

[**Mühendislik Fakültesi**](https://www.konya.edu.tr/muhendislikvemimarlik)

[**Bilgisayar Mühendisliği Bölümü**](https://www.konya.edu.tr/bilgisayarmuhendisligi)

**Yapay Zekâya Giriş Dersi Ön Raporu**

|  |
| --- |
| **Proje Başlığı** |
| California'daki Ev Fiyatları Tahmin Web Sitesi |

|  |  |
| --- | --- |
| **Öğrencilerin Bilgileri** | |
| **Öğrenci No** | 21100011033 |
| **Öğrenci Ad Soyad** | Büşra DANIŞMAN |
| **Öğrenci No** | 21100011030 |
| **Öğrenci Ad Soyad** | İrem Sıla SARIKAYA |

|  |
| --- |
| **DOÇ.DR.**  **Mehmet Hacıbeyoğlu** |

**Nisan 2024**

**Konya**

**İçindekiler**

[1. Proje Özeti ve Veri Seti Bilgileri 3](#_Toc51147498)

[2. Projenin Amacı ve Hedefi 4](#_Toc51147499)

[3. **Metodoloji** 5](#_Toc51147501)

[3.1. Veri Temizleme ve Ön İşleme 5](#_Toc51147502)

[3.2. Random Forest Modeli Oluşturma 5](#_Toc51147503)

[3.3. Web Sitesi Geliştirme 5](#_Toc51147502)

[4. Değerlendirme ve Sonuçlar 6](#_Toc51147504)

[5. Projede kullanılacak yöntem, donanımlar ve yazılımlar ile ilgili bilgiler 8](#_Toc51147504)

[6. Kaynaklar 9](#_Toc51147504)

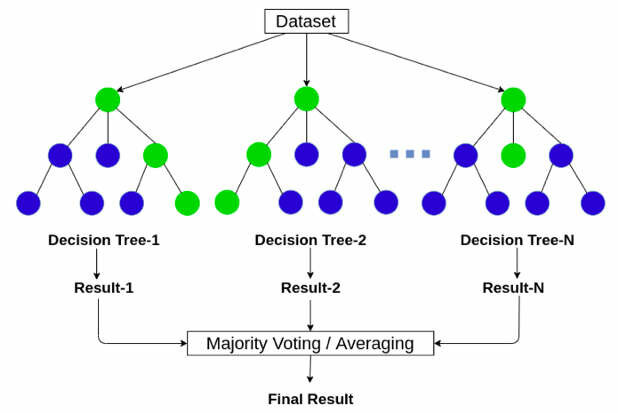
# 1. Proje Özeti ve Veri Seti Bilgileri

Bu proje California'daki ev fiyatlarını tahmin etmek amacıyla makine öğrenmesi yöntemlerini kullanmayı hedeflemektedir.

Ayrıca, kullanıcılar için interaktif bir web sitesi oluşturularak ev fiyatı tahminleri görsel olarak sunulacaktır. Projede elde edilen sonuçlar detaylı olarak raporlanacak ve projenin ilerleyen aşamaları için öneriler sunulacaktır. Bu proje, veri bilimi ve yapay zekâ alanında uygulamalı bir örnek sunarak ev fiyatı tahmini için makine öğrenmesi tekniklerinin kullanımını gösterecektir. Kaggle'dan temin edilen "California Housing Prices" (1) veri seti üzerinde Random Forest algoritması kullanılarak bir tahmin modeli geliştirilecektir.

* **Random Forest Regressor nedir?**

Random Forest Regressor, birden fazla karar ağacını bir araya getirerek tahminleme yapar. Her bir karar ağacı, rastgele seçilen bir alt veri seti ve rastgele seçilen özellikler ile eğitilir. Nihai tahmin, bu ağaçların tahminlerinin ortalamasıdır. (2)



Resim Random Forest Regresyon algoritması şeması

* **Algoritma nasıl işler?**

1. Eğitim verisi (özellikler ve hedef değişken) belirlenir.
2. Eğitim veri setinden rastgele N örnekleme seti oluşturun. Her set, orijinal veri setinin boyutuna sahiptir ancak örnekler tekrar edebilir.
3. Her örnekleme seti için bir karar ağacı eğitilir. Ağaç eğitilirken, her bir düğümde rastgele seçilen bir alt özellik seti kullanılır. Tahmin sonucu oluşan her değer için oylama gerçekleştirilir,
4. Birçok bağımsız karar ağacı eğitildikten sonra, bu ağaçların tahminleri bir araya getirilir.
5. Bu tahminler ortalanarak (ortalama alınarak) nihai tahmin elde edilir. (3)

