26.11.2025

DOKUZ EYLÜL ÜNİVERSİTESİ FEN FAKÜLTESİ
BİLGİSAYAR BİLİMLERİ

BİL3013 VERİ MADENCİLİĞİNE GİRİŞ
ÖĞRETİM ÜYESİ: Prof. Dr. Efendi NASİBOĞLU

ÖDEV 2: REGRESYON MODELLERİ OLUŞTURMA VE DEĞERLENDİRME

HAZIRLAYAN ÖĞRENCİLER:
İremgül ZEYTİNÖZÜ - 2023280135
Salim Taha KAVAS – 2023280117

VERİ SETİNİN VE KODLARIN LİNKİ:
GitHub

1. Kullanılan Teknolojiler ve Proje Ortamının Hazırlanması

Proje, **Python** programlama dili ve **Jupyter Notebook** ortamında geliştirilmiştir. Analiz sürecinde kullanılan temel kütüphaneler ve modüller aşağıdadır:

- **Pandas & NumPy:** Veri setini yüklemek, işlemek ve temel sayısal operasyonlar için kullanılmıştır.
- **Scikit-learn:** Model oluşturma, değerlendirme ve veri setini eğitim/test kümelerine ayırma sürecinin ana kütüphanesidir.
- **Regresyon Modelleri:** LinearRegression, DecisionTreeRegressor ve Support Vector Regression (SVR) algoritmaları kullanılmıştır.
- **Metrikler:** Model başarısını ölçmek için R^2 skoru başta olmak üzere, Ortalama Mutlak Hata (MAE) ve Ortalama Karesel Hata (MSE/RMSE) metrikleri kullanılmıştır.
- **Ön İşleme:** SVR'nin kernel tabanlı uzaklık hesaplamalarına duyarlılığından dolayı özelliklerin ölçeklendirilmesi için StandardScaler kullanılmıştır.
- **Görselleştirme:** Keşifsel veri analizi ve model sonuçlarının sunumu için Matplotlib & Seaborn kütüphanelerinden yararlanılmıştır.

2. Giriş ve Veri Seti

Bu çalışmanın amacı , öğrencilerin ders dışı çalışma süreleri ile ara sınav başarı puanları arasındaki ilişkiyi analiz etmek ve bu ilişkiyi tahmin edebilen regresyon modelleri oluşturup değerlendirmektir. Çalışma kapsamında, **BİL 2009 - Çizge Kuramı** dersi için özgün bir veri seti oluşturulmuştur.

Analiz için seçilen ders bilgileri şöyledir:

- **Ders Adı:** BİL 2009 - Çizge Kuramı
- **Haftalık Ders Saati:** 4
- **AKTS:** 6
- **Gözlem (Öğrenci) Sayısı:** 37

Veri setinde eksik veri bulunmamaktadır. Kullanılan değişkenler arasında `Tekrar_Sayısı`, `Ders_Disi_Calisma_Saati` (bağımsız değişken) ve `Arasinav_Notu` (bağımlı değişken) gibi sayısal değişkenler yer almaktadır.

