T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN EDEBİYAT FAKÜLTESİ

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

Can İPEK

LİSANS TEZİ

İstatistik Anabilim Dalı

Danışman

Arş. Gör. Dr. Coşkun PARİM

Mayıs, 2024

Danışmanım Arş. Gör. Dr. Coşkun PARİM sorumluluğunda tarafımca hazırlanan "Makine Öğrenmesi ve Sporda Veri Analitiği: Yapay Zeka Tabanlı Metriklerin Makine Öğrenmesi ile Performans Analizi ve Senaryo Tahmininde Kullanılması" başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Can İPEK

İmza

TEŞEKKÜR

Tez çalışmam boyunca bilgi birikimi ve deneyimleri ile bana yol gösteren değerli danışman hocam sayın Arş.Gör.Dr. Coşkun PARİM'e, desteklerini asla esirgemeyen aileme, özellikle tüm çalışmalarımdaki asıl gayemi bana öğretmiş olan babama ve Yıldız Teknik Üniversitesi İstatistik bölümündeki tüm arkadaşlarıma sonsuz teşekkürlerimi sunarım.

Can İPEK

İÇİNDEKİLER

KISALTMA LİSTESİ

ANN	Artificial Neural Networks
KKT	Karush–Kuhn–Tucker
KNN	K-Nearest Neighbors
MAE	Mean Absolute Error
MLB	Major League Baseball
MSE	Mean Squared Error
NBA	National Basketball Association
NP	Non-Penalty
OLS	Ordinary Least Squares
PPDA	Passes Allowed Per Defensive Action
RMSE	Root Mean Squared Error
SVM	Suport Vector Regressor
SVR	Support Vector Regressor
xA	Expected Assists
xG	Expected Goals

ŞEKİL LİSTESİ

TABLO LİSTESİ

ÖZET

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

Can İPEK

İstatistik Anabilim Dalı

Lisans Tezi

Danışman: Arş. Gör. Dr. Coşkun PARİM

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 yeni teknikler, ileriye dönük tahmin edici analizlere yeni bir boyut kazandırmıştır. Yeni nesil metrikler, makine öğrenmesi ile birlikte ileriye dönük analizler için kullanıldığında, elde edilen çıktıların değerinin artması kaçınılmazdır.

Bu çalışma, sporda veri analitiği alanında yapılmış bazı önemli çalışmalar ile başlıca makine öğrenmesi metotları hakkında literatür analizini içermesinin yanında, futbolda performans analizi ve oyun senaryosu tahminlemeyi amaçlayan makine öğrenmesi uygulamalarını içerir. Sporda veri analitiğinin farklı branşlarda

hangi amaçlarla kullanıldığına, geçmişte ne tür çalışmalar yapıldığında ve bu çalışmaların spor dünyasındaki etkilerinin yanı sıra, spor dışındaki disiplinler üzerindeki etkilerine de değinilir. Uygulama bölümündeki çalışmada, understat.com sitesinden elde edilen veriler kullanılmıştır. Veri seti İngiltere Premier Ligi 2023-2024 sezonunda oynanmış 380 maçı kapsamaktadır. Sezon geneli oyuncu ve takım istatistiklerinden dakikaya kadar değişen granülitlerdeki veriyi 5 farklı temel tabloda tutmaktadır. İçerdiği değişkenler arasında şut tipi, atak şekli gibi kategorik veya gol sayısı, kaleye yakın bölgedeki pas sayısı gibi sayısal istatistiklerin yanı sıra, son dönemlerde popülaritesi artan gol beklentisi (xG), asist beklentisi (xA) gibi yapay zekâ tabanlı yeni nesil metrikleri de içermektedir.

Çalışmadaki tüm işlemlerde Python dili kullanılmıştır. İlk olarak tanımlayıcı çıktılar elde etmek üzere tasarlanmış olan ham veri seti, gerekli veri ön işleme işlemleri ile tahmin edici analiz için uygun formda yeni veri setleri elde etmek için kullanılmıştır. Ardından, görselleştirmeler ile veri setindeki bazı değişkenlerin özellikleri ile ilgili çeşitli çıktılar elde edilmiş ve üzerine yorumlar yapılmıştır. Gol beklentisi istatistiğinin gerçek gol olasılığını temsil gücü test edilmiş, oyuncular arasında bireysel bazda tutarlılık farkları üzerine tartışılmıştır. Sonrasında ise maç içerisindeki bazı olayları önceden tahmin edebilmeyi amaçlayan; lineer regresyon, random forrest regresyonu, destek vektör regresyonu ve yapay sinir ağları gibi makine öğrenmesi modelleri kullanılmış ve elde edilen verilere göre modeller doğruluk bakımından kıyaslanmıştır. Tüm bu analizlerin teknik anlamda karar almaya sağlayabileceği katkılar üzerine tartışılmıştır.

Anahtar Kelimeler: Makine Öğrenmesi, Spor Analitiği, Performans Analizi, Senaryo Tahmini, Python, Gol Beklentisi, Lineer Regresyon, Regresyon Ağacı, Destek Vektör Regresyonu, Yapay Sinir Ağları

ABSTRACT

MACHINE LEARNING & SPORTS ANALYTICS: USAGE OF AI-BASED METRICS WITH MACHINE LEARNING ALGORITHMS IN PERFORMANCE ANALYSIS & SCENARIO PREDICTION

Can İPEK

Department of Statistics

Undergraduate Thesis

Supervisor: Danışman: Arş. Gör. Dr. Coşkun PARİM

The first concept that comes to mind in terms of modern analytics methods is machine learning. Today, with the dynamics changing every moment in the game, the sports sector is incomplete without machine learning—the success factor for making the earliest decision through fast predictions. Bluntly testing the predictive power of sports analytics, it has been diagnosed with the high development of varying momentum: from the time it initiated the journey to new techniques emerging with the development of artificial intelligence in recent times. Using latest-generation metrics combined with machine learning for forward-looking analyses thus further increases the value of the obtained outputs.

This study includes a literature analysis of some important studies in the field of sports data analytics and major machine learning methods. It also encompasses machine learning applications aimed at performance analysis and match events prediction in football. The study discusses the purposes for which sports data analytics is used in different branches, the types of studies conducted in the past, and the impacts of these studies not only on the sports world but also on other disciplines. In the application part of the study, data obtained from the

understat.com site were used. The dataset covers 380 matches played in the English Premier League 2023-2024 season. The dataset maintains data in 5 different fundamental data sets with granularity ranging from overall season player and team statistics to minute-by-minute data. Among the variables it includes are categorical statistics such as shot type and attack form, as well as numerical statistics like the number of goals and passes in close proximity to the goal. It also contains new generation metrics based on artificial intelligence, such as expected goals (xG) and expected assists (xA), which have recently gained popularity.

Python was used in all processes in this study. Initially designed to obtain descriptive outputs, the raw datasets were used to create new datasets in an appropriate form for predictive analysis with necessary data preprocessing steps. Subsequently, various outputs were obtained regarding the characteristics of some variables in the dataset through visualizations, and comments were made on these outputs. The representation power of the expected goal statistic for actual goal probability was tested, and differences in consistency were discussed on an individual basis among players. Later, machine learning models such as linear regression, random forest regression, support vector regression, and artificial neural networks were used to predict some events within the match, and the models were compared in terms of accuracy based on the obtained data. The contributions that all these analyses can provide in technical decision-making were discussed.

Keywords: Machine Learning, Sports Analytics, Performance Analysis, Event Prediction, Python, Expected Goals, Linear Regression, Regression Tree, Support Vector Regressor, Artificial Neural Networks

1
GİRİŞ

1.1 Literatür Taraması

Makine öğrenmesi, bilgisayar bilimi ve istatistiğin kesişimi konumundaki bir disiplin olarak, deneyim yoluyla öğrenen yazılımlar oluşturma sürecidir [31]. Büyük veri analitiği ise, yararlı bilgileri keşfetmek ve iletmek, sonuçlar önermek ve karar almayı desteklemek için büyük veriyi inceleme, temizleme, dönüştürme ve modelleme sürecidir [32]. Bu iki disiplin, özellikle büyük veri çağında spor bilimlerinde yenilikçi çözümler sunmaktadır. Sporda veri analitiği, sporcu performansının iyileştirilmesi, antrenörlük kararlarının alınması ve bu kararlarının anlaşılması gibi faydalar sağlar [45].

Performans analizi, oyunu anlamak, performans göstergelerini optimize etmek ve takım düzeni için daha iyi seçimler yapmak için takımların ve oyuncuların çeşitli metrikler kullanarak değerlendirilmesini içerir [50]. Doğrusal veya lojistik regresyon gibi makine öğrenmesi modelleri kullanarak faktörlerin performans üzerindeki etkisinin belirlenmesini ve tanımlayıcı istatistiklerin analiz edilmesini sağlar [7]. Gelecekteki sonuçları iyileştirmek için oyun davranışlarının daha iyi anlaşılmasını sağlamayı amaçlar [16].

Sakatlık önleme, sporcuların kariyerlerini sürdürebilmeleri için hayati öneme sahiptir. Sakatlık önlemede veri analitiği, yaygın görülen sakatlıkları ve bunların oyuncu ve takım performansı üzerindeki etkilerini belirlemeye yardımcı olur [54]. Özellikle geçmiş sakatlık verileri ve biyomekanik analizler kullanılarak gelecekteki sakatlıkların tahmin edilmesi mümkündür.

Takım sporlarında oyun stratejileri, maç sonuçlarını etkileyen önemli faktörlerdendir. Veri analitiği ve makine öğrenmesi, rakiplerin oyun stratejilerini analiz ederek, takımın stratejilerini optimize etmek için kullanılır [35]. Bu analizler, antrenörlerin ve analistlerin veri odaklı kararlar alabilmesine olanak tanır.

1.2 Sporda Veri Analitiği ve Bazı Farklı Sporlar Dallarındaki Gelişmeler

Sporda veri analitiği, sporcular, antrenörler ve hakemler hakkında saha düzeyinde analizleri, yönetimlerin kararlarını, ekonomi ve psikoloji alanlarındaki bir takım literatür analizini içerir [45].

Spor analitiği 2019 yılı itibarıyla 780 milyon dolarlık bir sektör haline geldi [51]. Elde edilen verilere göre, spor sektöründe veri analitiğinin kullanımının arttığı ve gelecekte de artmaya devam edeceği öngörülmektedir [55].

1.2.1 Beyzbol ve Veri Analitiği

Beyzbol, spor analitiği denildiğinde akla gelen ilk sporlardan biridir. 1960'lı yıllarda, matematikçi Earnshaw Cook'un "*Percentage Baseball*" adlı eserinin yayımlanması, beyzvolda veri analitiğinin erken gelişiminin bir göstergesidir [2]. 2003 yılında Michael Lewis'in yazdığı "*Moneyball: The Art of Winning an Unfair Game*" kitabı ve bu kitabın 2011 yılında sinemaya uyarlanması, Billy Beane'in Oakland Athletics takımında uyguladığı sabermetrik yaklaşımların geniş kitlelerce tanınmasını sağlamıştır [47].

Billy Beane, yönettiği düşük bütçeli takımında, kritik olduğunu tespit ettiği belli başlı özelliklere sahip oyuncuları tercih etmiş ve düşük maliyetli başlangıç atıcıları transfer ederek başarıya ulaşmayı başarmıştır [62]. Beane'in stratejileri, takımların koşuların optimizasyonu ve başlangıç atıcılarının düşük maliyetle transfer edilmesi gibi kararlarını içermektedir [10].

Moneyball yaklaşımı, beyzbol verilerinin analiz edilmesinin, oyuncu seçimi ve takım yönetiminde önemli avantajlar sağladığını, beyzbol veri analitiğinin organizasyonel bilgiye dayalı rekabet avantajı sağladığını ve bu bilginin kamuya açıldıktan sonra bile stratejik avantajını koruyabildiğini göstermiştir [52].

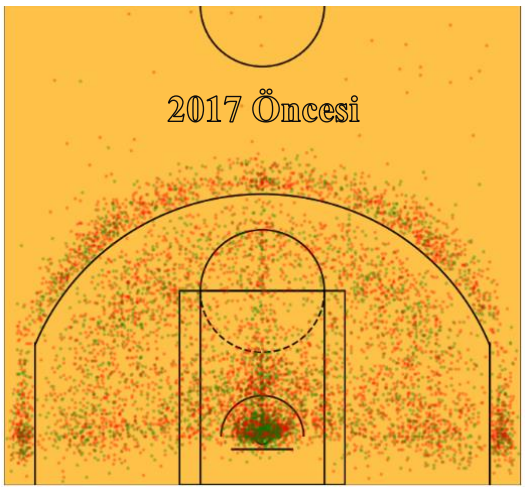
1.2.2 Basketbol ve Veri Analitiği

Basketbol tarihinde veri analitiği ilk dönemlerden itibaren kendine yer bulmuştur, ancak 1990'lara kadar kullanılan veriler kutu skorları, sayı, ribaunt ve asist gibi temel bilgileri sağlamaktaydı. Bu istatistikler oyunun yalnızca yüzeysel bir analizini sunarken, daha derinlemesine analizler için yeterli değildi [50].

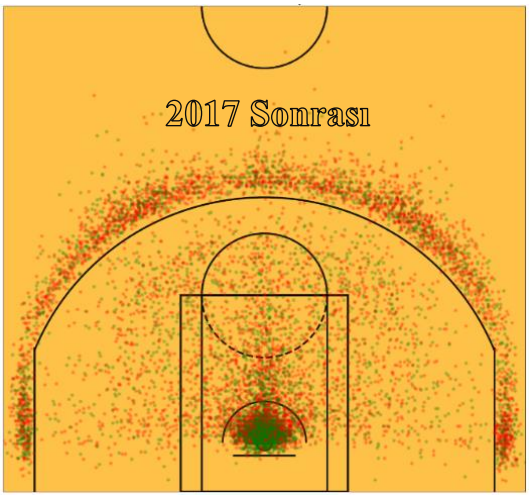
1990'ların sonlarından itibaren teknolojik gelişmeler ve veri bilimindeki yenilikler ile birlikte basketbolda daha detaylı ve ileri düzey analizler yapılmaya başlandı. Günümüzde, gelişmiş veri analitiği teknikleri sayesinde takımlar, oyuncu seçim süreçlerinden oyun stratejilerine kadar birçok alanda daha bilinçli ve veriye dayalı kararlar alabilmektedirler [56]. Sonuç olarak, basketbolda veri analitiği geçmişte daha çok tanımlayıcı analizlerle sınırlıyken, günümüzde daha ileri düzey analiz tekniklerine dayalıdır.

NBA seyircilerinin büyük bir bölümü, 2010'lu yıllarda oyun anlayışında ciddi değişimler yaşandığına inanmaktadırlar. Bu dönemden itibaren üç sayılık atışların oranı önemi hızla artmış ve bu değişim oyun planlarının merkezine yerleşmiştir. Oyuncular orta mesafeli atışlardan kaçınarak turnike ve üç sayılık atışlara daha fazla eğilim göstermeye başlamıştır.

Şekil 1.1 ve 1.2'de 1997-2020 yılları arasında NBA'de atılmış yaklaşık 5 milyon şutu içeren bir veri setinden, 2017 yılı öncesi ve sonrası için rastgele seçilmiş 10000 şut içeren örneklemlerin basketbol sahası üzerindeki görselleştirmeleri yer alıyor. Yeşil renkli noktalar isabetli atışları, kırmızı renkli noktalar isabetsiz atışları gösteriyor.



Şekil 1.1



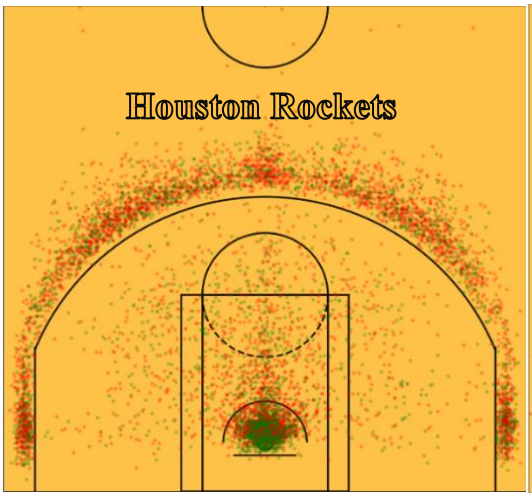
Şekil 1.2

Haritalardaki yoğun bölgeler karşılaştırıldığında, 2017 sonrası dönemde orta mesafeli atışlardan ziyade pota altı ve üç sayılık atışlara yönelik eğilim gözlemlenebilmektedir.

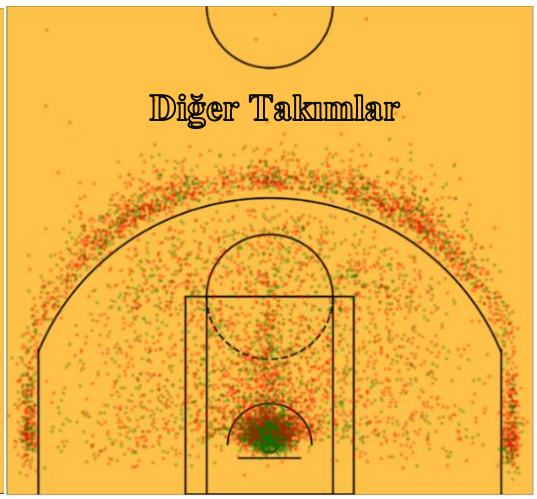
Daryl Morey, 2007'den 2020'ye kadar Houston Rockets'ın genel müdürü olarak görev yapmış bir istatistikçidir. Morey'nin istatistik ve veri analitiği konusundaki geçmişi, onun NBA'de yenilikçi stratejiler geliştirmesine olanak tanımıştır. Bu dönemde Houston Rockets, oyun tarzlarını istatistiksel analizlere dayandırarak belirgin bir şekilde 3 sayılık atışlara odaklanmıştır [40].

