**VERİ YAPILARI DERSİ FİNAL PROJESİ**

**ABDULHAMİT İNCİ**

**210404806**

ARAÇ DEĞERİ BULMA

Projemde, araç fiyat tahmini ve benzer araçların bulunmasını kolaylaştırmak amacıyla bir veri seti oluşturdum. Bu veri setini oluşturmak için arabam.com sitesinden veri çektim. Kullanıcı tarafından girilen araba marka, model ve sınıf bilgileri doğrultusunda veriler toplanarak Excel dosyasına kaydedildi. Bu Excel dosyasındaki verileri kullanarak, araç fiyat tahmini yapan bir model eğittim. Model, girilen yıl ve kilometre değerine göre araç fiyat tahmini yapmaktadır.

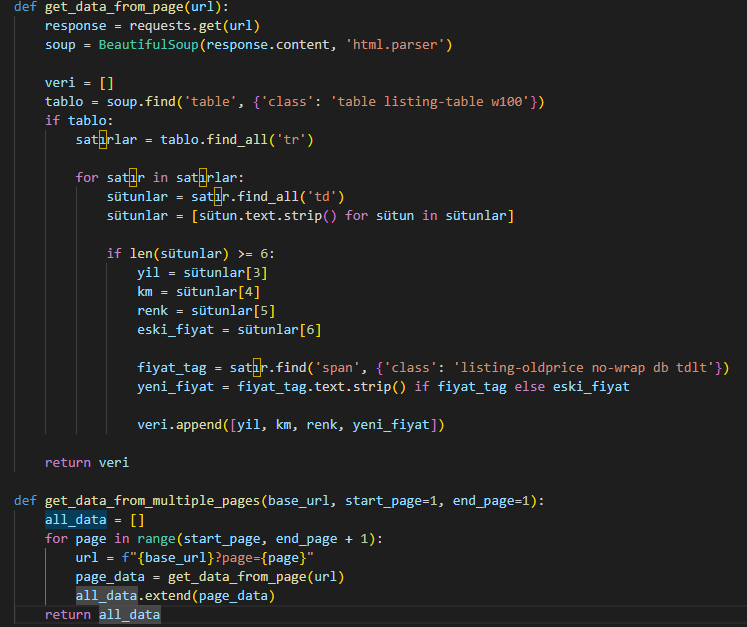
Projede ayrıca, araçların birbirine benzerliğini belirlemek için bir komşuluk matrisi oluşturulmuş ve bu matrisi farklı bir Excel dosyasına kaydedilmiştir. Komşuluk matrisinin amacı, beğendiğimiz araca yakın değerlere sahip diğer ilanları kolayca bulabilmeyi sağlamaktır. Kolaylık olması açısından, Excel dosyasında birbirine en yakın değerlere sahip araçlar otomatik olarak kırmızı renge boyanmıştır.

Projemin son aşamasında, doğrusal regresyon modeli kullanarak modelin performansını değerlendirdim. Bu değerlendirme, modelin tahmin doğruluğunu ve genel performansını ölçmeye yönelik olarak yapılmıştır.

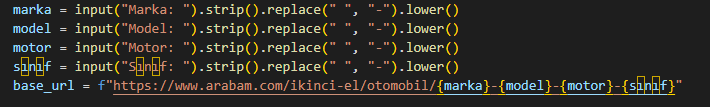
**Veri Toplama ve Hazırlama**

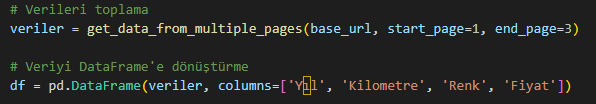
Projemde, arabam.com sitesinden araç bilgilerini çekerek bir veri seti oluşturdum. Bu veri toplama sürecini gerçekleştirmek için aşağıdaki Python kodunu kullandım. Kodun işleyişi şu şekildedir:

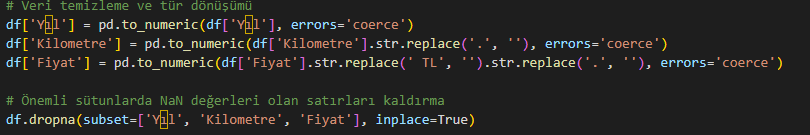
1. **Web Scraping ile Veri Toplama:**
   * **get\_data\_from\_page(url) Fonksiyonu:** Bu fonksiyon, belirli bir URL'den (sayfadan) araç bilgilerini çekmek için kullanılır. Fonksiyon, belirtilen sayfaya HTTP isteği gönderir ve sayfanın HTML içeriğini BeautifulSoup kütüphanesi ile parse eder. Araç bilgileri, HTML tablo elemanlarından çıkarılır ve listeye eklenir.
   * **get\_data\_from\_multiple\_pages(base\_url, start\_page, end\_page) Fonksiyonu:** Bu fonksiyon, birden fazla sayfadan veri toplamak için kullanılır. Belirtilen sayfa aralığında (start\_page'ten end\_page'e kadar) URL'ler oluşturulur ve her bir sayfanın verileri toplanarak birleştirilir.



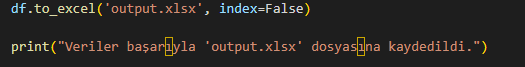
1. **Kullanıcıdan Girdi Alma:**
   * Kullanıcıdan marka, model, motor ve sınıf bilgileri alınır. Bu bilgiler, arabam.com'da ilgili sayfalara erişmek için gerekli URL'lerin oluşturulmasında kullanılır.



1. **Veri Çekme ve DataFrame'e Dönüştürme:**
   * Yukarıda belirtilen fonksiyonlar kullanılarak veriler toplanır ve pandas DataFrame formatına dönüştürülür. 
2. **Veri Temizleme ve Tür Dönüşümü:**
   * Toplanan verilerdeki gereksiz karakterler temizlenir ve veri tipleri uygun biçimlere dönüştürülür. Özellikle yıl, kilometre ve fiyat sütunları sayısal değerlere çevrilir.
   * NaN (eksik) değerler içeren satırlar veri setinden çıkarılır.

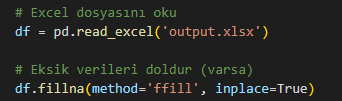


1. **Verinin Excel Dosyasına Kaydedilmesi:**
   * Temizlenmiş ve düzenlenmiş veri, output.xlsx adlı Excel dosyasına kaydedilir.



**FİYAT TAHMİNLEME**

 **Veri Okuma ve Ön İşleme:** Excel'den veri okunur (pd.read\_excel) ve eksik değerler bir önceki değerle doldurularak (fillna(method='ffill')) ön işleme yapılır.



 **Veriyi Ayırma:** Bağımsız değişkenler (X) ve bağımlı değişken (y) olarak veri ayrılır. Bağımsız değişkenler arasında aracın yılı ve kilometresi bulunurken, bağımlı değişken aracın fiyatıdır.



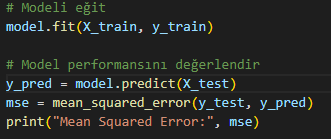
 **Eğitim ve Test Setlerine Ayırma:** Veri, eğitim ve test setlerine (train\_test\_split) ayrılır. Bu, genellikle modelin doğruluğunu değerlendirmek için kullanılır.



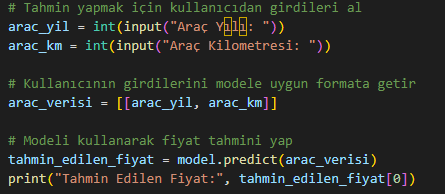
 **Model Oluşturma:** RandomForestRegressor kullanılarak bir rastgele orman regresyon modeli oluşturulur.



 **Model Eğitimi ve Performans Değerlendirmesi:** Model eğitilir (fit) ve performansı test seti üzerinde değerlendirilir. Performans, ortalama karesel hata (mean\_squared\_error) kullanılarak ölçülür.

