
Root Mean Squared Error	0,024
Mean Signed Difference	-0
Mean Absolute Percentage Error	NaN
Adjusted R^2	0,745

Model öğrendikten sonra veri tekrar incelendiğinde, waterfront değişkeninin 0 ve 1 değerlerine, view değişkeninin 0 ile 4 arasında 5 farklı değere ve condition değişkeninin 1 ile 5 arasında 5 farklı değere sahip olduğu görülmüştür. Buradaki değerler sürekli gözükmesine rağmen evet-hayır anlamında kategorik değişkenlerdir. Buradaki 1, 0'dan üstün değildir. Bunların evet-hayır şeklinde ikili değişkenlere dönüştürülmesi gerekir. Çünkü bir makine öğrenmesi modeli sayısal değerlerde büyük olan değerin daha önemli olduğuna göre hesaplamalar yapar. Bu şekilde modele katılması sonuçların hatalı olmasını sağlar. Mevcut değişkenleri ikili değişkene dönüştürmek için One to Many düğümü kullanılır. One to Many düğümü integer tipindeki değişkenlerle çalıştığı için NumberTo String düğümü ile rakam olan değişkenlerin tipleri değiştirilir. Daha önce yapılan modelleme kısmının aynısı yeni işlenmiş veri için yapılır. Tablo 6'da verildiği gibi test seti için R Kare değerine bakıldığında yaklaşık %70'lik bir performans değeri elde edilmiştir. Tablo 7'de verildiği gibi eğitim seti için R Kare değerine bakıldığında yaklaşık %75'lik bir performans değeri elde edilmiştir. Yine aynı şekilde eğitim ve test performansı birbirine yakın olduğu için aşırı öğrenme olmamıştır.

Tablo 6. Binary değişkenlere dönüştürüldükten sonra Yapay sinir ağları ile test performansının ölçülmesi

Metrikler	Sonuçlar
R^2	0,707
Mean Absolute Error	0,017
Mean Squared Error	0,001
Root Mean Squared Error	0,027
Mean Signed Difference	-0
Mean Absolute Percentage Error	0,387
Adjusted R^2	0,707

Tablo 7. Binary değişkenlere dönüştürüldükten sonra Yapay sinir ağları ile eğitim performansının ölçülmesi

Metrikler	Sonuçlar
R^2	0,753
Mean Absolute Error	0,016
Mean Squared Error	0,001
Root Mean Squared Error	0,024
Mean Signed Difference	-0
Mean Absolute Percentage Error	NaN
Adjusted R^2	0,753

Kıyaslama yapmak için dönüştürülen veri kullanılarak Polinomsal regresyon kullanılmıştır. İlk olarak 2.dereceden bir Polinom tasarlanmıştır. Performans durumuna göre Polinom derecesi yükseltilebilir. Ancak Polinomsal derecesini çok fazla yükseltmek aşırı öğrenmeye yol açar. Tablo 8'de verildiği gibi 2.dereceden bir Polinomsal performansı R Kare değeri ile ölçüldüğünde yaklaşık %67'lik bir performans değeri elde edilmiştir. Tablo 9'da verildiği gibi 3.dereceden bir Polinomsal performansı R Kare değeri ile ölçüldüğünde yaklaşık %62'lik bir performans değeri elde edilmiştir.

Tablo 8. İkinci dereceden bir Polinomsal regresyon ile performansın ölçülmesi

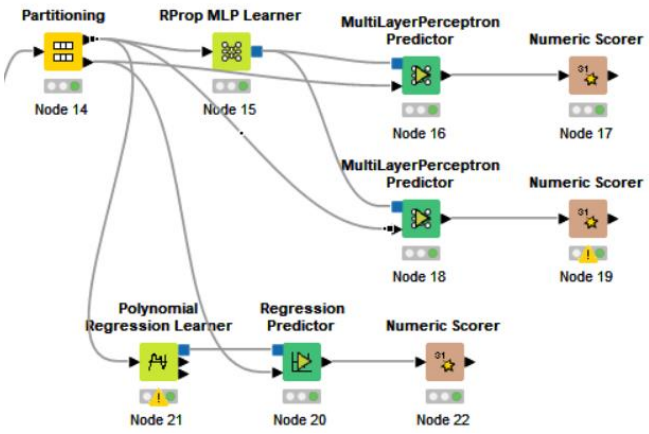
Metrikler	Sonuçlar
R^2	0,673
Mean Absolute Error	0,019
Mean Squared Error	0,001
Root Mean Squared Error	0,029
Mean Signed Difference	0,004
Mean Absolute Percentage Error	0,451
Adjusted R^2	0,673

Tablo 9. Üçüncü dereceden bir Polinomsal regresyon ile performansın ölçülmesi

Metrikler	Sonuçlar
R^2	0,629
Mean Absolute Error	0,018
Mean Squared Error	0,001
Root Mean Squared Error	0,031
Mean Signed Difference	-0,005
Mean Absolute Percentage Error	0,368
Adjusted R^2	0,629

Binary değişkenlere dönüştürülen Yapay sinir ağları ile ikinci dereceden bir Polinomsal regresyon için sonuçlara bakıldığında MAE, RMSE ve MAPE hata değerlerinin Yapay sinir ağları için daha düşük, R Kare ve Düzeltilmiş R Kare değerinin Yapay sinir ağları için daha büyük olduğu görülmüştür. Bu nedenle Yapay sinir ağları yönteminin Polinomsal regresyon yöntemine göre ev fiyatlarını daha yüksek doğrulukla doğru olarak tahmin ettiği görülmüştür.

