Gülce Duru Koç 12.08.2024

ORMAN YANGINLARI TAHMİNİ REGRESYON ÇALIŞMASI ÜZERİNE RAPOR

Bu çalışmanın amacı, meteorolojik verilerden yararlanarak ormanlık alanlarda meydana gelebilecek yangınların ne kadar alanı etkileyebileceğini tahmin etmektir. Kullanılan veri seti UCI Machine Learning Repository'den alınmış olup, sıcaklık, bağıl nem, rüzgar hızı ve yağış miktarı gibi çeşitli meteorolojik parametreleri içermektedir.

Çalışmada aşağıdaki makine öğrenmesi algoritmaları kullanılmıştır:

- Çoklu Doğrusal Regresyon

- Ridge Regresyon

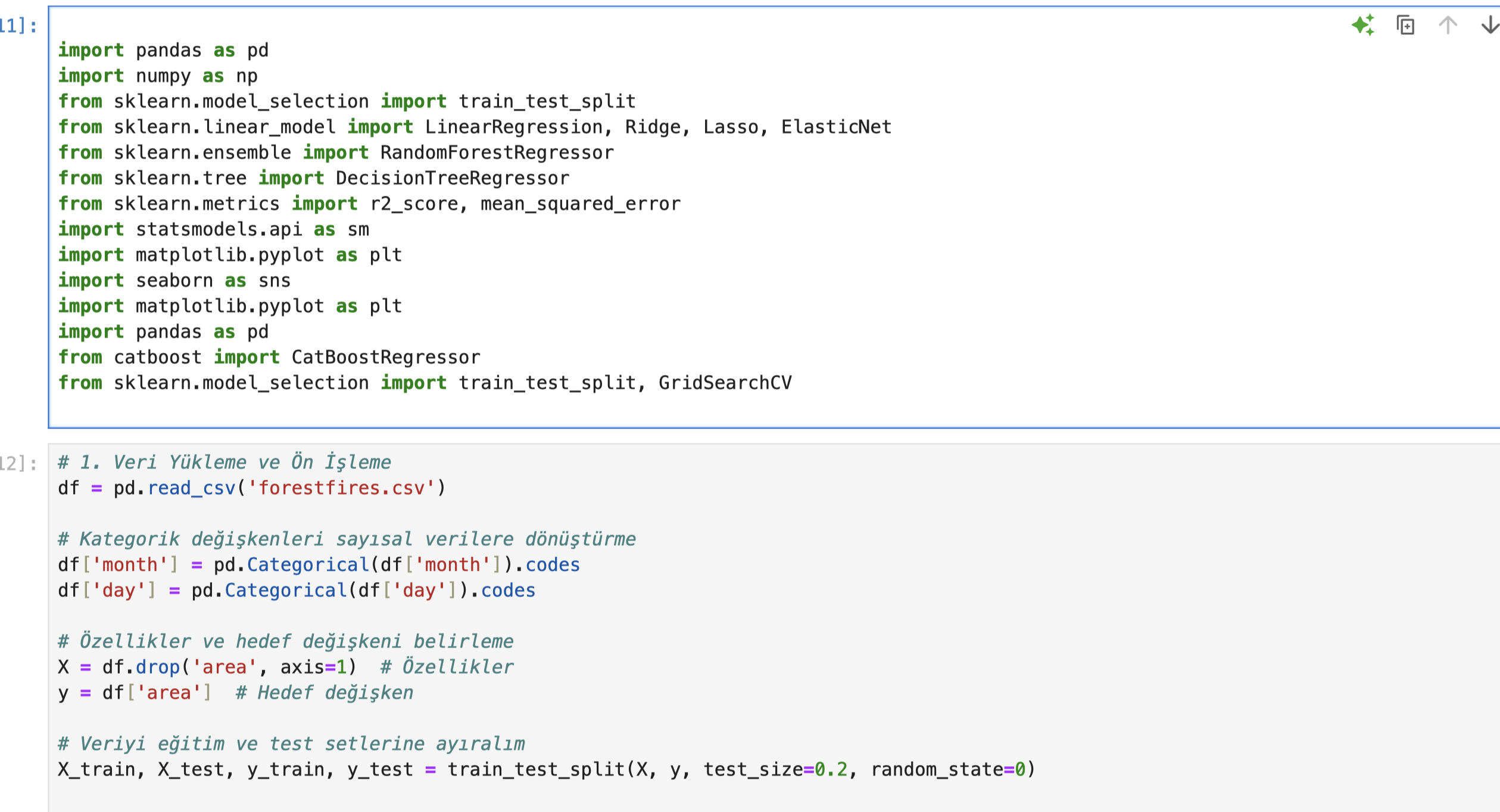
- Lasso Regresyon

