Teknoloji Fakültesi

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

MAKİNE ÖĞRENMESİ İLE NBA MVP TAHMİNİ

BİTİRME PROJESİ 1.ARA RAPORU

Bilgisayar Mühendisliği Bölümü

ÖĞRENCİLER

Davutcan KÖSEMEN 170421030

Ceyhun AY 170420844

DANIŞMAN

Dr. Öğr. Üyesi Eyüp Emre ÜLKÜ

İSTANBUL, 2025

MARMARA ÜNİVERSİTESİ
TEKNOLOJİ FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Marmara Üniversitesi Teknoloji Fakültesi Bilgisayar Mühendisliği Öğrencileri Davutcan Kösemen ve Ceyhun Ay tarafından "**MAKİNE ÖĞRENMESİ İLE NBA MVP TAHMİNİ**" başlıklı proje çalışması, xxx tarihinde savunulmuş ve jüri üyeleri tarafından başarılı bulunmuştur.

Jüri Üyeleri

Dr. Öğr. Üyesi Eyüp ÜLKÜ Marmara Üniversitesi	**(Danışman)**	(İMZA)……..
Prof. Dr. Xxx xxx Marmara Üniversitesi	(Üye)	(İMZA)……..
Prof. Dr. Xxx xxx Marmara Üniversitesi	(Üye)	(İMZA)……..

ÖNSÖZ

Bu proje çalışması fikrinin oluşması ve ortaya çıkmasındaki önerisiyle birlikte çalışmamız süresince karşılaştığımız bütün problemlerde okul içerisinde ve okul dışarısında sabırla yardım ve bilgilerini, maddi ve manevi desteklerini esirgemeyen, tüm desteğini sonuna kadar yanımızda hissettiğimiz değerli hocamız sayın Dr. Öğr. Üyesi Eyüp Emre Ülkü'ye en içten teşekkürlerimizi sunarız.

İÇİNDEKİLER

ÖZET

Bu araştırma projesi, NBA organizasyonunun en değerli oyuncu ödülüne yönelik tahmin sistemleri geliştirmek amacıyla oluşturulmuş yenilikçi bir makine öğrenmesi yaklaşımı sunmaktadır. Projede 1997-2022 yılları arasındaki NBA oyuncu istatistiklerini içeren geniş bir veri seti üzerinde hem klasik hem de modern makine öğrenmesi algoritmaları uygulanarak MVP adaylarının belirlenmesinde yüksek doğruluğa sahip bir model oluşturulması hedeflenmektedir. Mevcut çalışmalardan farklı olarak veri seti üzerinde detaylı bir veri analizi yapılmaktadır. Bu bağlamda, projenin özgünlüğü literatürde

daha az incelenmiş olan özellikleri, örneğin clutch time verilerinden türetilen stres temelli performans metriklerini MVP tahminine dahil ederek daha kapsamlı bir değerlendirme elde etmeye dayanıyor. Araştırma yöntemimiz veri temizleme, ön işleme, modelleme ve hiperparametre optimizasyonu gibi aşamaları içeren kapsamlı bir süreci kapsamaktadır. Bu süreçte Yapay Sinir Ağları, Random Forest, XGBoost, CatBoost gibi algoritmalar kullanılarak, model performansları karşılaştırılarak ve en iyi sonuçlar elde edilerek MVP adaylarının tahmin edilmesi sağlanacaktır. Sonuçların güvenilirliğini ve tekrarlanabilirliğini sağlamak için araştırma metodolojik açıdan dikkatlice yapılandırılmıştır. Geliştirilen model, spor kulüpleri tarafından oyuncu performans değerlendirmelerinde ve stratejik karar alma süreçlerinde kullanılabilir. Ayrıca, veri bilimi ve yapay zeka alanında ileri düzey araştırmalara altyapı sunacaktır.

ABSTRACT

This research project presents an innovative machine learning approach aimed at developing prediction systems for the MVP award in NBA. The project seeks to build a highly accurate model for identifying MVP candidates by applying both classical and modern machine learning algorithms on a comprehensive dataset containing NBA player statistics from 1997 to 2022. Unlike existing studies, this project conducts a detailed data analysis on the dataset. In this context, the originality of the project lies in incorporating less-studied features in the literature, such as stress-based performance metrics derived from clutch time data, into MVP prediction to achieve a more comprehensive evaluation. Our research methodology includes a comprehensive process encompassing data cleaning, preprocessing, modeling, and hyperparameter optimization. Algorithms such as Artificial Neural Networks, Random Forest, XGBoost, and CatBoost are utilized to compare model performances and achieve optimal results for predicting MVP candidates. To ensure the reliability and reproducibility of the results, the research has been methodologically structured with great attention to detail. The developed model can be utilized by sports clubs for player performance evaluations and strategic decision-making processes. Additionally, it provides a foundation for advanced research in data science and artificial intelligence.

