YTÜ FEN EDEBİYAT FAKÜLTESİ
İSTATİSTİK BÖLÜMÜ

MAKİNE ÖĞRENMESİ VE SPORDA VERİ ANALİTİĞİ: YAPAY ZEKA TABANLI METRİKLERİN MAKİNE ÖĞRENMESİ İLE PERFORMANS ANALİZİ VE SENARYO TAHMİNİNDE KULLANILMASI

Can İPEK – 20023072
Danışman: Arş. Gör. Dr. Coşkun PARİM

ÖZET

Modern analiz metotları dendiğinde akla ilk gelen kavram olan makine öğrenmesi, günümüzde sürekli olarak değişen oyun yapısına rağmen başarıya giden yolun en hızlı tahminleri yaparak en erken kararları almaktan geçtiği spor sektörünün önemli bir parçası haline gelmiştir. Kısa bir tarihe sahip olan spor analitiği ortaya çıktığı ilk dönemden itibaren yükselen ivme ile gelişmişken, son dönemlerde yapay zekânın gelişimiyle birlikte ortaya çıkan metriklerle yeni bir anlam kazanmaktadır. Bu çalışmada kullanılan veri, İngiltere Premier Ligi 2023-2024 sezonunda oynanmış 380 maçı kapsamaktadır. Veri seti sezon geneli oyuncu ve takım istatistiklerinden dakikaya kadar değişen granülitlerde veri içerir. İçerdiği değişkenler arasında şut tipi, atak şekli gibi kategorik veya gol sayısı, defansif aksiyon başına pas sayısı, gibi sayısal istatistiklerin yanı sıra, son dönemlerde popülaritesi artan xG, xA gibi yapay zekâ tabanlı yeni nesil metrikler de bulunmaktadır. Görselleştirmeler ile veri setindeki bazı değişkenlerin farklı kırınımlardaki ilişkileri ile ilgili çeşitli çıktılar elde edilmiş, performans analizleri ile ilgili uygulanabilir yenilikler tartışılmış ve üzerine yorumlar yapılmıştır. Sonrasında ise makine öğrenmesi kullanılarak lineer regresyon, rastgele orman, destek vektör makinesi ve yapay sinir ağları modelleri kullanılmış ve elde edilen verilere göre en doğru sonuçları veren modeller seçilmiştir. Tüm bu analizlerin teknik anlamda karar almaya sağlayabileceği katkılar üzerine tartışılmıştır. Uygulamadaki tüm çalışmalarda Python kullanılmıştır.

Sporda Veri Analitiği

Spor veri analitiği, sporcular, antrenörler ve hakemler hakkında saha düzeyinde analizleri, yöneticilerin kararlarını ve ekonomi ve psikoloji alanındaki analizleri içerir. [2] Bazı popüler spor branşlarında, saha düzeyinde analiz ile ilgili gelişmeler aşağıdaki gibidir.

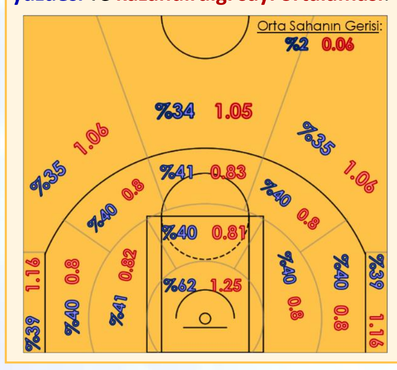
Beyzbol

Oyun yapısı itibariyle veri analitiğinin belirleyici konumda olduğu bir spor dalı olan beyzbol, aynı zamanda veri analitiğinde 1960'lı yıllara dayanan geçmişi ile bu alanda en eski spor dallarından biridir. 2003 yılında Michael Lewis'in kaleme aldığı, beyzbol kulübü Oakland Athletics takımında görev yapan Billy Beane'in, veri odaklı yönetimini konu alan '*Moneyball: The Art of Winning an Unfair Game kitabı*' ve bu kitabın 2011 yılında sinemaya uyarlanmasının yarattığı etki, sporda veri analitiği üzerine yapılan çalışmaların artmasında etkili olmuştur. Beane, oyundaki bazı ilişkileri dikkate alarak kurduğu düşük bütçeli kadrosuyla büyük başarılara imza atmıştır [5].

