T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

MAÇ SONUÇLARI İÇİN KLASİK ÇOK DEĞİŞKENLİ YÖNTEMLER VE MAKİNE ÖĞRENMESİ İLE İSTATİSTİKSEL ÖNGÖRÜ TAHMİNİ

Coşkun PARİM

DOKTORA TEZİ

İstatistik Anabilim Dalı

İstatistik Programı

Danışman

Prof. Dr. Ali Hakan BÜYÜKLÜ

Temmuz, 2021

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

MAÇ SONUÇLARI İÇİN KLASİK ÇOK DEĞİŞKENLİ YÖNTEMLER VE MAKİNE ÖĞRENMESİ İLE İSTATİSTİKSEL ÖNGÖRÜ TAHMİNİ

Coşkun PARİM tarafından hazırlanan tez çalışması 25.06.2021 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, İstatistik Programı **DOKTORA TEZİ** olarak kabul edilmiştir.

Prof. Dr. Ali Hakan BÜYÜKLÜ
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Prof. Dr. Ali Hakan BÜYÜKLÜ, Danışman
Yıldız Teknik Üniversitesi _____

Prof. Dr. Filiz KARAMAN, Üye
Yıldız Teknik Üniversitesi _____

Prof. Dr. Münevver TURANLI, Üye
İstanbul Ticaret Üniversitesi _____

Dr. Öğr. Üyesi Doğan YILDIZ, Üye
Yıldız Teknik Üniversitesi _____

Prof. Dr. Ünal Halit ÖZDEN, Üye
İstanbul Teknik Üniversitesi _____

Danışmanım Prof. Dr. Ali Hakan BÜYÜKLÜ sorumluluğunda tarafımca hazırlanan Maç Sonuçları için Klasik Çok Değişkenli Yöntemler ve Makine Öğrenmesi ile İstatistiksel Öngörü Tahmini 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.

Coşkun PARİM

İmza

Bu çalışma, Tübitak 2211-Yurt içi Doktora Burs Programı ile desteklenmiştir.

Aileme

ve

biricik eşime

TEŞEKKÜR

Tez çalışmamda bilgi ve tecrübelerinden yararlandığım ve desteğini esirgemeyen sayın hocam Prof. Dr. Ali Hakan BÜYÜKLÜ'ye, her zaman ailelerinin bir parçası olduğumu hissettiren ve bana mesleğimi sevdiren başta çok değerli hocalarım Prof. Dr. Münevver TURANLI ve Prof. Dr. Ünal Halit ÖZDEN olmak üzere İstanbul Ticaret Üniversitesi İstatistik Bölümü'ne, Yıldız Teknik Üniversitesi İstatistik Bölümü'nde bulunan tüm hocalarıma ve meslektaşlarıma, tezden üretilmiş yayında önemli katkıları bulunan sayın hocam Dr. Öğr. Üyesi Doğan YILDIZ ve değerli meslektaşım Arş. Gör. Mehmet Şamil GÜNEŞ'e teşekkürü bir borç bilirim.

Ayrıca, üzerimde çok büyük emekleri olan ve haklarını asla ödeyemeyeceğim saygıdeğer annem Serpil PARİM ile saygıdeğer babam Halil PARİM'e ve beni benden fazla düşünen, daima yanımda olan ve bana güç veren, can yoldaşım, çok kıymetli eşim Melda GÜNAY PARİM'e minnettarım.

Son olarak, doktora sürecim boyunca maddi destek sağlayan TÜBİTAK 2211 Yurt İçi Lisansüstü Burs Programına teşekkür ederim.

Coşkun PARİM

İÇİNDEKİLER

SİMGE LİSTESİ viii
KISALTMA LİSTESİ ix
ŞEKİL LİSTESİ x
TABLO LİSTESİ xi
ÖZET xii
ABSTRACT xiv
1 GİRİŞ 1
1.1 Literatür Özeti 1
1.2 Tezin Amacı 3
1.3 Hipotez 4
2 HİPOTEZ TESTLERİ VE KLASİK ÇOK DEĞİŞKENLİ İSTATİSTİKSEL YÖNTEMLER 5
2.1 Hipotez Testleri 5
2.1.1 Tek Yönlü Varyans Analizi (ANOVA) 5
2.1.1.1 Etki Büyüklüğü 8
2.1.1.2 Çoklu Karşılaştırma 8
- Tukey Testi 9
2.2 K-ortalamalar Kümeleme 10
2.3 Çok Boyutlu Ölçekleme 11
3 MAKİNE ÖĞRENMESİ 14
3.1 Genel Kavramlar 14
3.1.1 Denetimli ve Denetimsiz Öğrenme 14
3.1.2 Sınıflayıcının Performansının Değerlendirilmesi 15
3.1.3 Holdout Methodu 16
3.1.4 K-Katlı Çapraz Doğrulama 16
3.1.5 Bilgi Kriterleri 17
3.1.6 Hata Kareler Ortalaması 17
3.1.7 Aşırı Uyum 17
3.1.8 Karışıklık Matrisi ve Doğruluk Ölçütleri 18
3.1.9 ROC (Receiver Operating Characteristic) Eğrisi 20
3.2 Karar Ağaçları 20
3.2.1 Sınıflandırma Ağaçları 21

3.2.2 Bölünme Kriterleri 23
3.2.3 Genelleştirilmiş Gini İndeksi 24
3.3 Destek Vektör Makineleri 25
3.2.4 Ayrılabilir Sınıf Dağılımları 25
3.2.5 Çakışan Sınıf Dağılımları 28
3.2.6 Çekirdek (Kernel) Fonksiyonları 31
3.4 Yapay Sinir Ağları 31
3.4.1 Bayes Sinir Ağları ile Sınıflandırma 33
4 UYGULAMA **36**
4.1 Örneklem ve Değişkenler 36
4.2 İstatistiksel Analizler 37
5 SONUÇ VE ÖNERİLER **58**
KAYNAKÇA **61**
TEZDEN ÜRETİLMİŞ YAYINLAR **67**

SİMGE LİSTESİ

κ	Çekirdek (kernel) fonksiyonu
λ	Düzenleme parametresi
γ	Etkin parametre sayısı
ξ	Gevşek (Slack) değişken
G(p)	Gini indeksi
L(x)	Lagrangian fonksiyonu
ϕ(x)	Sabit bir özellik uzay dönüşümü

KISALTMA LİSTESİ

AIC	Akaike Bilgi Kriteri (Akakike Information Criteria)
ANOVA	Analysis of Variance
BIC	Bayesyen Bilgi Kriteri (Bayesian/Schwarz Information Criteria)
ÇBÖ	Çok Boyutlu Ölçekleme
DO	Doğruluk Oranı
DVM	Destek Vektör Makineleri
HKO	Hata Kareler Ortalaması
KO	Kareler Ortalaması
KT	Kareler Toplamı
ROC	Receiver Operating Characteristic
SD	Serbestlik Derecesi
YSA	Yapay Sinir Ağları

ŞEKİL LİSTESİ

Şekil 3.1 Aşırı uyum örnek görseli 18
Şekil 3.2 Örnek karar ağacı modeli 22
Şekil 3.3 Marj kavramının görsel olarak gösterilmesi 26
Şekil 3.4 Gevşek değişkenlerin gösterimi 29
Şekil 3.5 Çok katmanlı ve ileri beslemeli yapay sinir ağları örneği 32
Şekil 4.1 Maç sonucuna göre performans göstergelerinin radar grafiği 42
Şekil 4.2 Zayıf rakiplere karşı oynayan takımlar için ÇBÖ analizi sonuç grafiği .43
Şekil 4.3 Daha güçlü rakiplere karşı oynayan takımlar için ÇBÖ analizi sonuç grafiği 44
Şekil 4.4 Dengeli rakiplere karşı oynayan takımlar için ÇBÖ analizi sonuç grafiği 45
Şekil 4.6 Daha güçlü rakiplere karşı oynayan takımlar için karar ağacı 47
Şekil 4.7 Dengeli rakiplere karşı oynayan takımlar için karar ağacı 48
Şekil 4.8 Tek gizli katmanlı ve üç nöronlu YSA modeli için roc eğrisi (3 ihtimalli maç) 53
Şekil 4.9 İki gizli katmanlı ve 3-3 nöronlu YSA modeli için roc eğrisi (üç ihtimalli maç) 55
Şekil 4.10 Tek gizli katmanlı ve 2 nöronlu YSA modeli için roc eğrisi (iki ihtimalli maç) 56
Şekil 4.11 İki gizli katmanlı ve 1-1 nöronlu YSA modeli için roc eğrisi (iki ihtimalli maç) 57

TABLO LİSTESİ

Tablo 2.1 Anova tablosu 7
Tablo 2.2. Stress değerlerinin uyum iyiliği karşılıkları 12
Tablo 3.1 Karışıklık matris örneği 19
Tablo 3.2 Örnek veri seti 21
Tablo 4.1 Çalışmada kullanılan değişkenler ve açıklamaları 37
Tablo 4.2 Maç sonucuna göre tek yönlü ANOVA ve Tukey HSD test sonuçları...39
Tablo 4.3 Üç ihtimalli maç sonucu için destek vektör makineleri sonuçları 50
Tablo 4.4 Üç ihtimalli maç sonucu için kategori bazında doğruluk oranları.....51
Tablo 4.5 İki ihtimalli maç sonucu için destek vektör makineleri sonuçları.....52
Tablo 4.6 Üç ihtimalli maç sonucu için tek gizli katmanlı yapay sinir ağları sonuçları 53
Tablo 4.7 Üç ihtimalli maç sonucu için iki gizli katmanlı yapay sinir ağları sonuçları 54
Tablo 4.8 İki ihtimalli maç sonucu için tek gizli katmanlı yapay sinir ağları sonuçları 56
Tablo 4.9 İki ihtimalli maç sonucu için iki gizli katmanlı yapay sinir ağları sonuçları 57

ÖZET

Maç Sonuçları için Klasik Çok Değişkenli Yöntemler ve Makine Öğrenmesi ile İstatistiksel Öngörü Tahmini

Coşkun PARİM

İstatistik Anabilim Dalı

Doktora Tezi

Danışman: Prof. Dr. Ali Hakan BÜYÜKLÜ

Futbolda en önemli konulardan birisi asıl amaç olan maçın sonucunun tahmin edilmesidir. Çok fazla değişkene sahip olması nedeniyle iyi bir şekilde analiz edilmeli ve değerlendirilmelidir. Bu tez çalışmasının amacı, performans göstergelerini ve durumsal değişkenleri kullanarak galibiyet, beraberlik ve mağlubiyet olmak üzere maç sonucunu tahmin etmektir. Çalışmanın örneklemi, 2010/2021 sezonundan başlayarak, 2019/2020 yılına kadar olan Avrupa Şampiyonlar Ligi grup aşamasındaki tüm maçları kapsamaktadır. İlk olarak, performans değişkenlerinin maç sonucuna etkisini incelemek amacıyla, tek yönlü varyans analizi (ANOVA) uygulanmıştır. Analiz sonucunda, istatistiksel olarak anlamlı bulununan ve maç sonucu bakımından farklılık gösteren değişkenlerin hangi sonuç bakımından farklı olduğu Tukey HSD testi ile belirlenmiştir. Sonraki aşamada, maç sonucunu etkileyen performans göstergeleri kullanılarak veri seti k-ortalama kümeleme yöntemiyle, rakibin kalitesi bakımından zayıf, dengeli ve güçlü olmak üzere üç kümeye ayrılmıştır. Ayrıca, takımların rakiplerinin gücüne göre maç performanslarının nasıl farklılaştığını görselleştirmek için çok boyutlu ölçekleme yöntemi kullanılmıştır. Çok boyutlu ölçekleme analizinin sonuçlarını

desteklemek amacıyla, karar ağaçları analizinden faydalanılmıştır. Bu yöntem sayesinde kazanma, berabere kalma ve yenilme şansını azaltan ve arttıran değişkenler belirlenmiştir. Kullanılan tüm bu istatistiksel yöntemlerin sonucunda, performans göstergeleri ve durumsal değişkenler kullanılarak maç sonucu, galibiyet, beraberlik ve mağlubiyet olmak üzere destek vektör makineleri ve yapay sinir ağları kullanılarak sınıflandırılmıştır. Beraberlik sonucunun, sınıflandırma başarısını düşürürken, galibiyet ve mağlubiyet sonuçlarının iyi ayrıştığı saptanmıştır. Sonuç olarak, değişkenler doğru analiz edildiğinde, maç sonucunun yaklaşık %90 oranından doğru tahmin edildiği belirlenmiştir. Bu bulgular, antrenörlerin maç öncesinde veya sırasında rakibin kalitesine, performans göstergelerine ve durumsal değişkenlere göre farklı stratejiler geliştirmesine yardımcı olabileceğini göstermektedir.

Anahtar Kelimeler: Karar ağaçları, rakibin kalitesi, performans göstergesi, makine öğrenmesi, çok boyutlu ölçekleme

ABSTRACT

Statistical Prediction of Match Results with Classical Multivariate and Machine Learning Methods

Coşkun PARİM

Department of Statistics

Doctor of Philosophy Thesis

Advisor: Prof. Dr. Ali Hakan BÜYÜKLÜ

One of the most important issues in soccer is to predict the match outcome, which is the main goal. Because it has so many variables, it should be well analyzed and evaluated. The purpose of this thesis is to predict the match outcome, including win, draw and loss, using performance indicators and situational variables. The sample of the study covers all matches in the European Champions League group stage starting from the 2010/2021 season through to 2019/2020. First, one-way analysis of variance (ANOVA) was applied to examine the effect of performance variables on the match outcome. As a result of this analysis, Tukey HSD test was used to determine which groups were different for variables that were found to be statistically significant and differ in terms of match outcome. Second, the data set was divided into three clusters as weak, balanced and strong in terms of the quality of the oppositions, using the k-mean clustering method with the performance indicators affecting the match outcome. In addition, multidimensional scaling method was used to visualize how the match performances of the teams differ according to the quality of the opposition. Decision trees analysis was used to support the results of the multidimensional

scaling analysis. Thanks to this method, variables that decrease and increase the chances of winning, drawing, and losing have been determined. As a result of all these statistical methods used, using performance indicators and situational variables, the match outcome was classified using support vector machines and artificial neural networks as win, draw and loss. It has been determined that the drawing as an outcome decreases the classification accuracy, while the win and loss outcome are well differentiated. As a result, when the variables are analyzed correctly, it is determined that the match outcome is predicted correctly about 90%. These findings show that coaches can help develop different strategies according to the quality of the opposition, performance indicators and situational variables before or during the match.

Keywords: decision tree, quality of the opposition, performance indicators, machine learning, multidimensional scaling

1
GİRİŞ

1.1 Literatür Özeti

Futbol dünyada çapındaki en popüler sporlardan birisidir. Küresel hale gelen futbol, çok büyük bir pazardır ve ligler, takımlar, oyuncular, kulüp taraftarları, sponsorlar, yetenek avcıları, bahis şirketleri ve daha da fazlasından oluşmaktadır. Tüm bu paydaşların hepsi yoğun şekilde etkileşime girerek muazzam miktarda veri üretir [1].

Son yıllarda futbol takımları için performans göstergeleri ve durumsal değişkenler ile maç sonuçlarının tahmin edilmesi literatürde popüler araştırma konuları arasındadır [2-5]. Spor biliminin temel amaçlarından birisi, takımların performans anlamında hedeflerini belirlemek ve bu hedefe etki eden değişkenleri araştırmaktır [6, 7]. Bununla birlikte takım performansının istatistiksel olarak analiz edilmesi, antrenörlere ve futbolculara performanslarını yeniden değerlendirme fırsatı sağlayacaktır [3, 8]. Rakiplerin genel özelliklerini, güçlü ve zayıf yönlerini tespit etmede istatistiksel yaklaşımların kullanılması, takımın performansının değerlendirilmesi için çok önemlidir.

Futbolda maç sonucuna etki eden birçok değişken mevcuttur. Literatürde bu değişkenler genellikle notasyonel ve durumsal olmak üzere iki sınıfa ayrılır. Hava topu mücadelesi, adam geçme, top kaybı, pas, top çalma gibi değişkenler notasyonel değişkenlere örnek olarak verilebilir. Ayrıca, rakibin kalitesi, maçın yeri ve maç sonucu da durumsal değişken örnekleridir. Başka bir durumsal değişken örneği de atılan ilk goldür. Futbol düşük skorlu bir spor olduğu için ilk golün atılmasının maç sonucunu etkilediğini gösteren birçok çalışma vardır [2, 9, 10]. Önceki çalışmalarda kazanan, kaybeden veya berabere kalan takımların notasyonel değişkenlerinin araştırılmasında çeşitli pratik uygulamalar oluşturulmuştur. Buna örnek olarak, Lago-Peñas vd. [11] ve García-Rubio vd.

[10], kazanan takımların daha fazla isabetli şut attığını göstermiştir. Tenga vd. [12, 13], kontra atak yapanların normal ataklara kıyasla zayıf savunmalara karşı etkili olduğunu belirtmiştir. Ayrıca, topa sahip olma ve pas değişkenleri kazanan, kaybeden veya berabere kalan takımlar arasında belirleyici değişkenlerdir. Futbol maçının tahmini ile ilgili önceki çalışmalar, topa sahip olma [14-16], isabetli paslar [17, 18], kornerler, topla buluşma [2, 19] ve daha genel anlamda defansif ve ofansif performans değişkenlere [20, 21] odaklanmışlardır. Futbol takımlarının başarısını belirleyen birçok değişken bulunmaktadır. Bu durum, değişkenlerin görselleştirilmesini sağlayan istatistiksel tekniklerin kullanımını kaçınılmaz hale getirmektedir. Literatürde bu konuda yapılmış çok fazla çalışma bulunmamaktadır.

Bir diğer durumsal değişken olan rakibin kalitesi; güçlü, zayıf ve dengeli rakipler olmak üzere literatürde araştırılmıştır [2, 22-25]. Bilek ve Ulaş [2] ve Taylor vd. [25], kümeleme analizini kullanarak rakip takımları daha güçlü veya daha zayıf olarak sınıflandırmıştır. Bilek ve Ulaş [2] rakip takımların kalitesine göre İngiltere Premier Lig için maç sonuçlarını tahmin etmiştir. Futbolun karmaşık ve dinamik ortamı, takımın performansını ölçmede çeşitli zorluklar ortaya çıkarmaktadır [26]. Bu nedenle, performans göstergelerini birlikte görselleştirmek bu karmaşıklığın azaltılmasına yardımcı olacaktır. Bu tez çalışmasında, rakibin kalitesi özelinde performans göstergeleri değerlendirilerek çok boyutlu ölçekleme (ÇBÖ) analizi yardımıyla değişkenlerin birlikte nasıl etkileşim içerisinde olduğunu görselleştirmek amaçlanmıştır. Bu yanıyla tez çalışması literatüre katkı sağlayacaktır.

Maç sonucunun öngörüsü için son yıllarda makine öğrenmesi yöntemleri de kullanılmaktadır. Capobianco vd. [27] maç sonucunu tahmin etmek için Random Forest yöntemini kullanmışlardır. Timmaraju vd. [28] ise destek vektör makineleri yardımıyla İngiltere Premier Liginde maç sonucunu öngörmüşlerdir. Hucaljuk ve Rakipović [29] maç sonuçlarının tahmini için Naive Bayes, Bayesyen ağlar, LogitBoost, k-en yakın komşu algoritması, Random forest ve yapay sinir ağları modellerinin sonuçlarını karşılaştırmışlardır.

Önceki çalışmalar, takım performansının ve başarısının tahminini daha iyi hale getirmede modellemenin önemini ortaya koymaktadır. Böyle bir değerlendirme için, ÇBÖ ve kümeleme analizi gibi çok değişkenli istatistiksel yöntemlerle birlikte karar ağaçlarının kullanılması, yapısal sorunları anlamada önemli yöntemleri oluşturmaktadır [30]. Çok değişkenli istatistiksel yöntemler birden çok değişkenin davranışını ortaya çıkararak ve yapılar arasındaki benzer veya farklı kalıpları tespit ederek gruplar arasındaki çok değişkenli farklılıkları analiz etmek için kullanışlıdır [31]. Ek olarak, maç sonuçlarını etkileyen değişkenleri analiz etmek için karar ağaçları, ÇBÖ, k-ortalamalar kümeleme ve tek yönlü ANOVA kullanılmıştır. Değişkenlerin arasındaki etkileşimler daha iyi anlaşıldıktan sonra ise bayesyen yapay sinir ağı ve destek vektör makineleri kullanılarak galibiyet, mağlubiyet ve beraberlik olmak üzere üç ihtimalli ve galibiyet, mağlubiyet olmak üzere iki ihtimalli maç sonuçlarının öngörü tahminleri yapılmıştır.

Bu tezde amaçlanan; adam geçme denemesi, başarılı adam geçme, başarılı pas, başarılı top çalma, çalım yeme, defansif hava topu, faul, hava topu kazanma, isabetli şut, kilit pas, korner, ofansif hava topu, ofsayt, top çalma denemesi, top kapma, top kaybı, topla buluşma, toplam pas, toplam şut, uzaklaştırma olmak üzere oyun içi değişkenleri, rakibin kalitesi (daha güçlü, dengeli ve daha zayıf), maç yeri (ev sahibi ve deplasman) ve atılan ilk gol (attı ve atmadı) olmak üzere üç durumsal değişkeni kullanarak takımların galibiyet, mağlubiyet ve beraberlik durumlarını tahmin etmektir.

Literatürde bu konuda yapılan benzer çalışmalara rağmen maç sonucunu etkileyen performans göstergeleri arasındaki ilişkiyi belirlemek ve görselleştirmek için ÇBÖ analizinin kullanılması tez çalışmasının yenilikçi kısmını oluşturmaktadır. Ayrıca, değişkenler arasındaki dinamiklerin kullanılan analizler sayesinde bütünüyle ortaya konulması da literatüre katkı sağlayacaktır.

1.2 Tezin Amacı

Bu tez çalışmasının amacı, futbol maçlarının galibiyet, beraberlik ve mağlubiyet sonuçlarını durumsal değişkenler ve performans göstergeleri ile analiz etmektir. Bu nedenle, Avrupa Şampiyonlar Ligi'nde 2010/2011 ve 2019/2020 sezonları

arasındaki on yılı kapsayan grup aşaması maçlarından elde edilen verilerinden derlenen yirmi oyun içi değişkeni (adam geçme denemesi, başarılı adam geçme, başarılı pas, başarılı top çalma, çalım yeme, defansif hava topu, faul, hava topu kazanma, ilk golü atma, isabetli şut, kilit pas, korner, ofansif hava topu, ofsayt, top çalma denemesi, top kapma, top kaybı, topla buluşma, toplam pas, toplam şut ve uzaklaştırma) ve iki durumsal değişken (rakibin kalitesi ve maç yeri) kullanılarak bu değişkenlerin maç sonuçları üzerine etkileri araştırılmıştır. Maç sonucunu etkileyen performans göstergeleri arasındaki ilişkiyi belirlemek ve görselleştirmek için ÇBÖ analizinin kullanılması literatürdeki boşluğu katkı sağlayacaktır. Ayrıca, kullanılan değişkenler yardımıyla, makine öğrenmesi yöntemlerinin maç sonucunun sınıflandırılabilmesinde kullanılabileceği de kanıtlamak amaçlanmıştır.

1.3 Hipotez

Bu tezin amacı, futbol maçlarının galibiyet, beraberlik ve mağlubiyet sonuçlarını durumsal değişkenler ve performans göstergeleri ile analiz etmektir. Tezin hipotezleri: (i) Maçların sonucunu etkileyen performans göstergelerinin istatistiksel olarak belirlenmesi, (ii) Maçın kazanılması ya da kaybedilmesi olasılıklarını hangi değişkenlerin arttırdığının veya azalttığının tespit edilmesi, (iii) Maç sonucunu etkileyebileceği varsayımı altında rakibin kalitesinin (güçlü, dengeli ve zayıf) durumsal bir değişken olarak maç sonucunu etkileyen değişkenler arasında kullanılmasının gerekliliği, (iv) Maç sonucunu etkilediği belirlenen performans göstergelerinin rakibin kalitesine göre değişkenlik gösterdiğinin görsel olarak incelenmesi, (v) Kullanılan değişkenler yardımıyla maç sonucunun tahmin edilebileceği ya da sınıflandırılabileceği, şeklinde sıralanabilir. Bu bulgular, antrenörlerin maç öncesinde veya sırasında rakiplerin kalitesine, durumsal değişkenlere ve performans göstergelerine dayalı olarak farklı stratejiler geliştirmesine veya karar almalarına yardımcı olabilir.

