



VERİ ANALİTİĞİ PROJE RAPORU

PARİS AIRBNB İLANLARINDA GECELİK FİYAT TAHMİNİ

Regresyon Modelleri ile Karşılaştırmalı Bir Uygulama

2026

Hazırlayan: Feyza Nur Demirbaş

İÇİNDEKİLER

1. Giriş
 - 1.1 Projenin Amacı
 - 1.2 Projenin Kapsamı
2. Veri Kaynağı ve Yöntem
 - 2.1 Kullanılan Veri
 - 2.2 Kullanılan Değişkenler
 - 2.3 Analiz Akışı
3. Veri Temizleme ve Hazırlama
 - 3.1 Eksik Değerler
 - 3.2 Aykırı Değer Analizi
 - 3.3 Eğitim/Test Ayrımı ve Ön İşleme
4. Keşifsel Veri Analizi
 - 4.1 Fiyat Dağılımı
 - 4.2 Oda Tipi ve Fiyat
 - 4.3 Korelasyon Analizi
5. İstatistiksel Testler
6. Makine Öğrenmesi
 - 6.1 Temel Modeller
 - 6.2 Overfitting Kontrolü
 - 6.3 Veri Miktarı Deneyi
 - 6.4 GridSearchCV ve Final Model
7. Model Sonuçları
 - 7.1 Model Karşılaştırması
 - 7.2 Özellik Önemleri
8. Sonuç
9. Kısıtlamalar ve Öneriler
- Kaynakça

1. GİRİŞ

Kısa dönemli konaklama platformlarında ilan fiyatları; konum, konaklama kapasitesi, oda tipi, yorumlar ve müsaitlik gibi birçok değişkenden etkilenmektedir. Bu nedenle fiyat tahmini, tek bir değişkenle açıklanabilecek basit bir problem değildir. Bu projede Paris'teki Airbnb ilanları incelenmiş ve gecelik fiyatı tahmin etmek amacıyla farklı regresyon modelleri karşılaştırılmıştır.

Çalışmada yalnızca yüksek bir tahmin skoru elde etmek değil, modelin yeni verilere ne kadar iyi genelleştirebildiğini görmek de amaçlanmıştır. Bu nedenle eğitim ve test sonuçları birlikte değerlendirilmiş, özellikle ağaç tabanlı modellerde aşırı öğrenme (overfitting) kontrol edilmiştir.

1.1. Projenin Amacı

Projenin temel amacı, Paris Airbnb ilanlarının gecelik fiyatlarını açıklayan değişkenleri incelemek ve bu değişkenlerden yararlanarak fiyat tahmini yapabilen bir regresyon modeli oluşturmaktır. Bunun yanında farklı model türlerinin aynı veri üzerindeki performansları karşılaştırılmış ve en dengeli model belirlenmeye çalışılmıştır.

1.2. Projenin Kapsamı

Analiz Paris şehri ile sınırlandırılmıştır. Çalışmada fiyatı bulunmayan ilanlar modellemeye dahil edilmemiş; mahalle, konum, oda tipi, kapasite, yatak ve banyo bilgileri, yorum değişkenleri, minimum konaklama süresi ve yıllık müsaitlik gibi özellikler kullanılmıştır. Sonuçlar bu veri setinin kapsadığı dönem ve ilanlar için değerlendirilmelidir.

2. VERİ KAYNAĞI VE YÖNTEM

2.1. Kullanılan Veri

Veri kaynağı olarak Inside Airbnb tarafından yayımlanan Paris'e ait 16 Haziran 2026 tarihli detaylı listings veri seti kullanılmıştır. Ham veri 77.679 satır ve 90 sütundan oluşmaktadır. Tekrarlı tam satır bulunmamıştır. Fiyat alanı sayısal forma dönüştürüldükten ve fiyatı eksik ya da sıfırdan küçük/eşit kayıtlar çıkarıldıktan sonra analiz için 48.402 ilan kalmıştır.

Temizleme öncesinde veri setinde bazı sütunlarda belirgin eksiklikler bulunmaktadır. Özellikle reviews_per_month ve review_scores_rating değişkenlerinde yaklaşık %17, bedrooms değişkeninde %16 ve bathrooms değişkeninde %12 oranında eksik değer vardır.

2.2. Kullanılan Değişkenler

Değişken	Açıklama
price_eur	Hedef değişken; ilanın gecelik fiyatı (€)
neighbourhood_cleansed	İlanın bulunduğu Paris mahallesi/bölgesi
latitude / longitude	Coğrafi konum bilgileri
room_type	Konaklama türü
accommodates	İlanın ağırlayabildiği kişi sayısı
bedrooms / bathrooms / beds	Oda, banyo ve yatak bilgileri
number_of_reviews	Toplam yorum sayısı
reviews_per_month	Aylık ortalama yorum sayısı
review_scores_rating	Ortalama değerlendirme puanı
minimum_nights	Minimum konaklama süresi
availability_365	Yıl içindeki müsait gün sayısı

2.3. Analiz Akışı

İş akışı şu şekilde oluşturulmuştur: veri yükleme → temel temizlik → eksik değer analizi → aykırı değer analizi → keşifsel veri analizi → istatistiksel testler → eğitim/test ayrımı → ön işleme → modelleme → hiperparametre optimizasyonu → model karşılaştırması ve sonuçların yorumlanması.