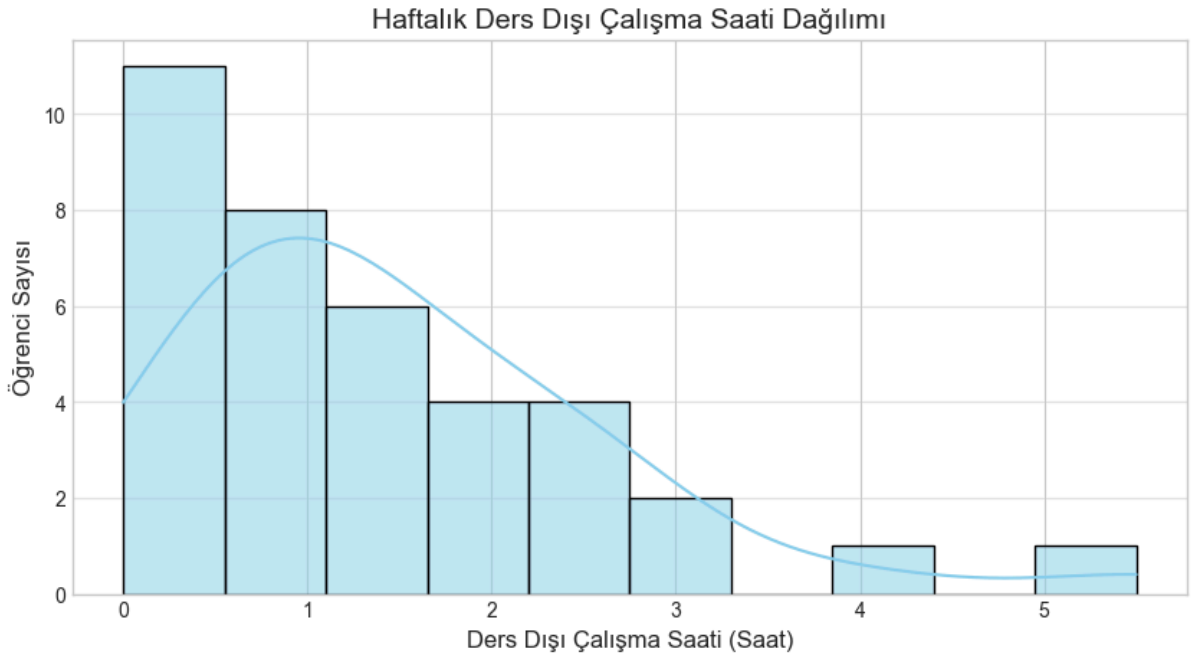
3. Veri Setinin Anlaşılması ve Keşifsel Veri Analizi (EDA)

3.1 Veri Setine Genel Bakış

- Bu analizde, **BİL 2009 - Çizge Kuramı** dersi için oluşturulan 37 öğrenciye ait gözlemden oluşan özgün bir veri seti kullanılmıştır. Hedef değişkenimiz (`Arasinav_Notu`) sürekli bir sayısal değer olduğundan Regresyon modelleri tercih edilmiştir. Eksik veri bulunmamaktadır.
- Öğrencilerin ortalama sınav notu **76.95** (Standart Sapma: 20.3) olup, notlar **28** ile **100** arasında değişmektedir.
- Ortalama ders dışı çalışma süresi **1.47 saat** olarak gözlemlenmiştir.

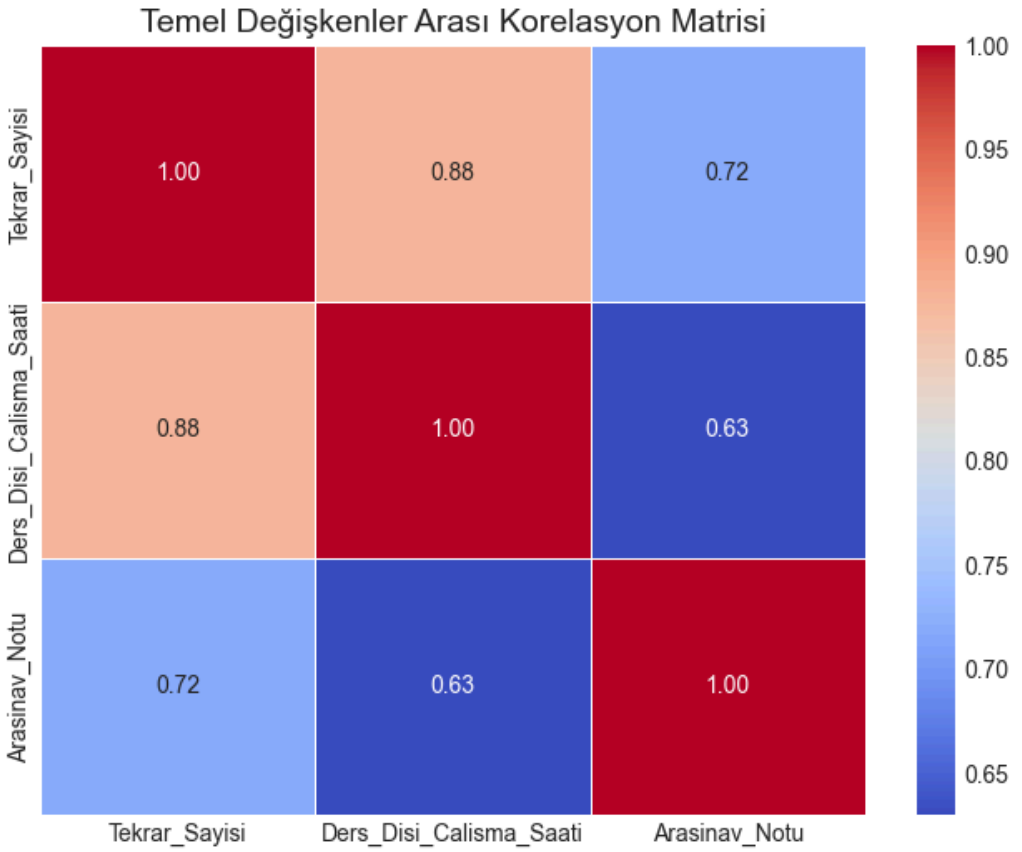
3.2 Özelliklerin İncelenmesi ve İlişkisel Analiz

