KARAR AĞAÇLARI:

Karar ağaçları, veri sınıflandırma ve regresyon problemlerini çözmek için kullanılan, dallanma yapısına sahip algoritmalardır. Her düğüm (node) bir özelliği (attribute) test eder, her dal (branch) bir test sonucunu temsil eder ve her yaprak düğüm (leaf node) bir sınıf etiketini veya sürekli bir değeri (regresyon için) temsil eder.

Yapısı

1. **Kök Düğüm (Root Node)**: Ağacın en üstündeki düğüm olup, ilk bölme burada yapılır.
2. **İç Düğüm (Internal Node)**: Özelliklere dayalı kararların verildiği düğümlerdir.
3. **Yaprak Düğüm (Leaf Node)**: Nihai kararın verildiği ve sınıf etiketinin ya da değerin bulunduğu düğümlerdir.
4. **Dallar (Branches)**: Karar ağaçlarında kök düğümden yaprak düğümlere giden yolları temsil eder.

Çalışma Prensibi

Karar ağacı algoritmaları, veri setindeki özellikler (attributes) arasındaki ilişkilere dayanarak bir ağaç yapısı oluşturur. Ağaç oluşturma süreci genellikle şu adımları içerir:

1. **Bölme Kriteri Seçimi**: Her düğümde veri, belirli bir kritere (örneğin Gini impüritesi veya bilgi kazancı) göre bölünür. Bu kriter, veri setindeki belirsizliği (entropy) en çok azaltan bölmeyi seçer.
2. **Bölme**: Seçilen kritere göre veri iki veya daha fazla alt gruba ayrılır.
3. **Tekrarlama**: Her alt grup için yukarıdaki adımlar tekrar edilir, ta ki durdurma kriterleri karşılanana kadar (örneğin, maksimum derinliğe ulaşıldığında veya daha fazla bölme yapılamadığında).

Avantajları

- **Kolay Anlaşılır ve Yorumlanabilir**: Ağaç yapısı görsel olarak kolayca anlaşılabilir.
- **Öznitelik Seçimi**: Karar ağaçları önemli özellikleri otomatik olarak seçer ve bu da özellik mühendisliğini kolaylaştırır.
- **Veri Ön İşleme Gerektirmez**: Genellikle ölçekleme veya normalizasyon gibi veri ön işleme adımlarına ihtiyaç duymaz.

Dezavantajları

- **Aşırı Uyum (Overfitting)**: Çok derin ağaçlar, eğitim verisine aşırı uyum sağlayarak genelleme kabiliyetini kaybedebilir.
- **Kararsızlık**: Küçük veri değişiklikleri büyük yapısal değişikliklere yol açabilir.
- **Verimsizlik**: Bazı durumlarda büyük veri setleri ile çalışmak zaman alıcı olabilir.

Kullanım Alanları

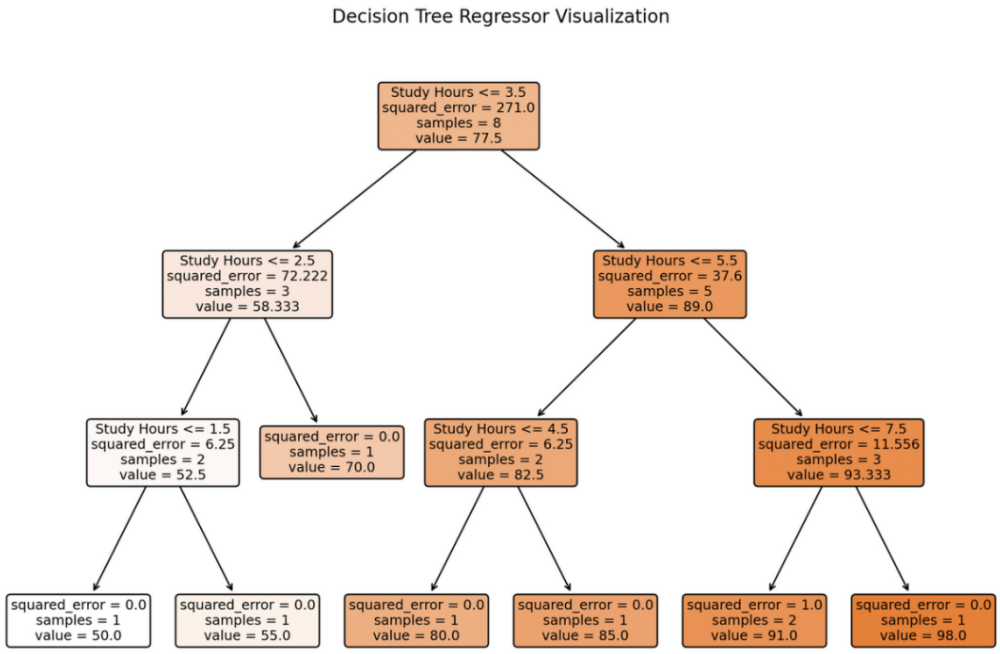
- **Sınıflandırma Problemleri**: Örneğin, spam tespiti, müşteri segmentasyonu.
- **Regresyon Problemleri**: Örneğin, ev fiyat tahmini, satış tahmini.

