Spor Bilimleri Araştırmaları Dergisi
Journal of Sport Sciences Researches
Cilt/Vol: 6, Sayı/Issue 2, Aralık, 2021
E-ISSN: 2548-0723
URL: http://www.dergipark.org.tr/jssr

Futbol Verilerinin Karar Ağaçları ve Lojistik Regresyon Yöntemleri ile İncelenmesi*

Duygu TOPCU[1], Özgül VUPA ÇİLENGİROĞLU[2†]
[1] Dokuz Eylül Üniversitesi, Fen Bilimleri Enstitüsü, İzmir.
[2] Dokuz Eylül Üniversitesi, Fen Fakültesi, İzmir.

Orijinal Makale **DOI:10.25307/jssr.978449**
Gönderi Tarihi: 04.08.2021 Kabul Tarihi: 16.12.2021 Online Yayın Tarihi: 31.12.2021

Öz

Futbol dünyada ve Türkiye'de en çok takip edilen sporlardan biridir. Futbolun bu yaygınlık durumu, bilgi teknolojilerinde kullanılmakta ve gelişen veri bilimi ile birlikte maç istatistikleri kolay bir biçimde saptanabilmektedir. Futbol müsabakalarında en çok ilgilenilen konu ise maç sonucudur. Maç sonucunu etkileyen birçok farklı kriter (atılan gol sayısı, takımın aldığı kart sayısı, hava durumu, deplasmanda oynamak vb.) bulunmaktadır. Bu çalışmada Türkiye Futbol Federasyonu Süper Ligi 2019-2020 ve 2020-2021 sezonlarında oynanan karşılaşmalardan elde edilen veriler kullanılmıştır. Takımların kazanma ve kaybetme durumları sınıflandırma ve karar ağacı yöntemleri ile modellenmesi ise çalışmanın temel amacını oluşturmaktadır. Oynanan maçlarda ev sahibi ve rakip takımın aldığı kırmızı veya sarı kartlar, takımlarda yer alan yabancı oyuncu sayıları ve atılan gol sayıları kategorik bir biçime getirilerek bağımsız değişkenler olarak belirlenmiştir. Bu değişkenlere bağlı olarak ev sahibi takımın kazanma veya kaybetme durumu Lojistik Regresyon ve Karar Ağacı (CART, QUEST ve CHAID) algoritmaları kullanılarak modellenmiştir. Çalışma kapsamında altı ayrı model oluşturulmuştur. Oluşturulan modellerin doğruluk yüzdeleri, duyarlılıkları, seçicilikleri ve F-skor değerleri karşılaştırılarak en iyi modelin karar ağaçlarından %67.6'lık doğruluk yüzdesi ile CART algoritması olduğuna karar verilmiştir. Bu modelde yer alan rakip kırmızı kart durumu ile ofansif ve defansif güçlerin takımın kazanmasında ya da kaybetmesinde önemli olduğu tespit edilmiştir. Ayrıca futbol verilerinin modellenmesinde makine öğrenim algoritmalarının kullanılabileceği de gösterilmiştir.
Anahtar kelimeler: Futbol, Lojistik Regresyon, Karar Ağaçları Algoritmaları

Analysis of Football Data with Decision Trees and Logistic Regression Methods

Abstract

Football is one of the most followed sports in the world and Turkey. This prevalence of football is used in information technologies and match statistics can be determined easily with the developing data science. The most important issue in football competitions is the match result. There are many different criteria (the number of goals scored, the number of cards the team has received, the weather, play away, etc.) that affect the match result. The data obtained from the matches played in the Turkish Football Federation Super League 2019-2020 and 2020-2021 seasons were used. The main purpose of the study is to model the winning and losing situations of the teams with classification and decision tree methods. In the matches played, the red or yellow cards received by the host and the rival team, the number of foreign players in the teams and the number of goals scored were determined as independent variables by bringing them into a categorical format. Depending on these variables, the winning or losing situation of the home team is modeled using Logistic Regression and Decision Tree (CART, QUEST and CHAID) algorithms. Six different models were created within the scope of the study. By comparing the accuracy percentages, sensitivities, specifity and F-score values of the models created, it was decided that the best model was the CART algorithm with an accuracy percentage of 67.6% from the decision trees. It has been determined that the rival's red card situation and offensive and defensive powers in this model are important for the team to win or lose. It has also been shown that machine learning algorithms can be used in modeling football data.
Keywords: Football, Logistic Regression, Desicions Tree Algorithms

* Bu çalışma, Özgül Vupa Çilengiroğlu danışmanlığında yürütülen Duygu Topcu'ya ait yüksek lisans tezinden üretilmiştir.
† **Sorumlu yazar:** Doç. Dr. Özgül Vupa Çilengiroğlu, E-posta: ozgul.vupa@deu.edu.tr

GİRİŞ

Tüm alanlarda ve günlük hayatta artık sıkça sözü geçen makine öğrenimi, veri biliminin geleceğinin de en belirleyici unsurlarından biridir. Bu bağlamda, veri biliminin en popüler alt dalını dünyanın en popüler spor dalına uyarlamak veri bilimciler için kaçınılmaz bir düşünce olmuştur. Geçmişten bugüne futbol popülerliğini sürdüren bir spor dalıdır. Bu nedenle futbol, ulaşılabilir olması ve hitap ettiği kitlenin fazlalığı nedeni ile diğer spor branşlarına oranla farklılık göstermektedir (Coşkuner, Büyükçelebi ve Kurak, 2020). Son yıllardaki bilimsel araştırmalarda maç analizi, müsabakalardaki olayların ortaya konması ve objektif değerlendirilmesi amacıyla oldukça popüler hale gelmiştir (Carling, Williams ve Reilly, 2005). Maç analizinde maç sonucunun tahmin edilmesi ise dikkat çeken diğer konulardan biri olmuştur. Maç sonucu tahminine, müsabaka esnasındaki pek çok faktör etki etmektedir. Bunlar arasında takımların ofansif ve defansif güçleri, saha koşulları (hava sıcaklığı, mevsim vb.), alınan kırmızı veya sarı kartlar yer almaktadır.