3. VERİ TEMİZLEME VE HAZIRLAMA

3.1. Eksik Değerler

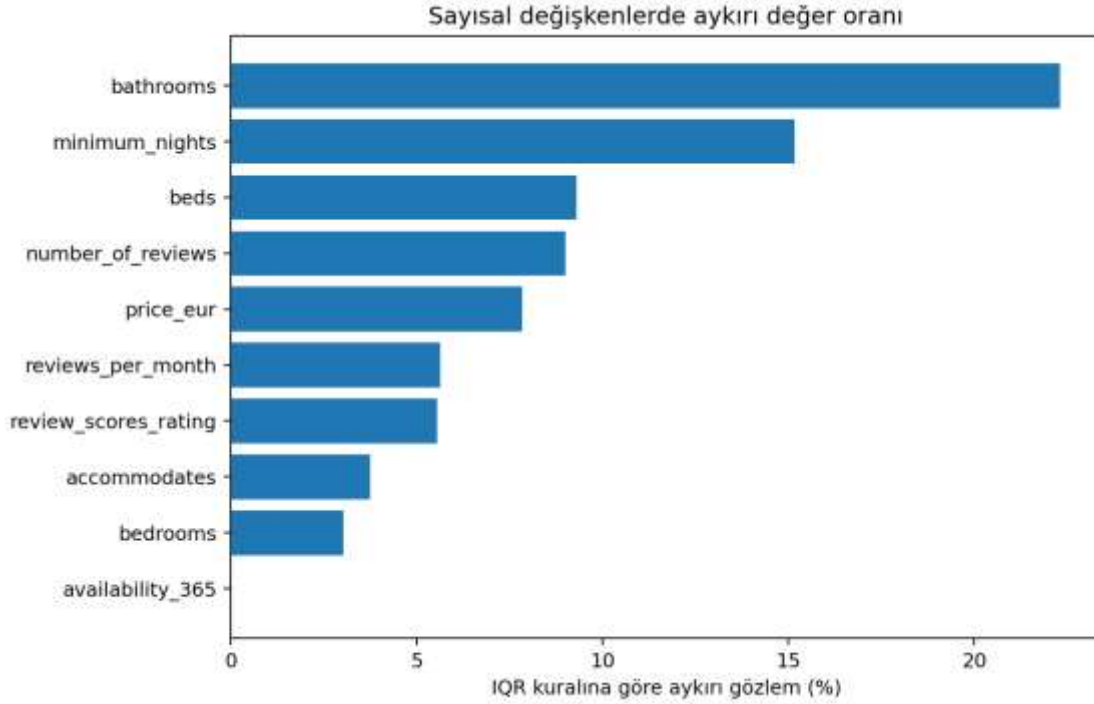
Eksik değerler doğrudan tüm veri üzerinden doldurulmamıştır. Veri eğitim ve test kümelerine ayrıldıktan sonra sayısal değişkenlerde medyan, kategorik değişkenlerde ise en sık görülen değer yalnızca eğitim verisinden öğrenilmiştir. Böylece test verisinin bilgisi modelin eğitim aşamasına taşınmamıştır.

3.2. Aykırı Değer Analizi

Sayısal değişkenlerde aykırı değerleri görmek için IQR yöntemi kullanılmıştır. Ancak IQR ile aykırı görünen her gözlem otomatik olarak silinmemiştir. Airbnb verisinde yüksek kapasiteli veya uzun minimum konaklama süresine sahip gerçek ilanlar bulunabildiği için, istatistiksel aykırılık ile veri hatası birbirinden ayrı değerlendirilmiştir.

Değişken	IQR Outlier %	Gözlenen maksimum
price_eur	7,85	97.003,05 €
accommodates	3,76	16
bedrooms	3,02	33
bathrooms	22,32	42
beds	9,31	32
number_of_reviews	9,00	4.589
reviews_per_month	5,64	55,49
minimum_nights	15,17	365
availability_365	0,00	365

Bathrooms değişkeni bu noktada dikkat çekmektedir. Q1 ve Q3 değerlerinin ikisi de 1 olduğu için IQR sıfır çıkmakta ve 1'den farklı birçok gerçek değer aykırı olarak işaretlenmektedir. Bu nedenle IQR sonuçları önce tanısal olarak incelenmiş, açıklayıcı değişkenlerde körlemesine IQR kırpması yapılmamıştır.



