**VERİ MADENCİLİĞİ PROJESİ**

**TUBA AYDIN**

Hisse Senedi Fiyat Tahmini: Sınıflandırma Algoritmaları Karşılaştırması

**1. Proje Konusu**

Bu projede, bir hisse senedinin (Tesla - TSLA) geçmiş fiyat hareketlerine bakılarak, bir sonraki gün fiyatın artacağı mı (1) yoksa azalacağı mı (0) sınıflandırması yapılmaya çalışılmıştır. Kullanılan veri zaman serisi türünde olup, bu tip verilerde hem trend hem de dönemsel desenler olabilir. Bu nedenle, veri sınıflandırması yapılmadan önce gerekli ön işlemler gerçekleştirilmiştir.

**2. Veri Seti ve Hazırlık Süreci**

* **Veri kaynağı**: Kaggle – [Stock Price Prediction Dataset](https://www.kaggle.com/datasets)
* **Kullanılan dosya:** TSLA.csv
* **Hedef değişken (bağımlı değişken):** Target → 1: Fiyat artışı, 0: Fiyat düşüşü
* **Bağımsız değişkenler:** Açılış, kapanış fiyatı, hacim, yüksek/düşük değerler gibi finansal göstergeler.

Veri seti sınıflandırma algoritmaları için uygun hale getirilmiş, eksik veriler kontrol edilmiş ve sınıf dengesizliği SMOTE yöntemi ile giderilmiştir.

**3. Temel Kavramlar: Confusion Matrix (Karışıklık Matrisi)**

Confusion Matrix, bir sınıflandırma modelinin başarı durumunu analiz etmek için kullanılan önemli bir araçtır.

|  |  |  |
| --- | --- | --- |
|  | Tahmin :0 | Tahmin :1 |
| Gerçek:0 | 20 | 19 |
| Gerçek:1 | 17 | 20 |

ekran görüntüsü, metin, dikdörtgen, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Açıklamaları:

* **True Positives (TP)** = 20 → Gerçekte fiyat artmış (1), model de artacak tahmin etmiş , Model gerçekten artış olan günü doğru tahmin etmiş – başarılı
* **True Negatives (TN)** = 20 → Gerçekte fiyat azalmış (0), model de azalacak tahmin etmiş, Model gerçekten düşüş olan günü doğru tahmin etmiş – başarılı
* **False Positives (FP)** = 19 → Gerçekte fiyat azalmış (0), ama model artacak tahmin etmiş, Model, düşüş olacak günü yanlışlıkla artış olacak sanmış – hata
* **False Negatives (FN)** = 17 → Gerçekte fiyat artmış (1), ama model azalacak tahmin etmiş, Model, artış olacak günü yanlışlıkla düşüş olacak sanmış – hata

Bu durumda modelin pozitif ve negatif sınıfları ayırt etmede orta seviye bir performans sergilediği görülmektedir. FP ve FN oranları biraz yüksektir, bu da modelin yanlış tahmin oranının yüksek olduğunu gösterir.

**4. Algoritmaların Sonuçları ve Yorumları**

1. **Logistic Regression**

* Accuracy (Doğruluk): 0.46
* Precision (Kesinlik): 0.45
* Recall (Duyarlılık): 0.54
* F1 Score: 0.49
* ROC AUC: 0.51

***Yorum:*** *Basit, doğrusal bir sınıflandırma yöntemidir. Fakat veri doğrusal olarak ayrılabilir olmadığı için sınıfları iyi ayıramamıştır. Bu nedenle başarı oranı düşüktür.*

1. **Decision Tree**

* Accuracy: 0.59
* Precision: 0.58
* Recall: 0.59
* F1 Score: 0.59
* ROC AUC: 0.59

***Yorum:*** *Verinin yapısını öğrenebilen, sezgisel bir modeldir. En iyi sonucu bu model vermiştir çünkü verideki örüntüleri iyi öğrenmiştir. Ağaç yapısı, karar verme sürecini net bir şekilde izah edebilir.*

1. **Random Forest**

* Accuracy: 0.53
* Precision: 0.51
* Recall: 0.54
* F1 Score: 0.53
* ROC AUC: 0.50

***Yorum:*** *Çok sayıda karar ağacının bir araya gelmesiyle oluşur. Genellikle güçlü bir modeldir ancak bu veride overfitting ya da feature seçimi problemleri sebebiyle beklenen performansı gösterememiştir.*

