Futbol Maç Sonucu Tahmini Yapma

Mehmet Furkan Koç
Elektrik-Elektronik Mühendisliği
TOBB ETÜ
Ankara, Türkiye
mehmetfurkan.koc@etu.edu.tr

***Özet* - Dünyanın yaklaşık 200'den fazla ülkesinde oynanan ve 250 milyondan fazla lisanslı oyuncusu bulunan futbolun, yıllık 500 milyar dolara ulaşan bahis sektörü bulunmaktadır. Bu da insanlara maçı kazanacak tarafı bilme isteğinin yüksek boyutlarda olduğuna yönelik bir durum olarak ortaya çıkmaktadır. Kullanılan veri seti, yıllar boyunca oynanan maçlardan toplanmış, yaklaşık 40 özelliği içeren bir veri setidir. Böylelikle yapay zeka tarafından maçların sonucunu tahmin etme üzerine bir araştırma yapılmıştır.**

I. GIRIŞ

Futbol, dünya genelinde en çok takip edilen spor dallarından biridir ve maç sonuçları üzerine yapılan tahminler, futbolseverler ve analistler için büyük bir ilgi alanı oluşturmaktadır. Özellikle son yıllarda, yapay öğrenme ve veri analitiği alanındaki gelişmeler, futbol maç sonuçlarının tahmin edilmesinde yeni yöntemler ve yaklaşımlar sunmuştur. Bu proje, Premier Lig'e ait 18 sezonluk futbol verilerini kullanarak, maç sonuçlarının tahmin edilmesini amaçlamaktadır.

Bu projede, çeşitli yapay öğrenme teknikleri kullanılarak takımların geçmiş performansları, gol istatistikleri, form durumları ve diğer önemli faktörler analiz edilmiştir. Veri setinden elde edilen bilgiler ışığında, belirli bir maçın sonucunun tahmin edilmesi hedeflenmiştir. Tahmin sürecinde, takımların ev sahibi ya da deplasman performansları, geçmiş maçlardaki form durumları ve diğer dinamikler dikkate alınarak, tahmin modelleri oluşturulmuştur.

Dokümanın ilerleyen bölümlerinde, veri setinin nasıl hazırlandığı, hangi özelliklerin seçildiği, kullanılan yapay öğrenme algoritmaları ve elde edilen sonuçlar detaylı bir şekilde ele alınacaktır. Bu çalışma, hem futbol maç sonucu tahminine ilgi duyanlar hem de yapay öğrenme yöntemlerini pratik bir problem üzerinde uygulamak isteyenler için kapsamlı bir rehber niteliğindedir.

II. VERI ÖZELLIKLERI

- **Div** = Lig Bölümü
- **Date** = Maç Tarihi (gg/aa/yy)
- **Time** = Maç Başlama Saati
- **HomeTeam** = Ev Sahibi Takım
- **AwayTeam** = Deplasman Takımı
- **FTHG** ve **HG** = Maç Sonu Ev Sahibi Takım Golleri
- **FTAG** ve **AG** = Maç Sonu Deplasman Takımı Golleri
- **FTR** ve **Res** = Maç Sonu Sonuç (H=Ev Sahibi Kazandı, D=Beraberlik, A=Deplasman Kazandı)
- **HTHG** = İlk Yarı Ev Sahibi Takım Golleri
- **HTAG** = İlk Yarı Deplasman Takımı Golleri
- **HTR** = İlk Yarı Sonuç (H=Ev Sahibi Kazandı, D=Beraberlik, A=Deplasman Kazandı)
- **Attendance** = Seyirci Katılımı

- **Referee** = Maç Hakemi
- **HS** = Ev Sahibi Takım Şutları
- **AS** = Deplasman Takımı Şutları
- **HST** = Ev Sahibi Takım İsabetli Şutları
- **AST** = Deplasman Takımı İsabetli Şutları
- **HHW** = Ev Sahibi Takım Direğe Çarpan Şutlar
- **AHW** = Deplasman Takımı Direğe Çarpan Şutlar
- **HC** = Ev Sahibi Takım Kornerleri
- **AC** = Deplasman Takımı Kornerleri
- **HF** = Ev Sahibi Takım Faulleri
- **AF** = Deplasman Takımı Faulleri
- **HFKC** = Ev Sahibi Takım Verilen Serbest Vuruşlar
- **AFKC** = Deplasman Takımı Verilen Serbest Vuruşlar
- **HO** = Ev Sahibi Takım Ofsaytları
- **AO** = Deplasman Takımı Ofsaytları
- **HY** = Ev Sahibi Takım Sarı Kartları
- **AY** = Deplasman Takımı Sarı Kartları
- **HR** = Ev Sahibi Takım Kırmızı Kartları
- **AR** = Deplasman Takımı Kırmızı Kartları
- **HBP** = Ev Sahibi Takım Kart Puanları (10 = sarı, 25 = kırmızı)
- **ABP** = Deplasman Takımı Kart Puanları (10 = sarı, 25 = kırmızı)