Şekil 1. IQR kuralına göre sayısal değişkenlerde aykırı değer oranları

Hedef fiyat için model kapsamı eğitim setinden hesaplanan IQR sınırına göre belirlenmiştir. Bu sınır yaklaşık 0–658,98 €'dur. Açıklayıcı değişkenlerde ise yalnızca çok uç değerlerin etkisini azaltmak için eğitim verisinin %99 üst yüzdelik sınırı kullanılmıştır.

3.3. Eğitim/Test Ayrımı ve Ön İşleme

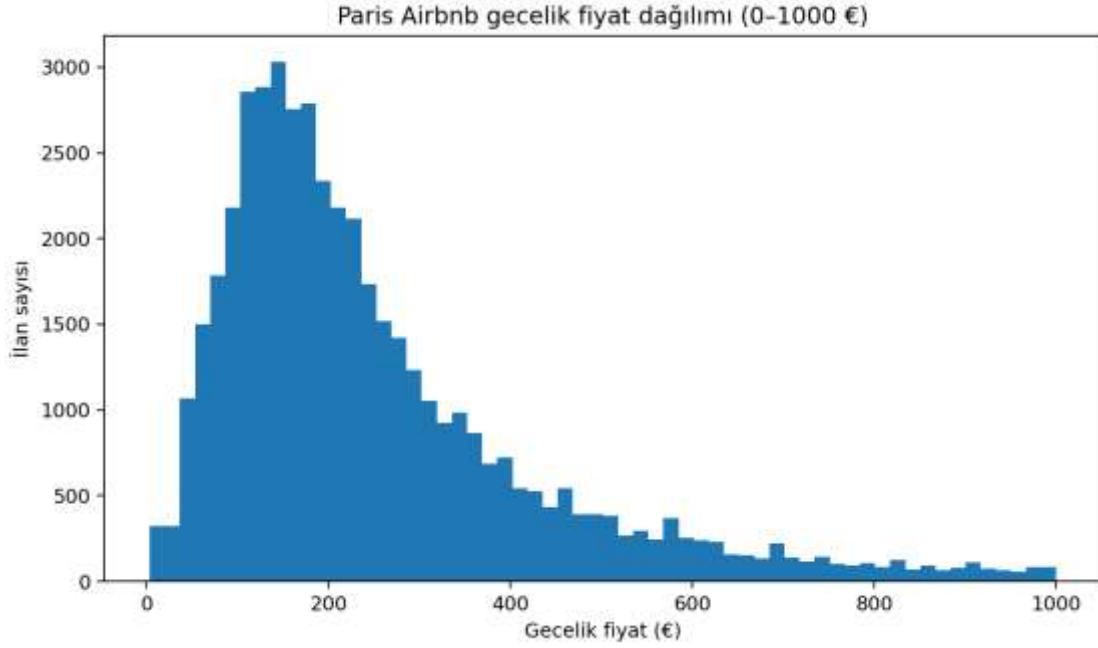
Veri %80 eğitim ve %20 test olacak şekilde, random_state=42 kullanılarak ayrılmıştır. Fiyat IQR kapsamı uygulandıktan sonra modelleme aşamasında 35.680 eğitim ve 8.926 test gözlemi kullanılmıştır. Ön işleme adımlarının tamamı eğitim setinde öğrenilmiş ve test setine aynı şekilde uygulanmıştır.

Kategorik değişkenler One-Hot Encoding yöntemiyle sayısal forma dönüştürülmüştür. Doğrusal ve düzenlenileştirilmiş modellerde StandardScaler kullanılmıştır. Random Forest ve Decision Tree gibi ağaç tabanlı modeller için ayrıca ölçekleme gerekmemiştir.