Morey'nin bu veri odaklı yaklaşımı bazı yorumcular ve taraftarlar tarafından eleştirilmiş olsa da Rockets önemli başarılar elde etmiştir. Örneğin, 2017-2018 sezonunda Rockets, sakatlıklara rağmen Batı Konferansı finallerine yükselmiş ve aynı sezonda bir sezonda atılan 3 sayılık atış sayısında tüm zamanların rekorunu kırarak Batı Konferansı'nı birinci sırada tamamlamıştır.

Şekil 1.3 ve 1.4'de 2017/ 2018 sezonunda Houston Rockets ve diğer NBA takımları için rastgele seçilmiş 8300 adet şut içeren örneklemlerin basketbol sahası üzerindeki görselleştirmeleri yer alıyor. Yeşil renkli noktalar isabetli atışları, kırmızı renkli noktalar isabetsiz atışları gösteriyor.



Şekil 1.3

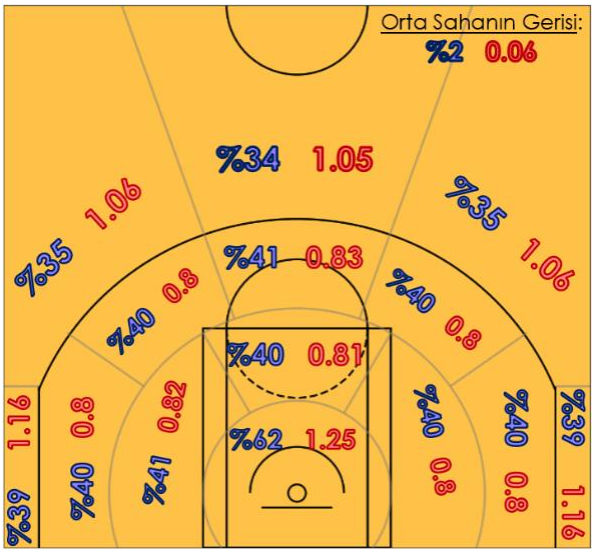


Şekil 1.4

Haritalar karşılaştırıldığında, Rockets'ın üç sayılık ve pota altı şutlarına yönelik eğiliminin, rakiplerinin eğilimi ile arasındaki farkı gözlemlenebilmektedir.

Morey'nin stratejik kararları ve Rockets'ın performansı üzerine yapılan akademik araştırmalar, genel müdürlerin teknik deneyimlerinin ve eğitimlerinin takım başarısı üzerinde önemli bir etkisi olduğunu göstermektedir [40].

Morey'in bu yaklaşımı tanımlayıcı istatistiklerin yorumlanmasıyla da açıklanabilir. Şekil 1.5'de NBA'deki 1997-2020 yılları arasında atılmış yaklaşık 5 milyon şuta dair bazı oranlar mevcuttur. Haritada mavi renkte gösterilen oranlar ilgili bölgeden atılan atışların başarılı olma oranını gösterirken, sağda kırmızı ile gösterilen sayı atış başına elde edilen ortalama sayıyı göstermektedir. Örneğin 3 sayılık bir atış bölgesinden atılan her üç şutun biri başarılı oluyor ise, atış başına elde edilen ortalama sayı 1 olur.



Şekil 1.5

Haritada genellikle pota ile atış noktası arasındaki mesafe azaldıkça atışın başarılı olma olasılığının arttığı görülse de sayılar işin içine girdiğinde orta mesafeli atışlardaki kayıp göze çarpıyor.

1.2.3 "Sıcak El" Tartışmaları

1985 yılında Thomas Gilovich, Robert Vallone ve Amos Tversky tarafından yapılan çalışmada, basketbol oyuncularının ardışık başarılı atışlar yapmasının sonraki atışların başarı olasılığını artırmadığı, yani bu anlama gelen "sıcak el" kavramının aslında bir yanılgı olduğu öne sürülmüştür. Philadelphia 76ers başta olmak üzere farklı takımlar üzerinde yapılan kontrollü deneyler sonucu ardışık atışlar arasında bu yönde hiçbir ilişkiye rastlanmamıştır [3]. Bu çalışma, araştırmacılar ve basketbol hayranları arasında günümüze kadar uzanan bir tartışmanın da başlangıcı olmuştur. Sıcak elin yanılsama olduğunu savunanlar istatistiksel kanıtların yanı sıra insan davranışları ile ilgili de kanıtlar sunmuşlardır. Bruce D. Burns, 2004 yılındaki çalışmasında sıcak eli "kumarbaz yanılgısı" ile

ilişkilendirmiş, insanların rastgele serilerdeki düzen arayışı ile bağdaştırmıştır [11]. Hilesiz bir bozuk para atışı beş kez tekrarlanmış ve tümü yazı gelmiş ise altıncı atışın da yazı gelmesi birçok insanı şaşırtabilir fakat aslında altıncı atışın yazı gelmesi %50'lik bir ihtimalin sonucudur. Burns'e göre sıcak el gibi daha karmaşık problemlerde insanlar bu gibi istatistiksel gerçekleri göz ardı etmeye eğilimlidir. Daha yeni araştırmalar, sıcak el fenomeninin oluşturduğu algıyı ve oyuncuların davranışlarını nasıl etkilediğini incelemiştir. Örneğin, bazı NBA oyuncularının, önceki atışlarının sonucuna bağlı olarak davranışlarında değişiklikler sergilediği bulunmuştur. Bir çalışmada, oyun içi atışlarda sıcak el etkisine dair kanıt bulunmamakla birlikte, serbest atış denemelerinde zayıf bir sıcak el etkisi gözlemlenmiştir [53].

Sıcak el ve kumarbaz yanılsaması algıları geçmişte ekonomi araştırmalarına konu olmuş, davranışsal ekonomide yatırımcıların çoğu zaman gerçek olmayan kalıpları görerek kötü kararlara sürüklendiklerinin kanıtı olarak ileri sürülmüştür [15]. "*The Gambler's and Hot-Hand Fallacies: Theory and Applications*" makalesi sıcak el yanılgısının teorisini ve potansiyel finansal uygulamalarını ortaya koymaktadır [19].

Tüm bu çalışmalara karşın gelişmiş analiz metotlarıyla son on yılda yapılan araştırmalar sıcak elin varlığına dair bazı bulgular sunmuştur.

Yaari ve Eisenmann tarafından Ekim 2011'de yayınlanan bir makalede, 300.000'den fazla NBA serbest atışından oluşan geniş bir veri kümesinin, bireysel düzeyde "sıcak el" fenomeni için "güçlü kanıtlar" gösterdiği bulundu. 2005'ten 2010'a kadar beş normal NBA sezonu boyunca kullanılan tüm serbest atışların analiz edilmesinin sonucu, iki atışlık bir seride oyuncuların ikinci şutu atma olasılıklarında birinciye kıyasla önemli bir artış olduğu bulunmuştur. Ayrıca, art arda iki atıştan oluşan bir sette, ikinci atışta isabet olasılığının, bir önceki atışın isabetli olması halinde daha yüksek olduğu da tespit edilmiştir [20].

Kasım 2013'te Stanford Üniversitesi'ndeki araştırmacılar MLB'den elde edilen verileri kullandılar ve beyzbolda sıcak elin on farklı istatistiksel kategoride var olduğuna dair "güçlü kanıtlar" olduğunu buldular [36].

2015 yılında, Joshua Miller ve Adam Sanjurjo tarafından 1985 yılında yapılan çalışmanın incelenmesi, 1985 çalışmasının metodolojisinde kusurlar buldu

ve aslında sıcak ellerin var olabileceğini gösterdi. Araştırmacılar bunun istatistiksel tekniklerin yanlış uygulanmasına atfedilebileceğini söyledi. Yazarlar, insanların basketbolda sıcak ellerin var olduğuna inanmakta haklı olduklarını savundular [37].

1986-2020 dönemindeki NBA Üç Sayı Yarışmalarından elde edilen verileri kullanan 2021 tarihli bir araştırma, "oyuncular arasında sıcak el fenomeninin varlığına dair önemli kanıtlar" buldu [57].

2014 yılında, üç Harvard mezununun Sloan Sports Analytics Konferansı'nda sunduğu ve ilk kez basketbol oyunlarında oyuncunun şut yeri ve defans oyuncusunun pozisyonu gibi değişkenleri kontrol edebilen ileri istatistikler kullanan bir makale, "küçük ama anlamlı" bir sıcak el etkisi gösterdi [26].

Ancak yakın zamanda yapılan diğer çalışmalarda sıcak el varlığına dair bir kanıt gözlemlenmedi [53]. Üstelik genel olarak elde edilen kanıtlar, oyuncuların yalnızca küçük bir alt kümesinin "sıcak el" gösterebileceğini ve bunun da etki büyüklüğünün küçük olma eğiliminde olduğunu göstermektedir [63].

Sonuç olarak, sıcak el fenomeni üzerine yapılan araştırmalar karışık sonuçlar vermektedir. Bazı çalışmalar bu olguyu desteklerken, diğerleri sıcak elin aslında bir yanılsama olduğunu öne sürmektedir. Basketboldaki genel yargı, seyircinin algıladığı etki düzeyindeki sıcak el etkisinin yanılsama olduğu fakat özellikle serbest atış veya üçlük yarışması gibi sabit şartlarda seri olarak yapılan atışlarda az miktarda sıcak el etkisinin var olduğu yönünde. Sıcak el tartışmaları, sporda veri analitiğinin karmaşık yüzünü ve psikoloji gibi farklı disiplinlerle ilişkisini ortaya koymaktadır.

1.2.4 Futbol ve Veri Analitiği

Futbolda veri analizinin kökeni, 1950'li yıllara kadar uzanmaktadır. Charles Reep, 1956 yılında Swindon Town ve Bristol Rovers arasındaki bir maçta, Swindon adına etkili çözüm önerileri üretmek için bir takım sayısal verileri not tutmaya başladı. Bu maçta, Swindon'ın toplam 147 hücum girişiminde bulunduğunu ve sadece tek bir gol atabildiğini gözlemledi. Reep, hücumların gelişimindeki pas sayısı ile gol şansı arasındaki negatif korelasyonu fark etti ve uzun pasların skora etkisinin daha olumlu yönde olduğunu belirledi. Reep, sürekli olarak maçlarda istatistiksel notlar tutarak profesyonel futbolun ilk performans analisti oldu.

Brentford'un menajeri Jackie Gibbons, Reep'i danışman olarak işe alarak takımı İkinci Lig'de tutmayı başardı. Reep'in analizleri sayesinde Brentford, maç başına gol oranını ikiye katladı [9]. Reep'in uzun top stratejisi, 1983'te İngiliz Futbol Federasyonu'nun eğitim ve koçluk direktörü olarak atanan Charles Hughes tarafından benimsendi. Hughes, "*The Winning Formula*" kitabında gollerin %85'inin beş veya daha az pasla geldiğini ve topu mümkün olan en kısa sürede "en yüksek şans bölgesi" olarak tanımladığı bölgeye taşımak gerektiğini savundu [46].

Reep ve Hughes'un yöntemleri, Graham Taylor tarafından da uygulandı. Taylor, Watford'u Division One'da ikinci sıraya taşıdı ve 1990'da İngiliz milli takımının baş antrenörü oldu. Ancak İngiltere'nin 1994 Dünya Kupası'na katılamaması üzerine Taylor istifa etti [9].

Reed'in anlayışı bu gelişmelerle geçerliliğini yitirmeye başlasa da Opta'nın 1996'da kurulmasıyla, futbol veri analitiği alanında büyük bir adım atıldı. Opta, maç istatistiklerini toplayarak analiz eden kapsamlı bir sistem geliştirdi ve bu verileri kulüplere, medyaya ve taraftarlara sundu [27].

İngiltere'de bu dönem teknik sorumlularının veri analitiğinin değerinin farkına varmaya başladığı bilinmektedir. 2001 yılında Alex Ferguson, defans oyuncusu Jaap Stam'ı aniden Lazio'ya satması basında büyük ölçüde istatistiklere dayalı ilk büyük transfer olarak yer almıştır. Hollandalı futbolcu Dennis Bergkamp, otobiyografisinde Fransız teknik adam Arsene Wenger ile, Wnger'in sayılara olan "takıntısından" dolayı aralarında geçen tartışmalardan bahsetmiştir [25].

Günümüzdeki gelişmiş olanaklar sayesinde sporda işlenebilir veriyi üretmek daha kolay hale gelmiştir. Dolayısıyla veriyi doğru şekilde kullanarak sakatlık tahminlerinden scouting'e, performans analizine kadar geniş bir yelpazedeki analizlerin yararlı sonuçlar üretebilmesi için doğru analiz metotlarının kullanılmasını kritik faktör haline getirmektedir. Ayrıca maçlardan elde edilen geleneksel kantitatif veriler de ısı haritaları gibi gelişmiş görselleştirme metotlarıyla eskisinden daha yararlı çıktılar sunmaktadır. Bu metotlar, sporcuların ve antrenörlerin oyun stratejilerini daha iyi anlamalarına ve geliştirmelerine olanak tanır. Tüm bu gelişmelerin yanı sıra, son yıllardaki teknolojik gelişmelerin bir meyvesi olarak ortaya çıkan "gol beklentisi" (xG) metriği, futbolda veri analitiği alanında çığır açmış ve kullanılmaya başladığı ilk andan itibaren performans

analizindeki en değerli ölçülerden biri haline gelmiştir. xG metriğinden bir sonraki bölümde daha detaylı olarak bahsedilecektir.

1.2.5 Yapay Zekâ Tabanlı Metrikler

Yapay zeka, makinelerin verimli çalışmasına ve karmaşık verileri analiz etmesine olanak tanıyan bir bilgisayar bilimi dalıdır [64]. Makine öğrenmesi ve derin öğrenme gibi alt dalları, Yapay Zeka'nın gelişmiş yeteneklerini destekleyen teknolojilerdir.

Spordaki yapay zekâ tabanlı metrikler de bu metodolojiler sayesinde ortaya çıkmıştır. Büyük miktarda veri kullanarak, önceki yüzbinlerce pozisyon kaydının sonuçlarına ve bu sonuçları etkileyen faktörlere dayanarak pozisyona dair herhangi bir sonucun gerçekleşme olasılığı öngörülebilmektedir.

Gol beklentisi (xG), bir futbol maçında atılan belirli özelliklere sahip bir şutun gol olma olasılığını veren bir metriktir. Şutun yapıldığı mesafe, açı, şut hızı ve diğer faktörler dikkate alınarak hesaplanır. Bu metrik, takımların ve oyuncuların performansını objektif bir şekilde değerlendirmek için farklı bir bakış açısı sunar [61].

Gol beklentisi metriğinden harektele, xA (asist beklentisi) npxG (penaltısız gol beklentisi) gibi farklı metrikler de türetilebilmektedir. Uygulama bölümünde, performans analizleri ve maç senaryoları üzerine tahminler yapmak üzere bu metriklerden faydalanılacaktır.

2
MAKİNE ÖĞRENMESİ

2.1 Makine Öğrenmesi Nedir?

Makine öğrenmesi, bilgisayar bilimi ile istatistiğin kesiştiği noktada, yapay zekâ ve veri biliminin merkezinde yer alan, günümüzün en hızlı büyüyen teknik alanlarından birisidir [31]. Genel bir tanım olarak makine öğrenmesi, bilgisayara bir işlem için özel olarak kodlanmadan bu işlemi yapabilme becerisinin kazandırılmasını sağlayan çalışma alanıdır [1]. 1959'da yapılmış bu tanım ile makine öğrenmesi kavramı ilk kez ortaya çıkmış olsa da bu kavram, son yıllarda önemli ölçüde anlam kazanmıştır. Bu durum, yeni öğrenme algoritmalarının ve teorilerinin geliştirilmesi ile çevrimiçi verilerin ve düşük maliyetli hesaplama gücünün artışıyla desteklenmiştir [31]. Özellikle derin öğrenme ve büyük veri analitiğindeki ilerlemeler, makine öğrenmesinin popülerliğini artırmıştır [41].

Makine öğrenmesi, analitik model oluşturma sürecini otomatikleştirir ve ilgilenilen probleme özel "train" (eğitim/öğrenme) verilerine dayalı olarak istenen sonucu ortaya koyar [59].

Makine öğrenmesi, makine tarafından okunabilen verilerden öğrenmek için algoritmalar kullanır. Bunlar denetimli (supervised) denetimsiz (unsupervised), yarı denetimli (semi-supervised) ve pekiştirmeli öğrenme (reinforcement learning) olmak üzere 4 ana başlık altında incelenir [48].

2.2 Denetimli Makine Öğrenmesi

Makine öğrenmesinin bilgisayarın elindeki veriden anlamlar çıkararak sonuca ulaşmasını sağlayan çalışma alanı olduğundan önceki bölümde bahsedilmiştir. Denetimli öğrenmede, algoritmanın eğitildiği veri, 'etiket' olarak adlandırılan beklenen sonuçları da içerir [48]. Bu etiketlerin diğer değişkenler ile arasındaki ilişkilerden faydalanılarak, bilinmeyen etiketler üzerine tahminler yapılır.

İki adet tipik denetimli öğrenme işlevi vardır; sınıflandırma ve regresyon. Her iki işlev için farklı algoritmalar mevcuttur. Bazı durumlarda regresyon algoritmaları sınıflandırma için veya tam tersi şeklinde kullanılabilir [48].