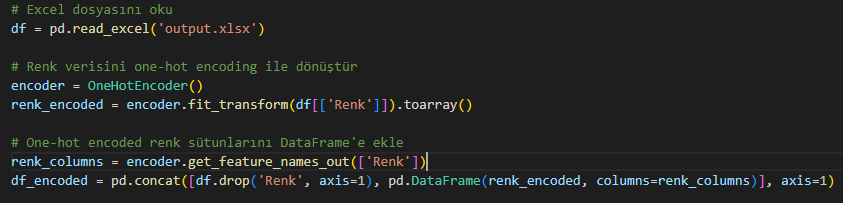


 **Tahmin Yapma:** Kullanıcıdan araçla ilgili bilgiler alınır. Bu bilgiler, modele uygun formata dönüştürülerek tahmin edilen fiyat elde edilir.

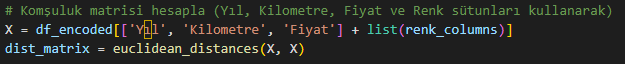


**KOMŞULUK MATRİSİ OLUŞTURMA**

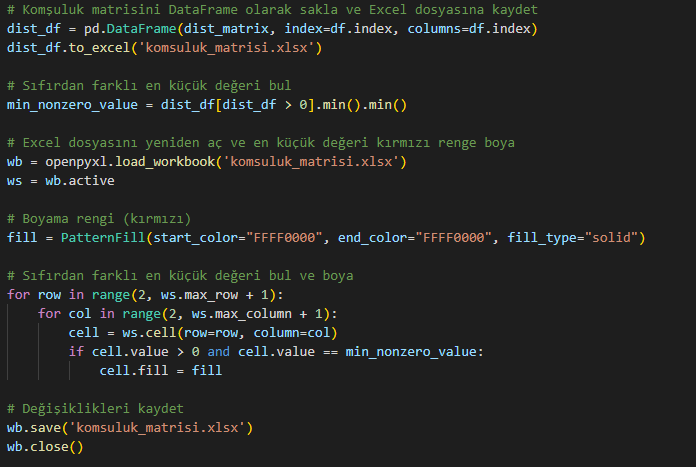
 **Veri Okuma ve Ön İşleme:** Excel dosyası okunur ve veri çerçevesine yüklenir. Ardından, araçların renk bilgisini One-Hot Encoding yöntemi ile dönüştürülür. Bu, renk bilgisini modele uygun hale getirir.



 **Komşuluk Matrisi Hesaplama:** Veri setindeki araçların özelliklerine dayalı olarak Euclidean mesafesi kullanılarak bir komşuluk matrisi hesaplanır. Bu matris, araçların birbirlerine olan benzerliklerini ölçer.



 **Excel Dosyasına Kaydetme ve Görselleştirme:** Komşuluk matrisi bir DataFrame olarak saklanır ve bir Excel dosyasına kaydedilir. Daha sonra, matrisin sıfırdan farklı en küçük değeri bulunur ve bu değer Excel dosyasında kırmızı renge boyanır.



**KOMŞULUK MATRİSİ OLUŞTURMA**

 **Veri Okuma ve Ön İşleme:** Excel dosyası okunur ve veri çerçevesine yüklenir.

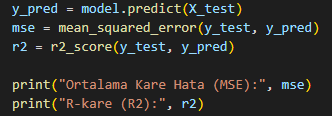


 **Verilerin Hazırlanması:** Bağımlı değişken (y) olarak araç fiyatı (Fiyat) ve bağımsız değişkenler (X) olarak aracın yılı (Yıl) ve kilometresi (Kilometre) seçilir. 

 **Verilerin Eğitim ve Test Setlerine Ayrılması:** Veri seti, eğitim ve test setlerine ayrılır. Bu, modelin doğruluğunu değerlendirmek için yapılır. 