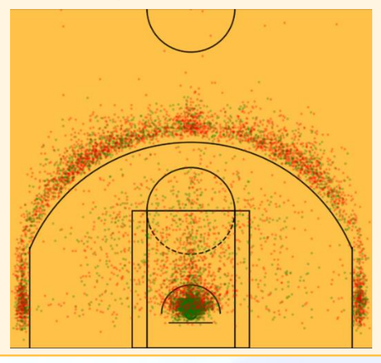
Basketbol

Basketbolda 90'lı yıllarda başkalaşım geçiren ve hız kazanan veri analitiği çalışmalarının, özellikle son yıllarda etkisi artmıştır. Houston Rockets takımında uzun yıllar yönetici görevi üstlenmiş Daryl Morey'in, orta mesafeli atışlardan kaçınmayı öneren veri odaklı yaklaşımı bazı yorumcular ve taraftarlar tarafından eleştirilmiş olsa da Rockets önemli başarılar elde etmiştir [7]. 2017-2018 sezonunda Rockets, sakatlıklara rağmen Batı Konferansı finallerine yükselmiş ve aynı sezon 3 sayılık atış rekorunu kırarak Batı Konferansı'nı 1. sırada tamamlamıştır.

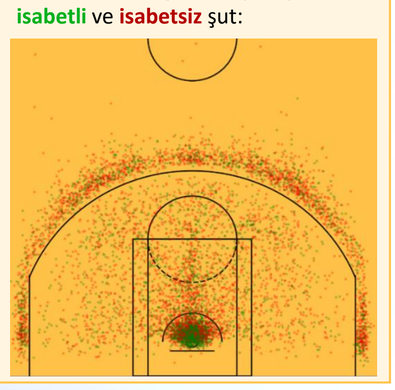
1997 – 2020 yılları arasında NBA'de atılmış tüm şutların bölgesel **isabet yüzdesi** ve **kazandırdığı sayı ortalaması**:



2017-18 sezonunda Houston Rockets'ın attığı 8 bin **isabetli** ve **isabetsiz** şut:



2017-18 sezonunda diğer takımlar arasından rastgele seçilmiş 8 bin **isabetli** ve **isabetsiz** şut:



'Sıcak El' Tartışması

1985 yılında yapılan bir çalışmada, basketbol oyuncularının ardışık başarılı atışlar yapmasının sonraki atışların başarı olasılığını artırmadığı öne sürülmüştür [8]. Bu çıkarım bazı basketbol takımlarının şutları ile ilgili toplanan verileri analiz ederek yapılmış ve aksi uzun yıllar bilimsel olarak ispat edilemese de, izleyici ve araştırmacılar arasında tartışma konusu olmuştur. Bu çalışma, psikoloji ve davranış ekonomisi gibi farklı alanlardaki çalışmalara dahi konu olmuştur [10].

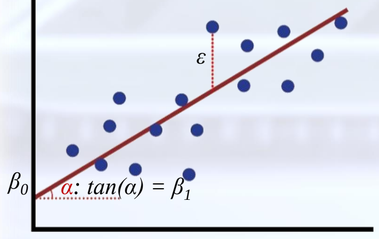
Futbol

1950'lerde Charles Reep'in atak öncesi yapılan pas sayısı ile atakların gol ile sonuçlanması arasındaki ilişkiyi saptayarak kurguladığı oyun modelinin geçerliliğini uzun yıllar sürdürmesiyle başlayıp, 90'lı yıllarda ve sonrasında çok kez değişip gelişen futbolda veri analitiği, geçmişte olduğu gibi bugün de farklı fikirlerin geçerli olduğu dönemler geçirmeye devam ederken, son yıllarda yapay zeka tabanlı metriklerin kullanılmaya başlamasıyla yeni bir anlam kazanmıştır [9].