* **Random Forest'ın Avantajları**

1. Çok sayıda karar ağacının ortalaması alındığı için genellikle yüksek doğruluk sağlar.
2. Her ağaç farklı veri setleri ve özellikler ile eğitildiği için model karar ağaçlarının en büyük problemlerinden olan overfitting azaltır.(Medium,2024)

( Overfitting (aşırı öğrenme), bir makine öğrenimi modelinin eğitim verileri üzerinde çok iyi performans gösterirken, yeni ve daha önce görülmemiş test verileri üzerinde düşük performans göstermesi durumudur.)

* **Random Forest'ın Dezavantajları**

1. Çok sayıda ağaç eğitmek ve tahmin yapmak zaman alıcı olabilir.
2. Modelin yorumlanabilirliği azalır; her ağacın ayrı ayrı yorumlanması zor olabilir. (4)

**Veri Seti Bilgisi**

Bu çalışmada, California eyaletindeki konut fiyatlarını tahmin etmek için "California Housing Prices" veri setini kullandık. Bu veri seti, Kaliforniya'nın farklı bölgelerindeki konutlara ait çeşitli özellikleri içermektedir.

## Veri setinin seçilme nedenleri:

* **Uygunluk**: Bu veri seti, konut fiyatlarını etkileyen çeşitli özellikleri içerir. Bu özellikler arasında konum, yaş, oda sayısı, gelir seviyesi gibi faktörler bulunmaktadır. Bu faktörlerin çeşitliliği, bir konutun fiyatını tahmin etmek için kullanışlı bir veri seti oluşturur.
* **Geniş Kapsam**: Veri seti, Kaliforniya eyaletinin farklı bölgelerinden toplanmış verilere dayanmaktadır. Bu, farklı coğrafi bölgelerdeki konut piyasalarının farklılıklarını ve eğilimlerini incelememize olanak tanır.
* **Uygulanabilirlik**: Konut fiyatlarının tahmini, gayrimenkul sektöründe birçok uygulama için önemlidir. Bu veri seti, emlak danışmanları, yatırımcılar ve konut alıcıları için gerçek dünya senaryolarına dayalı tahminler yapmamızı sağlar.

Bu nedenlerle, "California Housing" veri setini, konut fiyatlarını tahmin etmek için uygun bir veri kaynağı olarak seçtik.

### Veri setinde bulunan özellikler şunlardır:

* **longitude**: Konutun boylam koordinatı.
* **latitude**: Konutun enlem koordinatı.
* **housing\_median\_age**: Konutun medyan yaşı.
* **total\_rooms**: Konutun toplam oda sayısı.
* **total\_bedrooms**: Konutun toplam yatak odası sayısı. Bu sütunda 20433 non-null değeri bulunmaktadır, yani bazı veriler eksiktir.
* **population**: Konutun nüfusu.
* **households**: Konutun hane halkı sayısı.
* **median\_income**: Bölgedeki medyan gelir.
* **median\_house\_value**: Konutun medyan fiyatı.
* **ocean\_proximity**: Konutun okyanusa olan yakınlığı.

Veri setinde toplam 20640 veri örneği bulunmaktadır.

Ayrıca, kullanıcılar için interaktif bir web sitesi oluşturularak ev fiyatı tahminleri görsel olarak sunulacaktır. Projede elde edilen sonuçlar detaylı olarak raporlanacak ve. Bu proje, veri bilimi ve yapay zekâ alanında uygulamalı bir örnek sunarak ev fiyatı tahmini için makine öğrenmesi tekniklerinin kullanımını gösterecektir.

# 2.Projenin Amacı ve Hedefi

Bu projenin hedefi California'daki ev fiyatlarını tahmin etmektir. Amacımız, makine öğrenmesi yöntemlerini kullanarak Kaggle'dan temin edilen "California Housing Prices" veri seti üzerinde bir tahmin modeli geliştirmektir. Bu model, ev fiyatlarını özelliklerine dayanarak tahmin etmek için Random Forest algoritmasını kullanacaktır. Projede veri setinin temizlenmesi, eksik değerlerin doldurulması, aykırı değerlerin ele alınması ve özellik mühendisliği gibi ön işleme adımları gerçekleştirilecektir. Daha sonra, eğitim veri seti üzerinde Random Forest modeli eğitilecek ve modelin performansı değerlendirilecektir.