Makine öğrenimi yöntemleri kullanılarak maç sonucunun tahmin edilmesi ise literatürde pek çok çalışmaya konu olmuştur. Bilgisayar biliminin yapay zekada sayısal öğrenme ve model tanıma çalışmalarından geliştirilmiş bir alt dalı olan makine öğrenme sistemleri 1959 yılında ele alınmıştır (Özekes, 2003). Bu yöntemler esas olarak yığınlar halindeki verilerden yararlanarak tahmin ve sınıflama yapabilen algoritmalar bütünüdür. Makine öğrenim sistemindeki algoritmaların bir kısmı tahmin ve kestirim yaparken diğer bir kısmı da sınıflandırma yapabilme kapasitesine sahiptir (Michie, Spiegelhalter ve Taylor, 1994). Makine öğrenmesi denetimli, denetimsiz ve yarı-denetimli öğrenme olmak üzere üçe ayrılmaktadır. Denetimli öğrenme yöntemlerinde veri seti kurulan modeli eğitmektedir. Eğitilen model daha sonra bir test verisi ile test edilir. Çıkan sonuçlar geçerlilik yöntemleri ile değerlendirilir. Denetimli öğrenme yönteminde girdi verisindeki bağımlı değişkenin etiketi önceden belirlidir. Model bu bilgiyi kullanarak kendini eğitir. Denetimli öğrenme regresyon ve sınıflandırma başlığı altında; lojistik regresyon, doğrusal regresyon, karar ağaçları, k-en yakın komşu, yapay sinir ağları ve Navie Bayes gibi başlıca yöntemleri kapsamaktadır. Denetimsiz öğrenme algoritmaları denetimli öğrenmeden bu yönü ile ayrılır. Denetimsiz öğrenme yönteminde bağımlı değişkenin etiketi önceden belirli değildir. Algoritma bağımsız değişkenleri kullanarak verileri daha önceden belirlenmemiş gruplara ayırır. Denetimsiz öğrenme kümeleme ve birliktelik kuralları gibi yöntemleri kullanır. Başlıca denetimsiz öğrenme algoritmaları arasında, k-ortalamalar, k-medoids ve Apriori algoritması yer alır. Yarı denetimli öğrenme yöntemi ise denetimli ve denetimsiz metotları bir arada kullanarak sonuçlar türetir (Han, Kamber ve Pei 2012).

Spor verileri ile bu makine öğrenim yöntemlerini kullanan birçok araştırma bulunmaktadır. Özellikle ofansif ve defansif güç değerlerinin hesaplandığı, farklı değişkenlere göre maç sonuçlarının tahmin edildiği çalışmalar son yıllarda araştırmacılar tarafından çeşitli verilere uygulanmıştır. Takımların lig boyunca attıkları gol sayıları üzerinden hesaplanan ofansif ve defansif güç değerlerinin ve bu güç değerleri ile gelecek maçta atması beklenen ortalama gol sayısı bulunarak yapılan Karaoğlu'nun (2015) çalışmasına göre Avrupa'da yer alan 16 futbol ligi değerlendirilmiş ve makine öğrenimi yöntemlerinden Naive Bayes, Multilayer Perceptron, LogitBoost, BayesNet, Karar Ağacı, ZeroR ve C4.5 algoritmaları kullanılarak maç sonucu

tahmin edilmiştir. Yapılan çalışma sonucunda, değerlendirilen 16 ligin 11 tanesinde karar ağacı algoritması en yüksek başarıyı elde etmiştir. Bu 16 ligden 8'inde ZeroR, 6'sında BayesNet, 3'ünde Multilayer Perceptron ve 1 ligde ise LogitBoost algoritmaları en başarılı sonuçları vermişlerdir. NaiveBayes ve C4.5 algoritmaları ise değerlendirilen hiçbir futbol liginde başarılı olamamıştır.

Coşkuner ve diğerlerinin (2020) yaptığı Türkiye Süper Ligi'ndeki oyun içi değişkenlerin analizi çalışmasında, 2017-2018 sezonu için Türkiye Süper Ligi'nde yer alan 18 takım ve 612 karşılaşma incelenmiştir. Çalışmada atılan gol sayısı, toplam şut, isabetli şut, isabetsiz şut, top ile buluşma sayısı, topa sahip olma yüzdesi, toplam pas, isabetli pas, isabetli pas yüzdesi ve saha (ev sahibi/deplasman) değişkenleri ele alınmıştır. Bağımlı ve bağımsız değişkenler arasındaki ilişkiyi modellemek için çok değişkenli lojistik regresyon kullanılmıştır. 2017-2018 Türkiye Süper Ligi'nde futbol maçlarının sonucu %65 oranda başarılı bir şekilde tahmin edilmiştir.

Prasetio ve Harlili (2016), Premier Ligi maç sonuçlarını lojistik regresyon yöntemini kullanarak incelemişlerdir. Çalışmada bağımsız değişken olarak ev sahibi takımın defans değeri ile rakip takımın defans değeri kullanılmıştır. Model, 2010-2011 sezonundan 2015-2016 sezonuna kadar olan eğitim verilerinin varyasyonları kullanılarak oluşturulmuştur. 2015-2016 verisi üzerinde de test edilmiştir. 4 ayrı veri seti ile oluşturulan modelin tahmin sonuçlarının doğruluk yüzdeleri ise %68.005 ile %69.513 arasında bulunmuştur.

Hucalijuk ve Rakipovic (2011), Şampiyonlar Ligi maç sonuçlarını Naive Bayes, Bayessian net, LogitBoost, k-en yakın komşu, rassal orman ve yapay sinir ağları algoritmalarını kullanarak incelemişlerdir. Çalışmada son altı maçta elde edilen sonuçlara göre gösterilen mevcut takım formları, oyunu oynayan takımların bir önceki karşılaşmasının sonucu, sıralamadaki mevcut konumu, takımda sakatlanan oyuncu sayısı, maç başına atılan ve alınan ortalama gol sayısı gibi başlıklar altında toplamda 20 değişken kullanılmıştır. Ayrıca, her bir takımın kalitesinin bu alandaki uzmanlar tarafından öznel olarak değerlendirilmesine ek olarak, belirlenen değişkenlerin tümünü içeren başka bir veri seti oluşturulmuştur. Yapılan çalışma sonucunda çoğunlukla temel veri seti, uzmanlar tarafından oluşturulan veri setinden daha yüksek başarı göstermiştir. En başarılı algoritma ise %68.8 tahmin doğruluğu ile yapay sinir ağları olmuştur.

Bu çalışmanın temel amacı 2019-2021 yılları arasında olan 21 futbol takımının 526 karşılaşması incelenerek maçı kazanma ya da kaybetme durumlarının o takımın ofansif & defansif güçleri, kırmızı & sarı kart durumları, yabancı futbolcu sayısı gibi değişkenler ile modellenmesinin yapılması olarak belirlenmiştir. Bu amaçla çalışma kapsamında denetimli öğrenme yöntemlerinden, lojistik regresyon ve karar ağaçları algoritmalarından CART (Classification and Regression Tree), QUEST (Quick, Unbiased, Efficient Statistical Tree) ve CHAID (Chi-Squared Automatic Interaction Detector) yöntemleri Türkiye Süper Ligi verileri kullanılarak incelenmiştir.