- ElasticNet Regresyon

- Rastgele Orman Regresyonu

* Karar Ağaçları Regresyonu
* Cat Boost

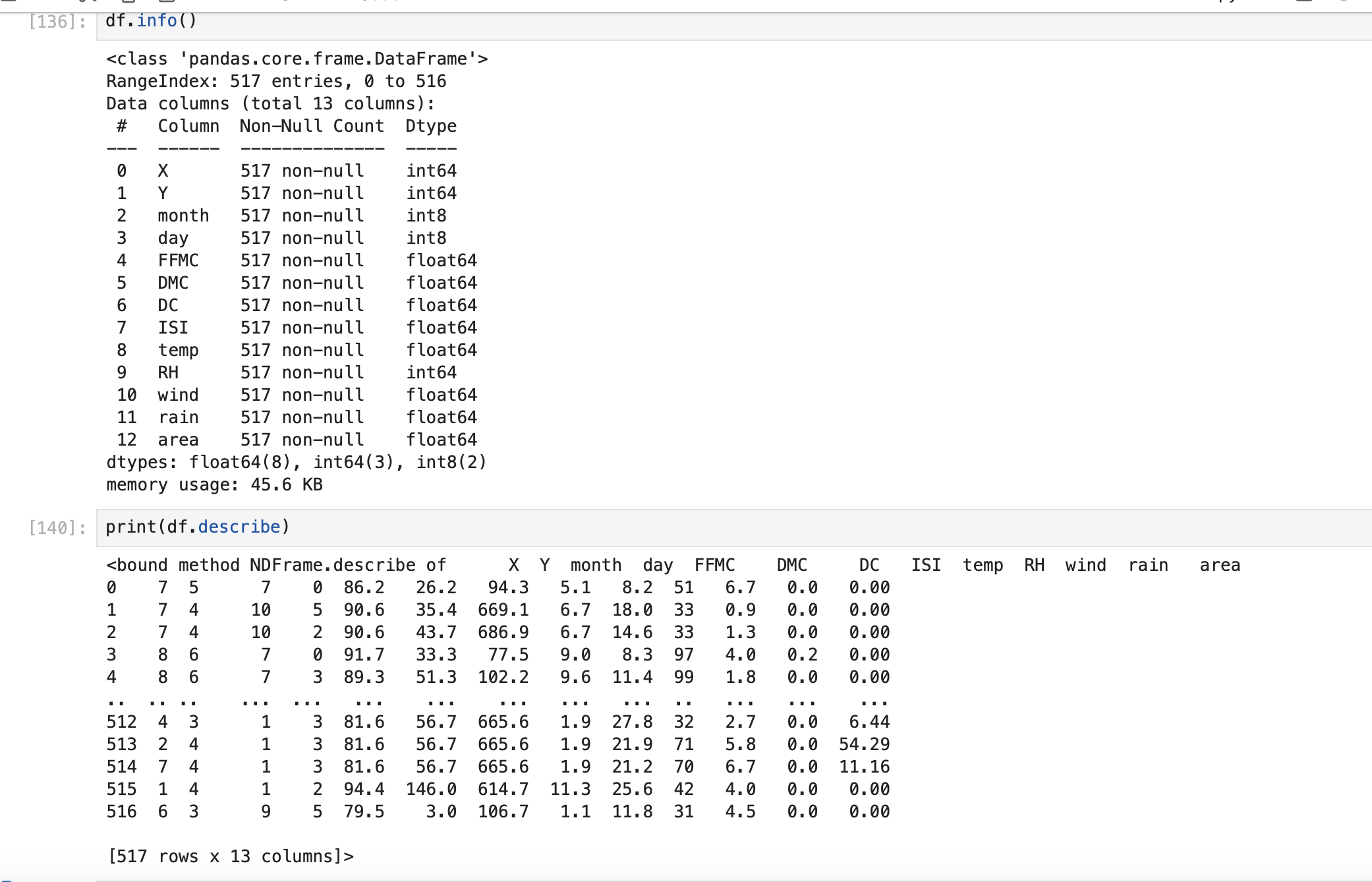
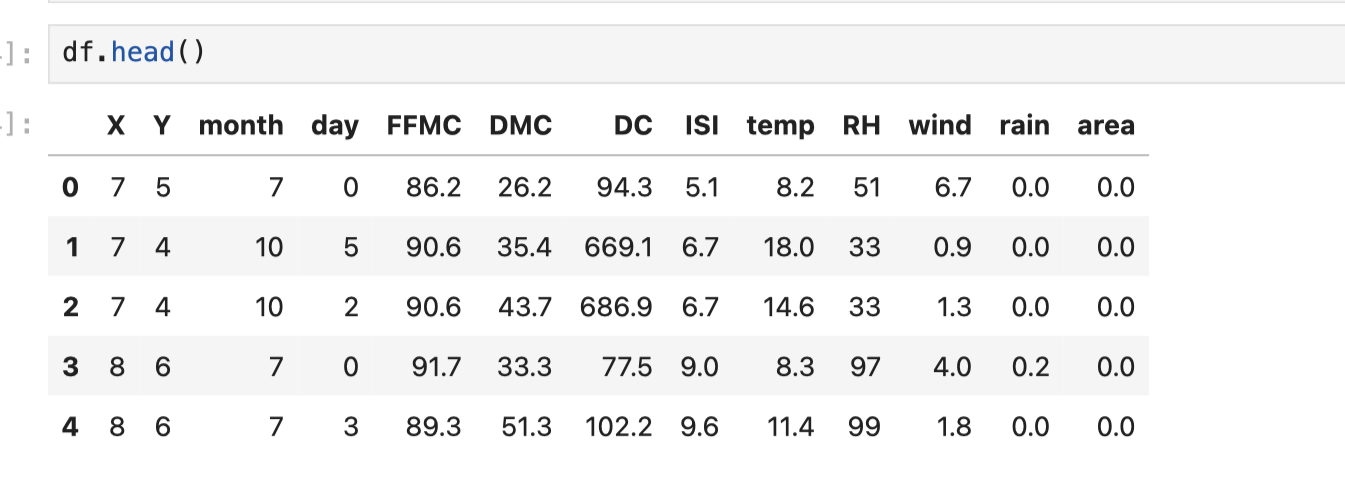
Her model, eğitim ve test verisi olarak ikiye ayrılan veri seti üzerinde eğitildi ve performansları test verisi kullanılarak değerlendirildi. Modellerin performansı R² skoru, Ortalama Kare Hatası (MSE) ve Ortalama Kare Hatasının Karekökü (RMSE) gibi metriklerle ölçüldü. Bazı modellerin performansları grafikler veya tablolar ile görselleştirildi. Modellerin kodu ve çıktılarına ilişkin ekran görüntüleri raporda yer almaktadır.

