KAHRAMANMARAŞ SÜTÇÜ İMAM ÜNİVERSİTESİ
MÜHENDİSLİK VE MİMARLIK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ
BİTİRME PROJESİ TEZİ

KONUT (EV) FİYAT TAHMİNİ

RAMAZAN ÖZER , 18110131027

Dr.Öğr.Üyesi ZEYNEP BANU ÖZGER

HAZİRAN 2022

KONUT (EV) FİYAT TAHMİNİ

Adı SOYADI

KAHRAMANMARAŞ SÜTÇÜ İMAM ÜNİVERSİTESİ
MÜHENDİSLİK VE MİMARLIK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ
Haziran 2022

ÖZET

Değişkenler arasındaki ilişkileri tespit etmek için kullanılan regresyon yöntemleri ve regresyon tabanlı yapay sinir ağı, pek çok farklı alanda uygulanmaktadır. Fiyat tahminine ihtiyaç duyulan uygulamalar bu yöntemlerin kullanıldığı alanlardandır. Bu alanda kullanılacak verilerin içeriğine uygun olarak en iyi sonucu verecek yöntemin tespit edilmesi gerekmektedir. Bu çalışmada konut fiyatlarının belirli özelliklere bağlı olarak, regresyon analizi yöntemleri ve regresyon tabanlı yapay sinir ağı aracılığıyla tahmin edilmesi sağlanmıştır. Bu amaçla doğrusal regresyon analizi, karar ağacı regresyonu ve rastgele orman regresyon ,ridge regresyon ,lasso regresyon , XGBoost regresyon ve regresyon tabanlı yapay sinir ağı yöntemleri gerçek veriler üzerinde test edilmiştir. Yapılan çalışma sonucunda XGBoost regresyon yönteminin konut fiyatlarının tahmini için modellemeyi diğer yöntemlere göre daha doğru olarak gerçekleştirdiği tespit edilmiştir.

Anahtar Kelimeler : Doğrusal regresyon , karar ağacı regresyonu ve rastgele orman regresyon ,ridge regresyon ,lasso regresyon , XGBoost regresyon , regresyon tabanlı yapay sinir ağı, fiyat tahmini

Sayfa Adedi : 41

Danışman : Dr.Öğr.Üyesi ZEYNEP BANU ÖZGER

TEŞEKKÜR

Bu çalışmam yürütülmesi sırasında her hafta projem ile ilgili sorularımı cevaplayan ve projem ile ilgili yol gösteren ve yardımını esirgemeyen danışmanım Kahramanmaraş Sütçü İmam Üniversitesinden (Dr.Öğr.Üyesi ZEYNEP BANU ÖZGER) çok teşekkür ederim.

İçindekiler

SİMGELER VE KISALTMALAR

Bu çalışmamda kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklamalar
KNN	K-en yakın komşuluk algoritması
MICE	Zincirli Denklemlerle Çok Değişkenli Imputation
KFT	Konut fiyat tahmin

1.GİRİŞ

Konut fiyat tahmini (KFT) projemde evlerin özellikleri ile fiyatları arasındaki ilişkiyi bulmak. Ve bu ilişki ile yeni evlerin fiyatlarını tahmin edebilmeyi öğrenen bir model tasarlamaktır.

1.1.Problemin tanımı / Konunun tanımı

Evlerin çeşitli özellikleri bilinen ve ev fiyatları belirlemede evlerin hangi özelliklerin etken olduğunun belirlenmesidir.

1.2.Tez ile Amaçlanan Nedir ?

KFT ile amaçlanan değişkenler arasındaki bağlantıları doğru bir model seçerek ev fiyatlarını tahmin etmeye çalışmaktır.

1.3.Tezin Önemi Nedir ?

Ev almak isteyen kişilerin evlerin çeşitli kriterlerine göre ne kadar fiyat verilmesini gerektiğini karşılaştırmak ve kendileri için en uygun evi , en optimum fiyatlara bulmaları için yapılan çeşitli regresyon yöntemleri ile en uygun şekilde evlerin fiyatlarını belirleyen bir model tasarlamaktır.

1.4.Tezin Kapsamı Nedir ?

Verilen verideki konutların özellikleri ile evlerin fiyatlarını belirlemektir. Bu belirleme işlemi içinde kullanılan yöntem yapay zeka ile modellemektir.

1.5.Tezin Özgün Değeri Nedir ?

Projemde çeşitli şekillerde ev tahmini için farklı yapay zeka yöntemler uygulanmıştır. Yöntemleri kullanan kişilerin veriyi tam kavramamasından gelen işlemler olduğunu ve veri ön işlemede çeşitli hataların olduğunu gözlemlenmiştir. Gözlemlediğim veri ön işlemede eksik verilerin silinmesi veya optimum değişkenler yerine veriyle uyuşmayan değerlerin atanmasıyla oluşan hatalardır. Kategorik değerlerinde ortalaması veya medyanın alınması gibi ,tarih kolonlarını içeren verilerin ayrıştırılması işlemleri uygulanmıştır.

2.LİTERATÜR ÖZETİ

Konut fiyatlarını belirlemeye yönelik olarak gerçekleştirilmiş pek çok çalışma bulunmaktadır.

Fan ve arkadaşları tarafından konut özellikleri ve fiyatları arasındaki ilişkiyi belirlemek için yeni bir karar ağacı regresyon modeli önerilmiştir [1].

Ecer tarafından Türkiye'deki konut fiyatlarını tahmin etmek için Hedonik Regresyon yöntemi ile yapay sinir ağları karşılaştırılmıştır. Çalışma sonucunda yapay sinir ağları kullanılarak geliştirilen modelin daha başarılı olduğu gösterilmiştir [2].

Özsoy ve Şahin tarafından İstanbul'daki konut fiyatları üzerinde etkili olan özellikleri tespit etmek amacıyla bir çalışma yapılmıştır. Analiz için sınıflandırma ve karar ağacı regresyon analizi yöntemleri kullanılmıştır [3].

Goldberg ve Harding tarafından yapılan çalışmada konut kredilerinin konut özellikleri ile ilişkileri belirlenmiştir [4].

Bin tarafından yarı parametrik regresyon yöntemi kullanılarak konut fiyatlarını tahmin etmeye yönelik bir çalışma yapılmıştır. Çalışma sonucunda yarı parametrik modellerin parametrik modellere göre %10-20 oranında daha doğru tahmin yaptığı belirtilmiştir [5].

Fitöz ve Öztürk tarafından yapılan başka bir çalışmada konut piyasasındaki belirleyici etmenleri tespit etmek amacıyla doğrusal regresyon yöntemi kullanılmıştır [6].

Ankara'daki konut fiyatlarının tahmini için Tuna ve arkadaşları tarafından bir çalışma yapılmıştır. Yapılan çalışma ile hem Ankara'daki konutların ortalama metrekare birim fiyatı hem de özellik bazında fiyat artışları tespit edilmiştir [7].

Çalışmada veri ön işleme kısmında kategorik verileri sayısallaştırma işlemi uygulanmıştır.

Eksik verileri için 3 farklı yöntemle eksik veri doldurma işlemi gerçekleşmiştir. Knn algoritması , Mice algoritması ve ortalama ile doldurma işlemi dahil ettim . Rastgele orman regresyonu ile modelimi başarı oranına baktım.

3.MATERYAL VE METOTLAR

3.1.Materyal

Bitirme projemi geliştirdiğim programlama dili Python ‘dur. Ayrıca Python’da kullandığım kütüphaneler şunlardır;

Numpy: Çok boyutlu dizileri ve matrisleri destekleyen, bu diziler üzerinde çalışacak üst düzey matematiksel işlevler ekleyen bir kitaplıktır.

Pandas: “ilişkisel” ve “etiketli” verilerle çalışmayı kolay ve sezgisel hale getirmek için tasarlanmış hızlı, esnek ve etkileyici veri yapıları sağlayan bir paketdir.

Matlotlib.pyplot: Veri görselleştirmesinde kullandığımız temel kütüphanedir.

Seaborn: Python’da ilgi çekici ve bilgilendirici istatistiksel grafikler yapmak için kullanılan bir kütüphanedir.

Csv: CSV dosyaları üzerinde işlem yapmak için kullanılan kütüphanedir.

Sklearn: Destek vektör makinesi, gradyan artırma,k-means,rassal ormanlar gibi algoritmaları içeren kütüphanedir.

Statsmodels: İstatistiksel modeller tahmini, istatistiksel testler yapma, vb. gibi istatistiksel veri analizi için birçok fırsat sağlayan bir Python modülüdür. Yardımı sayesinde birçok makine öğrenme yöntemini uygulayabilir ve farklı çizim olanaklarını keşfedebilirsiniz.

TensorFlow: Makine öğrenimi için ücretsiz ve açık kaynaklı bir yazılım kütüphanesidir . Bir dizi görevde kullanılabilir, ancak derin sinir ağlarının eğitimi ve çıkarımına özel olarak odaklanmaktadır.

3.2.Metot

Veri kümesinden aykırı verilerin silinme işlemi ve kategorik verileri sayısallaştırma işlemi yapılmıştır. Kayıp verilerin doldurma işlemleri 3 farklı yöntemle gerçekleştikten sonra normalizasyon işlemi yapılmıştır. Model eğitimi için çeşitli regresyon işlemleri ve regresyon tabanlı yapay sinir ağı kullanılmıştır.

3.2.1 Verileri Sayısallaştırma

Regresyon algoritmaları sayısal veriye ihtiyaç duymaktadır. Bu nedenle kategorik veriler sayısallaştırma ihtiyacı duyulmuştur.

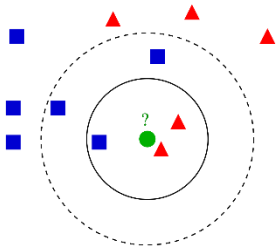
3.2.2 Eksik Verilerin Doldurulması

Verilerdeki (null) boş olan verileri çeşitli veri doldurma yöntemleri kullanılaraktan doldurma işlemi uygulanmıştır.

3.2.2.1 KNN (En Yakın Komşu)

Birçok veri bilimcisinin denediği popüler (hesaplama açısından en ucuz) yol, ortalama /medyan/mod kullanmak veya bu bir zaman serisi ise ,öncü veya gecikme kaydı kullanmaktır. Yaygın olarak tercih edilen KNN tabanlı Eksik Değer Tahmini budur.

İlk olarak Fix ve Hodges tarafından literatüre kazandırılmıştır [8]. Bu algoritmanın uygulama aşamasında Her örneğin eksik değerleri, eğitim setinde bulunan n_neighbors en yakın komşularından alınan ortalama değer kullanılarak hesaplanır. Hiçbirinde eksik olmayan özellikler yakınsa, iki örnek yakındır. Varsayılan olarak, en yakın komşuları bulmak için eksik değerleri destekleyen bir öklid uzaklık metriği olan nan_euclidean_distances kullanılır.



