Proje Dokümantasyonu

1. Projenin Amacı

Bu projenin amacı, çeşitli özelliklere dayalı olarak emlak fiyatlarını tahmin etmek için makine öğrenmesi yöntemlerini kullanmaktır. Emlak piyasasında fiyat tahmin sürecini otomatikleştirerek, doğru ve hızlı tahminler yapmayı hedeflemektedir.

---

2. Projenin Nasıl Yapıldığı (Giriş)

Projede, Kaggle platformundan elde edilen https://www.kaggle.com/c/house-prices-advanced-regression-techniques veri seti kullanılmıştır. Bu veri seti, emlak fiyatlarını etkileyen 79 özellik ve tahmin edilmesi gereken hedef değişken olan **"SalePrice"** içerir. Çalışma sürecinde uygulanan yöntemler ve adımlar şu şekilde özetlenmiştir:

1. Veri Analizi ve Temizleme:

- **Eksik veriler** tespit edilip uygun yöntemlerle doldurulmuştur. Sayısal veriler için ortalama, kategorik veriler için ise mod değeri kullanılmıştır.
- **Veri dağılımları** incelenerek, özellikle hedef değişken olan "SalePrice" üzerindeki etkisi olan özellikler belirlenmiştir.
- **Kategorik ve sayısal veriler** uygun şekilde ayrıştırılarak, sayısal verilere dönüştürülmesi gereken kategorik veriler **one-hot encoding** yöntemiyle işlenmiştir.

2. Özellik Seçimi:

- Modelin başarısını artırmak amacıyla, **korelasyon analizi** ve **özellik önem dereceleri** kullanılarak önemli özellikler belirlenmiştir.
- Korelasyon analiziyle, "SalePrice" ile güçlü ilişkisi olan özellikler belirlenmiş ve bunlar modele dahil edilmiştir.
- Ağaç tabanlı modeller (Random Forest, XGBoost) kullanılarak, her özelliğin model üzerindeki etkisi sıralanmıştır.

3. Makine Öğrenmesi Modelleri:

- **Doğrusal Regresyon (Linear Regression):** İlk model olarak kullanılmış ve temel performans değerlendirmeleri yapılmıştır.
- **Destek Vektör Makineleri (SVM):** Karmaşık ilişkileri modellemek için SVM kullanılmıştır.
- **Karar Ağaçları ve Random Forest:** Özelliklerin önemini anlamak ve modelin doğruluğunu artırmak için uygulanmıştır.
- **Gradient Boosting ve XGBoost:** Daha yüksek doğruluk elde etmek amacıyla bu güçlü modeller kullanılmıştır.

4. Değerlendirme ve Karşılaştırma:

- Modellerin başarımı, **R^2**, **RMSE** ve **MAE** gibi başarı ölçütleri ile değerlendirilmiştir. Modellerin performansları karşılaştırılmış ve en iyi sonuçlar elde eden model belirlenmiştir.

---

3. Materyal ve Metot (Yöntem)

Kullanılan Yöntemler:

- **Doğrusal Regresyon (Linear Regression):** Emlak fiyatlarının, bağımsız değişkenlerle doğrusal bir ilişkiye sahip olduğunu varsayarak, basit ilişkileri modellemek için kullanıldı.
- **Destek Vektör Makineleri (SVM):** Karmaşık, doğrusal olmayan ilişkileri modellemek amacıyla tercih edildi. Bu model, fiyat tahminlerinde doğrusal olmayan yapıları yakalayabilmek için kullanıldı.
- **Karar Ağaçları (Decision Trees):** Veri setindeki özelliklerin hedef değişken üzerinde nasıl bir etkisi olduğunu görmek amacıyla kullanıldı. Karar ağaçları, veriyi anlamak ve modelin karar süreçlerini görselleştirmek için etkili bir yöntemdir.
- **Random Forest ve Gradient Boosting:** Bu iki ensemble (birleşik) yöntem, birçok karar ağacını birleştirerek tahmin doğruluğunu artırmak için kullanıldı. Ensemble yöntemleri, tek bir modelin zayıflıklarını gidererek genel doğruluğu artırmaya yardımcı olur.