III. GEREKLI KÜTÜHANELER VE VERI SETININ KAYNAĞI

Gerekli kütüphaneler şunlardır:

- Pandas
- Numpy
- Matplotlib.pyplot
- Warnings
- Make_subplots
- Seaborn
- Sklearn.model

- Sklearn.ensemble
- Tensorflow

Veri seti kaggle platformundan ‘final_dataset.csv’ ismi ile indirilmiştir.

IV. Veri Seti Özellikleri ve Veri Ön İşleme

Veri seti 20 yıl boyunca İngiltere Premier Liginden alınan verilerden oluşmaktadır. 6839 satır ve 40 sütundan oluşmaktadır. Önemli bazı öznitelikler dokümanın devamında görselleştirilmiştir.

A. Maç Sonu Sonuç Dağılımı

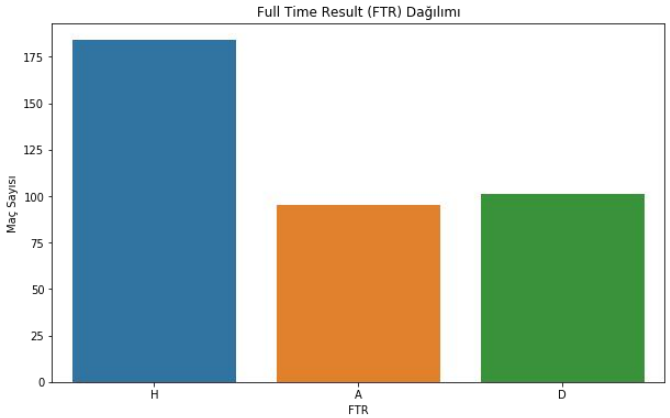

Figure 1 - Maç Sonucu

H=Ev Sahibi Kazandı, D=Beraberlik, A=Deplasman Kazandı, manasında kullanılmış olup maçların çoğunun ev sahibi tarafından kazanıldığı görülmektedir.

B. Takımlara Göre Gollerin Dağılımı

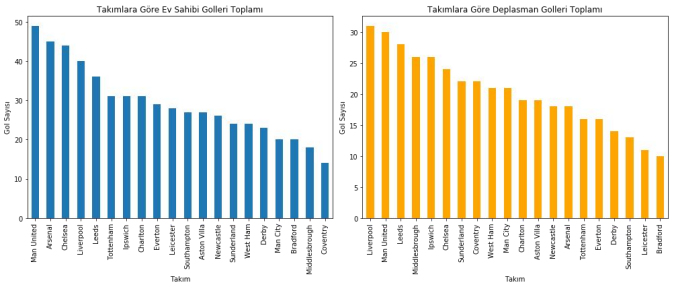

Figure 2 - Gol Dağılımı

Ev sahibi ve deplasman golleri toplamı en yüksek ve en düşük takımlar listelenmiştir.

C. Yıllara Göre Maç Sonuçları Dağılımı

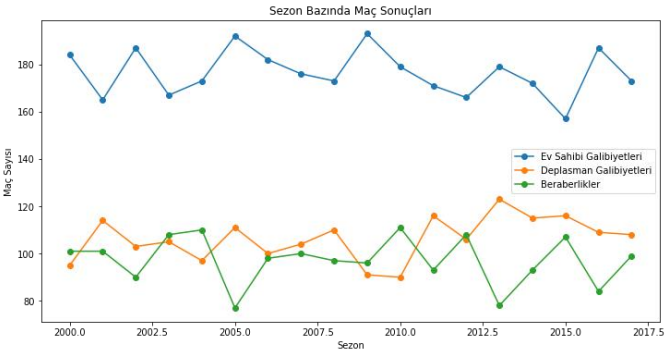

Figure 3 - Maç Sonuçları Dağılımı

Sezonlardaki maçların kazanma, kaybetme ve beraberlik sayıları listelenmiştir.