1. **Support Vector Machine (SVM)**

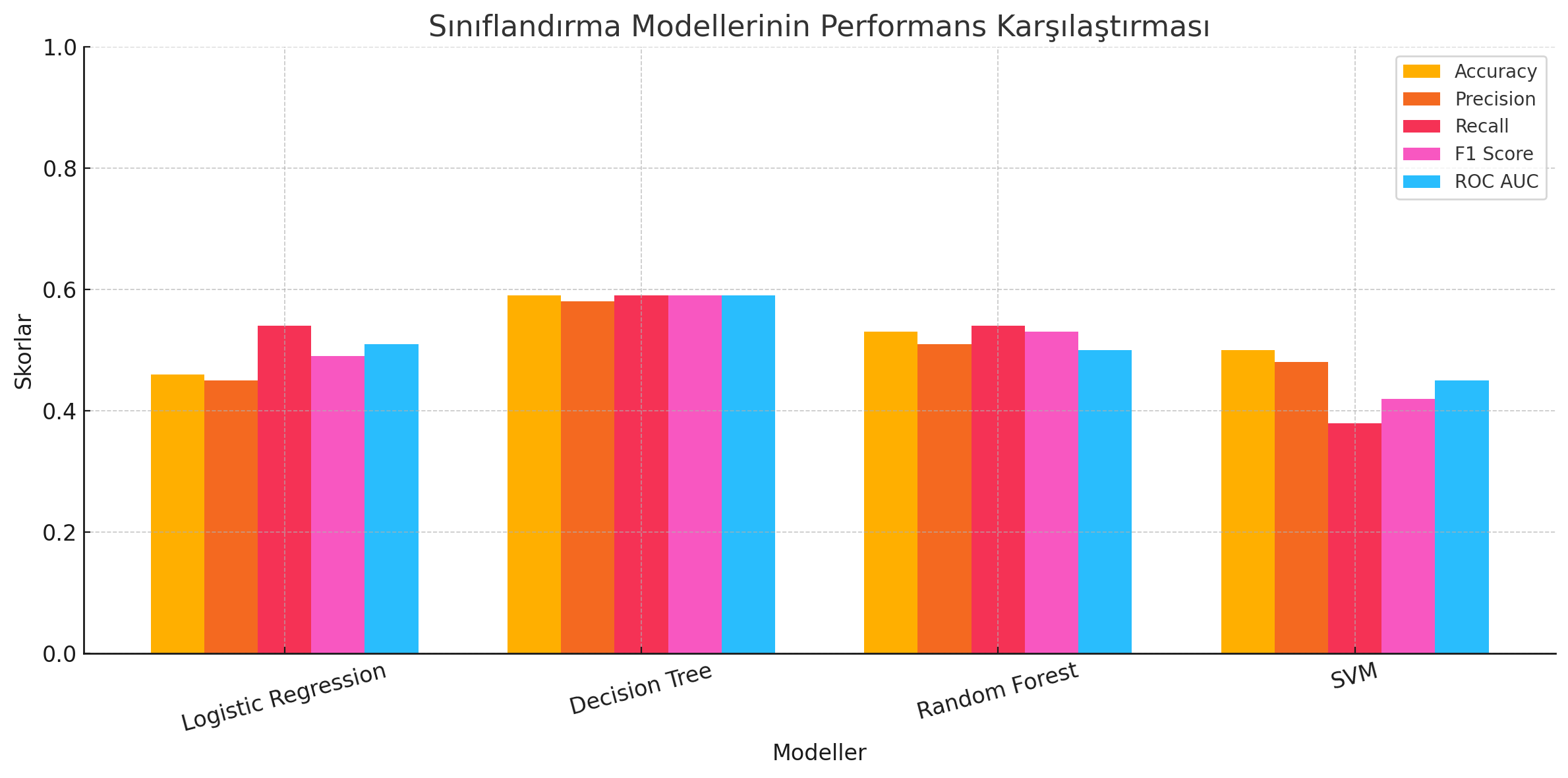
* Accuracy: 0.50
* Precision: 0.48
* Recall: 0.38
* F1 Score: 0.42
* ROC AUC: 0.45