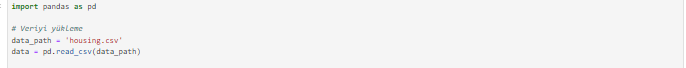
Projenin bir diğer amacı, kullanıcılar için interaktif bir web sitesi oluşturarak tahmin sonuçlarını görsel olarak sunmaktır. Web sitesi, kullanıcıların belirli ev özelliklerini girerek ev fiyatı tahminlerini almasını sağlayacaktır. Bu sayede, ev alıcıları veya satıcıları kolayca ev fiyatlarını değerlendirebilecek ve kararlarını destekleyici bilgilere sahip olabileceklerdir. Projede elde edilen sonuçlar detaylı olarak raporlanacak ve projenin ilerleyen aşamaları için öneriler sunulacaktır. Bu proje, veri bilimi ve yapay zekâ alanında uygulamalı bir örnek sunarak ev fiyatı tahmini için makine öğrenmesi tekniklerinin kullanımını gösterecektir.

# 3.Metodoloji

Bu çalışmada, California Housing Prices veri kümesi üzerinde konut fiyatlarını tahmin etmek için çeşitli makine öğrenimi algoritmaları kullanılacaktır. Aşağıda, çalışmanın adımları daha detaylı bir şekilde açıklanmıştır:

* **1.Veri Edinme ve Tanımı:**

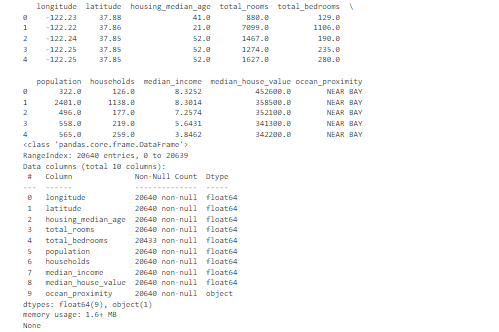
California Housing Prices veri kümesi, California'daki konutlarla ilgili çeşitli özellikleri içeren bir veri kümesidir. Veri kümesi, konum, yaş, oda sayısı, gelir seviyesi gibi özelliklerin yanı sıra konut fiyatlarını içerir.



housing.csv dosyasını data adlı bir DataFrame'e yüklüyoruz.

Verinin ilk birkaç satırını ve genel bilgi (satır sayısı, sütun adları, veri tipleri) gösteriyoruz.

Çıktımız:



* **2.Veri Ön İşleme:**

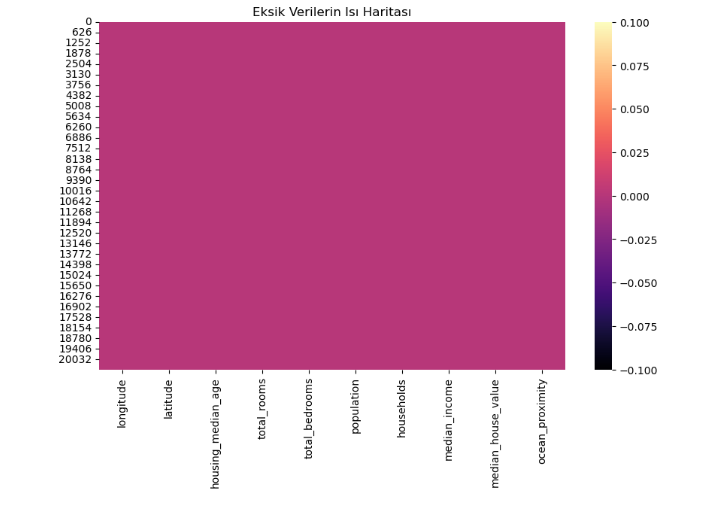
**2.1 Eksik Veri Kontrolü ve Doldurma:**

Veri setinde eksik değerler kontrol edilecek ve uygun bir doldurma stratejisi veya silme belirlenecektir. Eksik değerler, medyan veya ortalama değerlerle doldurulabilir veya silinebilir.



Her sütundaki eksik değerlerin sayısını kontrol ediyoruz. Eksik değerlere sahip satırları siliyoruz.

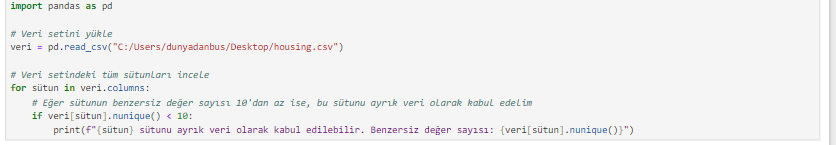
Burada ısı haritası yöntemiyle eksik değerimizin olup olmadığını kontrol edebiliriz..