4. KEŞİFSEL VERİ ANALİZİ

4.1. Fiyat Dağılımı

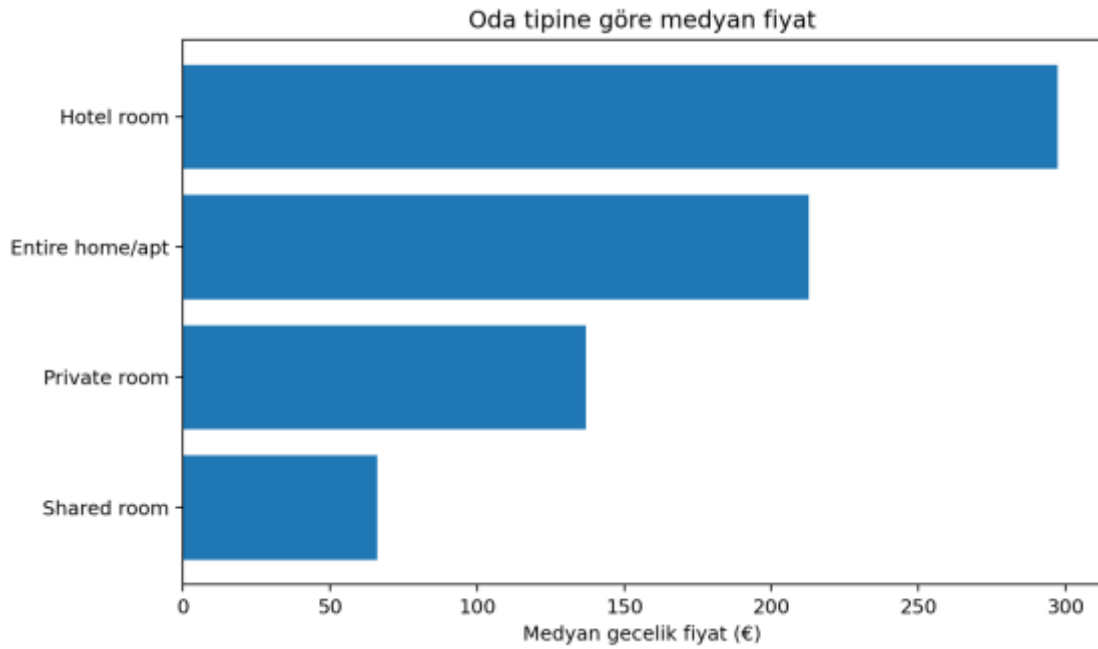
Ham fiyat değişkeni belirgin biçimde sağa çarpık bir dağılıma sahiptir. Medyan gecelik fiyat 205,50 €, ortalama ise 321,23 €'dur. Ortalama değer medyanın yüksek olması, az sayıda çok yüksek fiyatlı ilanların dağılımı yukarı çektiğini göstermektedir. Ham veride maksimum fiyat 97.003,05 €'ya kadar çıkmaktadır.



Şekil 2. Paris Airbnb gecelik fiyat dağılımı (grafik okunabilirliği için 0–1000 € aralığı gösterilmiştir)

4.2. Oda Tipi ve Fiyat

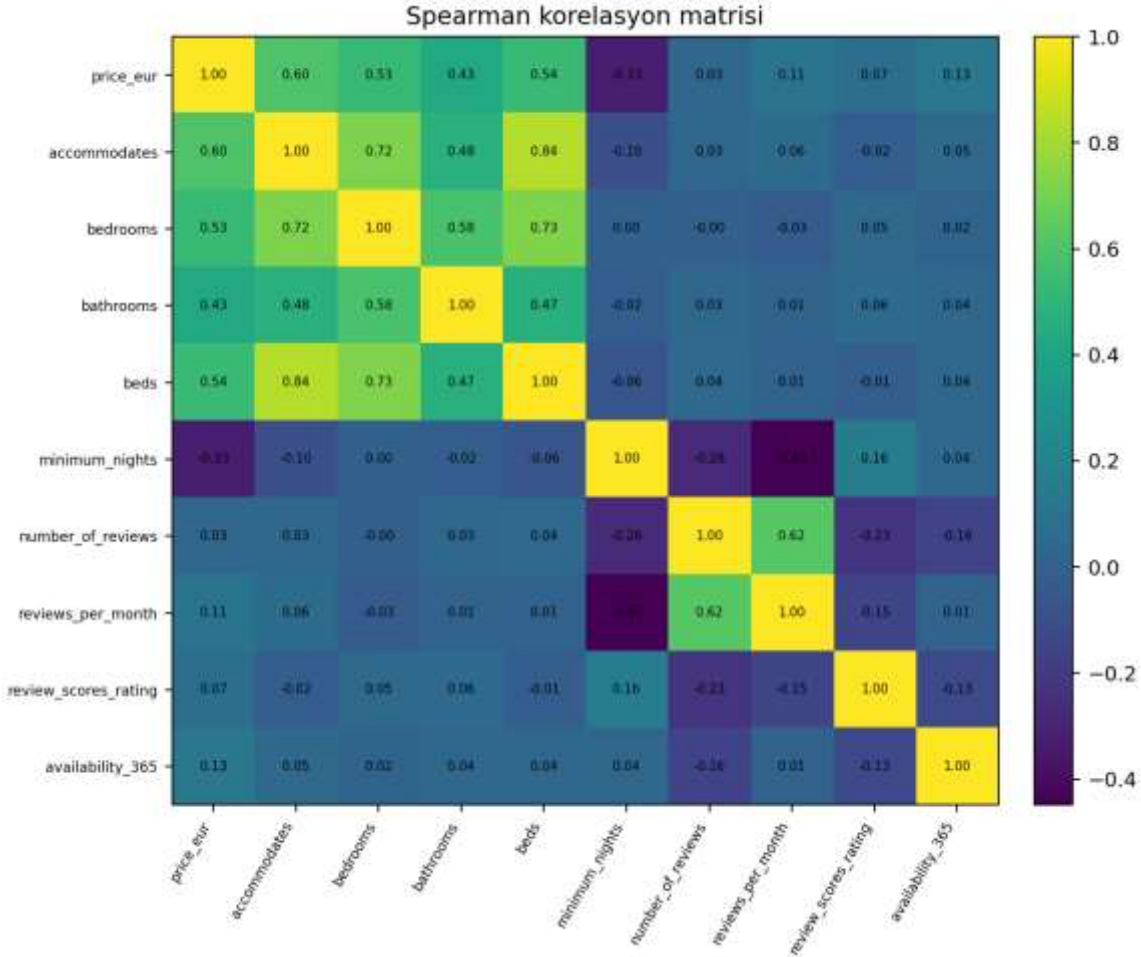
Veri setindeki ilanların büyük bölümü 'Entire home/apt' türündedir. Bu grupta 43.131 ilan bulunurken Private room 4.848, Hotel room 310 ve Shared room 113 ilandan oluşmaktadır. Oda tipleri arasında fiyat düzeylerinin farklılaştığı görülmektedir.