2
HİPOTEZ TESTLERİ VE KLASİK ÇOK DEĞİŞKENLİ İSTATİSTİKSEL YÖNTEMLER

2.1 Hipotez Testleri

2.1.1 Tek Yönlü Varyans Analizi (ANOVA)

Varyans analizi olarak da bilinen ANOVA (Analysis of Variance) ilk olarak Kruskal W. [32] tarafından önerilmiştir. Üç veya daha fazla kategorinin (grup veya örnek) anakütle (popülasyon) ortalamaları arasında istatistiksel olarak anlamlı bir fark olup olmadığının araştırılması için kullanılır. Tek yönlü ya da tek faktörlü denmesinin nedeni ise tekrarlı ölçümler olmadığında ve kategorilerin etkileşimlerini dikkate almadan hesaplanmasıdır. Tekrarlı ölçümler olduğunda veya kategori etkileşimleri de araştırılmak istendiğinde farklı isimler alabilmektedir.

Örneğin; araştırma sorusu, takımların yaptıkları pas sayıları maç sonucuna göre farklılık göstermekte midir? Şeklinde olabilir. Maç sonucunun galibiyet, beraberlik ve mağlubiyet olmak üzere üç kategorisi bulunmaktadır. Ayrıca, takımların pas sayıları da sayısal değerlere sahip, yani sürekli bir değişkendir. Analizde, kategorik olan değişken (maç sonucu) bağımsız değişken ya da faktör olarak isimlendirilir. Sürekli olan değişken (pas sayısı) ise bağımlı değişken olarak ifade edilir. Burada belirtilmesi gereken bir diğer durum da tek yönlü ANOVA'nın parametrik bir test olduğudur. İstatistikte dağılım bilindiğinde ve varyansların homojen olması durumunda parametrik testler kullanılmaktadır. Dağılım bilinmediğinde ve varyanslar homojen değilse, alternatif olarak parametrik olmayan testler kullanılmaktadır. Tek Yönlü ANOVA testinin parametrik olmayan durumlar için karşılığı ise Kruskall Wallis testidir.

İstatistiksel testler yapılırken, araştırma konusu ya da sorusu belirlendikten sonra öncelikle hipotez kurulması gerekmektedir. Hipotezin H_0 ve H_1 (ya da H_a) olmak

üzere iki seçeneği bulunmaktadır. H_0 hipotezi, genellikle hiçlik ifade etmektedir. Burada hiçlikten kasıt gruplar arasında fark olmamasıdır. H_0 hipotezi kabul edilirse değişen bir durum olmadığı söylenebilir. H_1 hipotezi ise, H_0 hipotezinin alternatifidir. Yani, gruplar arasında en az birinde farklılık olduğu anlamına gelir. Başka bir değişle, H_0 hipotezinin boş veya yanlış olmasıdır. Hipotezler sembolik olarak ifade edilirse,

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : \textit{En az iki ortalama birbirinden farklı}$$

Burada, μ_1, μ_2 ve μ_3 sırasıyla birinci, ikinci ve üçüncü kategorilerin anakütle ortalamalarını ifade etmektedir. Örneğin, μ_1 galip gelen takımların pas sayılarının, μ_2 berabere kalan takımların pas sayılarının ve μ_3 mağlup olan takımların pas sayılarının ortalaması olabilir.

Üç veya daha fazla grup için, sıfır hipotezi F oranıyla test edilir. F'in sıfır hipotezini test etmek için kendi örnekleme dağılımları ailesi vardır. Ortaya çıkan test, F testi olarak bilinir. Aslında F, paydaki tüm örnekleme ortalamaları (gruplar arası değişkenlik) ile payda teriminde tahmin edilen hata terimi veya varyans tahmini (grup içi değişkenlik) arasında gözlemlenen farklılıkların oranını yansıtır. F testi, sıfır hipotezi doğruysa, F oranının hem pay hem de paydasının yaklaşık aynı olma eğiliminde olacağı, ancak sıfır hipotezi yanlışsa, pay eğilimli olacağı yani paydadan daha büyük olması fikrine dayanmaktadır [33].

Varyans analizi, gruplar arasındaki ve gruplar içindeki değişkenliği ölçmek için örnek varyans tahminlerini kullanır. Örnek varyansı, herhangi bir gözlem seti arasındaki değişkenliği, ilk olarak kareler toplamını (KT), yani ortalamaları ile ilgili kare sapmaların toplamını bularak ölçer. Genel olarak ANOVA tablosu Tablo 2.1'deki gibi verilebilir.

ANOVA'daki hesaplamalar genellikle kareler toplamına yöneliktir. Burada KT_{ara} gruplar arası kareler toplamını, $KT_{iç}$ grup içi kareler toplamını ve bu ikisinin toplamı olan KT_{toplam} tüm veri için kareler toplamını ifade eder. Ek olarak, $KT_{iç}$ sadece benzer şekilde örnekler arasındaki değişkenliği yansıttığından, daha genel olarak rasgele hatanın karelerinin toplamı olarak adlandırılabilir ve KT_{hata} olarak

da sembolize edilebilir. Ayrıca, sd_{ara} gruplar arası serbestlik derecesini, $sd_{iç}$ grup içi serbestlik derecesini ve sd_{toplam} da tüm veri için serbestlik derecesini göstermektedir. N örneklemdeki gözlem sayısını ifade ederken, k grup sayısına karşılık gelmektedir.

Tablo 2.1 Anova tablosu

Değişimin kaynağı	Serbestlik Derecesi (sd)	Kareler Toplamı (KT)	Kareler Ortalaması (KO)	F hesaplanan
Gruplar arası	$sd_{ara} = k - 1$	$KT_{ara} = n\sum(\bar{X}_{grup} - \bar{X})^2$	$KO_{ara} = KT_{ara} / sd_{ara}$	$KO_{ara} / KO_{iç}$
Grup içi	$sd_{iç} = N - k$	$KT_{iç} = \sum(X - \bar{X}_{grup})^2$	$KO_{iç} = KT_{iç} / sd_{iç}$	
Toplam	$sd_{toplam} = N - 1$	$KT_{toplam} = \sum(\mathrm{X} - \bar{\mathrm{X}})^2$		

KO_{ara} gruplar arasındaki kareler ortalamasını göstermektedir. Gruplar arası kareler toplamının, gruplar arası serbestlik derecesine bölünmesiyle hesaplanabilir. Benzer şekilde, $KO_{iç}$ de grup içi kareler ortalamasını göstermektedir ve grup içi kareler toplamının, grup içi serbestlik derecesine bölünmesi ile hesaplanmaktadır. Bu iki kareler toplamının birbine oranı da F'in değerini vermektedir.

Sıfır hipotezi hakkında bir karar vermek için, hesaplanan F'in kritik bir F değeri ile karşılaştırılması gereklidir. Kritik F, F oranının payı ve paydasındaki ortalama kareler için serbestlik derecesi çifti aracılığıyla tanımlanır. F tablosunda, $sd_{iç}$'e eşit serbestlik derecelerine sahip satırla ve sd_{ara} 'ya eşit serbestlik derecelerine sahip sütunla kesişen hücre kritik F değerini göstermektedir. Hesaplanan F değeri kritik F değerini aştığında sıfır hipotezi reddedilebilir. Aksi takdirde, sıfır hipotezi reddedilemez. Kullanılan tablodaki anlamlılık seviyesi ne ise sonuç da o anlamlılık seviyesindedir. Örneğin %95 güven düzeyi için F tablosuna bakılırsa, sonucu da %5 anlamlılık seviyesinde değerlendirilmelidir.

Ayrıca p değeri hesaplanarak da karar verilebilir. %5 anlamlılık düzeyinde F (sd_{ara} , $sd_{iç}$, F hesaplanan) değerleri için hesaplanan p değeri 0.05'ten küçükse sıfır hipotezi reddedilir. Aksi takdirde sıfır hipotezi kabul edilir.

2.1.1.1 Etki Büyüklüğü

Eta kare iki grup arasındaki standartlaştırılmış farkı temsil etmek için kullanılan bir etki büyüklüğü ölçüsüdür ve ikiden fazla grup olduğunda Cohen'in d'si ile aynı etki büyüklüğü sınıflarına sahiptir. Bununla birlikte, sınıfların uzaklık değerleri Cohen'in d'sinden farklıdır. Eta kare, etki büyüklüğü şu şekilde yorumlanarak [0, 1] arasında bir değer alır: anlamlı değil (≤0,009), küçük (>0,01, <0,0588), orta (≥0,0589, <0,1379) ve büyük (≥ 0,138) [34]. Küçük bir etki boyutu, büyük örnek boyutları gerektirir. Büyük örnek boyutları, oldukça küçük ve muhtemelen önemsiz olan farklılıkları tespit edebilir [35].

İstatistiksel olarak anlamlı bir F, t testi gibi, yalnızca boş hipotezinin muhtemelen yanlış olduğunu gösterir; aksi takdirde, etki büyüklüğünün doğru bir tahminini sağlayamaz. Yeni bir etki boyutu tahmini, hem ANOVA'daki sıfır hipoteziyle ilişkili genel etkiyi yansıtmalı hem de örnek boyutlarından bağımsız olmalıdır [33]. Eta kare eşitlik (2.1)'deki gibi ifade edilebilir.

$$\eta^2 = \frac{KT_{ara}}{KT_{toplam}} \tag{2.1}$$

Burada η^2, açıklanan varyans oranını temsil eder. Ayrıca, KT_{ara} gruplar arası ve KT_{toplam} toplam kareler toplamını göstermektedir. Bu oran, popülasyon ortalamalarının farklılıklarını değil, tüm puanlar için toplam varyansın oranını gruplarının varyansı ile açıklayarak tahmin eder [33]. Bu oran ne kadar büyükse, bağımlı değişken üzerindeki genel etkisinin tahmini boyutu o kadar büyük olur.

Eta-kare olarak telaffuz edilen sembol η^2, genellikle kare eğrisel korelasyon katsayısı olarak anılır. Bu terminoloji, η^2'nin bağımsız ve bağımlı değişkenlerin değerleri arasındaki doğrusal olmayan korelasyonu tanımlayan bir sayıdan sadece bir karekök uzakta olduğunu gösterir [33].

2.1.1.2 Çoklu Karşılaştırma

ANOVA'da genel sıfır hipotezinin reddedilmesi, yalnızca tüm popülasyon ortalamalarının eşit olmadığını gösterir. Üç ve üçten fazla grup olması nedeniyle, farkın hangi grup ya da gruplardan kaynaklandığının tespiti için çoklu karşılaştırma testleri kullanılmalıdır. Literatürde, çok sayıda çoklu karşılaştırma

testi bulunmaktadır. Birden fazla karşılaştırma testi, yalnızca birini değil aynı zamanda popülasyon ortalamaları arasındaki bir dizi farklılığı değerlendirmek için tasarlanmıştır; örneğin, sıfır hipotezi reddedildiği takdirde, yukarıdaki verilen hipotezdeki gibi üç kategori için karşılaştırılması gereken üç farklı durum μ_1-μ_2, μ_1-μ_3 ve μ_2 - μ_3 şeklinde sıralanabilir.

- **Tukey Testi**

ANOVA'da genel boş hipotez reddedildikten sonra, Tukey'in HSD testi, ortalamalar çiftleri arasındaki tüm olası farklılıkları test etmek için kullanılabilir, ancak yine de tip I hatanın birikimli olasılığı hiçbir zaman belirtilen anlamlılık düzeyini aşmaz [33].

Genel sıfır hipotezi için önemli bir F verildiğinde, Tukey testi, her bir olası grup çifti için her bir farkın önemini değerlendirmek için tek bir kritik değer, bununla birlikte bir olasılık sağlar. Esasen HSD için kritik değer, en az tip I hataya maruz kalmanın artan kümülatif olasılığını dengelemek için karşılaştırılarak k grup ortalamalarının sayısı için yukarı doğru ayarlanır. Bu yukarı doğru ayarlamanın net etkisi, herhangi bir belirli popülasyon aracı çifti için sıfır hipotezini reddetmeyi zorlaştırmak ve yalnızca dürüstçe önemli (veya gerçek) farklılıkları tespit etme olasılığını artırmaktır [33].

Herhangi bir araç çifti arasındaki mutlak fark, HSD için kritik değere eşitse veya bu değeri aşarsa, bu belirli popülasyon aracı çifti için sıfır hipotezi reddedilebilir. HSD'yi belirlemek için eşitlik (2.2)'den yararlanılır.

$$Tukey - HSD = q\sqrt{\frac{KO_{iç}}{n}} \tag{2.2}$$

HSD, iki grup arasındaki herhangi bir fark için pozitif kritik değerdir; q, Tukey HSD q değerleri tablosundan elde edilen, teknik olarak Studentized Range İstatistiği olarak adlandırılan bir değerdir; $KO_{iç}$, genel ANOVA için grup içi değişkenliğin geleneksel ortalama karesidir; ve n, her gruptaki örneklem büyüklüğüdür [33].

ANOVA'daki grup sayısı ve grup içi serbestlik derecesi yardımıyla, Tukey HSD q değerleri tablosundan q değeri belirlenir ve Tukey-HSD değeri hesaplanır.

Sonrasında ikili grup ortalama farkları hesaplanır. Tukey HSD kritik değeri grup ortalama farkını aşıyorsa, o gruplar arasında istatistiksel olarak anlamlı bir fark olduğu söylenir. Bu karar p değeri ile de mümkündür. %95 güven düzeyinde p olasılığı 0.05'ten küçükse gruplar arası anlamlı bir fark olduğu söylenebilir.

2.2 K-ortalamalar Kümeleme

Kümeleme analizinin temel amacı, kümeler (gruplar) arasındaki benzerlikleri ve farklılıkları ortaya çıkarmaktır [36]. En tipik kümeleme algoritmalarından biri olan k-ortalamalar algoritması ilk olarak 1967'de MacQueen [37] tarafından önerilmiştir. Yöntemin amacı veriyi k farklı küme halinde sınıflandırmaktır. Dolayısıyla, oluşturulan kümeler kendi içinde homojen ve birbirlerinden bağımsızdır.

Algoritma iki ayrı aşamadan oluşur. İlk aşama, k değerinin önceden belirlendiği k merkezlerini rastgele seçer. İkinci aşama, her veri nesnesini en yakın merkeze atar. Genellikle her veri nesnesi ile küme merkezleri arasındaki mesafeyi belirlemek için Öklid mesafesi kullanılır. Tüm veri nesneleri bazı kümelere dahil edildiğinde ilk adım tamamlanır ve ilk gruplama yapılır. İşlem kriter fonksiyonu minimum olana kadar tekrar tekrar devam eder [38].

Burada, x_i hedef nesne ve c_j küme merkezi olmak üzere J kriter fonksiyonu eşitlik (2.3)'te tanımlanır.

$$J = \sum_{i=1}^{k} \sum_{i=1}^{n} \left| x_i - c_j \right|^2 \tag{2.3}$$

J, tüm nesnelerin hata karesinin toplamıdır. Bir x vektörü ile başka bir y vektörü arasındaki Öklid mesafesi, aşağıdaki gibi elde edilebilir:

$$d(x_i, y_i) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)} \tag{2.4}$$

Küme sayısı k olan ve n sayıda veri nesnesi içeren D veriseti için K-ortalamalar algoritma süreci aşağıdaki gibidir [38]:

- Başlangıç küme merkezleri olarak D veri kümesinden k veri nesnesi rastgele seçilir.

- Her veri nesnesi d_i ve tüm k küme merkezleri c_j arasındaki mesafe hesaplanır ve d_i veri nesnesi en yakın kümeye atanır.

- Kümelerin merkezinde hiçbir değişiklik olmayana kadar her j kümesi için, küme merkezi yeniden hesaplanır.

K-ortalamalı kümeleme algoritması her zaman yerel minimuma yakınsar. Yakınsamadan önce, mesafe ve küme merkezleri hesaplamaları yapılırken, döngüler birkaç kez yürütülür, burada pozitif tamsayı t, k-ortalamaları yinelemelerinin sayısı olarak bilinir [38].

2.3 Çok Boyutlu Ölçekleme

Çok boyutlu ölçekleme (ÇBÖ), bir dizi nesne üzerindeki benzerlik veya farklılık verilerinin analizi için kullanılan bir tekniktir. Bu tür veriler, test öğelerinin karşılıklı korelasyonları, siyasi adaylar üzerindeki benzerlik oranları veya bir dizi ülke için ticaret endeksleri olabilir. ÇBÖ, bu tür verileri geometrik bir uzaydaki noktalar arasındaki mesafeler olarak modellemeye çalışır. Bunu yapmanın ana nedeni, veri yapısının grafiksel bir görüntüsünü istemektir; bu, bir dizi sayıdan çok daha kolay anlaşılır ve dahası, verilerdeki temel bilgileri görüntüleyerek gürültüyü düzeltir.

Çok sayıda ÇBÖ çeşidi vardır. Verilerin haritalandırılmasının istendiği geometri türü, haritalama fonksiyonu, optimal veri görselini bulmak için kullanılan algoritmalar modellerdeki istatistiksel hatanın hesaplanması veya aynı anda sadece bir ya da birkaç benzerlik matrisi arasında temsil edilememe olasılığı bu türleri ayıran bazı yönlerdir. Diğer yönler, ÇBÖ'nin kullanıldığı farklı amaçlarla ilgilidir [39].

N madde için, farklı madde çiftleri arasında M=N(N-1)/2 sayıda benzerlik (uzaklık) vardır ve bu benzerlikler temel verileri oluşturur (benzerliklerin kolayca ölçülmediği durumlarda, benzerliklerin sıra sıraları temel verileri oluşturur). Hiçbir bağ olmadığı varsayıldığında, benzerlikler $s_{i_1k_1} < s_{i_2k_2} < \ldots < s_{i_Mk_M}$ şeklinde artan bir sırada düzenlenebilir. Burada, $s_{i_1k_1}$, M benzerliklerinin en küçüğüdür. Alt simge i_1k_1 en az benzer öğe çiftini, yani benzerlik sıralamasında birinci sıraya sahip

öğeleri belirtir. Diğer alt simgeler de aynı şekilde yorumlanabilir. N öğenin q boyutlu bir konfigürasyonu bulunmak istendiğinde, öğe çiftleri arasındaki $d_{ik}^{(q)}$ mesafeleri $s_{i_1k_1} < s_{i_2k_2} < ... < s_{i_Mk_M}$ 'deki sıralamayla eşleşir. Mesafeler bu sıralamaya uygun bir şekilde düzenlenirse, $d_{i_1k_1}^{(q)} > d_{i_2k_2}^{(q)} > ... > d_{i_Mk_M}^{(q)}$ olduğunda mükemmel bir eşleşme gerçekleşir. Yani, q boyutlarındaki mesafelerin azalan sıralaması, ilk benzerliklerin artan sıralamasına tam olarak benzer. Buradaki sıra korunduğu sürece, mesafelerin büyüklükleri önemsizdir [41].

Belirli bir q değeri için, ikili uzaklıkları orijinal benzerliklerle monoton olarak ilişkili olan noktaların bir konfigürasyonunu bulmak mümkün olmayabilir. Bu nedenle, Kruskal [40], stress adında bir ölçü önermiştir (2.5).

$$Stress(q) = \sqrt{\frac{\sum_{i<k} (d_{ik}^{(q)} - \hat{d}_{ik}^{(q)})^2}{\sum_{i<k} [d_{ik}^{(q)}]^2}} \qquad (2.5)$$

Stres formülündeki $\hat{d}_{ik}^{(q)}$'ler karşılığı bilinen sayılardır; yani, benzerliklerle monoton olarak ilişkilidirler. Ayrıca uzaklıklar değildir. Bunlar yalnızca, gözlemlenen $d_{ik}^{(q)}$ 'lerin tekdüzeliğini değerlendirmek için kullanılan referans numaralarıdır [41].

Buradaki fikir, stresin mümkün olduğu kadar küçük olacağı şekilde, maddelerin q boyutlarında noktalar olarak temsil edilmesidir. Kruskal [40], ayrıca stresin Tablo 2.2'deki yönergelere göre yorumlanmasını önermektedir.

Tablo 2.2. Stress değerlerinin uyum iyiliği karşılıkları

Stress	Uyum iyiliği
20	Zayıf
10	Orta
5	İyi
2.5	Çok İyi
0	Tam

Burada, uyum iyiliği, benzerlikler ve son uzaklıklar arasındaki tekdüze ilişkiyi ifade eder. Çok boyutlu ölçekleme algoritmasının tamamı şu adımlarda özetlenebilir [41]:

- N öğe için, farklı öğe çiftleri arasındaki M=N(N-1)/2 benzerlikleri (mesafeler) elde edilir ve benzerlikleri sıralanır. Eğer benzerlikler hesaplanamıyorsa, sıra sıraları belirtilir.

- q boyutlarında bir deneme yapılandırması kullanarak, $d_{ik}^{(q)}$ mesafesi ve $\hat{d}_{ik}^{(q)}$ sayılarını belirlenir ve stress en aza indirilir.

- $\hat{d}_{ik}^{(q)}$'leri kullanarak, gelişmiş bir yapılandırma elde etmek için noktaları hareket ettirilir. Yeni bir konfigürasyonda yeni $d_{ik}^{(q)}$ 'ler, $\hat{d}_{ik}^{(q)}$ 'ler daha küçük stress olacaktır. İşlem, en iyi (minimum stres) seviyesine ulaşana kadar tekrarlanır.

- Minimum stress (q) ile q 'nun grafiğini çizilir ve bu grafiğin yardımıyla en iyi boyut sayısı q^* seçilir.

3
MAKİNE ÖĞRENMESİ

3.1 Genel Kavramlar

3.1.1 Denetimli ve Denetimsiz Öğrenme

Denetimli öğrenme yöntemlerinde bağımlı değişken (Y) tanımlıdır ve kullanılır. Sınıflandırma ve regresyon olmak üzere iki ana tür denetimli öğrenmeden söz edilebilir. Öte yandan denetimsiz öğrenmede bağımlı değişken mevcut değildir. Bu durumda bağımsız değişkenlerin dağılımları üzerinden problem çözülmeye çalışılır [42]. Denetimli öğrenmedeki regresyonda problemleri için, bağımlı değişken sürekli veya sayısal bir değişkendir. Yani, $Y \epsilon R$ olmak üzere sürekli değişen bir gerçek sayıdır. Bu durumda, bağımsız X değişkeni veya değişkenleri yardımıyla bağımsız değişken tahmin edilmeye çalışılır. Regresyonda genellikle kayıp fonksiyonu olarak ortalama kare hatası kullanılır. Örneğin, bir futbol takımının pas sayısı reel bir sayıdır. Pas sayısı tahmin edilmeye çalışıldığında, tahmin değerleri ile gerçekleşen değerlerin arasındaki fark önemli hale gelir.

Denetimli öğrenmede sınıflandırma problemleri için, c sınıf sayısı olmak üzere $Y \epsilon \{1,2,\ldots,c\}$ olsun. Bağımlı değişken, sayısal bir anlamı olmayan ve sadece farklı kategorileri olan değişkendir. Burada kategoriler genellikle etiket olarak adlandırılır. Sınıflandırmada, $c = 2$ ise bağımlı değişken sadece iki sınıftan oluşmaktadır, yani iki farklı etiketi vardır. Örneğin, maç sonucunu gösteren ve ikili sınıfları (etiketi) sırasıyla sırasıyla "galibiyet" ve "mağlubiyet" olan bağımlı değişken olabilir. Bu durumda etiketler de 1 ve 2 olarak kodlanabilir. Bu durumda, bağımsız değişkeni veya değişkenleri yardımıyla kategorilerden oluşan bağımsız değişkeni sınıflandırılmaya çalışılır. Dikkat edilmesi gereken nokta, sınıflandırıcı yardımıya kategorilerin doğru sınıflandırılması ya da yanlış sınıflandırılmasıdır.

Öte yandan, denetimsiz öğrenmede Y mevcut olmaması nedeniyle hiçbir öngörü ve öngörü hatası yoktur. Bu nedenle, herhangi bir performans kriteri tanımlamak da zordur. [42].

3.1.2 Sınıflayıcının Performansının Değerlendirilmesi

Klasik istatistiksel yöntemlerden makine öğrenmesi yöntemlerine geçişte, temel fark aslında veri bölümlenmesidir. Genellikle, makine öğrenmesinde veri eğitim ve test olmak üzere iki parçaya veya eğitim, doğrulama ve test olmak üzere üç parçaya ayrılır. Bağımlı değişkenin kategorik olduğu bir durum için veri bölümlenmesini açıklamak gerekirse, eğitim, model parametrelerini oluşturmak ve etiketli bir veri seti kullanarak en iyi parametreleri seçmek için sistematik bir yaklaşım sağlamalıdır. Bu durumdaki bir etiketli veri setine eğitim veri seti denir. En iyi parametrelerin seçimi, tahmin edilen sınıf etiketleri ile eğitim veri setindeki gerçek sınıf etiketleri arasındaki hatanın en aza indirildiği anlamına gelir.