Grafikte tamamen morumsu renkli bir ısı haritası olduğu için, eksik veri olmadığını ve tüm hücrelerin dolu olduğunu söyleyebiliriz. Yani, veri setinde herhangi bir eksik veya boş değer bulunmamaktadır. (5)

**2.2 Ayrık Verilerin Dönüştürülmesi:**

Ayrık veri, belirli bir aralık veya kategoriden oluşan veri tipidir. Bu tür veriler, genellikle kategorik veya nominal olarak adlandırılan verilerdir ve kesikli veya sürekli olmayan değerler içerirler. Örneğin, cinsiyet (erkek, kadın), ülke (Türkiye, Amerika), renk (kırmızı, mavi) gibi özellikler.



C:\Users\dunyadanbus\AppData\Local\Microsoft\Windows\INetCache\Content.Word\7.PNG

"ocean\_proximity" sütununun 5 benzersiz değere sahip olduğunu ve bu sütunun kategorik (ayrık) veri içerdiğini belirtir



"ocean\_proximity" gibi kategorik veriler, sayısal değerlere dönüştürülecektir. Bu dönüşüm genellikle one-hot encoding veya label encoding yöntemleriyle yapılır.

* **3. Veri Analizi ve Görselleştirme:**

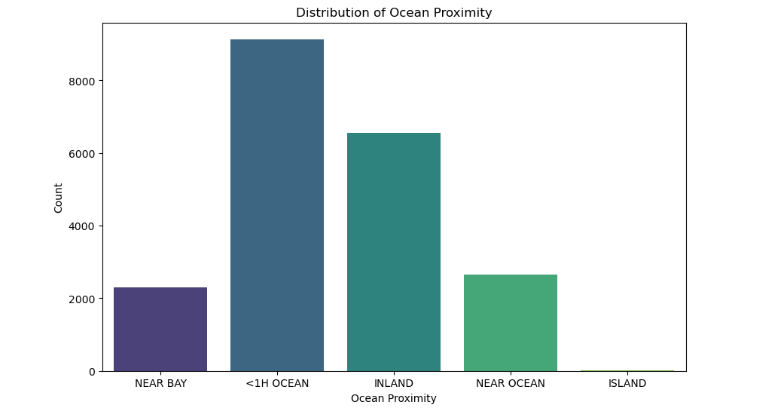
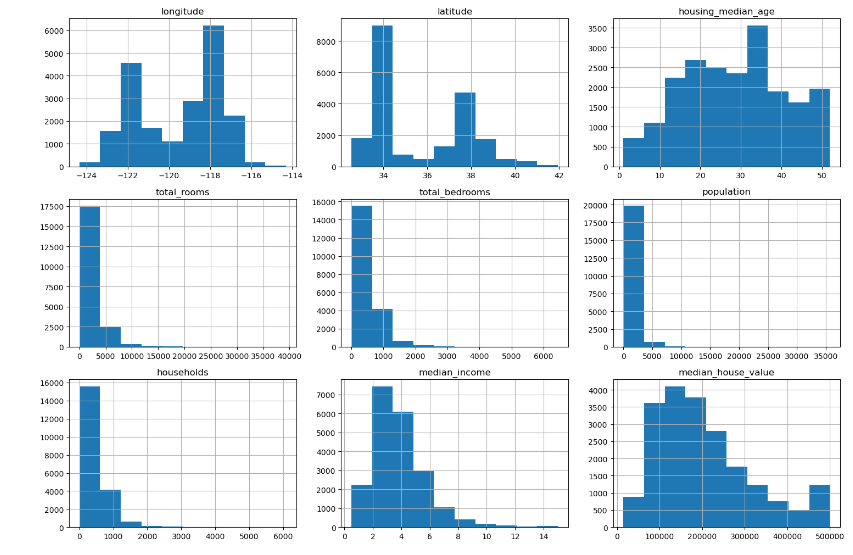
Özelliklerin dağılımları, histogramlar ve yoğunluk grafikleri ile incelenecek ve görselleştirilecektir. veri.hist() işlevi, yalnızca sayısal verilere sahip sütunları otomatik olarak çizer.



"ocean\_proximity" gibi kategorik bir sütun çizdirilmez. Bu tür veriler için farklı bir görselleştirme yöntemi kullanmanız gerekmektedir.