Şekil 3. Oda tiplerine göre medyan gecelik fiyat

4.3. Korelasyon Analizi

Fiyat dağılımının normal olmaması ve bazı değişkenlerde uç değerlerin bulunması nedeniyle sayısal ilişkiler Spearman korelasyonu ile incelenmiştir. Fiyat ile en belirgin pozitif ilişkiler accommodates ($\rho \approx 0,60$), beds ($\rho \approx 0,54$), bedrooms ($\rho \approx 0,53$) ve bathrooms ($\rho \approx 0,43$) değişkenlerinde görülmüştür. Minimum nights ile fiyat arasında ise orta düzeyde negatif ilişki vardır ($\rho \approx -0,33$).



Şekil 4. Sayısal değişkenler için Spearman korelasyon matrisi

Bu sonuçlar, konaklama kapasitesi ve fiziksel büyüklüğü temsil eden özelliklerin fiyatla daha güçlü ilişkili olduğunu göstermektedir. Yorum sayısı ve değerlendirme puanı gibi değişkenlerin tek başına fiyatla ilişkisi daha zayıftır.

5. İSTATİSTİKSEL TESTLER

Fiyat değişkeninin normal dağılıma uygunluğu Shapiro-Wilk testi ile 5.000 gözlemlik örneklem üzerinde kontrol edilmiştir. Büyük veri setlerinde testin küçük sapmalara dahi hassas olması nedeniyle örneklem kullanılmıştır. Sonuç fiyatın normal dağılmadığını göstermiştir. Bu nedenle sayısal değişkenler için Spearman korelasyonu, kategorik gruplar için Kruskal-Wallis testi tercih edilmiştir.

Oda tipi ve mahalle grupları ile fiyat arasında istatistiksel olarak anlamlı farklılıklar bulunmuştur. Sayısal değişkenlerde ise özellikle kapasite, yatak odası, banyo ve yatak sayısının fiyatla anlamlı ilişkisi vardır.

Bununla birlikte, büyük örneklem nedeniyle p-değerinin tek başına yeterli olmadığı dikkate alınmış ve ilişkinin gücü korelasyon katsayılarıyla birlikte yorumlanmıştır.

6. MAKİNE ÖĞRENMESİ

6.1. Temel Modeller

Fiyat tahmini için Linear Regression, Ridge, Lasso, ElasticNet, Decision Tree, Gradient Boosting ve Random Forest modelleri karşılaştırılmıştır. Modeller Test R^2 , RMSE ve MAE değerleriyle değerlendirilmiştir. Ayrıca ağaç tabanlı modellerde eğitim ve test R^2 farkı hesaplanarak aşırı öğrenme kontrol edilmiştir.

6.2. Overfitting Kontrolü

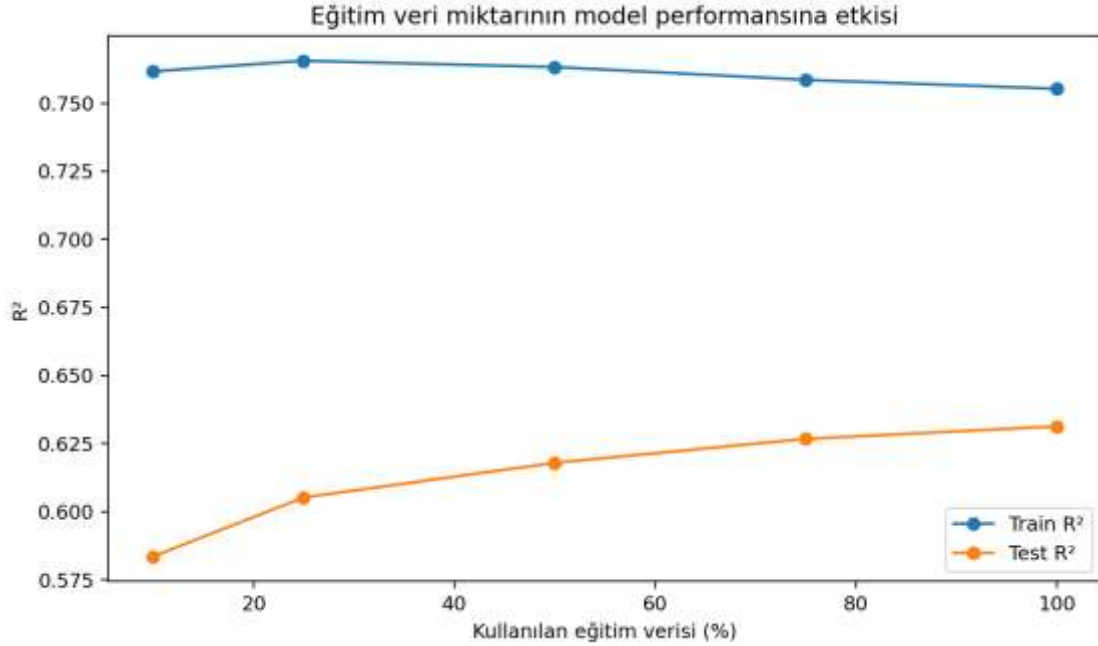
İlk denemelerde sınırlandırılmamış Decision Tree eğitim verisini neredeyse tamamen öğrenirken test performansı belirgin biçimde düşmüştür. Bu nedenle final çalışmada Decision Tree; max_depth=8, min_samples_split=10 ve min_samples_leaf=5 ile sınırlandırılmıştır. Kontrollü modelde Train R^2 yaklaşık 0,590, Test R^2 ise 0,547 olarak elde edilmiştir. Böylece eğitim-test farkı önemli ölçüde azaltılmıştır.