***Yorum:*** *SVM, genellikle doğrusal olmayan verilerde kernel fonksiyonları ile başarılı olur. Fakat bu veri setinde başarı oranı düşüktür. Özellikle recall değerinin düşük olması, pozitif sınıfların yeterince iyi tanınmadığını gösteriyor.*

# Sınıflandırma Modelleri Performans Tablosu

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy  (Doğruluk) | Precision  (kesinlik) | Recall | F1 Score | ROC AUC | Yorumlama |
| Logistic Regression | 0.46 | 0.45 | 0.54 | 0.49 | 0.51 | * **Doğruluk (Accuracy)**: %46 ile en düşük sonucu verdi. * **Precision ≈ Recall** olduğundan dengeli ama düşük performans var. * **Yorum**: Doğrusal karar sınırları oluşturur. Eğer veri doğrusal olarak ayrılabilir değilse (ki bu veri öyle görünüyor), başarı oranı düşer. |
| Decision Tree | 0.59 | 0.58 | 0.59 | 0.59 | 0.59 | * **Tüm metriklerde en iyi sonucu verdi.** * Özellikle **F1 Score = 0.59** ve **ROC AUC = 0.59** ile güçlü. * **Yorum**: Verideki karmaşık örüntüleri yakalayabiliyor. Ağaç yapısı sayesinde karar süreçleri yorumlanabilir ve görselleştirilebilir. Bu model en başarılısı olmuş. |
| Random Forest | 0.53 | 0.51 | 0.54 | 0.53 | 0.50 | * Normalde overfitting'e karşı güçlüdür, ancak burada tek ağaçtan (Decision Tree) **daha kötü** sonuç vermiş. **Muhtemel nedenler**: Parametre ayarları yeterince optimize edilmemiş olabilir (örneğin ağaç sayısı, derinlik). |
| SVM | 0.50 | 0.48 | 0.38 | 0.42 | 0.45 | * **Recall değeri = 0.38**, yani model pozitif sınıfları (fiyat artışı) yakalayamıyor. * **Yorum**: Kernel tipi, C parametresi gibi ayarlar önemli. Bu veride uygun kernel (örneğin RBF) ve ayarlar kullanılmadıysa performans düşebilir. Karmaşık yapılı verilerde SVM güçlü olabilir ama iyi ayarlanmalıdır. |

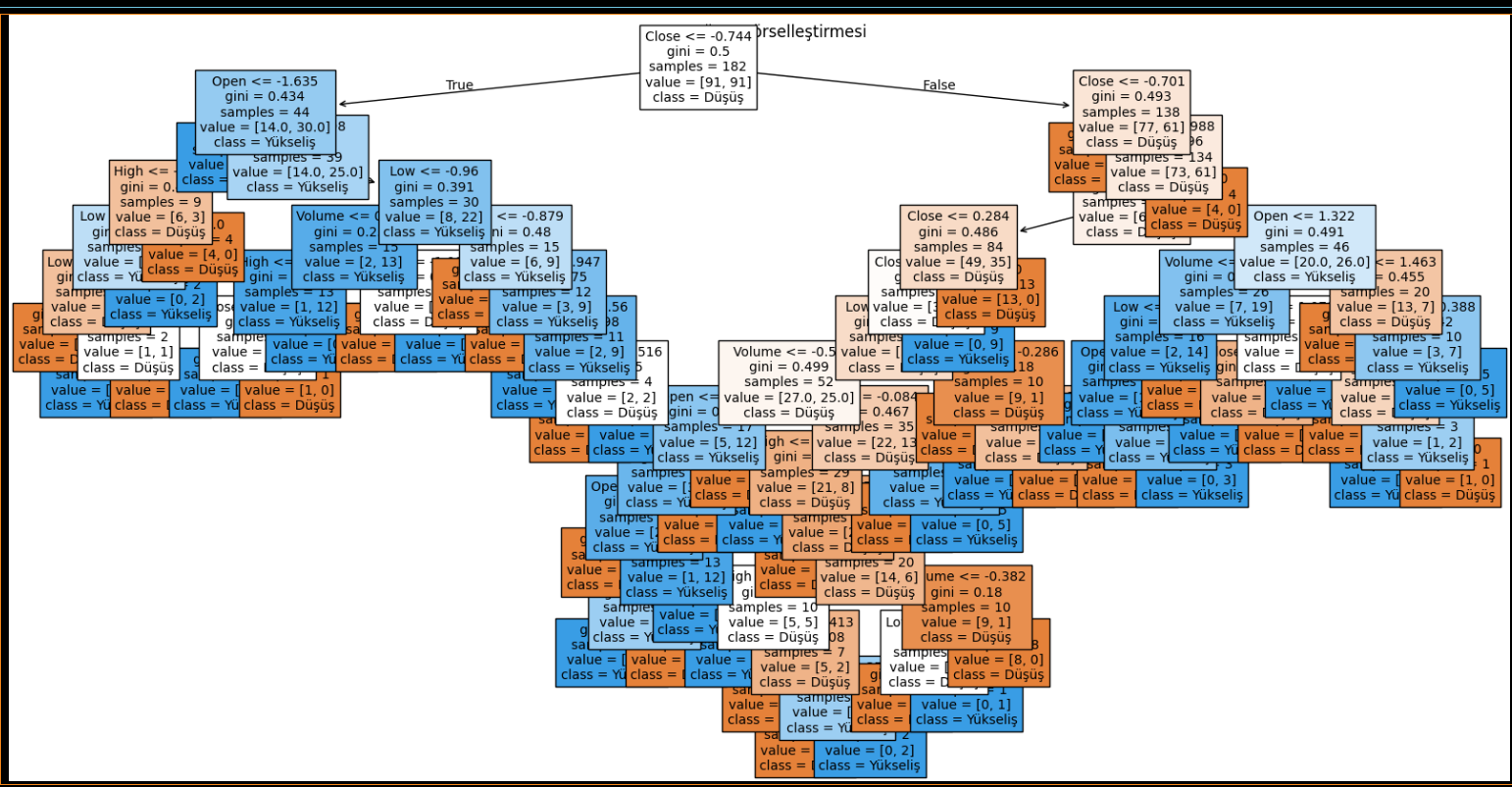
**5. Genel Karşılaştırma ve Değerlendirme**