Bu tür kategorik verileri görselleştirmek için countplot() yararlanabilirsiniz.



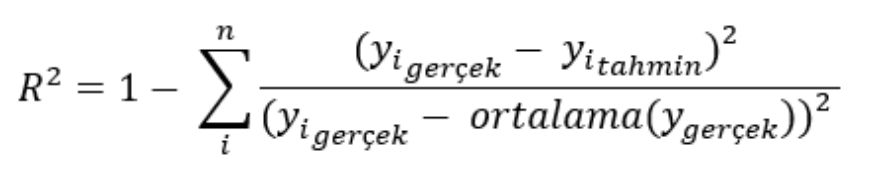


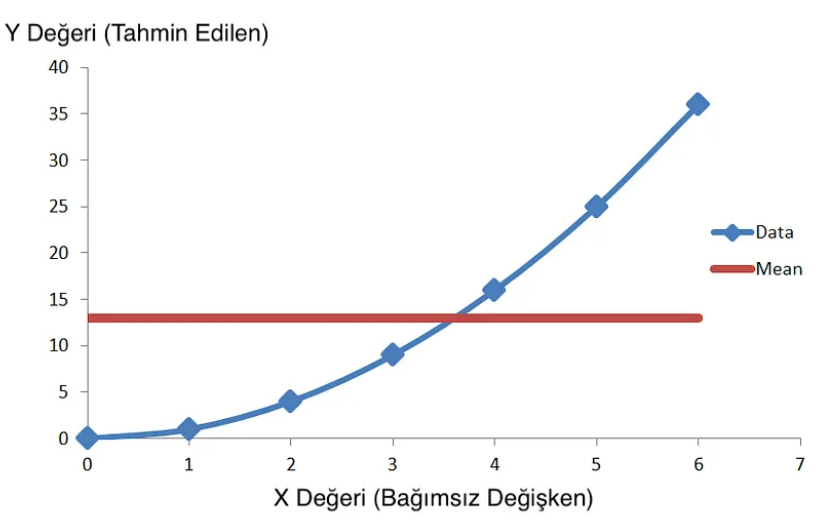
* **4. Model Eğitimi ve Değerlendirme:**

Modelin performansı, R-kare (R-squared), ortalama karesel hata (MSE) gibi metrikler kullanılarak değerlendirilecektir.

**R-Squared (R-Kare) :**

Bir regresyon modelinin verilere ne kadar iyi uyduğunu ölçen önemli bir istatistiksel ölçüdür. Bu değer genellikle 0 ile 1 arasında olup, modelin bağımsız değişkenler tarafından bağımlı değişkeni ne kadar başarılı bir şekilde açıkladığını gösterir. Yüksek bir R-Kare değeri, modelin verilere daha iyi uyduğunu ve tahminlerin güvenilir olduğunu işaret ederken, düşük bir değer modelin başarısız olduğunu gösterir.