Random Forest için de yalnızca en yüksek eğitim skoru hedeflenmemiştir. Model karmaşıklığı max_depth, min_samples_split, min_samples_leaf ve max_features parametreleriyle sınırlandırılmış; hiperparametre seçiminde çapraz doğrulama sonucu ile train-validation farkı birlikte değerlendirilmiştir.

6.3. Veri Miktarı Deneyi

Modelin veri miktarına duyarlılığını görmek için test seti sabit tutulmuş ve Random Forest eğitim verisinin %10, %25, %50, %75 ve %100'ü ile ayrı ayrı eğitilmiştir. Eğitim verisi arttıkça Test R^2 'nin düzenli biçimde yükseldiği görülmüştür.

Eğitim verisi	Train R^2	Test R^2
%10	0,7617	0,5835
%25	0,7656	0,6052
%50	0,7633	0,6179
%75	0,7586	0,6267
%100	0,7553	0,6313



Şekil 5. Eğitim veri miktarının Random Forest performansına etkisi

Bu deneyde veri küçüldükçe test başarısı artmamış, tersine azalmıştır. Dolayısıyla performans sorununun veri setinin fazla büyük olmasından kaynaklandığına dair bir bulgu elde edilmemiştir. Daha fazla eğitim verisi modelin genelleme başarısına katkı sağlamıştır.

6.4. GridSearchCV ve Final Model

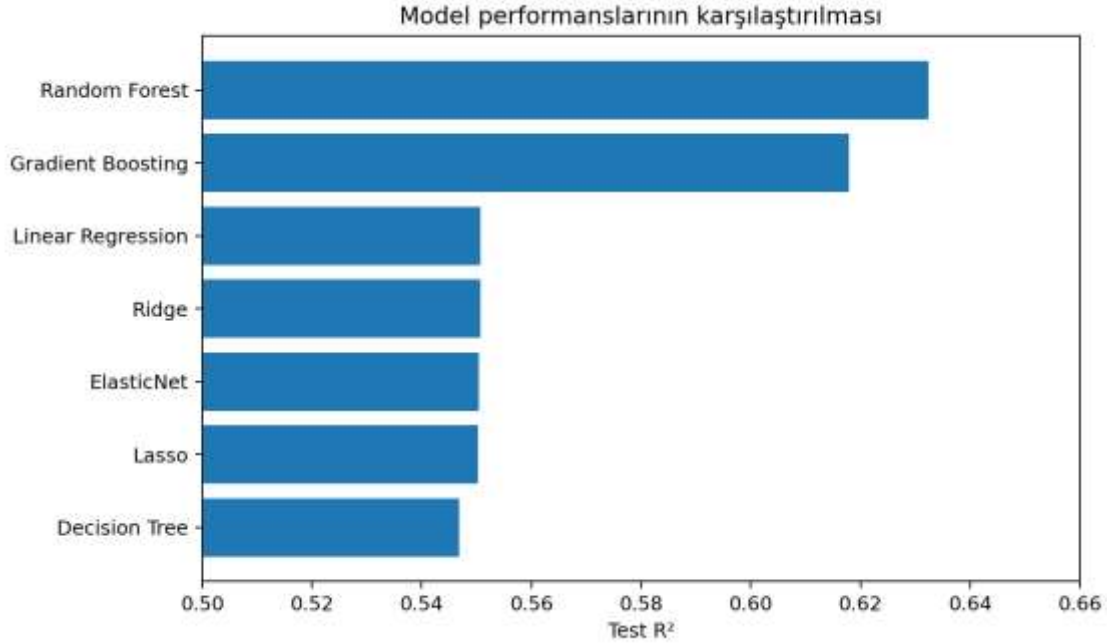
Random Forest hiperparametreleri 5 katlı çapraz doğrulama ile GridSearchCV kullanılarak araştırılmıştır. Hesaplama maliyetini kontrol altında tutmak için hiperparametre araması yalnızca eğitim setinden seçilen 10.000 gözlemlik sabit bir alt kümede yapılmış, test seti bu aşamada kullanılmamıştır. Seçilen parametrelerle final model tüm eğitim verisi üzerinde yeniden eğitilmiştir.