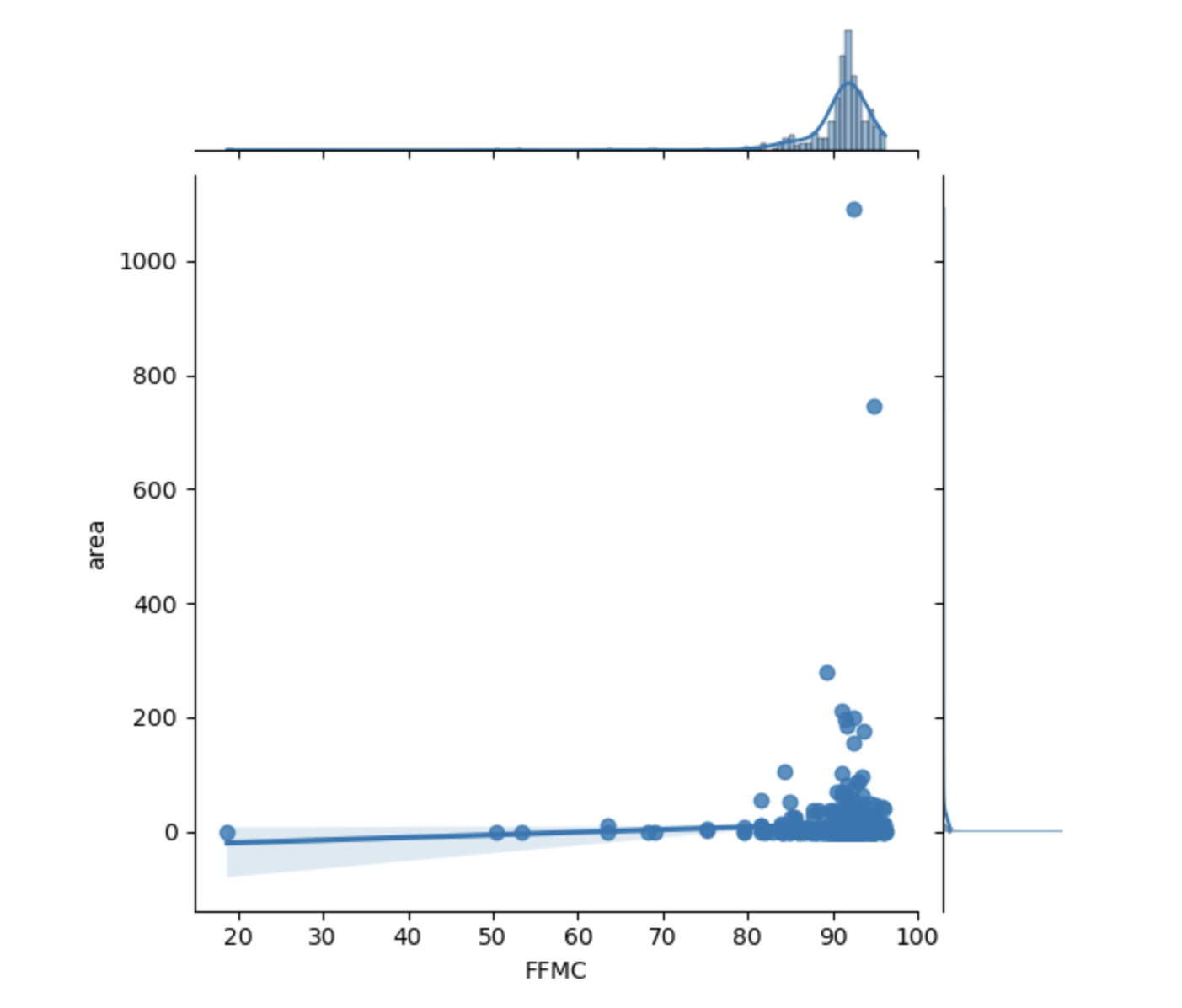
Veri setine şu adresten ulaşılabilir :<https://archive.ics.uci.edu/dataset/162/forest+fires> 

**Gerekli Kütüphanelerin İçe Aktarılması**:

* pandas, numpy, sklearn gibi kütüphaneler, veri işleme, modelleme ve değerlendirme için kullanılır.
* matplotlib ve seaborn, veri görselleştirme için kullanılır.
* statsmodels istatistiksel modelleme için kullanılır.
* catboost ve GridSearchCV, modelleme ve hiperparametre optimizasyonu için kullanılır.