Elde edilen sonuçlarda R Kare ve Düzeltilmiş R Kare değerlerinin hep aynı olduğu görülmüştür. Düzeltilmiş R Kare modele eklenen ek bağımsız değişkenleri cezalandırdığı için daha küçük olma eğilimindedir, ancak bu değerler genellikle birbirine yakın olmaktadır. Bu iki değerin arasının açılması gereksiz değişken kullanıldığı anlamına gelmektedir. Değerlerin eşit çıkması fazla değişken kullanılmadığı şeklinde yorumlanabilir. İki farklı yöntemin KNIME üzerinde iş akışı olarak modellenmesinin bir kısmı Şekil 6'da verilmiştir.



Şekil 6. KNIME üzerinde yapılan çalışma

4. Tartışma ve Sonuç

Bu çalışmada bir evin potansiyel fiyatını tahmin etmek için iki farklı yöntemin performansı incelenmiştir. Makine öğrenmesi yöntemleri olarak Yapay sinir ağları ve Polinomsal regresyon kullanılmıştır. Çalışma, KNIME Analytics Platform veri madenciliği programının 4.6.0 sürümünde uygulanmıştır. İncelenen yöntemlerin başarısını ölçmek için R Kare performans metriği kullanılmıştır. Çalışmada Yapay sinir ağları yönteminin Polinomsal regresyon yöntemine göre ev fiyatlarını daha yüksek doğrulukla doğru olarak tahmin ettiği görülmüştür. Aynı veri seti kullanılarak yapılan başka bir çalışmada, en uygun kümelenmiş modeli oluşturmak için birkaç alt model birleştirilerek stacked bir regresyon modeli oluşturulmuştur. Stacked model sonuçları Tablo 10 ve Tablo 11'de verilmiştir. Test ve eğitim setindeki sonuçların benzer olması, modellerin hiçbirinin verilere overfitting yapmadığını göstermektedir. Tablo 10'da gösterilen R Kare değerinin yüksek olması, modelin tahminleri ile gerçek ev fiyatları arasında güçlü bir korelasyon olduğunu göstermektedir. Model eğitim veri seti için günlük ev fiyatı varyasyonunun %88'ini açıklayabilmiştir. Stacked model performansı ile Binary değişkenlere dönüştürülen Yapay sinir ağları ve ikinci dereceden bir Polinomsal regresyon için sonuçlara bakıldığında, Stacked modelin R Kare değerinin daha yüksek, MAE ve RMSE değerlerinin daha düşük olduğu görülmüştür [22].

Tablo 10. Stacked model ile test performansının ölçülmesi [20].

Metrikler	Sonuçlar
R^2	0,87
Mean Absolute Error	0,14
Root Mean Squared Error	0,19

Tablo 11. Stacked model ile eğitim performansının ölçülmesi [20].

Metrikler	Sonuçlar
R^2	0,88
Mean Absolute Error	0,14
Root Mean Squared Error	0,18

Yapılan çalışmanın ev değerlendirilmesi için kullanılan uygulamaların geliştirilmesine ve bu alanda yapılan bilimsel çalışmalara katkı sağlayacağı düşünülmektedir. Çalışmada kullanılan veri setinin birçok sınırlaması vardır. En büyük sınırlama veri setinin potansiyel alıcılar ve satış ortamı hakkında herhangi bir bilgi içermemesidir. Ayrıca veriler bir yıllık bir süre boyunca toplanmıştır, bu nedenle mevsimselliği fazla yakalamaz ve ekonomik faktörleri dikkate almaz. Bu nedenle daha sonra yapılacak olan çalışmalarda ev özellikleri ve alıcıyla ilgili verilerin detaylı incelenmesi hedeflenmektedir.

Etik Beyanı

Bu çalışmada, "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" kapsamında uyulması gerekli tüm kurallara uyulduğunu, bahsi geçen yönergenin "Bilimsel Araştırma ve Yayın Etiğine Aykırı Eylemler" başlığı altında belirtilen eylemlerden hiçbirinin gerçekleştirilmediğini taahhüt ederiz.

Kaynakça

[1] Akay, E.Ç., Topal, K.H., Kizilarslan, S., Bulbul, H. 2019. Türkiye Konut Fiyat Endeksi Öngörüsü: Arıma, Rassal Orman ve Arıma-Rassal Orman. Istanbul Finance Congress, 10, 7-11.