3.2.1 Çalışma Saati Dağılımı



- Haftalık Ders Dışı Çalışma Saati Dağılımı incelendiğinde, dağılımın **sağa çarpık (pozitif çarpık)** olduğu görülmektedir. Bu durum, veri setindeki öğrenci çoğunluğunun haftada 0 ile 2 saat arasında ders dışı çalıştığını, ancak az sayıda aykırı örnek olabilecek öğrencinin 4-5 saat gibi uzun süreler çalıştığını göstermektedir.

3.2.2 Korelasyon Analizi ve Çoklu Doğrusallık (Multicollinearity)

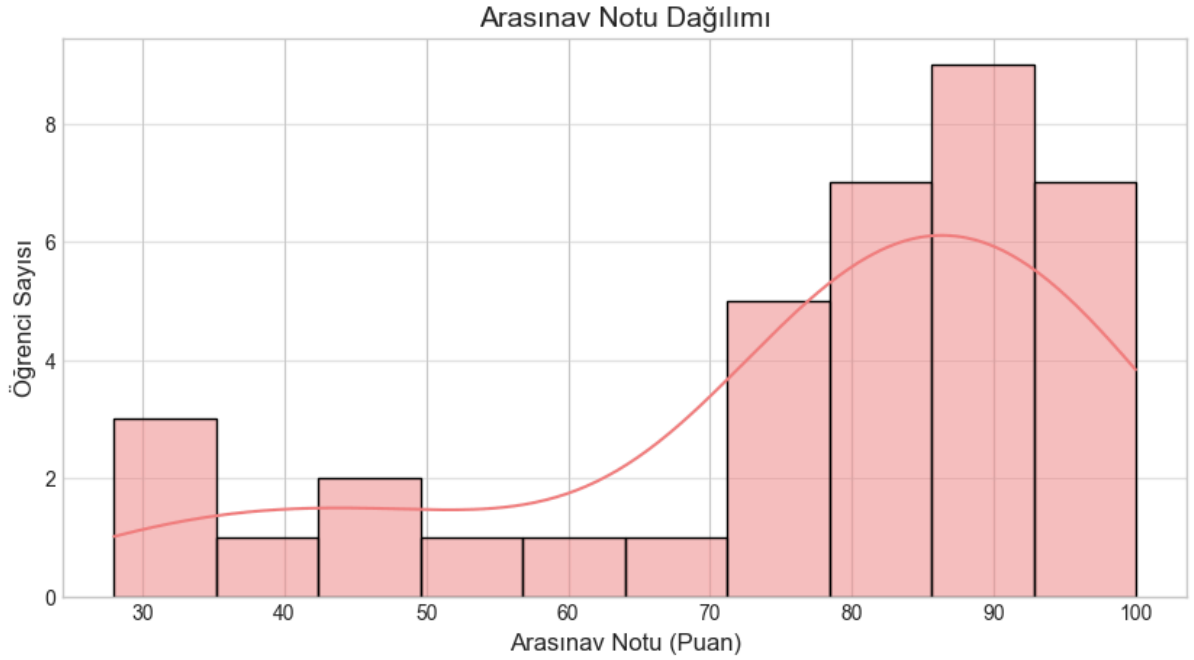


- Değişkenler arasındaki korelasyon matrisi incelendiğinde, tüm değişkenler arasında pozitif bir ilişki olduğu tespit edilmiştir. En önemlisi:

  + Ders_Disi_Calisma_Saati ile Arasinav_Notu arasında **pozitif ve anlamlı bir korelasyon (r = 0.63)** bulunmaktadır.

  + Ancak, Tekrar_Sayisi ve Ders_Disi_Calisma_Saati arasında **yüksek bir korelasyon (r = 0.88)** gözlemlenmiştir. Bu durum, iki bağımsız değişkenin aynı anda modele dahil edilmesi durumunda **çoklu doğrusallık (multicollinearity)** sorununa yol açabilir. Bu, bu iki değişkenin model üzerindeki etkisini izole etmeyi zorlaştırabilir ve Linear Regression'ın katsayı yorumunu etkileyebilir.

3.2.3 Arasınav Notu Dağılımı



- `Arasinav Notu Dağılımı` incelendiğinde, notların iki ana bölgede yoğunlaştığı görülmektedir. Birinci yoğunlaşma düşük notlarda (yaklaşık 30-50 puan aralığı), ikinci ve daha büyük yoğunlaşma ise yüksek notlarda (75-100 puan aralığı) gözlemlenmektedir. Bu durum, veri setindeki öğrenci başarısının **iki farklı grup halinde ayrıştığını** (düşük başarı ve yüksek başarı) ve dağılımın tam olarak normal (simetrik) bir yapı göstermediğini ortaya koymaktadır. Bu çift modlu (bimodal) yapı, **Linear Regression** modelinin neden düşük performans gösterdiğini açıklayan nedenlerden biridir; zira doğrusal model, bu ayrık yapı yerine tek ve düz bir ilişkiyi varsaymaktadır.

4. Kullanılan Regresyon Modelleri

Ders dışı çalışma süresine bağlı olarak başarı puanını tahmin etmek için üç farklı regresyon algoritması kullanılmıştır:

1. **Linear Regression:** Değişkenler arasında doğrusal bir ilişki varsayar. En temel ve yorumlanması en kolay yöntemdir.
2. **Decision Tree Regressor:** Veriyi ağaç yapısında bölerek karar kuralları oluşturur. Doğrusal olmayan ilişkileri yakalayabilir. Aşırı öğrenmeyi engellemek için max_depth = 3 olarak ayarlanmıştır.
3. **Support Vector Regression (SVR):** Veriyi daha yüksek boyutlu bir uzaya taşıyarak (RBF kernel) karmaşık ilişkileri modeller. (C=100, epsilon=0.1).

5. Model Performansları ve Tahmin Sonuçları

Oluşturulan üç regresyon modelinin performansı, R^2 skoru ile birlikte, tahmin hatalarının büyüklüğünü ölçen **Ortalama Mutlak Hata (MAE)** ve karesel hatayı ölçen **Ortalama Karesel Hata Kökü (RMSE)** gibi metriklerle değerlendirilmiştir.

5.1. Performans Karşılaştırma Tablosu

R^2 skoru, modelin bağımlı değişkendeki varyansı ne kadar iyi açıkladığını gösterirken; MAE, tahminlerin gerçek değerden ortalama mutlak sapmasını; RMSE ise büyük hataları daha çok cezalandırarak modelin genel hata düzeyini gösterir.