Sonuç olarak, karar ağaçları, veri sınıflandırma ve regresyon problemleri için güçlü ve anlaşılması kolay bir araçtır, ancak aşırı uyum gibi zorluklara dikkat edilmesi gerekir.

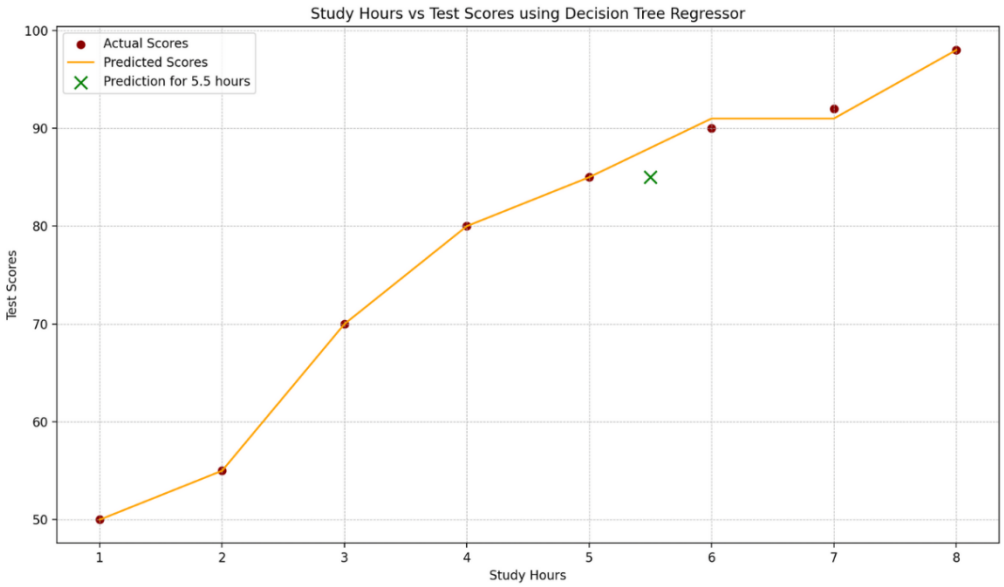
ÖRNEK: Çalışılan saat sayısı ile öğrencilerin aldığı puanlar arasındaki ilişki Alex'in ilgisini çekiyor. Alex, akranlarından çalışma saatleri ve ilgili sınav puanları hakkında veri topladı.

Şunu merak ediyor: Bir öğrencinin ders çalıştığı saat sayısına göre puanını tahmin edebilir miyiz? Bunu ortaya çıkarmak için Karar Ağacı Regresyonundan yararlanalım.

Teknik olarak, bağımsız bir değişkene (çalışma saatleri) dayalı olarak sürekli bir sonuç (test puanı) öngörüyoruz.



Görselleştirme, çalışma saatleri verileriyle eğitilmiş bir karar ağacı modelini göstermektedir. Her düğüm, test puanlarını en iyi tahmin eden koşullara dayalı olarak en üst kökten dallanarak çalışma saatlerine dayalı bir kararı temsil eder. İşlem, maksimum derinliğe ulaşana veya başka anlamlı bölünmeler kalmayana kadar devam eder. Alttaki yaprak düğümleri, regresyon ağaçları için o yaprağa ulaşan eğitim örneklerinin hedef değerlerinin ortalaması olan nihai tahminleri verir. Bu görselleştirme, modelin öngörücü yaklaşımını ve çalışma saatlerinin test puanları üzerindeki önemli etkisini vurgulamaktadır.



[Resim Kaynağı: Yazar] Karar ağacı regresörü kullanılarak çizilen çalışma saatleri ve test puanları

"Çalışma Saatleri ve Test Puanları" grafiği, çalışma saatleri ile ilgili test puanları arasındaki ilişkiyi gösterir. Gerçek veri noktaları kırmızı noktalarla gösterilirken, modelin tahminleri regresyon ağaçlarının özelliği olan turuncu adım işleviyle gösterilir. Yeşil bir "x" işareti, burada 5,5 saatlik bir çalışma süresini temsil eden yeni bir veri noktasına yönelik bir tahmini vurgulamaktadır. Kılavuz çizgileri, etiketler ve açıklamalar gibi olay örgüsünün tasarım öğeleri, gerçek ve beklenen değerlerin anlaşılmasını geliştirir.