Mail kutunuzdaki e-postaların spam olup olmadığının tespiti popüler bir sınıflandırma problemi örneğidir. E-posta sağlayıcınız bu işlemi spam ve spam olmayan mailleri içeren devasa boyutta bir veri kümesinden öğrenen bir sınıflandırma algoritması sayesinde yapar. Burada veri mailleriniz, etiket ise onların spam olup olmama durumudur.

Yüzlerce sporcunun piyasa değeri, yaşı, mevkii, maç istatistikleri gibi özelliklerinin bulunduğu bir veri setiyle eğitilerek ilgilenilen bir oyuncunun piyasa değerini tahmin edebilen bir algoritma ise regresyon algoritmalarına bir örnek teşkil etmektedir. Buradaki etiket de her bir oyuncunun piyasa değeridir.

Denetimli makine öğrenmesi bir matematik denklemine benzetilebilir. Girdiler ve elde edilen bir sonuç vardır. Denetimli makine öğrenmesinde farklı olarak bu sonuç, hesaplanarak bulunmaz, bunun yerine daha önce benzer girdiler ile elde edilmiş sonuçlardan yola çıkarak tahmin edilir. Sınıflandırmada eğitim verisinde belirli birkaç kategoriye ayrılan sonuç çeşitleri vardır. Amaç, bilinmeyen ve tahmin edilmeye çalışılan sonuçların bu farklı kategorilerden hangisi olduğunu tahmin etmektir. Regresyonda ise olası sayısal sonuçların sayısı sonsuzdur ve amaç tahmin edilmeye çalışılan sonuçların doğruya en yakın şekilde tahmin edilmesidir.

Denetimli öğrenmede tahmin etmeye çalışılan sonuç değişkeni ‘bağımlı değişken’, tahmin etmede kullanılan değişkenler ise ‘bağımsız değişken’ olarak adlandırılır.

Bu çalışmanın uygulama bölümünde de denetimli makine öğrenmesi algoritmaları kullanılmıştır. En çok kullanılan denetimli makine öğrenmesi algoritmalarının bazılarının tanımları alt başlıklarda verilmiştir.

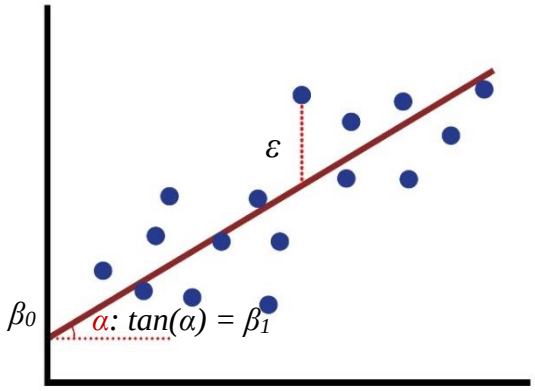
2.2.1 Lineer Regresyon

Lineer regresyon modeli, bağımlı değişken Y ile bir veya daha fazla bağımsız değişken (X) arasındaki doğrusal ilişkiyi tanımlar. Model genellikle $Y=\beta0+ \beta1X+ \varepsilon$ şeklinde ifade edilir [24].

$\beta0$: Sabit (slope), $\beta1$: Bağımsız değişken katsayısı (intercept) , ε: Hata (error)

Parametrelerin tahmini genellikle en küçük kareler yöntemi (Ordinary Least Squares - OLS) ile yapılır. Bu model, hataların karelerinin toplamını en aza indirerek doğrusal ilişkiyi belirler.

Şekil 2.1'de bahsedilen terimlerin bir basit lineer regresyon modeli üzerinde gösterimi mevcuttur. Mavi noktalar gözlemleri temsil eder. Kırmızı doğru ise regresyon modelini temsil eder. Bu doğru, bağımsız değişken ve bağımlı değişken arasındaki ilişkiyi en iyi şekilde tanımlayan doğrusal modeldir.



Şekil 2.1

Lineer regresyon, özellikle küçük veri setleri veya düşük sinyal-gürültü oranı olan durumlarda, daha karmaşık modellerden daha iyi performans gösterebilir. Ayrıca, veri dönüşümleriyle genişletilerek daha esnek hale getirilebilir [17].

2.2.2 Lojistik Regresyon

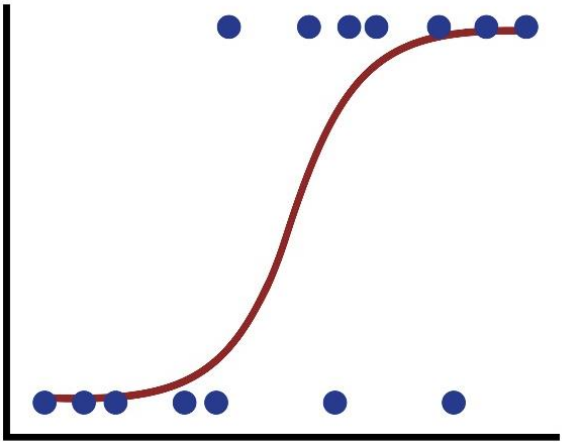
Lojistik regresyon, her bağımsız değişkenin benzersiz katkısını ölçerek bir grup bağımsız değişkenin ikili sonuç üzerindeki etkisini analiz etmenin etkili bir yoludur [21].

Bu yöntem, odds oranını, yani bir olayın gerçekleşme olasılığının gerçekleşmeme olasılığına oranını elde etmek için kullanılır ve birden fazla açıklayıcı değişkenin varlığında bile bu oranları belirleyebilir [28].

Maksimum likelihood yöntemi kullanılarak katsayılar tahmin edilir ve modelin uyumunun iyi olup olmadığı çeşitli iyilik-uyum ölçümleri ile değerlendirilir [21].

Şekil 2.2'de yalnızca iki değer alan gözlemlerin dağılımı ve lojistik regresyonun sigmoid fonksiyonu şeklindeki modeli görülmektedir. Sigmoid

fonksiyonunun y ekseninde 0 ile 1 arasında aldığı değer, ilgilenilen değerin üst gruba ait olma olasılığını gösterir.



Şekil 2.2

Lojistik regresyon, spor analitiğinde maç sonucu tahminleme veya gol/gol değil değişkenlerini bağımlı değişken olarak kullanarak gol beklentisi (xG) oranını elde etme gibi problemlerde kullanılmaktadır [65].

2.2.3 Destek Vektör Makinesi

Destek Vektör Makinesi (SVM), iki grup arasındaki sınıflandırma problemlerinde kullanılan bir algoritmadır. SVM, giriş vektörlerini doğrusal olmayan bir şekilde çok yüksek boyutlu bir özellik uzayına haritalar ve bu özellik uzayında doğrusal bir karar yüzeyi oluşturur. Bu yüzeyin özel özellikleri, algoritmanın yüksek genelleme yeteneğine sahip olmasını sağlar [4].

SVM'nin amacı, iki sınıfı ayıran ve maksimum marjini (iki sınıf arasındaki en geniş mesafeyi) sağlayan hiper düzlemi bulmaktır. Bu hiper düzlem, eğitim verileri tarafından belirlenen destek vektörlerine dayanır. Destek vektörleri, karar sınırını en yakın noktalar olarak belirler [12].

SVM, çekirdek fonksiyonları kullanarak doğrusal olmayan verileri doğrusal ayıran bir hiper düzleme dönüştürür. Çekirdek fonksiyonları, verileri daha yüksek boyutlu bir uzaya yansıtarak doğrusal olmayan sınırları doğrusal hale getirir. Yaygın kullanılan çekirdek fonksiyonları arasında polinomial, radyal tabanlı fonksiyon (RBF) ve sigmoid bulunur [13].

Karush-Kuhn-Tucker (KKT) Koşulları: SVM'nin uygulanmasında KKT koşulları önemli bir rol oynar. Bu koşullar, optimizasyon sürecinde çözümün bulunmasında kritik bir yere sahiptir ve diğer desen tanıma algoritmalarından farklı olarak, SVM algoritmalarını daha verimli hale getirir [38].

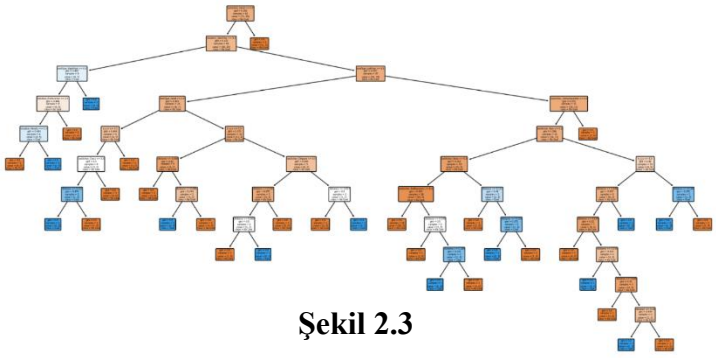
SVM'lerin, özellik uzayında doğrusal karar yüzeyleri oluşturma yeteneği, genelleme kabiliyetlerini artırır. Bu, özellikle doğrusal olarak ayrılabilir olmayan veriler için geçerlidir [4]. SVM regresyonu, eğitim verilerinden analitik parametre seçimi kullanılarak iyi bir genelleme performansı ile regresyonda da kullanılabilir [14].

2.2.4 Karar Ağaçları ve Rastgele Ormanlar

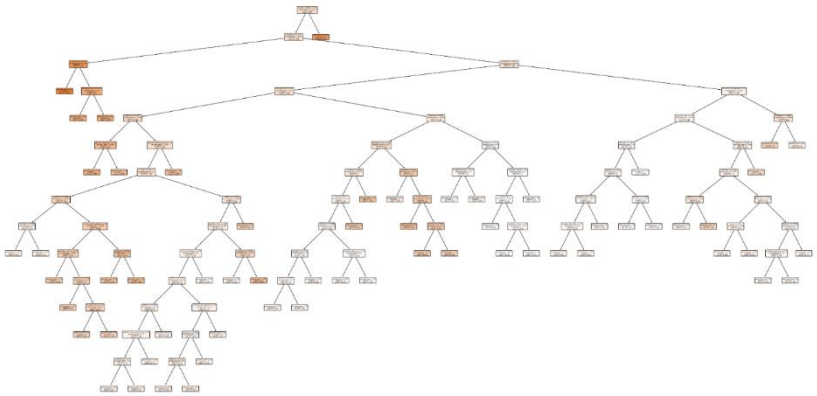
Karar ağaçları, ortak değişkenin alanını alt uzaylara bölen, her bir alt uzayı farklı bir tahmin fonksiyonu için kullanan tahmine dayalı modellerdir ve hem sınıflandırma hem regresyon görevleri için kullanılabilir. [39]. Sınıflandırma ağaçları belirli sayıda sırasız değerleri tahmin etmek için tasarlanırken, regresyon ağaçları sürekli veya sıralı ayrık değerler içindir; tahmin hatası, gözlemlenen ve tahmin edilen değerler arasındaki kare farkla ölçülür [22].

Rastgele ormanlar, her ağacın bağımsız olarak örneklenen ve ormandaki tüm ağaçlar için aynı dağılıma sahip rastgele bir vektöre bağlı olduğu ağaç tahminlerinin bir birleşimidir [8]. Farklı ağaçlardan elde edilen tahminlerin ortalamasını alarak bir araya getiren genel amaçlı bir sınıflandırma ve regresyon yöntemidir [33].

Şekil 2.3 ve 2.4'de İngiltere Premier Ligi 2023-2024 sezonunda En fazla gol atan oyuncu Erling Haland'ın sezon boyunca attığı şutların verisi kullanılarak, sırasıyla sınıflandırma ile gol/gol değil tahmini ve regresyon ile gol beklentisi tahmini yapan rastgele orman modelleri görselleştirilmiştir.



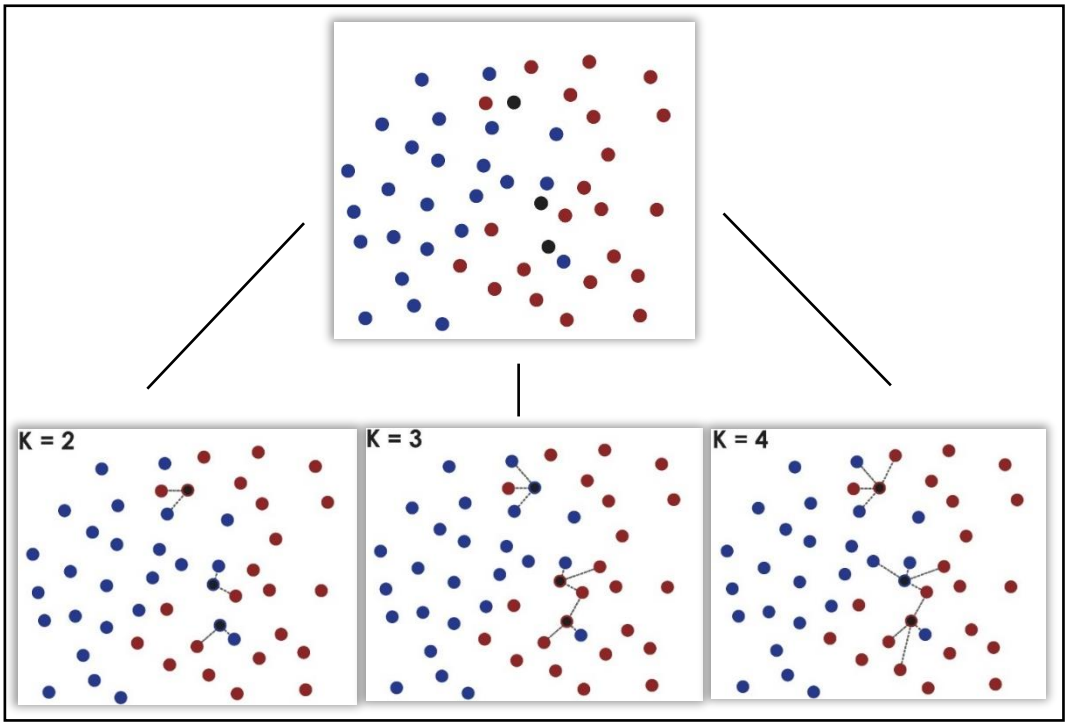
Şekil 2.3



Şekil 2.4

2.2.5 K-En Yakın Komşu

K En Yakın Komşu (kNN) yöntemi, veri madenciliği ve makine öğrenimi uygulamalarında sınıflandırma, regresyon ve eksik veri atama için kullanılan makine öğreniminin basit bir uygulamasıdır [42]. Burada her bir boyut, bağımlı değişkeni tahmin etmek için kullanılan bir bağımsız değişkeni ifade eder. Bağımlı değişken 'k' adet komşusu arasındaki sınıfların yoğunluğuna ve yakınlığına göre sınıflandırılır. Aşağıdaki grafikte iki boyutlu bir düzlemde, iki farklı sınıfa ait değerler ve üç adet sınıfı belli olmayan değer görselleştirilmiştir. Farklı 'K' değerlerine göre bilinmeyen değerlerin sınıflandırılması aşağıda temsili olarak görselleştirilmiştir.



Şekil 2.5

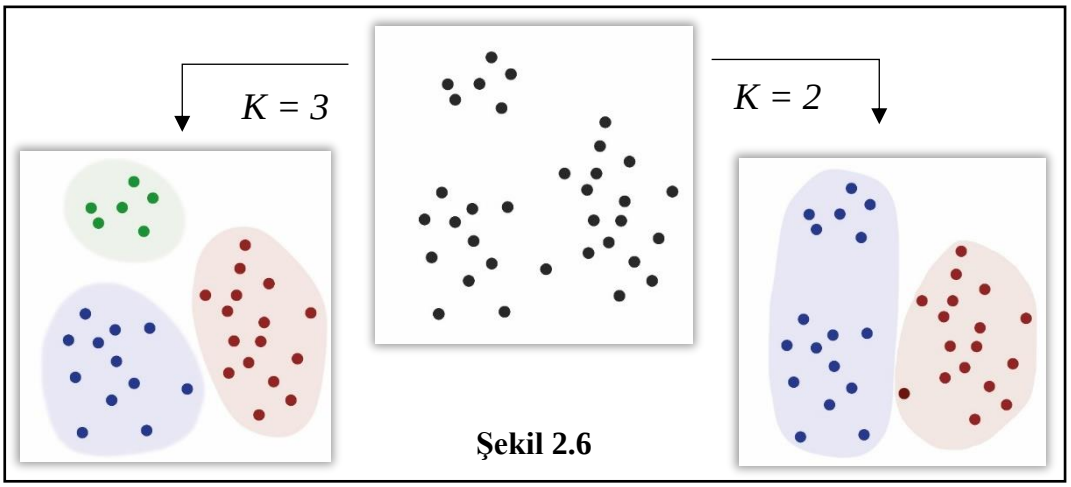
2.3 Denetimsiz Makine Öğrenmesi

Denetimsiz makine öğrenmesinde eğitim verileri etiketsizdir. Sistem belirli bir öğreten olmadan öğrenmeye çalışır [48]. Denetimsiz öğrenmenin kümeleme, anomali tespiti gibi çeşitli işlevleri vardır. Denetimsiz makine öğrenmesi, veri kümelerindeki gizli kalıpları ve grupları tanımlamak için kullanılan kümeleme ve ilişkilendirme kurallarını içerir [43]. Bu çalışmada denetimsiz öğrenme algoritmaları kullanılmamış olsa da çeşitli konularda denetimsiz makine öğrenmesi algoritmaları kullanılmaktadır [60]. Bir sonraki bölümde denetimsiz makine öğrenmesi algoritmalarına örnek olarak K-Means algoritması eke alınacaktır.