D. Pairplot ve Korelasyon Matrisleri

Bu veriler haricinde çok daha fazla öznitelik bulunmakla beraber bunların bir kısmı modelin öğrenmesini daha karmaşıklaştırmaktadır. Özniteliklerin ve verilerin görselleşmiş halini Figure-4 ve Figure-5’tedir.

Çift değişkenli ilişkileri göstermek için pairplot kullanılmıştır:

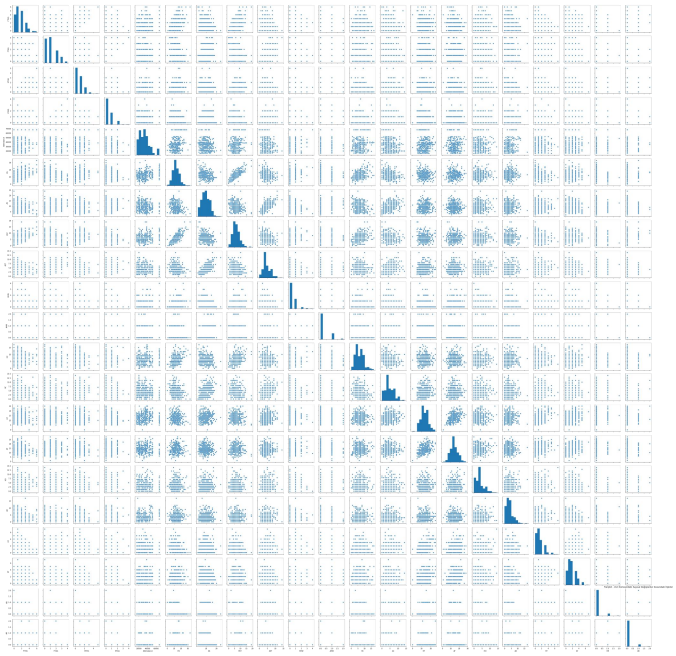

Figure 4 - Pairplot

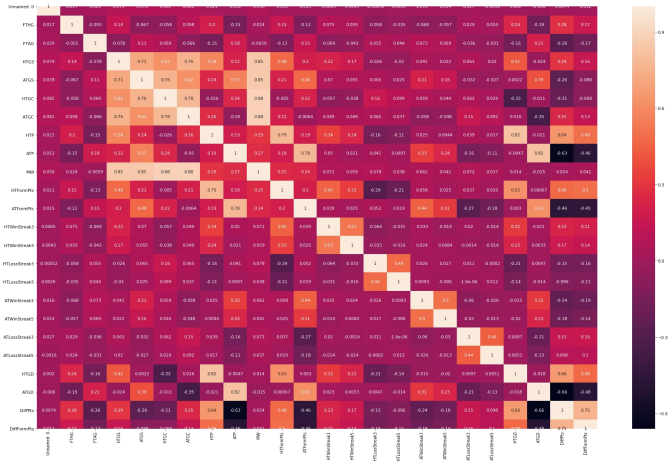

Figure 5 - Korelasyon Matrisi

Figure-4 ve Figure-5'te de görüleceği üzere fazla sayıda öznitelik vardır ve bazı öznitelikler modelin öğrenmesini zorlaştıracak cinstendir. Buna göre bazı öznitelikleri çıkartmak modelin öğrenmesi için çok daha iyi olacaktır. Bu bilgiler ışığında gerekli öznitelikleri alarak oluşturulan pairplot ve korelasyon matrisi Figure-6 ve Figure-7'deki gibidir.