KISALTMALAR

NBA : National Basketball Association

WNBA : Women's National Basketball Association

MVP : Most Valuable Player

RNN : Recurrent Neural Network

CNN : Convolutional Neural Network

GNN : Graph Neural Network

ANN : Artificial Neural Network

KNN : K-Nearest Neighbors

LSTM : Long Short-Term Memory

LRM : Linear Regression Model

PCA : Principal Component Analysis

FUCOM : Full Consistency Method

MLP : Multi-Layer Perceptron

HAC : Hierarchical Agglomerative Clustering

IQR : Interquartile Range

DEA : Data Envelopment Analysis

SMOTE : Synthetic Minority Over-sampling Technique

SVM : Support Vector Machine

DRB : Defensive Rebound

TPP : Three-Point Percentage

FT : Free Throw

TRB : Total Rebounds

PPG : Points Per Game

RPG : Rebounds Per Game

APG : Assists Per Game

VORP : Value Over Replacement Player

OWS : Offensive Win Shares

WS : Win Shares

FTA : Free Throw Attempts

PTS : Points

FG : Field Goals

2P : Two-Point Field Goal

ŞEKİL LİSTESİ

TABLO LİSTESİ

1. GİRİŞ

Takım sporları insanların hayatında tartışmasız bir şekilde önemli bir yere sahiptir. Her büyük şehirde profesyonel spor takımları bulunmaktadır ve taraftarlar genellikle tuttukları takımların sporcularını rol modelleri olarak görerek onlara büyük destek vermektedir. Bu sporların içinde barındırdığı rekabet ruhu, hangi takımın daha üstün olduğundan, çeşitli oyuncuların yeteneklerine kadar uzanan hararetli tartışmaları her zaman ateşlemiştir. Sezon sonunda şampiyon olan takım, diğer takımlara karşı olan üstünlüğünü kanıtlamış olur. Oyuncular ise sezon sonunda çeşitli ödüller kazanır. Aynı zamanda basketbol sezonunun sonunda, bireysel oyuncular için de çeşitli ödüller verilir. NBA dünya çapında en önde gelen profesyonel basketbol liglerinden biridir. Bu organizasyonda çeşitli bireysel ödüller bulunmasına rağmen, en önemli ödül tartışmasız şekilde MVP ödülüdür. MVP ödülünü kazanan oyuncuyu seçme konusundaki genel görüş, hem istatistiksel performansı hem de sezon boyunca takım başarısının genel görünümünü kapsayan kapsamlı bir bağlamı dikkate alması gerektiği yönündedir.

NBA normal sezonunun en değerli oyuncusunu tahmin etmek, spor analitiği ve makine öğrenimi alanlarını birleştiren ilgi çekici bir problemdir. Bu proje, veri ön işleme, özellik mühendisliği ve son teknoloji algoritmaları bir araya getirerek doğru ve güvenilir bir tahmin modeli geliştirmeyi amaçlamaktadır. Ana hedefimiz yalnızca yüksek tahmin doğruluğuna ulaşmak değil, aynı zamanda MVP seçimini etkileyen kritik faktörlere dair içgörüler sunmaktır.

Çalışmamızda birden fazla NBA sezonundaki oyuncu istatistiklerini kapsayan veri seti üzerinde mevcut çalışmalardan farklı olarak detaylı bir veri analizi yapılmaktadır. Ayrıca, veri setimize yenilikçi bir özellik olan “clutch time skoru”nu ekledik. Web kazıma (web scraping) ve özel bir formül kullanılarak elde edilen bu özellik, oyuncuların yüksek baskı anlarındaki performansını nicel olarak değerlendirmektedir ve genellikle geleneksel analizlerde göz ardı edilmektedir. Veri setimize eklenen bir diğer yenilikçi özellik ise “takım sıralaması”dır. Galibiyet-mağlubiyet yüzdesi yerine bu özelliğin eklenme nedeni, bir takımın sezonluk başarısının yalnızca galibiyet yüzdesiyle değerlendirilemeyeceği, ligin genelindeki göreceli performansının dikkate

alınması gerektiğinin kabul edilmesidir.Bu özellikler modele sağlam ve tarafsız katkılar sağlamıştır.

Yöntemimiz, Random Forest, Gradient Boosting ve Sinir Ağları gibi çeşitli makine öğrenimi tekniklerinin yanı sıra çapraz doğrulama ve hiperparametre optimizasyonu gibi ileri düzey doğrulama stratejilerini kullanmayı içermektedir. Modelimizin doğruluğunu artırmak için veri ön işleme aşamasında özellik seçimi (Feature Selection) ve aykırı değer temizleme (Outlier Detection) teknikleri uygulanmıştır. Bu yaklaşımları karşılaştırarak, hem doğruluk hem de hesaplama verimliliği açısından MVP tahmini için en etkili algoritmaları belirlemeyi hedefliyoruz. RNN, LSTM ve CNN gibi gelişmiş sinir ağı mimarilerinin entegrasyonu, modelin zamansal desenleri ve karmaşık oyuncu etkileşimlerini yakalama yeteneğini daha da artırarak spor analizinde yeni bir standart belirlemektedir. Ayrıca, bu çalışma, verilerdeki gizli yapıları belirlemek için PCA ve kümeleme yöntemleri gibi boyut azaltma tekniklerinin entegrasyonunu da keşfetmektedir.

Bu araştırmanın daha geniş etkileri, yalnızca MVP tahminiyle sınırlı değildir. Bu çalışmada elde edilen teknikler ve içgörüler, takım oluşturma stratejileri, performans değerlendirme metrikleri ve hatta profesyonel spor organizasyonlarındaki karar alma süreçlerini şekillendirebilir. Zengin veri kaynaklarıyla bir araya getirilen son teknoloji, bu proje kapsamında spor analitiği ve yapay zekânın büyüyen kesişimine katkıda bulunarak, sporda veri odaklı karar almanın dönüştürücü potansiyelini sergilemektedir.

Çalışma şu şekilde yapılandırılmıştır: 2. bölümde ilgili çalışmalar özetlenmiştir. 3. bölümde yöntem detaylandırılmıştır. 4. bölümde araştırmamızdan çıkarılan temel bulguları kapsayan kapsamlı bir tartışma sunulmuştur. 5. bölümde sonuçlar ile birlikte gelecekteki araştırmalar için fırsatlar özetlenmiştir. Bu yapılandırılmış analiz yoluyla, NBA MVP tahmini için kapsamlı bir çerçeve sunmayı ve bu alanda mevcut zorluklara ve fırsatlara değinmeyi amaçlıyoruz.