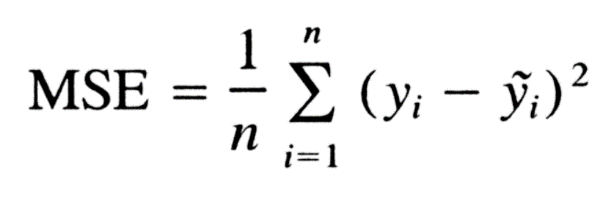


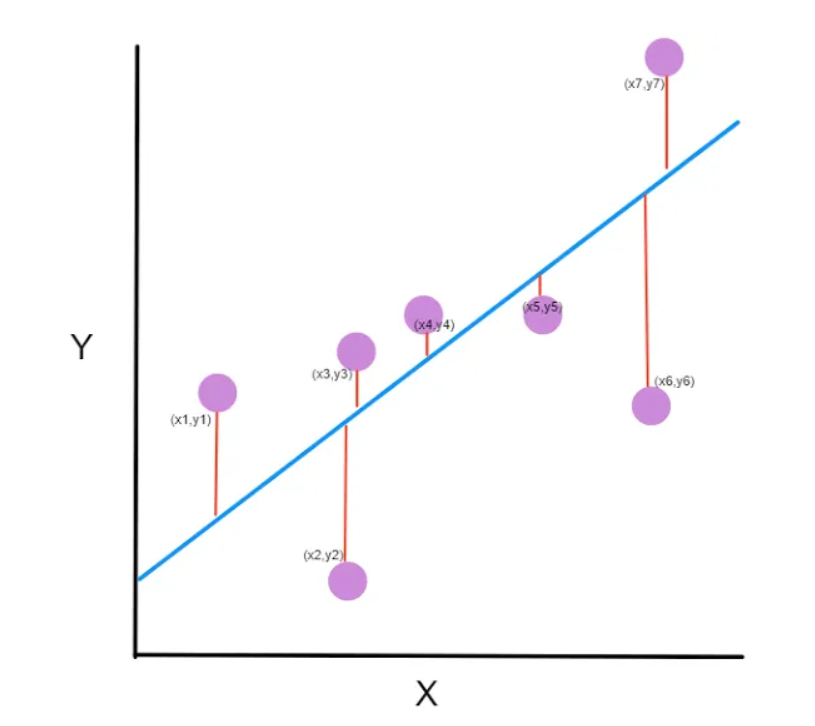
****

Resim R^2 grafiği

**Ortalama karesel hata (MSE)** **:**

Bir regresyon modelinin tahminlerinin gerçek değerlere ne kadar yakın olduğunu ölçen bir istatistiksel metriktir. MSE değeri, modelin her bir tahmininin gerçek değerden ne kadar sapma gösterdiğini ölçerek modelin performansını değerlendirir. Daha düşük bir MSE değeri, modelin daha iyi bir şekilde tahmin yaptığını ve gerçek verilere daha yakın olduğunu gösterir.



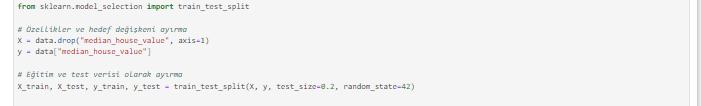


Resim MSE grafiği

Bu metrikler, modelin ne kadar iyi çalıştığını ve tahminlerinin ne kadar doğru olduğunu ölçmek için kullanılacaktır. Ayrıca, web sitesi kullanıcı geri bildirimleri ve kullanım istatistikleri de değerlendirilerek, web sitesinin kullanıcı deneyimi sürekli olarak iyileştirilecektir. Modelin performansı ve web sitesinin kullanıcı kabulü, projenin başarısını belirleyen kritik unsurlardır. Sonuçların değerlendirilmesi ve elde edilen sonuçların analizi, projenin önemli bir aşamasını oluşturacaktır.

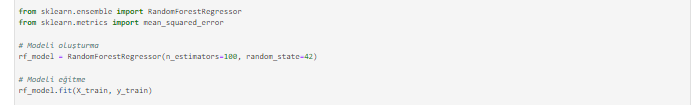
Bağımsız değişkenleri (X) ve bağımlı değişkeni (y) ayırıyoruz. median\_house\_value hedef değişkenimiz. Veriyi eğitim ve test setlerine ayırıyoruz. Verinin %20'si test için, %80'i eğitim için kullanılıyor.

(Veri setini eğitim ve test setlerine ayırırken kullanılan oran, projeye, veri setinin boyutuna ve diğer spesifik gereksinimlere bağlı olarak belirlenir.)

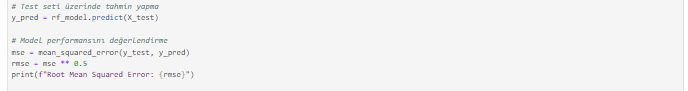


Bir RandomForestRegressor modeli oluşturuyoruz ve bunu eğitim verisi üzerinde eğitiyoruz. 100 adet ağaç (n\_estimators=100) kullanıyoruz.

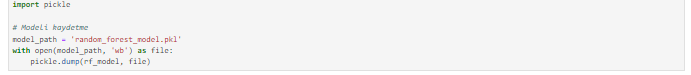
(100 adet ağaç kullanmak, RandomForestRegressor modelinin performansını artırmak ve genelleme yeteneğini sağlamak için iyi bir başlangıç noktasıdır. Bu sayı, hesaplama maliyetini makul düzeyde tutarken yeterli çeşitlilik ve stabilite sağlar.)