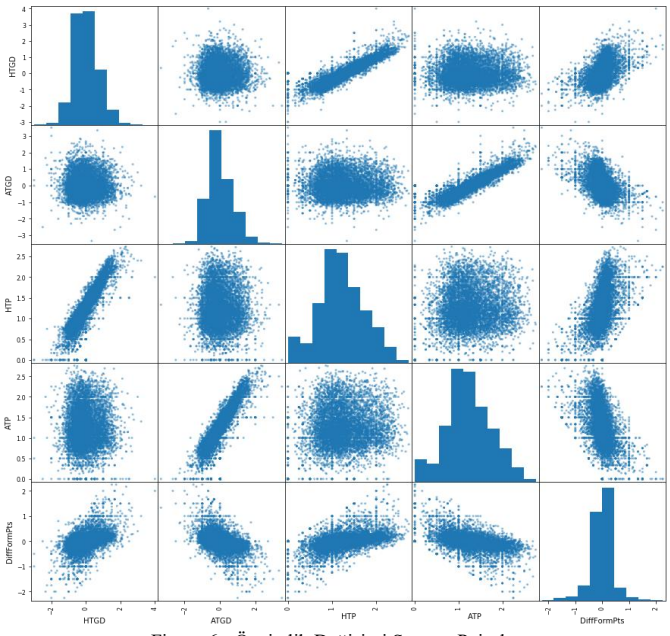

Figure 6 - Öznitelik Değişimi Sonrası Pairplot

Figure 7 - Öznitelik Değişimi Sonrası Korelasyon Matrisi

Bu noktada modeli eğitmek için gerekli öznitelik ve veriler elde edilmiştir. (Figure-6, Figure-7)

E. Veri Tipi, Eksik ve Yinelenen Veriler

Yapılan Explatory Data Analysis'e göre herhangi bir yinelene veriye rastlanmamıştır. Aynı zamanda eksik bir veri bulunamamıştır. Veri tipleri de int64 ve object olarak nitelendirilmektedir.

V. Model Eğitimi

İlk olarak veri setinin %25'i random olarak test verisini oluşturmuştur. Öncelikle kullanılan yöntemleri ve model eğitiminde neden seçildiğine bakalım.

A. Lineer Regresyon (Linear Regression)

Lineer regresyon, bağımlı değişken ile bağımsız değişkenler arasındaki doğrusal ilişkiyi modellemeyi amaçlayan bir regresyon tekniğidir. Bu yöntem, bağımsız değişkenlerin bir kombinasyonunu kullanarak bağımlı değişkenin değerini tahmin etmeye yönelik olarak kullanılır.

Neden kullanılabilir: Futbol maç tahmini problemlerinde, eğer sürekli bir çıktının tahmin edilmesi gerekiyorsa (örneğin, bir maçta atılacak gol sayısı gibi), lineer regresyon tercih edilebilir.

B. Random Forest

Random Forest, birden fazla karar ağacının ansambl olarak çalıştığı bir yöntemdir. Her bir ağaç, veri setinin rastgele bir alt kümesi üzerinde eğitim alır ve sonuçlar, ağaçlardan elde edilen tahminlerin ortalaması veya çoğunluk oyu ile belirlenir.

Neden kullanılabilir: Random Forest, değişkenler arasındaki karmaşık ilişkileri öğrenme yeteneği sayesinde futbol maçları gibi sınıflama problemlerinde yüksek doğruluk sağlayabilir. Ayrıca, aşırı uyum (overfitting) riskini azaltma kapasitesine sahip olduğundan, eksik verilerle başa çıkmada da etkili olabilir.

C. Destek Vektör Makineleri (Support Vector Machines - SVM)

SVM, verileri yüksek boyutlu bir uzaya dönüştürerek sınıflandırma ve regresyon problemlerini çözmeye yönelik bir yöntemdir. Bu model, sınıflar arasındaki en iyi ayrımı yapan hiperdüzlemi bulmayı hedefler.

Neden kullanılabilir: SVM, özellikle küçük ve orta ölçekli veri setlerinde yüksek doğruluk sağlamaktadır. Futbol maçları gibi karmaşık sınıflama problemlerinde, çeşitli özellikler

arasındaki ilişkileri öğrenmek için etkili bir seçenek olarak düşünülebilir.

D. Karar Ağaçları (Decision Trees)

Karar ağaçları, veriyi sınıflandırmak için bir dizi karar kuralı kullanan bir yöntemdir. Veri, her düğümde bir özelliğe göre bölünür ve yaprak düğümlerinde tahminler yapılır.

Neden kullanılabilir: Karar ağaçları, veri üzerinde kolayca yorumlanabilir kararlar sunduğundan, açıklayıcı analizlerde ve veriyi anlamaya çalışırken faydalı olabilir. Futbol maçları gibi verilerin çeşitli özelliklere göre sınıflandırılmasında etkili olabilir. Ancak, aşırı uyum riski taşır (overfit).

