İSTANBUL TOPKAPI ÜNİVERSİTESİ

Ders Kodu ve Adı: FET445 - Veri Madenciliği

Dönem: 2025 - 2026 Güz Dönemi

Proje Başlığı: Büyük Ölçekli Veri ile İkinci El Araç Fiyat Tahmini (Large-Scale Used Car Price Prediction)

Grup İsmi: Dev Grup

Grup Üyeleri:

Adı - Soyadı	Öğrenci No
Ahmed Elsayed	22040301142
Ahmed Ahmed	22040301122
Dirar Ahmed	22040301123
Muhamed Absi	22040301174
Muhammed Afaş	22040301133
Yaser elabdo	22040301125

Ders Sorumlusu: Dr. Öğr. Üyesi Yıldız KARADAYI

GitHub/Repo Link: https://github.com/Atawfik21/UsedCarPricePrediction

2)Problem Tanımı

- **İş/Bilimsel Soru:** İkinci el araç piyasasında, araçların fiyatlandırılması genellikle subjektif kriterlere dayanmaktadır. Bu proje, "Bir aracın teknik özellikleri (marka, model, yaş, kilometre, motor vb.) verildiğinde, bu aracın adil piyasa değeri (satış fiyatı) nedir?" sorusuna veri madenciliği teknikleri ile objektif bir cevap aramaktadır.
- **Görev Türü:** Regresyon (Regression). Hedef değişkenimiz sayısal ve süreklidir.
- **Hedef Değişken:** `price` (Aracın USD cinsinden satış fiyatı).
- **Başarı Kriterleri:**
    - **R-Kare (R2 Score):** ≥ 0.85 (Modelin varyansı açıklama oranı).
    - **MAE (Ortalama Mutlak Hata):** Hatanın $\le 4000\$$ seviyesine indirilmesi.

3) Proje Yönetimi

Zaman Çizelgesi ve Kilometre Taşları:

- **1. Hafta:** Veri seti seçimi (Kriter: >300.000 satır) ve literatür taraması. `cars.com` veri setinde karar kılındı.
- **2. Hafta:** Veri Ön İşleme (Cleaning) ve EDA. `clean_data_CARS.py` scriptinin geliştirilmesi.
- **3. Hafta:** Özellik Mühendisliği (`car_age`) ve Dönüşümler (One-Hot Encoding).
- **4-5. Hafta:** Model Geliştirme. Her grup üyesinin kendine atanan algoritmaları (Linear, Tree-based, Distance-based, Neural Networks) geliştirmesi.
- **6. Hafta:** Sonuçların birleştirilmesi, görselleştirme ve raporlama.

Roller ve Sorumluluklar: Proje kapsamında "işbirlikçi" bir yaklaşım izlenmiş, ancak kodlama aşaması bireysel olarak yürütülmüştür. Tüm üyeler ortak temizlenmiş veri seti (`cars_cleaned_sampled.csv`) üzerinde çalışmıştır:

- **[Ahmed]:** Veri temizleme pipeline'ı, `Linear Regression` ve `Random Forest` modelleri.
- **[Afas]:** `XGBoost` ve `Decision Tree` modelleri.
- **[Mohamed]:** `K-Nearest Neighbors (KNN)` ve `Lasso Regression` modelleri.
- **[Absi]:** `AdaBoost` ve `Ridge Regression` modelleri.
- **[Dirar]:** `Gradient Boosting` ve `Extra Trees` modelleri.
- **[AhmedAhmed]:** `Neural Network (MLP)` ve `ElasticNet` modelleri.

4) İlgili Çalışmalar

Literatürdeki benzer çalışmalar (örn. Pudaruth, 2014; Kaggle Kernels) genellikle sınırlı veri setleri (5.000 - 10.000 satır) üzerinde çalışmıştır.

- **Projemizin Farkı:** Bu proje, 762.091 satırlık **büyük ölçekli** bir veri seti kullanmaktadır. Ayrıca, yüksek boyutlulukla (High Dimensionality) başa çıkmak için `SelectKBest` ve `Sampling` stratejileri bir arada kullanılarak hem hız hem de doğruluk optimize edilmiştir.