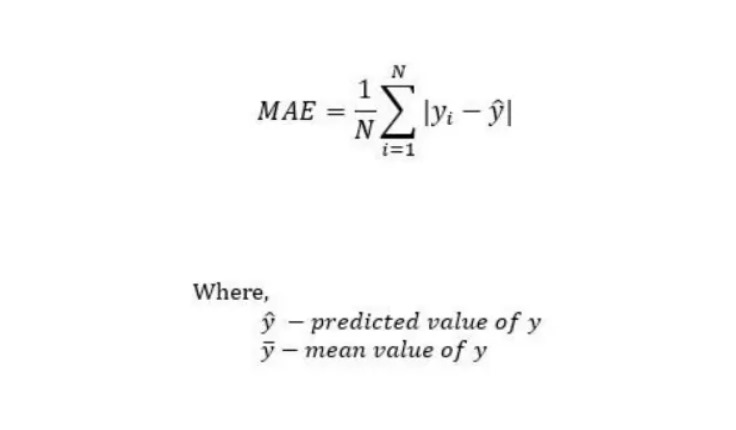
Test verisi üzerinde tahmin yapıyoruz. Tahminlerin performansını Mean Squared Error (MSE) ve Root Mean Squared Error (RMSE) kullanarak değerlendiriyoruz. RMSE, hatanın karekökünü alarak daha anlaşılır bir hata ölçüsü sağlar.



Eğittiğimiz modeli bir dosyaya (random\_forest\_model.pkl) kaydediyoruz ki daha sonra bu modeli kullanabilelim.



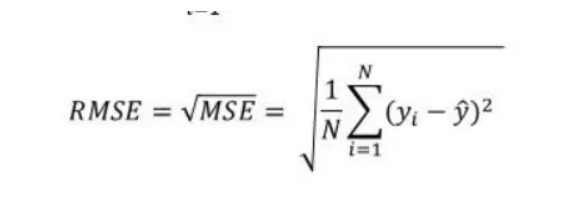
**Mean Absolute Error (MAE)** - Ortalama Mutlak Hata: Tahmin edilen değerlerin gerçek değerlerden ne kadar sapma gösterdiğini ortalama olarak ölçer.



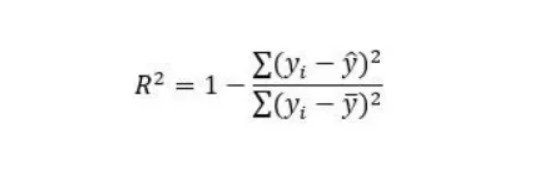
**Mean Squared Error (MSE)** - Ortalama Kare Hatası: Tahmin edilen değerlerin gerçek değerlerden ne kadar sapma gösterdiğini kareleri alınmış olarak ortalama ölçer.



**Root Mean Squared Error (RMSE)** - Kök Ortalama Kare Hatası: MSE'nin karekökü alınarak elde edilir. Hataların büyüklüklerini daha anlamlı bir ölçekte gösterir.



**R^2 Score (R-Kare Skoru):** Modelin bağımsız değişkenlerin bağımlı değişkeni açıklama oranını gösterir. 1'e ne kadar yakınsa, model o kadar iyi demektir.



Kendi veri setimizde bunları uyguladığımızda:



MAE ve RMSE: Her ikisi de modelin tahminlerinin gerçek değerlere ne kadar yakın olduğunu gösterir. Bu durumda, hem MAE hem de RMSE düşük değerlerde olduğu için modelin tahminlerinin gerçek değerlere oldukça yakın olduğu söylenebilir.

MSE: Daha büyük hataları daha fazla cezalandırdığı için önemli bir metriktir. Düşük bir MSE değeri, modelin büyük hatalardan kaçındığını gösterir.

R^2 Skoru: Modelin genel açıklayıcılığını gösterir. %80.47'lik bir R^2 skoru, modelin oldukça iyi bir performans sergilediğini ve bağımsız değişkenlerin büyük bir kısmını açıklayabildiğini gösterir. (6)

**4.Sonuçların Yorumlanması ve Raporlama**

**Sonuçlar:**

Toplam Örnek Sayısı: 20,640

Özellik Sayısı: 9

Hedef Değişken: median\_house\_value

**Veri Türleri**: Veri setindeki özelliklerin çoğu sayısal (numeric) olup, yalnızca 'ocean\_proximity' özelliği kategoriktir.

**Veri Ön İşleme:**

**Eksik Veri**: Veri setinde eksik değer bulunmamaktadır; dolayısıyla eksik veri temizleme adımına gerek duyulmamıştır.

Kategorik Değişkenlerin Kodlanması: 'ocean\_proximity' özelliği, LabelEncoder kullanılarak sayısal değerlere dönüştürülmüştür, böylece modelin kategorik verilerle çalışması sağlanmıştır.