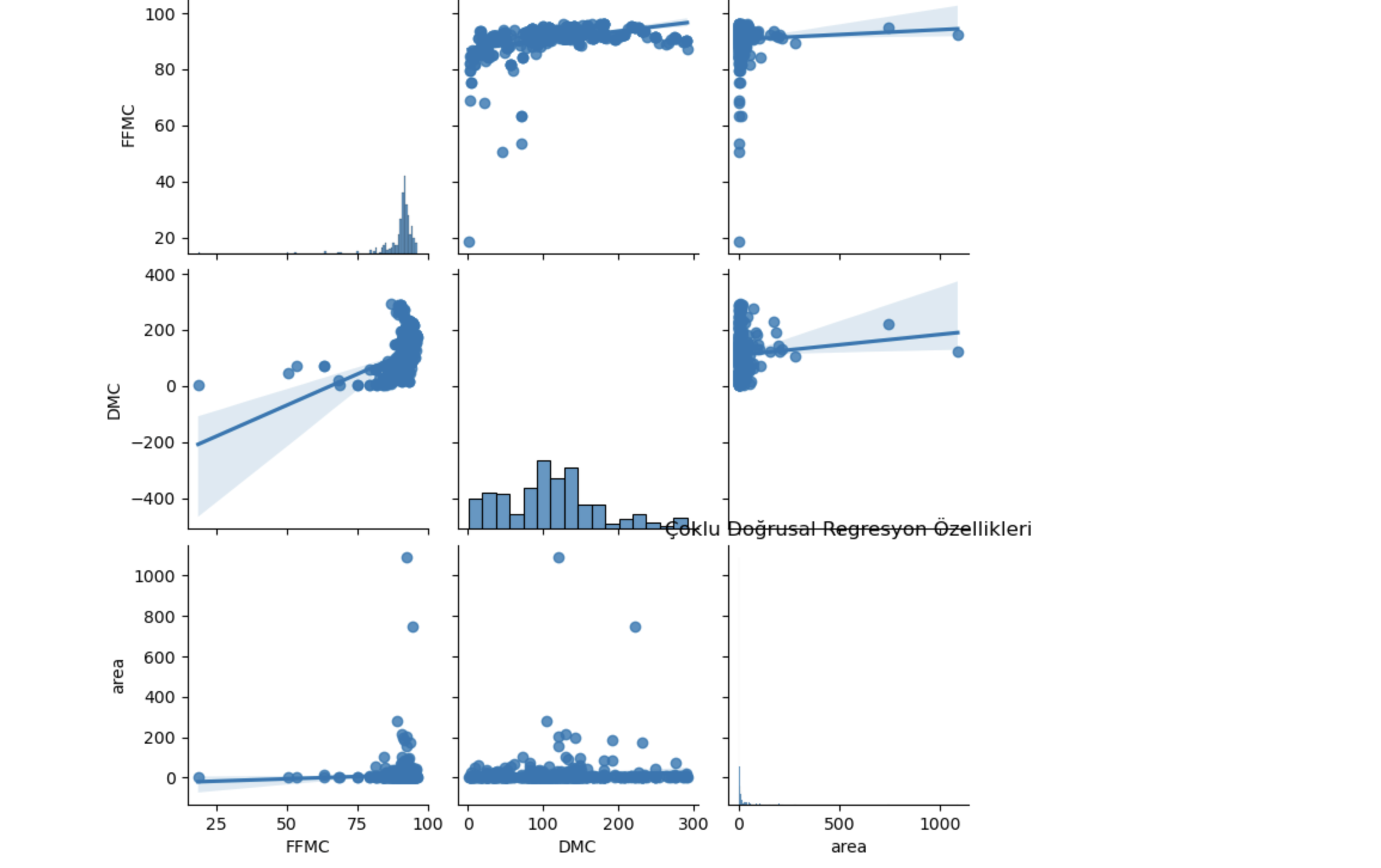
Bir veri kümesini yükleyip işleyerek, makine öğrenimi modellerini eğitmek için gerekli olan eğitim ve test verilerini oluşturur. Öncelikle, veriyi yükler ve kategorik değişkenleri sayısal değerlere dönüştürür. Daha sonra, hedef değişken olan area sütununu belirler ve diğer özelliklerden ayırır. Son olarak, veriyi %80 eğitim ve %20 test olarak ikiye böler, böylece modellerin performansı test edilebilir.