2. LİTERATÜR TARAMASI

Son yıllarda, ML ve AI teknikleri, özellikle NBA sezonunun en değerli oyuncusu gibi sonuçları tahmin etmek amacıyla spor analizlerinde giderek daha fazla uygulanmaktadır. NBA MVP'si, oyuncu performans metrikleri, takım başarısı ve oyuna genel etkisi gibi bir dizi faktöre dayalı olarak verilmektedir. Doğru tahmin modelleri, takımların, analistlerin ve taraftarların resmi ödül açıklanmadan önce potansiyel MVP adaylarını değerlendirmelerine yardımcı olabilir. Birçok çalışma, farklı makine öğrenimi yaklaşımlarını kullanarak MVP'yi tahmin etmeye çalışmış ve oyun başına puan, asist, ribaundlar ve ileri düzey istatistikler gibi geniş bir performans metriği yelpazesi üzerinde analizler yapmıştır.

Mason Chen [1] çalışmasında, MVP oyuncularını tahmin etmek için sezon boyunca oyuncuların bireysel performanslarına ve takım puanlarına dayalı olarak Uniform Model, Weighted Model, Power Model ve Discriminant Clustering Model olmak üzere 4 model değerlendirilmiştir. Power Model, takım başarısını da hesaba katarak %69 ile en yüksek doğruluğa sahip olmuştur. Yine Mason Chen ve Charles Cen [2], 2019-2020 NBA sezonu için MVP tahmini yapmak amacıyla Uniform, Weighted, Power ve Discriminant MVP Index modellerini kullanmıştır. Çalışmalarında oyuncu istatistiklerine Z-Dönüşümü (Z-Transformation) uygulayarak verileri normalize etmiş ve değişken seçiminde kollineariteyi minimize etmek için en önemli 6 bağımsız değişkeni belirlemişlerdir. Bu yöntemlerle, MVP tahmin modelinin doğruluğunu artırmayı hedeflemişlerdir. Zhai ve Xu [3] çalışmasında, NBA MVP tahmini için "Teammates" metriği önerilmiştir. Çalışmada, Random Forest, XGBoost ve LightGBM gibi makine öğrenimi modelleri kullanılarak MVP tahminleri yapılmış ve "Teammates" metriğinin belirli durumlarda tahmin doğruluğunu artırabileceği gösterilmiştir. Ancak, veri setinin yalnızca son beş yılı kapsaması, modelin genelleme kapasitesini sınırlamıştır. Chapman [4], MVP tahmini için LightGBM modelini Overlapping teknikleri ile birleştirerek %80,65 doğruluk oranı ile tahmin sonuçlarını elde etmiştir. Cheng [5] çalışmasında, NBA MVP tahmini için farklı makine öğrenimi algoritmaları karşılaştırılmış ve MVP oy payı (win share) tahmini yapılmıştır. 40 yıllık NBA verileri kullanılarak Linear Regression, Random Forest, XGBoost ve Neural Network Regression gibi modeller test edilmiş, en iyi sonuçları XGBoost Regression

Modeli (%63.99 R^2, %22.90 MAPE) vermiştir. Han ve Yu [6], çalışmasında, Random Forest algoritması kullanılarak NBA MVP tahmini yapılmıştır. 50 yıllık verilerden PPG, FG%, 3P%, RPG, APG ve TOV gibi karar faktörleri belirlenmiş ve Early NBA ile Small Ball dönemleri karşılaştırılmıştır. Model 2021-2023 MVP tahmini için uygulanmış ve %25 doğruluk oranına ulaşmıştır, ancak sonuçlar modelin daha fazla iyileştirilmesi gerektiğini göstermektedir. Malik [7] çalışmasında ANN, KNN ve LRM gibi farklı modeller test edilmiş ve LRM çerçevesine dayalı modelin en güvenilir tahminleri ürettiği belirlenmiştir. Özkir ve Değirmenci [8] çalışmasında, NBA MVP tahmini için çok kriterli karar verme yöntemi önerilmiştir. 535 oyuncunun 2022-2023 sezonu istatistikleri analiz edilerek, FUCOM ve Axiomatic Design yöntemleri kullanılmıştır. Çalışmada, Joel Embiid en düşük bilgi içeriğine sahip oyuncu olarak MVP seçilmiş, bu sonuç gerçek MVP seçimleriyle tutarlı bulunmuştur. Albert et al. [9] çalışmasında, NBA All-Star oyuncularını tahmin etmek için hibrit bir makine öğrenimi modeli önerilmiştir. Random Forest, AdaBoost ve MLP algoritmaları kullanılarak, farklı modellerin sonuçları bir ANN ile birleştirilmiş ve doğruluk oranı artırılmıştır. 1980-2021 yılları arasındaki 17.000 oyuncu verisi kullanılmış ve en yüksek başarı oranı MLP modeliyle %88.7 doğruluk ve %81 duyarlılık olarak elde edilmiştir. Ke et al. [10] çalışmasında, NBA ve WNBA için takım kadro optimizasyonu amacıyla denetimli ve denetimsiz makine öğrenimi yöntemlerini birleştiren bir çerçeve önerilmiştir. PCA ile boyut azaltma, kümeleme (HAC ve k-means), sinir ağı tabanlı tahmin modeli ve oyuncu derecelendirme sistemi kullanılarak, en iyi kadro kombinasyonu belirlenmiştir. Thabtah et al. [11] çalışmasında, NBA maç sonuçlarını tahmin etmek için makine öğrenimi yöntemleri kullanılmıştır. Naïve Bayes, ANN ve Karar Ağaçları (Decision Tree) algoritmaları test edilerek, en önemli özelliklerin belirlenmesi amaçlanmıştır. DRB en etkili faktör olarak öne çıkarken, TPP, FT ve TRB gibi özelliklerin de tahmin doğruluğunu artırdığı bulunmuştur. Li et al. [12] çalışmasında, NBA takım performansını tahmin etmek için DEA tabanlı veri odaklı bir yaklaşım geliştirilmiştir. Çalışmada, Golden State Warriors'un 2011-2015 sezonlarına ait verileri kullanılarak 2015-16 sezonu için tahminler yapılmıştır. Sonuçlar, DEA tabanlı yaklaşımların takım performansını tahmin etmekte başarılı olduğunu ve kazanma olasılığını artıracak içgörüler sağladığını göstermektedir. Chen et al [13] çalışmasında, NBA MVP tahmini için yapay sinir ağı tabanlı bir model