YÖNTEM

Araştırma Modeli

Bu çalışma, nicel araştırma çalışma tasarımlarından gözlemsel çalışmaya bağlı tanımlayıcı istatistiksel modeli ile yapılmıştır.

Örneklem ve Veri Toplama Aracı

Bu çalışmada Türkiye Futbol Federasyonu (TFF) Süper Ligi 2019-2020 ve 2020-2021 sezonu verileri kullanılmıştır. Sezonda yer alan 21 takımın yaptığı tüm maçlardan kazanılan ve kaybedilen karşılaşmalar dikkate alınarak toplamda 526 veri incelenmiştir. Verilerin hepsi TFF resmi internet sitesinde yer alan bilgilerden alınmıştır (TFF, 2021).

Verilerin Analizi

Spor verisinde sürekli değişkenler için tanımlayıcı istatistikler kullanılırken kesikli değişkenler için frekans tabloları kullanılmıştır. Y bağımlı rassal değişkenimiz maçın kaybedilmesi veya kazanılması olup ikili bir değişkendir. X bağımsız rassal değişkenlerimiz ise takımın ofansif gücü, rakibin defansif gücü, rakibin kırmızı kart alıp almaması, rakibin sarı kart alıp almaması, takımın kırmızı kart alıp almaması ve takımdaki yabancı oyuncu sayısıdır. Y rassal değişkeni ile kategorik olan X rassal değişkenleri arasındaki ilişkiyi tespit edebilmek için ki-kare test istatistiği kullanılmıştır. Ayrıca Y ile tüm X'ler arasında korelasyon matrisine (Spearman) bakılmıştır. İleri analizler için makine öğrenim yöntemleri kullanılmıştır.

Araştırmada Kullanılan İstatistiksel Yöntemler

Çalışma kapsamında makine öğrenim yöntemlerinden lojistik regresyon yöntemi, karar ağacı algoritmalarından CART, CHAID ve QUEST yöntemleri kullanılmıştır.

Lojistik regresyon, kategorik olan bağımlı değişkenin diğer bağımsız değişkenlerle neden sonuç ilişkisini belirlemede yararlanılan bir modelleme yöntemidir. Bu modelleme yöntemi ile bağımlı değişkenin tahmini değerleri olasılık olarak hesaplanarak sınıflandırma yapılır. Bu analiz sayesinde normallik, doğrusallık, süreklilik ve eşvaryanslık gibi varsayımları sağlamadan model kurulmaktadır (Tabachnick ve Fidell, 1996). Regresyon çözümlemesinde bağımlı değişken iki değer aldığında Bernoulli değişkeni ile ifade edilir. Bu nedenle koşullu dağılımın beklenen değeri olasılık değerine eşittir ($E(y|x) = p$). Bağımlı değişkenin koşullu dağılımının beklenen değerine ait regresyon denklemi 0 ve 1 arasında değer alacak şekilde kurulur ($y = b_0 + b_1 x_1$). Yapılan bu lojit dönüşüm ile lojistik fonksiyonu doğrusal hale getirilir.

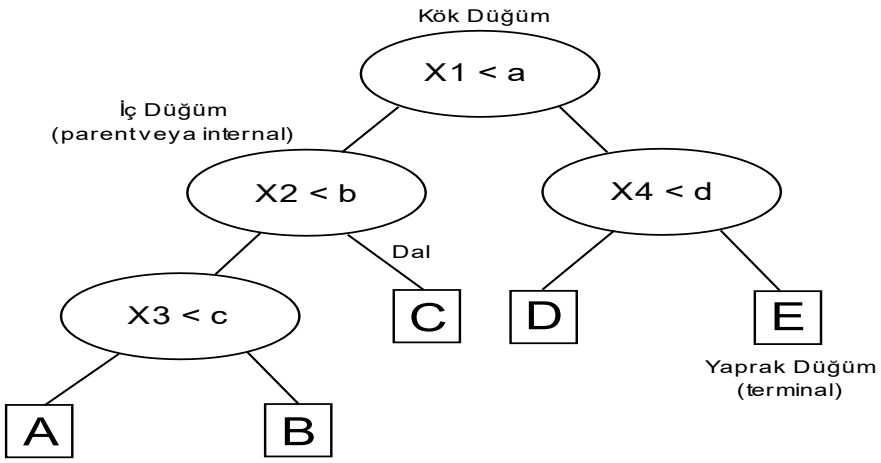
$$E(x|y) = p = \frac{1}{1+e^{-y}} \quad (1)$$

$$\ln(\frac{p}{1-p}) = = b_0 + b_1 x_1 \quad (2)$$

Lojistik regresyon modelinde "en çok olabilirlik" yöntemi ile değişkenlere ait katsayıların kestirimi yapılır. Kestirilen katsayıların anlamlılığı ise "olabilirlik oran testi" veya "Wald testi" ile yapılır. Katsayıların yorumu için odds ve odds oranından (OR) yararlanılır. Denklem 2'de $(\frac{p}{1-p})$ oranı odds olarak ifade edilir.

Lojistik regresyon analizinde odds, bir olayın olma olasılığının, o olayın olmama olasılığına oranı olarak gösterilmektedir (Şenel ve Alatlı, 2014). Lojistik regresyon modelinde katsayıların yorumu doğrusal regresyon modelinden farklıdır. Bunun sebebi lojistik regresyon modelindeki çıktı sonucunun bir olasılık değeri olmasıdır. Katsayılar olasılık değerini doğrusal olarak etkilemez. Ancak log(odds) fonksiyonu katsayıların doğrusal fonksiyonudur. Odds oranı ise ilgilenilen olayın odds değerinin referans olarak belirlenen olayın odds'una oranıdır. Odds oranı üç bant arasında hesaplanabilir. Bu oran 0 ve 1 değeri arasında olursa bağımsız değişkenin bağımlı değişkeni için "koruyucu" olduğu, tam olarak 1'e eşit olduğunda değişkenler arasında bir fark olmadığı ve son olarak 1'den büyük olursa değişkenler arasında fark olduğu ve bu farkın matematiksel olarak bir kat ile ifade edilebileceğidir (Yavuz ve Vupa Çilengiroğlu, 2020).