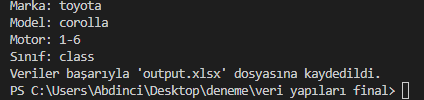
 **Lineer Regresyon Modelinin Oluşturulması ve Eğitilmesi:** Bir lineer regresyon modeli oluşturulur (LinearRegression) ve eğitilir (fit). Model, eğitim seti üzerinde eğitilir. 

 **Modelin Değerlendirilmesi:** Modelin performansı, test seti üzerindeki tahminlerin gerçek değerlerle karşılaştırılmasıyla değerlendirilir. Bu, ortalama karesel hata (MSE) ve R-kare (R2) metrikleri kullanılarak yapılır.

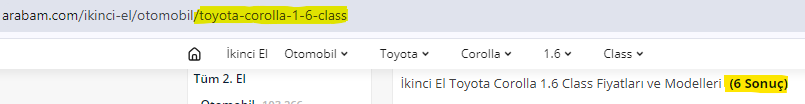


**ÖRNEK**

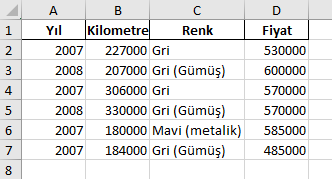
 **veri setinin hazırlanması:** kullanıcıdan alınan bilgiler sayesinde URL oluşur ve oluşan URL deki verileri çekip Excel dosyama kayıt eder.



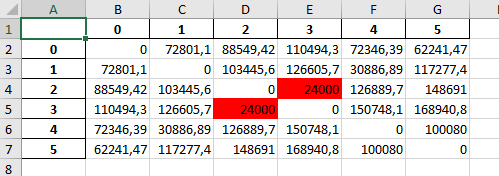
 Üstte verilen bilgiler sayesinde altta gördüğümüz siteye ulaşırız.



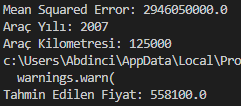
 Verilen URL deki sonuçlar Yıl, Kilometre, Renk ve Fiyat bilgileri Excel’de bir veri seti olarak kayıt olur.



 **komşuluk matrisi:** Oluşan veri setimdeki verilerin birbirine yakınlığını oluşan komşuluk matrisinden görebiliriz. Komşuluk matrisinde sıfırdan sonraki en küçük değeri boyayarak birbirine en yakın olan iki ilanı görmüş oluyoruz.



 **fiyat tahmini:** kullanıcıdan alınan bilgiler sayesinde URL oluşur ve oluşan URL deki verileri çekip Excel dosyama kayıt eder.



Bu çıktı, modelin kullanıcı tarafından verilen bilgilere dayanarak tahmin ettiği araç fiyatını gösterir. Kullanıcının girdiği araç için tahmin edilen fiyatı içerir.

**Mean Squared Error:** Bu çıktı, modelin test verisi üzerindeki performansını gösterir. Ortalama karesel hata, gerçek fiyatlar ile modelin tahmin ettiği fiyatlar arasındaki ortalama karesel farkı ölçer. Düşük bir ortalama karesel hata, modelin daha iyi performans gösterdiğini gösterir.

 **fiyat tahmini:** Bu kısım, bir lineer regresyon modeli kullanarak araç fiyatlarını tahmin etmek için eğitim ve test veri setlerine ayırır. Modelin performansını değerlendirmek için ortalama kare hata (MSE) ve R-kare (R2) metrikleri kullanılır.



**Ortalama Kare Hata (MSE):** Modelin gerçek fiyatlar ile tahmin ettiği fiyatlar arasındaki ortalama karesel farkı ölçer. MSE'nin düşük olması, modelin daha iyi performans gösterdiğini gösterir.

**R-kare (R2):** Bağımsız değişkenlerin bağımlı değişken üzerindeki varyansın yüzde kaçını açıkladığını gösterir. R2'nin değeri 1'e yaklaştıkça, modelin bağımsız değişkenlerin varyansının daha fazlasını açıkladığı anlamına gelir.