Test, eğitim aşamasında kurulan modelin başka bir etiketli veri kümesiyle verimli bir şekilde çalışıp çalışmadığını doğrulamak için sistematik bir yaklaşım sağlamalıdır. Bu durumdaki veri setine de test veri seti denir. Ayarlama parametrelerinin verimli bir şekilde çalıştığının doğrulanması, eğitilmiş modelin test veri seti ile yüksek sınıflandırma doğruluğu sağladığı anlamına gelir.

Doğrulama, modelin daha önce karşılaşılmayan veriler üzerinde verimli bir şekilde performans gösterecek şekilde iyileştirmek için modelleri farklı koşullar altında eğitmek ve test etmek için sistematik bir yaklaşım sağlamalıdır. Bunun için hem eğitim hem de doğrulama setlerinde bir eğitim yapılır [43].

Eğitim sırasında bir modelin genelleme hatasını tahmin etmek için çeşitli yöntemler mevcuttur. Tahmini hata, öğrenme algoritmasının aşırı uydurmaya duyarlı olmayan doğru karmaşıklıkta bir model bulmasına yani model seçimi yapmasına yardımcı olur. Model oluşturulduktan sonra, daha önce görülmemiş gözlemlerin sınıf etiketlerini tahmin etmek için test setine uygulanabilir.

Modelin performansını test setinde ölçmek genellikle önemlidir. Çünkü böyle bir ölçü, genelleme hatasının tarafsız bir tahminini sağlar. Test setinden hesaplanan

doğruluk veya hata oranı, farklı sınıflandırıcıların göreceli performansını karşılaştırmak için de kullanılabilir. Ancak bunun için, test kayıtlarının sınıf etiketleri bilinmelidir [44]. Veri setinin eğitim, test veya eğitim, test, doğrulama olmak üzere bölünmesi ile sınıflandırıcının performansını değerlendirmek için birkaç farklı yöntem mevcuttur. Holdout yöntemi ve k-katlı çapraz doğrulama yöntemi bunlara örnek olarak verilebilir.

3.1.3 Holdout Methodu

Holdout yönteminde, veri seti eğitim ve test seti olmak üzere ikiye bölünür. Daha sonra test seti sınıflandırma modeli için verisetinden çıkartılır ve eğitim seti kullanılarak model oluşturulur. Modelin performansı test setinde değerlendirilir. Eğitim ve test için ayrılan verilerin oranı genellikle analiz eden kişinin kararına göre değişkenlik gösterebilir. Eğitim seti için bu verisetinin %70'i ve test seti için %30'u rastgele seçilebilir. Sınıflandırıcının doğruluğu, test setindeki modelin doğruluğu ile ölçülür. Tipik olarak, bir veri kümesinin eğitim ve test kümelerine bölünmesi basit bir rastgele alt örnekleme sürecidir. Tüm veri noktalarının aynı olasılık dağılımından (her sınıfa göre) alındığı varsayılır [45].

3.1.4 K-Katlı Çapraz Doğrulama

K-katlı çapraz doğrulama, holdout yönteminin geliştirilmiş hali gibi düşünülebilir. Veri seti k alt gruba bölünür ve holdout yöntemi k kez tekrarlanır. Her seferinde k alt kümeden biri test kümesi olarak kullanılır ve diğer alt kümeler bir eğitim kümesi oluşturmak için bir araya getirilir. Daha sonra tüm k denemelerindeki ortalama hata hesaplanır. Bu yöntemin avantajı, verilerin nasıl bölündüğünün önemli olmamasıdır. Her veri noktası bir test setinde tam olarak bir kez ve eğitim setinde k-1 kez kullanılır. Sonuç tahmininin varyansı, k arttıkça azalır. Bu yöntemin dezavantajı, eğitim algoritmasının k kez yeniden çalıştırılması gerektiğidir, bu da bir değerlendirme yapmak için k kat daha fazla hesaplama gerektirdiği anlamına gelir [46].

3.1.5 Bilgi Kriterleri

Bilgi kriterleri model seçim kriterleri arasında gösterilebilir. Özellikle, bağımlı değişkeni aynı dereceden ölçülmemiş modellerin karşılaştırılmasında kullanılmaktadır. Ayrıca yapay sinir ağları modellerinde nöron sayısı farklı modelleri karşılaştırırken de sıklıkla kullanılmaktadır. En sık kullanılan bilgi kriterleri,

Akaike Bilgi Kriteri [47],

$$AIC = n\log(\mathrm{HKT}/n) + 2p \tag{3.1}$$

Düzeltilmiş Akaike Bilgi Kriteri [48],

$$AICc = n\log(HKT/n) + (n+p)/(1-(p+2)/n) \tag{3.2}$$

Schwarz (Bayesyen) Bilgi Kriteri [49],

$$BIC = n\log(HKT/n) + p\log(n) \tag{3.3}$$

Şeklinde verilebilir. Her üç bilgi kriteri için, n eğitim veri seti sayısını, p parametre sayısını, ve HKT de hata kareler toplamını göstermektedir. Bilgi kriterlerinin yorumlanmasında ise, model karşılaştırılırken sayısal olarak bilgi kriterinin küçük olması modelin daha iyi olduğu anlamını taşımaktadır. Bu yönüyle kriterlerin yorumlanması basittir.

3.1.6 Hata Kareler Ortalaması

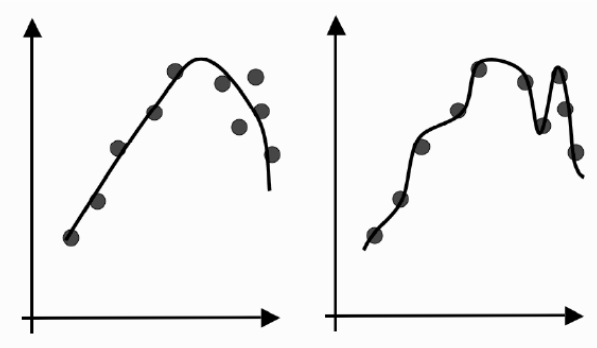
Hata kareler ortalaması (3.4) model hatasının karelerinin ortalamasını ifade etmektedir. Genellikle bağımlı değişken sürekli olduğu durumlarda kullanılır. Kategorik olduğu durumlarda da yer verilebilir fakat genellikle çok fazla tercih edilmez.

$$HKO = \frac{1}{n}\sum_{i=1}^{n}(Y-\hat{Y})^2 \tag{3.4}$$

3.1.7 Aşırı Uyum

Algoritmanın iyi bir genelleme yapması için yeterince eğitim yapılması gerekmektedir. Verinin az eğitilmesi istenmeyen bir durumsa da fazla eğitilmesi de sorun teşkil edebilmektedir. Çoğu makine öğrenimi algoritmasında çok fazla parametre bulunmaktadır. Aynı durum yapay sinir ağı için de söylenebilir ve çok

fazla ağırlık olabilir ve her biri değişebilir. Bu nedenle, modelin çok fazla eğitilmesi algoritmanın modeli çok fazla öğrenmesini ve karmaşık olmasını sağlayacaktır. Bu sebeple genelleme yapması zorlaşacaktır.



Şekil 3.1 Aşırı uyum örnek görseli [50]

Şekil 3.1 bir algoritmanın tahminlerini eğrisel olarak öğrenme sürecinde iki farklı kestirimini göstermektedir. Şeklin solundaki eğri, verilerin genel eğilimine iyi uymaktadır, ancak özellikle düşüşe geçtiği bölümde modelin hatasının sıfırdan farklı olduğu açıktır. Model eğitilmeye devam ettikçe, daha düşük bir eğitim hatasına sahip çok daha karmaşık bir model üretebilir, bu da eğitim verisini ezberlediği ya da çok iyi öğrendiği anlamına gelir. Bu duruma aşırı öğrenme veya aşırı uyum denir. Aşırı uyum olduğunda model veri setini çok iyi öğrendiği için yeni bir veri ile karşılaştığında model hatası yüksek olacaktır. Algoritma aşırı uyum sağlamadan öğrenme sürecini durdurmak gereklidir. Bu nedenle, bu amaçla kullanmak için üçüncü bir veri seti yani doğrulama kümesi kullanılır [50].

3.1.8 Karışıklık Matrisi ve Doğruluk Ölçütleri

Sınıflandırma problemlerinde tahmin edilmeye çalışılan durum bir kategori olması nedeniyle, eğitilen modelin başarısını değerlendirirken kategorilerin doğru ya da yanlış sınıflandırılması önemlidir. Kategorilerin ya da etiketlerin ne kadarının doğru nekadarının yanlış sınıflandırıldığını gösteren matrise karışıklık matrisi denir.

Tablo 3.1 örnek bir karışıklık matrisidir. Burada C_1, C_2 ve C_3 kategorilerinde oluşan bir bağımlı değişken bulunmaktadır. Satırlardaki değerler kategorilerin veri

setindeki gözlenen değerlerine, sütunlar ise algoritma sonunda tahmin edilen sınıf değerlerini göstermektedir. A_{11}, A_{22} ve A_{33} değerleri yani köşegen değerleri doğru sınıflandırılan gözlem değerlerinin sayısını göstermektedir. Matriste geriye kalan diğer kısımlar yanlış sınıflandırılan gözlemlerin sayısını göstermektedir.

Tablo 3.1 Karışıklık matris örneği

		Tahmin Edilen Değerler		
		C_1	C_2	C_3
Gözlenen Değerler	C_1	A_{11}	A_{12}	A_{13}
	C_2	A_{21}	A_{22}	A_{23}
	C_3	A_{31}	A_{32}	A_{33}

Örneğin A_{11} değeri, C_1 kategorisinde olan ve analiz sonucunda yine C_1 kategorisine atanan gözlemlerin sayısını göstermektedir. A_{12} değeri C_1 kategorisinde olan ve analiz sonucunda yanlış sınıflandırılıp, C_2 kategorisine atanan değerlerin sayısıdır. Benzer şekilde, A_{13} değeri de C_1 kategorisinde bulunan ve verinin eğitimi sonrasında C_3 sınıfında tahmin edilen, yani yanlış sınıflandırılan değerlerin sayısıdır. Matrisin diğer tüm elemanları da benzer şekilde değerlendirilmelidir. Karışıklık matrisi sayesinde, hangi sınıfların doğru ya da yanlış atandığı belirlenebilir ve birbirine karışan kategorilerin görünmesini kolaylaştırır.

Karışıklık matrisi kullanılarak bazı oranlar hesaplanabilir ve bunlar yorumlamayada yardımcı olurlar. Doğruluk oranı (DO), karışıklık matrisindeki tüm doğru sınıflandırılan grup değerlerini gösteren ve köşegende yer alan sayıların matrisin tüm elamanlarının toplam sayısına bölünerek elde edilebilir.

$$DO = \frac{A_{11} + A_{22} + A_{33}}{A_{11} + A_{12} + A_{13} + A_{21} + A_{22} + A_{23} + A_{31} + A_{32} + A_{33}} \tag{3.5}$$

Doğruluk oranı (3.5) bir yüzdedir. Fakat burada dikkat edilmesi gereken nokta genel bir oran olması nedeniyle, her bir etiketin nasıl sınıflandırıldığı hakkında bilgiyi vermez. Bu nedenle hassasiyet (precision) ve duyarlılık (recall) gibi doğruluk ölçütlerine de ihtiyaç duyulur. Karışıklık matrisi üzerinden her bir grup için hassasiyet formülleri aşağıda yer almaktadır.

$$C_1 \, için \, Hassasiyet = \frac{A_{11}}{A_{11} + A_{12} + A_{13}} \tag{3.6}$$

$$C_2 \text{ için } Hassasiyet = \frac{A_{22}}{A_{21} + A_{22} + A_{23}} \quad (3.7)$$

$$C_3 \text{ için } Hassasiyet = \frac{A_{33}}{A_{31} + A_{32} + A_{33}} \quad (3.8)$$

Hesaplanan eşitliklerden de anlaşılabileceği gibi hassasiyet kategorinin doğru sınıflandırılma oranını göstermektedir. Bir diğer doğruluk ölçütü de duyarlılıktır. Karışılık matrisi üzerinden hesaplanan her bir kategori için formülleri aşağıdaki gibidir.

$$C_1 \text{ için } Duyarlilik = \frac{A_{11}}{A_{11} + A_{21} + A_{31}} \quad (3.9)$$

$$C_2 \text{ için } Duyarlilik = \frac{A_{11}}{A_{11} + A_{21} + A_{31}} \quad (3.10)$$

$$C_3 \text{ için } Duyarlilik = \frac{A_{11}}{A_{11} + A_{21} + A_{31}} \quad (3.11)$$

Duyarlılık doğru sınıflandırılan bir kategorinin, o kategoriye sınıflandırılan tüm değerlerin sayısına olan oranı olarak tanımlanabilir.

3.1.9 ROC (Receiver Operating Characteristic) Eğrisi

Doğru sınıflandırma oranını için görsel bir grafik olan ROC eğrisinden de faydalanılabilinir. Tek bir model için fayda sağlamasının yanından birden fazla modelin karşılaştırılması için de kullanılabilir. ROC eğrisi, x eksenindeki bir grubun yanlış sınıflandırılmış gözlemlerine karşı y eksenindeki o grubun doğru sınıfılandırılmış yüzdesinin bir grafiğini gösterir. Bir sınıflandırıcının sonucu sol üst köşeye ne kadar yakınsa, sınıflandırıcı o kadar iyi performans göstermiştir. (0,0) ile (1,1) arasındaki çapraz çizgi üzerinde bulunan herhangi bir sınıflandırıcı, tam olarak şans düzeyinde davranır (pozitif ve negatif sınıfların eşit derecede yaygın olduğu varsayılarak). Eğri altındaki alanı hesaplanabilir [50]. Eğri altında kalan alan ne kadar fazla ise modelin başarısı da okadar yüksektir denilebilir.

3.2 Karar Ağaçları

Genellikle analizlerde bağımlı değişkenin türüne göre yöntemin değiştiği söylenebilir. Bağımlı değişken, literatürde hedef, çıktı, yanıt gibi farklı isimler de alabilmektedir. Karar ağaçları da bağımlı değişken sayısal (sürekli) olduğunda

regresyon ağaçları, kategorik bir değişken olduğunda da sınıflandırma ağaçları olarak adlandırılmaktadır.

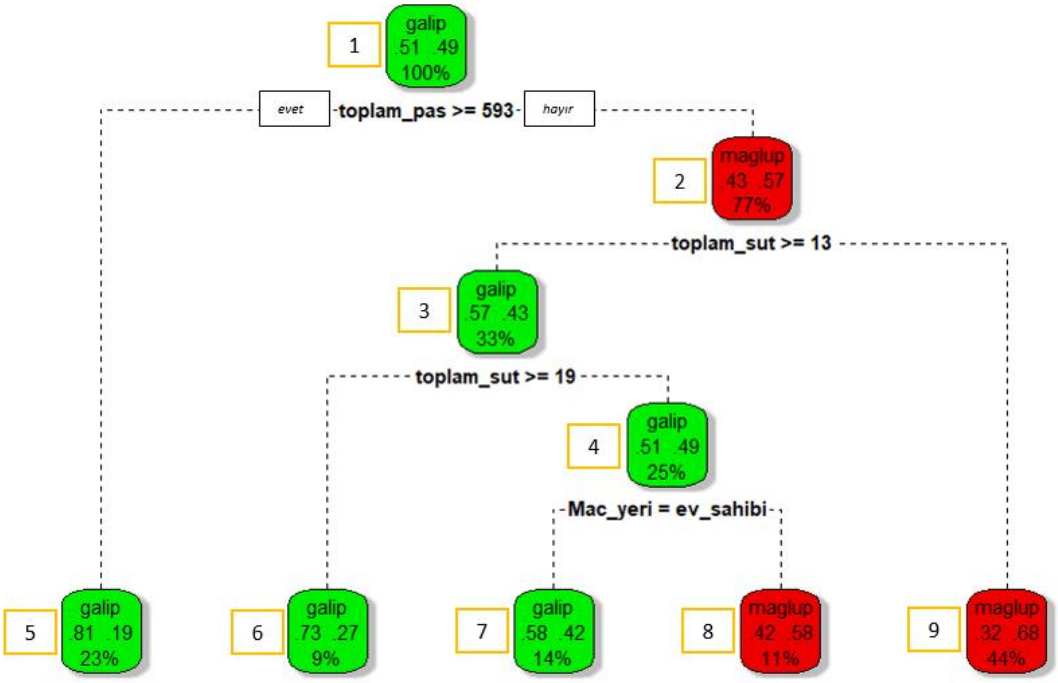
3.2.1 Sınıflandırma Ağaçları

Tablo 3.2 sınıflandırma ağaçlarını daha iyi anlatabilmek için kullanılacak örnek bir veri setini göstermektedir. Veri setinin 1484 gözlemi bulunmaktadır. Tezde kullanılan veri setinden düzenlenmiştir. Ayrıca maç sonucu, toplam şut, toplam pas ve maç yeri olmak üzere dört farklı değişken bulunmaktadır. Maç sonucu değişkeni galip ve mağlup olmak üzere iki kategoriden, maç yeri de ev sahibi ve deplasman olmak üzere yine iki kategoriden oluşmaktadır. Burada, toplam şut ve toplam pas değişkenleri sürekli değişkenlerdir.

Tablo 3.2'de verilen örnek veri setinden maç sonucu değişkeni ya da maç yeri değişkeni bağımlı değişken olarak alınarak karar ağacı uygulanmak istendiğinde değişkenlerin yapısı gereği bu sınıflandırma ağacı olacaktır. Fakat toplam şut ya da toplam pas değişkenleri kullanılarak oluşturulmak istenirse, regresyon ağacı olacaktır. Bu tez çalışmasında, karar ağaçları, sınıflandırma ağaçları olarak kullanıldığı için bu bölüm sınıflandırma üzerinden anlatılacaktır.

Tablo 3.2 Örnek veri seti

Maç Sonucu	Toplam şut	Toplam pas	Maç yeri
galip	9	618	ev sahibi
mağlup	10	555	deplasman
mağlup	6	574	ev sahibi
galip	20	555	deplasman
galip	10	674	ev sahibi
mağlup	14	369	deplasman
mağlup	13	572	ev sahibi
galip	17	737	deplasman
galip	20	796	ev sahibi
mağlup	7	360	deplasman
mağlup	6	289	ev sahibi
galip	15	549	deplasman
.	.	.	.
.	.	.	.
.	.	.	.



Şekil 3.2 Örnek karar ağacı modeli

Şekil 3.2'de verilen karar ağacı Tablo 3.2'de verilen örnek üzerinden hesaplanmıştır. Burada bağımlı değişken, yani sınıflandırılmaya çalışılan değişken maç sonucudur. Toplam şut, toplam pas ve maç yeri değişkenleri bağımsız değişkenlerdir. Bağımsız değişkenler literatürde, özellik, girdi, açıklayıcı gibi isimler de almaktadır. Daha iyi anlatabilmek için şekildeki kök düğüm, iç düğümler ve yapraklar (yaprak düğümler) 1'den 9'a kadar numaralandırılmıştır. Karar ağaçları kök düğüm ile başlar. Yani 1 numaralı düğüm kök düğümdür. Karar ağaçları kök düğümden başladığı için dallanmalarda kök düğüme giriş yoktur sadece çıkış vardır. Şekilde 2, 3 ve 4 numaralı düğümler iç düğümlerdir. Bu düğümlere hem giriş hem de çıkış vardır. Genellikle giriş olduktan sonra herhangi bir bağımsız değişken değerinin iki alt alana bölünmesiyle çıkış olur. Son olarak, şekilde 5, 6, 7, 8 ve 9 numaralı düğümler yaprak düğümlerini gösterir. Yaprak düğümleri de aslında karar düğümleridir ve sadece giriş olur, çıkış olmaz. Herhangi bir düğümdeki bağımsız değişken kategorik ise bölünme kategorilere göre olmalıdır. Sayısal bir bağımlı değişken var ise bölünmedeki koşul sayısal bir aralığı ifade eder. Ayrıca, şekilden görülebileceği gibi "evet" yazan yön yani sol

taraf düğümde verilen koşulun sağlandığı, “hayır” yazan kısım ise koşulun sağlanmadığı anlamına gelmektedir.

Şekil 3.2’de kök düğümde iki olasılık yazmaktadır. Bunlar 0,51 ve 0,49 ‘dur. Bağımlı değişkenin maç sonucu ve kategorilerinin “galip” ve “mağlup” olmasından kaynaklı bu iki olasılık kategorilerin olasılığıdır. Karar ağacındaki tüm düğümler yüksek olasılığın renkine boyanmıştır ve üst kısmında da kategorinin ismi yer almaktadır. Alt kısımda da koşulun verideki yüzdesi verilmiştir. Yani kök düğüm için konuşulacak olursa, verinin tamamı için galip gelme olasılığı %51 ve mağlup olma olasığı %49’dur. Yine kök düğüm için koşul olarak toplam pas “593 ve üzeri” ya da 593’ten daha az olarak görülmektedir. Koşul sağlandığında, yani toplam pas 593 ve fazlası olduğunda galip olma olasılığı %81’e çıkmaktadır. Bu durumda mağlup olma olasılığı %19’dur (5 numaralı düğüm). Diğer taraftan bakıldığında kök düğümdeki koşul sağlanmazsa, mağlup olma olasılığı 0,57’ye yükselmektedir. Toplam pas 593’ten az olduğunda, toplam şut da 13’ten az olduğu durumda takımın mağlup olma olasılığı %68’e yükselmektedir. Bu durumla karşılaşma olasılığı ise 0,44 ‘tür (9 numaralı düğüm). Toplam pas 593’ten az ise, $13 \leq$ toplam şut < 19 olduğu durumda galip olma olasılığı %51’dir. Bu koşullar sağlandığında takım, ev sahibi ise galip olma olasılığı %58’e çıkarken, eğer takım deplasmanda oynuyorsa bu oran %42’ye gerilemektedir.

3.2.2 Bölünme Kriterleri

Bir A düğümü sol (A_{SOL}) ve sağ (A_{SAG}) kol olarak ikiye ayrılırsa,

$$P(A_{SOL})r(A_{SOL}) + P(A_{SAG})r(A_{SAG}) \leq P(A)r(A) \tag{3.12}$$

Burada P(A), A’nın olasılığını ve r(A) ise A’ya atanmış sınıfı gösterir. Bunlar yardımıyla, bir karar ağacı elde etmek için, riski azaltarak yani Δr 'yi maksimize eden bölünmeyi seçmek gerekir. Bu durum çok fazla işlem gerektiren bir süreçtir. Bunun yerine, çeşitli safsızlık (impurity) veya çeşitlilik ölçülerinden faydalanılabilinir. Safsızlık fonksiyonu f olduğunda, A düğümünün safsızlığı (3.13) şu şekilde tanımlanabilir [51].

$$I(A) = \sum_{i=1}^{C} f(P_i A) \tag{3.13}$$

Burada, P_iA, i sınıfına ait olanların oranıdır. A safken I(A) değeri sıfır olması gerektiği için, f ile $f(0) = f(1) = 0$ içbükey olmalıdır. Bilgi indeksi f için $f(p) = -p\log(p)$ ve Gini indeksi $f(p) = p(1-p)$ olarak tanımlanabilir. Daha sonra bu bölünme maksimum safsızlık azaltımı ile kullanılırsa eşitlik (3.14) elde edilebilir [51].

$$\Delta I = p(A)I(A) - p(A_L)I(A_L) - p(A_R)I(A_R) \tag{3.14}$$

Diğer bir kriter, ikiye ayırmadır.

$$I(A) = \min_{C_1C_2}\left[f(pc_1) + f(pc_2)\right] \tag{3.15}$$

C sınıflarının iki ayrık kümeye bölünmesi C_1 ve C_2 olarak gösterilebilir. İkiye bölünme, f için olağan safsızlık endeksine eşdeğerdir. İkiye ayırmanın bir avantajı, çıktının kullanıcıya verilerin yapısıyla ilgili ek bilgi sağlayabilmesidir. C'nin iki üst sınıfa bölünmesi olarak görülebilir ve bunlar bir anlamda A'daki gözlemlerden en farklı olanıdır. Bazı problemler için bağımlı değişkenin kategorilerinin doğal bir sıralaması olabilir [51].