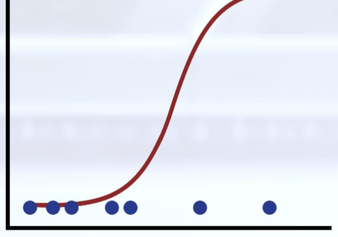
Makine Öğrenmesi

Makine öğrenmesi, bilgisayar bilimi ile istatistiğin kesiştiği noktada, yapay zekâ ve veri biliminin merkezinde yer alan, günümüzün en hızlı büyüyen teknik alanlarından biridir. Makine öğrenmesi, bilgisayara bir işlem için özel olarak kodlanmadan bu işlemi yapabilme becerisinin kazandırılmasını sağlayan çalışma alanıdır [6]. Makine öğrenmesi, denetimli, denetimsiz, yarı denetimli ve pekiştirmeli öğrenme olarak dörde ayrılır. Aşağıda bir denetimli öğrenme sınıfı olan regresyonda kullanılan bazı modellere örnekler verilmiştir.

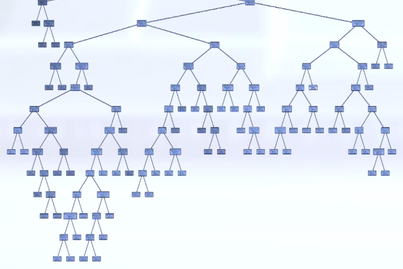
Lineer Regresyon



Lojistik Regresyon



Karar Ağaçları

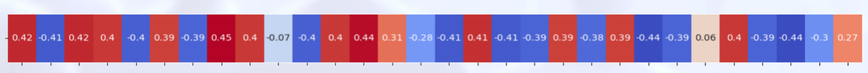


Uygulama

Uygulama bölümünde, İngiltere Premier Ligi 2023-2024 sezonuna ait veriler kullanılarak, makine öğrenmesi modelleri kullanılarak olabildiğince yüksek güçte çıkarımsal, tahmine yönelik sonuçlar elde edilmeye çalışılacak, xG istatistiği üzerine tartışılacak ve performans analizi için ne gibi çıkarımlar yapılabileceği incelenecektir.

1) Makine Öğrenmesi ile Senaryo Tahmini

Aşağıda '**Puan Beklentisi'** değişkeninin tüm değişkenler ile arasındaki korelasyon katsayıları ve değişkenlerden 'iç/dış saha', 'rakip lig geneli puan beklentisi', 'rakip lig geneli gol beklentisi' ve 'takım lig geneli puan beklentisi' değişkenleri ile tahmin edilmesi için kurulan regresyon modellerinin sonuçları mevcuttur.



MODEL	LİNEER REGRESYON		REGRESYON AĞACI		DESTEK VEKTÖR MAKİNESİ		YAPAY SİNİR AĞLARI	
Veri	Test	Eğitim	Test	Eğitim	Test	Eğitim	Test	Eğitim
Ortalama Kare Hatası	0.38	0.42	0.47	0.4	0.4	0.42	0.37	0.42
Ortalama Mutlak Hata	0.5	0.54	0.55	0.51	0.49	0.5	0.49	0.54
Karekök Ortalama Kare Hatası	0.62	0.65	0.69	0.63	0.63	0.65	0.61	0.65
R^2	**0.49**	**0.46**	**0.38**	**0.49**	**0.48**	**0.46**	**0.51**	**0.46**

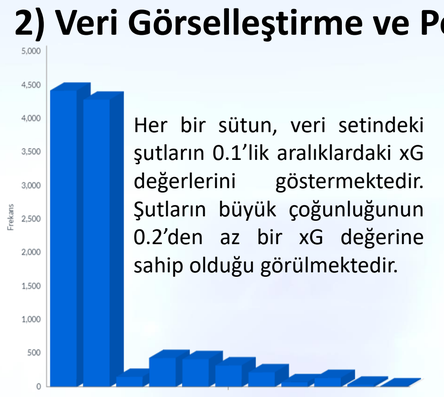
Diğer değişkenlerin tahmini için aynı metotla seçilen modeller ve test verisinden aldıkları R^2 değerleri:

Bağımlı Değişken	Seçilen Model	Test R^2
Kaleye Yakın Pas Sayısı	Yapay Sinir Ağları	**0.49**
Gol Beklentisi **(xG)**	Destek Vektör Makinesi	**0.35**
Gol Beklentisi Toleransı **(xGA)**	Lineer Regresyon	**0.34**