2.3.1 K-Ortalamalar

K-ortalamalar algoritması, benzer verileri aynı kümede sınıflandırmak için kullanılan basit bir kümeleme yöntemidir [44]. K-ortalamalar algoritması, deterministik bir küresel arama prosedürü yoluyla her seferinde bir küme merkezini dinamik olarak ekleyen, kümelemeye yönelik bir yaklaşımdır [23]. Belirlenen 'k' sayıda küme için birer merkez belirlenir, daha sonra her bir gözlem en yakın olduğu merkeze göre bir kümeye atanır. Yakınlık türleri çeşitlilik gösterebilir. Kümelemenin ardından yeni oluşan kümeler için yeni merkez noktası belirlenir. Sonrasında yeni merkez noktalarına göre kümeleme işlemi tekrar edilir. Bu işlem aynı sonuç elde edilene kadar tekrarlı biçimde devam eder ve eğitim setinde hakkında herhangi bir bilgi bulunmayan yeni kümeler oluşmuş olur [58].

Şekil 2.6'da bir veri grubunun K-ortalamalar algoritması ile Öklid uzaklığı kullanılarak 2 ve 3 k değerleri ile iki farklı şekilde kümelenmeleri temsili olarak görselleştirilmiştir.



Şekil 2.6

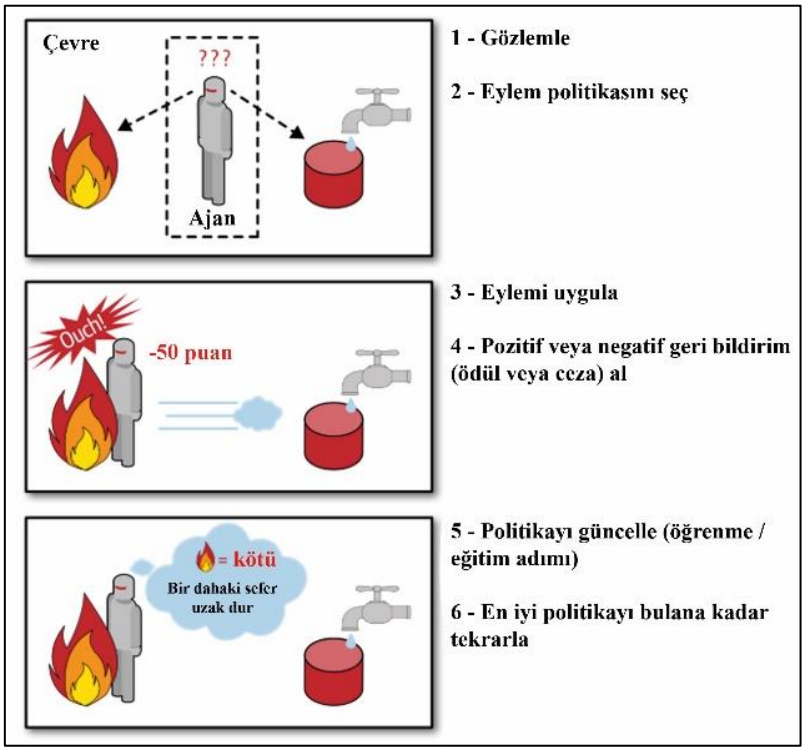
2.4 Yarı Denetimli Makine Öğrenmesi

Yarı denetimli öğrenme, etiketli verilerin az sayıda veya pahalı olduğu durumlarda denetimli öğrenme görevlerini geliştirmek için etiketli ve etiketsiz verilerin birlikte kullanıldığı makine öğrenmesi türüdür [18].

2.5 Pekiştirmeli Öğrenme

Yapay zekanın en aktif araştırma alanlarından biri olan pekiştirmeli öğrenme, ajan (agent) olarak tabir edilen makinanın karmaşık, belirsiz bir ortamla etkileşime girdiğinde aldığı toplam ödül miktarını en üst düzeye çıkarmaya çalıştığı, öğrenmeye yönelik hesaplamalı bir yaklaşımdır [6].

Pekiştirmeli öğrenmede bir yazılım ajanı çevresi ile etkileşime girer, eylemde bulunur ve yaptığı her eylem sonucunda negatif veya pozitif sonuçlar alır. Pozitif sonuçlar 'ödül' olarak, negatifler 'ceza' olarak adlandırılır. Ajan deneyimlerindeki eylemleri ve sonucunda elde ettiği ödüllerin verisini kaydeder ve aslında bir bakıma kendi eğitim verisini oluşturmuş olur. Sonrasında bu veriyi kullanarak ödül miktarını en optimal düzeye getirecek şekilde makine öğrenmesine benzer algoritmalar ile bir eylem planı (politika) oluşturur. Bu süreç şekil 2.7'de görsel olarak temsil edilmiştir.



Şekil 2.7

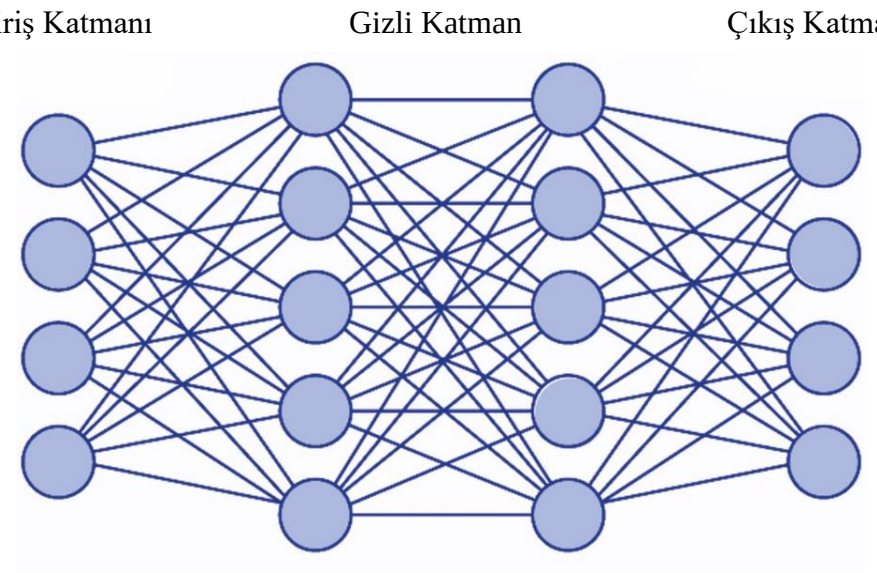
2017 yılında, o zamana kadar bilinen en başarılı satranç yapay zekâsı Stockfish ile ilk kez piyasaya sürülen AlphaZero karşı karşıya geldi. Stockfish, geçmişte insanların oynadığı binlerce satranç maçının verisi ile eğitilirken, AlphaZero pekiştirmeli öğrenme ile hiçbir satranç maçını görmeden, yalnızca birkaç saat içerisinde kendisine karşı tekrarlı maçlar yapıp elde ettiği ödül ve cezaları optimize edecek algoritmaları oluşturarak kendine hamle politikaları belirleyen bir pekiştirmeli öğrenme ajanıydı. İki yapay zekâ arasında oynanan maçların %80'den fazlası beraberlikle sonuçlanırken, geri kalan kısmın neredeyse tamamı AlphaZero zaferiyle sonuçlanmıştır.

2.6 Derin Öğrenme ve Yapay Sinir Ağları

Derin öğrenme, yüksek boyutlu verilerdeki görüntü tanıma ve konuşma tanıma gibi karmaşık işlevleri otomatik olarak keşfetmek için birden fazla temsil düzeyini kullanan bir makine öğrenmesi teknikleri sınıfıdır [34].

Yapay sinir ağları, giriş verilerindeki ilişkileri tanımlayabilen ve yeni sunulan veri kümelerindeki ilişkileri tahmin edebilen, biyolojik bir sinir sistemini örnek alarak modellenen, bilgi işlem öğelerinin dağıtılmış bir ağıdır [5].

Yapay sinir ağları, genellikle bir giriş katmanı, bir veya daha fazla gizli katman ve bir çıkış katmanından oluşur. Zaman içinde doğruluğu artırmak için ağırlıkların hata oranlarına göre ayarlanmasını içeren geri yayılımı eğitim için kullanırlar. Giriş katmanı, ham veriyi (örneğin, piksel değerleri, metin verisi) ağın geri kalanına ileten katmandır. Bu katman sadece veriyi alır ve bir sonraki katmana aktarır. Gizli katmanlar, veriyi işleyerek daha yüksek seviyede özellikler çıkarır. Bu katmanlar, non-lineer fonksiyonlar kullanarak girdiyi işleyip önemli bilgileri çıkarır. Bir yapay sinir ağında birden fazla gizli katman olabilir ve her katman belirli bir seviyede özellik çıkarımı yapar [32]. Çıkış katmanı, ağın son çıktısını üreten katmandır. Bu katman, gizli katmanlardan gelen bilgiyi alır ve belirli bir karar veya sınıflandırma yapar [48]. Yapay sinir ağlarındaki katman yapısını temsil eden şema, şekil 2.8'de mevcuttur.



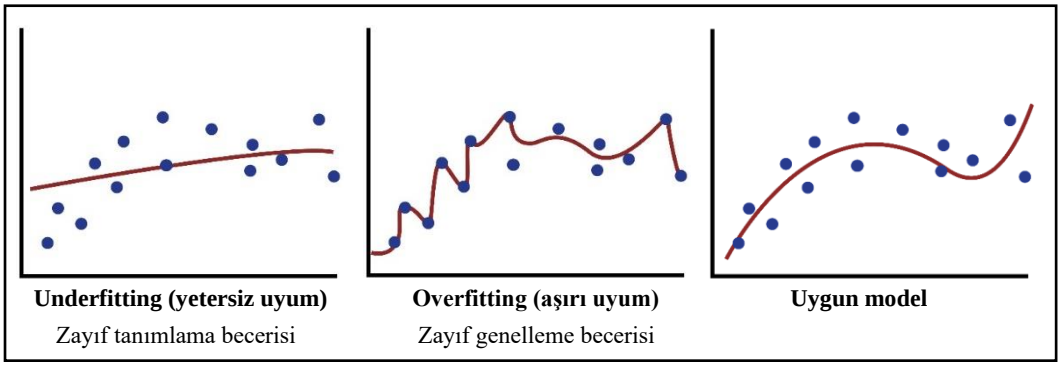
Şekil 2.8

Yapay sinir ağlarında genellikle eğitim için kapsamlı verilere ve hesaplama kaynaklarına ihtiyaç duyar. Milyonlarca parametre içerebilen derin öğrenme gibi karmaşık mimarileri ve eğitim süreçlerini içerirler. Bu karmaşıklık aşırı uyum gibi zorluklara yol açabilir, ancak bilgi işlem gücündeki ilerlemeler ve bırakma gibi teknikler bu sorunları hafifletmiştir [29].

2.7 Aşırı Uyum Durumu

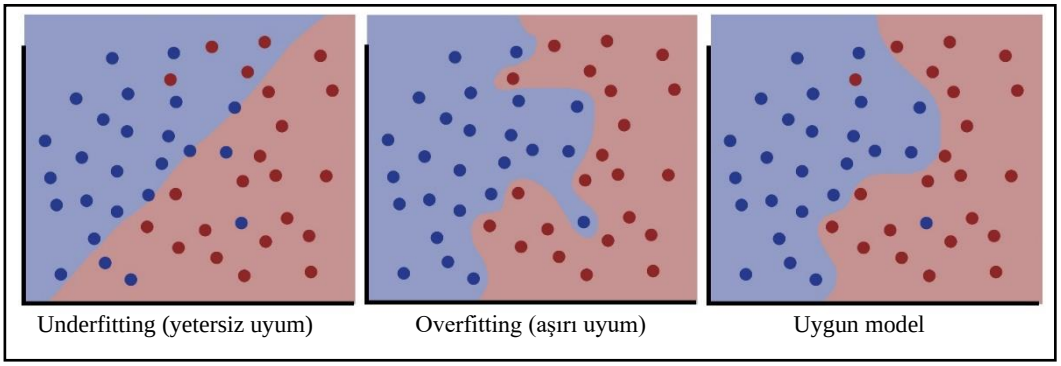
Genelleme, makine öğreniminde bir veya birkaç farklı ancak ilgili alandan öğrenebilecek modeller geliştirmek için temel bir gerekliliktir. Aşırı uyum, denetimli makine öğreniminde gürültü, sınırlı eğitim seti boyutu ve sınıflandırıcı karmaşıklığı nedeniyle modellerin gözlemlenen verilere ve görünmeyen verilere uyacak şekilde genelleştirilmesini engelleyen temel bir sorundur [49].

Şekil 2.9'da bir regresyon modeli için üç farklı durumun temsili grafikler mevcuttur.



Şekil 2.9

Şekil 2.10'da ise bu üç durum bir sınıflandırma modeli için temsili grafikler mevcuttur.



Şekil 2.10

Aşırı uyumu tespit etmek ve modelin genelleme kabiliyetini test edebilmek için hold-out metodu kullanılır. Bu metot, eğitilmek istenen model için kullanılacak veri seti eğitim ve test olarak ikiye bölünür. Büyük bir kısım eğitim için ayrılır. Model eğitim verisiyle eğitilir ve sonrasında test verisi üzerindeki tahminlerinin doğruluğu ölçülür. Eğer test ile elde edilen doğruluk eğitim ile elde edilen doğruluğa göre düşük ise bu bir aşırı uyum işaretidir [30].

2.8 Gol Beklentisi (xG) Nasıl Hesaplanır?

Tüm bu makine öğrenmesi konseptlerine değindikten sonra, yapay zekâ tabanlı metriklerden gol beklentisinin nasıl hesaplandığından bahsedebiliriz. Gol beklentisi bir futbol takımının veya oyuncunun belirli bir pozisyon kümesinde golü bulma olasılığını tahmin 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 [61], [65], [66]. Şutun kaleye olan uzaklığı, kalenin pozisyonu ve savunma oyuncularının yerleşimi gibi değişkenler, gol olasılığını belirlemede kullanılır. Makine öğrenimi algoritmaları, geçmiş maç verilerinden öğrenerek bu olasılıkları hesaplar [61]. Farklı değişkenlerin kullanıldığı xG modelleri mevcuttur. xG modelleri, oyuncu ve takım yetenekleri gibi daha önce test edilmemiş özelliklerle geliştirilebilir [66]. Yani bir başka deyişle, gol beklentisi için 0.25 gibi bir değer belirlemiş olması, değeri belirleyen yapay zekâ modelini eğitmek üzere basitçe rastgele olarak seçilmiş, yapay zekânın bu şut ile aynı özelliklere sahip olduğunu tespit ettiği şutların çeyreğinin gol ile sonuçlandığı anlamına gelir.

3
UYGULAMA

3.1 Çalışma Tanımı

Bu bölümde, understat.com sitesinden alınmış İngiltere Premier Ligi 2023-2024 sezonuna ait veriler kullanılarak, ilgili internet sitesinde tanımlayıcı istatistikler ve veri görselleştirme için kullanılan bu verilerden, performans analizi için ne gibi çıkarımlar yapılabildiği incelenecek, xG istatistiği üzerine tartışılacak ve makine öğrenmesi modelleri kullanılarak olabildiğince yüksek güçte tahminler elde edilmeye çalışılacaktır.

3.1.1 Kullanılan Araçlar

Tüm çalışmada Python dili kullanılmıştır. Python içerisinde kullanılan bazı kütüphaneler ve modüllerin tanımları aşağıdaki gibidir:

1- **Numpy**: Lineer cebir hesaplamaları için kullanıldı.
2- **Pandas**: Veri analizi ve manipülasyonu için kullanıldı.
3- **Matplotlib**: Grafikler ve haritalar oluşturmak için kullanıldı.
4- **Seaborn**: Grafikler oluşturmak için kullanıldı.
5- **Scipy.stats**: İstatistiksel analizler için kullanıldı.
6- **Scikit-learn**: Makine öğrenmesi modelleri için kullanıldı.
7- **Math**: Temel matematiksel işlemler için kullanıldı
8- **Json**: Json formatındaki verileri işlemek için kullanıldı.
9- **Requests**: Web Scrapping işleminde http talepleri oluşturmak için kullanıldı.
10- **BeautifulSoup**: Web Scrapping işleminde HTML ve XML dosyalarını ayrıştırmak için kullanıldı.
11- **MplSoccer:** Futbol verilerini görselleştirmek için kullanıldı.
12- **GridSearchCV**: Makine Öğrenmesi modellerinin hiperparametre optimizasyonu için kullanıldı.

3.1.2 Ham Veriyi Elde Etme

Veri, BeautifulSoup kütüphanesi kullanılarak veri kazıma (Web Scrapping) metodu ile elde edilmiştir. Bu işlem, understat.com sitesindeki farklı script indekslerine talepler oluşturarak 'json' formatında elde edilen verileri uygun forma getirip bir 'Pandas DataFrame' objesine dönüştürülerek yapıldı.

3.1.3 Tablolar ve İçerikleri

Web Scrapping işleminin sonucunda undertat.com sitesinden 5 adet tablodan oluşan bir veri seti elde edilmiştir. Tablolar sırasıyla takım_maçlar, maçlar, oyuncu_perf, oyuncular ve şutlar olarak isimlendirilmiştir. Tablolardaki sayısal verilerin tanımlayıcı istatistikleri ve tüm verilerin tanımları her bir tablo için alt başlıklarda belirtilmiştir.