3.2.3 Genelleştirilmiş Gini İndeksi

Bir nesnenin olasılıklara göre C sınıflarından birinden rastgele seçildiği ve aynı dağılım kullanılarak bir sınıfa rastgele atandığı varsayıldığında yanlış sınıflandırma olasılığı eşitlik (3.16) gibi olacaktır [51].

$$\sum_i \sum_{j\neq i} p_i p_j = \sum_i \sum_j p_i p_j - \sum_i p_i^2 = \sum_i p_i^2 \tag{3.16}$$

Bu aynı zamanda p için Gini indeksidir. L(i,j), j sınıfını aslında i sınıfına ait olan bir nesneye atamanın kaybı olduğunda, yanlış sınıflandırmanın beklenen maliyeti eşitlik (3.17) gibi olacaktır. Bu durum, genelleştirilmiş bir Gini safsızlık endeksinin tanımlanmasını açıklar [51].

$$G(p) = \sum_i \sum_j L(i,j) p_i p_j \tag{3.17}$$

Karşılık gelen bölme kriteri, değişkenin yanlış sınıflandırma maliyetlerini içeren uygulamalar için ümit verici görünmektedir. Ancak, ilk olarak safsızlık ölçümlerinin arkasındaki faktör olan olasılığın mutlaka içbükey bir fonksiyonu

değildir. Daha önemlisi, gini kullanmadan önce kayıp matrisini simetrikleştirir [51].

$$G(p) = (1/2)\sum\sum[L(i,\mathrm{j}) + \mathrm{L}(\mathrm{j},\mathrm{i})]\,p_i\,p_j \tag{3.18}$$

Özellikle, iki sınıflı problemler için, gini kayıp matrisini göz ardı eder.

3.3 Destek Vektör Makineleri

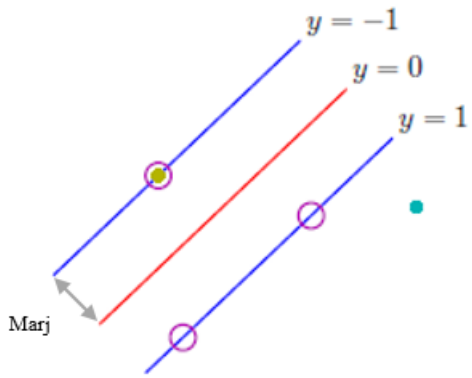
Destek Vektör Makineleri maksimum marjlı doğrusal diskriminantlara dayalı bir sınıflandırma yöntemidir. Yani amaç, sınıflar arasındaki boşluğu veya marjı en üst düzeye çıkaran optimal hiperdüzlemi bulmaktır. Ayrıca, bazı yüksek boyutlu "doğrusal olmayan" uzayda bir hiper düzleme karşılık gelen sınıflar arasındaki optimum doğrusal olmayan karar sınırını bulmak için çekirdek fonksiyonları da kullanılabilir [52].

3.2.4 Ayrılabilir Sınıf Dağılımları

Doğrusal modeller kullanarak iki sınıflı sınıflandırma problemi için denklem (3.19) aşağıdaki gibi verilebilir:

$$y(x) = w^T\phi(x) + b \tag{3.19}$$

Burada, $\phi(x)$ sabit bir özellik-uzay dönüşümünü ve b de bias parametresinin açık halini belirtir. Çekirdek (kernel) fonksiyonları cinsinden ifade edilen ikili bir gösterimdir. Eğitim veri seti, N giriş vektörü $[x_1,\ldots,x_N]$ olmak üzere, $t_N\epsilon\{-1,1\}$ koşulu ile $[t_1,\ldots,t_N]$ de hedef değişkene karşılık gelmektedir. Yeni x veri noktaları y(x)'in işaretine göre sınıflandırılır. Eğitim veri setinin doğrusal olarak ayrılabilir olduğu varsayılırsa, w ve b parametrelerinin en az bir seçimi vardır. Eğitim veri setini tam olarak sınıflandıran birden fazla çözüm varsa, o zaman en küçük genelleme hatasını verecek olan bulunmaya çalışılmalıdır. Destek vektör makineleri, Şekil 3.3'te gösterildiği gibi, karar sınırı ile herhangi bir örnek arasındaki en küçük mesafe olarak tanımlanan marj kavramı aracılığıyla çözüme ulaşmaya çalışır. Destek vektör makinelerinde, karar sınırı marjın maksimize edildiği sınır olarak seçilir. Maksimum marj çözümü, istatistiksel öğrenme teorisi kullanılarak yapılabilir [53, 54].



Şekil 3.3 Marj kavramının görsel olarak gösterilmesi [53]

Herhangi bir x noktasının y(x)=0 ile tanımlanan bir hiper düzleme olan dikey mesafesi |y(x)|/||w|| olarak hesaplanabilir. Tüm veri noktalarının doğru şekilde sınıflandırıldığı çözüm için, bir x_n noktasının karar yüzeyine olan uzaklığı denklem (3.2) gibi verilebilir.

$$\frac{t_n y(x_n)}{\|w\|} = \frac{t_n(w^T\phi(x_n)+b)}{\|w\|} \tag{3.20}$$

Marjın, veri setinden en yakın x_n noktasına dik mesafeyle verilir ve bu mesafeyi maksimize etmek için w ve b parametreleri optimize edilir. Böylece maksimum marj çözümü, denklem (3.21) çözülerek bulunur [53].

$$\arg\max_{w,b}\left\{\frac{1}{\|w\|}\min_n\left[t_n\left(w^T\phi(x_n)+b\right)\right]\right\} \tag{3.21}$$

Bu optimizasyon probleminin doğrudan çözümü çok karmaşık olacağından çözmesi çok daha kolay hale getirilebilir. Bu nedenle, w ve b kernel fonksiyonu cinsinden yeniden ölçeklendirilebilir. Bu durumda herhangi bir x_n noktasından karar yüzeyine $t_n y(x_n)/||w||$ ile verilen mesafenin değişmeyeceği söylenebilir. Bu durumda, yüzeye en yakın nokta için eşitlik (3.22) kullanılabilir [53, 54].

$$t_n\left(w^T\phi(x_n)+b\right)=1 \tag{3.22}$$

Tüm veri noktaları aşağıdaki kısıtlamayı (3.23) sağlamalıdır [53, 54].

$$t_n\left(w^T\phi(x_n)+b\right)\geq 1 \tag{3.23}$$

Bu durum hiper düzleminin kanonik temsili olarak bilinir [55]. Optimizasyon problemi basitçe $||w||^{-1}$ maksimize etmeye eşdeğer olan $||w||^2$ 'yi minimizime etmeyi gerektirir. Bu nedenle, denklem (3.24)'ün çözülmesi gereklidir [53].

$$\arg\min_{w,b} \frac{1}{2}\|w\|^2 \tag{3.24}$$

Bias parametresinin optimizasyonda yer almadığı için kısıtlamalar yoluyla dolaylı olarak belirlenir. Optimizasyon problemini çözmek için, Lagrangian fonksiyonundan (3.25) yararlanılır [56].

$$L(w,b,\mathrm{a}) = \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} a_n \left\{ t_n (w^T \phi(x_n) + b) - 1 \right\} \tag{3.25}$$

Burada, $(a = a_1,...,a_N)$ olmak üzere w ve b 'ye göre minimize ve α 'ya göre maksimize edilir. Fonksiyonun w ve b 'ye göre türevleri alınıp sıfıra eşitlenirse, koşul (3.26) ve (3.27) elde edilir [53, 56].

$$w = \sum_{n=1}^{N} a_n t_n \phi(x_n) \tag{3.26}$$

$$0 = \sum_{n=1}^{N} a_n t_n \tag{3.27}$$

Son durumda, denklem aşağıdaki gibi olacaktır.

$$\tilde{L}(\mathrm{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(x_n, x_m) \tag{3.28}$$

Denkleme ait kısıtlamalar ise,

$$a_n \geq 0, \quad n = 1,...,N \tag{3.29}$$

$$\sum_{n=1}^{N} a_n t_n = 0 \tag{3.30}$$

Çekirdek formülasyonu, Lagrangian fonksiyonunun sınırlanmasını sağlayarak iyi tanımlanmış bir optimizasyon problemine yol açar [53]. Eğitilmiş modeli kullanarak yeni veri noktalarını sınıflandırmak için y(x) eşitlik (3.31) gibi gösterilebilir.

$$y(x) = \sum_{n=1}^{N} a_n t_n k(x, x_n) + b \tag{3.31}$$

Bu durumda (3.32), (3.33) ve (3.34) kısıtlarını sağlaması gereklidir.

$$a_n \geq 0 \tag{3.32}$$

$$t_n y(x_n) - 1 \geq 0 \tag{3.33}$$

$$a_n \{t_n y(x_n) - 1\} = 1 \tag{3.34}$$

Böylece, her veri noktası için ya $\alpha_n = 0$ ya da $t_n y(x_n) = 1$ durumu geçerli olacaktır. Bu durumda $\alpha_n = 0$ olan herhangi bir veri noktası görünmeyecektir ve bu nedenle yeni veri noktaları için tahminlerde bulunmada hiçbir rol oynamayacaktır. Kalan veri noktaları, destek vektörleri olarak adlandırılır [54]. Koşul olarak, $t_n y(x_n) = 1$ 'i sağladıkları için, özellik uzayındaki maksimum marj hiper düzlemlerinde yer alan noktalara karşılık gelirler. Model eğitildikten sonra, veri noktalarının yalnızca destek vektörleri kullanılabilir [53]. Ayrıca, α için bir değer bulduktan sonra, herhangi bir x_n destek vektörünün $t_n y(x_n) = 1$ 'i sağladığı belirtilerek eşik parametresi b 'nin değeri eşitlik (3.35) gibi belirlenebilir.

$$t_n \left(\sum_{m \in S} a_m t_m k(x_n, x_m) + b \right) = 1 \tag{3.35}$$

Burada S, destek vektörlerinin indis kümesini gösterir. Bu denklem herhangi seçilen bir destek vektörü x_n kullanarak çözülebilir, ancak denklem (3.36)'daki gibi tüm destek vektörleri üzerinden ortalaması alınarak sayısal olarak daha kararlı bir çözüm elde edilebilir [53]. Burada N_S destek vektörlerinin toplamıdır.

$$b = \frac{1}{N_S} \sum_{n \in S} \left(t_n - \sum_{m \in S} a_m t_m k(x_n, x_m) \right) \tag{3.36}$$

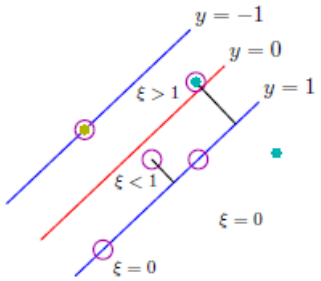
Alternatif modellerle karşılaştırma yapabilmek için, maksimum marj sınıflandırıcısını bir hata fonksiyonunun en aza indirilmesi açısından denklem (3.37) gibi ifade edilebilir [53].

$$\sum_{n=1}^{N} E_\infty (y(x_n) t_n - 1) + \lambda \|w\|^2 \tag{3.37}$$

3.2.5 Çakışan Sınıf Dağılımları

Eğitim veri noktaları özellik uzayında doğrusal olarak ayrılabilir olmayabilir. Destek vektör makinesine karşılık gelen karar sınırı doğrusal olmasa bile, eğitim verilerinin tam olarak ayrılmasını sağlayacaktır. Fakat sınıf dağılımları çakışması nedeniyle, eğitim verilerinin tam olarak ayrılması yanlış sınıflandırmalara neden olacaktır. Bu durumda, veri noktalarının marj sınırının yanlış tarafında olmasına

izin vererek, bu sınırdan uzaklaştıkça artan bir ceza vererek çözümlemeye çalışılabilir. Bu ceza mesafenin doğrusal bir fonksiyonu olarak tanımlanabilir. Bunun için gevşek (slack) değişkenler tanımlanır. Şekil 3.4 gevşek değişkenlerin gösterimini sağlar. Bunlar, doğru marj sınırı üzerinde veya içinde bulunan veri noktaları için $\xi_n=0$ ve diğer noktalar için $\xi_n=|t_n-y(x_n)|$ olarak tanımlanır. Bu değişkenler yardımıyla, karar sınırı $y(x_n)=0$ üzerindeki bir veri noktası $\xi_n=1$ olacak ve $\xi_n>1$ olan noktalar yanlış sınıflandırılacaktır. Bu durumda, kesin sınıflandırma kısıtlamaları $t_n-y(x_n)\geq 1-\xi_n$ olacaktır [53, 56].



Şekil 3.4 Gevşek değişkenlerin gösterimi [53]

$\xi_n=0$ olan veri noktaları doğru şekilde sınıflandırılır ve marjın doğru tarafında veya marjın üzerinde bulunur. Ayrıca $0\leq\xi_n\leq 1$ marjın içinde ancak karar sınırının doğru tarafında yer aldığı noktalar ve $\xi_n>1$ olan veri noktaları, karar sınırının yanlış tarafında yer alır ve yanlış sınıflandırılır. Bazı eğitim setleri, veri noktalarının yanlış sınıflandırılmasına izin verir. Gevşek değişkenleri örtüşen sınıf dağılımlarına izin verirken, aykırı değerlere duyarlıdır çünkü yanlış sınıflandırma cezası ξ ile doğrusal olarak artar [53]. Bunun çözümü için, Marj sınırının yanlış tarafında bulunan noktaları yumuşak bir şekilde cezalandırırken marjı maksimize etmeye çalışmaktır. Bu nedenle denklem (3.38) kullanılabilir [57].

$$C\sum_{n=1}^{N}\xi_n+\frac{1}{2}\|w\|^2 \tag{3.38}$$

Burada $C>0$ parametresi gevşek değişken cezası ile marj arasındaki ilişkiyi kontrol eder. Yanlış sınıflandırılan herhangi bir noktanın $\xi_n>1$ olması nedeniyle, $\Sigma_n\xi_n$ yanlış sınıflandırılan noktaların sayısının üst sınırını göstermektedir [53]. Karşılık gelen Lagrangian denklemi (3.39)'da verilmektedir.

$$L(w,b,a)=\frac{1}{2}\|w\|^2+C\sum_{n=1}^{N}\xi_n-\sum_{n=1}^{N}a_n\{t_n y(x_n)-1+\xi_n\}-\sum_{n=1}^{N}\mu_n\xi_n \quad (3.39)$$

α_n ve μ_n Lagrange çarpanlarıdır. Karşılık gelen kısıtlamalar ise n=1,…,N olmak üzere sırasıyla şu şekilde verilir [54, 55]. $\alpha_n \geq 0$, $t_n y(x_n)-1+\xi_n \geq 0$, $\alpha_n(t_n y(x_n)-1+\xi_n)=0$, $\mu_n \geq 0$, $\xi_n \geq 0$, $\mu_n \xi_n \geq 0$.

Doğrusal model formundaki y(x) kullanılarak w, b ve ξ_n aşağıdaki gibi optimize edilebilir [53, 57].

$$\frac{\partial L}{\partial w}=0 \rightarrow w=\sum_{n=1}^{N}a_n t_n \phi(x_n) \quad (3.40)$$

$$\frac{\partial L}{\partial b}=0 \rightarrow \sum_{n=1}^{N}a_n t_n=0 \quad (3.41)$$

$$\frac{\partial L}{\partial \xi_n}=0 \rightarrow a_n=C-\mu_n \quad (3.42)$$

Bu sonuçları kullanarak, ikili (dual) Lagrangian formu eşitlik (3.43)'teki gibi elde edilir [58].

$$\tilde{L}(\mathrm{a})=\sum_{n=1}^{N}a_n-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}a_n a_m t_n t_m k(x_n,x_m) \quad (3.43)$$

Bu nedenle, ikili değişkenlere göre α_n 'i en aza indirmesi gerekir. Kısıtlamalar $0 \leq \alpha_n \leq C$ ve $\Sigma\, \alpha_n t_n=0$ olarak verilebilir. Bunlar kutu kısıtlamaları olarak geçmektedir [57, 59]. Artık ortaya çıkan çözüm yorumlanabilir. Daha önce olduğu gibi, veri noktalarının bir alt kümesi $\alpha_n=0$ olabilir, bu durumda tahmin modeline katkıda bulunmazlar. Kalan veri noktaları, destek vektörlerini oluşturur. Bunlar $\alpha_n>0$ koşulundadır ve dolayısıyla eşitlik (3.44)'e uymalıdır [53].

$$t_n y(x_n)=1-\xi_n \quad (3.44)$$

Eğer bir $\alpha_n<C$ ise, $\xi_n=0$ 'dır ve $M_n>0$ olacaktır. Bu nedenle bu tür noktalar marjın üzerinde bulunur. $\alpha_n=C$ olan noktalar marjın içinde olabilir ve $\xi_n \leq 1$ ise doğru, $\xi_n>1$ ise yanlış sınıflandırılabilir. Parametre b 'yi belirlemek için, sayısal olarak kararlı bir çözüm için, destek vektörlerinin ortalamaları alınarak elde edilir [53, 57]. Burada, M veri noktalarının $0<\alpha_n<C$ durumunda indis kümesini gösterir.

$$b=\frac{1}{N_M}\sum_{n\in M}\left(t_n-\sum_{m\in S}a_m t_m k(x_n,x_m)\right) \quad (3.45)$$

3.2.6 Çekirdek (Kernel) Fonksiyonları

Eğer sınıflandırma problem doğrusal olarak ayırlabilir değilse, destek vektör makineleri uygun bir çekirdek fonksiyonu ile güçlendirilebilir. Bu durum DVM'nin çekirdek uzayında bir hiperdüzlem bulmasını sağlar ve bu sayede doğrusal olmayan bir ayrım yapabilir. Doğrusal olmayan kernel radyan tabanlı Gaussian denklem (3.46)'daki gibi verilebilir [60].

$$\kappa(x, x') = \exp(-\frac{\|x - x'\|^2}{2\sigma^2} = \exp(-\gamma \|x - x'\|^2) \quad (3.46)$$

Burada, σ kernel genişliğini gösterir ve eğer 0'a yakınsa, DVM aşırı uyuma meyillidir. Tam tersi durumda, yani kernel genişliğine büyük bir değer atanması durumunda, az uyuma neden olabilir ve bu durumda tek bir sınıfa atanma ile sonuçlanabilir. Bundan ötürü kernek genişliğinin seçimi önemlidir [60]. Ayrıca burada, γ'de çekirdek ölçek parametresini göstermektedir.

Radyal tabanlı Gaussian çekirdek ölçek parametresi (γ), üç farklı durum için tanımlıdır. Çekirdek ölçek parametresi $\sqrt{n/4}$ olaral alındığında fine Gaussian, $\sqrt{n}$ olarak alındığında medium Gaussian ve $4\sqrt{n}$ olarak alındığında ise, coarse Gaussian olarak isimlendirilir. Burada n, özellik veya boyut sayısıdır. Yani bağımsız değişkenlerin sayısı olduğu söylenebilir.

Diğer yaygın kullanılan kernel fonksiyonu da polinomsal fonsiyondur.

$$\kappa(x, x') = (1 + x^T x')^p \quad (3.47)$$

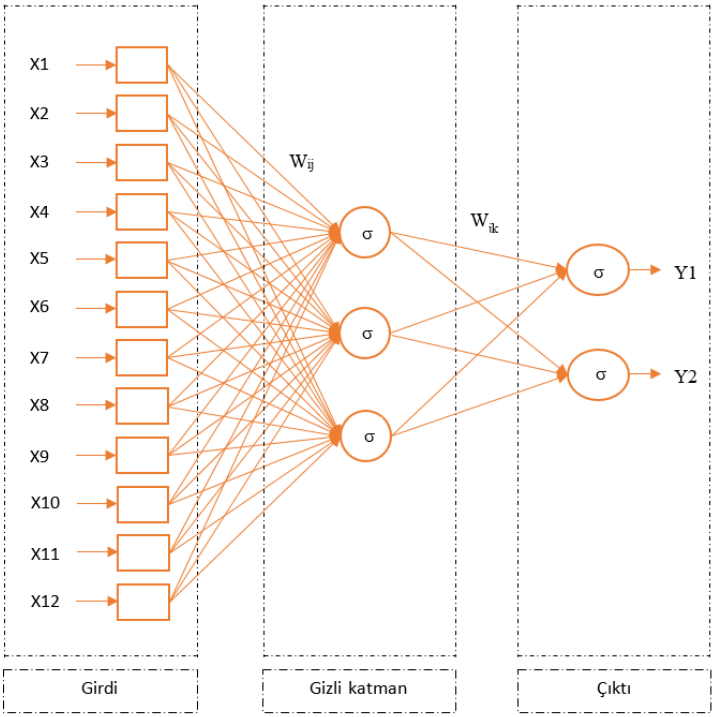
Burada, p polinomsal kernel fonksiyonunun derecesidir. En düşük dereceli polinomsal fonksiyon doğrusal kernel gösterir. Derecenin arttırılması sınıflayıcının esnekliğini arttırır. Son olarak, p değerinin 2 olması karesel (quadratic) ve 3 olması da kübik (cubic) olarak tanımlanır.

3.4 Yapay Sinir Ağları

Yapay sinir ağları, biyolojik sinir ağlarından esinlenilmiş yapay nöron ve sinaps ağlarıdır. Nöronlar, gelen tüm sinapslardan gelen girdiyi toplar, doğrusal olmayan bir işlev uygular ve hesaplamanın sonucunu tüm giden sinapslara verir. En temelde ileri beslemeli ağlar ve geri beslemeli ağlar olarak ikiye ayrılabilir. İleri

beslemeli bir ağın ilk katmanına giriş katmanı, son katmanına ise çıkış katmanı denir. Aradaki katmanlara gizli katmanlar denir. İşleme akışı girdiden çıktı katmanına gider, ancak tam tersi yönde hareket etmez [61].

Şekil 3.5'te çok katmanlı ve ileri beslemeli bir sinir ağı örneği verilmiştir. Burada, 12 bağımlı değişkenden oluşan bir girdi katmanı mevcuttur. Tek bir gizli katman bulunmaktadır ve 3 nörondan oluşmaktadır. Çıktı katmanında ise, iki kategoriden oluşan bir bağımlı değişken bulunmaktadır. σ işareti sigmoid fonksiyonu göstermektedir. Ayrıca, w_{ij} girdi katmanı ile gizli katman arasındaki, w_{ik} ise gizli katman ile çıktı katmanı arasındaki ağırlıkları göstermektedir.



Şekil 3.5 Çok katmanlı ve ileri beslemeli yapay sinir ağları örneği

İleri beslemeli sinir ağları temel olarak, bir giriş katmanı ve bir çıkış katmanından oluşan tek katmanlı algılayıcı, çok katmanlı algılayıcı (gizli katmanlı) ve radyal tabanlı fonksiyonlu ağlar olmak üzere üç kategoride incelenebilir [62]. Çok katmanlı ileri beslemeli ağlar genellikle her bir hesaplama birimi için eşik (threshold) fonksiyonunu veya sigmoid fonksiyonunu kullanır. Genel olarak konuşulacak olursa, temelde ağırlıkları belirlemek için geri yayılma kullanan bir

öğrenme algoritmasıdır [62]. Çok katmanlı algılayıcılar, standart geri yayılım algoritması ile eğitilmiş ileri beslemeli sinir ağları olmaları nedeniyle, eğitilmeleri için istenen bir yanıtı gerektirdikleri için, girdi verilerini istenen bir yanıta nasıl dönüştüreceklerini öğrenirler. Bu nedenle sınıflandırma problemleri için yaygın olarak kullanılırlar [63].

Yapay sinir ağlarında gizli katmanlarla ilgili olarak sinir ağında gerçekte kaç tane gizli katmana sahip olunacağı ve bu katmanların her birinde kaç nöron olacağı olmak üzere verilmesi gereken iki karar vardır. İki gizli katman gerektiren problemlere nadiren rastlanır. Bununla birlikte, iki gizli katmana sahip sinir ağları, herhangi bir şekle sahip fonksiyonları temsil edebilir. Şu anda ikiden fazla gizli katmana sahip sinir ağlarını kullanmak için teorik bir neden yoktur. Gizli katmanlarda çok fazla nöron kullanılması aşırı uyum problemine yol açabilir. Az nöron kullanılması ise az uyum sorununa yol açar. Bu nedenle, nöron sayısının belirlenmesi için önerilen birkaç durum vardır [63]. Bunlar şu şekilde sıralanabilir. (i) Gizli nöronların sayısı, girdi katmanının boyutu ile çıktı katmanının boyutu arasında olmalıdır. (ii) Gizli nöronların sayısı, girdi katmanının boyutunun 2/3'ü artı çıktı katmanının boyutu kadar olmalıdır. (iii) Gizli nöronların sayısı, giriş katmanının boyutunun iki katından az olmalıdır. Fakat yine de, bu öneriler sadece başlangıç noktasıdır. Deneme yanılma yoluyla kurulan model iyileştirilebilir.