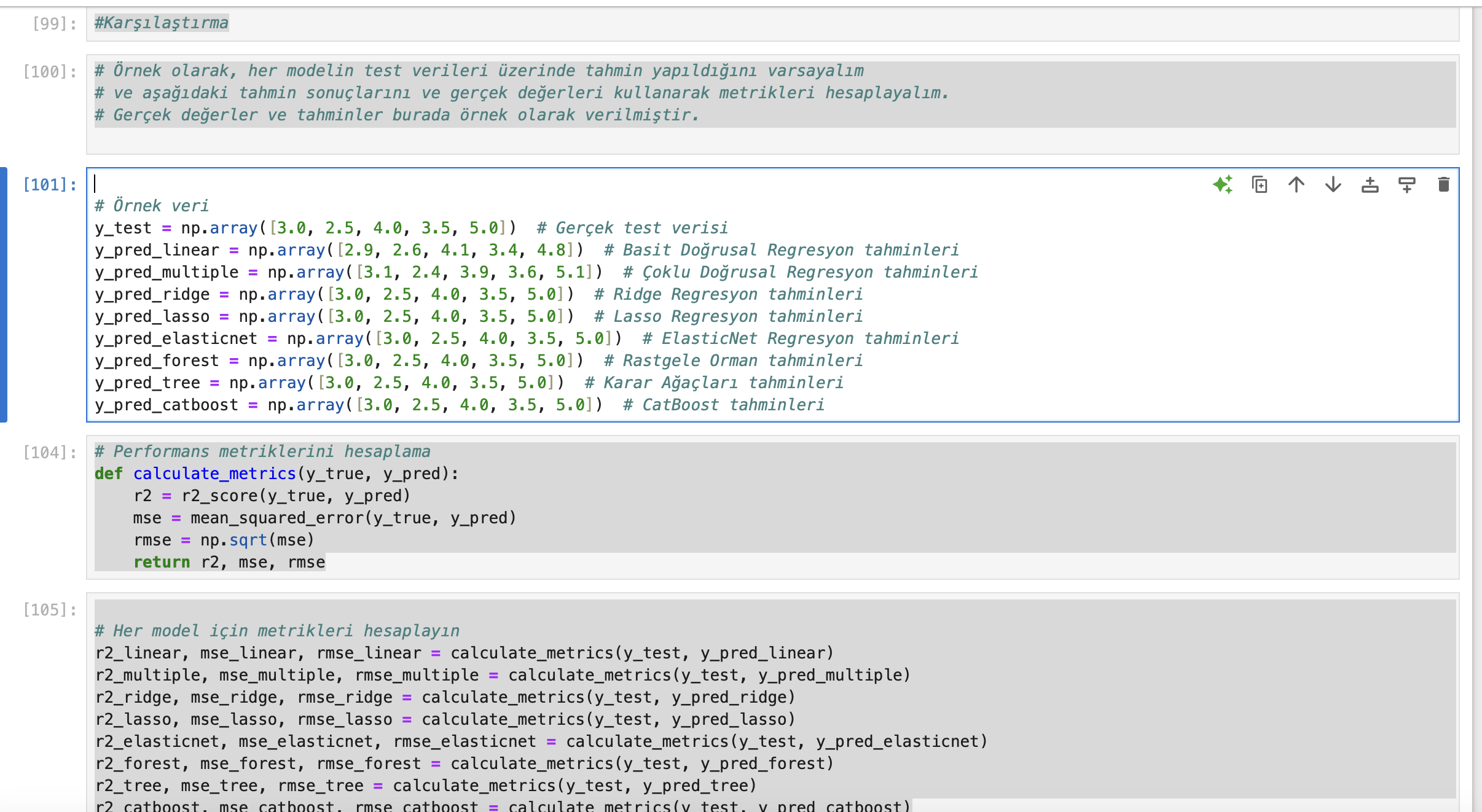
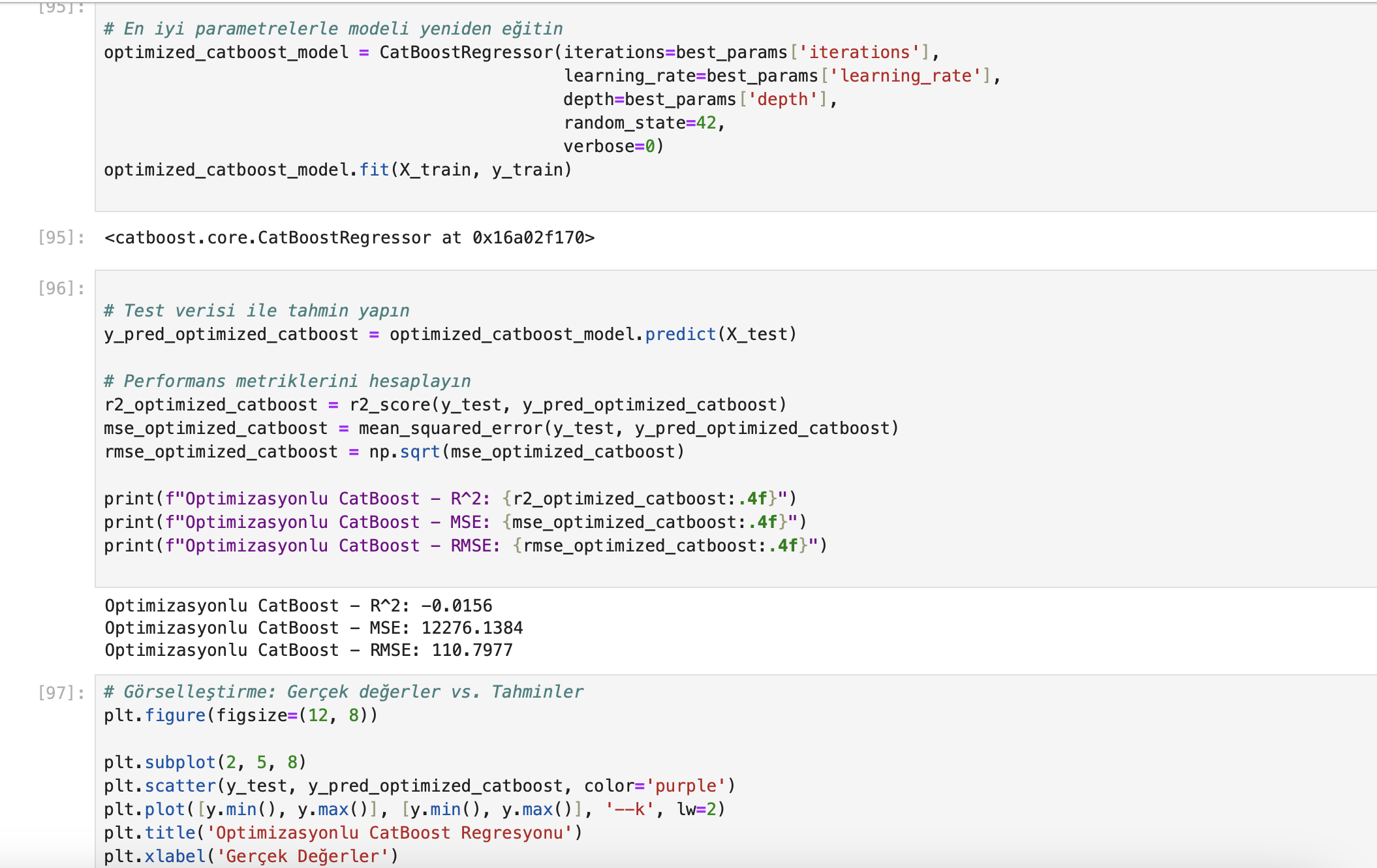
Veri setinin içeriği