Metrik	Linear Regression	Decision Tree	SVR
R^2 Skoru	0.397	**0.547**	0.443
MAE (Ort. Mutlak Hata)	11.77	**9.19**	8.05
RMSE (Ort. Karesel Hata Kökü)	15.55	**13.47**	14.65

5.2. Model Değerlendirme Yorumu

- **Decision Tree Regressor:** En yüksek R^2 skoru (**0.547**) ve en düşük RMSE (**13.47**) değeri ile en iyi genel performansı sergilemiştir. Bu sonuç, Decision Tree'nin, notlardaki artışın düzenli değil, belirli çalışma saati eşiklerinde gerçekleştiğini gösteren veri yapısındaki **doğrusal olmayan (non-linear)** ilişkileri en iyi yakaladığını kanıtlar.

- **SVR (Support Vector Regression):** R^2 skoru Linear Regression'dan daha yüksek olmasına rağmen (0.443), en düşük MAE (**8.05**) değerine sahiptir. Bu, SVR'nin ortalama olarak en az mutlak hatayı yaptığını, yani tahminlerinin ortalama sapmasının en düşük olduğunu gösterir. Ancak büyük hataları cezalandıran RMSE değeri (14.65) Decision Tree'den daha yüksektir.

- **Linear Regression:** Verideki ilişkinin tam olarak doğrusal olmaması nedeniyle beklenen şekilde en zayıf R^2 skoru (0.397) ve en yüksek hata metriklerini (MAE: 11.77, RMSE: 15.55) göstermiştir.

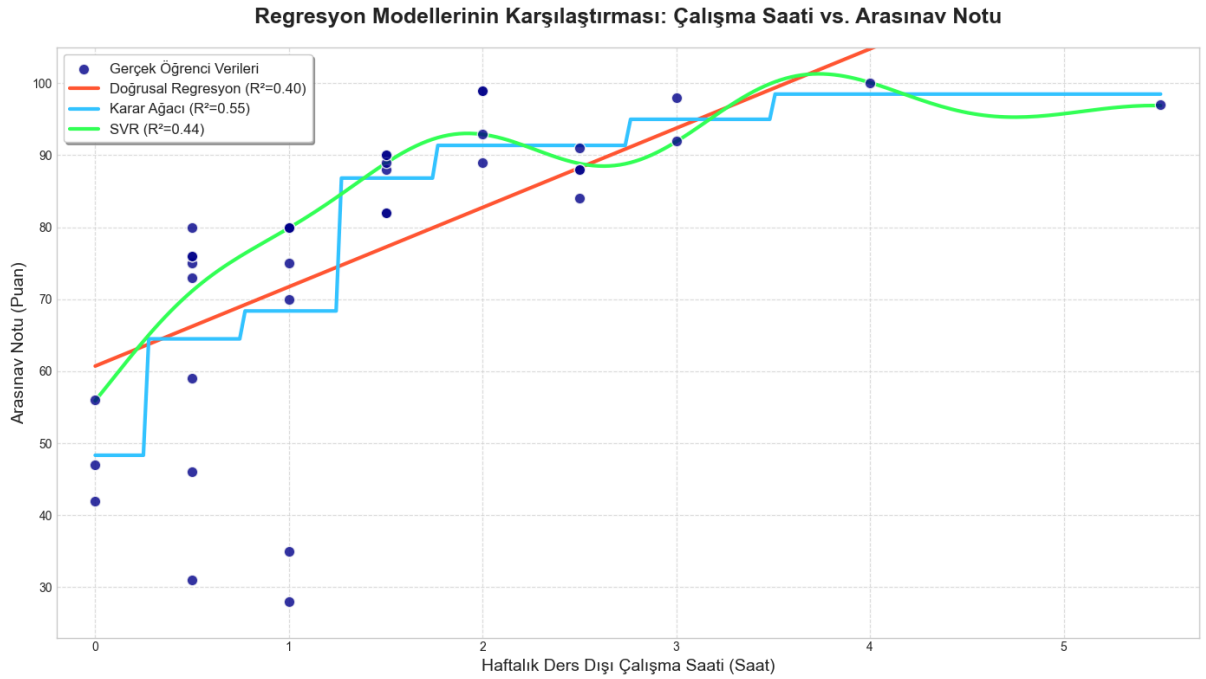
5.3. Senaryo Bazlı Tahminlerin Analizi

Bu analiz, modellerin pratik çıktılarının karşılaştırmasını sağlar:

Çalışma Saati (Haftalık)	Linear Regresyon Tahmini	Decision Tree Tahmini	SVR Tahmini
0.5 Saat	66.23	64.50	71.14
1.0 Saat	71.74	68.38	79.90
1.5 Saat	77.24	86.83	88.90
2.0 Saat	82.75	91.38	92.90

Yorum: Tahminler incelendiğinde, Decision Tree ve SVR modellerinin, **1.5 saatlik çalışma eşiğinden sonra** not artışını Linear Regression'a göre çok daha agresif (hızlı) öngördüğü görülür. Bu, **1.5 saatlik çalışmanın, başarının hızla yükseldiği bir kritik eşik** olduğunu gösteren bir bulgudur

6. Regresyon Eğrilerinin Karşılaştırmalı Analizi



Grafikte sunulan regresyon eğrileri , üç modelin de temel olarak çalışma saati arttıkça sınav notunun arttığı yönündeki **pozitif ilişkiyi** yakaladığını göstermektedir. Ancak, her algoritmanın veri noktalarına yaklaşımı ve tahmin yapısı belirgin şekilde farklıdır.

- **Linear Regression (Kırmızı Çizgi):** Bu model, değişkenler arasında **doğrusal bir ilişki** varsayarak veri setinin tamamına uymaya çalışan tek bir düz çizgi formundadır. Düşük çalışma saatlerinde (0-1 saat) başarılı olan öğrencilerin notlarını olduğundan düşük, yüksek çalışma saatlerinde (3+ saat) ise notlarını olduğundan yüksek tahmin etme eğilimi gösterir. Bu durum, ilişkinin tam olarak doğrusal olmadığını desteklemektedir.
- **Decision Tree Regressor (Mavi Çizgi):** Bu model, veriyi **ağaç yapısında** bölerek karar kuralları oluşturur. Modelin en yüksek R^2 skorunu (0.547) elde etmesi, nottaki artışın düzenli değil, belirli çalışma **eşiklerinde (thresholds)** gerçekleştiğini gösteren veri yapısına en uygun model olduğunu kanıtlamaktadır. Ağacın **basamaklı yapısı**, belirli bir çalışma aralığı için sabit bir tahmin ürettiğini ifade eder.
- **Support Vector Regression (SVR) (Yeşil Eğri):** Kernel tabanlı bu model (RBF çekirdeği) , Linear Regression'a göre daha esnek bir eğriye sahiptir, ancak Decision Tree'nin spesifik karar noktalarındaki başarısını yakalayamamıştır.