**Modelleme**:

Kullanılan Model: Random Forest Regressor

Model Parametreleri:

Ağaç Sayısı (n\_estimators): 100

Rastgelelik Kontrolü (random\_state): 42

Bu parametreler, modelin ağaç sayısını ve rastgeleliğini kontrol eder.

**Model Performansı**:

Root Mean Squared Error (RMSE): 0.3497

RMSE, modelin test setindeki tahminlerinin gerçek değerlerden ortalama olarak 0.3497 birim kadar uzaklıkta olduğunu göstermektedir.

**Modelin Ayrıntılı Performansı**:

Eğitim Seti RMSE: 0.2357

Test Seti RMSE: 0.3497

**Model Değerlendirme**:

Modelin eğitim setindeki RMSE değeri test setindeki değerden daha düşüktür, bu da modelin eğitim setine aşırı uyum (overfitting) yapmış olabileceğini gösterebilir.

**Model Kaydedildi**:

Model, 'random\_forest\_model.pkl' adıyla diske kaydedildi.

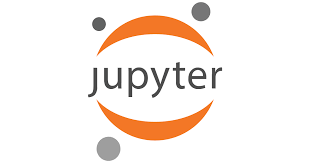
**Web Sitesi** :Oluşturulan model streamlit kullanılarak bir web sitesi oluşturuldu.



Resim 4 Web Sitesi görümü

**5.** **Projede kullanılacak yöntem, donanımlar ve yazılımlar ile ilgili bilgiler**

* **Jupyter:** Jupyter, interaktif hesaplamalar için kullanılan açık kaynaklı bir web tabanlı uygulamadır. Jupyter, bilimsel hesaplamalar, veri analizi, model geliştirme ve eğitimi gibi işlemleri gerçekleştirmek için kullanılan bir araçtır. Web tarayıcısı üzerinden erişilebilen ve Python gibi programlama dillerini kullanarak interaktif çalışmaları destekleyen bir ortamdır. Kullanıcılar, metin hücrelerinde açıklamalar yapabilir, kod hücrelerinde kodları çalıştırabilir ve anlık çıktıları görebilirler.
* **Kaggle:** Kaggle, veri bilimi ve yapay zekâ alanındaki projelerin geliştirilmesi ve yarışmaların düzenlendiği bir platformdur. Kaggle, veri bilimcilerin, araştırmacıların ve öğrencilerin gerçek dünya verileri üzerinde çalışarak model geliştirmelerine, deney yapmalarına ve bilgi paylaşmalarına olanak tanır.
* **Phyton:** Python, genel amaçlı, yüksek seviyeli, açık kaynaklı bir programlama dilidir. Guido van Rossum tarafından geliştirilen Python, basit ve okunabilir sözdizimiyle bilinir ve geniş bir kullanıcı kitlesine sahiptir. Python, çeşitli platformlarda (Windows, macOS, Linux) çalışabilir ve birçok farklı alanda kullanılmaktadır, özellikle veri bilimi, yapay zeka, web geliştirme, otomasyon, bilgisayar grafikleri ve oyun geliştirme gibi alanlarda popülerdir.
* **Streamlit:** Python programlama dili kullanılarak hızlı ve kolay bir şekilde veri uygulamaları ve dashboard'lar (gösterge panoları) oluşturmak için kullanılan açık kaynaklı bir kütüphanedir. Streamlit, veri bilimcileri ve geliştiricilere, karmaşık web uygulamaları geliştirme bilgisine ihtiyaç duymadan interaktif web uygulamaları oluşturma imkanı sağlar. Temel özellikleri arasında kolay kullanım, anında geri bildirim, interaktif bileşenler, kolay entegrasyon ve hızlı prototip oluşturma yer alır. Python kodu yazarak web uygulamaları oluşturabilir ve bu uygulamaları anında çalıştırabilirler. Kodda yapılan değişiklikler anında görülebilir ve uygulama kodu her kaydedildiğinde otomatik olarak yeniden yüklenir.



# Kaynakça

1. *https://www.kaggle.com/search?q=california+housing+prices.*

**2. *Miraç ÖZTÜRK / 13 Nisan 2022 / Python ile Sınıflandırma Analizleri – Rastgele .***

**3. *https://ece-akdagli.medium.com/makine-%C3%B6%C4%9Frenmesinde-random-forest-algoritmas%C4%B1-a79b044bbb31.***

**4. *https://www.slideshare.net/SezerFidanc/random-forest-algoritmas.***

**5. *https://jupyter.org/.***

**6. *https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e.***