3.4.1 Bayes Sinir Ağları ile Sınıflandırma

İki sınıflı bir sınıflandırma problemine karşılık gelen tek bir lojistik sigmoid çıktısına sahip bir ağ düşünüldüğünde log olabilirlik fonksiyonu eşitlik (3.48)'deki gibi verilebilir [64]:

$$\ln p(D|w) = \sum_{n} 1^{N} \{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} \tag{3.48}$$

Burada, $t_n \epsilon \{0,1\}$ hedef değerlerdir. Bu fonsiyonda β olmamasının sebebi, veri noktalarının doğru etiketlendiğinin varsayılmasından kaynaklanır. Laplace çerçevesinin bu modele uygulanmasının ilk aşaması α hiperparametresinin başlatılması ve sonrasında log sonsal (posterior) dağılımını maksimize ederek parametre vektörü w'yi belirlemektir. Bu, düzenlenmiş hata fonksiyonunu (3.49) en aza indirmeye eşdeğerdir [53, 64].

$$E(w) = -\ln p(D|w) + \frac{\alpha}{2} w^T w \tag{3.49}$$

Standart optimizasyon algoritmalarıyla birlikte hata geri yayılımı kullanılarak elde edilebilir. Ağırlık vektörü için bir w_{MAP} (MAP: maksimum sonsal) çözüm bulduktan sonraki adım, negatif log olabilirlik fonksiyonunun ikinci türevlerini içeren Hessian matrisini değerlendirmektir. Bu kesin yöntemi veya dış çarpım yaklaşımı kullanılarak yapılabilir. Negatif log sonsal dağılımının ikinci türevleri biçiminde ve sonsal Gauss yaklaşımı ile verilebilir. Hiperparametre α 'yı optimize etmek için, marjinal olasılığı yeniden maksimize ederek denklem (3.50)'deki gibi gösterilebilir [53, 64].

$$\ln p(D|\alpha) \simeq -\mathrm{E}(w_{MAP}) - \frac{1}{2}\ln|A| + \frac{W}{2}\ln\alpha + sabit \tag{3.50}$$

Düzenlenmiş hata fonksiyonu da eşitlik (3.51)'deki gibi tanımlanır.

$$E(w_{MAP}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} + \frac{\alpha}{2} w_{MAP}^T w_{MAP} \tag{3.51}$$

Bu kanıt fonksiyonunu α 'ya göre maksimize etmek denklem (3.52) tahmin denklemine yol açar.

$$\alpha = \frac{\gamma}{w_{MAP}^T w_{MAP}} \tag{3.52}$$

Etkin parametre sayısını γ gösterir ve eşitlik (3.53)'de gösterildiği gibi tanımlanır:

$$\gamma = \sum_{i=1}^{W} \frac{\lambda_i}{\alpha + \lambda_i} \tag{3.53}$$

Burada, λ düzenleme parametresidir. Ayrıca, eşitlik (3.54)'te verilen tahmini dağılıma ihtiyaç vardır. Bu birleşim, ağ fonksiyonunun doğrusal olmaması nedeniyle hesaplanması açısından zorlayıcı olacaktır [53, 64].

$$p(t \mid \mathrm{x}, \mathrm{D}) = \int p(t \mid x, w)\, q(w \mid D)\, dw \tag{3.54}$$

MAP maksimum sonsal (maximum posterior) yaklaşımını göstermektedir [53, 64]. Sonsal dağılımın dar olduğu varsayılırsa formül eşitlik (3.55)'deki gibi yazılabilir.

$$p(t \mid \mathrm{x}, \mathrm{D}) \simeq \mathrm{p}(\mathrm{t} \mid \mathrm{x}, \mathrm{w}_{MAP}) \tag{3.55}$$

Bununla birlikte, sonsal dağılımın varyansı kullanılarak iyileştirme yapılabilir. Bu nedenle, formdaki çıktı birimi aktivasyonu için doğrusal bir yaklaşımla eşitlik (3.56)'daki gibi yapılabilir.

$$a(x,w) \simeq a_{MAP}(x) + b^T(w - w_{MAP}) \tag{3.56}$$

Buradaki denklem geri yayılımla bulunabilir. Artık w üzerinden sonsal dağılım için bir Gauss yaklaşımı ve w 'nin doğrusal fonksiyonu olan α için bir model olduğundan, ağ ağırlıkları üzerinden dağılımın neden olduğu çıktı birimi aktivasyon değerlerinin dağılımı denklem (3.57)'deki şekilde olacaktır [64, 65].

$$p(a \mid x, D) = \int \delta(a - a_{MAP}(x) - b^T(\mathrm{x})(\mathrm{w} - \mathrm{w}_{MAP}))\, q(w \mid D)\, dw \tag{3.57}$$

Bu dağılımın ortalaması $\alpha_{MAP}=\alpha(x,w_{MAP})$ ve varyansı $\sigma_\alpha^2(x) = b^T(x)A^{-1}b(x)$ olan Gauss dağılımına sahip olduğu söylenebilir. Son olarak, tahmine dayalı dağılımı elde etmek için, denklem (3.57) eşitlik (3.58)'deki gibi α için marjinalleştirilebilir [53, 64].

$$p(t = 1 \mid x, D) = \int \sigma(a)\, p(a \mid x, D)\, da \tag{3.58}$$

Bir normal dağılımın lojistik sigmoid ile birleşiminin hesaplanması zor olmasın nedeniyle aşağıdaki MacKay [66] tarafından önerilen yaklaşımı denklem (3.59)'daki gibi gösterilebilir.

$$p(t = 1 \mid x, D) = \sigma(\kappa(\sigma_a^2) b^T w_{MAP}) \tag{3.59}$$

Burada κ(.) eşitlik (3.60) şekilde tanımlanır.

$$\kappa(\sigma^2) = \sqrt{(1 + \pi\sigma^2 / 8)} \tag{3.60}$$

Bu aşamadan sonra sınıflandırma kararları için, elde edilen en olası ağırlık vektörüne karşılık gelen çıktı ile elde edilen marjinalleştirilmiş çıktılar kullanılarak karşılaştırılır. Çıktı, yanlış sınıflandırma olasılığını en aza indirecek şekilde ağ girdisini sınıflandırmak için kullanılırsa, karar sınırı 0.5'lik bir ağ çıkışına karşılık gelir. En olası çıktı için, lojistik sigmoid aktivasyon fonksiyonunun, 0.5 için $\alpha(x,w_{MAP})=0$ 'a karşılık geldiği söylenebilir [64].

4
UYGULAMA

4.1 Örneklem ve Değişkenler

Çalışmanın verileri Avrupa Şampiyonlar Ligi 2010/2011 sezonundan başlayarak 2019/2020 sezonuna kadar grup aşamasında oynanan maçlardan oluşmaktadır. Veriler whocored.com internet sitesinden derlenmiştir. Şampiyonlar Ligi'nde grup aşamasında her sezon 8 grupta 4'er takım mücadele etmektedir. Takımlar birbirileriyle kendi sahasında ve deplasmanda olmak üzere ikişer maç yapar. Bu nedenle, her bir takım grupta 6 maç yapar ve her grupta toplam 24 takım istatistiği oluşmaktadır. Yılda 8 grup olduğu düşünüldüğünde toplamda 192 takım istatistiği vardır. Bu nedenle, 10 yıl ile birlikte verinin gözlem sayısı 1920'dir. Sadece grup maçlarının alınmasının sebebi, eleme turlarında maçların grup aşamasından daha farklı bir ortamda oynamasındandır. Çünkü eleme turları iki ayaklı maçlardan oluşur ve takımlar gruptaki maçlarda oynadıkları oyundan daha farklı bir oyun oynayabilirler.

ÇBÖ ve k-ortalamalar kümeleme analizi için R programındaki 'Stats' ve 'MASS' kitaplıkları kullanılmıştır [67]. Radar grafiği, 'ggradar' kütüphanesi ile R programında uygulandı. Karar ağacı için 'Rpart' ve 'rpart.plot' fonksiyonları kullanılmıştır [68, 69]. Yapay sinir ağları ve destek vektör makineleri için Matlab programının 2019b sürümü kullanılmıştır.

Çalışmada kullanılan değişkenler ve açıklamaları Tablo 4.1'de verilmiştir. Değişkenlerin yapısına göre ikiye ayrıldığı söylenebilir. Bu değişkenler, performans göstergeleri ve durumsal değişkenler olarak ifade edilebilir. Performans göstergesi olarak alınan değişkenler sırasıyla adam geçme denemesi, başarılı adam geçme, başarılı pas, başarılı top çalma, çalım yeme, defansif hava topu, faul, kafa topu kazanma, isabetli şut, kilit pas, korner, ofansif hava topu, ofsayt, top çalma denemesi, top kapma, top kaybı, topla buluşma, toplam pas,

toplam şut ve uzaklaştırma değişkenleridir. Durumsal değişkenler ise çalışmanın bağımlı değişkeni maç sonucu (galibiyet, mağlubiyet ve beraberlik), maç yeri (ev sahibi ve deplasman), ilk golü atma (evet ve hayır), k- ortalama kümeleme analizi sunucunda oluşturulan rakibin kalitesi (güçlü, zayıf ve dengeli) değişkenlerinden oluşmaktadır.

Tablo 4.1 Çalışmada kullanılan değişkenler ve açıklamaları

Değişkenler	Açıklamalar
Adam geçme denemesi	Takımdaki oyuncuların toplam adam geçme denemesi sayısı
Başarılı adam geçme	Takımdaki oyuncuların toplam başarılı adam geçme sayısı
Başarılı pas	Takım tarafından yapılan başarılı pas sayısı
Başarılı top çalma	Takım tarafından yapılan başarılı top çalma sayısı
Çalım yeme	Takımdaki oyuncuların toplam yedikleri çalım sayısı
Defansif hava topu	Takım tarafından kazanılan defansif hava topu sayısı
Faul	Rakip takım tarafından yapılan faul sayısı
Hava topu kazanma	Takım tarafından kazanılan hava topu sayısı
İsabetli şut	Takım tarafından çekilen isabetli şut sayısı
Kilit pas	Takımdaki oyuncuların yaptığı toplam kilit pas sayısı
Korner	Takım tarafından kazanılan korner sayısı
Maç sonucu	Takıma göre maçın sonucu : galibiyet, beraberlik, ve mağlubiyet
Maç yeri	Takımın maçı oynadığı yer: kendi sahası, deplasman
İlk golü atma	Takımın ilk golü atıp atmadığı: evet, hayır
Ofansif hava topu	Takım tarafından kazanılan toplam ofasif hava topu sayısı
Ofsayt	Takımın düştüğü ofsayt sayısı
Rakibin kalitesi	Rakip takımın kalitesi: güçlü, zayıf, ve dengeli
Top çalma denemesi	Takım tarafından gerçekleştirilen toplam top çalma denemesi
Top kapma	Takımdaki oyuncular tarafından toplam kapılan top sayısı
Top kaybı	Takımın yaptığı toplam top kaybı sayısı
Topla buluşma	Takımdaki oyuncuların maç içerisindeki toplam topla buluşma sayısı
Toplam pas	Takım tarafından yapılan toplam pas sayısı
Toplam şut	Takım tarafından çekilen toplam şut sayısı
Uzaklaştırma	Takım tarafından uzaklaştırılan top sayısı

4.2 İstatistiksel Analizler

İlk olarak, bağımlı değişken maç sonucu 3 kategoriye sahip bir değişken olmasından ötürü, performans göstergelerinin maç sonucuna göre farklılık gösterip göstermediğini tespit etmek için tek yönlü varyans analizi (ANOVA) testi uygulanmıştır. Bu test aynı zamanda performans göstergesinin, maç sonucunu etkileyip etkilemediğini de gösterecektir. Tek yönlü ANOVA sonuçları Tablo 4.2'de verilmiştir. Galibiyet, beraberlik ve mağlubiyet sonuçları için performans

göstergelerinin ortalamaları ve standart sapmaları gösterilmiştir (Tablo 4.2). Aynı zamanda, ANOVA testi için p olasılık değerleri de verilmiştir. Adam geçme denemesi, başarılı adam geçme, başarılı pas, çalım yeme, defansif hava topu, faul, hava topu kazanma, isabetli şut, kilit pas, korner, ofansif hava topu, ofsayt, top çalma denemesi, topla buluşma, toplam pas, toplam şut ve uzaklaştırma değişkenlerinin galibiyet, beraberlik ve mağlubiyet olan maç sonuçları bakımından istatistiksel olarak anlamlı ($p<0,05$) olduğu belirlenmiştir. Ayrıca, başarılı top çalma, top kapma ve top kaybı değişkenlerinin maç sonucu bakımından istatistiksel olarak anlamsız olduğu saptanmıştır. Yani, bir bakıma bu değişkenlerin maç sonucu üzerinde bir etkisi olmadığı söylenebilir. Bu sebeple, ANOVA testi sonucunda istatistiksel olarak anlamlı olmayan bu değişkenler sonraki analizlerde kullanılmayacaktır.

Tek yönlü ANOVA testi, farkın hangi grup ya da gruplardan kaynaklandığını göstermediği için bir çoklu karşılaştırma testi olan Tukey HSD testinden faydalanılmıştır. Ortalama değerlerinin (Tablo 4.2) yanında yer alan işaretler (‡,†,●) %95 güven seviyesinde Tukey HSD testinin sonuçlarını göstermektedir. Herhangi bir kategorinin altında bulunan işaret, diğer kategorinin altında da bulunuyorsa bu iki kategori arasında istatistiksel olarak anlamlı bir fark olduğunu gösterir. Ayrıca, iki kategori arasında fark varsa, ortalaması yüksek olan değişkenin diğerine göre pozitif olarak farklılık yarattığı söylenebilir.

Adam geçme denemesi performans göstergesi için, galibiyetin altında iki işaret bulunması (‡,†) beraberlik ve mağlubiyetten farklı olduğunu göstermektedir. Beraberlik kategorisi'nin altında tek bir işaret bulunması (‡) ve aynı işaretin galibiyette de olması beraberliğin sadece galibiyetten farklı olduğunu belirtir. Mağlubiyet kategorisinin altında da galibiyetin altında olan tek bir işaret (†) bulunmaktadır. En nihayetinde, adam geçme denemesi bakımından galip gelen takımların berabere kalan ve mağlup olan takımlardan farklı olduğu söylenebilir. Berabere kalan ve mağlup olan takımlar adam geçme denemesi bakımından birbirlerinden istatistiksel olarak farklı değillerdir.

Tablo 4.2 Maç sonucuna göre tek yönlü ANOVA ve Tukey HSD test sonuçları

Performans göstergesi	Galibiyet (ort, std. sapma)	Beraberlik (ort, std. sapma)	Mağlubiyet (ort, std. sapma)	*p*	Etki büyüklüğü
Adam geçme denemesi	18,16‡,†; 7,26	16,86‡; 6,74	16,25†; 6,58	0,00*	0,015 (küçük)
Başarılı adam geçme	10,26‡,†; 5,03	9,10‡; 4,45	8,51†; 4,51	0,00*	0,027 (küçük)
Başarılı pas	451,49‡,†; 155,10	385,23‡,•; 125,73	355,08†,•; 110,59	0,00*	0,095 (orta)
Başarılı top çalma	19,40; 8,49	19,24; 5,77	18,87; 5,76	0,33	0,001 (anlamsız)
Çalım yeme	8,51†; 4,51	9,10•; 4,45	10,28†,•; 5,05	0,00*	0,027 (küçük)
Defansif hava topu	9,77†; 8,33	10,34•; 8,58	8,71†,•; 7,56	0,00*	0,011 (küçük)
Faul	12,49†; 4,18	13,00; 4,16	13,26†; 4,38	0,00*	0,013 (küçük)
Hava topu kazanma	12,62†; 5,74	13,24•; 6,12	11,12†,•; 5,94	0,00*	0,021 (küçük)
İsabetli şut	6,48‡,†; 2,93	4,56‡,•; 2,52	3,31†,•; 2,12	0,00*	0,232 (büyük)
Kilit pas	11,77‡,†; 4,88	10,02‡,•; 4,89	8,12†,•; 4,12	0,00*	0,108 (orta)
Korner	5,49†; 3,04	5,20•; 3,11	4,46†,•; 2,95	0,00*	0,023 (küçük)
Ofansif hava topu	8,75‡,†; 7,87	10,33‡; 8,56	9,77†; 8,35	0,00*	0,025 (küçük)
Ofsayt	2,72‡,†; 2,07	2,42‡; 2,17	2,27†; 1,88	0,00*	0,010 (küçük)
Top çalma denemesi	27,70†; 7,95	28,34; 8,11	29,16†; 8,68	0,00*	0,025 (küçük)
Top kapma	15,82; 6,29	16,61; 6,57	15,99; 6,03	0,10	0,002 (anlamsız)
Top kaybı	11,01; 4,39	11,51; 4,45	11,44; 4,39	0,09	0,003 (anlamsız)
Topla buluşma	714,83‡,†; 148,61	658,77‡,•; 124,00	614,62†,•; 114,91	0,00*	0,102 (orta)
Toplam pas	535,35‡,†; 150,45	472,71‡,•; 122,22	439,54†,•; 110,30	0,00*	0,097 (orta)
Toplam şut	15,57‡,†; 5,93	13,50‡,•; 6,12	11,04†,•; 5,15	0,00*	0,109 (orta)
Uzaklaştırma	21,06†; 11,01	23,60‡,•; 11,58	21,66•; 9,37	0,00*	0,010 (küçük)

Başarılı pas değişkeni için; galibiyet; beraberlik ve mağlubiyet ortalamaları arasında %5 anlamlılık seviyesinde istatistiksel olarak anlamlı bir fark bulunmaktadır ($p<0,05$). Galibiyet (‡;†); beraberlik (‡;●) ve mağlubiyet (†;●) sonuçlarının tümünün yanında iki işaret yer almaktadır (Tablo 4.2). Burada; ‡ işareti galibiyet ve beraberlik arasındaki farkı; † işareti galibiyet ve mağlubiyet arasındaki farkı ve son olarak; ● işareti beraberlik ve mağlubiyet arasındaki

anlamlı farkı göstermektedir. Ortalamalar incelendiğinde; galip gelen takımların (451,49) beraber kalan (385,23) ve mağlup olan takımlardan (355,08) ortalama olarak daha fazla başarılı pas yaptığı söylenebilir. Ayrıca; berabere kalan takımlar (385,23) da mağlup olan takımlardan (355,08) daha fazla başarılı pas yapmaktadır. Diğer tüm performans göstergeleri de benzer şekilde yorumlanmalıdır.

Başarılı pas; isabetli şut; kilit pas; topla buluşma; toplam pas ve toplam şut değişkenleri galibiyet; beraberlik ve mağlubiyet için ortalamaları bakımından %95 güven düzeyinde ortalama olarak birbirlerinden farklı oldukları söylenebilir. Ayrıca; galip gelen takımların berabere kalan ve mağlup olan takımlardan; berabere kalan takımların da mağlup olan takımlardan ortalama olarak ve %5 anlamlılık seviyesinde daha fazla başarılı pas yaptığı; daha fazla şut çektiği; çektiği şutları da fazla isabet ettirdiği; daha fazla kilit pas verdiği; daha fazla topla buluştuğu ve daha fazla pas çevirdiği söylenebilir.

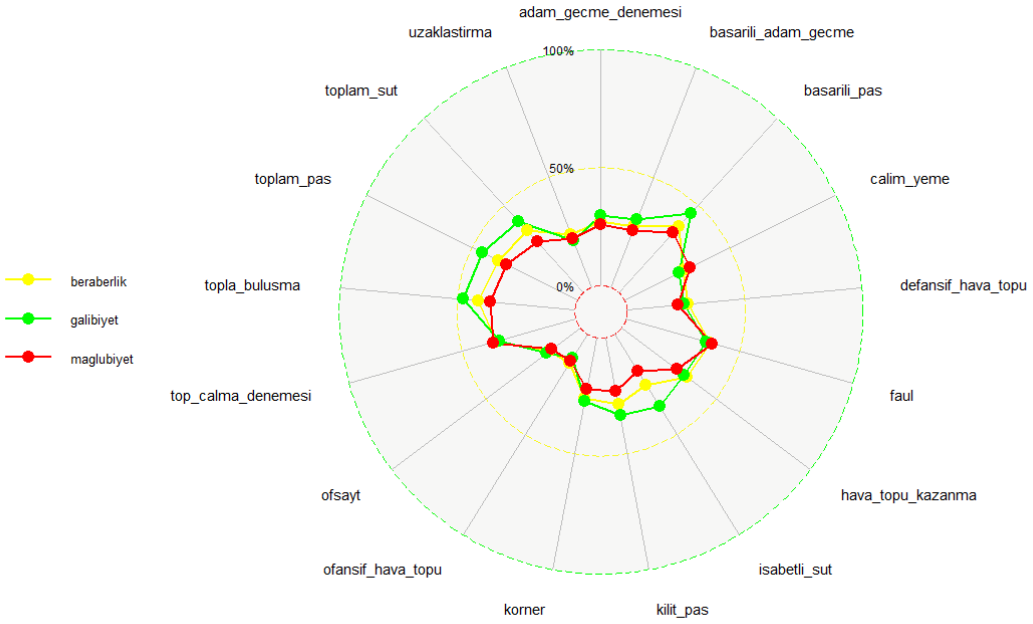
Berabere biten maçların; adam geçme denemesi bakımından mağlup olunan maçlardan; başarılı adam geçme bakımından mağlup olunan maçlardan; çalım yeme bakımından galip gelinen maçlardan; defansif hava topu kazanma bakımından galip gelinen maçlardan; yapılan faul bakımından hem galip hem de mağlup olunan maçlardan; hava topu kazanılması bakımından kazanılan maçlardan; atılan korner bakımından galip gelinen maçlardan; ofansif hava topu kazanma bakımından mağlubiyetle sonuçlanan maçlardan; ofsayt bakımından kaybedilen maçlardan ve top çalma denemesi bakımından hem kazanılan hem de kaybedilen maçlardan ortalama olarak herhangi bir istatistiksel farka sahip olmadığı belirlenmiştir. Bu da berabere kalınan maçların; adam geçme denemesi; başarılı adam geçme; faul; ofansif hava topu; ofsayt ve top çalma denemesi göstergeleri yönünden mağlup olunan maçlar ile; çalım yeme; faul; hava topu kazanma; korner; top çalma denemesi bakımından da galip gelinen maçlar ile benzer özellikler gösterdiği söylenebilir.

Ayrıca; performans göstergeleri için Tablo 4.2'de etki büyüklüğü değerleri de verilmiştir. ANOVA testi sonucunda anlamsız olan başarılı top çalma; top kapma

ve top kaybı değişkenleri etki büyüklüğü bakımından da anlamsızdır. Adam geçme denemesi; başarılı adam geçme; çalım yeme; defansif hava topu; faul; hava topu kazanma; korner; ofansif hava topu; ofsayt; top çalma denemesi ve uzaklaştırma performans göstergelerinin küçük etki büyüklüğüne sahip olduğu söylenebilir. Büyük örneklemler için küçük etki büyüklükleri kabul edilebilirdir. Bu nedenle bu değişkenler de çalışmada kullanılmıştır. Başarılı pas; kilit pas; topla buluşma; toplam pas ve toplam şut değişkenlerinin de orta etki büyüklüğüne sahip olduğu söylenebilir. Büyük etki büyüklüğüne sahip olan tek değişken isabetli şut değişkenidir.

Ayrıca; performans göstergelerini daha iyi anlatabilmek ve aradaki farkı net bir şekilde ortaya koymak için radar grafiğinden yararlanılmıştır. Şekil 4.1'de maç sonucuna göre performans göstergelerinin radar grafiği yer almaktadır. Bu grafikten de görülebileceği üzere çalım yeme; faul ve top çalma denemesi mağlup olan takımlar için daha fazladır. Diğer tüm değişkenler için (performans göstergeleri) galip gelen takımların daha üstün olduğu söylenebilir. Galip gelen takımların daha fazla adam geçmesi; daha fazla başarılı pas yapması; daha çok hava topu kazanması; daha fazla topla buluşması ve daha fazla pas yapması sonuçların mantıklı olduğunu da göstermektedir. Ayrıca; mağlup olan takımların daha fazla çalım yemesi; daha fazla faul yapması ve daha fazla top çalma denemesinde bulunması mantıksal olarak da beklenen bir durumdur.