Şekil 3.1. Knn algoritmasının çalışma mantığı

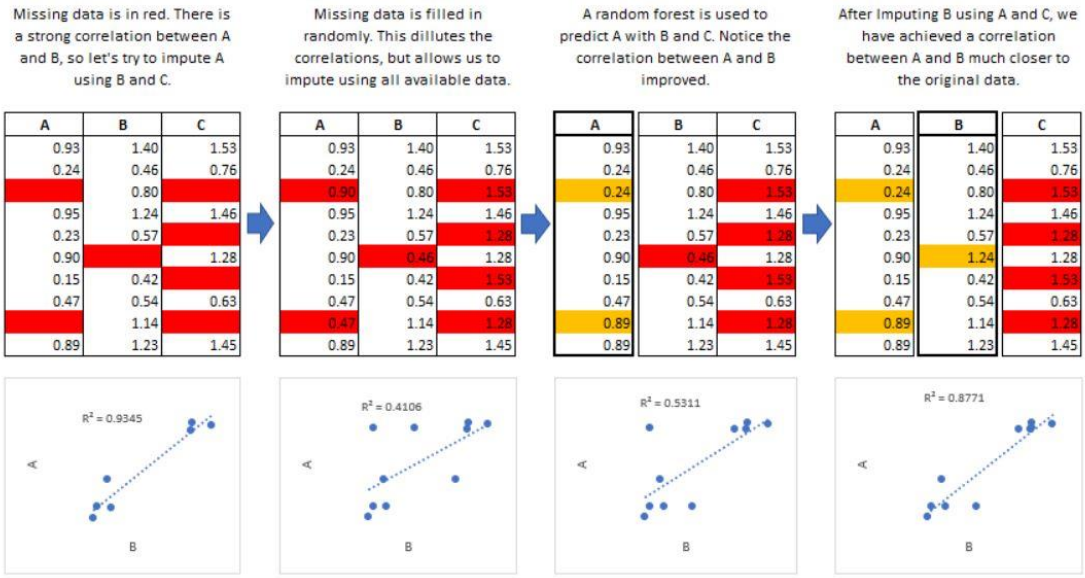
Şekil 3.1 ‘de görüldüğü gibi bir Öklid uzaklığı alınarak eksik olan veriyi hangi gruptan olduğu o uzaklık içindeki grupların fazlalıklarına bakılaraktan en fazla bulunan gruptan seçiliyor.

Bu algoritmanın güçlü yanlarından bazıları; uygulamanın basit ve verimli olması, algoritmanın verinin dağılımı hakkında varsayım yapmaması ve eğitim sürecinin hızlı olmasıdır. Zayıf yönleri ise sınıflandırmanın yavaş yapılması, yüksek bilgisayar hafızası gerektirmesidir.

3.2.2.2 MICE

Zincirleme Denklemlerle Çoklu Atama, veri kümelerindeki eksik verilerle başa çıkmak için sağlam, bilgilendirici bir yöntemdir. Prosedür, yinelemeli bir dizi tahmine dayalı model aracılığıyla bir veri kümesindeki eksik verileri 'doldurur' (imumlar). Her yinelemede, veri

kümesindeki belirtilen her bir değişken, veri kümesindeki diğer değişkenler kullanılarak atanır. Bu yinelemeler, yakınsama karşılandığı görülene kadar çalıştırılmalıdır.



Şekil 3.2.Mice algoritmasının çalışma mantığı

Belirtilen tüm değişkenler atanana kadar bu işleme devam edilir. Ortalama empoze edilen değerlerin yakınsamadığı görülürse ek yinelemeler çalıştırılabilir, ancak genellikle 5'ten fazla yineleme gerekli değildir. Tahminlerin doğruluğu, veri kümesindeki bilgi yoğunluğuna bağlı olacaktır. Hiçbir korelasyonu olmayan tamamen bağımsız değişkenlerden oluşan bir veri seti, doğru tahminler sağlamayacaktır (miceRanger) .Kullanıcının tahminlerin ne kadar geçerli olabileceğini belirlemesine izin veren tanılama grafikleri mevcuttur .Verilerdeki boş verilere için en uyumlu tahmin etmeye çalışarak bir veri kümesindeki eksik değerleri zahmetsizce uygulayabildiğimiz bir teknik olan Zincirli Denklemlerle Çok Değişkenli İmputasyon algoritması anlamına gelir.

3.2.2.3 Ortalama ile doldurma

Bu yöntem, veri setinde kayıp verinin olduğu alandaki diğer verilerin ortalamasını alarak kayıp olan verilerin yerine bu değeri yazarak doldurmaya yarayan yöntemdir. Veri aralığı düşük olan verilerde kullanıldığında yararlı olabilir.

3.2.3 Normalizasyon

Normalizasyon, verileri 0 ve 1 arasında yeniden ölçekler. Bu işlem tüm parametrelerin aynı pozitif ölçeğe sahip olması gereken bazı durumlarda yararlı olabilir. Ancak aykırı değerlerin (outliers) kaybolmasına yol açar. Verileri ölçeklemek için sayısal özellikleri normalize

edilerek 0-1 aralığına indirgenmiştir. Normalizasyon için azami normalleştirme (Min-Max Normalisation) yöntemi kullanılmıştır.

3.2.4 Öznitelik Seçimi

Öznitelik bir veri seti içerisinde bulunan ve hedeflenen model çıktısının oluşturmamızı sağlayacak olan her kolon/sütundur.

Öznitelik seçimi, veri seti içerisinde en yararlı öznitelikleri seçme ve bulma sürecidir. Bu işlem makine öğrenmesi modelinin performansını çok fazla etkilemektedir.

Gereksiz öznitelikler;

Modelinizin eğitim süresini arttırabilmektedir. Modelimizin basit ve açıklanabilir olmasını isteriz. Fazla sayıdaki öznitelik modelinizin yorumlana bilirliğini azaltılabilmektedir. Model başarışımın eğitim veri setinde aşırı öğrenme(overfitting) nedeniyle yüksek ancak test veri setinde ise düşük olmasına sebep olabilmektedir. Test veri setinde gelecek olan kayıtlar eğitim veri setindeki kayıtlar ile benzerlik göstermediği durumda modelde hata oranı yüksek olacaktır. Model tasarımınızda en önemli ve ilk adım mutlaka veri temizleme ve öznitelik seçimi olmalıdır. Öznitelik seçim metotlarının büyük bir kısım 3 ana kategoriye ayrılabilir.

3.2.4.1 Filtre Yöntemleri

Öznitelikleri önemini hesaplamak için öznitelik ile hedef değişken arasındaki ilişkiyi dikkate alan yöntemlerdir. Yapılan işlemler sonucunda veri setimizde filtreleme yaparız ve ilgili özellikleri seçerek bir alt küme oluştururuz. En sık kullanılan yöntemler ise Pearson Correlation ve Ki-Kare Yöntemidir.

3.2.4.2 Embedded Metotları

Bu yöntem öğrenme algoritmalarının bir parçası olduğundan gömülü kelimesi ile ifade edilir. Değişken seçimini algoritmanın kendisi yapar. Hem sınıflandırma için hem de regresyon için kullanılabilecek gömülü yöntemleri uygulayan algoritmalar yer almaktadır. Bu yöntem girdi değişkenler ve çıktı değişkeni için algoritma inşası aşamasında girdi değişkenlerin modele olan etkisini azaltarak ya da artırarak modelin değişken yapısını ortaya koyar. Böylece hangi değişkenlerin daha az değerli olduğunu görebiliriz. Bunun yanı sıra bu değişkenleri otomatik olarak belirlediği için ekstra bir efor harcamamıza gerek kalmayabilir.

3.2.4.3 Sarmalayıcı Yöntemler(Wrapper-Based)

Aslında istatistik alanında bilgi sahibi olan herkesin bildiği gerçeğin makine öğrenmesi alanında da sanki hiç daha önce yokmuş gibi servis edilmesi ile canlandırılan bir yöntemler dizidir. Burada fark olabilecek tek bir nokta belki eskiden yani henüz makineler bu kadar öğrenmeye aç olmadığı ya da yeterince güçlü olmadığı zamanlarda istatistikçiler bu işlemleri

elle yapılmaktaydı. Ancak makine öğrenmesindeki döngüler içerisine bu yöntemlerin girmesi ile çok yüksek sayıdaki değişkenlerin seçimi makinelere kalmış oldu. Bu yöntemler temel olarak modelleme aşmasında kullanılırlar. Model değişkenlerin seçimine göre tekrar ve tekrar kurulabilir. Ta ki en iyi sonuç veren model için değişkenler seçilene kadardır.

Yukarıdaki modelleme adımlarından yola çıkarak numaralandırılmış yerleri açıklayarak wrapper yöntemlerinin mantığını anlamaya çalışalım: Modelleme için kullanacağımız veri setinin ham hali.

Bu aşamada veri seti ile ilgili çeşitli işlemler yapılabilir. Bunlar betimleyici istatistikleri yorumlamak, kukla (dummy) değişken üretmek, kayıp (missing) değerleri doldurmak ya da kaldırmak gibi yöntemlerin yanı sıra değişken seçiminin bir diğer türü olan filtreleme yöntemleri de bu aşamada kullanılabilirler.

Tüm bu ön hazırlık aşamasından sonra makine öğrenmesinde seçilen algoritma ile modelleme çalışması yapılır.

Bu aşamada oluşturulan model performansı değerlendirilir. Öyle ki buradaki değerleme veri setinde yer alan değişkenlerin model üzerindeki etkisi göz önünde bulundurulur.

Beşinci aşamada performansı iyi/kötü etkileyen değişkenler seçilir ya da çıkarılır. Böylece 3. aşamaya yeni seçilen/eksiltilen değişkenlerin yer aldığı veri seti ile devam edilir. Performansı en iyi yansıtan değişken seçimi olana kadar üçüncü, dördüncü ve beşinci aşamalar kendini tekrar eder.

Son aşamada en iyi performans ile sonuçlandığı düşünülen, döngüden çıkan (3-4-5) model kesinleşmiş olur.

Peki ya bu yöntemler neler diye sormadan önce aşağıdaki notuma aklınızda tutmanızda fayda var.

Düşük etkisi görünen bir girdi değişkeninin modelden çıkarılması her zaman model performansının artacağı anlamına gelmez. Bu değişkenin başka bir girdi değişken ile olan ilişkisinden dolayı çıktı değişken etkileniyor olabilir. Bu yüzden farklı alt kümeler ile denemeler yapılmaktadır.

Regresyon ya da sınıflandırma algoritmalarında kullanılan temelde üç tip sarmal (wrapper) yöntemi bulunmaktadır bunlar:

İleri Doğru Seçim = Step Forward Selection

Geriye doğru seçim = Step Backwards Selection

Adımsal Seçim = Stepwise (Exhaustive) Selection

3.2.4.3.1 Forward/İleri Arama

Boş bir öznitelik kümesi ile başlıyarak her seferinde gruptaki öznitelikleri tek tek ekleyerek özniteliklerin kalitesi test edilir.

N öznitelikli bir veri setinde;

İlk adımda en iyi tahminleme yapan tek öznitelik seçilir.

İkinci adımda ilk seçilen öznitelik ile beraber en iyi tahminleme yapan 2.öznitelik belirlenir.Böylece en iyi tahminleme yapan 2'li alt grup oluşturulmuş olur.

Bu işlemler en iyi tahminleme yapan "m" adet öznitelik kombinasyonu bulunana kadar devam eder.

