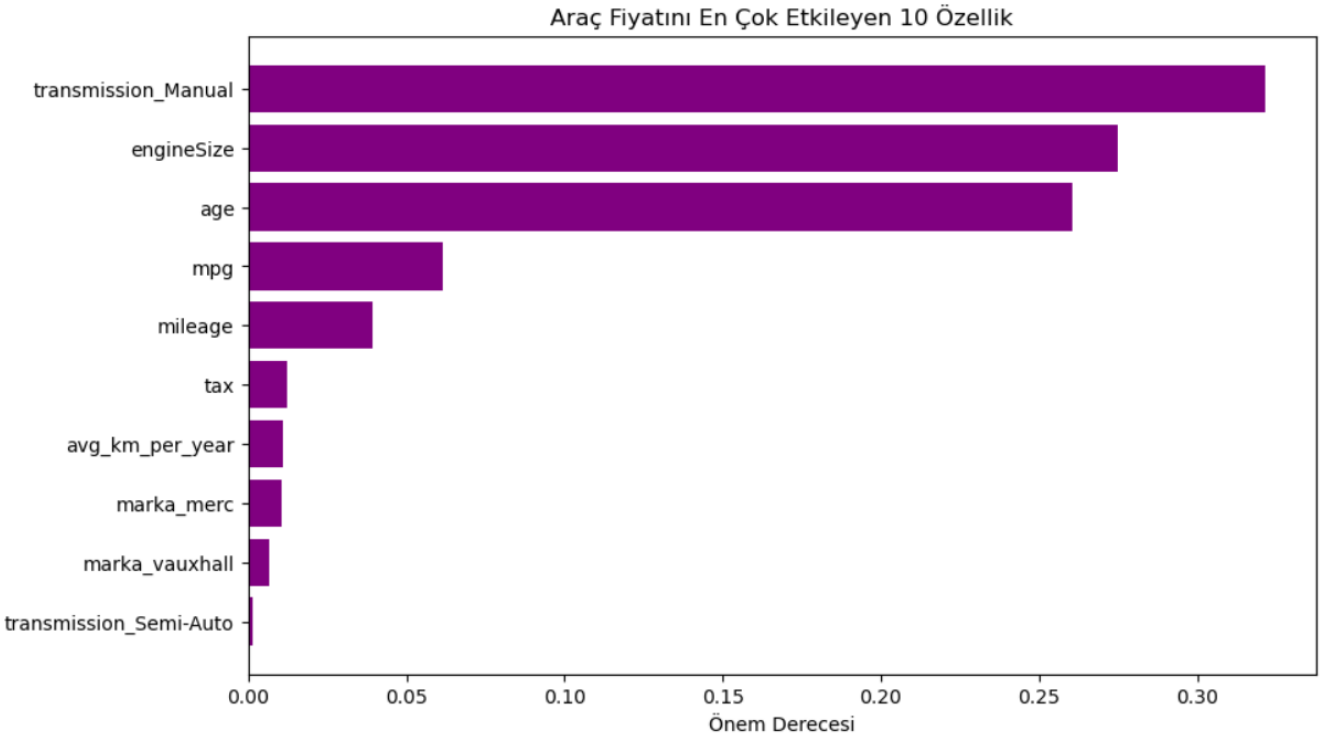
## 6. Veri Ön İşleme ve Hazırlık (Data Preprocessing)

- **Veri Temizliği ve Tekilleştirme:** Kodlama sürecinde pandas kütüphanesiyle veriyi okuduktan sonra ilk iş olarak veri setinin tutarlılığını kontrol ettik. 301 adet tekrar eden kayıt tespit ettik ve bunları sildik. Ayrıca "price", "mileage" gibi sayısal olması gereken sütunlardaki para birimi işaretlerini (£, $) ve virgülleri Regex fonksiyonları kullanarak temizledik.
- **Aykırı Değer (Outlier) Yönetimi:** Fiyat sütunundaki uç değerlerin analizi bozmaması için mantıksal filtrelemeler yaptık. 100£ altındaki hurda sayılabilecek araçlar ile 200.000£ üzerindeki ekstrem lüks araçları ve 1995 yılından eski modelleri veri setinden çıkardık.
- **Eksik Veri Stratejisi (Imputation):** Veri setindeki eksiklikler için kod içerisinde SimpleImputer kullandık. Eksik alanları, veri dağılımını bozmadan sütun ortalamalarıyla doldurarak modelin hata vermesini engelledik.
- **Öznitelik Mühendisliği (Feature Engineering):** Modelin daha iyi öğrenebilmesi için mevcut verilerden yeni özellikler türettik. 2025 yılını baz alarak "Year" sütunundan araç yaşını ("Age") hesapladık. Ayrıca kilometreyi yaşa bölerek aracın yıllık ortalama kullanımını gösteren "Avg_Km_Per_Year" adlı yeni bir sütun oluşturduk.