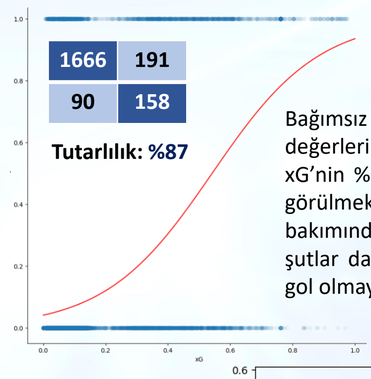
xG İstatistiği

'xG', bir futbol takımının veya oyuncunun belirli bir pozisyonda gol atma olasılığını ifade eder. Bu model, bir şutun gol olma olasılığını değerlendirmek için şutun mesafesi, açısı, şutun yapıldığı pozisyon gibi faktörleri kullanır. Hesaplama, genellikle lojistik regresyon veya diğer makine öğrenimi modelleriyle yapılır [3][4]. Bir şutun 0.2 xG değerine sahip olması, bu şut 100 kez denendiğinde ortalama 20 tanesinin gol ile sonuçlanacağının tahmin edildiği anlamına gelir.

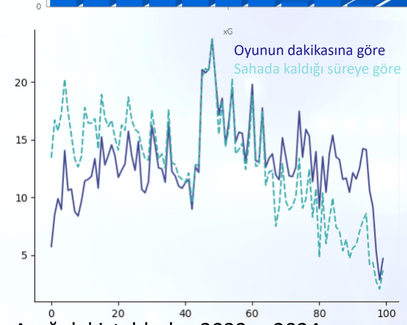
2) Veri Görselleştirme ve Performans Analizi



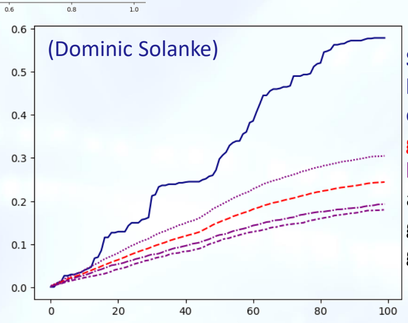
Her bir sütun, veri setindeki şutların 0.1'lik aralıklardaki xG değerlerini göstermektedir. Şutların büyük çoğunluğunun 0.2'den az bir xG değerine sahip olduğu görülmektedir.



Bağımsız ve bağımlı değişken olarak sırasıyla xG ve gol değerlerinin kullanıldığı lojistik regresyon modeliyle, xG'nin %87 oranında gol olma durumunu tanımladığı görülmektedir. Buna karşın, gol olmayan şutlar xG bakımından sağa çarpık dağılım gösterirken, gol olan şutlar daha tekdüze dağılmışlardır. Bu yüzden model, gol olmayan şutları tahmin etmede daha başarılıdır.



Oyuncuların xG katkısının maçın dakikasına ve oyunda kalma süresine göre değişimi yer yandaki grafikte yer almaktadır. Oyuncuların sahada kalma süresi arttıkça ortalama xG katkısının genel ortalamanın altına düştüğü görülmektedir.

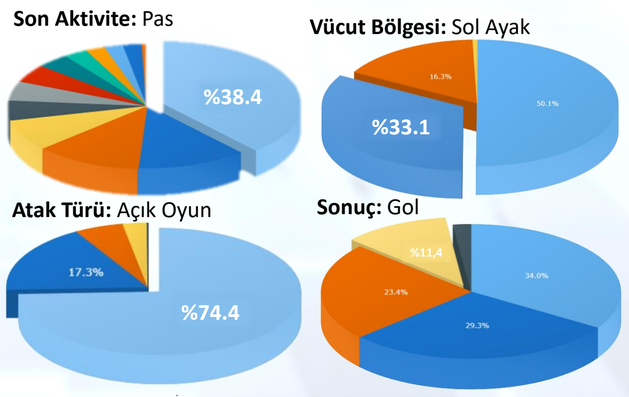


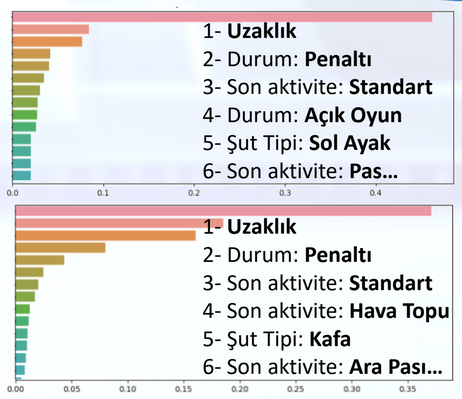
Seçilen herhangi bir oyuncunun ortalama xG artışı, genel ve **mevki bazında** ortalama artışlar yandaki grafikte görselleştirilmiştir.