3.2.4.3.2 Backward /Geri Arama(Recursive Feature Elemination)

Bu yöntem en iyi öznitelik alt kümesi bulunanaa kadar tüm öznitelik kümesinden başlayarak adım adım en kötü performans gösten öznitelikler elenir.

"n" adet öznitelikli bir veri setinde;

Tüm veri setindeki öznitelikler alınır ve en az performans gösteren öznitelik kaldırılır.

İkinci adımda ilk adımda belirlenen alt gruptaki yine en az performans gösteren öznitelik kaldırılır.

Bu işlemler en iyi tahminleme yapan "m" adet öznitelik kombinasyonu bulunana kadar devam eder.

3.2.4.3.3 Adımsal Seçim (Stepwise Selection)

Adımsal seçim, ileri ve geri seçimin avantajlarını birleştiren bir teknik olarak önerilmiştir. Aramanın herhangi bir noktasında, tek bir tahmin değişkeni eklenebilir veya silinebilir. Genellikle, başlangıç alt kümesi boş kümedir. Bir tahmin değişkeni alt kümesinin bit gösterimi açısından düşündüğümüzde, aşamalı seçim, aramanın herhangi bir noktasında gösterimdeki bir bitin çevrilmesine izin verir. Bu nedenle, bir alt kümenin bit gösterimi N bit içerdiğinden, adım adım seçim için arama grafiğindeki her alt kümenin N komşusu vardır. Hiçbir bit bir kereden fazla çevrilmezse, adım adım seçim sona ermeden önce en fazla N^2 alt küme değerlendirilir. Bununla birlikte, her bitin en fazla bir defa çevrileceğinin garantisi yoktur.

İleri veya geri seçim ile adımsal seçimin etkinliğini karşılaştırmak için güçlü teorik sonuçlar mevcut değildir. Adımsal seçim, diğer iki teknikten daha fazla alt kümeyi değerlendirir, bu nedenle pratikte daha iyi alt kümeler üretme eğilimindedir [15].

3.2.5 Regresyon Analizleri

Regresyon, genellikle finans ve yatırımda kullanılan, bir bağımlı değişken ile bir dizi bağımsız değişken arasındaki ilişkinin gücünü ve niteliğini belirlemeye çalışan istatistiksel bir terimdir.

Regresyon analizi, bağımlı değişken ile bağımsız değişken arasındaki ilişkiyi tahmin etmek için kullanılan istatistiksel bir araçtır. Daha spesifik olarak, bağımlı değişkenin bağımsız

değişkenlerdeki değişikliklere göre nasıl değiştiğine odaklanır. Ayrıca değişkenler arasındaki gelecekteki ilişkinin modellenmesine de yardımcı olur.

Regresyon analizi, doğrusal, doğrusal olmayan ve çoklu doğrusal olmak üzere çeşitli türlerden oluşuyor.

3.2.5.1 Rastgele Orman Regresyon

Rastgele orman regresyonu yöntemi pek çok karar ağacından oluşan bir topluluk tabanlı öğrenme yöntemidir. 1995 yılında Kam Ho tarafından ortaya atılmıştır [9].

Yöntemde oluşturulacak olan ağaçların sayısı için literatürde kesin bir değer belirtilmemiştir. Bu nedenle kullanılan veri kümesine uygun olarak *s* sayıda farklı ağaç kullanılarak hata değerindeki artışa bakılmalıdır. Analizi yapılacak veri kümesinin *M* özellik ve *N* satırdan oluştuğunu kabul edersek, algoritmada yer alacak her bir alt ağacı elde etmek için aşağıdaki adımlar uygulanır.

Ağacın her bir düğümünde karar vermek için kullanılacak değişkenlerin sayısı m ile ifade edilirse, m < M olmalıdır.

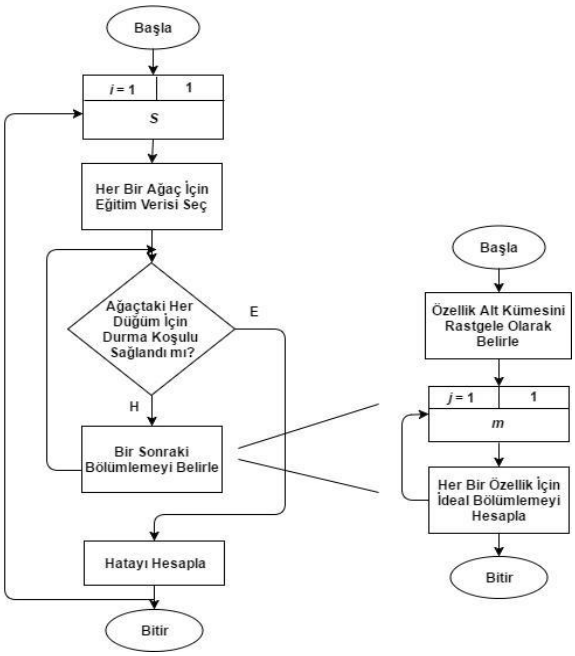
N satırdan oluşan eğitim setinden rastgele olarak n satır seçilir. Kalanlar ise oluşan ağacın hatasını tahmin etmek için kullanılır.

Ağacın her bir düğümü için o düğümdeki kararın dayandığı m değişken rastgele olarak seçilir.

Eğitim seti kullanılarak bu m değişken için en iyi bölümlemeler hesaplanır.

Her ağaç veri kümesi kullanılarak tüm yapraklar elde edilene kadar budama yapılmadan oluşturulur.

Rastgele orman regresyonu algoritmasının akış diyagramı Şekil 3. 2'de detaylı olarak gösterilmiştir.



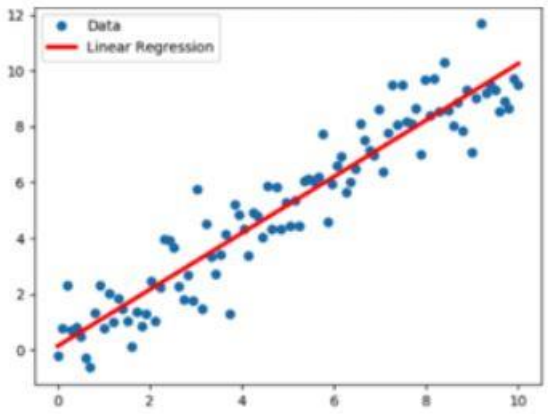
Şekil 3.2. Random Forest algoritması akış diyagramı

Rastgele orman regresyonu algoritması büyük veri kümeleri üzerinde kolaylıkla işlem yapabilen bir yöntemdir. Alt ağaçları kullanarak işlem yapması sebebiyle özellik indirgemesine ihtiyaç duymadan işlemleri gerçekleştirebilir. Aynı zamanda farklı özellik seçimlerine bağlı olarak işlem yaptığı için sınıf etiketinin belirlenmesinde etkisi yüksek özelliklerin tespit edilmesini de sağlamaktadır.

3.2.5.2 Doğrusal Regresyon (Lineer Regression)

Doğrusal korelasyon ve basit doğrusal regresyon, iki değişken arasındaki doğrusal ilişkiyi inceleyen istatistiksel yöntemlerdir. Burada şu farklılığı vurgulamakta fayda var: Korelasyon, iki değişkenin ne kadar ilişkili olduğunu gösterirken, doğrusal regresyon, iki değişken arasındaki ilişkiye dayanarak birinin değerini diğerinden tahmin etmeyi sağlayan bir denklem (model) oluşturmayı içerir.

Doğrusal regresyon, bir dizi noktaya en uygun düz çizgiyi veya hiper düzlemi bulmak için kullanılmaktadır. Bir diğer ifadeyle doğrusal regresyon, en uygun düz çizgi (regresyon çizgisi olarak da bilinir) kullanarak bağımlı değişken (Y) ile bir veya daha fazla bağımsız değişken (X) arasında bir ilişki kurar.



Şekil 3.3. kırmızı çizgi en uygun düz çizgi olarak adlandırılır.

3.2.5.3 Extreme Gradyan Artırma (Extreme Gradient Boost ,XGBoost)

Ekstrem Gradyan Artırma (XGBoost) algoritması sonuçların performansı ve çalışma hızı açısından gradyan artırma algoritmasının geliştirilmiş halidir. Makine öğrenmesi problemlerini çözmek için kullanılan bir algoritma paketidir. Milyonlarca örneklemiaz miktarda kaynak kullanarak çözebilen verimli bir algoritmadır. Tianqi Chen tarafından 2014

yılında geliştirilip açık kaynak kodlu makine öğrenmesi algoritmalarının bulunduğu DMLC (Distributed Machine Learning Community) adlı kütüphaneye eklenmiştir [10].

XGBoost algoritması milyonlarca veri setini kısa zamanda ve diğer algoritmalara göre daha az işlemci gücü ve geçici belleğe (ram) ihtiyaç duyarak analiz edebildiği için bir çok makine öğrenmesi ve veri analizi problemlerinin çözümünde kullanılmıştır. Örneğin, Avrupa Nükleer Araştırma Merkezi (CERN), Büyük Hadron Çarpıştırıcısı'nın ürettiği yıllık 3 petabayt (PB, 1024 terabayt) veriyi analiz etmek için XGBoost algoritmasını kullanmıştır [11].

3.2.5.4 Ridge Regresyon(L2 regularization)

Çok değişkenli regresyon verilerini analiz etmede kullanılır. Amaç hata kareler toplamını minimize eden katsayıları, bu katsayılara bir ceza uygulayarak bulmaktır. Over-fittinge karşı dirençlidir. Çok boyutluluğa çözüm sunar. Tüm değişkenler ile model kurar, ilgisiz değişkenleri çıkarmaz sadece katsayılarını sıfıra yaklaştırır. Modeli kurarken alpha (ceza) için iyi bir değer bulmak gerekir.

3.3.5.5 Lasso Regresyon(L1 regularization)

Ürettiği modelin tahmin doğruluğunu ve yorumlanabilirliğini arttırmak için hem değişken seçimi hem de regularization yapar. Aynı ridge regresyonda olduğu gibi amaç hata kareler toplamını minimize eden katsayıları, katsayılara ceza uygularayarak bulmaktır. Fakat ridge regresyondan farklı olarak ilgisiz değişkenlerin katsayılarını sıfıra eşitler.

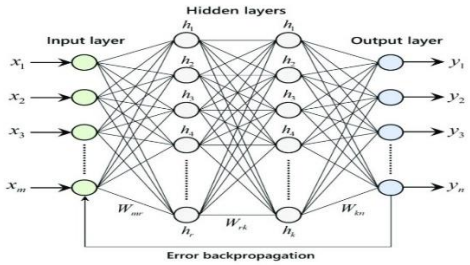
3.3.5.6 Karar Ağaçları Regresyon (Decision Trees Regression)

Karar ağaçlarının amacı veri setinde değişkenleri kullanarak hedeflenen değeri tahmin edebilecek bir ağaç modeli oluşturmaktır. Karar ağaçları bir kök ile başlar. Bütün veriler düğüm adı verilen karar verme noktalarından geçerek dallara ayrılır ve yapraklara ulaşır.Düğüm sonucunda verilen karar çeşitliliğine göre dallar oluşturulur. Son olarak dallar yapraklara yani sınıflandırılan ya da sayısal değer atanan bir çıktı değerlerine ulaşır. Dallar birbirine karışmazlar, düğüm aşamalarından geçerek yukarıdan aşağıya doğru ağaç biçimine benzer bir yapı oluştururlar. Yukarıdaki örnek basit bir karar ağacı regresyonu örneğidir. Karar ağacı regresyon yönteminin; sonuçların rahat yorumlanabilmesi, eğitim ve uygulama sürecinin hızlı olması ve anlaşılır görseller sunabilmesi gibi avantajları vardır [12] . Karar ağacı yöntemindeki en önemli nokta kök düğümlerden itibaren bütün düğümlerin hangi kritere göre ve nasıl dallanacağına karar vermektir. Bu noktada literatürde sıkça kullanılan ID3, C4.5, CART, CHAID, C5.0, gibi algoritmalar bulunmaktadır [13,14].