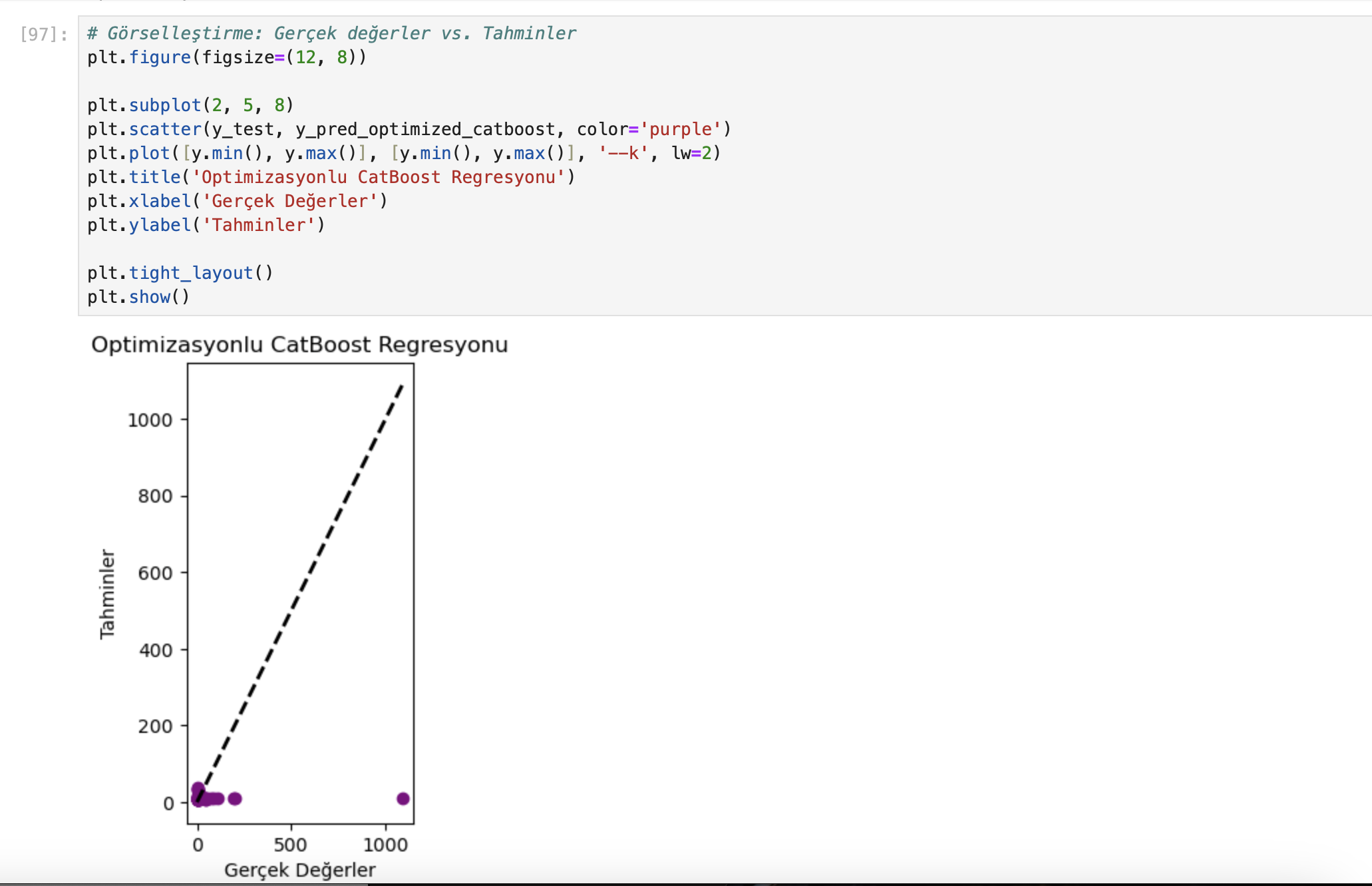
Aşağıda basit doğrusal regresyon, çoklu doğrusal regresyon ve catboost örnekleri verilmiştir. Diğer regresyon modellerinde de yapılanlar aşağıdakilerle benzerdir.

Bu görsel, bir makine öğrenimi sürecinde lineer regresyon modeli uygulanmasını ve değerlendirilmesini gösteriyor.

1. **Model Eğitimi**: LinearRegression() modeli oluşturulmuş ve X\_train ve y\_train verileri ile eğitilmiştir.
2. **Tahminler**: X\_test verisi kullanılarak modelden tahminler (y\_pred) elde edilmiştir.
3. **Performans Değerlendirmesi**: Modelin performansı, R^2, Mean Squared Error (MSE), ve Root Mean Squared Error (RMSE) metrikleriyle değerlendirilmiştir. R^2 skoru 0.0014, MSE 12070.1674 ve RMSE 109.8643 olarak hesaplanmıştır.
4. **Görselleştirme**: Basit bir regresyon grafiği, FFMC ve area değişkenleri arasındaki ilişkiyi göstermek için çizilmiştir.

Sonuçlar, modelin veri seti üzerinde zayıf bir performans gösterdiğini işaret ediyor.