ANOVA testi sonucunda istatistiksel olarak anlamlı olan performans göstergelerinin maç sonucuna etki ettiklerini söylemek yanlış olmayacaktır (Tablo 4.2). Galip gelen takım; berabere kalan takımdan ve mağlup olan takımdan daha fazla pas yapıyorsa; berabere kalan takım da mağlup olan takımdan daha fazla nasıl yapıyorsa toplam pas değişkeninin maç sonucuna etkisi olduğu yadsınamaz. Aynı yorum anlamlı olan diğer tüm değişkenler için yapılabilir.

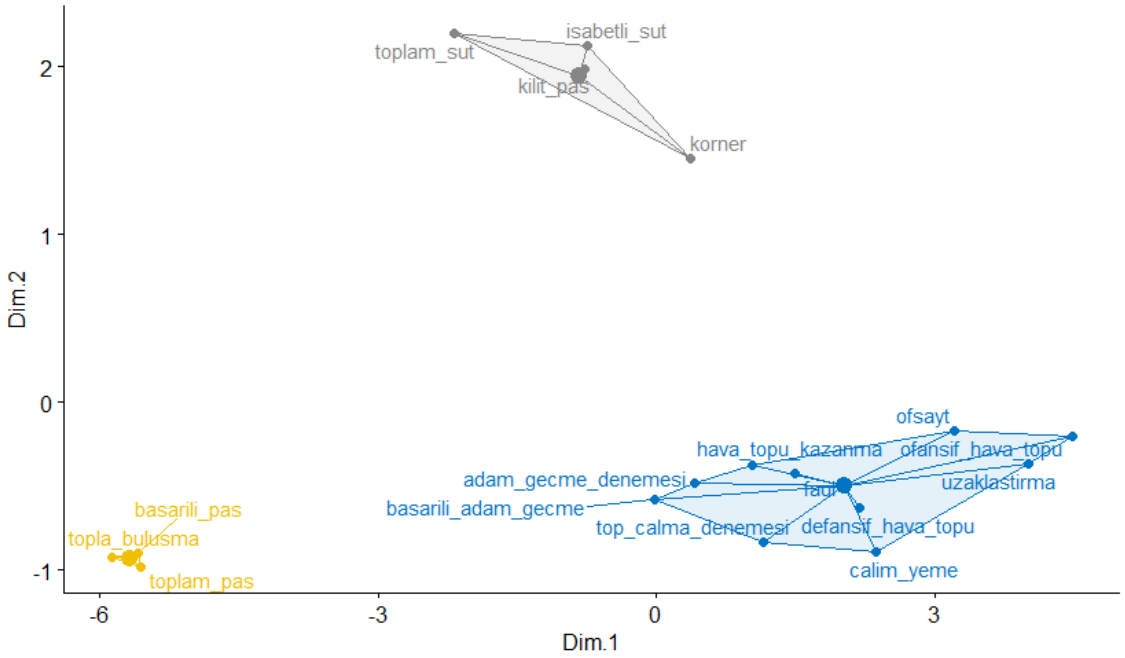


Şekil 4.1 Maç sonucuna göre performans göstergelerinin radar grafiği

Futbol maçları karşılıklı iki takım arasında oynandığı için her iki takımın performansı da maç sonucunun belirlenmesi açısından önemlidir. Bu yüzden; takımlar kendi performanslarının yanında rakip takımların performanslarına göre de değerlendirilmelidir. Yani; güçsüz bir rakiple oynayan bir takımın performansı ile güçlü bir rakiple oynayan takımın performansının karşılaştırılması çok da sağlıklı olmayacaktır. Bu nedenle de; performans göstergeleri kullanılarak k-ortalama kümeleme analizi (k=3) yardımıyla veri seti üç gruba ayrılmıştır. Analiz sonucunda; küme merkezine olan uzaklığı 1,224'ten fazla olan 338 gözlemin zayıf rakiplere karşı; küme merkezine olan uzaklığı 1,8 ile 1,223 arasında olan 787 gözlemin dengeli rakiplere karşı ve son olarak; küme merkezine olan uzaklığı 0,179'dan az olan 795 gözlemin güçlü rakiplere karşı oynadığı belirlenmiştir.

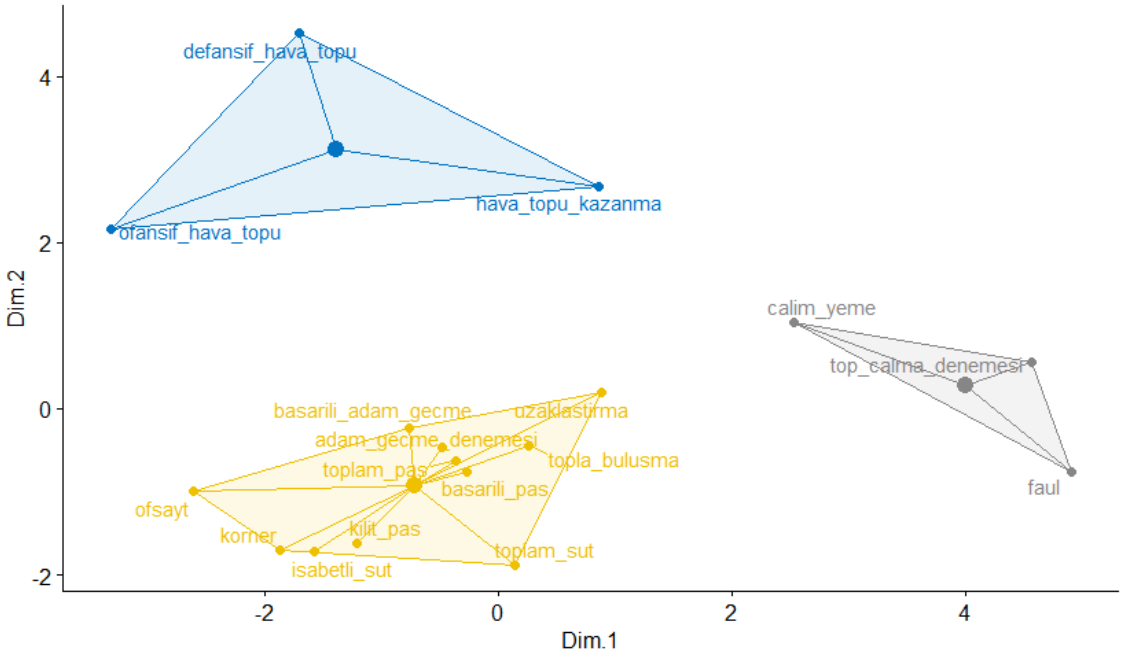
Rakibin kalitesine göre performans göstergelerinin nasıl bir yapıya sahip olduğunu anlamak için çok boyutlu ölçekleme analizi kullanılmıştır. Bu nedenle zayıf rakiplere karşı oynayan takımlar; dengeli rakiplere karşı oynayan takımlar ve güçlü rakiplere karşı oynayan takımlar olmak üzere 3 farklı çok boyutlu ölçekleme haritası görselleştirilmiştir. İki boyutlu görselleştirmede çok boyutlu ölçekleme sonuçlarının daha anlaşılır olması için benzerlikleri bakımından üç gruba

toplanmıştır. Haritalardaki renkler benzerlikleri fazla olan grupları göstermektedir.



Şekil 4.2 Zayıf rakiplere karşı oynayan takımlar için ÇBÖ analizi sonuç grafiği

Şekil 4.2 zayıf rakiplere karşı oynayan takımlar için çok boyutlu ölçekleme analizi sonuçlarını göstermektedir. Topla buluşma; başarılı pas ve toplam pas göstergeleri zayıf rakiplere karşı oynayan takımlarda benzer özellikler taşımaktadır; yani zayıf olan bir takım daha fazla topla buluşup; daha fazla pas yapıp; yaptığı paslardan daha fazla başarı elde etmektedir. Farklı bir bakış açısı ile bakıldığında; topa sahip olmayı işaret eden bu üç değişkenin benzer özellikler taşıması sonuçların tutarlı olduğunu da göstermektedir. Burada diğer bir grup; kilit pas; korner; toplam şut ve isabetli şut değişkenlerinin oluşturduğu gruptur. Bu değişkenler de rakip zayıf olduğunda benzer özelliklere sahiptir. Başka bir deyişle; bu değişkenler takımın hücum özelliklerini göstermektedir. Rakibi zayıf olan takımlar hücum özellikleri bakımından da birbirlerine benzerdir. Geriye kalan tüm değişkenler tek bir grupta toplanmıştır.

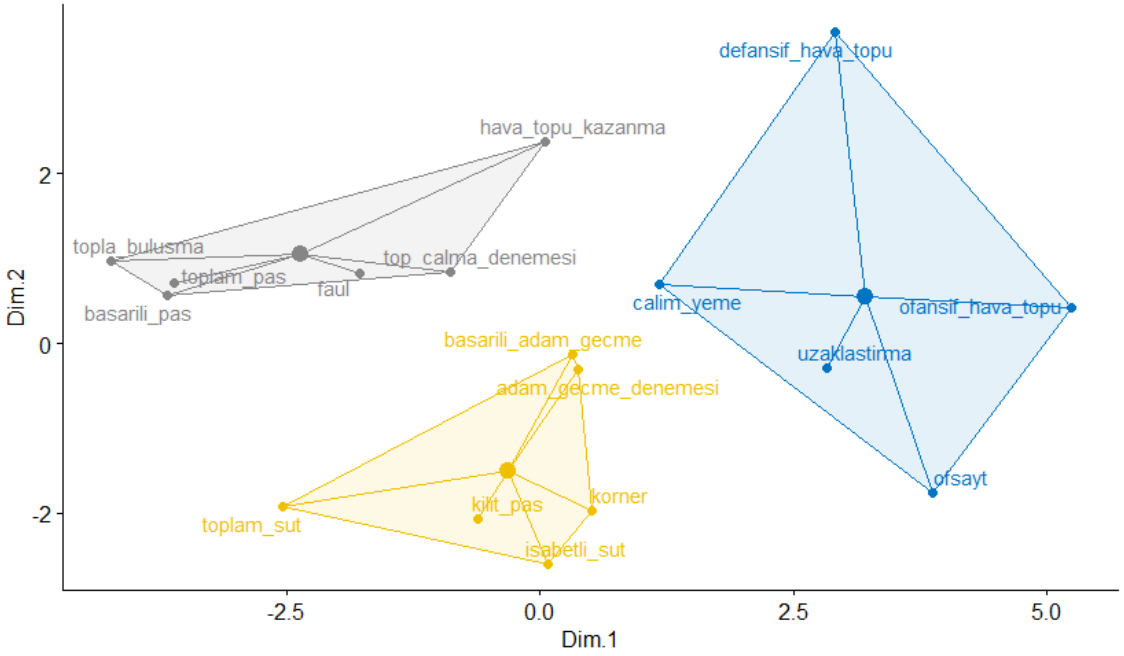


Şekil 4.3 Daha güçlü rakiplere karşı oynayan takımlar için ÇBÖ analizi sonuç grafiği

Güçlü rakiplere karşı oynayan takımlar için çok boyutlu ölçekleme analiz sonuç grafiği Şekil 4.3'te verilmiştir. Burada defansif hava topu; ofansif hava topu ve hava topu kazanma değişkenleri aynı grupta yer alır ve bu değişkenlerin aynı zamanda hava hakimiyeti ile ilgili olduğu söylenebilir. Yani rakip güçlü olduğunda takımların hava hakimiyetleri birbirlerine benzer özellikler taşımaktadır. Benzer şekildeki diğer bir grup ise çalım yeme; top çalma denemesi ve faul'dür. Radar grafiğinde (Şekil 4.1); kaybeden takımların daha fazla faul yaptığı; daha fazla çalım yediği ve daha fazla top çalma denemesinde bulunduğu belirtilmiştir. Bu üç performans göstergesinin güçlü rakiplere karşı oynayan takımlar için olan çok boyutlu ölçekleme haritasında çıkması; kaybeden takımların güçlü rakiplere karşı oynayan takımlarla ortak özellikler taşıdığının da bir göstergesidir. Zaten; genellikle rakibi güçlü olan takımların büyük oranda kaybetmesi beklenir. Geriye kalan tüm performans göstergeleri aynı grupta yer almaktadır.

Dengeli rakiplere karşı oynayan takımlar için olan çok boyutlu ölçekleme analizi Şekil 4.4'te verilmiştir. Rakibi zayıf (Şekil 4.2) ve rakibi güçlü (Şekil 4.3) olan takımları gösteren şekillere nazaran daha dengeli bir yapıya sahiptir. Topla

buluşma; toplam pas ve başarılı pas göstergelerinin yanında faul; hava topu kazanma ve top çalma denemesi göstergeleri de eklenmiştir; yani tüm bu değişkenler tek bir gruptadır. Topa sahip olmayı ve defansif özellikleri gösteren performans göstergelerinin birlikte yer alması takımların daha dengeli bir yapıda oynadığını gösterir. Şekildeki diğer bir grup; toplam şut; başarılı adam geçme; adam geçme denemesi; kilit pas; korner; isabetli şut değişkenlerinden oluşur. Geriye kalan tüm performans göstergeleri de bir grupta yer alır. Aslında; bu şekil göstermektedir ki dengeli takımlar mücadele ettiğinde performans göstergeleri de daha dengeli bir yapıya sahip olmaktadır.

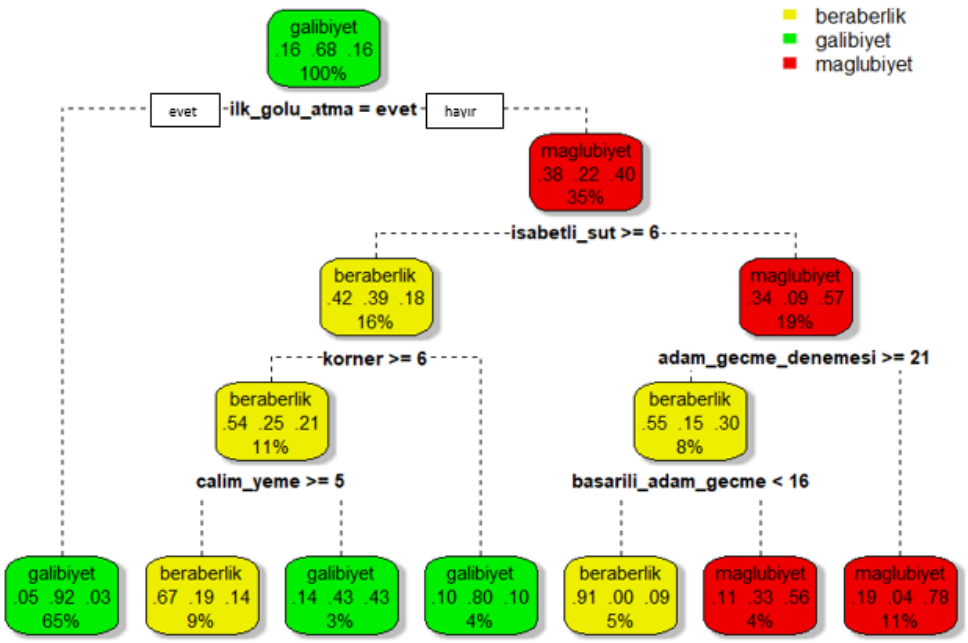


Şekil 4.4 Dengeli rakiplere karşı oynayan takımlar için ÇBÖ analizi sonuç grafiği

Performans göstergelerinin yapıları incelendikten sonra maç sonucu üzerinde etkilerini incelemek için yine rakibin kalitesine göre üç farklı karar ağacı modeli oluşturulmuştur. Burada performans göstergelerine ek olarak durumsal değişkenler de modelde yer almaktadır. Bağımlı değişken galibiyet; mağlubiyet ve beraberlik kategorilerinden oluşan maç sonucudur. Yorumlamaya geçmeden önce bilinmesi gereken birkaç durum mevcuttur. Her bir düğümün içerisinde 3 farklı olasılık mevcuttur. Bunlar sırasıyla beraberlik; galibiyet ve mağlubiyet olasılıklarıdır. Bu olasılıkların altında yer alan yüzde değeri düğümde

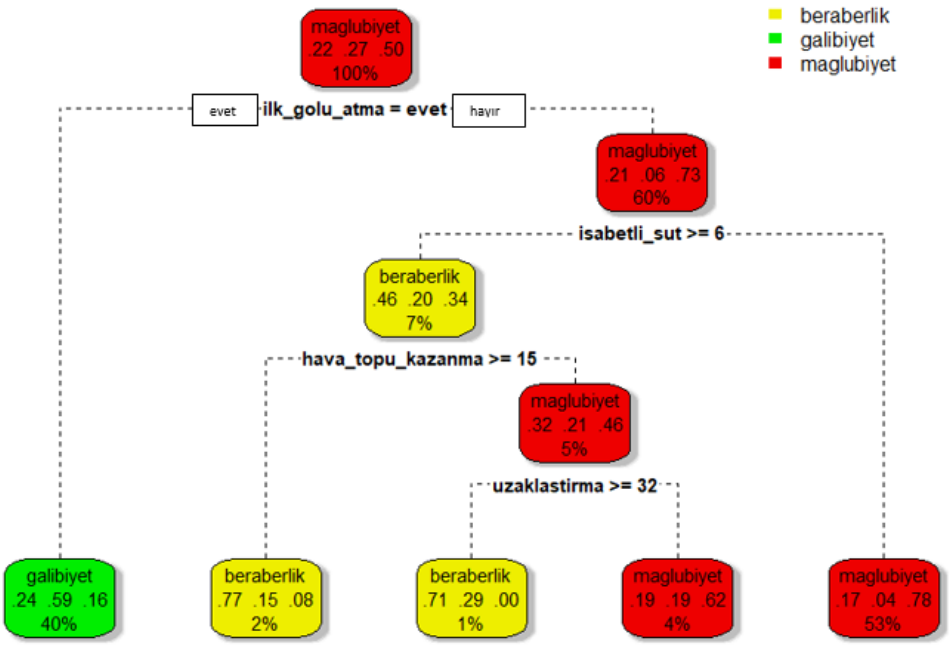
yer alan koşulun örneklemde yüzde ne kadar gözüktüğünü gösterir. Düğüm rengi kategorilerden hangisinin olasılığı daha yüksekse o renktir. Düğümlerin altındaki bölünmelerde sol taraf; düğümdeki koşulun sağlandığı anlamına gelen eveti gösterir. Sağ taraf ise koşulun sağlanmadığı anlamını taşır ya da daha iyi bir açıklama ile koşulun dışındaki olasılığı göstermektedir.

Zayıf rakiplere karşı oynayan takımlar için oluşturulan karar ağacını Şekil 4.5'te verilmiştir. Kök düğüm incelendiğinde; beraberlik şansı %16; galibiyet şansı %68 ve mağlup olma olasılığı yüzde %16'dır. Başka bir deyişle; rakip zayıf olduğunda kazanma şansı %68'dir. İlk gol atıldığında; takımın maçı kazanma şansı artarak 0,92 olmaktadır. Ayrıca; maçı kaybetme olasılığı da azalarak 0,03'e düşmektedir. Takım ilk golü atamadığında; galibiyet şansı sadece 0,22'dir. Buna ek olarak; 6 ve 6'dan daha fazla isabetli şut çekildiğinde; galibiyet şansı %39'a çıkartılabilir. Rakibi zayıf olan takımlar ilk golü atmadığında 6 ve 6'dan fazla isabetli şut çekip; 6'dan daha az korner kullanarak galibiyet şansını %80'e çıkartabilir. Benzer şekilde; takımın ilk golü atamadığı durumda; 6'dan daha az isabetli şut çekilirse; mağlup olma olasılığı %57'ye çıkar; bunun koşulun yanında adam geçme denemesi 21 ve altında olursa; mağlup olma olasılığı 0,78'e yükselir.



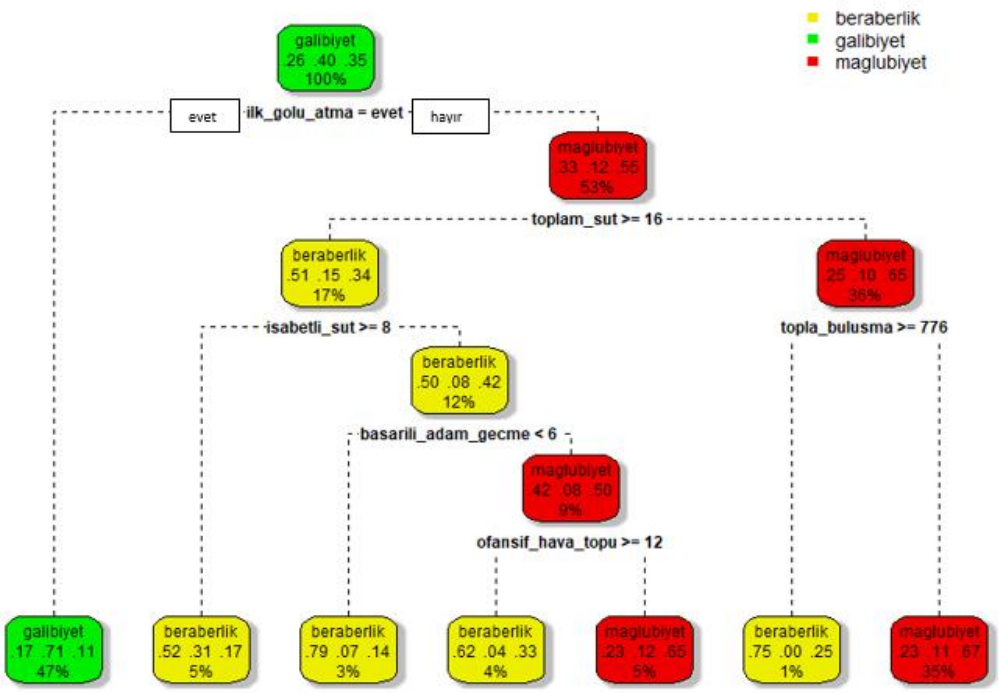
Şekil 4.5 Daha zayıf rakiplere karşı oynayan takımlar için karar ağacı

Daha güçlü rakiplere karşı oynayan takımlar için oluşturulan karar ağacı Şekil 4.6'da verilmiştir. Kök düğüm incelendiğinde; rakibi güçlü olan takımın maçı kaybetme olasılığı 0,50'dir. Ayrıca; maçı kazanma şansı 0,27 ve berabere kalma şansı da 0,22'dir. Takım ilk golü atarsa; takımın kazanma olasılığı 0,59'a çıkar; kaybetme olasılığı da 0,16'ya geriler. Takımın ilk golü atamadığı durumda; kaybetme olasılığı %73'e olur. Bu durumda; maçı kazanma olasılığı sadece %6'dır. Takım ilk golü atamaz ve 6'dan daha az isabetli şut çekerse; %78 olasılıkla maçı kaybedecektir. İlk gol rakip takım tarafından atılırsa; 6 ve 6'dan fazla isabetli şut çekerse; beraberlik şansı %46'ya; buna ek olarak 15 ve daha fazla hava topu kazanırsa; beraberlik şansı %77'ye yükselir. Benzer şekilde; ilk gol rakip tarafından atılırsa; 6 ve 6'dan fazla şut çekip; 15'ten daha az hava topu kazanılırsa mağlubiyet şansı 0,46; buna ek olarak takım 32 ya da daha fazla topu uzaklaştırmayı başarabilirse mağlup olma olasılığı neredeyse sıfıra düşer. Bu durum; beraberlik şansını da yüzde %71'e yükseltmektedir. Topu uzaklaştırmada 32 keze ulaşamama durumunda ise %62 olasılıkla takım yenilecektir.



Şekil 4.6 Daha güçlü rakiplere karşı oynayan takımlar için karar ağacı

Dengeli rakiplere karşı oynayan takımlar için oluşturulan karar ağacı şekil 4.7'den incelenebilir. Kök düğüme bakıldığında; dengeli rakiplerin oynadığı maçlarda beraberlik şansı 0,26; galibiyet şansı 0,40 ve mağlubiyet olasılığı 0,35'tir. Eğer bir takım ilk golü atarsa; galip gelme olasılığı yüzde 71'e yükselecektir. İlk golün atılamaması durumunda; galibiyet olasılığı sadece %12'dir. Bu durumda mağlup olma olasılığı %55'dir. Rakibi dengeli olan bir takım gol atamaz; 16 ve daha fazla şut çekerse; galip gelme olasılığı 0,15; çektiği bu şutlardan 8 ve daha fazlasını isabet ettirebilirse galip gelme olasılığı 0,30'dur. Bu durumda berabere kalma olasılığı 0,52'dir. Bir takım ilk golü atamaz; 16 ve daha fazla şut çeker; 8'den az isabet sağlar ve 6'dan daha az başarılı adam geçme ile oynarsa beraberlik şansı %79 olur. Benzer şekilde ilk golün rakip tarafından atılması durumda; 16'dan daha az şut çekilir; 776'dan daha az topla buluşulursa; takım %67 oranla mağlup olacaktır. Bu durumda; 776 ve daha fazla topla buluşma gerçekleşirse takımın galip gelme olasılığı neredeyse sıfırdır; fakat beraberlik şansı %75'e çıkar.