3.2.6 Regresyon Tabanlı Yapay Sinir Ağı

Regresyon tabanlı yapay sinir ağı ,yapay sinir ağ modelinin regresyon problemleri için kullanılan bir model olarak karşımıza çıkıyor. Yapay Sinir Ağları, insan beynindeki nöronların çalışmalarını simüle eden derin öğrenme algoritmalarından biridir.

3.2.6.1 Yapay Sinir Ağı Yapısı



Şekil 3.4. Yapay sinir ağı modeli

Yukarıdaki Şekil 3.4 'te yapay sinir ağlarının girdi katmanı, gizli katmanlar, çıktı katmanından oluştuğunu görülüyor. Gizli katman birden fazla olabilir. Her katman n sayıda nörondan oluşur. Her katman, nöronların her biri ile ilişkili bir aktivasyon fonksiyonuna sahip olacaktır. Aktivasyon işlevi, ilişkide doğrusal olma yanlığın ortaya çıkmasından sorumlu olan işlevdir. Bizim durumumuzda, çıktı katmanı lineer bir aktivasyon fonksiyonu içermelidir. Her katman, kendisiyle ilişkilendirilmiş düzenleyicilere de sahip olabilir. Düzenleyiciler, fazla takmayı önlemekten sorumludur.

Yapay Sinir Ağları iki aşamadan oluşur,

İleri Yayılım

Geriye Yayılım

İleri yayılım, her bir özellik ile ağırlıkların çarpılması ve eklenmesi işlemidir. Önyargı da sonuca eklenir. Geriye yayılma, modeldeki ağırlıkların güncellenmesi işlemidir. Geriye yayılma, bir optimizasyon işlevi ve bir kayıp işlevi gerektirir.

Regresyon tabanlı yapay sinir ağlarının kullanılmasının amacı, regresyon modellerinin sadece özellikler ve hedef arasındaki lineer ilişkiyi öğrenebilmesi ve dolayısıyla karmaşık lineer olmayan ilişkiyi öğrenememesidir. Öznitelikler ve hedef arasındaki karmaşık doğrusal olmayan ilişkiyi öğrenmek için başka tekniklere ihtiyacımız var. Bu tekniklerden biri de Yapay Sinir Ağlarını kullanmaktır. Yapay Sinir Ağları, her katmanda aktivasyon fonksiyonunun varlığından dolayı öznitelikler ve hedef arasındaki karmaşık ilişkiyi öğrenme yeteneğine sahiptir.

3.2.6 Çapraz Doğrulama (Cross-Validation)

Çapraz doğrulama, makine öğrenimi modellerinin performansını (veya doğruluğunu) tahmin etmek için kullanılan istatistiksel bir yöntemdir. Özellikle veri miktarının sınırlı olabileceği bir durumda, tahmine dayalı bir modelde fazla uydurmaya karşı koruma sağlamak için kullanılır. Çapraz doğrulamada, verileri sabit sayıda katlar (veya bölümler) yapar, analizi her kat üzerinde çalıştırır ve ardından genel hata tahmininin ortalamasını alırsınız.

Bir Makine Öğrenimi göreviyle uğraşırken, size en iyi puanı verebilecek en uygun algoritmayı seçebilmeniz için sorunu doğru bir şekilde tanımlamanız gerekir. Ama modelleri nasıl karşılaştıracağız?

Diyelim ki modeli mevcut veri seti ile eğittiniz ve şimdi modelin ne kadar iyi performans gösterebileceğini bilmek istiyorsunuz. Bir yaklaşım, modeli, üzerinde eğittiğiniz veri kümesi üzerinde test etmeniz olabilir, ancak bu iyi bir uygulama olmayabilir.

Öyleyse, modeli eğitim veri setinde test etmenin nesi yanlış bunu yaparsak, eğitim verilerinin gerçek dünyanın tüm olası senaryolarını temsil ettiğini ve kesinlikle böyle olmayacağını varsayıyoruz. Ana hedefimiz, eğitim veri kümesi aynı zamanda gerçek dünya verileri olmasına rağmen, modelin gerçek dünya verileri üzerinde iyi çalışabilmesidir, oradaki tüm olası veri noktalarının (örneklerinin) küçük bir kümesini temsil eder.

Bu yüzden modelin gerçek puanını bilmek için daha önce hiç görmediği veriler üzerinde test edilmesi gerekir ve bu veri kümesine genellikle test kümesi denir. Ancak verilerimizi eğitim verileri ve test verileri olarak bölersek, test veri kümesinin tutabileceği bazı önemli bilgileri kaybederiz bu sebeple çapraz doğrulama ile verilerin içinde olan önemli verileri kullanmak için çapraz doğrulama kullanılır.

4.SİSTEM TASARIMI

Veri kümesi Kaggle sitesinden alınmıştır.(https://www.kaggle.com/c/house-prices-advanced-regression-techniques)

4.1 Veri Kümesi hakkında Ön Bilgi

Deneysel çalışma için kullanılan veri kümesi, tahmin modelleme ve analitik yarışmalarında kullanılan ve önemli bir platform olan Kaggle ’den elde edilmiştir. Veri kümesi; konutlara ait özellikler üzerinden bu konutlara ait satış fiyatlarının regresyon yöntemleri ile tahmin edilmesi amacıyla oluşturulmuştur.

Veri kümesi 3 bölüm halindedir. İlk bölüm (train)olan ve 1460 sütun ve 81 kolona sahip olan verimizdir. İkinci bölüm’ de (test) olan ve 1459 sütun ve 80 kolona sahiptir. Üçüncü bölümde (sample_submission) ise ikinci bölümdeki (test) evlerin fiyatları bulunmaktadır. Toplam veri kümesi 2919 farklı konut için 80 farklı özellik ve bu özelliklere göre de konutların fiyatlarını içermektedir. Bu 80 özelliğin 43 tanesi kategorik değer alırken geriye kalan 38 tanesi sayısal değer almaktadır. Aşağıdaki Çizelge 4.1 ‘de görüldüğü üzere kategorik verilerin içeriğini göstermektedir.

4.1.2 Kategorik Veriler

Çizelge 4.1 Öznitelikleri hakkında bilgi

Öznitelik Adı	Değişken Sayısı	Değişkenler	Eksik veri sayısı
MSZoning (Satışın genel imar sınıflamasını tanımlar.)	5	RL: konut düşük yoğunluk RM: Konut orta yoğunluğu FV: Köy konutu RH: Konut yüksek yoğunluğu C(all): Reklam	4
Street: Mülke yol erişim türü	2	Grvl: Çakıl Pave: Döşeli	0
Alley: Mülke geçit erişimi türü	2	Grvl: Çakıl Pave: Döşeli	2727
LotShape: Mülkün genel şekli	4	Reg: Düzenli IR1: biraz düzensiz IR2: Orta derecede Düzensiz IR3: Düzensiz	0
Utilities: Kullanılabilir yardımcı program türleri	2	AllPub: Tüm kamu hizmetleri (E, G, W, &S) NoSeWa: Elektrik, Gaz ve Su (Feptik Tank)	2
LandContour: Mülkün düzlüğü	4	Lvl: Daire/Seviyeye Yakın Bnk: Banked - Sokak seviyesinden binaya hızlı ve önemli artış HLS: Hillside - Bir yandan diğer yana önemli eğim	0

		Low: Çöküntü	
LotConfig: Parti yapılandırması	5	Inside: iç kısım Corner: köşe partisi CulDSac: Çıkmaz sokak FR2: Taşınmazın 2 cephesi FR3: Taşınmazın 3 cephesi	0
LandSlope: Mülkün eğimi	3	Gtl: Hafif eğim Mod: Orta Eğim Sev: Şiddetli Eğim	0
HouseStyle: konut tarzı	8	1Story: Bir Katlı 1.5Fin: Bir buçuk kat: 2. seviye tamamlandı 1.5Unf: Bir buçuk kat: 2. seviye tamamlanmamış 2Story: İki kat 2.5Fin: İki buçuk kat: 2. seviye tamamlandı 2.5Unf: İki buçuk kat: 2. seviye bitmemiş SFoyer: Bölünmüş Fuaye SLvl: Bölünme düzeyi	0
Neighborhood: Ames şehir sınırları içindeki fiziksel konumlar	25	Blmngtn: Bloomington Tepeleri Blueste: Bluestem BrDale: Briardale BrkSide: Brookside ClearCr: Clear Creek: Creek CollgCr: Kolej Crawfor: Crawford Edwards: Edwards Gilbert: Gilbert IDOTRR: Iowa DOT ve Demir Yolu MeadowV: Meadow Köy Mitchel: Mitchell Names: North Ames NPkVill: Northpark Villa NWAmes: Northwest Ames NoRidge: Northridge NridgHt: Northridge Heights OldTown: Old Town SWISU: South & West of Iowa State University Sawyer: Sawyer SawyerW: Sawyer West Somerst: Somerset StoneBr: Stone Brook Timber: Timberland Veenker: Veenker	0
Condition1: Çeşitli koşullara yakınlık	9	Artery: Ana caddeye bitişik Feedr: Besleme caddesine bitişik Norm: Normal RRNn: Kuzey-Güney Demiryolunun 200' içinde RRAn: Kuzey-Güney Demiryoluna Bitişik PosN: Yakın pozitif tesis dışı özellik- park, yeşil kuşak, vb. PosA: Pozitif site dışı özelliğinin bitişiğinde RRNe: Doğu-Batı Demiryolunun 200' içinde RRAe: Doğu-Batı Demiryoluna Bitişik	0
Condition2: Çeşitli koşullara	8	Artery : Ana caddeye bitişik Feedr : Besleme caddesine bitişik Norm : Normal	0

yakınlık (birden fazla varsa)		RRNn : Kuzey-Güney Demiryolunun 200' içinde RRAn : Kuzey-Güney Demiryoluna Bitişik PosN: Yakın pozitif tesis dışı özellik - park, yeşil kuşak PosA: Pozitif site dışı özelliğinin bitişiğinde RRAe: Doğu-Batı Demiryoluna Bitişik	
BldgType: Konut tipi	5	1Fam: Tek Aile Müstakil 2FmCon: İki Aile Dönüşümü; aslen tek aile konutu olarak inşa edilmiş Duplex: dubleks TwnhsE: Konak Bitiş Birimi Twnhs: Konak İç Ünitesi	0
RoofStyle: Çatı tipi	6	Flat: Düz Gable: Üçgen çatı Gambrel: Kumarbaz (Ahır) Hip: Çatı dış açısı Mansard: Tavan arası Shed: Baraka	0
RoofMatl: Çatı materyal	8	ClyTile: Kil veya Çini CompShg: Standart (Kompozit) Shingle Membran: Membrane Metal: Metal Roll: Rulo Tar&Grv: Çakıl ve Katran WdShake: Odun sallar WdShngl: Ahşap Zona	0
Exterior1st: Evin dış kaplaması	15	AsbShng: Asbest Zona AsphShn: Asfalt Zona BrkComm: Tuğla Ortak BrkFace: Tuğla Yüz CBlock: Kül Blok CemntBd: Çimento levha HdBoard: Sert Tahta ImStucc: taklit sıva MetalSd: Metal Dış Cephe Kaplaması Plywood: kontrplak Stone: Taş Stucco: Sıva VinylSd: Vinil Dış Cephe Kaplaması Wd Sdng: Ahşap Dış Cephe Kaplaması WdShing: Ahşap Zona	1
Exterior2nd: Evin dış kaplaması (birden fazla malzeme varsa)	16	AsbShng: Asbest Zona AsphShn: Asfalt Zona BrkComm: Tuğla Ortak BrkFace: Tuğla Yüz CBlock: Kül Blok CemntBd: Çimento levha HdBoard: Sert Tahta ImStucc: Taklit sıva MetalSd: Metal Dış Cephe Kaplaması Other: Başka Plywood: kontrplak	1