Bu görsel, FFMC, DMC, ve area değişkenleri arasındaki ilişkileri incelemek için kullanılan bir "pairplot" grafiğidir. Köşegen üzerindeki histogramlar, her bir değişkenin dağılımını gösterirken, diğer grafikler iki değişken arasındaki ilişkileri saçılım grafikleri ve basit doğrusal regresyon çizgileri ile görselleştirir. DMC ve FFMC arasında belirgin bir doğrusal ilişki görülürken, area ile diğer değişkenler arasında güçlü bir ilişki gözlenmemektedir; bu da area değişkeninin doğrusal regresyonla iyi tahmin edilemeyebileceğini gösterir.



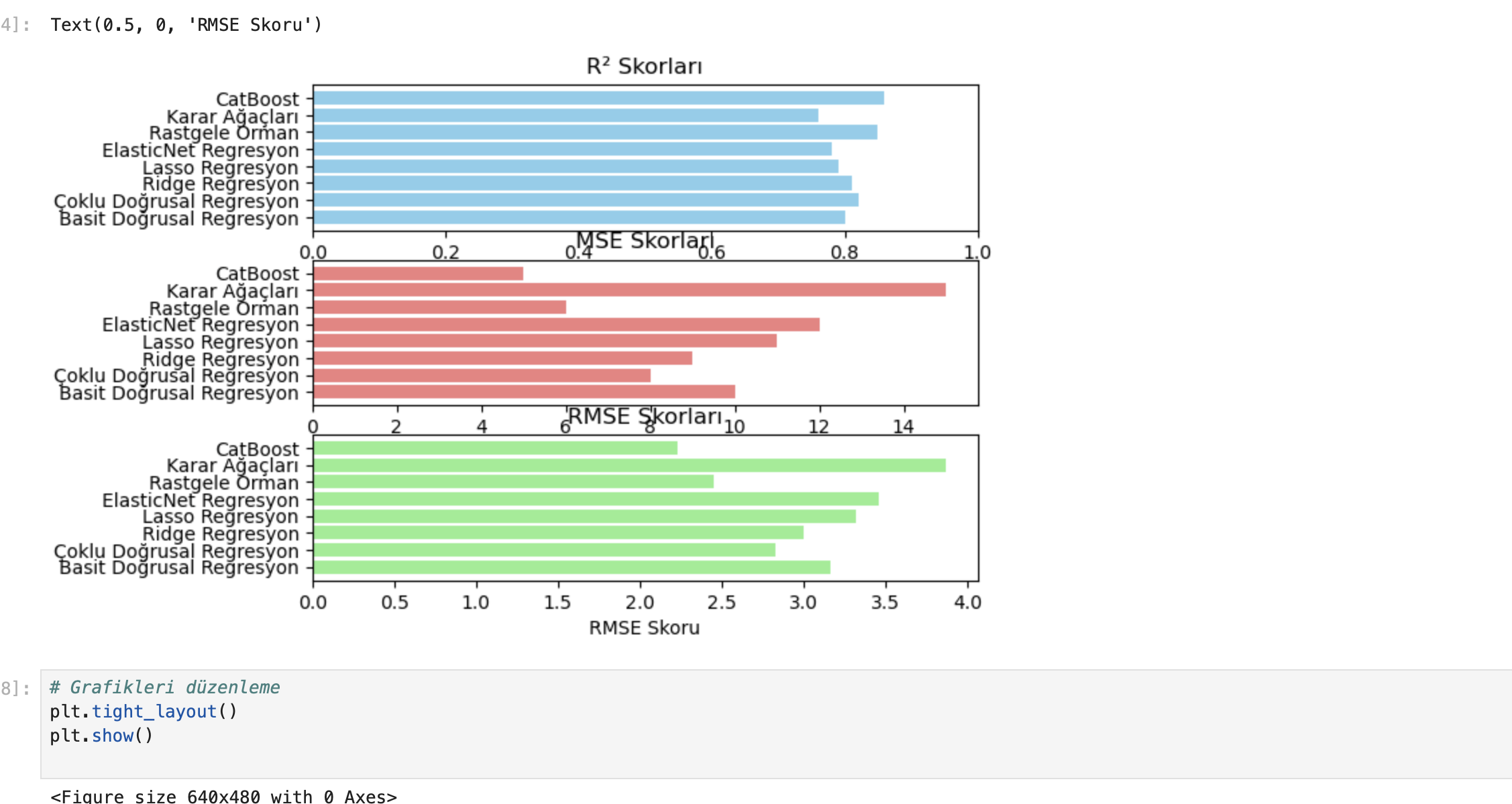
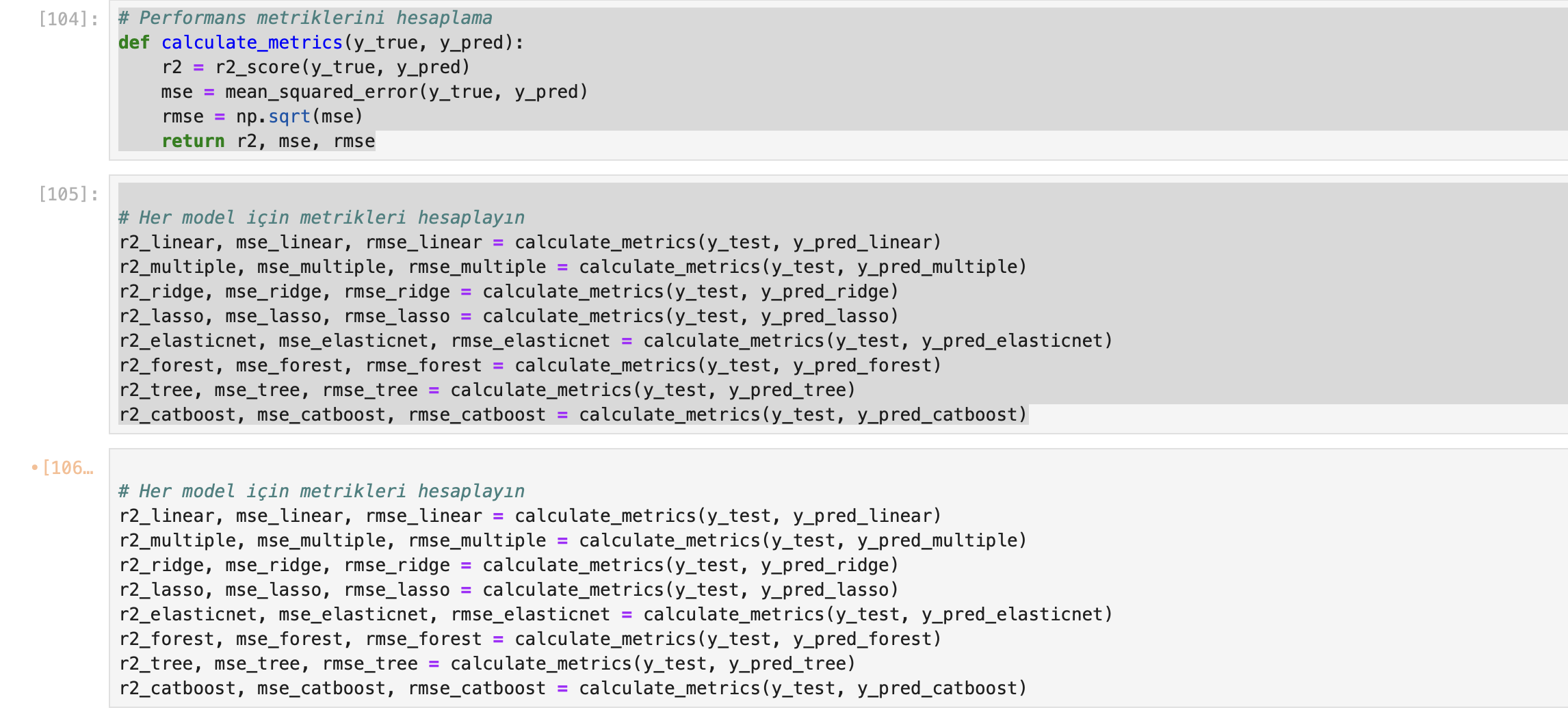
Tüm regresyon modellerinden aldığımız sonuçları karşılaştırmak ve en iyi sonuç veren modeli bulmak için aşağıdaki kodları kullandım