geliştirmiştir. 1997-2019 sezonları arasındaki oyuncu performans verileriyle eğitilen model, 2009-2010 ve 2016-2017 sezonlarından rastgele seçilen test verileri ile başarıyla test edilmiş ve sırasıyla LeBron James ve Russell Westbrook'un MVP seçileceğini doğru tahmin etmiştir. Çalışma, optimize edilmiş özellikler kullanarak sezon bazlı MVP tahmini için makine öğrenimi modellerinin uygulanabilirliğini göstermektedir. Son olarak Mertz et al. [14] ise çalışmasında, en iyi NBA oyuncularını sıralamak için doğrusal regresyon temelli bir model önerilmiştir. Çalışmada oyuncuların performanslarını analiz etmek amacıyla PPG, RPG, APG gibi temel değişkenler ile kazanılan NBA şampiyonlukları gibi faktörler değerlendirilmiştir.

Mevcut çalışmalar, kazananları tahmin etme konusunda değerli içgörüler sağlasa da, çalışmamız birkaç önemli noktada farklılaşmaktadır. Çoğu önceki çalışmanın sezon genelindeki istatistiklere odaklanmasının aksine, modelimiz "clutch-time performansını" da dikkate alan ilk çalışmalardan biridir. Bu özellik, oyuncuların maçların son anlarında, yüksek baskı altında nasıl performans gösterdiğini ölçen bir metrik olarak geliştirilmiştir. Özel bir formül kullanılarak ve web kazıma (web scraping) yöntemiyle elde edilen bu veri, oyuncuların kritik anlardaki etkinliklerini sayısal olarak ifade etmektedir. Modelimize dahil edilen bu özellik, yalnızca bireysel istatistiklerden ziyade oyuncuların maç sonucuna doğrudan etkisini ölçmeye yönelik bir katkı sağlamaktadır. Buna ek olarak, mevcut literatürde takım başarısını ölçmek için genellikle "win percentage" (galibiyet yüzdesi) metriği kullanılırken, çalışmamızda "team ranking" (takım sıralaması) metriği tercih edilmiştir. Bu farklılığın nedeni, galibiyet yüzdesinin her sezon farklı bağlamlara sahip olmasıdır. Örneğin, bir takım bir sezonda %80 galibiyet oranıyla şampiyon olabilirken, başka bir sezonda %75 galibiyet oranı ligi lider tamamlamak için yeterli olabilir. Bu yaklaşım, oyuncuların bireysel istatistiklerinin takımlarının genel başarısı içindeki bağlamını daha doğru bir şekilde analiz etmeye olanak tanımaktadır. Ayrıca modelimizde boyut azaltma teknikleri kullanılmaktadır. Özellikle PCA yöntemiyle yüksek boyutlu veri setimizdeki en önemli bileşenleri seçerek, modelin daha genel geçer sonuçlar üretmesini sağlamaktayız. Literatürdeki çalışmaların büyük bir kısmı PCA veya diğer boyut azaltma yöntemlerini kullanmamış, bu da modelin fazla gürültü içeren veri ile eğitilmesine sebep olmuştur. Modelin doğruluğunu artırmak için ise kümeleme

yöntemlerini entegre ettik. Kümeleme yöntemleri, oyuncuları farklı gruplara ayırarak modelin her grupta daha iyi genelleme yapmasını sağlamaktadır. MVP tahmin çalışmalarında genellikle oyuncular tek bir büyük grup olarak ele alınırken, biz oyuncuları istatistiksel olarak benzer özelliklere sahip olan kümelere ayırarak modelin farklı oyuncu profillerine göre daha hassas tahminler yapmasını sağladık. Son olarak, model doğrulama aşamasında çapraz doğrulama tekniklerini titizlikle uyguladık. Literatürdeki çalışmalar genellikle modeli tek bir test seti üzerinde değerlendirirken, biz K-Fold Cross-Validation gibi yöntemler kullanarak modelin farklı veri bölümlerinde nasıl performans gösterdiğini detaylı bir şekilde analiz ettik. Böylece, modelimizin farklı sezonlar ve oyuncu grupları arasında genelleme yeteneğini artırdık. Özetle, çalışmamız MVP tahmin sürecine literatürde eksik kalan birçok önemli unsuru ekleyerek, daha kapsamlı ve genelleştirilebilir bir model ortaya koymaktadır.

3. MATERYAL VE YÖNTEM

3.1 Araştırma Tasarımı

Bu çalışmanın amacı basketbol alanında MVP tahmini yapabilecek bir yapay zeka modeli geliştirmektir. Bu amaca ulaşmak için farklı makine öğrenmesi ve derin öğrenme modelleri test edilmiş, en başarılı olanlar belirlenerek optimize edilmiştir.

Veri seti 1997-2022 yılları arasındaki NBA oyuncularının istatistiklerini içermekte olup, 1997-2017 arası eğitim ve test verisi, 2018-2022 arası ise tahmin verisi olarak ayrılmıştır. Öğrenme performansını test etmek için farklı veri bölme oranları kullanılmış ve en uygun olanları belirlenmiştir.