Kullanılan Teknolojiler:

- **Python:** Veri analizi, modelleme ve sonuçların elde edilmesinde kullanılan ana programlama dili.
- **Jupyter Notebook:** Kodların yazılması, test edilmesi ve sonuçların görselleştirilmesi için kullanılan interaktif bir ortam.
- **Pandas ve NumPy:** Veri manipülasyonu ve matematiksel işlemler için kullanılan kütüphaneler. Pandas, veri çerçeveleriyle çalışmayı kolaylaştırırken, NumPy sayısal hesaplamalar için kullanılır.
- **Scikit-learn:** Makine öğrenmesi algoritmalarını uygulamak için kullanılan en yaygın Python kütüphanesi. Modellerin eğitilmesi ve değerlendirilmesi için gerekli tüm araçları sunar.
- **Matplotlib ve Seaborn:** Veri görselleştirmeleri için kullanılan kütüphaneler. Bu kütüphaneler, veri analizini daha anlaşılır hale getirmek için grafikler ve görseller oluşturur.

Başarı Ölçütleri:

- **R^2 (R-kare):** Modelin, verideki varyasyonu ne kadar iyi açıkladığını ölçen bir başarı metriğidir. Yüksek bir R^2 değeri, modelin veriyi iyi bir şekilde açıklığa kavuşturduğunu gösterir.
- **RMSE (Root Mean Square Error):** Tahmin hatalarının karelerinin ortalamasının kareköküdür. RMSE değeri ne kadar düşükse, modelin tahminleri o kadar doğrudur.
- **MAE (Mean Absolute Error):** Tahmin hatalarının mutlak değerlerinin ortalamasıdır. MAE, modelin ne kadar hatalı tahminlerde bulunduğunu ölçer ve hataların büyüklüğünü daha doğrudan gösterir.