		Stone: Taş Stucco: Sıva VinylSd: Vinil Dış Cephe Kaplaması Wd Sdng: Ahşap Dış Cephe Kaplaması WdShing: Ahşap Zona	
MasVnrType: Duvar kaplama tipi	4	BrkCmn: Tuğla Ortak BrkFace: Tuğla Yüz None: Yok Stone: Taş	23
ExterQual: Dış cephedeki malzemenin kalitesini değerlendirir	4	Ex: Harika Gd: İyi TA: Ortalama/Tipik Fa: Makul	0
ExterCond: Malzemenin mevcut durumunu dış cephede değerlendirir	5	Ex: Harika Gd: İyi TA: Ortalama/Tipik Fa : Makul Po: kötü	0
Foundation: Temel türü	6	BrkTil: Tuğla ve Kiremit CBlock: Kül Blok PConc: Dökülmüş beton Slab: Döşeme Stone: Taş Wood: Odun	0
BsmtQual: Bodrumun yüksekliğini değerlendirir	4	Ex: Mükemmel (100+ inç) Gd: İyi (90-99 inç) TA: Tipik (80-89 inç) Fa: Fuar (70-79 inç) NA: Bodrum Yok	0
BsmtCond: Bodrumun genel durumunu değerlendirir	4	Gd: İyi TA: Tipik- hafif neme izin verilir: Fa: Orta- nem veya biraz çatlama veya çökme Po: Kötü- Şiddetli çatlama, çökme veya ıslaklık NA: Bodrum yok	0
BsmtExposure: Grev veya bahçe seviyesindeki duvarları ifade eder	4	Gd: İyi Pozlama Av: Ortalama Pozlama (bölünmüş seviyeler veya fuayeler genellikle ortalama veya üzeri puan alır) Mn: Minimum Pozlama No: Pozlama Yok NA: Grev veya bahçe yok	0
BsmtFinType1: Bodrum bitmiş alanın değerlendirmesi	6	GLQ: İyi Yaşam Alanları ALQ: Ortalama Yaşam Alanları BLQ: Ortalama Yaşam Alanlarının Altında Rec: Ortalama Dinlenme Odası LwQ: Düşük kalite Unf: bitmemiş NA: Bodrum yok	0
BsmtFinType2: Bodrum bitmiş alanın	6	BsmtFinType1 ile aynı değerler	0

değerlendirmesi (birden fazla tip varsa)			
Heating: Isıtma türü	6	Floor: Yer Fırını GasA: Gaz zorla sıcak hava fırını GasW: Gaz sıcak su veya buhar ısısı Grav: Yerçekimi fırını OthW: Gaz dışında sıcak su veya buhar ısısı Wall: Duvar fırını	0
HeatingQC: Isıtma kalitesi ve durumu	5	Ex : Harika Gd: İyi TA: Ortalama/Tipik Fa: Makul Po: kötü	0
CentralAir: Merkezi klima	2	N: Hayır Y: Evet	0
Electrical: Elektrik sistemi	5	SBrkr: Standart Devre Kesiciler ve Romex FuseA: AMP üzerinde Sigorta Kutusu ve tüm Romex kabloları (Ortalama) FuseF: AMP Sigorta Kutusu ve çoğunlukla Romex kablolama (Fair) FuseP: AMP Sigorta Kutusu ve çoğunlukla düğme ve boru kablolaması (zayıf) Mix :Karışık	1
KitchenQual: Mutfak kalitesi	4	Ex: Harika Gd: İyi TA: Ortalama/Tipik Fa: Makul	1
Functional: Ev işlevselliği (Kesintiler garanti edilmediği sürece tipik olduğunu varsayın)	7	Typ: Tipik İşlevsellik Min1: Küçük Kesintiler 1 Min2: Küçük Kesintiler 2 Mod: Orta Kesintiler Maj1: Büyük Kesintiler 1 Maj2: Büyük Kesintiler 2 Sev: Ağır hasarlı	2
FireplaceQu: Şömine kalitesi	6	Ex: Mükemmel- Olağanüstü duvar şömine Gd: İyi- Ana seviyede yığma şömine TA: Ortalama- Ana yaşam alanında prefabrik şömine veya bodrum katında yığma şömine Fa: Makul- Bodrum prefabrik şömine Po: Zayıf – soba NA: Şömine yok	0
GarageType: Garaj konumu	6	2Types: Birden fazla garaj türü Attchd : Eve bağlı Basment: Bodrum garajı BuiltIn: Yerleşik (Evin garaj kısmı- genellikle garajın üzerinde oda bulunur) CarPort: Yanları açık garaj Detchd: Evden müstakil NA:Garaj yok	0

GarageFinish: Garajın iç dekorasyonu	3	Fin: Bitmiş RFn: Kaba bitmiş Unf: Bitmemiş NA: Garajın iç dekorasyonu yok	0
GarageQual: Garaj kalitesi	5	Ex: Harika Gd: İyi TA: Tipik/Ortalama Fa: Makul Po: Zayıf NA: Garaj kalitesi yok	0
GarageCond: Garaj durumu	5	Ex: Harika Gd: İyi TA: Tipik/Ortalama Fa: Makul Po: Zayıf NA: Garajın durumu yok	0
PavedDrive: Asfalt yol	3	Y: Döşeli P: Kısmi Kaplama N: Dirt/Gravel	0
PoolQC: Havuz kalitesi	3	Ex: Harika Gd: İyi Fa: Makul	2909
Fence: Çit kalitesi	4	GdPrv: İyi Gizlilik MnPrv: Minimum Gizlilik GdWo: İyi odun MnWw: Minimum Ahşap/Tel	2348
MiscFeature: Diğer kategorilerde kapsanmayan çeşitli özellikler	4	Gar2: 2. Garaj (garaj bölümünde anlatılmamışsa) Othr: Başka Shed: Kulübe (100 SF'den fazla) TenC: Tenis kortu	2814
SaleType: Satış türü	9	WD: Garanti Belgesi – Konvansiyonel CWD: Garanti Belgesi – Nakit New: Ev yeni yapıldı ve satıldı COD: Mahkeme Memuru Tapu/Emlak Con: Sözleşme %15 Peşinat normal şartlar ConLw: Sözleşme düşük peşinat ve düşük faiz ConLI: Sözleşme düşük faiz ConLD: Sözleşme düşük aşağı Oth: Diğer	1
SaleCondition: Satış durumu	6	Normal: Normal Satış Abnorml: Anormal Satış- ticaret, haciz, açığa satış AdjLand: Bitişik Arazi Alımı Alloca: Tahsis- ayrı tapulu iki bağlantılı mülk, tipik olarak bir garaj ünitesine sahip apartman dairesi Family: Aile üyeleri arasında satış Partial: Ev, en son değerlendirildiğinde tamamlanmadı (Yeni Evlerle ilişkili)	1

Sayısal verilerin içinde bulunan iki tane özniteliğin kategorik değerden sayısal değere dönüştürülmüş halde bulunuyor.

Çizelge 4.2 Kategorik öznitelikler iken sayısal olan öznitelikler

Öznitelik Adı	Özniteliklerin Değeri
OverallQual: GenelKalite: Evin genel malzemesini ve bitişini değerlendirir	10: Çok mükemmel 9: Harika 8: Çok güzel 7: İyi 6: Ortalamanın üstü 5: Ortalama 4: Ortalamanın altında 3: makul 2: kötü 1: Çok kötü
OverallCond: Genel Durum: Evin genel durumunu değerlendirir	

Yukarıdaki Çizelge 4.2' de görüldüğü üzere özniteliklerin içerikleri gösteriliyor. Diğer sayısal veriler ise Şunlardır:

Lot Frontage: Parsel Cephesi: Mülke bağlı caddenin doğrusal ayakları

Lot Area: Parti Alanı: Metrekare cinsinden arsa büyüklüğü

YearBuilt: Yapım Yılı: Orijinal yapım tarihi

YearRemodAdd: Yıl Tadilat Ekleme: Tadilat tarihi (tadilat veya ilave yapılmamışsa inşaat tarihi ile aynı)

MasVnrArea: Metre kare cinsinden duvar kaplama alanı

BsmtFinSF1: Tip 1 bitmiş fit kare

BsmtFinSF2: Tip 2 bitmiş fit kare

TotalBsmtSF: Bodrum alanının toplam metrekaresi

Index	count	mean	std	min	25%	50%	75%	max
Id	2919	1460	842.787	1	730.5	1460	2189.5	2919
MSSubClass	2919	57.1377	42.5176	20	20	50	70	190
LotFrontage	2433	69.3058	23.3449	21	59	68	80	313
LotArea	2919	10168.1	7887	1300	7478	9453	11570	215245
OverallQual	2919	6.08907	1.40995	1	5	6	7	10
OverallCond	2919	5.56458	1.11313	1	5	5	6	9
YearBuilt	2919	1971.31	30.2914	1872	1953.5	1973	2001	2010
YearRemodAdd	2919	1984.26	20.8943	1950	1965	1993	2004	2010
MasVnrArea	2896	102.201	179.334	0	0	0	164	1600
BsmtFinSF1	2918	441.423	455.611	0	0	368.5	733	5644
BsmtFinSF2	2918	49.5822	169.206	0	0	0	0	1526
BsmtUnfSF	2918	560.772	439.544	0	220	467	805.5	2336
TotalBsmtSF	2918	1051.78	440.766	0	793	989.5	1302	6110
1stFlrSF	2919	1159.58	392.362	334	876	1082	1387.5	5095
2ndFlrSF	2919	336.484	428.701	0	0	0	704	2065
LowQualFinSF	2919	4.69442	46.3968	0	0	0	0	1064
GrLivArea	2919	1500.76	506.051	334	1126	1444	1743.5	5642
BsmtFullBath	2917	0.429894	0.524736	0	0	0	1	3
BsmtHalfBath	2917	0.0613644	0.245687	0	0	0	0	2
FullBath	2919	1.568	0.552969	0	1	2	2	4
HalfBath	2919	0.380267	0.502872	0	0	0	1	2
BedroomAbvGr	2919	2.86023	0.822693	0	2	3	3	8
KitchenAbvGr	2919	1.04454	0.214462	0	1	1	1	3
TotRmsAbvGrd	2919	6.45152	1.56938	2	5	6	7	15
Fireplaces	2919	0.597122	0.646129	0	0	1	1	4