Sınıflandırma ve karar ağacı yöntemlerinden CART algoritması, 1984 yılında Breiman ve arkadaşları tarafından Morgan ve Sonquist'in AID (Automatic Interaction Detection) adlı karar ağacı algoritmasının devamı kapsamında önerilmiştir (Breiman, Freidman, Olshen ve Stone, 1984). CART algoritması, sınıflandırma ve regresyon problemlerinde kullanılır. Karar ağaçları çok sayıda veriyi belirli bölme kriterlerine göre daha küçük veri kümelerine ayıran bir yöntemdir. Ağaç yapısı kök, iç düğüm (parent veya internal), dal ve yaprak düğümlerden (terminal) oluşmaktadır. Kök düğüm veri setindeki tüm elemanları içerir. Daha sonra ağaç dallanarak yeni düğümler oluşturur. Her dalda yeni bir düğüm veya artık bölünme olmayacaksa yapraklar oluşur (Şekil 1).



Şekil 1. Karar ağacı kök, düğüm ve dal yapıları

Karar ağacı algoritmalarından CART, ikili olarak tekrarlı bir biçimde bölünen bir yapıya sahiptir. CART'ın "büyük ağacı oluşturma", "budama" ve "en iyi (optimum) ağacı tespiti" başlığı altında üç temel adımı vardır. CART ağacı, kuruluş aşamasında herhangi bir durma kuralı olmaksızın sürekli olarak bölünerek büyük ağacı oluşturmaktadır (Zheng, Chen, Han, Zhao ve Ma, 2009). Büyük ağaç oluşturulurken dallanma kriteri olarak Gini indeksinden veya Twoing algoritmasından yararlanılır. Kategorik bağımlı değişkenler için Gini indeksi kullanılırken sıralayıcı bağımlı değişkenler için Twoing algoritması kullanılır (Oğuzlar, 2010). Herhangi bir düğümde veriler kategorilere eşit biçimde dağıldığında Gini indeksi maksimum

değerini alır ve bölümleme eşit biçimde olur. Bir düğümde veriler aynı kategoriye ait bir biçimde yer aldığında ise Gini indeksi 0 değerini alır ve bölümlenme olmaz. Twoing algoritması ise, Gini indeksine göre daha dengeli bir yapı sunar. Bunun nedeni ise her defasında kök ve iç düğümlerin %50′ sini içermeye çalışmasıdır (SPSS, 2001). Dallanma sırasında artık yeni bir bölünmenin gerçekleşmeyeceği durumda uçtan kök düğüme doğru budama işlemi başlar. Budama işlemi ile ağacın sınıflandırma doğruluğuna etkisi olmayan veya katkı sağlamayan kısımları çıkarılır. Bu sayede az karmaşıklığa sahip ve daha anlaşılabilir bir ağaç elde edilmiş olur. Olası en iyi karar ağacı, her budama işlemi sonrası bağımsızca seçilmiş bir test verisi ile değerlendirme yapılarak tespit edilmeye çalışılır (Wu ve Kumar, 2009). Bu şekilde budama işleminden sonra en iyi (optimum) ağaç yapısına ulaşılmış olunur.

Bir diğer karar ağacı algoritması olan CHAID algoritması ilk olarak 1975 yılında Kass tarafından tanıtılmıştır. CHAID algoritması CART algoritması ile benzerlikler gösterirken, iki algoritmayı birbirinde ayıran en önemli özellik ise ağacı dallara ayırırken kullandıkları bölünme yöntemidir. CART algoritması bölünme için Gini veya Twoing kullanırken, CHAID algoritması ki-kare testini uygulayan bir yöntemdir. Bu algoritma ile hem kategorik hem de sürekli değişkenler kullanılmasının yanında ağaçta her düğümün ikiden fazla alt grubu bölünebilme özelliğinin varlığı bu algoritmayı popüler bir hale getirmiştir (Díaz-Pérez ve Cejas, 2016).

QUEST algoritması 1997 yılında Loh and Shih tarafından geliştirilmiştir (Çalış ve Kayapınar, 2014). Bu algoritmanın en önemli özelliği CART ve CHAID algoritmalarının aksine, dallanma sırasında optimum bölünmeyi verecek bağımsız değişkene ve optimum bölünmenin sağlanacağı noktanın karar verilmesine ayrı zamanlar ayırmasıdır. Bu durum da modelin daha hızlı sonuç vermesini sağlamaktadır. Bir dezavantajı ise karar değişkeninin sürekli olduğu durumlarda kullanılamamasıdır. QUEST algoritması bölünme için eğer bağımsız değişken kesikli ise ki-kare, sürekli ise F testini kullanmaktadır (Kuzey, 2012).

BULGULAR

Bu çalışmada Türkiye Futbol Federasyonu 2019-2020 ve 2020-2021 Süper Ligi maç sonuçlarına göre kazanma ve kaybetme durumlarının hangi değişkenlere bağlı olarak değiştiği makine öğrenme algoritmaları kullanılarak gösterilmiştir. Türkiye Futbol Federasyonu 2019-2020 ile 2020-2021 Süper Ligi maç sonuçları ile toplanan veri setinde takımların attıkları gol sayıları, kazanan takım, görülen kart sayıları, yabancı oyuncu sayıları, saha durumu, hava durumu ve sıcaklığı, maçın saati, mevsim, maçın oynandığı hafta gibi birçok veri toplanmıştır.

Veri seçimi için literatürde yer alan gol sayıları üzerinden hesaplanan ofansif ve defansif güç değerleri, toplam şut, isabetli şut, isabetsiz şut, topla buluşma sayısı, topa sahip olma yüzdesi, toplam pas, isabetli pas, isabetli pas yüzdesi, saha (ev sahibi/deplasman), mevcut takım formları, oyunu oynayan takımların bir önceki karşılaşmasının sonucu, sıralamadaki konumu ve takımda sakatlanan oyuncu sayısı gibi değişkenlerin daha önce kullanıldığı tespit edilmiştir. Literatürdeki bu değişkenlerden farklı olarak çalışmamızda görülen kart sayıları, yabancı oyuncu sayıları, hava durumu, hava sıcaklığı, maçın saati ve mevsim gibi değişkenler

