Maç Sonucu Tahmini İçin Veri Seti Oluşturma

Ensar Akbaş
Bilişim Sistemleri Mühendisliği
Kocaeli Üniversitesi
ensarr.akbas@gmail.com

***Özet*—Bu projede, futbol maç sonucu tahmin modelleri için kapsamlı bir veri seti oluşturulmuştur. Veri, Maçkolik sitesinden web scraping yöntemiyle toplanmıştır. Liverpool, Manchester City, Manchester United, Arsenal, Chelsea ve Tottenham takımlarına ait bilgiler derlenmiştir. Ham verilerde yer alan temel istatistiklere ek olarak, özellik mühendisliği (feature engineering) teknikleriyle yeni değişkenler üretilmiştir. Eksik veriler uygun yöntemlerle doldurulmuş, korelasyon analizi ve görselleştirmeler ile veri yapısı detaylı olarak incelenmiştir. Proje, maç istatistiklerinden öğrenebilen modeller için kaliteli, temiz ve anlamlı bir veri seti oluşturmayı hedeflemektedir.**

***Keywords*—Web scraping, Feature engineering, Veri seti, Veri ön işleme**

I. GIRIŞ

Futbol, yalnızca bir spor dalı olmanın ötesinde, büyük bir veri kaynağı ve analiz alanı haline gelmiştir. Bu projede amaç, geçmiş futbol maçlarından elde edilen istatistiksel verileri kullanarak ileriye dönük tahminler yapılabilecek nitelikte bir veri seti oluşturmaktır. Özellikle İngiltere Premier Lig'in önde gelen takımlarına ait uzun dönemli maç verileri incelenmiş, her bir maçın istatistiksel özellikleri detaylı şekilde toplanarak analiz edilebilir bir yapıya dönüştürülmüştür. Proje kapsamında web scraping, veri temizleme, veri ön işleme ve feature engineering işlemleri gerçekleştirilmiştir.

II. VERI TOPLAMA

Veriler, Maçkolik web sitesinden *Python* programlama dili kullanılarak çekilmiştir. Kodlama işlemleri *PyCharm* editörü üzerinde yürütülmüş, verileri çekmek için *Selenium* kütüphanesi tercih edilmiştir. Ayrıca verilerin işlenmesi ve analizinde *pandas, numpy* ve görselleştirme için *matplotlib* kütüphanelerinden yararlanılmıştır. Süreç boyunca altı farklı takımın (Liverpool, Manchester City, Manchester United, Chelsea, Arsenal, Tottenham) her biri için yaklaşık 700 maçlık veri toplanmıştır. Bu işlemler sırasında Google Chrome Driver kullanılmış ve eri çekimi sırasında Maçkolik platformu herhangi bir bot koruması veya erişim engeli uygulamadığı için süreç kesintisiz ve verimli bir şekilde tamamlanmıştır.

III. VERI SETI

Projenin veri seti, 2013-2014 sezonu ile 2024-2025 sezonu arasındaki Premier Lig maçlarını kapsamaktadır. Liverpool, Manchester City, Manchester United, Chelsea, Arsenal ve Tottenham olmak üzere 6 farklı takım için oluşturulan bu veri setinde, her takım adına yaklaşık 700 maç verisi bulunmaktadır. Bu kapsamda, toplam 32 özellik (*feature*) yer almaktadır. Özelliklerin bir bölümü doğrudan maç istatistiklerinden elde edilirken; bazıları ise özellik mühendisliği (*feature engineering*) süreçleriyle türetilmiştir. Bu yapı, takımların geçmiş performanslarını daha anlamlı şekilde modelleyebilmek için tasarlanmıştır.

A. Features (Özellikler)

Tarih	Rakip Takım	Takım ID
Is_Home	Sonuç	Gol
Rakip Gol	Topla Oynama(%)	Şut
İsabetli Şut	Başarılı Pas	Pas Başarısı(%)
Korner	Faul	Ofsayt
Rakip Topla Oynama (%)	Rakip Şut	Rakip İsabetli Şut
Rakip Başarılı Pas	Rakip Pas Başarısı(%)	Rakip Korner
Rakip Faul	Rakip Ofsayt	

Tablo 1. *Maçkolik üzerinden çekilen özellikler*

- **Takım ID:** Takıma özgü kimlik numarasıdır. (Örneğin: Liverpool=1, Manchester City=2)
- **Is_Home:** Maçın ev sahibi olarak oynanıp oynanmadığını belirtir. (Ev Sahini=1, Deplasman=0)
- **Sonuç:** Galibiyet: 1, Beraberlik: 0, Mağlubiyet: -1 olarak tutulmaktadır.