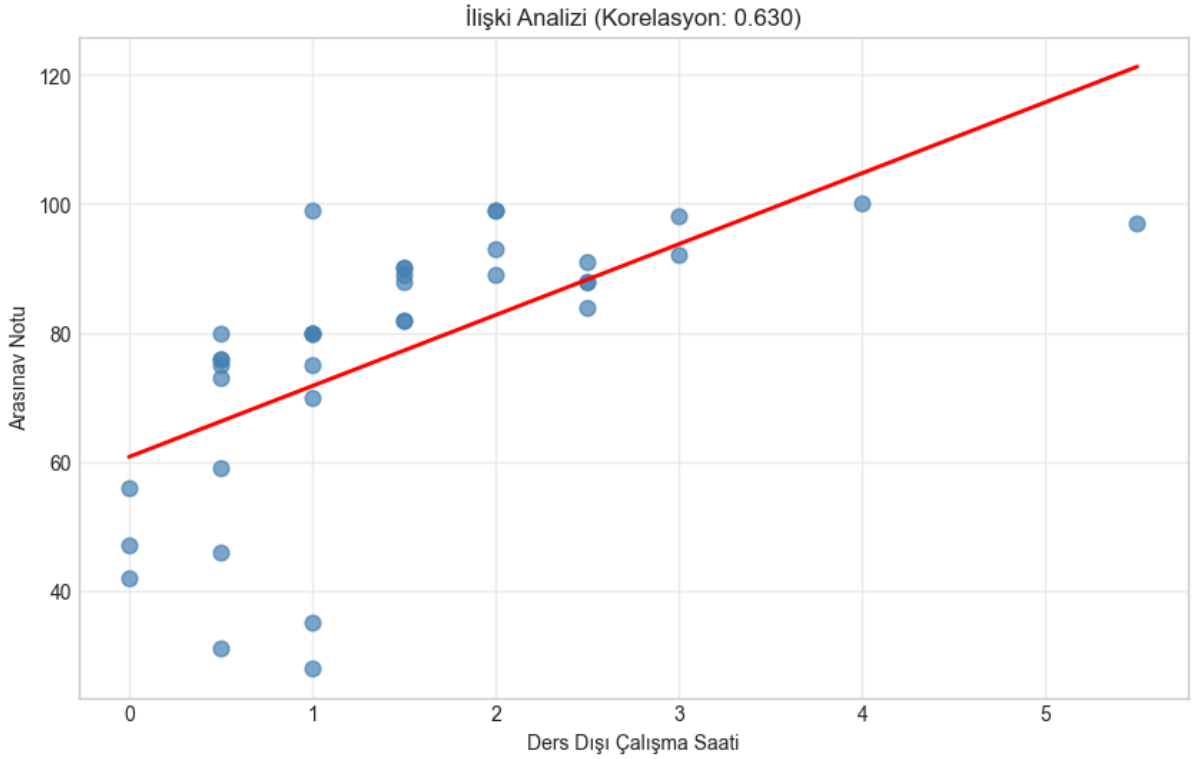
7. Araştırma Soruları ve Analiz Sonuçları

Yapılan analizler sonucunda ödev kapsamında sorulan kritik sorular şu şekilde cevaplanmıştır:

Soru 1: Ders dışı çalışma saatleri miktarı ile sınav puanı arasında bir ilişki var mı?

Cevap: Evet, ders dışı çalışma saatleri ile sınav puanı arasında **pozitif yönlü ve anlamlı bir ilişki** tespit edilmiştir. Analiz sonuçlarına göre, öğrenciler derse ayırdıkları ekstra zaman arttıkça sınav başarı puanları da yükselmektedir. Korelasyon analizleri ve regresyon eğrileri bu ilişkiyi doğrulamaktadır.

Aşağıdaki grafikte bu ilişki net bir şekilde görülmektedir:



İlişki Analizi Grafiği (Korelasyon: 0.63)

- **X ekseni:** *Ders dışı çalışma saati*
- **Y ekseni:** *Sınav notu*

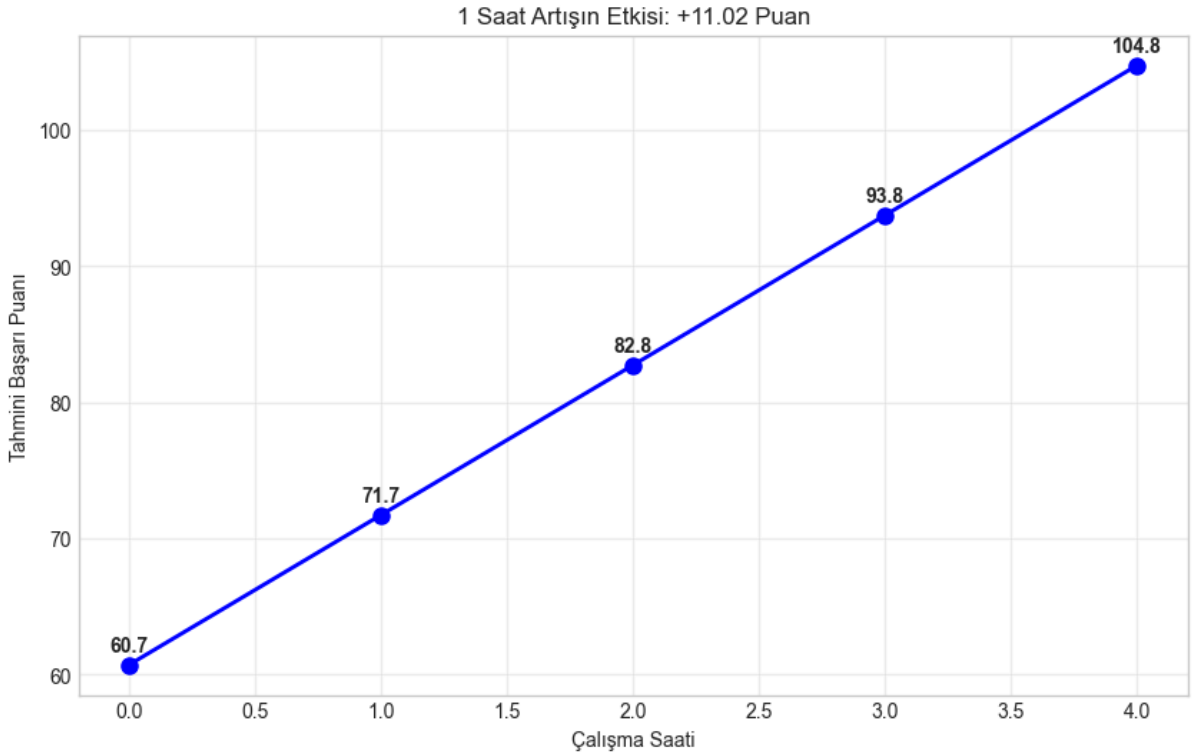
- Korelasyon 0.63, **orta-üst düzeyde pozitif bir ilişki** olduğunu gösteriyor.
  → Yani çalışma süresi arttıkça sınav notu genelde yükseliyor.