toplanmıştır. Bu değişkenlerin takımın kazanmasına veya kaybetmesine olan etkisini incelemek için ki kare p-değerlerine ve korelasyon matrisine bakılmıştır. Analiz sonucunda hava durumu, hava sıcaklığı, maç saati ve mevsim değişkenleri ki-kare sonucuna göre anlamlı olmadığından karar ağacı modellerine alınmamıştır. Bununla birlikte maçın kazanılması veya kaybedilmesine etkisi olan kart (kırmızı veya sarı) görme durumu ve yabancı oyuncu sayısı değişkenleri anlamlı bulunmuş ve modellere eklenmiştir. Takım ya da rakibin kart görme durumu da etkili bir performans kriteridir. Çünkü takımların kart görme durumları saha içindeki hakimiyetlerini değiştireceği için takımların kazanma veya kaybetme sonucunu etkilemektedir. Anlamlı bulunan değişkenler içerisinde görülen kart sayıları kategorik hale getirilmiştir (kart gördü ise 1, kart görmedi ise 0). Yabancı oyuncu sayısı değişkeni 7 ve altı ile 8 ve üzeri şeklinde iki kategoriye ayrılmıştır. Maçta atılan ve yenilen gol sayısı değişkeni kullanılarak Karaoğlu'nun (2015) çalışmadaki yer alan ofansif ve defansif güçler hesaplanmıştır. Bu hesaplama denklem 3 ve denklem 4 ile verilmiştir.

$$O_i = OAG_i \div LOAG \quad (3)$$

$$D_i = OYG_i \div LOAG \quad (4)$$

O_i - i takımının ofansif gücü
D_i - i takımının defansif gücü
OAG_i - i takımının ortalama attığı gol sayısı
OYG_i - i takımının ortalama yediği gol sayısı
$LOAG$ - Ligde bir takımın attığı/yediği ortalama gol

$$LOAG = TG \div (TS * N) \quad (5)$$

LOAG - Ligde bir takımın attığı ortalama gol
TG - Ligde atılan toplam gol
TS - Ligdeki takım sayısı
N - Değerlendirmeye alınan hafta sayısı

İlgili hesaplamalar için sezondaki tüm maçlar dikkate alınmıştır. Her takımın ofansif ve defansif gücü ile elde edilen yeni değişken değerleri ortalama üstü ve ortalama altı olarak kategorik hale getirilmiştir. Kategorik hale gelen bu değişkenlerden maçtaki ev sahibi takımın ofansif gücü ve rakip takımın defansif gücü değişkenleri modellerde kullanılmıştır. Anlamlı bulunan değişkenler rakip takımın sarı kart görmesi (rakip sarı), rakip takımın kırmızı kart görmesi (rakip kırmızı), kazanan takımın kırmızı kart görmesi (takım kırmızı), kazanan takımın yabancı oyuncu sayısı (yabancı oyuncu), kazanan takımın ofansif gücü (Otakım) ve rakip takımın defansif gücü (Drakip) olarak belirlenmiştir. Bu değişkenlerin frekans, yüzde, ki-kare p değerleri ve Spearman korelasyon katsayıları Tablo 1'de verilmiştir.

Tablo 1. Değişkenlerin tanımlayıcı istatistikleri (Frekans, yüzde, ki-kare p değerleri ve korelasyon katsayıları)

Değişkenler	Kategori	n(%)			p-değeri	Korelasyon Katsayıları
		Kazandı	Kaybetti	Toplam		
		307 (%100)	219 (%100)	526 (%100)		
Rakip Sarı	**Yok**	22 (%7.2)	24 (%11)	46 (%8.7)	0.129	0.066
	Var	285 (%92.8)	195 (%89)	480 (%91.3)		
Takım Kırmızı	**Yok**	283 (%92.2)	179 (%81.7)	64 (%12.2)	0.000	-0.158
	Var	24 (%7.8)	40 (%18.3)	462 (%87.8)		
Rakip Kırmızı	**Yok**	251 (%81.8)	199 (%90.9)	450 (%85.6)	0.003	0.128
	Var	56 (%18.2)	20 (%9.1)	76 (%14.4)		
Yabancı Oyuncu	**7 ve altı**	155 (%50.5)	129 (%58.9)	284 (%54)	0.056	-0.002
	8 ve üzeri	152 (%49.5)	90 (%41.1)	242 (%46)		
Takım Ofansif Gücü	**Ort. Altı**	131 (%42.7)	154 (%70.3)	285 (%54.2)	0.000	0.274
	Ort. Üstü	176 (%57.3)	65 (%29.7)	241 (%45.8)		
Rakip Defansif Gücü	**Ort. Altı**	124 (%40.4)	140 (%63.9)	264 (%50.2)	0.000	0.232
	Ort. Üstü	183 (%59.6)	79 (%36.1)	262 (%49.8)		

"Rakip sarı" değişkeninde sezon boyunca oynanıp kazanılan 307 maçın 22'inde (%7.2) rakip takım hiç sarı kart görmemiştir. Aynı değişken için kaybedilen 219 maçın %11'inde rakip takımın hiç sarı kartı yokken, %89'unda ise en az bir sarı kartı vardır. Aynı şekilde "takım kırmızı" değişkeni için kazanılan maçın %92.2'sinde kırmızı kart yoktur. "Rakip kırmızı" değişkeninde ise bu oran biraz daha düşük olup %81.8 olarak bulunmuştur. "Yabancı oyuncu" değişkeninde "7 ve altı" ile "8 ve üzeri" şeklinde iki kategori olarak belirlenmiştir. Buna göre "7 ve altı" kategorisinde kazanılan oran %50.5 olarak tespit edilmiştir. "Otakım" değişkeninde kazanılan maçların %57.3'ü ortalamanın üstünde iken "Drakip" değişkeninde kazanılan maçların %59.6'sı ortalamanın üstündedir.

Bununla birlikte tüm bu bağımsız değişkenler ile bağımlı değişken olan takımın "kazanma&kaybetme" durumu arasında ilişkinin incelendiği ki-kare testine göre %90 güvenle ilişki tespit edilmiştir (p-değerleri<0.10). Korelasyon katsayılarına bakıldığında rakip sarı kart, rakip kırmızı kart, takımın ofansif gücü ve rakibin defansif gücü ile kazanma&kaybetme değişkeni arasında pozitif yönlü bir ilişki tespit edilmiştir. Takımın kırmızı kart görmesi ile kazanma&kaybetme değişkeni arasında ise negatif yönlü bir ilişki görülmüştür. Ayrıca takım kırmızı, rakip kırmızı, takımın ofansif gücü ve takımın defansif gücü değişkenleri ile kazanma&kaybetme arasında daha güçlü bir ilişki olduğu da tespit edilmiştir.