B. Feature Engineering (Özellik Mühendisliği)

Şut Verimliliği (İsabetli Şut/Şut)	Gol Farkı	Sezon
Ay	Haftanın Günü	Son 5 Maçın Gol Ortalaması
Son 5 Maçın Galibiyet Oranı		

Tablo 2. *Feature engineerimg sonucu oluşan yeni özellikler*

- **Sezon/Ay:** Maçın hangi yıl ve ayda yapıldığını ifade eder. (Örneğin: Sezon=2024, Ay=4)
- **Haftanın Günü:** Maçın hangi gün yapıldığını ifade eder. (Örneğin: 1=Pazartesi, 2=Salı)

Bu veri seti, sonraki aşamalarda analiz, görselleştirme ve makine öğrenmesi modelleri için temel kaynak olarak kullanılacaktır.

IV. VERI ÖN İŞLEME VE TEMIZLEME

Toplanan ham veriler, analiz ve modelleme süreçlerinde kullanılmadan önce çeşitli temizleme ve ön işleme adımlarından geçirilmiştir. Bazı maçlarda istatistiklerin eksik olması nedeniyle, ilgili sütunlarda boş (*null*) değerler tespit edilmiştir. Bu durum özellikle "*Topla Oynama*", "*Pas Başarısı*", "*Başarılı Pas*" gibi istatistiklerde görülmüştür. Bu eksik verileri doldurmak amacıyla iki farklı yaklaşım kullanılmıştır:

Sütun Ortalamasıyla Doldurma: Eksik değerler, ilgili sütunun genel ortalaması ile doldurularak, verinin genel yapısı korunmuştur.

Önceki Verilere Yakınlaştırma: Bazı durumlarda, eksik değerler bir önceki veya benzer maçlardaki istatistikler göz önünde bulundurularak manuel olarak doldurulmuştur. Böylece ani sapmaların önüne geçilmiştir.

Ayrıca, verilerin bütünlüğünü korumak adına tarih biçimleri, takım ID'leri ve sonuç değerleri gibi temel alanlarda format düzeltmeleri ve tutarlılık kontrolleri yapılmıştır. Bu adımlar, modelin güvenilir sonuçlar üretebilmesi için temel bir hazırlık sürecidir.

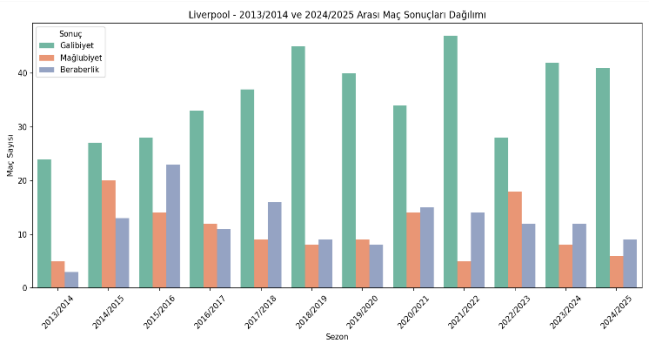
V. VERI GÖRSELLEŞTIRME

Veri analizi sürecinde, istatistiksel içgörülerin daha anlaşılır ve yorumlanabilir hale gelmesi için veri görselleştirme önemli bir adımdır. Bu projede, görselleştirme işlemleri hem genel eğilimleri ortaya koymak hem de değişkenler arasındaki ilişkileri incelemek amacıyla gerçekleştirilmiştir. Özellikle Liverpool takımı üzerinde yapılan analizlerle örneklemeler sunulmuştur.