Season	Player	Pos	Age	Tm	G	MP	FG	3P	3PA	2P
1997	mahmoud abdul-rauf	PG	27	SAC	75	28,4	5,5	1,3	3,3	4,2
1997	shareef abdur-rahim	PF	20	VAN	80	35	6,9	0,1	0,3	6,8
1997	rafael addison	SF	32	CHH	41	8,7	1,2	0,2	0,5	1
1997	cory alexander	PG	23	SAS	80	18,2	2,4	1,2	3,2	1,3
1997	jerome allen	SG	24	TOT	76	12,4	1	0,4	1,2	0,6
1997	ray allen	SG	21	MIL	82	30,9	4,8	1,4	3,6	3,3
1997	derrick alston	C	24	ATL	2	5,5	0	0	0	0

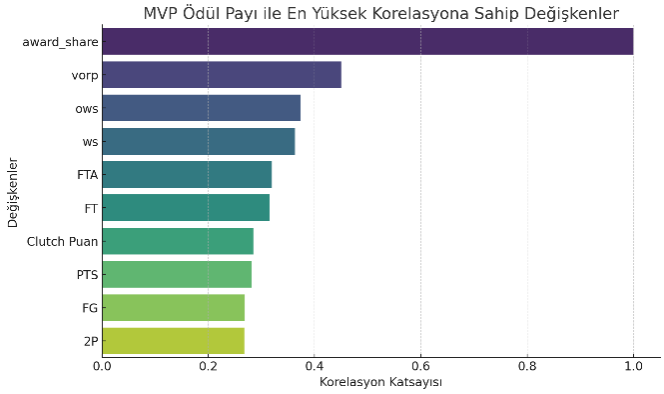
Tablo 3.1.1 : Verisetinden Örnek

MVP ödülü ile en yüksek korelasyona sahip değişkenler şunlardır:

1. **VORP (0.45)** - Value Over Replacement Player, oyuncunun yerine ortalama bir oyuncu konduğunda takımın nasıl etkileneceğini ölçer.
2. **OWS (0.37)** - Offensive Win Shares, hücum katkısıyla kazanılan maç sayısını gösterir.
3. **WS (0.36)** - Win Shares, oyuncunun toplam katkısını ölçer.
4. **FTA (0.32)** - Serbest atış denemeleri, oyuncunun faul alıp serbest atış çizgisine gitme sıklığını gösterir.
5. **FT (0.32)** - İsabetli serbest atış sayısı.
6. **Clutch Puan (0.28)** - Kritik anlarda atılan sayılar.

7. **PTS (0.28)** - Oyuncunun maç başına ortalama sayı üretimi.
8. **FG (0.27)** - İsabetli saha içi atış sayısı.
9. **2P (0.27)** - İsabetli iki sayılık atış sayısı.

Bu değişkenler, MVP ödülünü kazanmada en belirleyici faktörler olarak öne çıkmaktadır.



Şekil 3.1.1: Hedef Sütun ile Korelasyon

3.2 Veri Kümesi ve Önişleme

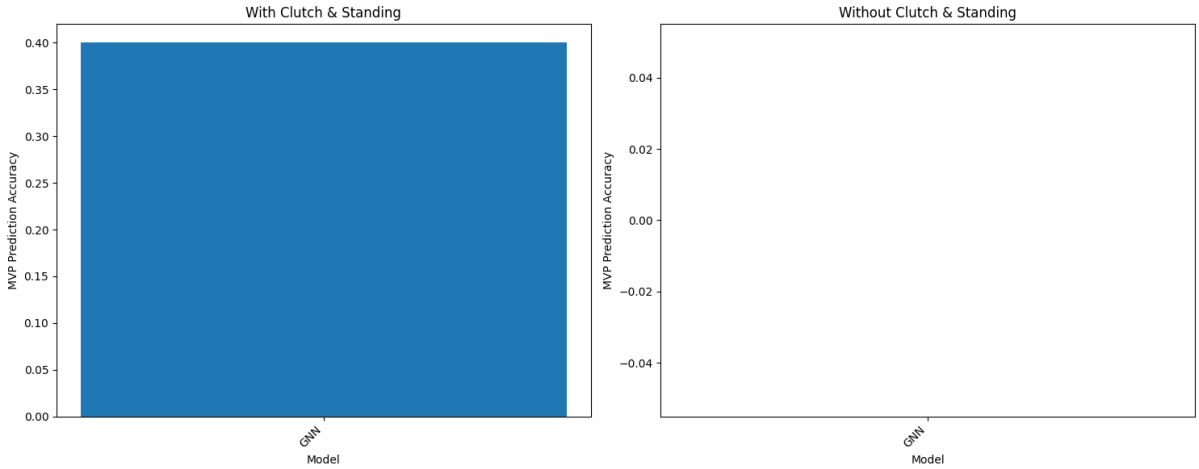
3.2.1 Veri Kaynağı

Çalışmada kullanılan veri seti, 12.000 satırdan oluşan NBA istatistiklerini içermektedir. Aşağıdaki temel değişkenleri kapsamaktadır:

- **Genel oyuncu bilgileri**: Sezon, oyuncu ismi, pozisyon, yaş, takım.
- **Temel istatistikler**: G, GS, MP, FG%, 3P%, ribaundlar, asist, top çalma, blok, top kaybı.
- **Gelişmiş istatistikler**: PER, BPM, VORP, WS, Clutch Puan, Takım Derecesi.