Şekil 4.1. Sayısal verilerin özellikleri (ortalama, minimum, maksimum) değerleri

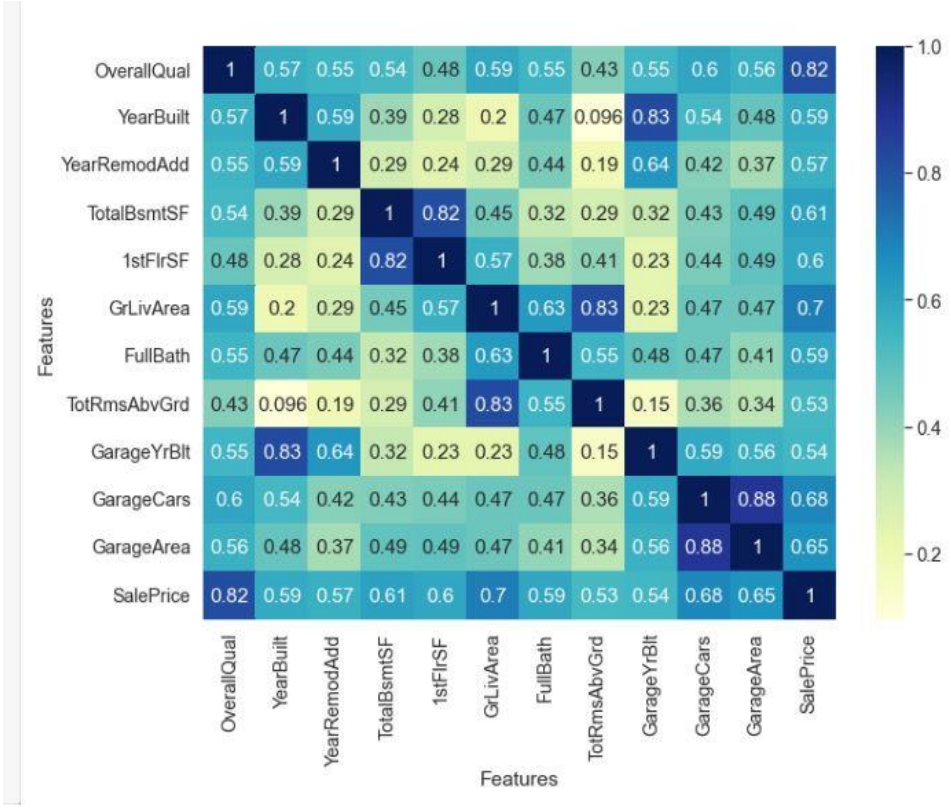
Index	count	mean	std	min	25%	50%	75%	max
1stFlrSF	2919	1159.58	392.362	334	876	1082	1387.5	5095
2ndFlrSF	2919	336.484	428.701	0	0	0	704	2065
LowQualFinSF	2919	4.69442	46.3968	0	0	0	0	1064
GrLivArea	2919	1500.76	506.051	334	1126	1444	1743.5	5642
BsmtFullBath	2917	0.429894	0.524736	0	0	0	1	3
BsmtHalfBath	2917	0.0613644	0.245687	0	0	0	0	2
FullBath	2919	1.568	0.552969	0	1	2	2	4
HalfBath	2919	0.380267	0.502872	0	0	0	1	2
BedroomAbvGr	2919	2.86023	0.822693	0	2	3	3	8
KitchenAbvGr	2919	1.04454	0.214462	0	1	1	1	3
TotRmsAbvGrd	2919	6.45152	1.56938	2	5	6	7	15
Fireplaces	2919	0.597122	0.646129	0	0	1	1	4
GarageYrBlt	2760	1978.11	25.5743	1895	1960	1979	2002	2207
GarageCars	2918	1.76662	0.761624	0	1	2	2	5
GarageArea	2918	472.875	215.395	0	320	480	576	1488
WoodDeckSF	2919	93.7098	126.527	0	0	0	168	1424
OpenPorchSF	2919	47.4868	67.5755	0	0	26	70	742
EnclosedPorch	2919	23.0983	64.2442	0	0	0	0	1012
3SsnPorch	2919	2.60226	25.1882	0	0	0	0	508
ScreenPorch	2919	16.0624	56.1844	0	0	0	0	576
PoolArea	2919	2.2518	35.6639	0	0	0	0	800
MiscVal	2919	50.826	567.402	0	0	0	0	17000
MoSold	2919	6.21309	2.71476	1	4	6	8	12
YrSold	2919	2007.79	1.31496	2006	2007	2008	2009	2010
SalePrice	2919	180053	57381.6	34900	154795	176735	191896	755000

Şekil 4.2. Sayısal verilerin özellikleri (ortalama, minimum, maksimum) değerleri

Yukarıdaki Şekil (4.1 ve 4.2) ‘de görüldüğü üzere sayısal verilerin özellikleri tablo şeklinde gösterilmektedir.

4.2 Veri Ön işleme

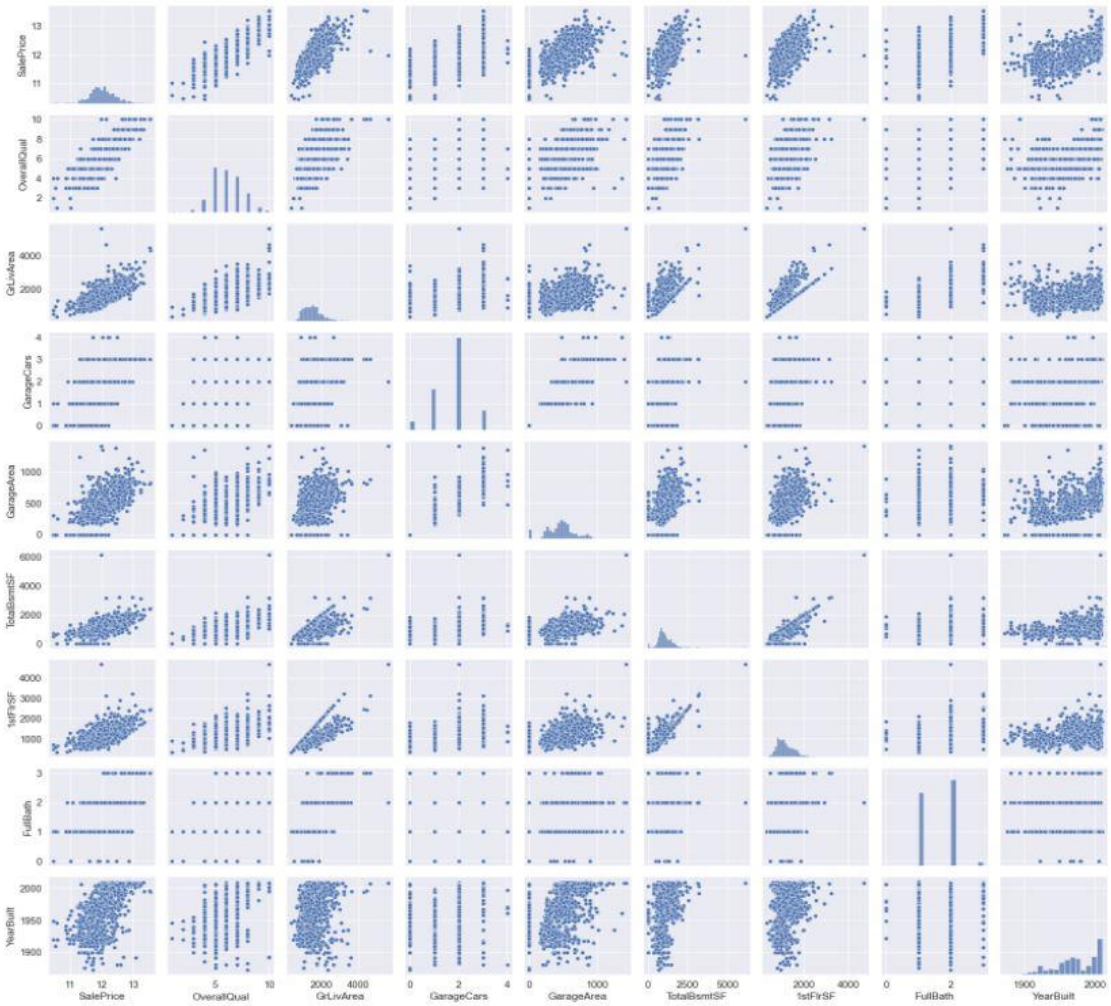
Proje üzerinde gerekli işlemler yapmak için ilk olarak gerekli kütüphaneler eklenmiştir. Veri kümesi kaggle sitesinden 3 farklı (sample_submissions ,test ve train) olarak 3 parça şeklinde eklenmiştir. Veri kümesinde ilk olarak (train) verisindeki korelasyona baktım.



Şekil 4.6. Korelasyon Tablosu

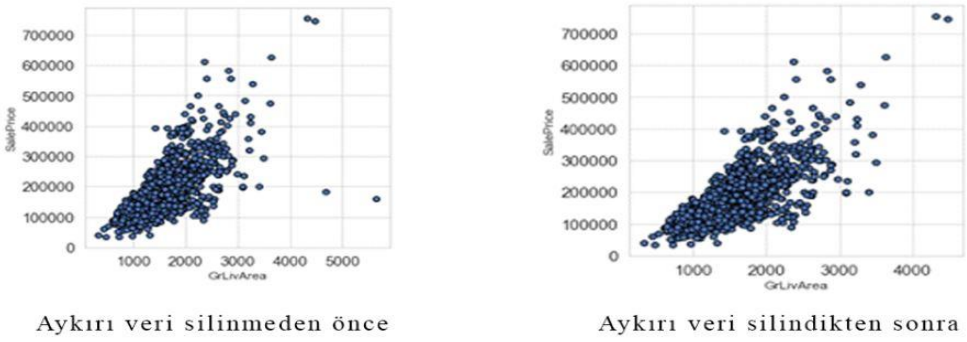
Yukarıdaki Şekil 4.6 ‘da (train) verisinin korelasyon tablosu vardır. Bağımlı değişken olan (SalePrice) evlerin fiyatları olan özniteliğin ile hangi öznitelikler ile arasındaki bağın ne derecede olduğunu gösteriyor.

Veri kümesindeki korelasyon değeri yüksek olan kolonlara aşağıdaki Şekil 4.7 ‘de dağılım tablosu oluşturulmuştur.