Değişken analizinden sonra maçın "kazanma&kaybetme" değişkeni ile diğer değişkenler arasında makine öğrenim yöntemlerinden lojistik model ile CART, CHAID ve QUEST algoritmaları ile altı model kurulmuştur. Modeller %80 eğitim %20 test verisi şeklinde ayrılarak kurulmuştur. Eğitim verisi ile model eğitilmiş, test verisi ile ise tahminleme yapılmıştır. Bu modellere ait bağımsız değişkenler Tablo 2'de gösterilmiştir.

Tablo 2. Model değişkenleri

Modeller	Değişkenler			
	1. Değişken	2. Değişken	3. Değişken	4. Değişken
Model 1	Takım Kırmızı	Otakım	Drakip	
Model 2	Rakip Sarı	Otakım	Drakip	
Model 3	Rakip Kırmızı	Otakım	Drakip	
Model 4	Yabancı Oyuncu	Otakım	Drakip	
Model 5	Takım Kırmızı	Rakip Kırmızı	Yabancı Oyuncu	
Model 6	Takım Kırmızı	Yabancı Oyuncu	Otakım	Drakip

Karar ağacı algoritmalarından CART yöntemi tüm modellerde kullanılırken diğer yöntemler farklı modellerde Tablo 3'teki gibi kullanılmıştır.

Tablo 3. Modeller ve kullanılan yöntemler

Model	CART	CHAID	QUEST	LOJ. REG
Model 1	√	√	√	√
Model 2	√	X	√	X
Model 3	√	√	√	√
Model 4	√	√	√	√
Model 5	√	√	√	√
Model 6	√	X	X	√

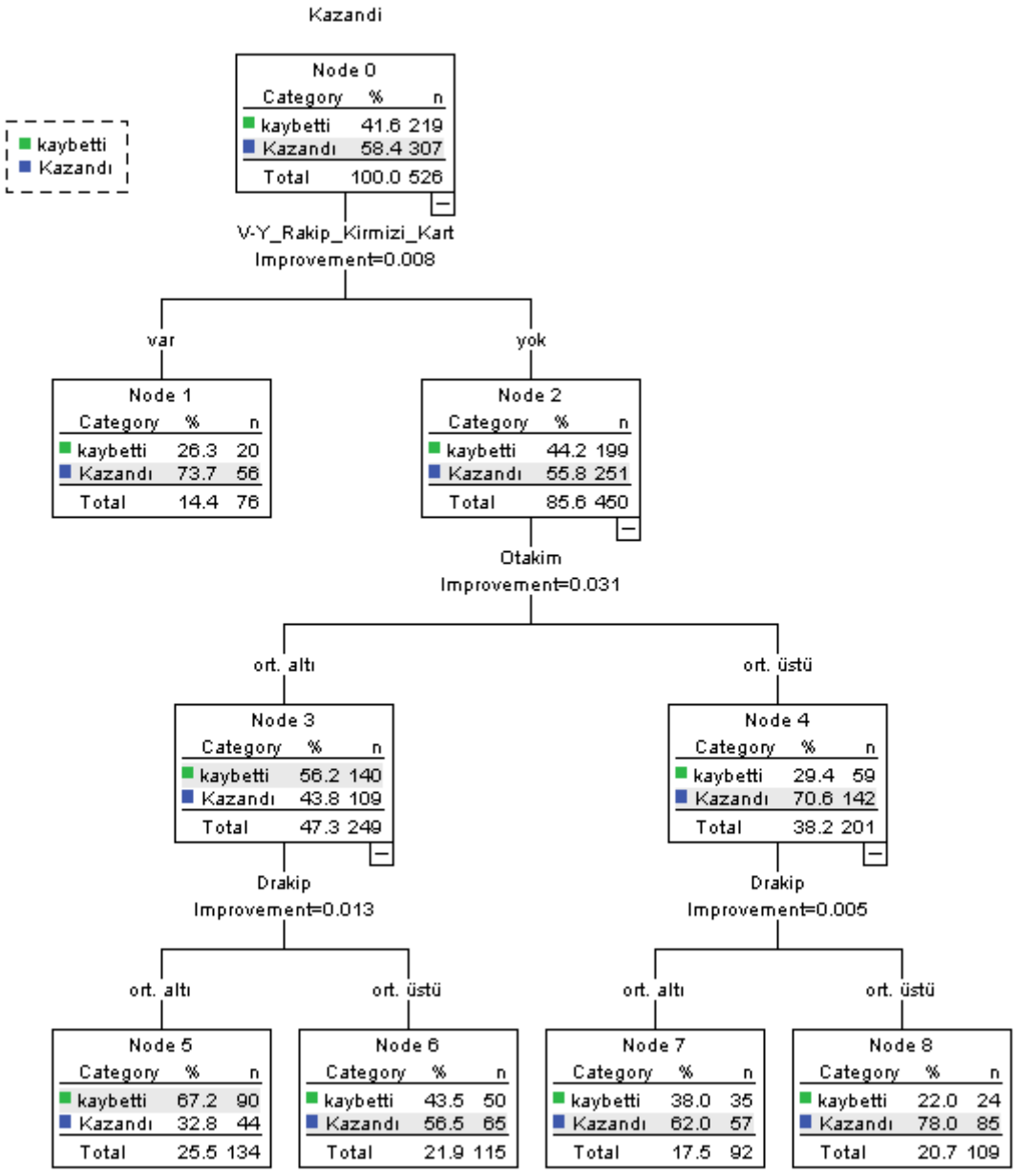
%80 eğitim %20 test verisi ile kurulan bu modellerde karşılaştırma yapabilmek için tekrarlı ölçümler yapılmıştır. Farklı örneklemler seçilerek 30 kez tekrarlı bir şekilde yapılan modellere ait doğruluk, duyarlılık, seçicilik ve F-skor değerleri hesaplanmıştır. Hesaplanan bu değerlerin ortalamaları alınarak oluşturulan Tablo 4'te en iyi modelin karar ağacı algoritmaları ile Model-3 olduğu tespit edilmiştir.