3.1.3.1 "oyuncular"

- Bu veri seti her bir oyuncu için bir kayıt tutar.
- Toplamda 570 kayıt ve 18 değişken vardır.
- Sayısal değişkenlerin tanımlayıcı istatistikleri tablo 3.1'deki gibidir:

	GAMES	TIME	GOALS	XG	ASSİSTS	XA	SHOTS	KEY PASSES	YELLOW CARDS	RED CARDS	NPG	NPXG	XG CHAİN	XG BUİLDUP
COUNT	570	570	570	570	570	570	570	570	570	570	570	570	570	570
MEAN	19,97	1318,6	2,1	2,29	1,5	1,66	18,38	13,75	2,79	0,1	1,93	2,15	6,64	4,08
STD	11,74	1036,22	3,64	3,72	2,37	2,35	23,08	17,96	2,78	0,32	3,21	3,32	6,83	4,37
MİN	1	1	0	0	0	0	0	0	0	0	0	0	0	0
25%	10	360	0	0,09	0	0,06	2	1	1	0	0	0,09	1,22	0,59
50%	21	1205,5	1	0,88	0	0,79	9	7	2	0	1	0,87	4,65	2,81
75%	31	2154	3	2,68	2	2,16	26	19,75	4	0	2	2,62	10,14	5,92
MAX	38	3420	27	31,65	13	13,34	122	113	13	2	20	25,56	31,88	26,29

Tablo 3.1

- Değişkenler ise aşağıdaki gibidir:

1. id: İlgili oyuncunun kodu

2. player_name: İlgili oyuncunun adı

3. games: İlgili oyuncunun çıktığı toplam maç sayısı

4. time: İlgili oyuncunun toplam aldığı süre

5. goals: İlgili oyuncunun attığı toplam gol sayısı

6. xG: İlgili oyuncunun toplam gol beklentisi

7. assists: İlgili oyuncunun toplam asist sayısı

8. xA: İlgili oyuncunun toplam asist beklentisi

9. shots: İlgili oyuncunun toplam şut sayısı

10. key_passes: İlgili oyuncunun toplam kilit pas sayısı

11. yellow_cards: İlgili oyuncunun toplam sarı kart sayısı

12. red_cards: İlgili oyuncunun toplam kırmızı kart sayısı

13. position: İlgili oyuncunun mevkii

14. team_title: İlgili oyuncunun takımı

15. npg: İlgili oyuncunun penaltılar hariç toplam gol sayısı

16. npxG: İlgili oyuncunun penaltılar hariç toplam gol beklentisi

17. xGChain: İlgili oyuncunun dahil olduğu tüm pozisyonlardaki toplam gol beklentisi

18. xGBuildup: İlgili kilit pas ve şutlar hariç oyuncunun dahil olduğu tüm pozisyonlardaki toplam gol beklentisi

3.1.3.2 “oyuncu_perf”

- Bu veri seti her bir oyuncunun sahaya çıktığı her bir maç için bir kayıt tutar.
- Toplamda 11384 kayıt ve 21 değişken vardır.
- Sayısal değişkenlerin tanımlayıcı istatistikleri 3.2’deki gibidir:

	GOALS	OWN GOALS	SHOTS	XG	TİME	YELLOW CARD	RED CARD	KEY PASSES	ASSİSTS	XA	XG CHAİN	XG BUİLDUP
COUNT	11384	11384	11384	11384	11384	11384	11384	11384	11384	11384	11384	11384
MEAN	0,11	0	0,92	0,11	66,02	0,14	0,01	0,69	0,08	0,08	0,33	0,2
STD	0,35	0,07	1,31	0,25	33,1	0,35	0,07	1,1	0,29	0,19	0,42	0,31
MİN	0	0	0	0	1	0	0	0	0	0	0	0
25%	0	0	0	0	38	0	0	0	0	0	0,03	0
50%	0	0	0	0	90	0	0	0	0	0	0,16	0,07
75%	0	0	1	0,1	90	0	0	1	0	0,07	0,5	0,29
MAX	4	1	12	2,93	90	1	1	9	4	1,78	3,64	2,89

Tablo 3.2

- Değişkenler ise aşağıdaki gibidir:

1. id: İlgili oyuncunun ilgili maça özgü kodu
2. goals: İlgili maçta ilgili oyuncunun attığı gol sayısı
3. own_goals: İlgili maçta ilgili oyuncunun kendi kalesine attığı gol sayısı
4. shots: İlgili maçta ilgili oyuncunun çektiği şut sayısı
5. xG: İlgili maçta ilgili oyuncunun gol beklentisi
6. time: İlgili maçta ilgili oyuncunun aldığı süre (dakika)
7. player_id: İlgili oyuncunun kodu
8. team_id: İlgili oyuncunun takımının kodu
9. position: İlgili maçta ilgili oyuncunun mevkii
10. player: İlgili oyuncunun adı
11. h_a: İlgili maçta ilgili oyuncunun tarafı (iç saha/deplasman)
12. yellow_card: İlgili maçta ilgili oyuncunun gördüğü sarı kart sayısı
13. red_card: İlgili maçta ilgili oyuncunun gördüğü kırmızı kart sayısı
14. roster_in: Çıkan oyuncu yerine giren oyuncunun giriş kaydının kodu
15. roster_out: Giren oyuncu yerine çıkan oyuncunun çıkış kaydının kodu
16. key_passes: İlgili maçta ilgili oyuncunu attığı kilit pas sayısı
17. assists: İlgili maçta ilgili oyuncunun asist sayısı
18. xA: İlgili maçta ilgili oyuncunun ilgili maçtaki asist beklentisi

19. xGChain: İlgili maçta ilgili oyuncunun dahil olduğu pozisyonlardaki toplam gol beklentisi

20. xGBuildup: İlgili maçta ilgili kilit pas ve şutlar hariç oyuncunun dahil olduğu tüm pozisyonlardaki toplam gol beklentisi

21. positionOrder: İlgili maçta ilgili oyuncunun oynadığı pozisyonun sıralaması

3.1.3.3 "şutlar"

- Bu veri İngiltere Premier Lig 2023-2024 sezonu boyunca çekilmiş her bir şut için bir kayıt tutar.
- Toplamda 10524 kayıt ve 17 değişken vardır.
- Sayısal değişkenlerin tanımlayıcı istatistikleri tablo 3.3'deki gibidir:

	MİNUTE	X	Y	XG	PLAYER_İD	SEASON	H_GOALS	A_GOALS
COUNT	10524	10524	10524	10524	10524	10524	10524	10524
MEAN	49,2	0,86	0,5	0,12	6806,28	2023	1,84	1,51
STD	27,03	0,09	0,12	0,17	3672,31	0	1,36	1,28
MİN	0	0	0,03	0	65	2023	0	0
25%	26	0,8	0,42	0,03	4105	2023	1	1
50%	49	0,87	0,5	0,06	7322	2023	2	1
75%	72	0,92	0,59	0,11	9948	2023	3	2
MAX	105	1	1	0,97	12549	2023	6	8

Tablo 3.3

- Değişkenler ise aşağıdaki gibidir:

1. id: Şutun kodu

2. minute: Şutun çekildiği dakika

3. result: Şutun sonucu (kurtarış, kaçış, gol vb...)

4. xG: Şutun gol beklentisi

5. player: Şutu çeken oyuncunun adı

6. h_a: Şutu çeken taraf (iç saha/deplasman)

7. player_id: Şutu çeken oyuncunun kodu

8. situation: Şutun çekildiği durum (Açık oyun, serbest vuruş vb…)

9. shotType: Şutun çeşidi (kafa, sağ ayak...)

10. match_id: Şutun çekildiği maçın kodu

11. h_team: İç saha takımı

12. a_team: Deplasman takımı

13. date: Tarih

14. player_assisted: Şut pasını veren oyuncu

15. lastAction: Son aktivite (kafa pası, top çalma...)

16. X: Şutun çekildiği bölgenin sahadaki X koordinatı

17. Y: Şutun çekildiği bölgenin sahadaki Y koordinatı

3.1.3.4 "maçlar"

- Bu veri seti oynanan her bir maç için ayrı bir kayıt tutar.
- Toplamda 380 kayıt ve 12 değişken vardır.
- Sayısal değişkenlerin tanımlayıcı istatistikleri aşağıdaki gibidir:

	GOALS.H	GOALS.A	XG.H	XG.A
COUNT	380	380	380	380
MEAN	1,8	1,48	1,92	1,46
STD	1,37	1,28	1,03	0,9
MİN	0	0	0,09	0,03
25%	1	1	1,11	0,8
50%	2	1	1,76	1,29
75%	3	2	2,63	1,93
MAX	6	8	6,67	5,11

Tablo 3.4

- Değişkenler ise aşağıdaki gibidir:

1. id: Maçın kodu

2. datetime: Tarih

3. h.id: Ev sahibi takım kodu

4. h.title: Ev sahibi takım ismi

5. h.short_title: Ev sahibi takım kısaltması

6. a.id: Deplasman takım kodu

7. a.title: Deplasman takım ismi

8. a.short_title: Deplasman takım kısaltması

9. goals.h: Ev sahibi takım gol sayısı

10. goals.a: Deplasman takım gol sayısı

11. xG.h: Ev sahibi gol beklentisi

12. xG.a: Deplasman gol beklentisi

3.1.3.5 "takım_maçlar"

- Bu veri seti her bir takımın oynadığı 38 maç için ayrı bir kayıt tutar.
- Toplamda 760 kayıt ve 21 değişken vardır.

- Sayısal değişkenlerin tanımlayıcı istatistikleri aşağıdaki gibidir:

	XG	XGA	NP XG	NP XGA	DEEP	DEEP ALLOWED	SCORED	MİSSED	XPTS	WİNS	DRAWS	LOSES	PTS	NP XGD	OPPDA	PPDA
COUNT	760	760	760	760	760	760	760	760	760	760	760	760	760	760	760	760
MEAN	1,69	1,69	1,58	1,58	8,3	8,3	1,64	1,64	1,4	0,39	0,22	0,39	1,39	0	13,64	13,64
STD	1	1	0,91	0,91	5,52	5,52	1,33	1,33	0,88	0,49	0,41	0,49	1,35	1,47	11,07	11,07
MİN	0,03	0,03	0,03	0,03	0	0	0	0	0	0	0	0	0	-4,93	2,57	2,57
25%	0,93	0,93	0,9	0,9	4	4	1	1	0,65	0	0	0	0	-0,94	8,19	8,19
50%	1,51	1,51	1,42	1,42	7	7	1	1	1,35	0	0	0	1	0	11,57	11,57
75%	2,28	2,28	2,16	2,16	11	11	2	2	2,12	1	0	1	3	0,94	15,62	15,62
MAX	6,67	6,67	5,66	5,66	37	37	8	8	3	1	1	1	3	4,93	193	193

Tablo 3.5

- Değişkenler ise aşağıdaki gibidir:

1. h_a: İlgili maçta ilgili takımın tarafı (iç saha/deplasman)

2. xG: İlgili maçtaki ilgili takımın toplam gol beklentisi

3. xGA: İlgili maçtak rakip takımın gol beklentisi

4. npxG: İlgili maçta ilgili takımın penaltılar hariç gol beklentisi

5. npxGA: İlgili maçta rakip takımın penaltılar hariç gol beklentisi

6. deep: İlgili maçta ilgili takımın kaleye 20 metreden az mesafede yaptığı pas sayısı

7. deep_allowed: İlgili maçta rakibin kaleye 20 metreden az mesafede yaptığı pas sayısı

8. scored: İlgili maçta ilgili takımın attığı gol

9. missed: İlgili maçta ilgili takımın kaçırdığı gol

10. xpts: İlgili maçta ilgili takımın puan beklentisi

11. result: İlgili maçta ilgili takım adına sonuç (galibiyet / beraberlik / mağlubiyet)

12. date: Maçın oynandığı tarih

13. wins: İlgili takım galibiyet (0/1)

14. draws: İlgili takım beraberlik (0/1)

15. loses: İlgili takım mağlubiyet (0/1)

Şekil 3.1

16. pts: İlgili maçta ilgili takımın kazandığı puan

17. npxGD: İlgili maçta beklenen penaltısız gol farkı

18. team_id: İlgili takımın kodu

19. team_title: İlgili takımın adı

20. ppda: Rakip takımın birinci ve ikinci bölgesinde yaptığı pas sayısının, aynı bölgede yapılan defansif aksiyonlara oranı. Bu metrik takımların önde pres gücünü ölçmek için kullanılır, bu nedenle belirtilen bu oranın sonucunda çıkan sayı ne kadar küçükse, takımın önde pres gücü o kadar iyi olarak değerlendirilir

21. oppda: İlgili takımın birinci ve ikinci bölgede yaptığı pas sayısının, aynı bölgede rakibin yaptığı defansif aksiyonlara oranı. (Rakibin ppda değeri)

3.1.4 Veri İşleme ile Türetilmiş Veriler

‘şutlar’ tablosundaki ‘result’ değişkeni ‘Gol’ değerini aldığı satırlarda 1, diğer satırlarda 0 değerini alacak şekilde bir ‘isGoal’ değişkeni oluşturulmuştur.

Taraf bilgisi içeren tüm tablolardaki iç/dış saha değişkenleri iç saha olması durumunda 1, deplasman olması durumunda 0 değerini alacak şekilde değiştirilmiştir.

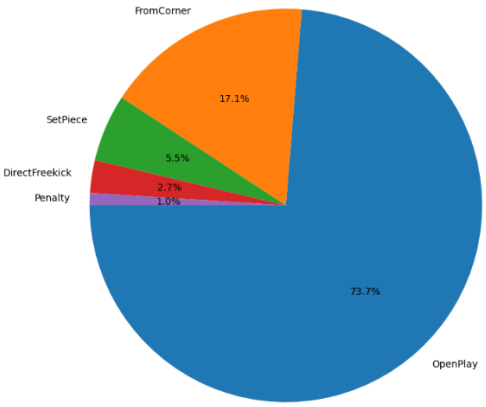
Oyuncuların oyuna başladığı dakikayı gösteren bir ‘entered’ değişkeni oluşturulmuştur. ‘oyuncu_perf’ tablosundaki ‘roster_out’ değişkeninin dolu olduğu satırlarda 0, boş olduğu satırlarda ‘time’ değeri X olmak üzere 90-X değerleri lambda fonksiyonu ile yeni oluşturulan değişkenin üzerine yazılmıştır.

takım_maçlar tablosundaki değişkenler takımlara göre gruplanarak sayısal değişkenlerin ortalaması alınmıştır. Bu sayede maç bazındaki sayısal istatistikler için her bir takımın tek maç ölçeğinde sezonluk ortalamasını içeren, yani her bir takım için bir kayıt buunduran 20 satırlık bir tablo oluşturulmuştur. Bu tablo, sonrasında one-to-many ilişki ile maçlar_hist tablosu ile yeniden birleştirilmiştir. Bu sayede her bir satırın bir maç içerdiği yeni tabloda, takımların o maç özelinde elde ettiği sayısal istatistiklerin yanı sıra her iki tarafın sezonluk ortalamalarının da yer aldığı ‘games_reg_match’ adında yeni bir tablo oluşturulmuştur. Bu işlem sezonluk ortalamaların makine öğrenmesi maç özelindeki istatistikleri tahmin edilmesinde kullanılmasına olanak sağlamıştır.

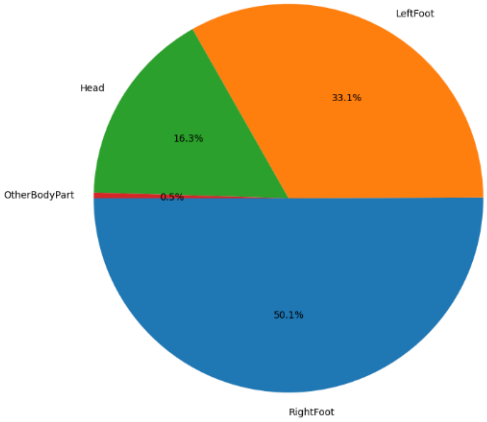
‘games_reg_match’ tablosu ile oluşturulan modellerin performansını ilgilenilen maç özelinde takımda oynayacak olan oyuncuların biliniyor olması durumunda güçlendirebilmek adına kadro özelinde sayısal istatistikler içeren df_xg adında yeni bir tablo oluşturulmuştur. Bunun için oyuncu_perf tablosundan, roster_out değişkeninin boş olduğu satırlardaki ilk 11 oyuncularına göre elde edilmiş sezonluk ortalama değişkenleri oluşturulmuştur. Bu işlem ile ilgilenilen takımın sezon genelindeki ortalamaları yerine, takımın o maç için sahaya sürdüğü kadronun sezon genelindeki ortalamaları kullanılarak daha iyi performans gösteren modeler oluşturulması amaçlanmıştır.

3.2 Veri Görselleştirme ve Performans Analizi

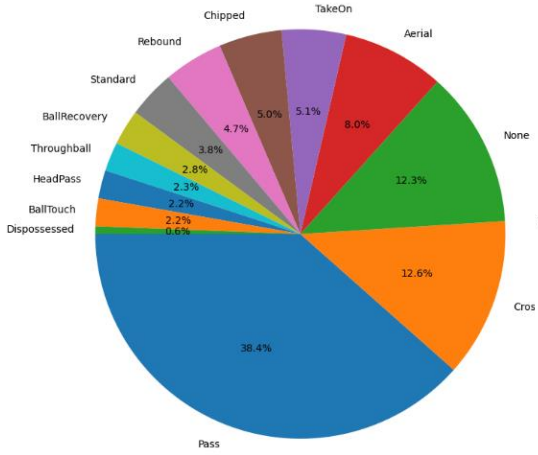
'şutlar' tablosundaki ligde çekilmiş tüm şutların niteliklerini ifade eden dört farklı kategorik değişkenin oransal dağılımlarını incelemek adına pasta grafiği oluşturmak için öncelikle 'value_counts()' fonksiyonu kullanılarak frekans tablosu oluşturulmuştur. Sonrasında matplotlib kütüphanesi ile elde edilen grafikler aşağıdaki gibi görünmektedir.