Literatür taramasında görüldüğü gibi bu alandaki çalışmalar istenen performans sonuçlarının elde edilmesinde yetersiz kalmaktadır. Saf performans verileri üzerine istatiksel yaklaşımlar ve makine öğrenmesine dayalı yaklaşımların performanslarının yetersiz kaldığı görülmektedir. Bunun yanı sıra NBA Reference.com üzerinden elde edilen veri kümesine ek olarak Clutch skoru, TeamStanding değeri gibi değerleri ekleyip veri kümesinin zenginleştirilmesi ve modelin eğitimindeki katkısının ölçülmesi planlanmaktadır. Clutch skoru, bir oyuncunun maçın kritk anlarında yaptığı başarıya itaf eden NBA tarafından belirlenen skordur. TeamStanding skoru ise NBA'in bir oyuncunun

takım içerisindeki katkısını belirten skordur.



Şekil 3.2.1.1 : GNN Modelinin Öğrenimi

Şekil 3'te GNN yapay sinir ağı modeli 100 epoch boyunda eğitildikten sonraki accuracy değerleri görülmektedir. Team Standing ve Clutch Score sütunları eklendikten ve eklenmeden önceki eğitim değerleri görülmektedir. Bu değerlere göre modelin öğrenimini kolaylaştırıldığı görülmektedir.

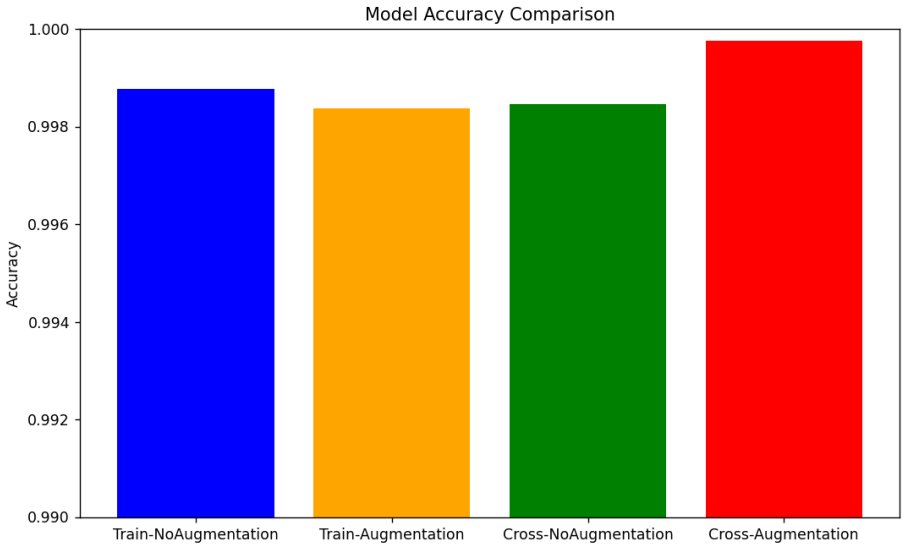
3.2.2 Veri Önişleme Teknikleri

Veri seti üzerinde farklı önişleme teknikleri uygulanmıştır:

- **Eksik verilerin tamamlanması:** SimpleImputer kullanılarak eksik veriler uygun istatistiksel değerlerle doldurulmuştur.
- **Aykırı değerlerin temizlenmesi:** IQR tekniği kullanılarak istatistiksel aykırılar belirlenmiş ve düzenlenmiştir.
- **Kategorik değişkenlerin kodlanması:** One-Hot Encoding ile kategorik veriler sayısal hale getirilmiştir.
- **Veri ölçeklendirme:** MinMaxScaler kullanılarak tüm sayısal değişkenler belirli bir aralığa normalize edilmiştir.
- **Augmentation :** Görsel verilerde kullanılsada sayısal verilerde resampling ile yeni eklenen veriler test edildi.

Literatürdeki çalışmalarda yoğunlukla istatiksel ve SMOTE dayalı sentetik veri üretimi kullanıldığı görülmektedir. Modelin güvenirlik seviyesini etkileyen bu faktörlerin kullanılmasının sonucu istenmeyen yönde etki edeceği için augmentation işlemi yapılmadan eğitim yapılması planlanmaktadır.

Veri seti içerisindeki dengesizlik durumunu çözmek için her özelliğin dağılım histogramları belirlendi ve pozitif sınıfların bu özelliklerinin değerleri incelenmektedir. Pozitif sınıftaki oyuncuların özelliklerinin diğer oyunculara kıyasla üst sınırı aştığı durumlar görülmektedir. Buradaki dengesizliği kaldırmak için sadece alt sınıf ve belirli özelliklere belirli trashold değerleri koyulması planlanmaktadır.



Şekil 3.2.2.1 : *Augmentation*

3.3 Kullanılan Modeller

Farklı makine öğrenmesi ve derin öğrenme modelleri kullanılarak veri seti üzerinde öğrenimler gerçekleştirilmektedir. Random Forest, karar ağaçlarını kullanarak veri setindeki örüntüleri öğrenen bir topluluk öğrenme modeli olup, değişkenler arasındaki etkileşimleri etkili bir şekilde yakalayabilmektedir. KNN, komşuluk ilişkilerine dayalı olarak çalışan parametrik olmayan bir sınıflandırma algoritmasıdır. SVM, verileri en iyi ayıran hiper düzlemi belirleyerek sınıflandırma yapan bir model olarak öne çıkmaktadır.