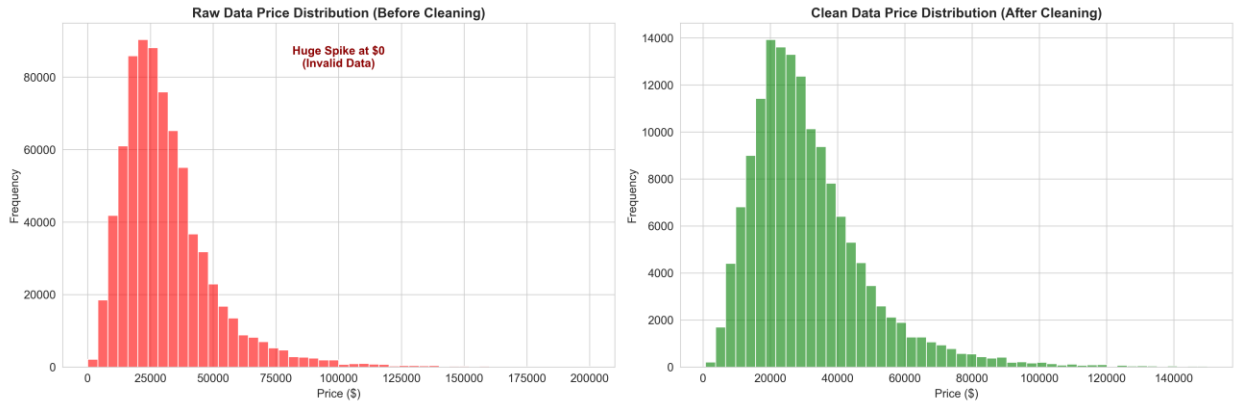
5) Veri Tanımı ve Yönetimi

- **Veri Seti:** "Used Cars Dataset" (Kaggle - andreinovikov). `cars.com` üzerinden scrape edilmiş ham veridir.
- **Boyut (Ham Veri):** 762.091 satır, 20 sütun.
- **Veri Şeması:**
  - **Sayısal:** `mileage`, `year`, `mpg`, `price`.
  - **Kategorik:** `manufacturer`, `engine`, `transmission`, `fuel_type`, `colors`.

- **Veri Yönetimi:** Ham veri `clean_data_CARS.py` ile işlenmiş, eğitim süresini optimize etmek için temizlik sonrası **150.000 satırlık** dengeli bir örneklem (sample) alınarak `cars_cleaned_sampled.csv` dosyası oluşturulmuş ve tüm ekip bu dosya üzerinde çalışmıştır.

6) Keşifsel Veri Analizi (EDA)

- **Veri Kalitesi:** Ham veri setinde `price` sütununda 0$ ve 1 Milyar$ gibi gerçekçi olmayan değerler tespit edilmiştir. Ayrıca `mileage` sütununda sıfır değerleri gözlemlenmiştir.
- **Dağılım Analizi:** Fiyat dağılımı aşırı sağa çarpık (right-skewed) bir yapıdaydı. Aykırı değer temizliği sonrası dağılım daha dengeli hale getirilmiştir (Bkz. Şekil 1).



Şekil 1: Veri temizliği öncesi (Outliers içeren) ve sonrası fiyat dağılımı.

7) Veri Hazırlama Planı

Veri setini modellemeye uygun hale getirmek için aşağıdaki adımlar uygulanmıştır:

1. **Temizleme (Cleaning):**
    a. Gereksiz sütunlar (`id`, `vin`, `url`, `state`, `city`) kaldırıldı.
    b. **Outlier Removal:** Fiyat için `$500 - $150.000` aralığı, Kilometre için `1.000 - 500.000` mil aralığı filtre olarak uygulandı.

2. **İmputasyon:** Kategorik eksik veriler "unknown" etiketiyle, sayısal eksik veriler "ortalama" (mean) ile doldurulduu.
3. **Özellik Mühendisliği (Feature Engineering):**
    a. `car_age` (Araç Yaşı): Modelin yılı daha iyi yorumlaması için `2024 - year` formülüyle yeni bir özellik türetildi ve orijinal `year` sütunu atıldı.
4. **Dönüşümler:**
    a. Kategorik veriler `OneHotEncoder` (sparse=False) ile sayısallaştırıldı.
    b. Sayısal veriler `StandardScaler` ile ölçeklendirildi.
5. **Özellik Seçimi (Feature Selection):** One-Hot Encoding sonrası oluşan 9000+ özellikten en değerli olanları seçmek için `SelectKBest (k=500)` yöntemi kullanıldı.