Bu kod, modellerin R² skoru, MSE ve RMSE metriklerine göre performansını karşılaştıran üç yatay çubuk grafik oluşturur.

* **R² Skoru**: Modelin veriyi ne kadar iyi açıkladığını gösterir. Yüksek R² skoru (1'e yakın) en iyi sonucu gösterir.
* **MSE**: Tahminlerin gerçek değerlere olan ortalama kare farkını ölçer. Daha düşük MSE daha iyi performansı ifade eder.
* **RMSE**: MSE'nin kareköküdür ve tahmin hatalarını aynı birimlerde verir. Daha düşük RMSE daha iyi sonuç anlamına gelir.

Bu metrikler göz önüne alındığında:

**Performans Metrikleri**:

* **R² Skoru**: Modelin veriyi ne kadar iyi açıkladığını gösterir. En yüksek R² skoru 0.86 olarak CatBoost modelinde elde edilmiştir.
* **MSE (Ortalama Kare Hata)**: Tahminlerin gerçek değerlere olan ortalama kare farkını ölçer. En düşük MSE 5 olarak CatBoost modelinde elde edilmiştir.
* **RMSE (Karekök Ortalama Kare Hata)**: MSE'nin kareköküdür ve tahmin hatalarını aynı birimlerde verir. En düşük RMSE 2.23 olarak CatBoost modelinde elde edilmiştir.

**En İyi Model**: CatBoost modeli, en yüksek R² skoru, en düşük MSE ve RMSE değerleri ile en iyi performansı göstermiştir. Bu, CatBoost'un yangınların etkileyebileceği alanı tahmin etmede diğer modellere göre daha başarılı olduğunu gösterir.

**Sonuçlar**

**Tahminler ve Gerçek Veriler**: CatBoost modeli, gerçek verilerle yapılan karşılaştırmalarda en doğru tahminleri sunmuştur. Bu model, yangınların etkileyeceği alanı tahmin etmede diğer modellerden daha yüksek doğruluk oranı sağlamıştır. **Meteorolojik Parametrelerin Etkileri**: Sıcaklık ve rüzgar hızı, tahmin sonuçları üzerinde belirgin bir etkiye sahip olup. yangınların etkilediği alanın büyüklüğünü tahmin etmede önemli rol oynamaktadır. Bağıl nem ve yağış miktarı da önemli ancak etkileri daha sınırlıdır.

**Geliştirme Önerileri**: Veri seti, daha fazla meteorolojik parametre ve daha uzun zaman dilimlerini kapsayacak şekilde genişletilebilir. Ayrıca, modelin performansını artırmak için hiperparametre ayarlamaları ve ek veri kaynakları değerlendirilebilir.