NBA Maç Sonucu Tahmini - Proje Raporu

1. Giriş

1.1. Veri Seti

Bu projede, Kaggle platformundan alınan **Historical NBA Data** veri seti kullanılmıştır. Veri seti, NBA maçlarının detaylı istatistiklerini içermektedir:

- **Games.csv**: 72,470 maçın genel bilgileri (17 kolon)
    - Maç tarihi, ev sahibi takım, deplasman takımı, skor bilgileri vb.
- **TeamStatistics.csv**: 144,940 takım istatistiği kaydı (48 kolon)
    - Takımların maç başına istatistikleri (sayı, asist, rebound, şut yüzdesi vb.)

1.2. Problem Tanımı

Bu projede, bir **Sınıflandırma (Classification)** problemi çözülmüştür.

Hedef Değişken: Ev sahibi takımın maçı kazanıp kazanmayacağı (`HomeWin`)

- `HomeWin = 1`: Ev sahibi takım kazandı
- `HomeWin = 0`: Ev sahibi takım kaybetti

Bu problem, makine öğrenmesi modelleri kullanılarak, maç öncesi mevcut olan bilgilerle (takım performansları, geçmiş maç istatistikleri vb.) tahmin edilmeye çalışılmıştır.

2. Veri Analizi

2.1. Veri İnceleme

Veri seti incelendiğinde:

- Toplam 72,470 maç verisi mevcut
- Takım istatistikleri verileri her maç için hem ev sahibi hem de deplasman takımı için ayrı ayrı kaydedilmiştir
- Veri setinde eksik veri bulunmamaktadır (veya gerekli temizleme işlemleri yapılmıştır)

2.2. Feature Engineering

Tahmin için kullanılacak özellikler, maç öncesi mevcut olan bilgilerden oluşturulmuştur:

- **Rolling Averages (Yuvarlanan Ortalamalar)**: Her takım için son 10 maçın ortalaması hesaplanmıştır
    - Şut yüzdesi (fieldGoalsPercentage)
    - Üç sayı yüzdesi (threePointersPercentage)
    - Serbest atış yüzdesi (freeThrowsPercentage)
    - Toplam rebound (reboundsTotal)
    - Asist (assists)
    - Top çalma (steals)
    - Blok (blocks)
    - Top kaybı (turnovers)
    - Boya içi sayılar (pointsInThePaint)

Bu özellikler, takımların güncel form durumlarını yansıtmaktadır ve maç sonucunu tahmin etmede önemli bilgiler sağlamaktadır.

2.3. Görselleştirmeler

Proje kapsamında aşağıdaki görselleştirmeler yapılmıştır:

1. **Eksik Veri Analizi**: Veri setinde eksik veri olup olmadığı kontrol edilmiştir
2. **Hedef Değişken Dağılımı**: Ev sahibi takımın kazanma/kaybetme dağılımı görselleştirilmiştir
3. **Sayısal Özellik Dağılımları**: Önemli özelliklerin histogram grafikleri oluşturulmuştur
4. **Aykırı Değer Analizi**: Boxplot grafikleri ile aykırı değerler tespit edilmiştir
5. **Korelasyon Analizi**: Özellikler arasındaki korelasyonlar heatmap ile görselleştirilmiştir

3. Yöntem

3.1. Veri Ön İşleme

1. **Veri Birleştirme**: Games ve TeamStatistics verileri birleştirilmiştir
2. **Özellik Seçimi**:
    - Tahmin için gerekli özellikler seçilmiştir
    - Sadece sayısal (numeric) kolonlar modelleme için kullanılmıştır
    - String ve tarih kolonları (gameDateTimeEst, takım isimleri, şehirler vb.) ve ID kolonları (gameId, teamId vb.) hariç tutulmuştur
    - Hedef değişken (homeWin) ve gereksiz ID kolonları da özellik setinden çıkarılmıştır
3. **Normalizasyon**: StandardScaler kullanılarak özellikler normalize edilmiştir (Logistic Regression, KNN, SVM için)
4. **Train-Test Ayrımı**: Veri seti %80 eğitim, %20 test olarak ayrılmıştır

3.2. Kullanılan Modeller

Projede aşağıdaki dört farklı makine öğrenmesi algoritması kullanılmıştır:

1. **Logistic Regression**
    - Doğrusal sınıflandırma modeli
    - İkili sınıflandırma problemleri için uygun
    - Yorumlanabilirliği yüksek
2. **K-Nearest Neighbors (KNN)**
    - K=15 parametresi ile kullanılmıştır
    - Örnek tabanlı (instance-based) öğrenme yöntemi
    - Non-parametrik model
3. **Support Vector Machine (SVM)**
    - Linear kernel ile kullanılmıştır
    - Yüksek boyutlu verilerde etkili
    - Sınır optimizasyonu yapar
4. **Decision Tree**
    - Max depth=5 parametresi ile sınırlandırılmıştır
    - Ağaç yapısı ile karar kuralları oluşturur
    - Yorumlanabilir model

3.3. Model Seçimi Gerekçesi

Farklı algoritma türlerinden modeller seçilerek:

- Doğrusal ve doğrusal olmayan modellerin karşılaştırılması sağlanmıştır
- Basit ve karmaşık modellerin performansları incelenmiştir
- En uygun modelin belirlenmesi amaçlanmıştır

4. Sonuçlar

4.1. Model Performansları

Modellerin test seti üzerindeki başarı oranları (Accuracy) aşağıdaki gibidir:

Model	Accuracy	Açıklama
Logistic Regression	**62.9%**	En yüksek başarı oranı
SVM	61.4%	İkinci en iyi performans
KNN	56.3%	Orta düzey performans
Decision Tree	49.9%	En düşük performans (rastgele tahmin seviyesine yakın)

4.2. Confusion Matrix Analizi

Tüm modeller için Confusion Matrix görselleştirmeleri yapılmış ve analiz edilmiştir. Confusion Matrix'ler, modellerin:

- True Positive (TP): Doğru tahmin edilen kazanma sayısı
- True Negative (TN): Doğru tahmin edilen kaybetme sayısı
- False Positive (FP): Yanlış tahmin edilen kazanma sayısı
- False Negative (FN): Yanlış tahmin edilen kaybetme sayısı

değerlerini göstermektedir.

4.3. Bulgular ve Yorumlar

1. **En İyi Model**: Logistic Regression modeli %62.9 accuracy ile en yüksek başarı oranına ulaşmıştır. Bu, ev sahibi takım avantajı (home court advantage) gibi doğrusal ilişkilerin bu problemde önemli olduğunu göstermektedir.

2. **SVM Performansı**: SVM modeli %61.4 accuracy ile ikinci sırada yer almıştır. Linear kernel kullanılmasına rağmen, Logistic Regression'dan biraz daha düşük performans göstermiştir.

3. **KNN Performansı**: KNN modeli %56.3 accuracy ile orta düzey bir performans sergilemiştir. Komşu örnekler arasındaki benzerliklerin bu problemde yeterince güçlü olmadığı söylenebilir.

4. **Decision Tree Performansı**: Decision Tree modeli %49.9 accuracy ile en düşük performansı göstermiştir. Bu sonuç, ağaç yapısının bu problemde yeterince bilgi çıkaramadığını veya overfitting/underfitting sorunları yaşadığını düşündürmektedir.

5. **Ev Sahibi Avantajı**: Tüm modeller %50'nin üzerinde bir başarı oranı elde etmiştir (Decision Tree hariç, bu da rastgele tahmin seviyesinde). Bu, ev sahibi takım avantajının gerçekten var olduğunu ve modellerin bu ilişkiyi yakalayabildiğini göstermektedir.

4.4. Sınırlamalar ve İyileştirme Önerileri

1. **Daha Fazla Özellik**: Oyuncu yaralanmaları, takım dinlenme süreleri, sezon içi performans trendleri gibi ek özellikler eklenebilir.

2. **Model Tuning**: Hyperparameter optimizasyonu (Grid Search, Random Search) ile modellerin performansları artırılabilir.

3. **Ensemble Yöntemleri**: Random Forest, XGBoost, Gradient Boosting gibi ensemble yöntemleri deneyebilir.

4. **Daha Derin Analiz**: Feature importance analizi ile hangi özelliklerin daha önemli olduğu belirlenebilir.

5. **Zaman Serisi Analizi**: Maçların kronolojik sırası daha detaylı analiz edilebilir.

5. Sonuç

Bu projede, NBA maç sonuçlarını tahmin etmek için dört farklı makine öğrenmesi modeli eğitilmiş ve karşılaştırılmıştır. Logistic Regression modeli %62.9 accuracy ile en iyi performansı göstermiştir.

Proje, ham veriden anlamlı özellikler çıkarma, farklı makine öğrenmesi algoritmaları uygulama ve sonuçları değerlendirme süreçlerini içermektedir. Elde edilen sonuçlar, ev sahibi takım avantajının gerçekten var olduğunu ve makine öğrenmesi modelleri ile tahmin edilebilir olduğunu göstermektedir.

https://github.com/CenkAk/VeriAnalizi_Proje