[2] Oral, M., Okatan, E., Kırbaş, İ. 2021. Makine Öğrenme Yöntemleri Kullanarak Konut Fiyat Tahmini Üzerine Bir Çalışma: Madrid Örneği. 3 rd International Young Researchers Student Congress, 263-272.

[3] Aydemir, E., Aktürk, C., Yalçınkaya, M.A. 2020. Yapay Zekâ ile Konut Fiyatlarının Tahmin Edilmesi. Turkish Studies – Information Technologies and Applied Sciences, 15(2), 183-194.

[4] Gülağız, F.K., Ekinci, E. 2017. Farklı Regresyon Analizi Yöntemleri Kullanılarak Ev Fiyatlarının Tahmini. Conference: International Symposium on Industry 4.0 and Applications, 203-207.

[5] Yılmazel, Ö., Afşar, A., Yılmazel, S. 2018. Konut Fiyat Tahmininde Yapay Sinir Ağları Yönteminin Kullanılması. UİİD-IJEAS, 20, 285-300.

[6] Hadavandi, E., Ghanbari, A., Mirjani, S.M., Abbasian, S. 2011. An Econometric Panel Data-Based Approach for Housing Price Forecasting in Iran. International Journal of Housing Markets and Analysis, 4(2), 70-83.

[7] Phan, D. 2018. Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), 35-42.

[8] Park, B., Bae, J.K. 2015. Using Machine Learning Algorithms for Housing Price Prediction: The Case of Fairfax County, Virginia Housing Data. Expert Systems with Applications, 42(19), 2928-2934.

[9] Teoh, E.Z., Yau, W.C., Ong, T.S., Connie, T. 2022. Explainable Housing Price Prediction with Determinant Analysis. International Journal of Housing Markets and Analysis, 15(5), 1-25.

[10] Aksan, C.E. 2022. KNIME Nedir ?. https://ceaksan.com/tr/knime-nedir (Erişim Tarihi: 10.06.2022).

[11] Kobaner, C. 2022. KNIME. https://sistek.com.tr/tr/knime-veri-analitigi-platformu-cozum-ortagi/ (Erişim Tarihi: 11.06.2022).

[12] Yıldırım, E. 2020. Yapay Sinir Ağı (Artificial Neural Network) Nedir?. https://www.veribilimiokulu.com/yapay-sinir-agiartificial-neural-network-nedir/ (Erişim Tarihi: 28.11.2022).

[13] GTech. Yapay Sinir Ağları ve Uygulamaları-1. https://www.gtech.com.tr/yapay-sinir-aglari-ve-uygulamalari-1/ (Erişim Tarihi: 28.11.2022).

[14] Şirin, E. 2022. Polinomsal Regresyon: Python ile Uygulama-1. https://www.veribilimiokulu.com/Polinomsal-regresyon-python-uygulama-1/ (Erişim Tarihi: 10.06.2022).

[15] Şener, Y. 2022. Polinomsalsal (Polynomial) Regresyon ve Python Uygulaması. https://yigitsener.medium.com/Polinomsalsal-polynomial-regresyon-ve-python-uygulamas%C4%B1-f742fb61a158 (Erişim Tarihi: 18.06.2022).

[16] Köseoğlu, B. 2022. Model Performansını Değerlendirmek: Regresyon. https://medium.com/yaz%C4%B1l%C4%B1m-ve-bili%C5%9Fim-kul%C3%BCb%C3%BC/model-performans%C4%B1n%C4%B1-de%C4%9Ferlendirmek-regresyon-48b4afec8664 (Erişim Tarihi: 18.06.2022).

[17] Özcan, E. 2007. Kükürt Giderme İşlemi için Kullanılan Malzeme Miktarının Makine Öğrenme Yöntemleri ile Tahmini. Karabük Üniversitesi, Lisansüstü Eğitim Enstitüsü, Yüksek Lisans Tezi, 61s, Karabük.

[18] Sevinç, A., Kaya, B. 2021. Derin Öğrenme Yöntemleri ile Sıcaklık Tahmini: Diyarbakır İli Örneği. Journal of Computer Science, 217-225.

[19] Kaggle. 2022. House Sales in King County, USA veri seti. https://www.kaggle.com/datasets/harlfoxem/housesalesprediction (Erişim Tarihi: 14.06.2022).

[20] Alasmary, F. 2022. linear-regression-numpy. https://github.com/farisalasmary/linear-regression-numpy/blob/master/kc_house_train_data.csv (Erişim Tarihi: 14.06.2022).

[21] Fisher, A. 2022. Predicting King County House Prices with Multiple Linear Regression. https://medium.com/analytics-vidhya/predicting-king-county-house-prices-with-multiple-linear-regression-84de5feeafb2 (Erişim Tarihi: 19.06.2022).

[22] Farrell, S. 2018. Comparison of Data Mining Models to Predict House Prices. ACADEMIA, 1-9.