Şekil 4.7 Verilerin dağılım tablosu

Yukarıdaki Şekil 4.7'de dağılım tablosundaki özniteliklerin bağımlı değişken olan (SalePrice) evlerin fiyatı olan öznitelik ile orantılı dağılımını gösteriyor. Bu dağılımda aykırı verilere sahip olan (GrLivArea) özniteliğin aykırı verisi silinme işlemi yapıldı.(GrLivArea) özniteliğinin aykırı verisi silinmeden önceki hali ve aykırı veri silindikten sonraki hali aşağıdaki Şekil 4.8'te görülüyor.

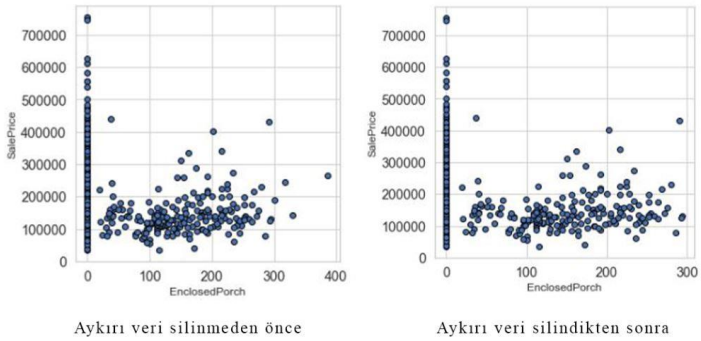


Şekil 4.8. (GrLivArea) özniteliğin grafiği

Veri kümesindeki en iyi korelasyona sahip olan özniteliklerin 5 tanesi tablosundan öznitelikler içinden aykırı veriye sahip olan özniteliği düzenlenmiştir. Daha sonra veri kümesindeki en kötü korelasyon değeri alan 6 tane özniteliğin tablosu aşağıdaki Şekil 4.9'da verilmiştir.

Şekil 4.9 Verilerin dağılım tablosu

Yukarıdaki tabloda aykırı verilere sahip olan (EnclosedPorch) özniteliğin aykırı verisi silinme işlemi yapıldı.(EnclosedPorch) özniteliğinin aykırı verisi silinmeden önceki hali ve aykırı veri silindikten sonraki hali aşağıdaki Şekil 4.10'da görülüyor.



Şekil 4.10. (EnclosedPorch) özniteliğin grafiği

4.4 Veri sayısallaştırma

Veri kümesindeki aykırı verileri sildikten sonra (test ve train) verilerini ayrı bir şekilde kategorik verilerini sayısallaştırma yapılmıştır. Verileri birleştirmeden önce ayrı ayrı yapılmasının nedeni ise (test ve train) 'deki kategorik verilerin yazımında ikisi arasında bazı özelliklerde farklılık olduğu için fazladan değişken saymasından dolayı ayrı ayır işlem yapılmıştır. Verileri kategorik verilerden arındırdıktan sonra verileri birleştirme işlemi yapılışmıştır.

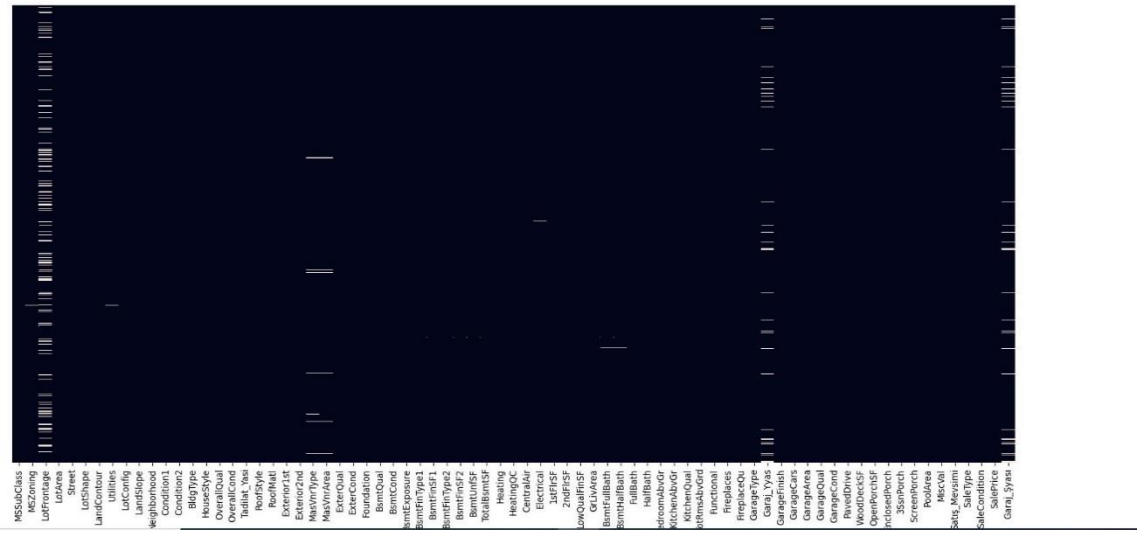
Kategorik Özellikler		Sayısallaştırılmış Özellikler	
Index	Condition1	Index	Condition1
Norm	2498	2	2498
Feedr	162	1	162
Artery	90	0	90
RRAn	50	6	50
PosN	38	4	38
RRAe	28	5	28
PosA	20	3	20
RRNn	9	8	9
RRNe	6	7	6
Index	MSZoning	Index	MSZoning
3	2250	RL	2250
4	458	RM	458
1	139	FV	139
0	25	C (all)	25
2	25	RH	25
Index	Street	Index	Street
Pave	2889	1	2889
Grvl	12	0	12

Şekil 4.12. Veri kümesindeki kategorik özniteliklerden bir kesit

Yukarıdaki şekil 4.12 'de veri kümesinden bazı kategorik özelliklerin sayısal hale çevrilmiş halini gösteren bir tablo gösterilmektedir.

4.5 Eksik verilerin silinmesi

Veri kümesini bileştirme işleminden sonra aşağıdaki Şekil 4.13 ‘ de veri kümesinin ısı haritası ile eksik verilerin hangi özelliklere ait olduğunu görülüyor. Veri kümesinde eksik veri sayısı 2000’den fazla olan özellikler (Alley , PoolQC, Fence, MiscFeature) özellikleri veri kümesinden silinme işlemi yapılmıştır. Veri kümesinde kolonlar ile bağlantısı olmayan (Id) özelliği veri kümesinden silinmiştir.



Şekil 4.13. eksik verileri ısı haritası ile gösterimi

4.6 Özellik Çıkartma

Veri kümesinde tarihler ile olan özellikler regresyon analizinde direkt kullanması uygun olmadığı için o özellik çıkarımı yapılmıştır. İlk olarak(YearBuilt) evin yapılış yılından (YearRemodAdd) evin tadilat yılından çıkartılarak evin tadilat yaşını buldum. Evin tadilat yaşını yeni özellik olarak (TadilatYaşı) diye tanımlandı. Veri kümesinden (GarageYrBlt) olan özellik ise evin garajının yapılış yılını tutuyor. Evin yapılış yılı olan (YearBuilt) den evin garajının yapılış yılından (GarageYrBlt) çıkartılarak evin garajının yaşını buldum. Evin garajının yaşını yeni özellik olarak (GarajYeniY) diye tanımlandı. Veri kümesinden yıl verisi olarak (YearBuilt, YearRemodAdd, GarageYrBlt) verilerini çıkarılmıştır.

4.7 Boş verileri doldurma işlemi

Veri kümesindeki boş verileri doldurmak için 3 farklı yöntem ile veri doldurma işlemi yapılmıştır.

Bunlar (mice, knn) algoritmaları ve ortalama ile boş olan veriler doldurulmuştur.

KNN

Index	MSSubClass	MSZoning	LotFrontage
6	20	3	75
7	60	3	84.0588
8	50	4	51
9	190	3	50
10	20	3	70
11	60	3	85
12	20	3	70.1765
13	20	3	91
14	20	3	83.6471
15	45	4	51
16	20	3	78.1765
17	90	3	72

MİCE

Index	MSSubClass	MSZoning	LotFrontage
6	20	3	75
7	60	3	78.5831
8	50	4	51
9	190	3	50
10	20	3	70
11	60	3	85
12	20	3	67.2793
13	20	3	91
14	20	3	81.1076
15	45	4	51
16	20	3	77.6827
17	90	3	72

ORTALAMA

Index	MSSubClass	MSZoning	LotFrontage
6	20	3	75
7	60	3	68.9988
8	50	4	51
9	190	3	50
10	20	3	70
11	60	3	85
12	20	3	68.9988
13	20	3	91
14	20	3	68.9988
15	45	4	51
16	20	3	68.9988
17	90	3	72

Şekil 4.14. Veri kümesinden bir kesit 3 farklı yöntem ile eksik veri doldurma

Yukarıdaki Şekil 4.14'den veri kümesinden bir kesit alarak (LotFrontage) özelliğindeki boş olan verileri 3 farklı yöntem ile farklı sonuçlar ile doldurulmuş olduğunu görüyoruz.

Veri kümesinde boş olan sütunların haricinde tamamıyla dolu olan sütunlarında ayrı şekilde veri olarak alınmıştır. Dolu olan veri seti 1744 satır ve 74 kolona sahipken, (knn,mice, ortalama) ile doldurulan veriler ise 2901 satır ve 74 kolondan oluşmaktadır.

4.9 Normalizasyon İşlemi

Veri kümesinde çok farklı aralıkta değer alan özellikler bulunmaktadır. Bu nedenle verileri normalizasyon işlemine tabi tutularak 0-1 aralığına indirgenmiştir. Ama bu işlemi kategorik sütunlar ile tarih sütunlarını eklenmemiştir.

4.10 Öznitelik Seçimi

Öznitelik seçimi için sarmal (wrapper) yönteminden adımsal seçim (stepwise (exhaustive) selection) kullanılmıştır. Veri kümemizden 74 tane özniteliğin içinden tahmin edeceğimiz evlerin fiyatları hariç (GrLivArea, OverallQual, LotArea, TotalBsmtSF, ExterQual, BldgType, BsmtFinSF1, MiscVal, BsmtQual, LandContour, RoofMatl, Functional, MasVnrType, MasVnrArea, KitchenAbvGr, TotRmsAbvGrd, GarageArea, PoolArea, MoSold, OverallCond, TadilatYaşı, KitchenQual, Heating, GarajYeniY, GarageQual, Foundation, GarageCond, PavedDrive) öznitelikler yer alıyor. Seçilen özniteliklerden (TadilatYaşı ,GarajYeniY) öznitelik çıkarımından oluşturduğumuz özniteliktir. Toplam 28 tane öznitelik seçimi yapmıştır.

4.11 Performans Ölçütleri

Proje üzerinde veri kümesinin modeller üzerinde değerlendirilmesi için ortalama mutlak hata (Mean absolute error) ve r kare hatası(R Squared, R^2) ölçütleri kullanılmıştır.

4.11.1 R Kare Hata (R Squared, R^2)

R kare hatası bu metrik bağımsız değişkenler ile bağımlı değişkenler etkilediği durumların yüzdesel hacmini açıklıyor. R kare hatasının değeri 0 ile 1 arasında değişebilir. Modelin tahminlemesi 1' e yaklaşması daha iyi performans gösterir.