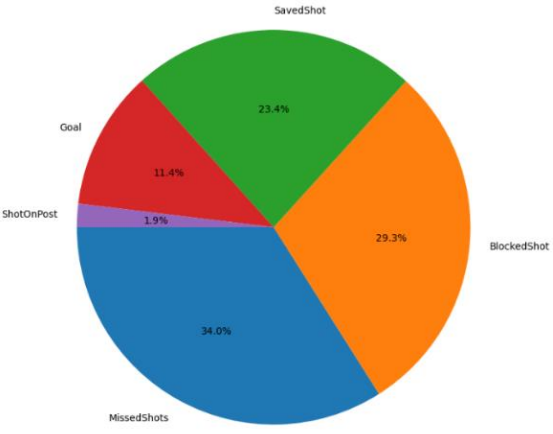
Şekil 3.2



Şekil 3.3

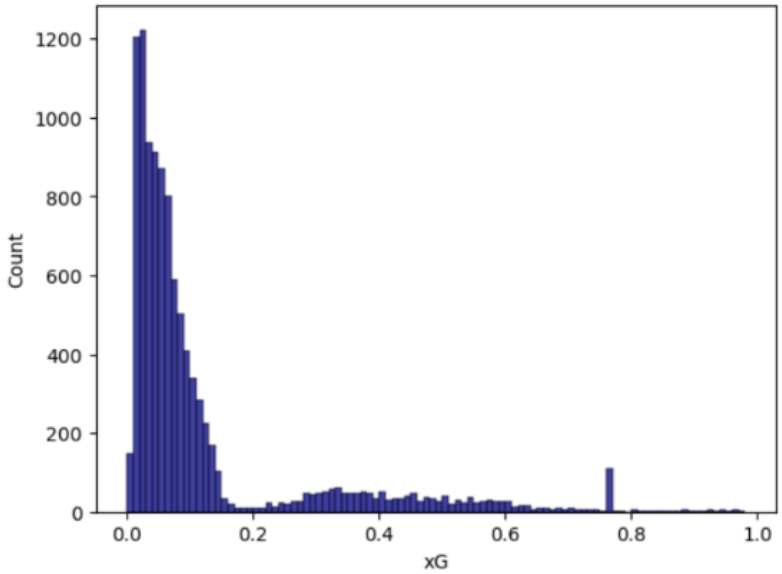


Şekil 3.4



Şekil 3.5

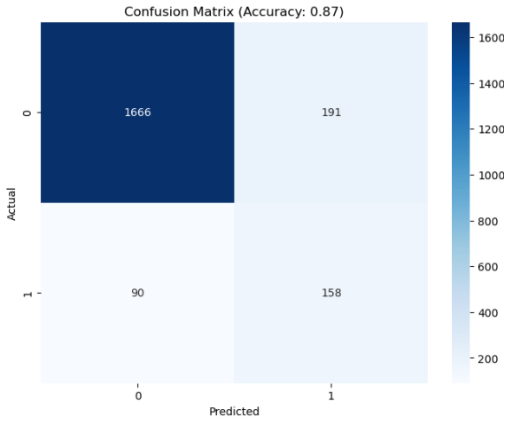
Gol beklentisi istatistiğine göre şutların frekans dağılımı şekil 3.6'daki histogram grafiğinde görselleştirilmiştir.



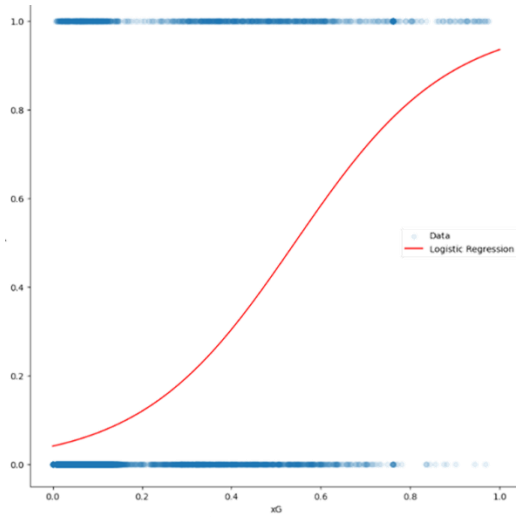
Şekil 3.6

Seaborn içerisindeki histplot fonksiyonu ile yapılan grafikte parametreler her bir sütun, 0.01 gol beklentisini ifade edecek şekilde ayarlanmıştır. 0.2 değerinin altındaki gol beklentisine sahip şutların yoğunlukta olduğu dikkat çekerken, penaltı atışlarındaki gol beklentisinin daima 0.76 olmasından dolayı bu noktadaki sütundaki artış göze çarpmaktadır.

Veri setindeki gol beklentisinin, gerçek gollerin dağılımını temsil etme kabiliyetini ölçmek için Scikit-learn kütüphanesi ile bir lojistik regresyon modeli oluşturulmuştur. Sonrasında matplotlib kütüphanesi kullanılarak verinin dağılımı modelin sigmoid fonksiyonu ile birlikte görselleştirilmiştir (Şekil 3.8). 'X' ekseninde tahmin edilen ve 'Y' ekseninde gerçek değerlerin frekansını gösteren karmaşıklık matrisi (confusion matrix) ise Seaborn kütüphanesi ile görselleştirilmiştir (Şekil 3.7).



Şekil 3.7

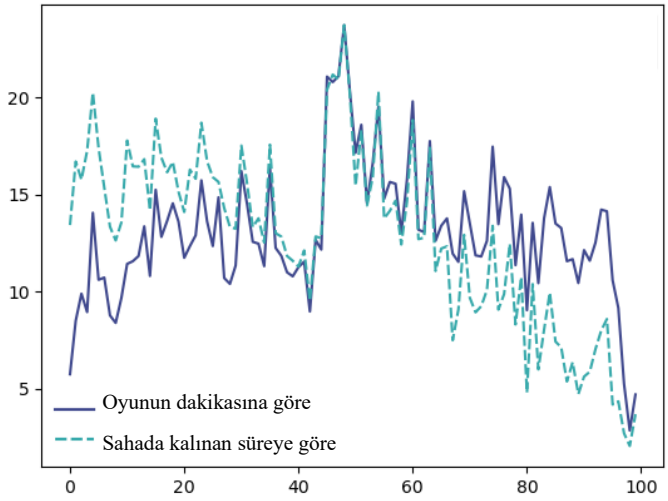


Şekil 3.8

Veri setinde gol olan ve olmayan şutların dağılımı dengesiz olduğundan (gol olmayan şutlar yaklaşık 9 kat daha fazla), lojistik regresyon modeli çalıştırılırken azınlık olan ‘gol’ sınıfındaki örnekleri çoğaltan SMOTE tekniği kullanılmıştır.

Elde edilen xG ile gol tahmin modeli %87 doğruluk ile çalışırken, doğru tahminlerin daha büyük bir bölümü çoğunluğu oluşturan ‘gol olmayan’ şutlara ait. Gol olmayan şutlardaki doğruluk %90 iken, gol ile sonuçlanan şutlardaki doğruluk %63 seviyesinde kalıyor. Şekil 3.8’de bulunan dağılım ve sigmoid fonksiyon grafiğine de bakılacak olursa, gol olmayan şutlar düşük gol beklentisi düzeylerinde yoğunlaşmaya daha eğilimliyken, gol ile sonuçlanan şutlar daha homojen bir dağılım göstermektedir.

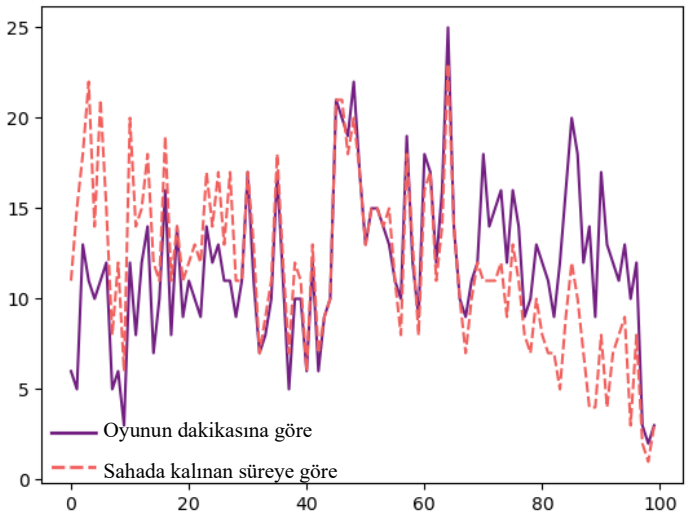
Şekil 3.9’da oyuncuların oyunda kalma sürelerine göre dakika başına toplam gol sayıları ve maç dakikalarına göre toplam gol sayıları görselleştirilmiştir. Bu işlem için ‘entered’ ve ‘time’ değişkenlerinin aldıkları her bir sayı değeri için ayrı ayrı toplam gol sayıları alınmıştır. X ekseninde bu iki değişkenin aldığı değerler ve Y ekseninde toplam gol sayıları yer almaktadır. Kesikli çizgi oyuncuların oyunda kalma süresine göre, düz çizgi maç dakikalarına göre gol sayısını göstermektedir.



Şekil 3.9

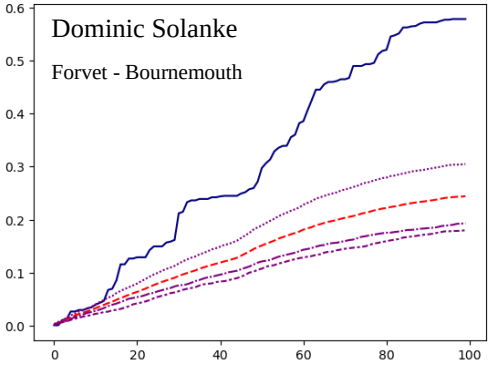
Oyuncular süre aldıkça gol sayılarının azaldığı ve maç dakikalarına göre gol sayısının altına düştüğü gözlemlenmektedir.

Şekil 3.10'da aynı parametreler, bu kez gol sayısı yerine gol beklentisi bakımından karşılaştırılmıştır. Benzer ilişki gol sayısı yerine gol beklentisi değişkeni kullanıldığında da geçerlidir.

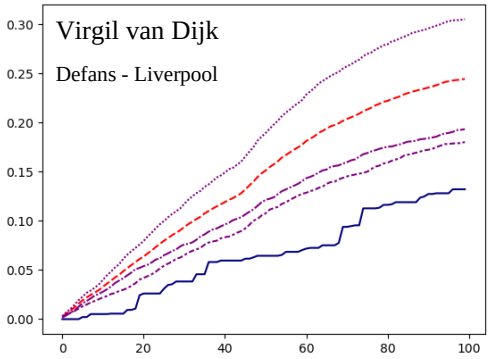


Şekil 3.10

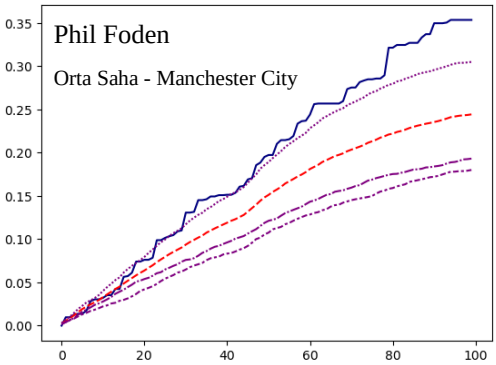
Aşağıdaki grafiklerde bazı oyuncuların, ortalama bir maç boyunca kümülatif gol beklentisi artışı görselleştirilmiş, genel ve mevki bazında ligin kümülatif gol beklentisi artışlarıyla kıyaslanmıştır. Bu işlem için ise dakika bazında gol beklentisi grafiği için kullanılan metot, aynı şekilde mevki ve oyuncuya göre filtrelendikten sonra kullanılmıştır. Her bir dakikadaki gol beklentisi, önceki dakikalar ile toplanarak kümülatif bir dağılım elde edilmiştir. Sonrasında maç bazındaki ortalamalar elde edilmek için her bir grup, kendine ait toplam dakikaya göre ölçeklendirilmiştir. Lacivert renkli çizgiler ilgili oyuncunun kümülatif gol beklentisi artışını, mor renkliler üç farklı mevkideki (forvet, orta saha, defans) oyuncuların kümülatif gol beklentisi artışını ve kırmızı renkli çizgi genel kümülatif gol beklentisi artışını göstermektedir.



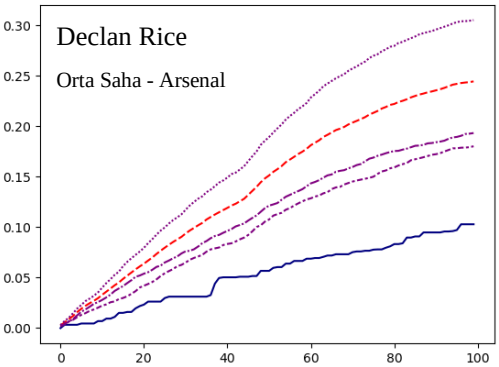
Şekil 3.11



Şekil 3.12



Şekil 3.13



Şekil 3.14

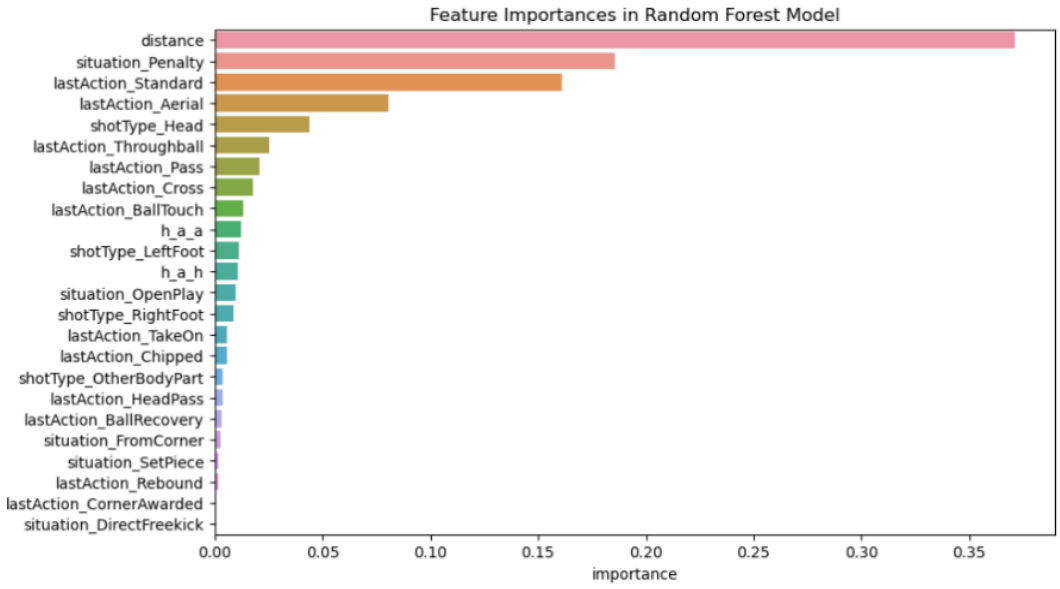
İngiltere Premier Lig'de 2023-2024 sezonunu en çok gol ile tamamlayan 5 oyuncu ve attıkları gol ile gol beklentisi arasındaki farklar oyuncular tablosuna uygun filtreler uygulanarak tablo 3.6'da göründüğü şekilde listelenmiştir

OYUNCU ADI	GOL	XG	TAKIM	GOL XG FARKI
ERLİNG HAALAND	27	31.653997	Manchester City	-4.653997
COLE PALMER	22	17.832245	Chelsea, Manchester City	4.167755
ALEXANDER ISAK	21	22.074266	Newcastle United	-1.074266
DOMİNİC SOLANKE	19	21.406831	Bournemouth	-2.406831
PHİL FODEN	19	11.307983	Manchester City	7.692017

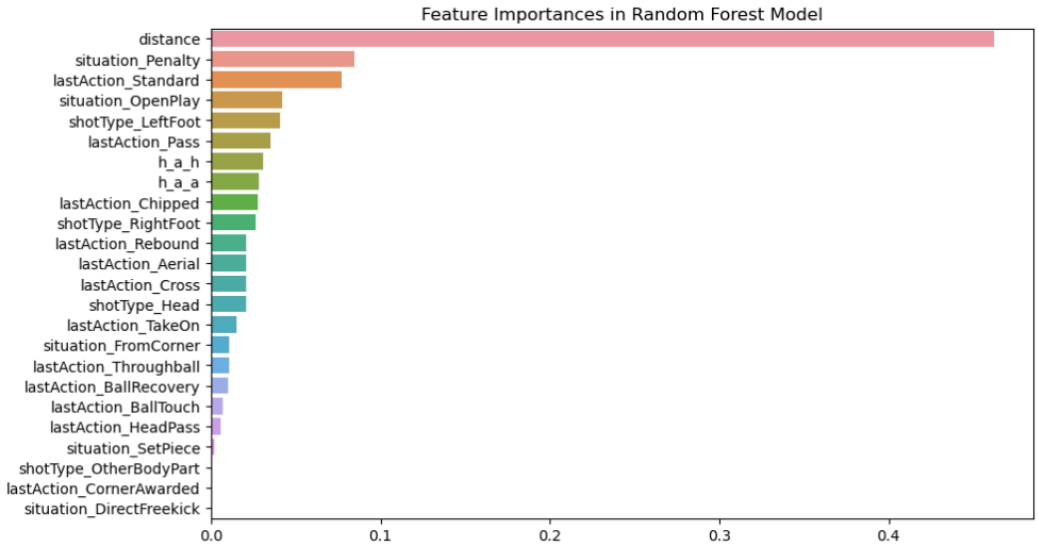
Tablo 3.6

Tabloya göre gol krallığı yarışını 27 gol ile zirvede tamamlayan oyuncu Erling Haaland'ın, lig boyunca attığı şutlardaki toplam gol sayısı, gol beklentisine göre 4.65 eksiktir. Bu, oyuncunun elde ettiği şut imkanlarını ortalamanın altında bir başarıyla değerlendirdiğini gösterir. Bu, ligi zirvede tamamlamış bir oyuncu için çarpıcı bir gerçektir.