- Dağılım çok dağınık değil; kırmızı regresyon çizgisi yukarı doğru eğimli.
  → Bu, çalışma saatinin başarı üzerinde anlamlı bir etkisi olduğunu gösteriyor.

Soru 2: Haftada fazladan 1 saat ders dışı çalışma süresi başarı puanını ortalama ne kadar etkiliyor?

Cevap: Doğrusal regresyon modelinin katsayı analizi (coefficient analysis) sonucunda; diğer değişkenler sabit kalmak koşuluyla, haftalık ders dışı çalışmaya eklenen her **1 saatin**, başarı puanını ortalama **11.02 puan** artırdığı hesaplanmıştır.

Bu etkiyi gösteren analiz grafiği aşağıda sunulmuştur:



"1 Saat Artışın Etkisi" Grafiği

- Regresyon modeline göre **çalışma saatine 1 saat eklenmesi başarıyı yaklaşık +11 puan artırıyor**.

- Noktalar (0,1,2,3,4 saat) için tahmini başarı puanları çizilmiş:

    - 0 saat → ~60.7
    - 1 saat → ~71.7
    - 2 saat → ~82.8
    - 3 saat → ~93.8
    - 4 saat → ~104.8

- Bu da doğrusal modelin güçlü bir artış öngördüğünü gösteriyor.

8. Sonuç ve Pratik Çıkarımlar

Bu çalışmada, **BİL 2009 - Çizge Kuramı** dersi için oluşturulan veri seti üzerinde regresyon analizi yapılmış ve ders dışı çalışmanın akademik başarı üzerindeki etkisi araştırılmıştır

- **Model Üstünlüğü:** Elde edilen metrikler ve görsel analizler, **Decision Tree Regressor** modelinin, verideki doğrusal olmayan yapıyı en iyi yakalayarak diğer modellere göre daha başarılı olduğunu (R^2 = 0.547) kanıtlamıştır.

- **Temel Çıkarım:** Elde edilen bulgular, ders dışı çalışmanın akademik başarı üzerinde **doğrudan ve pozitif** bir etkisi olduğunu (saat başına ortalama 11 puan) kanıtlamıştır.

- **Pratik Öneri:** En başarılı Decision Tree modelinin tahminleri, öğrencilerin 1.5 saatlik çalışma eşiğinden sonra notlarında keskin bir artış yaşandığını gösterir. Bu bulgu, öğrencilere, başarının maksimize edilmesi için **haftalık 1.5 saatlik ders dışı çalışma süresinin** kritik bir eşik olarak hedeflenmesini önermektedir.