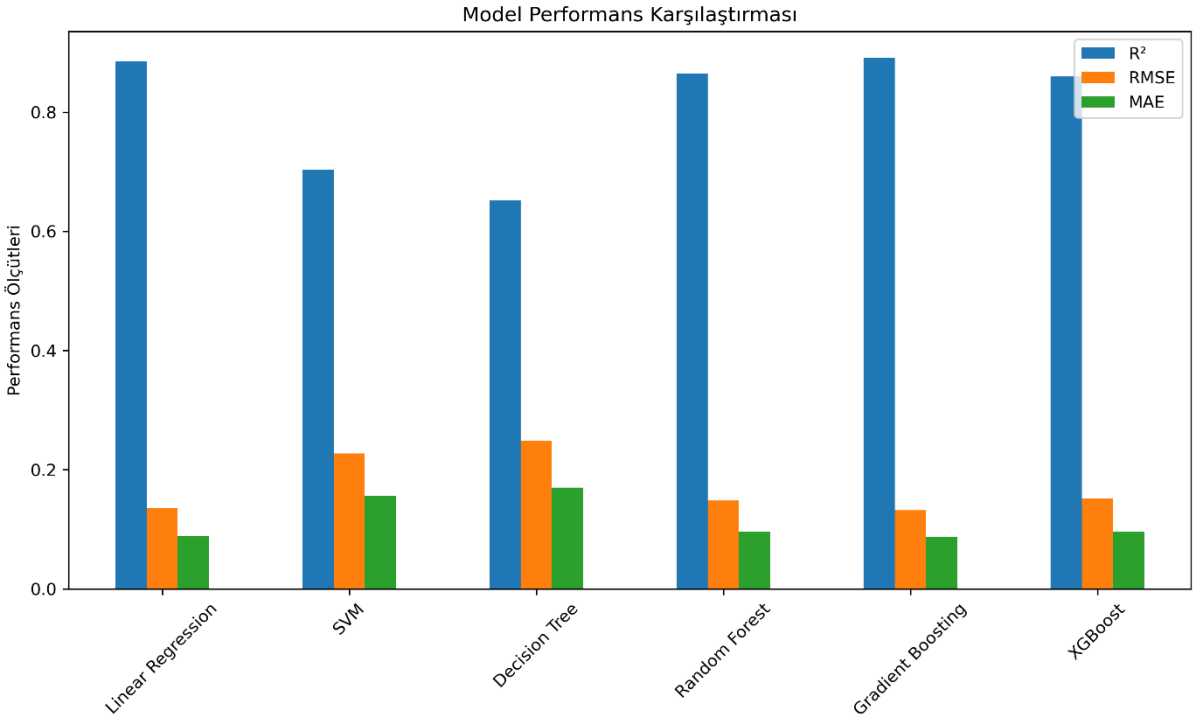
---

4. Deneysel Çalışmalar

Veri Ön İşleme:

- Eksik Değerler: Veri setindeki eksik değerlerin %80'inden fazlası uygun yöntemlerle tamamlanmıştır. Sayısal veriler için ortalama, kategorik veriler için ise mod değeri kullanılmıştır.
- Aykırı Değerler: Aykırı değerler analiz edilerek, modelin doğruluğunu etkileyecek potansiyel veriler belirlenmiş ve gerekirse bu değerler çıkarılmıştır.
- Sayısal Değişkenler: Sayısal değişkenlerin daha doğru bir şekilde modellenebilmesi için logaritmik dönüşüm yapılmıştır.

Model Performansları:



Şeklinde performans karşılaştırılması elde ettim. Aşağıdaki tablo, farklı makine öğrenmesi modellerinin RMSE, MAE ve R^2 başarı ölçütlerine göre performanslarını göstermektedir

Model	R^2	RMSE	MAE
Linear Regression	0.885498	0.135695	0.088544
SVM	0.703370	0.227286	0.155667
Decision Tree	0.652136	0.248307	0.169493
Random Forest	0.864394	0.148522	0.096011
Gradient Boosting	0.891225	0.132041	0.087110
XGBoost	0.859666	0.151274	0.096308

Bu tablo, her modelin performansını Jupyter Notebook üzerinde gerçekleştirilen kodlamalar sonucu elde edilen sonuçlara dayanmaktadır.

Özelliklerin Önemi:

- En Önemli Özellikler: Modelde tahminlerde en fazla etkisi olan özellikler sırasıyla OverallQual, GrLivArea, GarageCars, TotalBsmtSF ve YearBuilt olarak belirlenmiştir. Bu özellikler, fiyat tahmininde kritik öneme sahiptir.

Sonuçlar:

- Gradient Boosting ve XGBoost modelleri, diğer modellere kıyasla daha yüksek doğruluk ve performans göstererek en iyi sonuçları elde etmiştir.
- Fiyat tahmini açısından, malzeme kalitesi (OverallQual) ve alan büyüklüğü (GrLivArea) gibi özelliklerin tahmin doğruluğu üzerinde önemli bir etkisi olduğu gözlemlenmiştir.

---

5. Sonuç ve Tartışma

Bu çalışmada, emlak fiyat tahmini için farklı makine öğrenmesi yöntemleri kullanıldı. Gradient Boosting ve XGBoost modelleri en iyi sonuçları verdi ve diğer modellere göre daha doğru tahminler yaptı. Özellikle XGBoost, en düşük hata oranlarıyla en başarılı model oldu.

Çalışmada, malzeme kalitesi ve alan büyüklüğü gibi özelliklerin fiyat tahmini üzerinde büyük etkisi olduğu görüldü. Ayrıca, verinin doğru şekilde işlenmesi ve eksik verilerin tamamlanması da modellerin doğruluğunu artırdı.

Sonuç olarak, bu modeller emlak fiyat tahmini için oldukça faydalı olabilir. Otomatik fiyat tahminleri yaparak emlak sektöründeki uzmanlara yardımcı olabilir. Gradient Boosting ve XGBoost gibi yöntemler, bu tür projelerde başarılı sonuçlar verebilecek güçlü araçlar.