Makine Öğrenmesi bölümündeki rastgele orman örneklerine geri dönelim, burada sembolik olarak Erling Haaland'ın gol ve gol beklentisi değerlerini tahmin eden bir algoritma oluşturmuştuk. Bu algoritmalarda kullanılan bağımsız değişkenlerin önem sırasına göre sıralanmış halini kıyaslamak gerekirse, şekil 3.15'de gol beklentisi tahmini için kullanılan rastgele orman regresyon modelindeki bağımsız değişkenlerin önem derecesini belirten histogram grafiği mevcuttur.

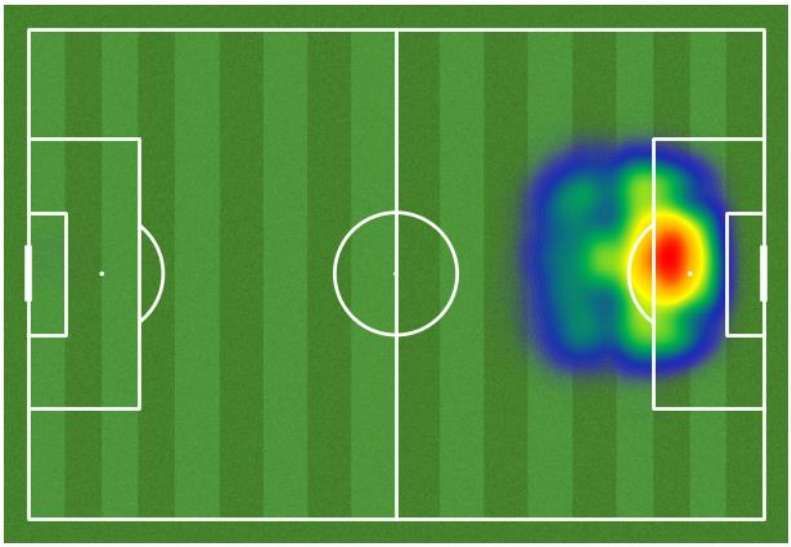


Şekil 3.16'de ise, gol tahmini için kullanılan rastgele orman sınıflandırma modeli için kullanılan bağımsız değişkenlerin önem derecesini belirten histogram grafiği mevcuttur.



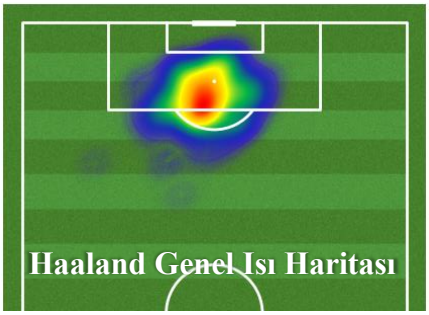
Şekil 3.16

Her iki grafikteki değişkenlerin etki düzeylerine göre oyuncunun oyun tarzı adına daha fazla fikir edinilebilir. Yüksek derecede etki eden faktörlerin etkisinin yönü saptanarak spesifik olarak optimize edilmeye çalışılabilir. Değişkenlerin etkisi, ısı haritaları gibi alternatif tanımlayıcı metotlarla da desteklenip detaylandırılabilir. Örneğin şekil 3.17'deki ısı haritası Mplsoccer kütüphanesi ile oluşturulmuş olup, Sezon boyunca İngiltere Premier Lig'de atılmış tüm şutların ısı haritasını göstermektedir.

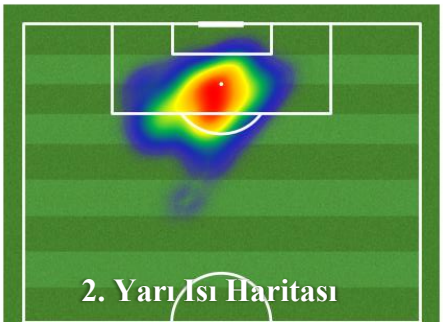


Şekil 3.17

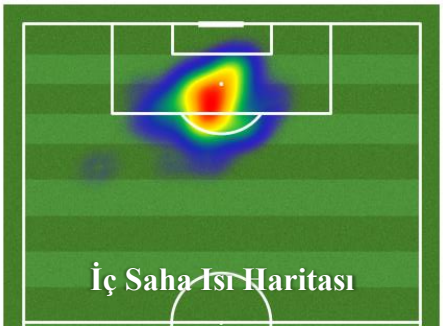
Beklendiği üzere çekilen 10000'i aşkın şutun hedefe göre hemen hemen simetrik ve homojen bir dağılımı vardır. Fakat veri, pandas kullanılarak oyuncular özelinde veya çekilen şutun belli başlı özelliklerine göre filtrelendiğinde, ısı haritasının ayırt edici özellikleri ortaya çıkmaktadır. Haaland'ın gol fırsatlarını değerlendirebilme durumunu etkileyen faktörlerin etki derecesini saptamak için kullandığımız rastgele orman modellinden elde edilen çıktılar hakkında konum ile ilgili bilgi edinmek için bu haritalar kullanılabilir. Aşağıdaki şekillerde Erling Haaland'ın farklı özelliklere sahip sınırsız sayıda kombinasyonla filtrelenebilecek şutlarının ısı haritalarına bazı örnekler mevcuttur.



Şekil 3.18



Şekil 3.19



Şekil 3.20

Şekil 3.21

Gol ve xG için oluşturulan rastgele orman modellerinin değişken etkilerini gösteren iki tablo kıyaslandığında ise, değişkenlerin önem sıralamaları ve dereceleri aralarındaki farklılıklar göze çarpmaktadır. Örneğin gol beklentisi algoritmasında, uzaklık faktörünün etkisi daha az olarak görünürken, Haaland'ın pozisyonları gole çevirebilmesindeki etkisi daha fazla olduğu görülüyor.

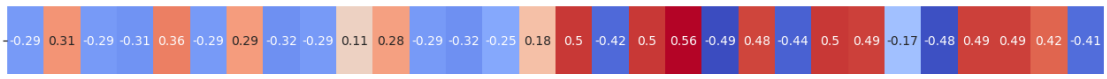
Buradan genel geçer olarak kullanılan gol beklentisi istatistiğinin, temsil gücünün oyuncudan oyuncuya değişebileceği sonucuna varılabilir. Bu çalışmada kullanılan veri seti, her bir oyuncu için bir sezonda atılmış şutlardan oluşan küçük bir veri grubunu içeriyor olsa da antrenman verileri veya daha geriye dönük detaylı veriler toplandığında oyunculara özgü gol beklentisi algoritmaları geliştirilebilir, bu algoritmaları en pozitif yönde etkileyen faktörlerin optimizasyonu üzerine çalışılarak performans analizi alanında verimli sonuçlar elde edilebilir.

3.3 Makine Öğrenmesi ile Maç Senaryosu Tahmini

Bu bölümde, bir maç öncesi teknik anlamda yararlı bilgiler sağlayacak nitelikte senaryo tahminleri için lineer regresyon, karar ağaçları, destek vektör makineleri ve yapay sinir ağları ile dört farklı regresyon modeli oluşturulmuştur. Bağımsız değişken seçiminde değişkenlerin tanımlarından ve korelasyon katsayılarından faydalanılırken, farklı kombinasyonlarla yapılan denemelerde elde edilen sonuçlar kıyaslanmıştır. Her bir modelin performansını tanımlamak için MSE (Hata kareler ortalaması), MAE (Mutlak hata ortalaması), RMSE (Ortalama karekök sapması) R^2 (Tanımlayıcılık katsayısı) kriterleri incelenmiştir.

3.3.1 Kaleye Yakın Pas Aksiyonlarının Tahmini

Rakibin kaleye 20 metre yakında yaptığı pas sayısı değişkeninin, ilgilenilen takım ve rakip takıma ait verilerle arasındaki doğrusal ilişkiyi belirten korelasyon katsayıları şekil 3.25'deki gibidir.



Şekil 3.22

Kullanılan bağımsız değişkenler şu şekildedir:

1. İlgilenilen takımın geçmiş rakiplerinin kaleye 20 metre yakında yaptığı ortalama pas sayısı
2. İlgilenilen takımın ortalama gol beklentisi toleransı
3. İlgilenilen takımın ortalama puan beklentisi
4. İlgilenilen takımın galibiyet oranı
5. Rakip takımın geçmiş maçlarda kaleye 20 metre yakında yaptığı pas sayısı
6. Rakip takımın penaltısız ortalama gol beklentisi

7. Rakip takımın ortalama puan beklentisi
8. İlgilenilen takımın tarafı (iç saha/deplasman)

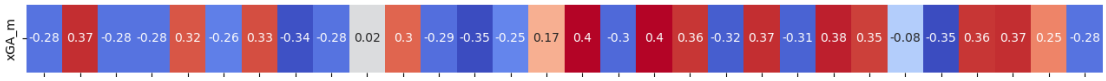
Oluşturulan regresyon modellerinin tutarlılığına dair istatistikler tablo 3.7'deki gibidir.

	LINEAR REGRESSİON		REGRESSION TREE		SVR		ANN	
	Test	train	test	train	test	train	test	train
MSE	12.66	17.78	13.59	16.09	11.91	17.86	12.64	17.8
MAE	2.78	3.18	3.0	2.94	2.7	3.05	2.84	3.18
RMSE	3.56	4.22	3.69	4.01	3.45	4.23	3.56	4.22
R^2	0.46	0.45	0.42	0.5	0.49	0.44	0.46	0.45

Tablo 3.7

3.3.2 Gol Beklentisi Toleransı Tahmini

Rakibin gol beklentisine gösterilen tolerans değişkeninin, ilgilenilen takım ve rakip takıma ait verilerle arasındaki doğrusal ilişkiyi belirten korelasyon katsayıları şekil 3.26'deki gibidir



Şekil 3.23

Kullanılan bağımsız değişkenler şu şekildedir:

1. İlgilenilen takımın ortalama gol beklentisi toleransı
2. Rakip takımın ortalama gol beklentisi
3. Rakip takımın ortalama puan beklentisi
4. İlgilenilen takımın tarafı (iç saha/deplasman)

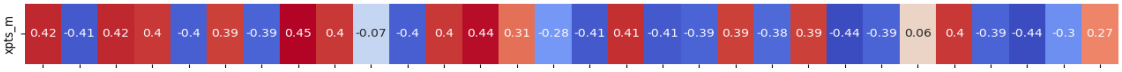
Oluşturulan regresyon modellerinin tutarlılığına dair istatistikler tablo 3.8'deki gibidir.

	LİNEAR REGRESSİON		REGRESSİON TREE		SVR		ANN	
	test	train	test	train	test	train	test	train
MSE	0.78	0.62	0.88	0.58	0.86	0.61	0.79	0.61
MAE	0.69	0.62	0.71	0.59	0.71	0.59	0.69	0.62
RMSE	0.88	0.79	0.94	0.76	0.93	0.78	0.89	0.78
R²	0.34	0.34	0.25	0.38	0.27	0.35	0.33	0.34

Tablo 3.8

3.3.3 Puan Beklentisi Tahmini

Maç bazındaki puan beklentisi değişkeninin, ilgilenilen takım ve rakip takıma ait sezon bazındaki ortalama verileriyle arasındaki doğrusal ilişkiyi belirten korelasyon katsayıları şekil 3.24'deki gibidir.



Şekil 3.24

Kullanılan bağımsız değişkenler şu şekildedir:

1. İlgilenilen takımın ortalama puan beklentisi
2. Rakip takımın ortalama gol beklentisi
3. Rakip takımın ortalama puan beklentisi
4. İlgilenilen takımın tarafı (iç saha/deplasman)

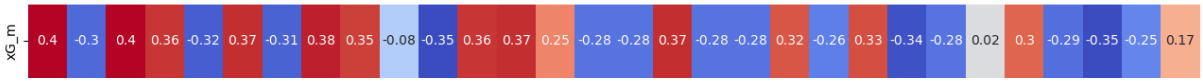
Oluşturulan regresyon modellerinin tutarlılığına dair istatistikler tablo 3.9'daki gibidir.

	LİNEAR REGRESSİON		REGRESSİON TREE		SVR		ANN	
	test	train	test	train	test	train	test	train
MSE	0.38	0.42	0.47	0.4	0.4	0.42	0.37	0.42
MAE	0.5	0.54	0.55	0.51	0.49	0.5	0.49	0.54
RMSE	0.62	0.65	0.69	0.63	0.63	0.65	0.61	0.65
R²	0.49	0.46	0.38	0.49	0.48	0.46	0.51	0.46

Tablo 3.9

3.3.4 Gol beklentisi Tahmini

Maç bazındaki gol beklentisi değişkeninin, ilgilenilen takım ve rakip takıma ait sezon bazındaki ortalama verileriyle arasındaki doğrusal ilişkiyi belirten korelasyon katsayıları şekil 3.25'deki gibidir.



Şekil 3.25

Kullanılan bağımsız değişkenler şu şekildedir:

1. İlgilenilen takımın ortalama gol beklentisi
2. İlgilenilen takımın ortalama puan beklentisi
3. Rakip takımın ortalama gol beklentisi toleransı
4. Rakip takımın ortalama puan beklentisi
5. İlgilenilen takımın tarafı (iç saha/deplasman)

Oluşturulan regresyon modellerinin tutarlılığına dair istatistikler tablo 3.10'daki gibidir.

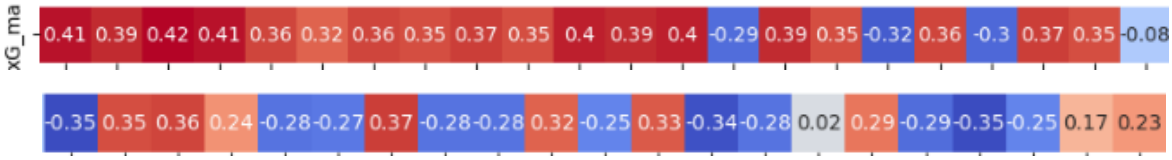
	LİNEAR REGRESSİON		REGRESSİON TREE		SVR		ANN	
	test	train	test	train	test	train	test	train
MSE	0.57	0.67	0.73	0.68	0.58	0.69	0.58	0.69
MAE	0.61	0.65	0.67	0.64	0.61	0.64	0.61	0.66
RMSE	0.75	0.82	0.86	0.82	0.76	0.83	0.76	0.83
R^2	0.35	0.34	0.16	0.33	0.34	0.33	0.33	0.32

Tablo 3.10

3.3.4.1 Gol Beklentisi Tahmin Modelini İyileştirmek Üzere Veri İşleme

Veri ön işleme ile eldeki bilgiden faydalanarak daha verimli metriklerle daha iyi performans gösteren modeler oluşturmak mümkündür. Bir önceki bölümde maç bazındaki gol beklentisini tahmin etmek üzere oluşturulan modelde tahmin işlemi her iki tarafın sezonluk ortalamaları ile yapılıyordu. Ne var ki bazı durumlarda sahada bulunan oyuncular, sezon genelinden farklı olabilmektedir. Öyleyse sahaya çıkacak oyuncuların biliniyor olması, modeli iyileştirmeye fayda sağlayabilir. Bu

bölümde sezonluk ortalamaların yanı sıra, sahaya çıkan kadrodaki oyunculara özel olarak oluşturulmuş sezonluk gol beklentisi değişkeni kullanılmıştır. Yeni değişkenlerle birlikte, maç bazındaki gol beklentisi değişkeninin diğer tüm değişkenlerle arasındaki doğrusal ilişkiyi belirten korelasyon katsayıları şekil 3.26'daki gibidir.



Şekil 3.26

Kullanılan bağımsız değişkenler şu şekildedir:

1. **İlgilenilen takımın sahaya çıktığı kadronun ortalama gol beklentisi**
2. İlgilenilen takımın ortalama puanı
3. Rakip takımın ortalama gol beklentisi toleransı
4. Rakip takımın ortalama puanı
5. İlgilenilen takımın tarafı (iç saha/deplasman)

Oluşturulan regresyon modellerinin tutarlılığına dair istatistikler tablo 3.11'deki gibidir.

	LİNEAR REGRESSİON		**REGRESSİON TREE**		**SVR**		**ANN**	
	test	train	test	train	test	train	test	train
MSE	0.58	0.72	0.65	0.74	0.59	0.68	0.59	0.73
MAE	0.62	0.67	0.62	0.68	0.6	0.63	0.61	0.67
RMSE	0.76	0.85	0.81	0.86	0.77	0.83	0.77	0.86
R^2	0.36	0.34	0.29	0.32	0.35	0.37	0.36	0.33

Tablo 3.11

3.3.5 Sonuç Değerlendirme

Makine öğrenmesi sayesinde oynanmamış maçlarda yaşanacak olan olaylar hakkında tahminler yapmak mümkündür. Understat.com sitesinde tanımlayıcı istatistikler sunmak ve veri görselleştirmede kullanılmak üzere tutulan veriler ile, %30 ila %50 arası tanımlayıcılık düzeylerine sahip makine öğrenmesi modelleri oluşturulmuştur. Bu modellerin, eldeki bilgiler doğrultusunda doğru veri mühendisliğiyle iyileştirilebildiği görülmüştür. Derin sektör bilgisi ile amaca uygun olarak oluşturulmuş daha büyük veri setleri ve daha kapsamlı ön işleme metotlarıyla daha tutarlı modeller elde edilebilmesi mümkündür. Bu modeller, saha içi yöneticilerine değerli öngörüler sağlayabilir.