Bununla birlikte, gradyan artırımlı modeller arasında yer alan XGBoost, yüksek doğruluk oranı ile dikkat çeken ağaç tabanlı bir yöntemdir. LightGBM, büyük veri

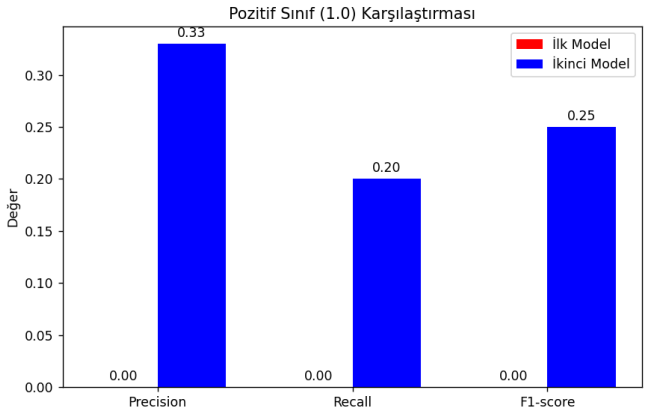
setlerinde daha hızlı eğitim sağlayan bir gradyan artırımlı karar ağaçları modeli olarak öne çıkmaktadır. CatBoost ise özellikle kategorik değişkenleri daha etkin kullanabilen bir gradyan artırımlı modeldir.

Derin öğrenme modelleri kapsamında, ANN çok katmanlı ileri beslemeli sinir ağları ile tahminleme yaparken, KNN genellikle görsel verilerde başarılı olsa da, veri setine uygulanarak desen tanıma yetenekleri test edilmiştir. Son olarak GNN, oyuncular arasındaki ilişkileri graf tabanlı modelleme ile analiz eden bir yaklaşım olarak değerlendirilmiştir. Bu farklı model yaklaşımları sayesinde veri setindeki örüntüler detaylı bir şekilde incelenmiş ve en iyi performans gösteren model belirlenmiştir.

3.4 Model Eğitimi ve Değerlendirme (Model Training and Evaluation)

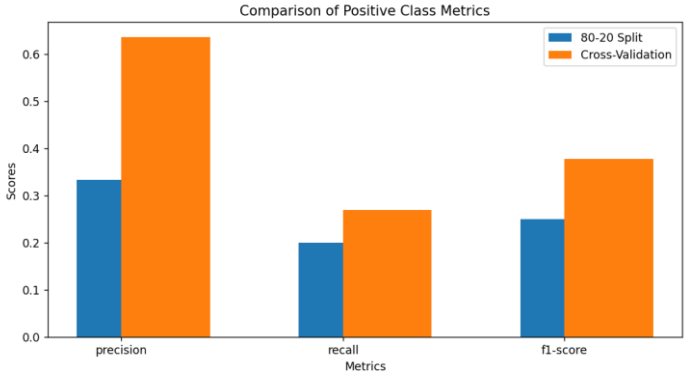
3.4.1 Veri Setinin Bölünmesi

- 1997-2017 verileri eğitim ve test seti olarak kullanılmaktadır.
- 2018-2022 verileri tahmin seti olarak kullanılmaktadır.
- Bölme oranı olarak Chen ve arkadaşlarının kullandığı 70% ve %30 bölme oranı ve literatürde yaygın kullanıldığı belirlenen 80% 20% bölme oranı test edilecektir. [5][15].



Şekil 3.4.1.1: RandomForest'ın Farklı Bölme Oranlarındaki Classification Raporu

- Cross-Validation yöntemlerinden biri olan SimpleKfold ile modellerin öğrenimi karşılaştırılmaktadır.



Şekil 3.4.1.2 : Catboost ML'nin TrainTestSplit ve CrossValidation a göre Eğitimlerindeki Sonuçları

3.4.2 Model Eğitim Teknikleri

Aşırı uyumu (overfitting) engellemek amacıyla Early Stopping yöntemi uygulanmıştır. Modelin eğitim süreci boyunca performans takibini sağlamak için verbose ayarı etkinleştirilmiştir. Ayrıca, farklı model eğitim stratejilerini değerlendirebilmek adına cross-validation ve train-test split teknikleri kullanılarak veri seti farklı şekillerde bölünmüş ve modelin genelleme kabiliyeti test edilmiştir.

3.4.3 Model Optimizasyonu

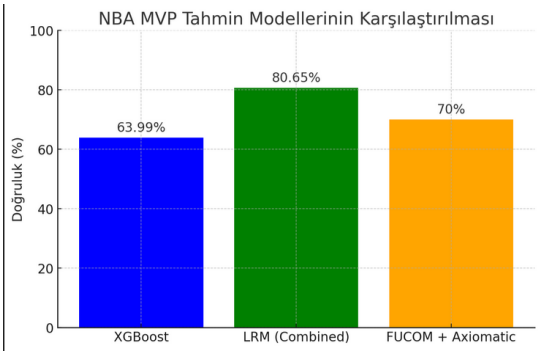
Modelin en iyi performansı göstermesi için GridSearch ve Parameter Grid fonksiyonları kullanılarak kapsamlı bir hiperparametre optimizasyonu gerçekleştirilmiştir. Farklı modeller için epoch, iteration, sample, depth, layer ve optimizasyon algoritmaları gibi çeşitli hiperparametreler üzerinde hyperparameter tuning yapılmıştır. Her modelin en iyi versiyonunu belirlemek ve genel performansını artırmak amacıyla, en uygun hiperparametreler belirlendikten sonra modellerin yeniden eğitilmesi planlanmıştır.

CatBoost modeli için iterations, depth ve learning rate, Random Forest algoritması için n_estimators, max_depth ve min_samples_split, KNN algoritması için n_neighbors ve weight, SVM algoritmasında C ve kernel, XGBoost modeli için n_estimators, learning_rate ve max_depth gibi kritik hiperparametreler optimize edilmiştir. Ayrıca, derin öğrenme modellerinde katman sayısı ve her katmandaki nöron sayısı gibi parametreler değerlendirilmiştir. Bu süreçte geniş bir parametre aralığı test edilerek, her model için en iyi yapılandırmanın belirlenmesi ve optimum başarıya ulaşılması hedeflenmiştir.