Aşağıdaki tabloda, 2023 - 2024 sezonunu 27 gol ile en fazla gol atan oyunu olarak tamamlayan Erling Haaland'ın şutlarının toplam xG değerinin, gollerinin toplamından 4,65 gol daha fazla olduğu görülmektedir.

Oyuncu Adı	Gol	xG	Takım	Gol - xG Farkı
Erling Haaland	27	31.65	Manchester City	**-4.65**

Bu, oyuncunun elde ettiği şut imkanlarını ortalamanın altında bir başarıyla değerlendirdiğini gösterir. Yani yüksek gol sayısı, pozisyonları gole çevirebilme kabiliyetinden daha çok, fazla pozisyona girme becerisiyle ilişkilidir. Bu durumun sebepleri çeşitli analiz ve görselleştirme metotlarıyla incelenebilir. Aşağıda örnek olarak ligdeki tüm şutların, 4 farklı kategori grubundan seçilmiş birer kategorinin, tüm kategorilerin frekansına göre dağılımının pasta grafiği ve ilgili oyuncu Erling Haaland'ın bu kategorilere ait şutlarının ısı haritası ile grafiği mevcuttur.





Oyuncunun xG değerinin gol sayısından daha yüksek olması oyuncu ile ilgili olması olası olsa da, bu sebep birçok farklı tipte oyuncunun attığı şutların verisi ile eğitilmiş xG modelinin ilgili oyuncunun gol olasılıklarını tahmin etmede genelde olduğu kadar başarılı olamaması da olabilir.

Yandaki grafiklerde ilgili oyuncu Erling Haaland'ın attığı şutların yukarıdakine benzer çeşitli özelliklerini bağımsız değişken olarak kullanarak oluşturulmuş, xG istatistiğini tahmin etmek için rastgele orman regresyonu ve gol değişkenini tahmin eden rastgele orman sınıflandırması algoritmalarının bağımsız değişkenlerinin önem sıralaması mevcuttur. Buradan, Haaland'ın attığı şutların gol olma olasılığını etkileyen faktörlerin, xG istatistiğinin derecesini etkileyen faktörlerden farklı olduğu görülmektedir. Sonuç olarak antrenman verileri gibi verileri kullanarak xG metriklerinin oyunculara özel olarak tasarlanması, oyuncu bazında tutarlılığı arttıracak ve analizleri geliştirecektir.

KAYNAKÇA

[1] Geron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd ed.). O'Reilly Media.

[2] Morgulev, E., Azar, O.H. & Lidor, R. Sports analytics and the big-data era. Int J Data Sci Anal 5, 213–222 (2018). https://doi.org/10.1007/s41060-017-0093-7

[3] Hewitt, J. H., & Karakuş, O. (2023). A Machine Learning Approach for Player and Position Adjusted Expected Goals in Football (Soccer). *arXiv preprint*. https://doi.org/10.48550/arXiv.2301.13052

[4] Mead, J., O'Hare, A., & McMenemy, P. (2023). Expected goals in football: Improving model performance and demonstrating value. *PLOS ONE, 18*(4), e0282295. https://doi.org/10.1371/journal.pone.0282295

[5] Lewis, M. (2003). *Moneyball: The art of winning an unfair game*. W. W. Norton & Company.

[6] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3(3), 210-229.

[7] web.archive.org/web/20150325075449/http://www.houstonpress.com/2007-11-01/news/rocket-science/

[8] Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17*(3), 295–314. https://doi.org/10.1016/0010-0285(85)90010-6

[9] Pollard, R. (2002). Charles Reep (1904-2002): pioneer of notational and performance analysis in football. *Journal of Sports Sciences*, *20*(10), 853–855. https://doi.org/10.1080/026404102320675684

[10] Rabin, Matthew and Vayanos, Dimitri, The Gambler's and Hot-Hand Fallacies: Theory and Applications (February 2007). CEPR Discussion Paper No. 6081, Available at SSRN: https://ssrn.com/abstract=1004563