Şekil 4.7 Dengeli rakiplere karşı oynayan takımlar için karar ağacı

Bu aşamaya kadar; maç sonucunun etkileyen değişkenler belirlenmiş; bu değişkenlerin rakibin kalitesine göre nasıl değişkenlik gösterdikleri

görselleştirilmiş ve son olarak bu performans göstergelerinin ve durumsal değişkenlerin maç sonucunu ne derecede etkiledikleri araştırılmıştır. Bu aşamadan sonra; performans göstergeleri ve durumsal değişkenler kullanılarak galibiyet mağlubiyet ve beraberlik durumundan oluşan maç sonucu (bağımlı değişken) tahmin edilmeye çalışılacaktır. Bağımlı değişken kategorilerden oluştuğu için (etiket) problemin bir sınıflandırma problemi olduğu söylenebilir. Bu nedenle bunun çözümü için makine öğrenmesi ve yapay sinir ağları yöntemleri kullanılacaktır. Veri seti %70 eğitim ve %30 test olarak ayrılmıştır. Yorumlara geçmeden önce tablolarda bulunan performans ölçüm kriterlerinden bahsedilmesi gerekir. Burada 3 adet bilgi kriteri yer almaktadır. Bunlar; Akaike (AIC); Düzeltilmiş Akaike (AICc) ve Bayesyen (BIC) bilgi kriterleridir. Modeller karşılaştırılırken bilgi kriter değeri sayısal olarak küçük olan bilgi kriteri daha iyi bir modeli gösterir. Her ne kadar ortalama hata kare değeri sürekli olan değişkenler için daha uygun olsa da; kontrol sağlama amacı ile tablolarda ortalama hata kare değerine de yer verilmiştir. Ortalama hata kare değeri sıfıra ne kadar yakınsa; modelin o kadar iyi olduğu söylenebilir. Son olarak doğruluk oranları da tablolarda verilmiştir. Doğruluk oranı modelin yüzde ne kadar doğru sınıflandırıldığını göstermektedir. Bir diğer önemli nokta; eğitim setinde modelin ne kadar iyi eğitildiği ya da kestirildiğidir. Yani veri setinin %70'i kullanılarak model eğitilir (kestirilir). Burada daha önemli olan durum; test için ayrılan kısımdır. Eğitim veri setinde öğrenilen model daha önce hiç karşılaşmadığı test verisinde sınanır. Bu nedenle test veri setinin başarısı; aslında modelin gerçek başarısını göstermektedir.

Doğrusal; karesel; kübik; ince normal; orta normal ve kalın normal olmak üzere altı farklı destek vektör makineleri yönteminin sonuçları Tablo 4.3'te verilmiştir. Doğrusal destek vektör makineleri incelendiğinde; modelin eğitim başarısı 0,68'dir. Bu modeldeki test verisi için doğruluk oranı da yine 0,68'dir. Yani oluşturulan model hiç görmediği bir veri seti karşısına geldiğinde yine benzer bir doğruluk oranına sahiptir. Veri seti ile ilişkilendirilecek olursa kullanılan performans göstergeleri ve durumsal değişkenler yardımıyla maç sonucunu %68

oranında doğru tahmin edilebilmektedir. Bunun aslında çok da yüksek olmadığı söylenebilir.

Karesel destek vektör makineleri incelendiğinde (Tablo 4.3); bu kez eğitim başarısı 0,75'e çıkmıştır. Fakat test başarısı yine 0,68'dir. Model veriyi daha iyi öğrenmiştir ancak yeni veri ile karşılaştığında doğruluk oranı 0,75'e ulaşamamıştır. Kübik destek vektör makineleri eğitim doğruluk oranı %92'dir. Model eğitim verisini daha iyi öğrenmiştir ancak test doğruluk oranı yüzde 0,65'e düşmüştür. İnce normal destek vektör makineleri eğitim veri seti için doğruluk oranı %100'dür. Bu modelin eğitim verisetini çok iyi öğrendiği anlamını taşır. Fakat test için doğruluk oranı 0,58'dir. Farklı modellerde eğitim veri seti için oranlar yükselmektedir; fakat test veriseti için aynı durum söz konusu değildir. Bu durum makine öğrenmesi yöntemlerindeki aşırı uyum durumuna tipik bir örnektir. Model eğitim veri setini o kadar iyi öğrenir ki; yeni bir veri seti ile karşılaştığında başarısı düşer. Bilgi kriterleri bakımından; İnce normal destek vektörleri modeli için eğitim veri seti %100 sınıflandırma yapması nedeniyle bilgi kriterleri hesaplanamamıştır. Çünkü modelin herhangi bir hatası yoktur aynı şekilde ortalama hata kare değeri eğitim veri seti için burada sıfırdır. Bu modeler arasından herhangi bir model seçilmesi gerekirse; bu doğrusal destek vektör makineleri olmalıdır. Çünkü eğitilen modelin başarısı test setinde de kanıtlanmıştır. Burada verilen doluluk oranları bağımlı değişkenin kategorilerini yansıtmamaktadır ve genel doğruluk oranlarıdır. Her bir kategorinin sınıflandırılma başarısı da modeli daha iyi anlamak açısından önemlidir.

Tablo 4.3 Üç ihtimalli maç sonucu için destek vektör makineleri sonuçları

	Bilgi Kriterleri			MSE		Doğruluk Oranı		
Model	AIC	AICc	BIC	Eğitim	Test	Eğitim	Test	Tüm
Doğrusal	-614,27	732,43	-510,21	0,61	0,55	0,68	0,68	0,68
Karesel	-984,74	361,95	-880,68	0,47	0,55	0,75	0,68	0,73
Kübik	-2610,77	-1264,07	-2506,70	0,14	0,58	0,92	0,65	0,84
Fine Gaussian	-	-	-	0,00	1,01	1,00	0,58	0,87
Medium Gaussian	-903,64	443,06	-799,58	0,56	0,74	0,74	0,68	0,72
Coarse Gaussian	-657,26	689,44	-553,19	0,60	0,68	0,68	0,66	0,67

Destek vektör makinelerinin kategori bazında doğruluk oranları Tablo 4.4'te verilmiştir. Buradan da görülebileceği gibi mağlubiyet ve galibiyet sınıflandırılma doğruluk oranları yüksek olmasına rağmen beraberlik için bu oranlar çok düşüktür. Bu durum; modelin berabere kalınan maçları yanlış sınıflandırıldığı anlamına gelmektedir. Örneğin; doğrusal modelin genel doğruluk oranı 0.68 olarak verilmişti (Tablo 4.3). Fakat Tablo 4.4 göstermektedir ki aslında model mağlubiyeti 0,81; galibiyeti 0,88 oranında doğru sınıflandırabilmektedir; bunun yanında berabere biten maçların yalnızca 0,11 olarak doğru sınıflandırabilmektedir. İnce normal destek vektör makinelerine bakıldığında; beraberlik sonucunun doğruluk oranı sıfırdır. Yani model; maç sonucu berabere olan hiçbir maçı doğru olarak tahmin edememiştir. Diğer modeler için de durum farklı değildir. Diğer bir açıdan bakılacak olursa; modellerin berabere kalınan maçları galibiyet veya mağlubiyet olarak tahmin ettiği söylenebilir. Aslında; performans göstergeleri ve durumsal değişkenler kullanılarak; maç sonucu tahmininde beraberlik sonucunun galibiyet ve mağlubiyet ile iyi ayrışamadığı da sonuçlardan anlaşılabilir.

Tablo 4.4 Üç ihtimalli maç sonucu için kategori bazında doğruluk oranları

Maç Sonucu	Doğruluk Oranları					
	Doğrusal	Karesel	Kübik	Fine	Medium	Coarse
Mağlubiyet	0,82	0,81	0,73	0,59	0,83	0,83
Beraberlik	0,11	0,18	0,34	0,00	0,13	0,03
Galibiyet	0,88	0,86	0,75	0,91	0,86	0,86

Beraberliğin model başarısını düşürmesi; berabere kalınan maçlar olmadan bir sınıflandırılma yapma durumunu akla getirebilir. Bu nedenle; Tablo 4.5 galibiyetin ve mağlubiyetin olduğu maçların performans göstergeleri ve durumsal değişkenler ile nasıl sınıflandırılabileceğini göstermek için tasarlanmış destek vektör makineleri sonuçlarıdır. Bu durumda; doğrusal model eğitim seti için doğruluk oranı 0,86'dır. Test verisi için bu oran 0,87'dir. Karesel model için eğitim doğruluk oranı 0,88; test doğruluk oranı 0,90'dır. Üç ihtimalli maç sonucu için oluşturulan modelden daha iyi sonuçlar elde edildiği açıktır. Burada en iyi modelin karesel destek vektör makineleri olduğu söylenebilir çünkü test verisi için

doğruluk oranı %90'dır. Bu aynı zamanda galip gelme ya da mağlup olma sonucu için %90 oranında doğru bir tahmini gösterir.

Tablo 4.5 İki ihtimalli maç sonucu için destek vektör makineleri sonuçları

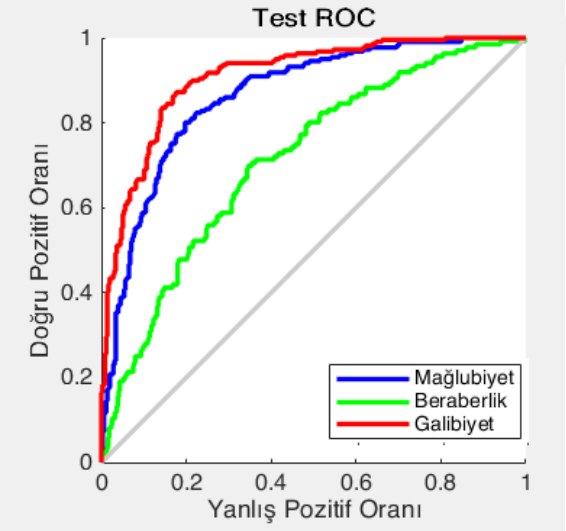
Model	Bilgi Kriterleri			MSE		Doğruluk Oranı		
	AIC	AICc	BIC	Eğitim	Test	Eğitim	Test	Tüm
Doğrusal	-572,91	469,00	-473,99	0,55	0,53	0,86	0,87	0,86
Karesel	-753,72	288,19	-654,80	0,47	0,40	0,88	0,90	0,89
Kübik	-1878,15	-836,24	-1779,23	0,16	0,56	0,96	0,86	0,93
Fine Gaussian	-	-	-	0,00	1,05	1,00	0,74	0,92
Medium Gaussian	-771,04	270,87	-672,12	0,46	0,49	0,89	0,88	0,88
Coarse Gaussian	-580,15	461,76	-481,23	0,55	0,54	0,86	0,87	0,86

Destek vektör makineleri ile karşılaştırma yapabilmek için bayesyen yapay sinir ağları da kullanılmıştır. Tablo 4.6; üç ihtimali olan maç sonucu için tek gizli katmanlı yapay sinir ağları sonuçlarını göstermektedir. Burada; 2 nörondan 10 nörona kadar dokuz farklı modelin sonuçları bulunmaktadır. Her modelin iterasyon sayıları da tabloda verilmiştir. Görülebileceği gibi nöron sayısı arttıkça eğitim verisi için doğruluk oranın da arttığı söylenebilir; yani nöron sayısı arttırıldığında modelin veriyi öğrenmesi de iyileşmektedir (Tablo 4.6). Fakat aynı durumu test verisi için söylemek mümkün değildir. Nöron sayısı arttıkça test setinin doğruluk oranı düşmektedir. Burada en iyi modelin 3 nörona sahip olan bayesyen yapay sinir ağı olduğu söylenebilir. Çünkü eğitim veri setinin doğruluk oranı 0,71 iken test setinin doğruluk oranı 0,70'tir. Destek vektör makinelerinde bulunan 3 ihtimalli maç sonucu için en iyi model; test verisi için doğruluk oranı 0,68 iken; bayesyen yapay sinir ağlarında doğru sınıflama oranı bir miktar artmıştır. Destek vektör makinelerinde olduğu gibi yapay sinir ağlarında da benzer bir durum söz konusudur. Burada da kategorilerin doğru sınıflandırma oranlarını da incelemek modeli anlamak açısından önemlidir.

Kategorilerin ya da modelin başarısını gösteren bir diğer yöntem de roc eğrisinin çizilmesidir. Şekil 4.8'de tek gizli katmanlı ve 3 nöronlu yapay sinir ağları modeli için roc eğrisi verilmiştir. Şekilde eğrilerin altında kalan alan kategorilerin doğruluk oranlarını göstermektedir. Destek vektör makinelerinde de olduğu gibi burada da beraberlik kategorisi model başarısını düşürmektedir.

Tablo 4.6 Üç ihtimalli maç sonucu için tek gizli katmanlı yapay sinir ağları sonuçları

Nöron	İter.	Bilgi Kriterleri			MSE		Doğruluk Oranı		
		AIC	AICc	BIC	Eğitim	Test	Eğitim	Test	Tüm
2	40	-1081,20	269,06	-815,83	0,138	0,141	0,69	0,69	0,69
3	192	-1049,95	305,29	-659,69	0,137	0,140	0,71	0,70	0,71
4	246	-1053,52	308,73	-538,38	0,131	0,144	0,72	0,69	0,71
5	393	-1123,83	247,60	-483,81	0,120	0,170	0,76	0,64	0,73
6	390	-1075,01	307,89	-310,11	0,120	0,162	0,76	0,67	0,73
7	471	-1180,53	216,29	-290,75	0,107	0,207	0,82	0,63	0,76
8	1000	-1398,85	14,48	-384,18	0,088	0,215	0,85	0,65	0,79
9	433	-1448,17	-15,58	-308,62	0,082	0,253	0,87	0,61	0,79
10	915	-1589,41	-134,62	-324,98	0,071	0,274	0,89	0,58	0,80



Şekil 4.8 Tek gizli katmanlı ve üç nöronlu YSA modeli için roc eğrisi (3 ihtimalli maç)

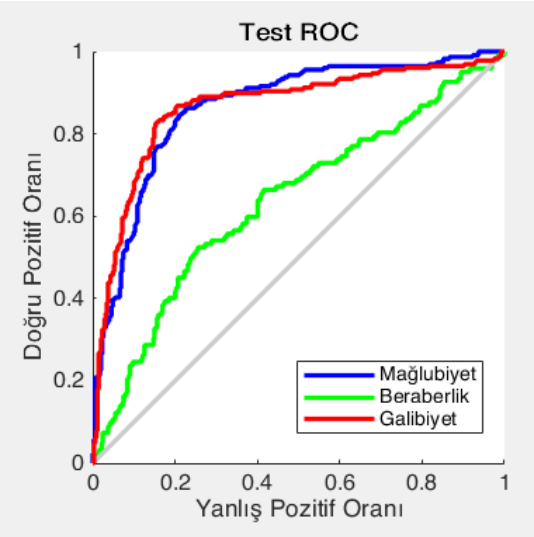
Yapay sinir ağları modelleri tek gizli katmandan oluşabileceği gibi çok gizli katmandan da oluşabilir. Gizli katman sayısının artması modeli karmaşık hale getirir ve bu durum aynı zamanda derin öğrenme yöntemlerinin de çıkış noktasını oluşturmaktadır. İki gizli katmandan oluşan bayesyen yapay sinir ağları model sonuçları Tablo 4.7'de verilmiştir. Tek katmanlı yapay sinir ağı modelinde 10

nörona kadar olan modeller karşılaştırıldığı için burada da 2 gizli katmanda toplam 10 nöron olacak şekilde modeller oluşturulmuştur ve her iki gizli katmanda da eşit sayıda nöronlar alınarak modeller kurgulanmıştır. Bu durumda da; en iyi model birinci gizli katmanda 3 ve ikinci gizli katmanda 3 nörona sahip olan modeldir. Eğitim veri setinde %71 doğruluk oranı ve test setinde de %71 doğruluk oranına sahiptir. Dikkat edilmesi gereken durum; yine önceki modellerle benzerdir. Birinci gizli katmanda 5 ve ikinci gizli katmanda da 5 nörona yer verildiğinde eğitim veri setinin doğruluk oranı %84 olmaktadır. Nöron sayısı arttıkça; yani model karmaşıklığı artınca algoritmaların veriyi çok iyi öğrendiği açıktır. Fakat bunun aksine modelin daha önce hiç karşılaşmadığı bir veriyi tahmin etmesi çok da başarılı değildir. Çünkü; test veri seti için doğruluk oranı %61'dir.

Tablo 4.7 Üç ihtimalli maç sonucu için iki gizli katmanlı yapay sinir ağları sonuçları

Gizli Katman 1.	Gizli Katman 2.	İter.	Bilgi Kriterleri			MSE		Doğruluk Oranı		
Nöron			AIC	AICc	BIC	Eğitim	Test	Eğitim	Test	Tüm
1	1	41	-1095,80	251,62	-944,90	0,141	0,142	0,68	0,69	0,68
2	2	64	-1063,06	288,26	-766,47	0,139	0,137	0,70	0,71	0,70
3	3	1000	-1078,33	280,15	-625,64	0,131	0,152	0,71	0,71	0,71
4	4	1000	-1206,43	163,32	-587,22	0,114	0,175	0,78	0,65	0,74
5	5	1000	-1472,78	-86,63	-676,66	0,089	0,240	0,84	0,61	0,77

Şekil 4.9 birinci gizli katmanda 3 nöron ve ikinci gizli katmanda da 3 nöron olan iki gizli katmanlı yapay sinir ağı modeli için roc eğrisini göstermektedir. Roc eğrisi; bu modelin de diğer modellerde olduğu gibi beraberlik kategorisini doğru sınıflandırmadığını ve bu nedenle model başarısının düştüğünü göstermektedir. Berabere kalınan maçlar dahil edilmediğinde; performans göstergelerinin ve durumsal değişkenlerin maç sonucunu tahmin etmeden ne kadar başarılı olduğu incelenebilir.



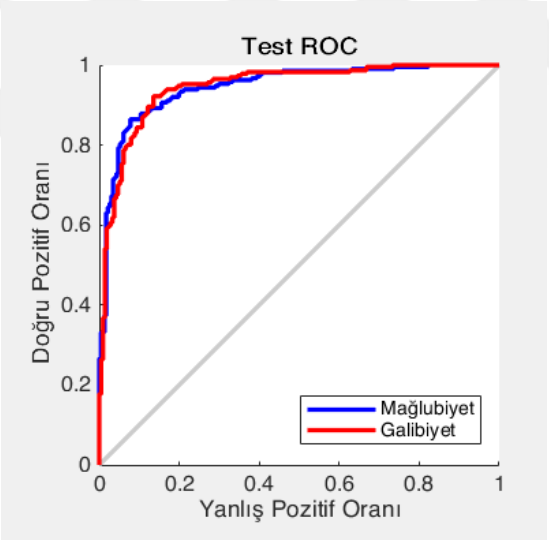
Şekil 4.9 İki gizli katmanlı ve 3-3 nöronlu YSA modeli için roc eğrisi (üç ihtimalli maç)

İki ihtimalli maç için tek gizli katmanlı yapay sinir ağı sonuçları Tablo 4.8'de gösterilmiştir. Yine üç ihtimali maç sonucu gibi 2 nörondan 10 nörona kadar dokuz farklı model incelenmiştir. Nöron sayısı artarken eğitim seti doğruluk oranı da artmaktadır; yani modelin veri setini iyi öğrendiğini göstermektedir. Fakat yukarıda verilen diğer sonuçlar gibi yine test veri seti için model başarısının düştüğü görülmektedir. Burada en iyi model; 2 nörona sahip olan eğitim veri setinde 0,88; test setinde 0,89 doğruluk oranına olan modeldir. Bilgi kriterleri eğitim veri seti üzerinden hesaplandığı için tüm bilgi kriterlerindeki en küçük değerlerin 10 nöronlu modelde bulunduğu söylenebilir. Bu durum eğitim veri seti için en iyi modelin 10 nöronlu model olduğunu gösterir zaten model doğruluk oranına bakıldığında 0,99 ile en iyi model eğitim seti için 10 nöronlu modeldir. Ancak bu model test veriseti için %78 doğruluk oranına sahiptir.

Şekil 4.10 iki ihtimalli maç sonucu tek gizli katmanlı ve 2 nöronlu bayesyen yapay sinir ağları modeli için roc eğrisini göstermektedir. Hem galibiyet hem de mağlubiyet durumları için sınıflandırma doğruluk oranları neredeyse aynıdır. Beraberlik maçları olmadığında performans göstergeleri ve durumsal değişkenler kullanarak maç sonucunu tahmin etmek daha gerçekçi sonuçlar vermektedir.

Tablo 4.8 İki ihtimalli maç sonucu için tek gizli katmanlı yapay sinir ağları sonuçları

Nöron	İter.	Bilgi Kriterleri			MSE		Doğruluk Oranı		
		AIC	AICc	BIC	Eğitim	Test	Eğitim	Test	Tüm
2	329	-1633,68	-587,73	-1396,27	0,095	0,087	0,88	0,89	0,88
3	530	-2001,33	-949,45	-1650,16	0,064	0,113	0,93	0,87	0,91
4	370	-2250,88	-1190,53	-1785,95	0,048	0,160	0,95	0,83	0,91
5	348	-2391,59	-1320,06	-1812,90	0,040	0,150	0,96	0,85	0,93
6	855	-2447,57	-1361,93	-1755,13	0,036	0,140	0,96	0,85	0,93
7	544	-2617,03	-1514,11	-1810,83	0,029	0,180	0,97	0,81	0,92
8	689	-2853,00	-1729,38	-1933,04	0,022	0,175	0,98	0,82	0,93
9	716	-2805,36	-1657,33	-1771,64	0,022	0,180	0,98	0,80	0,93
10	1000	-3436,38	-2259,92	-2288,91	0,012	0,200	0,99	0,78	0,93



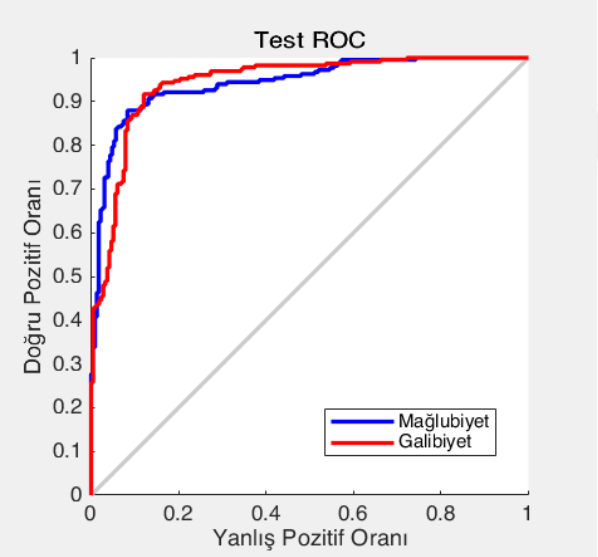
Şekil 4.10 Tek gizli katmanlı ve 2 nöronlu YSA modeli için roc eğrisi (iki ihtimalli maç)

İki ihtimalli maç sonucu için iki gizli katmanlı yapay sinir ağları sonuçları Tablo 4.9'da gösterilmektedir. En iyi modelin birinci gizli katmanda 1 ve ikinci gizli katmanda 1 nörona sahip olan bayesyen yapay sinir ağları modeli olduğu söylenebilir. Eğitim veri seti 0,89 doğruluk oranına sahip ve test seti için de 0,89 doğruluk oranına sahiptir.

Tablo 4.9 İki ihtimalli maç sonucu için iki gizli katmanlı yapay sinir ağları sonuçları

Gizli Katman Nöron 1.	2.	İter.	Bilgi Kriterleri AIC	AICc	BIC	MSE Eğitim	Test	Doğruluk Oranı Eğitim	Test	Tüm
1	1	335	-1670,31	-627,71	-1536,77	0,095	0,089	0,89	0,89	0,89
2	2	1000	-1887,26	-839,99	-1620,18	0,073	0,112	0,92	0,87	0,90
3	3	946	-2262,90	-1206,94	-1852,39	0,048	0,157	0,95	0,83	0,91
4	4	1000	-2455,88	-1385,98	-1892,04	0,038	0,152	0,96	0,83	0,92
5	5	1000	-2445,71	-1355,16	-1718,65	0,036	0,182	0,96	0,82	0,92

İki gizli katmanlı ve birinci katmanında 1 nöron ve ikinci katmanında da 1 nöron bulunan model için olan roc eğrisini göstermektedir. Galibiyet ve mağlubiyet sonuçları için doğruluk oranları benzerdir.