Final Random Forest modelinde $n_estimators=80$, $max_depth=15$, $min_samples_split=12$, $min_samples_leaf=6$ ve $max_features=0,7$ kullanılmıştır. Bu yapı, daha karmaşık modellere göre test başarısını büyük ölçüde korurken eğitim-test farkını daha kontrollü tutmuştur.

7. MODEL SONUÇLARI

7.1. Model Karşılaştırması

Model	Train R²	Test R²	RMSE (€)	MAE (€)
Linear Regression	0,5590	0,5507	90,33	66,31
Ridge	0,5590	0,5507	90,33	66,31
Lasso	0,5586	0,5503	90,36	66,30
ElasticNet	0,5586	0,5504	90,36	66,32
Decision Tree	0,5899	0,5468	90,72	64,98
Gradient Boosting	0,6372	0,6179	83,30	59,78
Random Forest	0,7562	0,6324	81,71	58,06



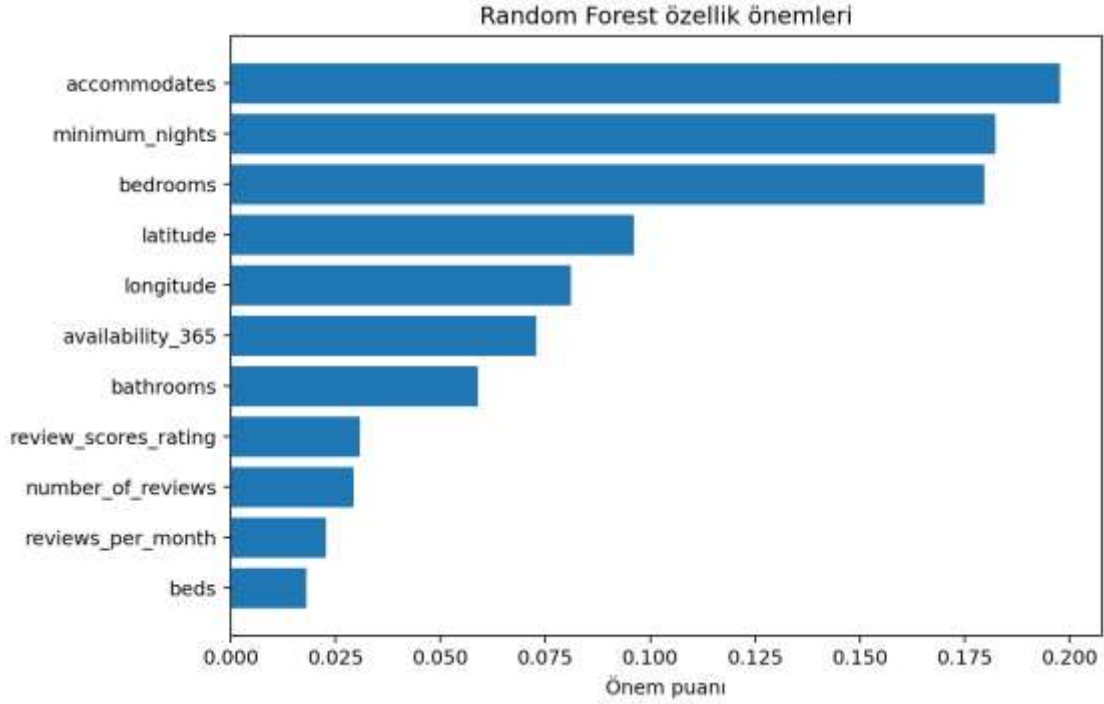
Şekil 6. Modellerin Test R^2 değerleri

En yüksek test performansı Random Forest modelinde elde edilmiştir. Test $R^2=0,6324$ olması, modelin fiyat değişkenindeki değişimin yaklaşık %63'ünü kullanılan özelliklerle açıklayabildiğini göstermektedir. MAE değerinin 58,06 € olması ise test setinde tahminlerin gerçek fiyattan ortalama mutlak sapmasının yaklaşık 58 € olduğunu göstermektedir.

Random Forest'ın Train R^2 değeri 0,7562, Test R^2 değeri 0,6324'tür. Aradaki yaklaşık 0,124'lük fark modelde hâlâ bir miktar eğitim-test ayrışması olduğunu gösterse de, ilk denemelerdeki daha belirgin overfitting'e göre önemli ölçüde azaltılmıştır. Bu nedenle final model seçiminde yalnızca en yüksek eğitim skoru değil, test başarısı ve genelleme dengesi birlikte dikkate alınmıştır.