Liverpool Üzerinden Görselleştirme Örnekleri

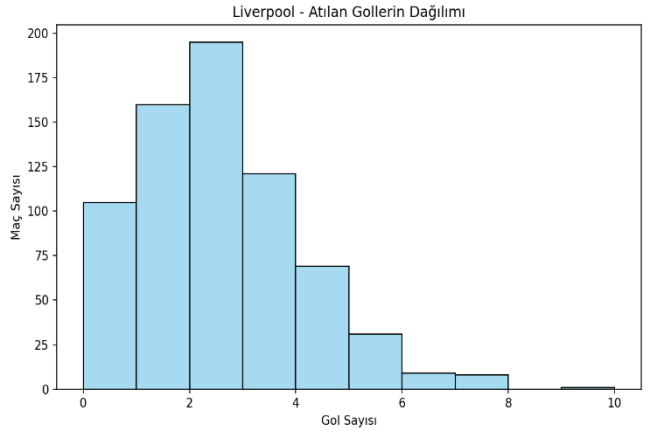
A. *Sezonlara Göre Maç Sonuçları Dağılımı*

Bu grafik ile her sezon boyunca kazanılan, kaybedilen ve berabere biten maçların dağılımı görselleştirilmiştir. Böylece takımın dönemsel performans dalgalanmaları net bir şekilde gözlemlenmiştir.



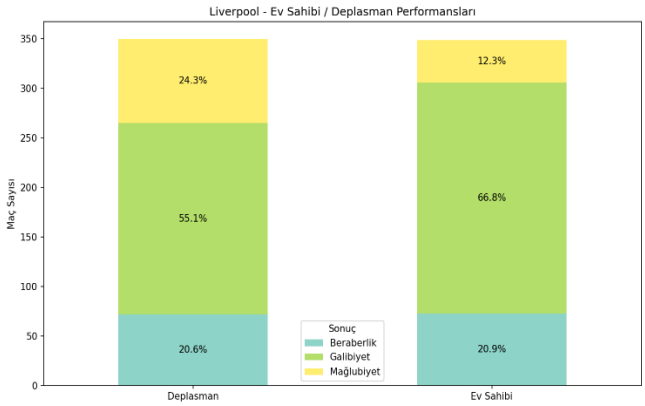
B. *Atılan Gollerin Dağılımı*

Takımın maç başına attığı gollerin frekansı incelenmiş, hangi gol aralığında en çok maç oynandığı grafiklerle gösterilmiştir.



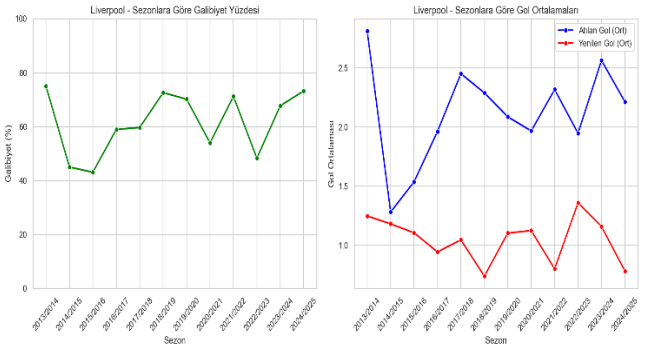
C. *Ev Sahibi / Deplasman Performansları*

Takımın iç saha ve dış saha maçlarındaki galibiyet oranları karşılaştırılarak performans farkları ortaya konmuştur.



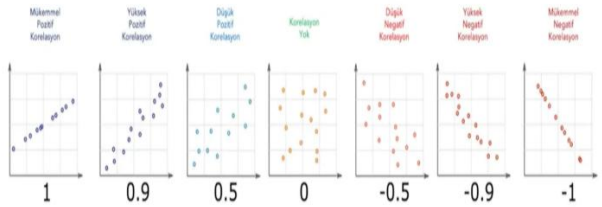
D. *Sezonlara Göre Galibiyet Yüzdesi ve Gol Ortalamaları*

Her sezonun galibiyet yüzdesi ile maç başına düşen gol ortalaması birlikte incelenerek takımın genel başarımı değerlendirilmiştir.