E. K-En Yakın Komşu (K-Nearest Neighbors - KNN)

KNN, sınıflandırma ve regresyon problemlerinde kullanılan basit bir algoritmadır. Belirli bir veri noktasının sınıfı, en yakın komşularının sınıflarına dayanarak belirlenir.

Neden kullanılabilir: KNN, özellikle sınıflar arasında belirgin sınırların olmadığı durumlarda etkili olabilir. Futbol maçları gibi verilerde, benzer maçların sonuçlarına bakarak tahmin yapma yeteneği sunar. Ancak, büyük veri setlerinde hesaplama maliyeti yüksek olabilir.

F. Gradient Boosting Classifier

Gradient Boosting, zayıf öğrenicileri (genellikle karar ağaçları) ardışık olarak eğiten bir ansambl yöntemidir. Her zayıf öğrenici, önceki öğrenicilerin hatalarını düzeltmeye çalışır.

Neden kullanılabilir: Gradient Boosting, yüksek doğruluk ve güçlü performans sağlayabilme kapasitesine sahiptir. Futbol maçları gibi karmaşık verilerle çalışırken, modelin hata payını azaltma yeteneği sunar. Model, verilerdeki karmaşık ilişkileri öğrenmede etkili olabilir.

Bunlara ek olarak bu yöntemleri birbirleri ile de kullanabiliriz. Her bir modelin avantajları ve sınırlamaları göz önünde bulundurulduğunda, futbol maç tahmini gibi bir problem için **Random Forest** ve **Gradient Boosting** gibi yöntemler yüksek performans sağlayabilir. **SVM** ve **KNN** gibi yöntemler de belirli veri setleri ve özellikler için uygun olabilir. **Karar Ağaçları** ve **Lineer Regresyon** ise veri üzerinde hızlı analiz yapma ve yorumlama olanağı sunar.

G. Modellerin Değerlendirilmesi

Model	Başarı Oranı (%)	Egitim Süresi (s)
Gradient Boosting Classifier	%100	0.5
Support Vector Machines	%52.81	5.7
Logistic Regression	%64.62	0.4
Random Forest	%63.80	3.3
KNN	%50.64	0.6
Decision Tree	%57.25	0.5

Table 1 - Yöntem Başarı Oranları ve Eğitim Süreleri

VI. Sonuç

Başlangıçta futbol maçlarının sonuçlarını tahmin etme üzerine düşünülen bu projede, 40 adet öznitelik ve 20 sezonluk maçlardan toplanan diğer veriler kullanılarak, modeli eğitmek için bazı yöntemler kullanılmıştır. Bu yöntemler arasında benzer sonuçlar olmakla birlikte beklenen bazı değerler de ortaya çıkmıştır. Artık dünyada futbol maçlarının sonucunu tahmin etmenin çok büyük bir endüstri olduğu göz önünde bulundurulduğunda kullanılan yöntemlerden logistic regression ve random forest'ın ortalamanın üstünde bir başarı gösterdiği söylenebilir. Diğer yöntemler de onlara yakın ancak daha vasat performanslar göstermişlerdir.

VII. Literatür Taramasi

1. "Predicting football match outcomes using artificial neural networks" (2022) Yiannis B. and Foteini S., Journal of Sports Analytics

Derin öğrenme yöntemleri, özellikle sinir ağları kullanarak futbol maçlarının sonuçlarını tahmin eder. Özellikle, veri ön işleme ve özellik mühendisliği konularına odaklanır.

2. "A Comparison of Machine Learning Models for Football Match Outcome Prediction" (2023) Markus K., Hannah G., IEEE Transactions on Computational Intelligence and AI in Games

Çeşitli makine öğrenme modelleri (karar ağaçları, random forest, gradient boosting vb.) karşılaştırılmış ve hangi modellerin daha iyi performans gösterdiği analiz edilmiştir.

Linkler

https://github.com/furkan-koc/Yap470

https://drive.google.com/file/d/1lYoTJNDWtpXRn7GBr-IMRZSA44w1QX4v/view?usp=drive_link

References

[1] https://www.kaggle.com/code/saife245/football-match-prediction/input