Tablo 4. %80 eğitim ve %20 test verisinden oluşan model sonuçları

Model	Doğruluk	Duyarlılık	Seçicilik	F-skor
CART Model-1	0.677^{*2}	0.774	0.626^{*3}	0.736^{*6}
CHAID Model-1	0.674^{*4}	0.774	0.615^{*6}	0.737^{*5}
QUEST Model-1	0.686^{*1}	0.766	0.622^{*5}	0.743^{*4}
Lojistik Model-1	0.659	0.742	0.614	0.713
CART Model-2	0.648	0.791	0.608	0.722
QUEST Model-2	0.646	0.787	0.605	0.720
CART Model-3	0.676^{*3}	0.849^{*6}	0.648^{*2}	0.757^{*1}
CHAID Model-3	0.672^{*5}	0.850^{*5}	0.66^{*1}	0.752^{*3}
QUEST Model-3	0.671^{*6}	0.857^{*4}	0.66^{*1}	0.754^{*2}
Lojistik Model-3	0.603	0.556	0.516	0.620
CART Model-4	0.620	0.721	0.571	0.683
CHAID Model-4	0.615	0.719	0.565	0.680
QUEST Model-4	0.602	0.699	0.551	0.661
Lojistik Model-4	0.638	0.628	0.560	0.667
CART Model-5	0.592	0.866^{*3}	0.594	0.702
CHAID Model-5	0.618	0.896^{*1}	0.625^{*4}	0.731
QUEST Model-5	0.594	0.887^{*2}	0.562	0.716
LOJİSTİK Model-5	0.590	0.571	0.513	0.614
CART Model-6	0.669	0.769	0.608	0.732
Lojistik Model-6	0.659	0.674	0.596	0.692

(*1 en iyi değer) (*2 en iyi ikinci değer) (*3 en iyi üçüncü değer)
(*4 en iyi dördüncü değer) (*5 en iyi beşinci değer) (*6 en iyi altıncı değer)

Model-3'e ait karar ağacı da Şekil 2'deki gibi gösterilmiştir. CART, CHAID ve QUEST algoritmaları için kurulan modeller aynıdır. CART, CHAID ve QUEST modelleri için değişkenler, rakip takımın kırmızı karta sahip olup&olmaması durumu ile takımın ofansif ve defansif güçleri olarak belirlenmiştir. Bu algoritmalar için toplam 9 düğüm vardır bu 9 düğümün 5 tanesi terminal düğümdür. Üç tane dallanma gerçekleşmiştir. Bu algoritmalara göre rakip takımın kırmızı kartı olduğu durumda takımın kazanma oranı %73.7 olarak bulunmuştur. Rakip takımın kırmızı kartı yok iken takımın ofansif gücü ortalamamın üzerinde ve rakibin defansif gücü de ortalamanın üzerinde ise o takımın kazanma oranı %78'dir. Aynı şekilde rakip takımın kırmızı kartı yok iken takımın ofansif gücü ortalamamın altında ve rakibin defansif gücü de ortalamanın altında ise o takımın kazanma oranı %32.8 olarak elde edilmiştir (Şekil 2).



Şekil 2. CART, CHAID ve QUEST modellerine ait karar ağacı

TARTIŞMA VE SONUÇ

Günümüzde olduğu gibi gelecekte de futbol maçları ve maç sonuçlarının tahmini gibi çalışmalar devam edecektir. Futbol verileri ile yapılan bu çalışmada makine öğrenim yöntemleri (lojistik, CART, CHAID ve QUEST) kullanılarak maç sonucunun tahmini yapılmıştır. Model performansları duyarlılık, seçicilik, doğruluk ve F skoruna göre hesaplanmış ve bu modellerden veri setinde en iyi sonucu veren algoritmanın CART olduğu tespit edilmiştir. Ayrıca model kurmada önemli olan bir diğer konu ise bağımsız değişkenlerin seçimidir. Bağımsız değişkenlerin farklı kombinasyonları ile kurulan 6 ayrı CART modelinde en başarılı olan model, rakibin kırmızı karta sahip olup&olmaması, takımın ofansif gücü ve rakibin defansif gücü değişkenlerinin kullanıldığı modeldir. Modelin doğruluk yüzdesi %67.6 olarak bulunmuştur. Bu çalışmanın literatüre katkısı, maç sırasında görülen kart değişkenin modele dahil edilmesi ve aslında maç sonucunu etkileyen önemli bir değişken olduğunun tespit edilmiş olmasıdır.

Bağımsız değişken seçiminde maç sırasında görülen kartların maç sonucuna etkili olduğu görülmüştür. Başarılı olan modelde yer alan rakibin kırmızı kart görme durumu etkili bir değişkendir. Rakibin kırmızı kart görmesi o takımın performansını ve maça hakimiyetini azaltacağından ev sahibi takımın kazanma olasılığını artıracaktır.

Kurulan CART modelinde olası sonuçlarla da tahmin yapılmıştır. Örneğin Şekil 2'ye göre rakip takımın maç içerisinde kırmızı kart görmesi, rakip takımın maça daha az hakim oynamasına neden olacağından ev sahibi takımın kazanma yüzdesini artırmıştır (%73.7). Aynı zamanda rakip takımın kırmızı kartı yoksa hakimiyetini kaybetme olasılığı çok yüksek olmayacağından ev sahibi takımın kazanma yüzdesi bir önceki duruma göre daha düşük bulunmuştur (%55.8). Sonuç olarak maç sırasında rakibin kart görmesi durumu kazanma ve kaybetmeye etkili olacağı için maç stratejisi yapılırken bu durum dikkate alınmalıdır.

Literatürde futbol verisi kullanılarak maç sonucunun tahminlenmesi üzerine yapılan çalışmaların sonuçları değerlendirildiğinde; Karaoğlu (2015), 16 ligi değerlendirdiği çalışmasında ofansif ve defansif güçler dikkate alınarak modeller kurmuş ve en başarılı olarak karar ağacı algoritmasını elde etmiştir. Modelinin doğruluk yüzdesi %50.86 ile Türkiye 1. Ligi olmuştur. Bu çalışmada Karaoğlu'nun çalışmasına ek olarak takımın ofansif ve defansif gücü kategorik hale getirilmiş ve yanına takımın kırmızı kart görmesi değişkeni eklenmiştir. Sonuç olarak model ileri taşınmış ve doğruluk yüzdesi %67.6 olarak artırılmıştır. Ayrıca Karaoğlu bu çalışmasında ofansif ve defansif güç değerlerinin kazanıp&kaybetme değişkeni arasındaki ilişkisini ki-kare testi ile yapmış ve anlamlı bulmuştur. Bu çalışmada da ofansif ve defansif güç değerleri ki-kare testi ile anlamlı bulunmuştur (Tablo1).

Coşkuner ve diğerleri (2020), Süper Lig maçlarını ele almış ve çalışmasında kullandığı lojistik regresyon modeli ile %65'lik bir doğruluk yüzdesi elde etmiştir. Bu çalışmada kullanılan modellerden biri olan lojistik regresyon modeli ise %65.9'luk bir doğruluk yüzdesine sahiptir.