* **En başarılı model:** Decision Tree
* **En zayıf model:** SVM, Logistic Regression
* **Öneri:** Daha fazla özellik mühendisliği (feature engineering) yapılabilir. Örneğin RSI, MACD, EMA gibi teknik indikatörler dahil edilerek modeller güçlendirilebilir.

Yukarıdaki görselleştirme, dört farklı sınıflandırma modelinin performans metriklerini karşılaştırmalı olarak sunmaktadır. Her bir metrik için model bazında ne kadar başarılı oldukları açıkça görülebilir. En dengeli ve yüksek performansı **Decision Tree** modeli göstermiştir.

Bu görsel, bir **karar ağacı (decision tree)** sınıflandırma modelinin grafiksel temsilidir. Ağaç, finansal verileri (muhtemelen hisse senedi verileri gibi: Open, Close, High, Low, Volume) kullanarak **"Yükseliş"** veya **"Düşüş"** şeklinde sınıflandırma yapmaktadır.

**Karar Ağacının Genel Yapı**

****

Bu karar ağacı, hisse senedi verilerine göre (özellikle Close, Open, Volume, Low, High) bir günün sonunda fiyatın **"Yükseliş"** mi yoksa **"Düşüş"** mü olacağını sınıflandırmak için oluşturulmuştur. Kök düğümde Close <= -0.744 koşulu ile veri ikiye ayrılır. **Sol tarafta genellikle Yükseliş, sağ tarafta ise Düşüş tahmini yapılmaktadır. Renkler, sınıf tahminini ve bu tahminin gücünü (koyu renk = yüksek güven) gösterir.** Ağaçta gini değerlerinin düşük olması, modelin veriyi iyi ayırabildiğini gösterirken, çok sayıda dal olması **aşırı öğrenme (overfitting)** riskine işaret edebilir. Model, finansal göstergelerle sınıflandırma yaparak kararlarını verir.

**6. Sonuç**

Bu çalışmada hisse senedi yön tahmini için sınıflandırma algoritmaları karşılaştırılmıştır. Verinin küçük boyutlu olması, zaman serisi özelliği taşıması ve sınıf dengesizliği modellerin başarısını etkilemiştir. En başarılı model karar ağacı olmuştur. Bu proje, finansal verilerde sınıflandırma uygulamaları açısından temel bir örnek niteliği taşımaktadır.