3.4.5 Model Performans Değerlendirmesi (Model Performance Evaluation)

Modellerin başarımlarını değerlendirmek için makine öğrenmesi algoritmalarının başarımlarını ölçen metrikler kullanılmıştır. Doğruluk kararlarının değerini ölçen precision, Kararlarının doğruluğunu ölçen recall ve ikisinin dengesini belirten F1 score metrikleri başarımı ölçmek için kullanılmıştır.

Literatürde bu probleme olan yaklaşımlardan elde edilen sonuçlar aşığıdaki gibidir.



Şekil 3.4.5.1 : Sonuç Karşılaştırması

Cheng tarafından yapılan çalışmada 0.63 oranlarında r2 elde edildiği görülmüştür [5]. Jordan tarafından yapılan araştırmada ise %80 oranında başarıya sahip bir modelin geliştirilmesi sağlandığı görülmektedir [7]. Fucom+Axiomatic modelinin kullandığında ise %70 oranında başarı elde edildiği görülmektedir. Bu sonuçlara ve modellerin tahmin ettiği adaylar arasındaki başarı durumuna göre geliştirilen modelin başarısı ölçülecektir.

Classification Report kullanılarak elde edilen metrikler bir modelin doğru veya yanlış olarak verdiği kararların aslındaki cevaplara oranını vermektedir. Bunun için sklearn kütüphanesindeki ClassificationReport fonksiyonu kullanmaktadır. Metriklerin yanı sıra modellerin tahmin sonuçları güvenlirliklerini belirlediği için tahmin için ayrılan 5 yıllık veri kümesindeki mvp olacak oyuncunun tahmini ve bahsedilen metriklerin değerlendirilmesi ile modellerin başarısı test edilecektir.

4. BULGULAR VE TARTIŞMA

Proje araştırmamızda NBA maçları içerisindeki MVP oyuncuyu bulma aşamasında veri seti işleme hazır hale getirilmiştir. Proje kapsamında kullanılacak olan modellerin belirlenmesi ve probleme en iyi çözümü sağlayacak hale getirilmesi ile ilgili çalışmalar devam etmektedir. Araştırma sonucunda ulaşılacak metrikler ve modelin değerlendirilmesi ile ilgili ölçütler belirlenmiş, probleme dayalı bağımlı ve bağımsız değişkenlerin ilişkisi açıklanmıştır. Bu ilişkiye bağlı kalarak modellerin eğitilmesi, en iyi hallerine getirilmesi için araştırmalara ve eğitimlerine devam edilmesi planlanmaktadır. Sonuçların çıkmasının ardından geriye dönük iyileştirmeler yapılması, elde edilen sonuçların ve model başarılarının raporlanması planlanmaktadır.

5. SONUÇLAR

KAYNAKLAR

[1] M. Chen, "Predict NBA Regular Season MVP Winner," in *Proc. Int. Conf. Ind. Eng. Oper. Manag.*, Bogota, Colombia, Oct. 2017, pp. 44–51.

[2] M. Chen and C. Chen, "Data mining computing of predicting NBA 2019–2020 regular season MVP winner," in *Proc. Int. Conf. Ind. Eng. Oper. Manag.*, Bogotá, Colombia, Oct. 2020, pp. 1–9.

[3] Y. Zhai and T. Xu, "Novel metric to predict NBA regular season MVP," in *Proc. 2024 IEEE 10th Int. Conf. High Perform. Smart Comput. (HPSC)*, Beijing, China, 2024, pp. 36–42.

[4] A. L. Chapman, "The application of machine learning to predict the NBA regular season MVP," M.S. thesis, Dept. of Data Science, Utica Univ., Utica, NY, USA, May 2023.

[5] Cheng, Z. (2024). *A Comparison of Machine Learning Algorithms for National Basketball Association (NBA) Most Valuable Player (MVP) Vote Share Prediction*. In *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, pp. 262-267. SCITEPRESS.

[6] J. Han and Z. Yu, "Random forest prediction of NBA regular season MVP winners

based on metrics optimization," *Inf. Knowl. Manag.*, vol. 4, no. 4, pp. 53–62, 2023.

[7] J. M. McCorey, "Forecasting most valuable players of the National Basketball Association," M.S. thesis, Dept. of Engineering Management, Univ. of North Carolina at Charlotte, Charlotte, NC, USA, 2021.

[8] V. Özkir and A. Değirmenci, "A novel multiple criteria ranking approach for determining the most valuable player (MVP) of a sport season: A numerical study from NBA league," *J. Soft Comput. Decis. Anal.*, vol. 1, no. 1, pp. 265–272, Oct. 2023.

[9] A. A. Albert, L. F. de Mingo López, K. Allbright, and N. Gómez Blas, "A hybrid machine learning model for predicting USA NBA All-Stars," *Electronics*, vol. 11, no. 1, p. 97, Dec. 2021.

[11] F. Thabtah, L. Zhang, and N. Abdelhamid, "NBA game result prediction using feature analysis and machine learning," *Ann. Data Sci.*, vol. 6, no. 1, pp. 103–116, Jan. 2019.

[12] Yongjun, L., Lizheng, W., & Feng, L.. A data-driven prediction approach for sports team performance and its application to National Basketball Association. Omega, 2021.

[13] Y. Chen, J. Dai, and C. Zhang, "A Neural Network Model of the NBA Most Valued Player Selection Prediction", In Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence (PRAI '19), Association for Computing Machinery, New York, NY, USA, pp. 16–20, 2019

[14] J. Mertz, L. D. Hoover, J. M. Burke, D. Bellar, M. L. Jones, B. Leitzelar, and W. L. Judge, "Ranking the greatest NBA players: A sport metrics analysis," *Int. J. Perform. Anal. Sport*, vol. 16, no. 3, pp. 737–759, 2016.