4.11.2 Ortalama Mutlak Hata (Mean absolute error)

Ortalama mutlak hata iki sürekli değişken arasındaki farkın ölçüsüdür. MAE, her gerçek değer ile veriye en iyi uyan çizgi arasındaki ortalama dikey mesafedir. MAE aynı zamanda her veri noktası ile en iyi uyan çizgi arasındaki ortalama yatay mesafedir. MAE değeri kolay yorumlanabilir olduğu için regresyon ve zaman serisi problemlerinde sıkça kullanılmaktadır. MAE, yönlerini dikkate almadan bir dizi tahmindeki hataların ortalama büyüklüğünü ölçen, tüm tekil hataların ortalamada eşit olarak ağırlıklandırıldığı doğrusal bir skordur. MAE değeri 0'dan ∞'a kadar değişebilir. Negatif yönelimli puanlar yani daha düşük değerlere sahip tahminleyiciler daha iyi performans gösterir.

4.12 Deneysel Çalışma ve Sonuçlar

Bu çalışmada ilk olarak sarmal(wrapper) yönteminden adımsal seçim (Stepwise Selection) kullanılmadan ve kullanıldıktan sonraki modelin doğruluk değerlerini çeşitli veri doldurma yöntemleri (knn,mice ,ortalama) veriler ile ve boş olan sütunların çıktığı dolu veriler üzerinden görmek için doğrusal (lineer) regresyon yöntemi uygulanmıştır.

Çizelge 4.1 Veriler üzerinde adımsal seçim (Stepwise Selection) yöntem kullanılmamış sonuçlar

	Ortalama	Knn	Mice	Dolu Veri
Test(R Kare)	**0.4768**	0.4710	0.4680	0.4288
Eğitim(R Kare)	**0.5472**	0.5396	0.5379	0.5595
Test(Ortalama Mutlak Hata)	**0.0033**	0.0033	0.0034	0.0039
Eğitim(Ortalama Mutlak Hata)	**0.0028**	0.0029	0.0029	0.0032

Çizelge 4.2 Veriler üzerinde adımsal seçim (Stepwise Selection) yöntem kullanılmamış sonuçlar

	Ortalama	Knn	Mice	Dolu Veri
Test(R Kare)	**0.4952**	0.4898	0.4871	0.4679
Eğitim(R Kare)	**0.5327**	0.5244	0.5240	0.5337
Test(Ortalama Mutlak Hata)	**0.0032**	0.0032	0.0032	0.0037
Eğitim(Ortalama Mutlak Hata)	**0.0029**	0.0030	0.0030	0.0034

Yukarıdaki çizelge 4.1 ve 4.2 'de görüldüğü üzere veri kümesinde adımsal seçim (Stepwise Selection) yöntemi kullanıldığı ve kullanılmadığındaki farkları doğrusal(lineer) regresyon modeli ile değerlendirilmiştir.

Çizelge 4.3 Çeşitli regresyon yöntemleri kullanılmış sonuçlar

		Ortalama	Knn	Mice	Dolu Veri
Lineer	Test (R Kare)	0.5022	0.4956	0.4922	0.4534
	Eğitim (R Kare)	0.5308	0.5233	0.5212	0.5399
	Test (Ortalama Mutlak Hata)	0.0572	0.0576	0.0578	0.0602
	Eğitim(Ortalama Mutlak Hata)	0.0541	0.0545	0.0546	0.0585
Ridge	Test (R Kare)	0.5022	0.4956	0.4922	0.4534
	Eğitim (R Kare)	0.5308	0.5233	0.5212	0.5399
	Test (Ortalama Mutlak Hata)	0.0572	0.0576	0.0578	0.0602
	Eğitim(Ortalama Mutlak Hata)	0.0541	0.0545	0.0546	0.0636
Lasso	Test (R Kare)	0.4404	0.4386	0.4368	0.4466
	Eğitim (R Kare)	0.4374	0.4359	0.4351	0.4567
	Test (Ortalama Mutlak Hata)	0.0607	0.0608	0.0609	0.0600
	Eğitim(Ortalama Mutlak Hata)	0.0592	0.0593	0.0593	0.0636
Karar Ağacı(Decision Tree)	Test (R Kare)	0.6952	0.6747	0.7523	0.7124
	Eğitim (R Kare)	0.9999	0.9999	0.9999	1.0
	Test (Ortalama Mutlak Hata)	0.0448	0.0463	0.0404	0.0437
	Eğitim(Ortalama Mutlak Hata)	0.0001	0.0001	0.0001	0.0
Rastgele Orman((Random Forest)	Test (R Kare)	0.8762	0.8640	0.8801	0.8470
	Eğitim (R Kare)	0.9802	0.9813	0.9812	0.9778
	Test (Ortalama Mutlak Hata)	0.0285	0.0299	0.0281	0.0318
	Eğitim(Ortalama Mutlak Hata)	0.0110	0.010	0.0108	0.0128
XGBOOST	Test (R Kare)	**0.8930**	0.8905	0.8889	0.8146
	Eğitim (R Kare)	**0.9699**	0.9670	0.9525	0.9517
	Test (Ortalama Mutlak Hata)	**0.0265**	0.0268	0.0270	0.0350
	Eğitim(Ortalama Mutlak Hata)	**0.0137**	0.0143	0.0172	0.0189

Değerlendirme sonucunda adımsal seçim (Stepwise Selection) yöntemlerin daha iyi sonuç alındığı görülmüştür. Yukarıda Çizelge 4.3 ‘te Sarmal (wrapper) yönteminler’inden adımsal seçim (Stepwise Selection) yöntemi kullanılarak (Doğrusal (Lineer), Ridge, Lasso, Karar Ağacı (Decision Tree), Rastgele Orman ((Random Forest)) Regresyonları uygulanmıştır. Çeşitli regresyon modellerinin çeşitli veri doldurma yöntemleri (knn,mice ,ortalama) veriler ile ve boş olan sütunların çıktığı dolu verilerin sonuçları görülüyor.

Çalışma üzerinde çeşitli regresyon modellerinden sonra regresyon tabanlı yapay sinir ağı uygulanmıştır.

Regresyon tabanlı yapay sinir ağı modelini uygulanmıştır.

Çizelge 4.4 Regresyon Tabanlı Yapay Sinir Ağı kullanılmış sonuçlar

	Ortalama	Knn	Mice	Dolu Veri
R Kare	0.8071	**0.8710**	0.8451	0.7169
Ortalama Mutlak Hata	0.3228	**0.01906**	0.0235	0.0265

Yukarıdaki Çizelge 4.4’de görüldüğü üzere regresyon tabanlı yapay sinir ağının model üzerindeki başarı oranları verilmiştir.

5.SONUÇLAR VE BULGULAR

Evlerin fiyatlarının tahmin edilmesi günümüzde hala güncelliğini koruyan, pek çok farklı disiplin için önemli olan bir çalışma alanıdır. Bu çalışmada, farklı makine öğrenmesi algoritmaları ve regresyon tabanlı yapay sinir ağı ile evlerin fiyatları tahmini yapılmıştır. Farklı modeller geliştirilerek en iyi performans gösteren modeller uygulanmıştır. Çalışma kapsamında konut fiyatlarının tahmin etmek amacıyla doğrusal regresyon, (lasso) regresyon, (ridge) regresyon, Karar ağacı regresyon, rastgele orman regresyon, (XGBoost)regresyon ve regresyon tabanlı yapay sinir ağı kullanılmıştır. Kullanılan modeller 3 farklı veri doldurma (knn,mice,ortalama) verisi ve eksik verilerin olmadığı dolu veri üzerinde denenmiştir. Veriler içinde en yüksek başarılı veri Ortalama ile doldurulan veri olduğu görülmüştür. Modeller içinde en yüksek doğruluğa sahip olan (XGBoost) Regresyon modeli üzerinde en az hata ile en başarılı yöntem olduğu görülmüştü

KAYNAKLAR

1. G. Z. Fan, S. E. Ong, and H. C. Koh. (2006). *Determinants of house price: A decision tree approach.*.
2. F.Ecer. (2014). *Türkiye'deki konut fiyatlarının tahmininde hedonik regresyon*. basım yeri bilinmiyor : Proceedings of the International Conference on Eurasian Economies, Cilt no 7. 1-10 .
3. Şahin, O. Özsoy and H. (2009). *"Housing price determinants in Istanbul, Turkey: An application of the classification and regression tree model International Journal of Housing Markets and Analysis.* , Cilt vol. 2(2),167-178.
4. Harding, G. M. Goldberg and J. P. (2003). *Investment characteristics of low- and moderate-income mortgage loans.*, basım yeri bilinmiyor : Journal of Housing Economics, , Cilt vol. 12, 51-160.
5. Bin, O. (2004). *A prediction comparison of housing sales prices by parametric versus semi-parametric regressions.* basım yeri bilinmiyor : Journal of Housing Economics, , Cilt vol. 13(1) ,68-84.
6. Fitöz, N. Öztürk and E. (2009). *Türkiyede konut piyasasının belirleyicileri: Ampirik bir uygulama.* basım yeri bilinmiyor : ZKU Journal of Social Sciences,, Cilt vol. 5(10), 21-46.
7. Türk, M. F. Tuna and T. (2010). *Lineer regresyon ve coğrafi bilgi sistemleri yardımıyla ev fiyatlarının tahmin edilmesi.* basım yeri bilinmiyor : O. Kitapçı, 10-22.
8. Fix, E, ve J. Hodges. (1951).*An important contribution to nonparametric discriminant analysis and density estimation.* basım yeri bilinmiyor : International Statistical Review , Cilt 3(57), 233-238
9. Ho, T. K. (1995). *Random decision forests. proceedings of the third international conference on document analysis and recognition.* 278-282.
10. Chen, T. ve C. Guestrin. (2016). *Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* basım yeri bilinmiyor : ACM.
11. Sundaram, R. B. (2019). *Understanding the Math behind the XGBoost Algorithm.*
12. Swetapadma, A. ve A. Yadav. (2017). *A Novel Decision Tree Regression-Based Fault Distance Estimation Scheme for Transmission Lines.* basım yeri bilinmiyor : IEEE Transactions on Power Delivery, 234-245.
13. Breiman, L. (2017). *Classification and regression trees, Routledge*.
14. Loh, W. Y. (2011). *Classification and regression trees." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* . 14-23.
15. Miller, AJ. (1990). *Subset Selection in Regression , Chapman & Hall.*

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : Özer, Ramazan
Uyruğu : Türk
Doğum tarihi ve yeri : 01.01.2000 Gaziantep
Telefon : +(90)5433701917
e-mail : ramazan107127@gmail.com

Eğitim

Derece	**Eğitim Birimi**	**Mezuniyet Tarihi**
Lisans	Kahramanmaraş Sütçü İmam Üniversitesi	
Lise	Hasan Ali Yücel Anadolu Lisesi	
Ortaokul	Alparslan OrtaOkulu	

Yabancı Dil

İngilizce

Açık Rıza Beyanı	**İmza**
Bu tezde verdiğim kişisel bilgilerimin Kahramanmaraş Sütçü İmam Üniversitesi ve birimlerince işlenmesine açık bir şekilde rıza gösterdiğimi kabul ve beyan ederim.	☐ (Pandemi süreci için kutuyu onaylamanız yeterlidir)


T.C.
1992
KAHRAMANMARAŞ
SÜTÇÜ İMAM ÜNİVERSİTESİ