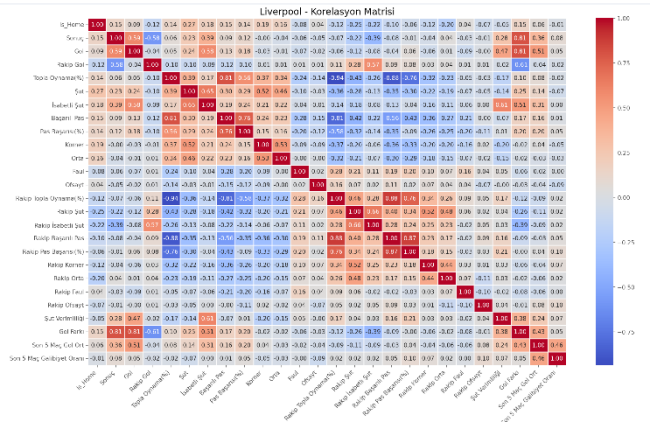
Korelasyon Analizi

Korelasyon analizi, iki değişken arasındaki doğrusal ilişkinin yönünü ve gücünü ölçen istatistiksel bir yöntemdir. Bu projede, maç istatistikleri arasındaki ilişkileri incelemek amacıyla korelasyon analizinden yararlanılmış ve sonuçlar aşağıdaki başlıklar altında görselleştirilmiştir.



E. *Korelasyon Matrisi*

Tüm sayısal değişkenler arasındaki korelasyon katsayıları ısı haritası ile gösterilerek, hangi değişkenlerin birbirini daha fazla etkilediği görsel olarak sunulmuştur.



F. *İsabetli Şut ve Gol İlişkisi*

Orta düzeyde pozitif bir ilişki gözlemlenmiştir. Bu sonuç, daha fazla isabetli şut çeken takımların gol atma olasılığının önemli ölçüde arttığını göstermektedir.

G. *Topla Oynama ve Sonuç İlişkisi*

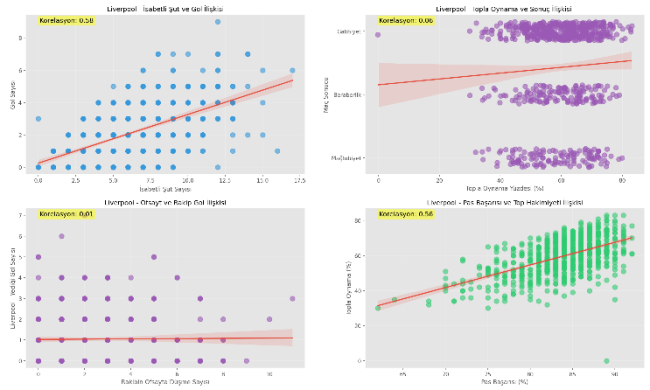
Bu oldukça düşük korelasyon, topa sahip olma yüzdesinin maç sonucunu doğrudan belirlemediğini, başka faktörlerin daha etkili olduğunu ortaya koymaktadır.

H. *Ofsayt ve Rakip Gol İlişkisi*

Neredeyse sıfıra yakın bir ilişki bulunmuştur. Bu, hücumdaki ofsayt hatalarının savunma performansı veya rakip gol sayısıyla bağlantılı olmadığını göstermektedir.

İ. *Pas Başarısı ve Top Hakimiyeti*

Orta düzeyde güçlü bir pozitif ilişki gözlemlenmiştir. Başarılı pas oranı arttıkça topa sahip olma yüzdesinin de arttığı açıkça görülmektedir.



Bu görselleştirmeler, hem istatistiksel değerlendirme hem de modelleme aşamasında önemli ipuçları sağlamış ve veriyle ilgili çıkarımların doğruluğunu artırmıştır.

KAYNAKÇA

[1] Taşpınar, Y. S., Çınar, İ., & Koklu, M. (2021). Improvement of Football Match Score Prediction by Selecting Effective Features for Italy Serie A League. MANAS Journal of Engineering, 9(1), 1-9. https://doi.org/10.51354/mjen.802818

[2] Zaveri, Nilay, et al. "Prediction of football match score and decision making process." *International Journal on Recent and Innovation Trends in Computing and Communication* 6.2 (2018): 162-165.

[3] Igiri, Chinwe Peace, and Enoch Okechukwu Nwachukwu. "An improved prediction system for football a match result." *IOSR journal of Engineering* 4.12 (2014): 12-20.

[4] https://arsiv.mackolik.com/Teams/OptaList.aspx

[5] https://medium.com/machine-learning-t%C3%BCrkiye/korelasyon-nedir-python-ile-korelasyon-analizi-fbd24deb20dd

[6] https://gokerguner.medium.com/machine-learning-2-korelasyon-matrisi-%C3%B6zellik-se%C3%A7imi-s%C4%B1n%C4%B1flar%C4%B1n-dengesizli%C4%9Fi-karar-a%C4%9Fa%C3%A7lar%C4%B1-af993bd8ea66

- Google Drive Linki
- GitHub Linki