7.2. Özellik Önemleri

Random Forest özellik önemlerine göre fiyat tahminine en fazla katkı sağlayan değişkenler accommodates, minimum_nights ve bedrooms olmuştur. Bu üç değişkenin ardından konum bilgileri (latitude ve longitude), availability_365 ve bathrooms gelmektedir.



Şekil 7. Final Random Forest modelinde özellik önemleri

Accommodates ve bedrooms değişkenlerinin üst sıralarda olması keşifsel analizde görülen korelasyonlarla uyumludur. Minimum_nights değişkeninin de yüksek önem göstermesi, basit korelasyonda negatif yönde görülen ilişkinin model içinde diğer değişkenlerle birlikte daha belirleyici hale gelebildiğini düşündürmektedir.

8. SONUÇ

Bu çalışmada Paris Airbnb ilanlarının gecelik fiyatları farklı istatistiksel ve makine öğrenmesi yöntemleriyle incelenmiştir. Ham veri 77.679 ilandan oluşmuş, fiyat bilgisi kullanılabilir olan 48.402 ilan üzerinden veri analizi gerçekleştirilmiştir. Fiyat dağılımında belirgin sağ çarpıklık ve birçok sayısal değişkende aykırı değerler görülmüştür.

Aykırı değer analizinde her istatistiksel outlier'ın veri hatası olmadığı dikkate alınmıştır. Bu nedenle IQR yöntemi tespit amacıyla kullanılmış; modelleme aşamasında hedef fiyat için eğitim verisinden öğrenilen IQR kapsamı, açıklayıcı değişkenlerde ise yalnızca çok uç değerleri sınırlayan %99 yüzdelik yaklaşımı tercih edilmiştir.

Model karşılaştırmasında Random Forest en iyi sonucu vermiştir. Final model Test $R^2=0,6324$, RMSE=81,71 € ve MAE=58,06 € değerlerine ulaşmıştır. Eğitim ve test skorları arasında hâlâ fark bulunmakla birlikte, model karmaşıklığının sınırlandırılması ve çapraz doğrulama kullanılmasıyla başlangıçtaki overfitting belirgin biçimde azaltılmıştır.

Eğitim veri miktarı deneyi de önemli bir sonuç vermiştir. Veri küçültüldüğünde modelin test performansı iyileşmemiş, veri miktarı arttıkça Test R^2 yükselmiştir. Bu nedenle bu veri setinde daha az veri kullanmanın overfitting sorununu çözmediği, asıl olarak model karmaşıklığının kontrol edilmesinin daha etkili olduğu görülmüştür.

Genel olarak fiyat üzerinde konaklama kapasitesi, yatak odası sayısı, minimum konaklama süresi ve konum değişkenleri öne çıkmaktadır. Bununla birlikte Test R^2 'nin yaklaşık %63 seviyesinde kalması, fiyatın veri setinde yer almayan başka faktörlerden de etkilendiğini göstermektedir.

9. KISITLAMALAR VE ÖNERİLER

- Çalışma yalnızca Paris ilanlarını ve kullanılan veri setinin kapsadığı dönemi temsil etmektedir; sonuçlar farklı şehir ve dönemlere doğrudan genellenmemelidir.
- Airbnb fiyatları dönemsel talep, etkinlikler, tatiller ve rezervasyon tarihine göre değişebilir. Veri setinde zaman boyutu sınırlı olduğundan bu etkiler modele tam olarak yansımamaktadır.
- Fiyat değişkenindeki çok yüksek uç değerlerin bir kısmı gerçek lüks ilanlar olabilir. Bu nedenle yapılan outlier işlemleri model kapsamını daha istikrarlı hale getirmek için uygulanmıştır; tüm pazarın fiyat dağılımını temsil ettiği düşünülmemelidir.
- Yorum metinleri, ilan açıklaması, fotoğraf kalitesi, ev sahibi özellikleri ve turistik noktalara uzaklık gibi değişkenler modele dahil edilmemiştir. Gelecek çalışmalarda bu özellikler eklenerek tahmin başarısı artırılabilir.
- Random Forest en iyi sonucu vermesine rağmen eğitim ve test skorları arasında tamamen sıfırlanmamış bir fark vardır. Daha geniş hiperparametre aramaları veya farklı boosting yöntemleri gelecekte ayrıca denenebilir.

KAYNAKÇA

[1] Inside Airbnb. Paris listings veri seti, 16 Haziran 2026.

[2] Scikit-learn Documentation. Regression, ensemble methods, preprocessing ve model selection modülleri.

[3] Ders kapsamında kullanılan Keşifsel Veri Analizi, Decision Tree, Random Forest, Ridge-Lasso-ElasticNet ve anomali tespiti notları.