Şekil 4.11 İki gizli katmanlı ve 1-1 nöronlu YSA modeli için roc eğrisi (iki ihtimalli maç)

5
SONUÇ VE ÖNERİLER

Tezin amacı; futbolda en prestijli turnuvalardan biri olan Şampiyonlar Ligi grup aşaması için maç sonucu öngörüsünde bulunmaktır. Bu nedenle; performans göstergeleri ve durumsal değişkenler kullanılarak rakibin kalitesine göre yeni bir görselleştirme yaklaşımı olarak ÇBÖ analizi uygulanmıştır. Literatürdeki araştırmalardan farklı olarak örneklem sayısının on yıllık bir dönemi içermesi 1920 gözlemden oluşması daha ayrıntılı incelenmesi açısından önemlidir. [70-72]. İlk olarak; tek yönlü ANOVA; Tukey çoklu karşılaştırma testi; k-ortalamalar kümeleme; ÇBÖ ve karar ağacı analizi kullanılmıştır. Daha sonrasında; destek vektör makineleri ve bayes yapay sinir ağları uygulanmıştır. Ayrıca; kullanılan değişkenlerin takımların kalitesi dikkate alınarak incelenmesi; ÇBÖ yardımıyla daha iyi yorumlamaya katkıda bulunabilir.

İlk golün atılması tez çalışmasındaki en önemli buldulardan biridir ve literatürde çoğunlukla maç sonucuna etkisi araştırılmıştır [9; 73; 74]. Futbol veya buz hokeyi gibi maç skorunun düşük olduğu sporlar; ilk golü atmanın önemini göstermektedir [10; 75]. Avrupa'daki dört büyük ligde (Premier Lig; Bundesliga; Calcio ve La Liga) maç başına atılan gol sayısı 2,66'dır [76]. Armatas vd. [9] ve Molinuevo ve Barmejo [74] da ilk golü atan takımların maç kazanma olasılığının daha yüksek olduğu sonucuna varmışlardır. Bu çalışmada da daha güçlü; dengeli ve daha zayıf rakiplere karşı oynayan takımların; ilk golü atmaları halinde maçı kazanma şansı sırasıyla 0,59; 0,71 ve 0,92 olarak belirlenmiştir. UEFA Şampiyonlar Ligi'ne [10]; beş büyük Avrupa ligine [77] ve İngiltere Premier Ligi'ne [2] odaklanan diğer çalışmalar bulgularımızı desteklemektedir. Bu tez çalışmanın sonuçları; UEFA Şampiyonlar Ligi takımları için ilk golü atmanın önemini ortaya koymaktadır.

Maç sonucunu etkileyen bir diğer önemli performans göstergesi; rakiplerin kalitesine göre oluşturulan ve her üç karar ağacında da yer alan isabetli şutlardır. Rakibi zayıf olan takımlar için isabetli şut 6 ve 6'dan fazla olduğunda kazanma

şansı 0,14; dengeli rakiplere karşı oynayan takımlar için isabetli şut 8'den fazla ise kazanma şansı 0,16 ve güçlü rakiplere karşı isabetli şut 6'dan fazla olduğunda kazanma şansı 0,17 olasılıkla artmaktadır. İsabetli şutun kazanma şansını arttırdığı literatürdeki diğer çalışmalar da bu sonuçları desteklemektedir [23; 71; 78; 79]. İsabetli şutun kazanma şansını arttırdığı göz önünde bulundurulursa; antrenörlerin bu durum için özel çalışmalar yapması önerilebilir [79].

Güçlü rakiplere karşı oynayan takımlar için hava topu kazanma ve uzaklaştırma değişkenleri kazanma olasılığını artırırken; zayıf rakiplere karşı oynayan takımlar için adam geçme denemesinin kazanma şansını artırdığı söylenebilir. Ayrıca; korner; çalım yeme ve başarılı adam geçme değişkenleri kazanma şansı üzerinde olumsuz bir etkiye sahiptir. Önceki çalışmalar bulguları desteklemektedir [22; 25; 80]. Güçlü; zayıf ve dengeli takımlara karşı oynayan takımlar için maç yeri (ev sahibi ve deplasman) karar ağacı analizine dahil edilmiş ancak herhangi bir etki gözlenmemiştir. Literatürdeki birçok çalışmada maç yeri kazanmayı etkileyen önemli değişkenlerden biri olarak gösterilmiştir [25; 77]. Bunun sebebi; UEFA Şampiyonlar Ligi'nin yerel liglerle kıyaslandığında daha rekabetçi olmasından kaynaklanıyor olabilir; çünkü takımlar bu turnuvada mümkün olan en yüksek seviyede performans göstermeye çalışmaktadır [72].

Tez çalışmasında; istatistiksel olarak anlamlı bulunan performans göstergeleri; atılan ilk gol değişkeni ve ÇBÖ sonucunda maç sonucuna göre farklı olduğu görülen rakibin kalitesi değişkeni kullanılarak maç sonucu destek vektör makineleri ve yapay sinir ağları ile tahmin edilmiştir. Üç ihtimali bulunan (galibiyet; mağlubiyet ve beraberlik) maç sonucu için destek vektör makinelerinde en iyi model için doğruluk oranı %68 ve yapay sinir ağlarında en iyi model doğruluk oranı %70 olarak tespit edilmiştir. Hucaljuk ve Rakipović [103] çalışmasında; küçük bir örneklem kullanarak farklı makine öğrenmesi yöntemleri uyguladığında; en iyi doğruluk oranını %65 olarak bulmuştur. Bir başka çalışma da ise bu oran en iyi model için %55 olarak belirlenmiştir [104]. Buradaki farkın sebebi; bu tez çalışmasında daha büyük bir örneklem kullanılması ya da kullanılan değişkenlerin seçimlerinden kaynaklanıyor olabilir. Analizler sonunda galibiyet ve mağlubiyet sonuçlarının doğruluk oranlarının yüksek fakat beraberlik sonucunun

doğruluk oranının düşük olduğu belirlendi. Bu nedenle; veri setinden beraberlik sonucu çıkarılarak yapay sinir ağları ve destek vektör makineleri tekrar uygulandı. Burada doğruluk oranları sırasıyla; yapay sinir ağları için en iyi modelde 0,89 ve destek vektör makinaları için en iyi modelde 0,90 olarak belirlendi. Prasetio [2016] çalışmasında lojistik regresyon analizi kullanarak en iyi model için doğruluk oranını 0,70 olarak belirlemiştir. Bu tez çalışması göstermektedir ki iki ihtimalli maç sonucu için maçın kaybedilip kazanılacağı %90 oranından doğru tahmin edilebilmektedir.

Rakibi kalitesinin; performans göstergelerinin ve maç sonucunun UEFA Şampiyonlar Ligi performansı üzerindeki etkisinin çok değişkenli istatistiksel yaklaşımlar kullanarak ayrıntılı değerlendirmesi ve karşılaştırması; antrenörler için fayda sağlayabilir. Takımların performansla ilgili göstergeleri ile görsel analizleri oluşturularak; rakipler incelenebilir ve sonuçlar farklı taktiklerin geliştirilmesine ve rakibin kalitesine göre oyuncunun maçtan önce seçilmesine yardımcı olabilir. Tüm performans göstergelerinin toplu yapısı rakibin kalitesine göre ÇBÖ analizi ile görselleştirilebilir. Bu durumun antrenörlerin farklı taktikler geliştirmesi ve rakibin kalitesine göre maçtan önce en iyi oyuncuları seçmesi açısından önemli olacağı yadsınamaz.

Mevcut tez çalışmasının sonuçlarının pratikliğini geliştirmek için gözden geçirilmesi gereken ve gelecekteki uygulamalarda dikkate alınması gereken bazı sınırlamaları vardır. Bu çalışma takım performans göstergeleri üzerine kurulmuştur. Bireysel oyuncu performansları üzerinden de maç sonucunu öngörmek mümkün olabilir. Fakat bu haliyle bile geçmiş takım performansları üzerinden maç sonucunun tahminin (kazanılıp kaybedileceğinin) %90 oranında tahmin edilmesi mümkün olabilmektedir.

KAYNAKÇA

[1] M. Burch; S. L. Angelescu; G. Wallner; ve P. Lakatos; "Visual Analysis of FIFA World Cup Data;" in International Conference Information Visualisation; 2020: IEEE Computer Society.

[2] G. Bilek ve E. Ulas; "Predicting match outcome according to the quality of opponent in the English premier league using situational variables and team performance indicators;" International Journal of Performance Analysis in Sport; cilt 19; sayı 6; s. 930-941; 2019.

[3] A. Kubayi ve A. Toriola; "Differentiating African Teams from European Teams: Identifying the Key Performance Indicators in the FIFA World Cup 2018;" Journal of Human Kinetics; cilt 73; sayı 1; s. 203-208; 2020.

[4] C. M. Young; W. Luo; P. Gastin; J. Tran; ve D. B. Dwyer; "The relationship between match performance indicators and outcome in Australian Football;" Journal of Science and Medicine in Sport; cilt 22; sayı 4; s. 467-471; 2019.

[5] C. J. Zhou; A. L. Calvo; S. Robertson; ve M. A. Gomez; "Long-term influence of technical; physical performance indicators and situational variables on match outcome in male professional Chinese soccer;" Journal of Sports Sciences; cilt 39; sayı 6; s. 598-608; Mar 2021.

[6] T. McGarry; "Applied and theoretical perspectives of performance analysis in sport: Scientific issues and challenges;" International Journal of Performance Analysis in Sport; cilt 9; sayı 1; s. 128-140; 2009.

[7] A. Rizvandi; M. Taghipour Gharbi; M. Esmaeili; ve F. Ashraf Ganjoee; "The Evaluation of Performance Indicators of Coaches in Football Development;" Journal of Humanities Insights; cilt 3; sayı 04; s. 246-252; 2019.

[8] D. G. Liebermann; L. Katz; M. D. Hughes; R. M. Bartlett; J. McClements; ve I. M. Franks; "Advances in the application of information technology to sport performance;" Journal of Sports Sciences; cilt 20; sayı 10; s. 755-769; 2002.

[9] V. Armatas; A. Yiannakos; S. Papadopoulou; ve D. Skoufas; "Evaluation of goals scored in top ranking soccer matches: Greek “Super League” 2006-07;" Serbian Journal of Sports Sciences; cilt 3; sayı 1; s. 39-43; 2009.

[10] J. Garcia-Rubio; M. A. Gomez; C. Lago-Penas; ve S. J. Ibanez; "Effect of match venue; scoring first and quality of opposition on match outcome in the UEFA Champions League;" International Journal of Performance Analysis in Sport; cilt 15; sayı 2; s. 527-539; 2015.

[11] C. Lago-Penas; J. Lago-Ballesteros; A. Dellal; ve M. Gomez; "Game-related statistics that discriminated winning; drawing and losing teams from the Spanish soccer league;" Journal of Sports Science and Medicine; cilt 9; sayı 2; s. 288-293; 2010.

[12] A. Tenga; I. Holme; L. T. Ronglan; ve R. Bahr; "Effect of playing tactics on achieving score-box possessions in a random series of team possessions from Norwegian professional soccer matches;" Journal of Sports Sciences; cilt 28; sayı 3; s. 245-255; 2010.

[13] A. Tenga; I. Holme; L. T. Ronglan; ve R. Bahr; "Effect of playing tactics on goal scoring in Norwegian professional soccer;" Journal of Sports Sciences; cilt 28; sayı 3; s. 237-244; 2010.

[14] R. Aquino ve diğerleri; "Comparisons of ball possession; match running performance; player prominence and team network properties according to match outcome and playing formation during the 2018 FIFA World Cup;" International Journal of Performance Analysis in Sport; cilt 19; sayı 6; s. 1026-1037; 2019.

[15] A. Kubayi ve A. Toriola; "The Influence of Situational Variables on Ball Possession in the South African Premier Soccer League;" Journal of Human Kinetics; cilt 66; sayı 1; s. 175-181; 2019.

[16] C. Lago-Penas ve J. Lago-Ballesteros; "Game location and team quality effects on performance profiles in professional soccer;" Journal of Sports Science and Medicine; cilt 10; sayı 3; s. 465-471; 2011.

[17] A. Kubayi ve A. Toriola; "Match Performance Indicators that Discriminated Between Winning; Drawing and Losing Teams in the 2017 AFCON Soccer Championship;" Journal of Human Kinetics; cilt 72; sayı 1; s. 215-221; 2020.

[18] J. Lago-Ballesteros; C. Lago-Penas; ve E. Rey; "The effect of playing tactics and situational variables on achieving score-box possessions in a professional soccer team;" Journal of Sports Sciences; cilt 30; sayı 14; s. 1455-1461; 2012.

[19] M. Siegle ve M. Lames; "Game interruptions in elite soccer;" Journal of Sports Sciences; cilt 30; sayı 7; s. 619-624; 2012.

[20] C. H. Almeida; A. P. Ferreira; ve A. Volossovitch; "Effects of Match Location; Match Status and Quality of Opposition on Regaining Possession in UEFA Champions League;" Journal of Human Kinetics; cilt 41; sayı 1; s. 203-214; 2014.

[21] S. I. Ismail ve H. Nunome; "The key performance indicators that discriminate winning and losing; and successful and unsuccessful teams during 2016 FIFA Futsal World Cup;" Science and Medicine in Football; cilt 4; sayı 1; s. 68-75; 2020.

[22] C. Lago; "The influence of match location; quality of opposition; and match status on possession strategies in professional association football;" Journal of Sports Sciences; cilt 27; sayı 13; s. 1463-1469; 2009.

[23] H. Y. Liu; M. A. Gomez; B. Goncalves; ve J. Sampaio; "Technical performance and match-to-match variation in elite football teams;" Journal of Sports Sciences; cilt 34; sayı 6; s. 509-518; 2016.

[24] P. O'Donoghue; "Interacting performances theory;" International Journal of Performance Analysis in Sport; cilt 9; sayı 1; s. 26-46; 2009.

[25] J. B. Taylor; S. D. Mellalieu; N. James; ve D. A. Shearer; "The influence of match location; quality of opposition; and match status on technical performance in professional association football;" Journal of Sports Sciences; cilt 26; sayı 9; s. 885-895; 2008.

[26] L. Vilar; D. Araujo; K. Davids; ve C. Button; "The Role of Ecological Dynamics in Analysing Performance in Team Sports;" Sports Medicine; cilt 42; sayı 1; s. 1-10; 2012.

[27] G. Capobianco; U. Di Giacomo; F. Mercaldo; V. Nardone; ve A. Santone; "Can machine learning predict soccer match results?" in ICAART (2); 2019; s. 458-465.

[28] A. S. Timmaraju; A. Palnitkar; ve V. Khanna; "Game ON! Predicting English Premier League Match Outcomes;" Stanford University; 2013.

[29] J. Hucaljuk ve A. Rakipović; "Predicting football scores using machine learning techniques;" in 2011 Proceedings of the 34th International Convention MIPRO; 2011; s. 1623-1627: IEEE.

[30] F. A. Moura; L. E. B. Martins; ve S. A. Cunha; "Analysis of football game-related statistics using multivariate techniques;" Journal of Sports Sciences; cilt 32; sayı 20; s. 1881-1887; 2014.

[31] P. E. Green; Mathematical tools for applied multivariate analysis. Academic Press; 2014.

[32] W. H. Kruskal ve W. A. Wallis; "Use of ranks in one-criterion variance analysis;" Journal of the American statistical Association; cilt 47; sayı 260; s. 583-621; 1952.

[33] R. Witte ve J. Witte; Statistics; 11th Edition ed. Wiley; New York.; 2017.

[34] M. Tomczak ve E. Tomczak; "The need to report effect size estimates revisited. An overview of some recommended measures of effect size;" Trends in Sport Sciences; cilt 21; sayı 1; 2014.

[35] G. M. Sullivan ve R. Feinn; "Using effect size—or why the P value is not enough;" Journal of graduate medical education; cilt 4; sayı 3; p. 279; 2012.

[36] B. S. Everitt; S. Landau; M. Leese; ve D. Stahl; "Cluster Analysis. –John Wiley & Sons;" Ltd.; New York; p. 330; 2011.

[37] J. MacQueen; "Some methods for classification and analysis of multivariate observations;" in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability; 1967; cilt 1; sayı 14; s. 281-297: Oakland; CA; USA.

[38] S. Na; L. Xumin; ve G. Yong; "Research on k-means clustering algorithm: An improved k-means clustering algorithm;" in 2010 Third International Symposium on intelligent information technology and security informatics; 2010; s. 63-67: IEEE.

[39] I. Borg ve P. J. Groenen; Modern multidimensional scaling: Theory and applications. Springer Science & Business Media; 2005.

[40] J. B. Kruskal; "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis;" Psychometrika; cilt 29; sayı 1; s. 1-27; 1964.

[41] R. A. Johnson ve D. W. Wichern; Applied multivariate statistical analysis (sayı 8). Prentice hall Upper Saddle River; NJ; 2002.

[42] U. Braga-Neto; "Fundamentals of Pattern Recognition and Machine Learning."

[43] S. Suthaharan; "Machine learning models and algorithms for big data classification;" Integr. Ser. Inf. Syst; cilt 36; s. 1-12; 2016.

[44] P.-N. Tan; M. Steinbach; ve V. Kumar; Introduction to data mining. Pearson Education India; 2016.

[45] S. Raschka; "Model evaluation; model selection; and algorithm selection in machine learning;" arXiv preprint arXiv:1811.12808; 2018.

[46] K. Polat; S. Güneş; ve S. Tosun; "Diagnosis of heart disease using artificial immune recognition system and fuzzy weighted pre-processing;" Pattern Recognition; cilt 39; sayı 11; s. 2186-2193; 2006.

[47] H. Akaike; "Information theory and an extension of the maximum likelihood principle;" in Selected papers of hirotugu akaike: Springer; 1998; s. 199-213.

[48] C. M. Hurvich ve C.-L. Tsai; "Regression and time series model selection in small samples;" Biometrika; cilt 76; sayı 2; s. 297-307; 1989.

[49] G. Schwarz; "The annals of statistics;" Estimating the dimension of a model; cilt 6; s. 461-464; 1978.

[50] S. Marsland; Machine learning: an algorithmic perspective. CRC press; 2015.

[51] T. M. Therneau ve E. J. Atkinson; "Mayo Foundation. An introduction to recursive partitioning using the RPART routines;" ed: Technical Report. 2019.

[52] M. J. Zaki ve W. Meira Jr; Data Mining and Machine Learning: Fundamental Concepts and Algorithms. Cambridge University Press; 2019.

[53] C. M. Bishop; Pattern recognition and machine learning. Springer; 2006.

[54] C. J. Burges; "A tutorial on support vector machines for pattern recognition;" Data mining and knowledge discovery; cilt 2; sayı 2; s. 121-167; 1998.

[55] V. Vapnik; "Statistical learning theory new york;" NY: Wiley; 1998.

[56] V. Vapnik; The nature of statistical learning theory. Springer science & business media; 2013.

[57] K.-R. Müller; S. Mika; K. Tsuda; ve K. Schölkopf; "An introduction to kernel-based learning algorithms;" Handbook of Neural Network Signal Processing; s. 4-1-4-40; 2018.

[58] J. P. Hwang; S. Park; ve E. Kim; "Dual margin approach on a Lagrangian support vector machine;" International Journal of Computer Mathematics; cilt 88; sayı 4; s. 695-708; 2011.

[59] C.-P. Lee ve D. Roth; "Distributed box-constrained quadratic optimization for dual linear SVM;" in International Conference on Machine Learning; 2015; s. 987-996: PMLR.

[60] C. Savas ve F. Dovis; "The impact of different kernel functions on the performance of scintillation detection based on support vector machines;" Sensors; cilt 19; sayı 23; p. 5219; 2019.

[61] J. Beyerer; M. Richter; ve M. Nagel; Pattern recognition: introduction; features; classifiers and principles. Walter de Gruyter GmbH & Co KG; 2017.

[62] A. J. Tallón-Ballesteros; J. C. Riquelme; ve R. Ruiz; "Semi-wrapper feature subset selector for feed-forward neural networks: Applications to binary and multi-class classification problems;" Neurocomputing; cilt 353; s. 28-44; 2019.

[63] G. Panchal; A. Ganatra; Y. Kosta; ve D. Panchal; "Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers;" International Journal of Computer Theory and Engineering; cilt 3; sayı 2; s. 332-337; 2011.

[64] C. M. Bishop; Neural networks for pattern recognition. Oxford university press; 1995.

[65] T. Auld; A. W. Moore; ve S. F. Gull; "Bayesian neural networks for internet traffic classification;" IEEE Transactions on neural networks; cilt 18; sayı 1; s. 223-239; 2007.

[66] D. J. MacKay; "Bayesian interpolation;" Neural computation; cilt 4; sayı 3; s. 415-447; 1992.

[67] R. C. Team; "R: a language and environment for statistical computing. R Found;" Stat. Comput. Vienna; Austria; 2017.

[68] S. Milborrow; "rpart. plot: Plot'rpart'Models: An Enhanced Version of'plot. rpart'version 2.1. 2; 2017;" ed.

[69] T. M. Therneau; B. Atkinson; ve B. Ripley; "rpart: Recursive partitioning;" R package version; cilt 3; sayı 3.8; 2010.

[70] M. D. Bush; D. T. Archer; R. Hogg; ve P. S. Bradley; "Factors Influencing Physical and Technical Variability in the English Premier League;" ; International Journal of Sports Physiology and Performance; cilt 10; sayı 7; s. 865-872; 2015.

[71] H. Y. Liu; W. G. Hopkins; ve M. A. Gomez; "Modelling relationships between match events and match outcome in elite football;" ; European Journal of Sport Science; cilt 16; sayı 5; s. 516-525; 2016.

[72] Q. Yi; M. A. Gomez; H. Y. Liu; B. H. Gao; F. Wunderlich; ve D. Memmert; "Situational and Positional Effects on the Technical Variation of Players in the

UEFA Champions League;" ; Frontiers in Psychology; cilt 11; p. 10; Jun 2020; sayı 1201.

[73] Y. Michailidis; C. Michailidis; ve E. Primpa; "Analysis of goals scored in European Championship 2012;" 2013.

[74] J. S. Molinuevo ve J. P. Bermejo; "El efecto de marcar primero y la ventaja de jugar en casa en la liga de fútbol y en la liga de fútbol sala de España;" Revista de psicología del deporte; cilt 21; sayı 2; s. 301-308; 2012.

[75] B. M. Jones; "Scoring first and home advantage in the NHL;" International Journal of Performance Analysis in Sport; cilt 9; sayı 3; s. 320-331; 2009.

[76] C. Anderson ve D. Sally; The numbers game: Why everything you know about soccer is wrong. Penguin; 2013.

[77] C. Lago-Penas; M. Gomez-Ruano; D. Megias-Navarro; ve R. Pollard; "Home advantage in football: Examining the effect of scoring first on match outcome in the five major European leagues;" International Journal of Performance Analysis in Sport; cilt 16; sayı 2; s. 411-421; 2016.

[78] L. J. Mao; Z. F. Peng; H. Y. Liu; ve M. A. Gomez; "Identifying keys to win in the Chinese professional soccer league;" International Journal of Performance Analysis in Sport; cilt 16; sayı 3; s. 935-947; 2016.

[79] C. J. Zhou; S. L. Zhang; A. L. Calvo; ve Y. X. Cui; "Chinese soccer association super league; 2012-2017: key performance indicators in balance games;" International Journal of Performance Analysis in Sport; cilt 18; sayı 4; s. 645-656; 2018.

[80] H. Y. Liu; M. A. Gomez; C. Lago-Penas; ve J. Sampaio; "Match statistics related to winning in the group stage of 2014 Brazil FIFA World Cup;" Journal of Sports Sciences; cilt 33; sayı 12; s. 1205-1213; 2015.

TEZDEN ÜRETİLMİŞ YAYINLAR

Makale

1. Parim; C.; Güneş; M. Ş.; Büyüklü; A. H.; ve Yıldız; D. “Prediction of match outcomes with multivariate statistical methods for the group stage in the UEFA Champions League” Journal of Human Kinetics; cilt 79; 2021. (SCI Expanded)

Ödüller

1. TÜBİTAK Yurtiçi Doktora Başarı Bursu