4
SONUÇ

Sporda veri analitiğinin 50 yılı aşkın geçmişi olduğu bilinse de teknolojideki gelişmelerle hızla değişen, yükselen ivmeyle ilerleyen bir çalışma alanı haline gelmiştir. Geniş kitleleri ilgilendiren bir çalışma alanı olmasından doğan veri yoğunluğu spor sektörünü de veri bilimi için değerli bir konuma getirmiştir. Öyle ki spor analitiğindeki bazı çalışmalar psikoloji, davranış ekonomisi gibi farklı alanlardaki çalışmalara konu olmuştur. Spor özelinde ise sakatlık tahminlerinden saha dışı yönetimine kadar geniş bir kullanım alanına sahip veri biliminin teknik anlamdaki oyun anlayışına teshiri özellikle 2000'li yılların ardından geniş kitlelerin takip ettiği spor dallarında etkisini güçlü şekilde hissettirmiştir.

Son dönemlerde yapay zekâ tabanlı metriklerin futbol başta olmak üzere çeşitli spor dallarında kullanılmaya başlanması spor analitiğinde yeni bir sayfa açmıştır. Sonuç odaklı bir yaklaşım olarak gol beklentisinin, gole giden yolda yüksek tutarlılıkta tahmin gücüne sahip bir metrik olduğu çalışmadaki bulgularda saptanmıştır. Bu gibi metriklerin diğer spor dallarında da kullanımının yaygınlaşması mümkündür. 2010'lu yıllarda basketbolda Daryl Morey gibi isimlerin yeni oyun anlayışı yaklaşımı, futboldaki gol beklentisi metriğine benzer olarak bölge bazlı sayı beklentisinin doğru analiz edilmesine bağlı bulguların sonucudur. Bu bağlamda tıpkı gol beklentisinin şutun çeşitli özelliklerine bağlı

olarak değişmesi gibi, basketbol oyuncularının son dönemdeki analizler sonucu orta mesafeli atışlardan üç sayılık atışlara ve turnikelere yönelmesi, bir pekiştirmeli makine öğrenmesi modelinin elde ettiği sonuçları optimize etmek üzere yeni davranış politikaları belirlemesinden farksızdır.

Gol beklentisi metriği artık her ne kadar futbolda aktif olarak kullanılan bir metrik haline gelmiş olsa da çalışmada farklı amaçlar için daha verimli formlarda kullanılabileceğine dair bulgular elde edilmiştir. Ortalamaya göre yüksek seviyede gol sayısına ulaşmayı başaran oyuncuların gol sayısı, gol beklentisinin altında kalabilmektedir. Bunun, oyuncunun yakaladığı fırsatları değerlendirebilme becerisi ile ilişkili olabilmesi mümkün olsa da aynı zamanda birçok farklı seviyedeki oyuncunun şutları ile eğitilmiş gol beklentisi metriğini ortaya çıkaran modelin tutarlılığının oyuncudan oyuncuya değişebildiği anlamına da gelebilir. Çalışmada elde edilen bulgulara göre, yüksek gol sayısına sahip oyuncuların pozisyonları gole çevirebilmelerini etkileyen faktörler ile gol beklentisini yükselten faktörler arasında farklılıklar tespit edilmiştir. Bu durum gol beklentisi metriğini ortaya çıkaran modelin, oyuncuya, oyuncunun oyun tarzına veya lige özgü, sayı olarak yeterli miktarda olmak şartıyla daha özgün veriler ile eğitildiği takdirde tahmin gücünün artacağını ortaya koymuştur. Bu alanda ileride daha geniş veri setleri ile yapılacak çalışmalar, oyunculara uygun taktiksel planlar geliştirmede, eksik noktaların tespit edilmesinde ve skora yönelik sonuçları optimize etmede fayda sağlayacaktır.

Gol beklentisinin yanı sıra bir futbol maçından çıkarılabilecek sayısız sayısal değişken, makine öğrenmesi ile maç senaryolarını tahmin etmede kullanılabilmektedir. Bu çalışmada “kalenin yakınında yapılacak pas sayısı” gibi somut veya gol beklentisi gibi türetilmiş istatistiklerin, %50’ye varan tutarlılık katsayıları ile tahmin edilmiş olmasının yanı sıra, sonuçların daha derin teknik bilgi ile detaylandırılmaya açık uygun veri mühendisliği teknikleri ile iyileştirilebildiği görülmüştür. Bu alanda ileride yapılacak çalışmalar, saha içi yöneticilerine teknik anlamda daha kapsamlı öngörüler sunacaktır.

KAYNAKÇA

[1] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, 3(3), 210-229. https://doi.org/10.1147/rd.33.0210

[2] Cook, E. (1966). *Percentage Baseball*. MIT Press.

[3] 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

[4] Cortes, C., Vapnik, V. Support-vector networks. Mach Learn 20, 273–297 (1995). https://doi.org/10.1007/BF00994018

[5] Itchhaporia, D., Snow, P., Almassy, R., & Oetgen, W. (1996). Artificial neural networks: current status in cardiovascular medicine.. Journal of the American College of Cardiology, 28 2, 515-21 . https://doi.org/10.1016/0735-1097(96)00174-X.

[6] Sutton, R.S., & Barto, A.G. (1998). Introduction to Reinforcement Learning.

[7] Atkinson, G., & Nevill, A. (2001). Selected issues in the design and analysis of sport performance research. Journal of Sports Sciences, 19, 811 - 827. https://doi.org/10.1080/026404101317015447.

[8] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[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] Lewis, M. (2003). *Moneyball: The Art of Winning an Unfair Game*. W.W. Norton & Company.

[11] Burns, B. D. (2004). Heuristics as beliefs and as behaviors: The adaptiveness of the "hot hand". *Cognitive Psychology, 48*(3), 295–331. https://doi.org/10.1016/j.cogpsych.2003.07.003

[12] Cortes, C., & Vapnik, V.N. (2004). Support-vector networks. Machine Learning, 20, 273-297.

[13] Smola, A.J., Schölkopf, B. A tutorial on support vector regression. Statistics and Computing 14, 199–222 (2004). https://doi.org/10.1023/B:STCO.0000035301.49549.88

[14] Cherkassky, V., & Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. Neural networks : the official journal of the International Neural Network Society, 17 1, 113-26.

[15] Sundali, J., & Croson, R. (2006). Biases in casino betting: The hot hand and the gambler's fallacy. Judgment and Decision Making, 1(1), 1–12. https://doi.org/10.1017/S1930297500000309

[16] McGarry, T. (2009). Applied and theoretical perspectives of performance analysis in sport: Scientific issues and challenges. International Journal of Performance Analysis in Sport, 9, 128 - 140. https://doi.org/10.1080/24748668.2009.11868469.

[17] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.
[18] Zhu, X., & Goldberg, A.B. (2009). Introduction to Semi-Supervised Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning.
[19] RABIN, M., & VAYANOS, D. (2010). The Gambler's and Hot-Hand Fallacies: Theory and Applications. The Review of Economic Studies, 77(2), 730–778. http://www.jstor.org/stable/40587644
[20] Yaari, G., & Eisenmann, S. (2011). The hot (invisible?) hand: Can time sequence patterns of success/failure in sports be modeled as repeated random independent trials? PLOS ONE, 6(10), e24532. https://doi.org/10.1371/journal.pone.0024532
[21] Stoltzfus, J. (2011). Logistic regression: a brief primer.. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine, 18 10, 1099-104 . https://doi.org/10.1111/j.1553-2712.2011.01185.x.
[22] Loh, W. (2011). Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1.
[23] Xie, J., Jiang, S., Xie, W., & Gao, X. (2011). An Efficient Global K-means Clustering Algorithm. J. Comput., 6, 271-279.
[24] Su, X., Yan, X. and Tsai, C.-L. (2012), Linear regression. WIREs Comp Stat, 4: 275-294. https://doi.org/10.1002/wics.1198
[25] Bergkamp, D., & Winner, D. (2013). Stillness and Speed: My Story. Simon & Schuster UK.
[26] Cohen, B. (2014, February 27). Does the 'hot hand' exist in basketball? Wall Street Journal. Retrieved May 6, 2016, from https://www.wsj.com/
[27] Coutts, A. J. (2014). Evolution of football match analysis research. Journal of Sports Sciences, 32(20), 1829–1830. https://doi.org/10.1080/02640414.2014.985450
[28] Sperandei S. (2014). Understanding logistic regression analysis. Biochemia medica, 24(1), 12–18. https://doi.org/10.11613/BM.2014.003
[29] Schmidhuber, J. (2014). Deep learning in neural networks: An overview. Neural networks : the official journal of the International Neural Network Society, 61, 85-117 .
[30] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2014). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005
[31] Jordan, M., & Mitchell, T. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349, 255 - 260. https://doi.org/10.1126/science.aaa8415.
[32] Min Cao, Roman Chychyla, Trevor Stewart; Big Data Analytics in Financial Statement Audits. *Accounting Horizons* 1 June 2015; 29 (2): 423–429. https://doi.org/10.2308/acch-51068.
[33] Erwan Scornet. Gérard Biau. Jean-Philippe Vert. "Consistency of random forests." Ann. Statist. 43 (4) 1716 - 1741, August 2015. https://doi.org/10.1214/15-AOS1321

[34] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. Nature, 521, 436-444. https://doi.org/10.1038/nature14539.
[35] R. Stanojevic and L. Gyarmati, "Towards Data-Driven Football Player Assessment," *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, Barcelona, Spain, 2016, pp. 167-172, doi: 10.1109/ICDMW.2016.0031.
[36] Green, B., & Zwiebel, J. (n.d.). The hot hand fallacy: Cognitive mistakes or equilibrium adjustments? Evidence from baseball. Stanford Graduate School of Business. Retrieved May 6, 2016, from https://www.gsb.stanford.edu/
[37] Miller, J. B., & Sanjurjo, A. (2016). Surprised by the gambler's and hot hand fallacies? A truth in the law of small numbers. IGIER Working Paper, 552. https://doi.org/10.2139/ssrn.2627354
[38] Mammone A., Turchi M., Cristianini N. Support vector machines. Wiley Interdisciplinary Reviews: Computational Statistics 2009; 1(3): 283-289.
[39] Rokach, L. (2016). Decision forest: Twenty years of research. Information Fusion, 27, 111-125. https://doi.org/10.1016/j.inffus.2015.06.005
[40] Juravich, M., Salaga, S., & Babiak, K. (2017). Upper Echelons in Professional Sport: The Impact of NBA General Managers on Team Performance. Journal of Sport Management, 31, 466-479.
[41] Lee, J.H., Shin, J., & Realff, M.J. (2017). Machine learning: Overview of the recent progresses and implications for the process systems engineering field. Comput. Chem. Eng., 114, 111-121.
[42] Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning k for kNN Classification. ACM Transactions on Intelligent Systems and Technology (TIST), 8, 1 - 19. https://doi.org/10.1145/2990508.
[43] Hodeghatta, U. R., & Nayak, U. (2017). Business Analytics Using R - A Practical Approach. Apress.
[44] Yu, S., Chu, S., Wang, C., Chan, Y., & Chang, T. (2017). Two improved k-means algorithms. Appl. Soft Comput., 68, 747-755. https://doi.org/10.1016/j.asoc.2017.08.032.
[45] Morgulev, E., Azar, O., & Lidor, R. (2018). Sports analytics and the big-data era. International Journal of Data Science and Analytics, 5, 213-222. https://doi.org/10.1007/s41060-017-0093-7.
[46] Pollard, R. (2019). Invalid interpretation of passing sequence data to assess team performance in football: Repairing the tarnished legacy of Charles Reep. The Open Sports Sciences Journal, 12(1), 17-21. https://doi.org/10.2174/1875399X01912010017
[47] Duquette, C. M., Cebula, R. J., & Mixon, F. G. (2019). Major league baseball's *Moneyball* at age 15: a re-appraisal. *Applied Economics*, *51*(52), 5694–5700. https://doi.org/10.1080/00036846.2019.1617399
[48] Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.
[49] Ying, X. (2019). An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series, 1168.

[50] Sarlis, V., & Tjortjis, C. (2020). Sports analytics - Evaluation of basketball players and team performance. Inf. Syst., 93, 101562. https://doi.org/10.1016/j.is.2020.101562.
[51] Hoege, J., Lansing, M., Nelson, S., Ungerleider, D., Iyer, R., Rhodes, C., Metzger, B., Worcester, P., Chandra, A., Leonard, J., Kreitzer, R., & Scherer, W. (2020). An Interdisciplinary Approach to Sports Analytics in a University Setting. 2020 Systems and Information Engineering Design Symposium (SIEDS), 1-6. https://doi.org/10.1109/SIEDS49339.2020.9106647.
[52] Elitzur, R. (2020). Data analytics effects in Major League Baseball. *Omega*, 90, Article 102021. https://doi.org/10.1016/j.omega.2018.11.010
[53] McNair, B., Margolin, E., Law, M., & Ritov, Y. (2020). The hot hand and its effect on the NBA. arXiv:2010.15943 [stat.AP].
[54] Sarlis, V., Chatziilias, V., Tjortjis, C., & Mandalidis, D. (2021). A Data Science approach analysing the Impact of Injuries on Basketball Player and Team Performance. Inf. Syst., 99, 101750. https://doi.org/10.1016/J.IS.2021.101750.
[55] Watanabe, N. M., Shapiro, S., & Drayer, J. (2021). Big Data and Analytics in Sport Management. Journal of Sport Management, 35(3), 197-202. Retrieved May 27, 2024, from https://doi.org/10.1123/jsm.2021-0067
[56] Branga,V.(2021).Big Data Analytics in Basketball Versus Business. Studies in Business and Economics,16(3) 24-31. https://doi.org/10.2478/sbe-2021-0042
[57] Miller, J. B., & Sanjurjo, A. (2021). Is it a fallacy to believe in the hot hand in the NBA three-point contest? European Economic Review, 138, 103771. https://doi.org/10.1016/j.euroecorev.2021.103771
[58] Cebeci Z., Yıldız, F., & Kayaalp G., (2015). K-Ortalamalar Kümelemesinde Optimum K Değeri Seçilmesi . 2. Ulusal Yönetim Bilişim Sistemleri Kongresi (pp.231-242). Erzurum, Turkey
[59] Janiesch, C., Zschech, P. & Heinrich, K. Machine learning and deep learning. Electron Markets 31, 685–695 (2021). https://doi.org/10.1007/s12525-021-00475-2
[60] Koshkina, M., Pidaparthy, H., & Elder, J.H. (2021). Contrastive Learning for Sports Video: Unsupervised Player Classification. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 4523-4531.
[61] Umami, I., Gautama, D., & Hatta, H. (2021). implementing the Expected Goal (xG) model to predict scores in soccer matches. International Journal of Informatics and Information Systems, 4(1), 38-54. https://doi.org/10.47738/ijiis.v4i1.76
[62] Pinheiro, R., & Szymanski, S. (2022). All Runs Are Created Equal: Labor Market Efficiency in Major League Baseball. Journal of Sports Economics, 23(8), 1046-1075. https://doi.org/10.1177/15270025221085712
[63] Pelechrinis, K., & Winston, W. (2022). The hot hand in the wild. PLOS ONE, 17(1), e0261890. https://doi.org/10.1371/journal.pone.0261890

[64] Bhattamisra, S., Banerjee, P., Gupta, P., Mayuren, J., Patra, S., & Candasamy, M. (2023). Artificial Intelligence in Pharmaceutical and Healthcare Research. Big Data Cogn. Comput., 7, 10. https://doi.org/10.3390/bdcc7010010.

[65] Hewitt, J. H., & Karakuş, O. (2023). A machine learning approach for player and position adjusted expected goals in football (soccer). Frontiers in Artificial Intelligence and Operations Research, 2, 100034. https://doi.org/10.1016/j.fraope.2023.100034

[66] Mead J, O'Hare A & McMenemy P (2023) Expected goals in football: Improving model performance and demonstrating value. Muazu Musa R (Editor) PLOS ONE, 18 (4), Art. No.: e0282295. https://doi.org/10.1371/journal.pone.0282295

Ek Kaynaklar

- Zhang, T. (2001). An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. AI Mag., 22, 103-104.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Baca, A., & Kornfeind, P. (2012). Stability analysis of motion patterns in biathlon shooting. *Human movement science*, *31*(2), 295–302. https://doi.org/10.1016/j.humov.2010.05.008
- Provost, F., & Fawcett, T. (2013). Data Science and Its Relationship to Big Data and Data-Driven Decision Making. *Data Science Journal*, 1(1), 51-59.
- Sampaio, J. E., Lago, C., Gonçalves, B., Maçãs, V. M., & Leite, N. (2014). Effects of pacing, status and unbalance in time motion variables, heart rate and tactical behaviour when playing 5-a-side football small-sided games. *Journal of science and medicine in sport*, *17*(2), 229–233. https://doi.org/10.1016/j.jsams.2013.04.005
- Batra, R., & Verma, S. (2020). A study of the adoption of artificial intelligence by the marketing professionals of India. *Omega*, *90*, 102021. https://doi.org/10.1016/j.omega.2018.09.005
- https://understat.com/