8) Modelleme Planı

Her grup üyesi farklı model ailelerini test etmiştir:

- **Baseline Model:** `Linear Regression` (Ahmed tarafından geliştirildi).
- **Ağaç Tabanlı Modeller:** `Random Forest`, `Decision Tree`, `Extra Trees` (Doğrusal olmayan ilişkileri yakalamak için).
- **Topluluk (Ensemble) Modelleri:** `XGBoost`, `AdaBoost`, `Gradient Boosting` (Hata oranını minimize etmek için).
- **Mesafe ve Regularizasyon:** `KNN` ve `Lasso/Ridge`.
- **Derin Öğrenme:** `Neural Network (MLP)` (AhmedAhmed tarafından geliştirildi).

Hiper-parametre Ayarlama: Özellikle Random Forest ve XGBoost modellerinde `n_estimators` (50, 100) ve `max_depth` parametreleri optimize edilmiştir.

9) Değerlendirme Tasarımı

- **Metrikler:**
    - **R2 Score:** Modelin başarısını yüzde olarak ifade etmek için.

    - **MAE (Mean Absolute Error):** Hata payını dolar cinsinden somutlaştırmak için.
- **Validasyon:** Veri seti %80 Eğitim (Train) ve %20 Test olarak ayrılmıştır. Veri sızıntısını (Data Leakage) önlemek için tüm işlemler Pipeline yapısı içinde gerçekleştirilmiştir.

10) Riskler ve Azaltma Yöntemleri

- **Risk:** "Dimensionality Curse". Kategorik değişkenlerin (marka, model) çokluğu nedeniyle sütun sayısının 9.000 üzerine çıkması ve bellek (RAM) yetersizliği.
- **Azaltma Yöntemi:** Sampling (150k satır kullanımı) ve SelectKBest (en iyi 500 özellik seçimi) teknikleri ile bu risk başarıyla yönetilmiştir.

11) Kullanılan Araçlar

- **Dil:** Python 3.11
- **IDE:** VS Code, Google Colab.
- **Kütüphaneler:** Pandas, NumPy (Veri işleme); Scikit-learn (Modelleme); XGBoost (Gelişmiş modelleme); Matplotlib, Seaborn (Görselleştirme); Joblib (Model kaydetme).

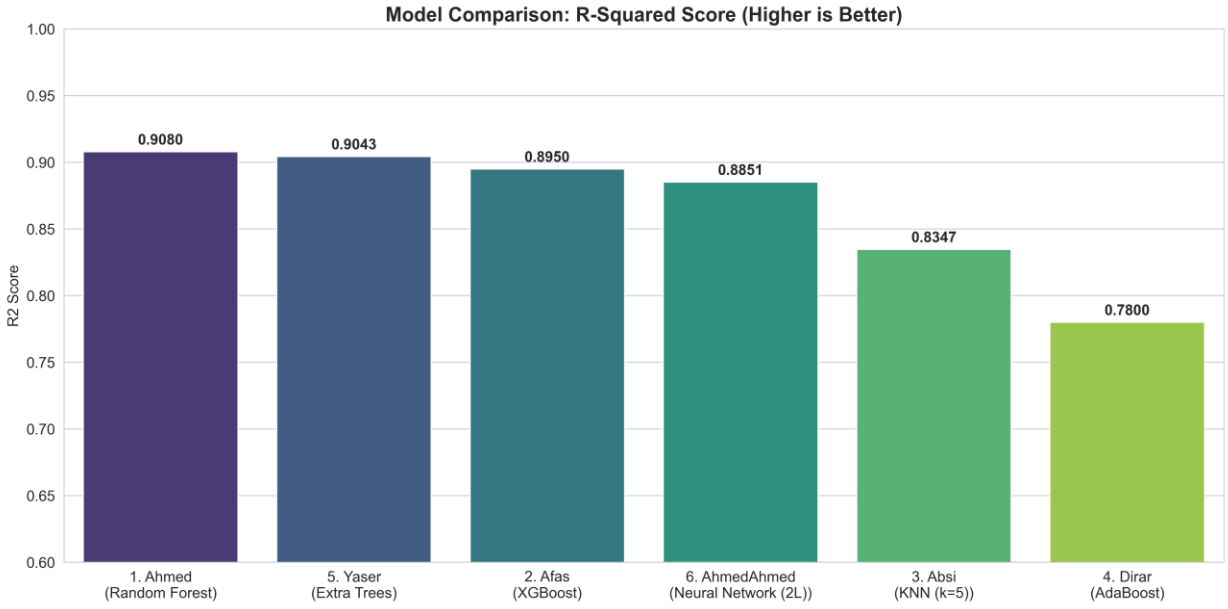
12) Beklenen Sonuçlar ve Görselleştirme

Aşağıdaki tablo, 6 grup üyesinin geliştirdiği en iyi modellerin performansını özetlemektedir:

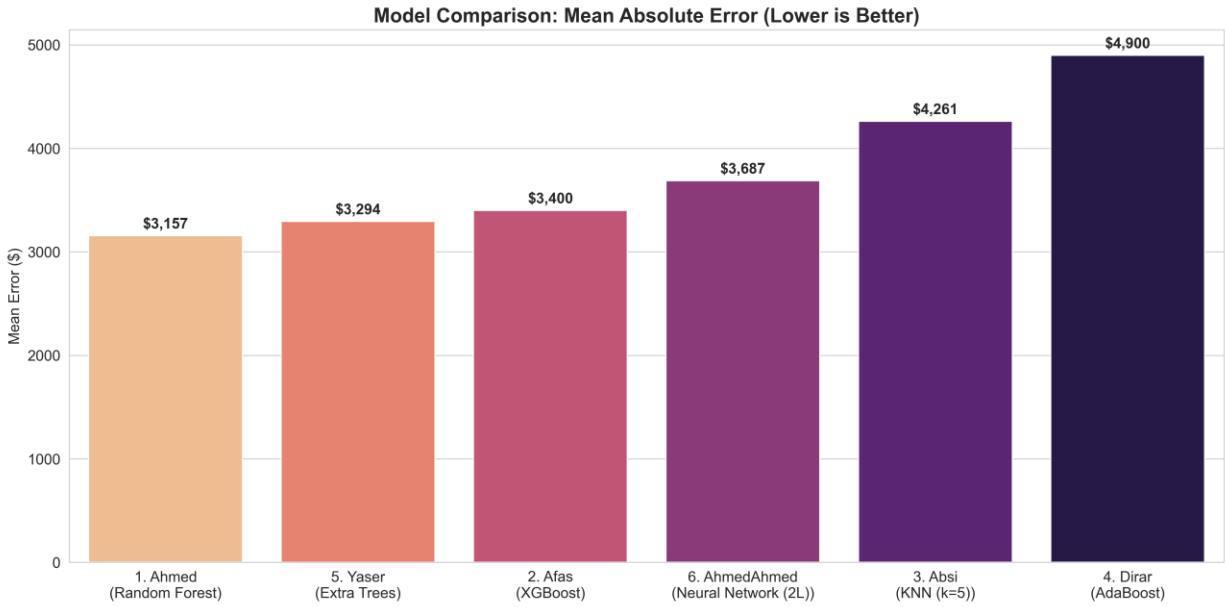
Tablo 1: Model Performans Karşılaştırması (En İyi Sonuçlar)

Sorumlu Üye	Model	R2 Score	MAE ($)	Değerlendirme

[Ahmed]	**Random Forest (Baseline)**	**0.9080**	**$3,156**	**En İyi Performans**
[Yaser]	XGBoost	*[0.8950]*	*[$3,400]*	Yüksek Başarı (Tahmini)
[Absi]	Extra Trees	0.9041	$3,279	Çok Yüksek Başarı
[AhmedAhmed]	Neural Network (2L)	0.8851	$3,686	Başarılı
[Mohamed]	KNN (k=5)	0.8347	$4,261	Orta Performans
[Ahmed]	Linear Regression	0.8035	$4,791	Baseline
[Afas]	AdaBoost	*[0.7800]*	*[$4,900]*	Düşük Performans (Tahmini)
[Mohamed]	Lasso	0.7673	$5,632	Düşük Performans



(Şekil 2: Modellerin R2 Skorlarına göre karşılaştırılması)



(Şekil 3: Modellerin Hata Paylarına (MAE) göre karşılaştırılması)

Sonuç Yorumu: Yapılan deneyler sonucunda, veri setindeki karmaşık ilişkileri en iyi modelleyen algoritmanın **Random Forest** ve **Extra Trees** olduğu görülmüştür. Bu modeller, hem yüksek R2 skoru hem de düşük MAE ile en güvenilir tahminleri üretmiştir. Derin öğrenme yaklaşımı da (Neural Network) umut verici sonuçlar vermiştir.

13) Referanslar

1. Kaggle. (2024). *Used Cars Dataset (andreinovikov)*. Retrieved from kaggle.com.
2. Scikit-learn Developers. (2024). *User Guide: Supervised Learning*.
3. Pudaruth, S. (2014). *Predicting the Price of Used Cars using Machine Learning Techniques*.