## 7. Modelleme Stratejisi (Modeling)

### A. Regresyon (Fiyat Tahmini) Yaklaşımı

- **Kodlama (Encoding):** Kategorik veriler (Marka, Model, Vites vb.) için kodda pd.get_dummies fonksiyonunu kullanarak One-Hot Encoding yöntemini uyguladık. Bu sayede kategorik değişkenleri modelin anlayabileceği 0 ve 1'lerden oluşan sayısal matrislere dönüştürdük. Sütun sayısı arttı ama modelin markalar arasındaki ilişkiyi yanlış yorumlamasının önüne geçtik.
- **Öznitelik Seçimi (Feature Selection):** One-Hot Encoding sonrası sütun sayısı çok arttığı (200'ün üzerine çıktığı) için SelectKBest algoritmasını devreye soktuk. İstatistiksel olarak fiyatı en çok etkileyen en iyi 10 özelliği (örneğin; motor hacmi, mpg, araç yaşı vb.) seçerek modeli sadeleştirdik.




- **Model Mimarisi ve Optimizasyon:**
    - **Base Model (Referans):** İlk etapta herhangi bir ayar yapmadan Decision Tree ve Bayesian Ridge modellerini ham haliyle eğittik.
    - **Final (Optimize) Model:** Karar Ağacı modelinin performansını artırmak ve ezberlemeyi (overfitting) önlemek için GridSearchCV ile hiperparametre optimizasyonu yaptık. Ağaç derinliği (max_depth) ve yaprak başına düşen örnek sayısı gibi parametreleri 5 katlı çapraz doğrulama ile test ettik.

**B. Kümeleme (Segmentasyon) Yaklaşımı**

- **Boyut İndirgeme (PCA):** Veri setini görselleştirebilmek ve gürültüyü azaltmak için PCA (Temel Bileşen Analizi) uyguladık. Veriyi %95 varyans koruyacak şekilde daha az boyuta indirgedik.
- **Algoritma Seçimi:** Araçları özelliklerine göre gruplamak için yoğunluk tabanlı bir algoritma olan DBSCAN'i kullandık. Bu algoritma, araçları belirli merkezlere zorlamak yerine yoğunluklarına göre doğal kümeler oluşturduğu için tercih edildi.

**8. Değerlendirme ve Sonuçlar (Evaluation)**

- **Regresyon Sonuçları ve Optimizasyon:**

- **Base Model Performansı:** İlk kurduğumuz modellerden Decision Tree, 0.89 R2 skoru ile oldukça yüksek bir başarı gösterdi. Buna karşılık Bayesian Ridge modeli 0.71 seviyesinde kalarak daha düşük performans sergiledi.
- **Optimize Modelin Başarısı:** SelectKBest ile 200 küsur özellikten sadece en önemli 10 tanesini seçip GridSearchCV ile optimize ettiğimizde, modelimiz 0.88 R2 skorunu korumayı başardı. Bu sonuç, çok daha az veri ve işlem gücüyle neredeyse aynı başarıyı yakaladığımızı, yani daha verimli bir model kurduğumuzu gösteriyor.

| Yöntem<br>Model | 1. Tüm Özellikler (Scaled) | 2. SelectKBest (10 Özellik) | 3. PCA (İndirgenmiş) |
|---|---|---|---|
| Bayesian Ridge | 0.8675 | 0.7546 | 0.7946 |
| Decision Tree | 0.8972 | 0.8333 | 0.8761 |

| | Model | Yöntem | R2 Score | MAE (Hata) | Özellik Sayısı |
|---|---|---|---|---|---|
| 0 | Decision Tree | 1. Tüm Özellikler (Scaled) | 0.897239 | 1907.356593 | 202 |
| 2 | Decision Tree | 3. PCA (İndirgenmiş) | 0.876126 | 2024.951868 | 173 |
| 3 | Bayesian Ridge | 1. Tüm Özellikler (Scaled) | 0.867499 | 2294.246005 | 202 |
| 1 | Decision Tree | 2. SelectKBest (10 Özellik) | 0.833338 | 2384.847213 | 10 |
| 5 | Bayesian Ridge | 3. PCA (İndirgenmiş) | 0.794551 | 2911.229283 | 173 |
| 4 | Bayesian Ridge | 2. SelectKBest (10 Özellik) | 0.754571 | 3205.366119 | 10 |

- **Model Açıklanabilirliği ve Güvenilirlik:**
  - **Feature Importance:** Kod çıktılarına göre fiyatı etkileyen en önemli faktörlerin başında "Motor Hacmi" (engineSize), "Araç Yaşı" (age) ve "Kilometre" (mileage) geldiğini gördük.
  - **Hata Analizi (Residual Analysis):** Tahmin hatalarını görselleştirdiğimizde (Residual Plot), hataların sıfır noktası etrafında toplandığını ve normal dağılıma yakın bir çan eğrisi oluşturduğunu gözlemledik. Bu da modelimizin güvenilir olduğunu kanıtlamaktadır.
- **Segmentasyon Çıkarımı:**
  - DBSCAN algoritması ile yaptığımız analiz sonucunda, araç piyasasında belirgin özelliklere sahip 3 ana küme (segment**)** tespit ettik. Ayrıca herhangi bir

gruba uymayan yaklaşık 29 adet aracı "aykırı değer" (noise) olarak belirledik. Bu kümeleme, pazarın genel yapısını anlamamıza yardımcı oldu.