Prasetio ve Harlili (2016), Premier Ligi maç sonuçlarını tahmin ederken 2010 ve 2016 yılları arasındaki sezonları dikkate almışlar ve sadece bir tane lojistik regresyon modeli kurmuşlardır.

Elde edilen sonuçta veri setlerinin doğruluk yüzdesi %68-69 civarındadır. Bu çalışmadaki tüm lojistik modellerin doğruluk yüzdesi %60-66 arasında bulunmuştur.

Hucalijuk ve Rakipovic (2011), benzer bir futbol veri seti kullanarak Naive Bayes, Bayessian net, LogitBoost, k-en yakın komşu, rassal orman ve yapay sinir ağları algoritmalarını kullanarak modeller oluşturmuş ve en iyi sonucu yapay sinir ağları algoritması ile elde etmişlerdir. Bu algoritmanın doğruluk yüzdesi ise %68.8 olarak tespit edilmiştir. Çalışmanın verisi, yapısı ve yöntemleri bu çalışmadan farklı olmasına rağmen elde edilen doğruluk yüzdeleri benzer olarak bulunmuştur.

Çalışmanın kısıtlarından biri sadece iki sezon alınmasıdır. Sezon sayısının artırılması ile veri seti genişletilip modele dahil edilemeyen (hava durumu, mevsim, maç saati, oynanan hafta vb.) bağımsız değişkenler modele dahil edilip algoritmalar yeniden çalıştırılır ise elde edilen doğruluk yüzdelerinin artacağı öngörülmektedir. Ayrıca daha sonraki çalışmalarda genişletilmiş veri seti ile farklı makine öğrenim yöntemleri kullanılarak modellerin performans ölçütlerinin değerlendirilmesinin yapılması da planlanmaktadır. Böylece maç sonucunun tahmininde değişkenlerin yüzde veya oran olarak modelde nasıl ve ne kadar bir paya sahip olduğu basit bir şekilde kolaylıkla anlaşılabilecektir.

Çıkar Çatışması: Bu çalışma ile ilgili herhangi bir finansal açıklama veya sorumluluk reddi beyanı ve de yazarlar için çıkar çatışması bulunmamaktadır.

Araştırmacıların Katkı Oranı Beyanı: Araştırma Dizaynı- DT; ÖVÇ, İstatistik analiz- DT; ÖVÇ, Makalenin hazırlanması- DT; ÖVÇ, Verilerin Toplanması- DT.

Etik Kurul: Çalışma kapsamında tüm etik konular yazarlar tarafından dikkate alınmıştır. Verilerin toplanması etik kurul raporuna dahil değildir.

KAYNAKLAR

Breiman, L., Freidman, J.H., Olshen, R. A. & Stone, C.J. (1984). *Classification and regression trees (1. baskı)*. Boca Raton, USA: Taylor&Francis Group, Chapman and Hall.

Carling, C., Williams, A. M. & Reilly, T. (2005). *Handbook of soccer match analysis: A Systematic approach to improving performance* (1. baskı). New York, USA: Routledge.

Coşkuner, Z., Büyükçelebi, H. ve Kurak, K. (2020). Türkiye Süper Ligi'ndeki oyun içi değişkenlerin analizi. *Germenica Beden Eğitimi ve Spor Bilimleri Dergisi, 1*(1), 46-54.

Çalış, A. ve Kayapınar, S. (2014). Veri madenciliğinde karar ağacı algoritmaları ile bilgisayar ve internet güvenliği üzerine bir uygulama. *Endüstri Mühendisliği Dergisi, 25*(3-4), 2-19.

Díaz-Pérez, F. & Cejas, B. (2016). CHAID algorithm as an appropriate analytical method for tourism market segmentation. *Journal of Destination Marketing & Management, 5*(3), 275-282. https://doi.org/10.1016/j.jdmm.2016.01.006

Han, J., Kamber, M. & Pei, J. (2012). *Data mining: concepts and techniques (3. baskı).* MA, USA: Morgan Kaufmann Publishers.

Hucaljuk, J., & Rakipović, A. (2011, May). Predicting football scores using machine learning techniques. In *2011 Proceedings of the 34th International Convention MIPRO (pp. 1623-1627).* IEEE.

Karaoğlu, B. (2015). Makine öğrenmesi ile spor karşılaşmalarının modellenmesi. *EMO Bilimsel Dergi, 5*(9), 1-5.

Kuzey, C. (2012). *Veri madenciliğinde destek vektör makinaları ve karar ağaçları yöntemlerini kullanarak bilgi çalışanlarının kurum performansı üzerine etkisinin ölçülmesi ve bir uygulama*. Yayımlanmamış Doktora Tezi, İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul.

Michie, D., Spiegelhalter, D.J. & Taylor, C.C. (1994). *Machine learning, neural and statistical classification (1. baskı)*. New York, USA: Ellis Horwood series, Prentice Hall.

Oğuzlar, A. (2010). CART analizi ile hanehalkı işgücü anketi sonuçlarının özetlenmesi. *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi, 18*(3-4), 79-90.

Özekes, S. (2003). Veri madenciliği modelleri ve uygulama alanları. *İstanbul Ticaret Üniversitesi Dergisi, 2*(3), 65-82.

Prasetio, D. & Harlili, D. (2016). Predicting football match results with logistic regression. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)* (pp. 1-5). IEEE.

SPSS. (2001). *Statistical package for the social sciences.* USA: SPSS Inc.

Şenel, S. ve Alatlı, B. (2014). Lojistik regresyon analizinin kullanıldığı makaleler üzerine bir inceleme. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 5*(1), 35-52.

Tabachnick, B. G. & Fidell, L. S. (1996). *Using multivariate statistics (3. baskı)*. New York, USA: Harper Collins College Publishers.

TFF. (2021). Türkiye Futbol Federasyonu, https://www.tff.org/, Son Erişim Tarihi: 3 Ağustos 2021.

Wu, X. & Kumar, V. (2009). *CART: Classification and regression trees, top ten algorithms in data mining (1. baskı)*. New York: Chapman and Hall.

Yavuz, A. ve Vupa Çilengiroğlu, Ö. (2020). Lojistik regresyon ve CART yöntemlerinin tahmin edici performanslarının yaşam memnuniyeti verileri için karşılaştırılması. *Avrupa Bilim ve Teknoloji Dergisi, (18),* 719-727.

Zheng H., Chen L., Han X., Zhao X. & Ma Y. (2009). Classification and regression tree (CART) for analysis of soybean yield variability among fields in Northeast China: The importance of phosphorus application rates under drought conditions. *Agriculture, Ecosystems & Environment, 132*, 98-105.