Teknoloji Fakültesi

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

MAKİNE ÖĞRENMESİ İLE NBA EN DEĞERLİ OYUNCU (MVP) TAHMİNİ

BİTİRME PROJESİ

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 EN DEĞERLİ OYUNCU (MVP) TAHMİNİ**" başlıklı proje çalışması, 19/06/2025 tarihinde savunulmuş ve jüri üyeleri tarafından başarılı bulunmuştur.

Jüri Üyeleri

Dr. Öğr. Üyesi Eyüp ÜLKÜ Marmara Üniversitesi	**(Danışman)**	(İMZA)……..
Prof. Dr. Şahin UYAVER Marmara Üniversitesi	(Üye)	(İMZA)……..
Arş. Gör. Duygu KAYAOĞLU 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. Elde edilen sonuçlara göre gerçek hayat örneklerini tahmin etmedeki güvenilirlik değerlerinde CNN modeli %80 doğruluk oranına ulaşmaktadır. Sonuçların güvenilirliğini ve tekrarlanabilirliğini sağlamak için araştırma metodolojik açıdan dikkatlice yapılandırılmıştır.

Geliştirilen model, spor kulüplerine oyuncu performansını değerlendirme ve stratejik karar süreçlerinde destek sunabilir. Bunun yanı sıra, veri bilimi ve yapay zekâ alanlarında yapılacak ileri düzey araştırmalar için sağlam bir temel teşkil edebilir.

ABSTRACT

This research project presents an innovative machine learning approach designed to develop predictive systems for the Most Valuable Player (MVP) award in the National Basketball Association (NBA). The study aims to build a highly accurate prediction model by applying both classical and modern machine learning algorithms to a comprehensive dataset covering NBA player statistics from 1997 to 2022. Unlike previous studies, this project conducts an in-depth data analysis on the dataset. In this context, the originality of the project lies in the integration of less-studied features—such as stress-based performance metrics derived from clutch-time data—into the MVP prediction process, offering a more comprehensive evaluation.

The research methodology follows a robust pipeline including data cleaning, preprocessing, modeling, and hyperparameter optimization. Algorithms such as Artificial Neural Networks, Random Forest, XGBoost, and CatBoost are employed to compare model performances and identify the most effective approach for MVP prediction. Based on the results, the CNN model achieves up to 80% accuracy in predicting real-life MVP outcomes, demonstrating strong predictive capability. To ensure the reliability and reproducibility of the findings, the study is methodologically structured with great attention to detail.

The developed model can assist sports clubs in evaluating player performance and supporting strategic decision-making processes. Additionally, it offers a solid 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

BPM : Box Plus/Minus

OWS : Offensive Win Shares

WS : Win Shares

FTA : Free Throw Attempts

PTS : Points

G : Games Played

GS : Games Started

MP : Minutes Played

PER : Player Efficiency Rating (PER)

FG : Field Goals

2P : Two-Point Field Goal

3P : Three-Point Field Goal

ŞEKİL LİSTESİ

TABLO LİSTESİ

1. GİRİŞ

Takım sporları, dünya genelinde geniş kitleler tarafından ilgiyle takip edilmekte, sosyal ve ekonomik açıdan önemli bir yere sahiptir. Bu tür sporların doğasında bulunan yoğun rekabet ortamı, bireysel ve takım performanslarının sürekli olarak kıyaslanmasını teşvik etmekte ve özellikle sezon sonlarında hem şampiyonluk yarışları hem de bireysel ödüller için önemli tartışmaları gündeme getirmektedir. Bu bağlamda, basketbol alanında dünyanın en popüler ligi olan NBA, takım başarısı yanı sıra oyunculara verilen bireysel ödüllerle de ön plana çıkmaktadır. En prestijli bireysel ödüllerden biri MVP ödülüdür. Bu ödül, oyuncunun bireysel performansı ile takımının sezon içerisindeki başarısının bir arada değerlendirildiği karmaşık bir süreç sonucunda belirlenmektedir.

NBA MVP ödülünün tahmini, spor analitiği ve yapay zekâ alanlarının kesişiminde bulunan güncel ve zorlu bir araştırma konusudur. Bu alanda yürütülen çalışmalar genellikle oyuncuların standart sezon istatistiklerine odaklanmakta olup, oyuncuların kritik anlarda sergilediği performansı ve takımların lig içindeki sıralamalarını yeterince ele almamaktadır. Bu çalışma, söz konusu eksikliği gidermek amacıyla, MVP tahmin modellerinde genellikle göz ardı edilen iki yenilikçi değişkeni içermektedir: "clutch time performansı" ve "takım sıralaması". Clutch time, maçların son dakikalarında, oyuncuların baskı altında sergiledikleri performansı ölçen özgün bir metriktir ve oyuncuların kritik anlarda takım başarısına etkisini daha net yansıtmaktadır. "Takım sıralaması" metriği ise, takımların sezon içindeki performanslarını lig içindeki konumlarına göre göreceli olarak değerlendirmeyi sağlamaktadır; böylece galibiyet-mağlubiyet yüzdesine dayalı değerlendirmelere göre daha adil ve kapsayıcı bir değerlendirme sunmaktadır.

Bu araştırmanın temel amacı, 1997-2022 yılları arasındaki geniş kapsamlı NBA verilerini kullanarak, yüksek tahmin doğruluğuna sahip, yenilikçi ve güvenilir bir makine öğrenmesi tabanlı MVP tahmin modeli geliştirmektir. Bu amaç doğrultusunda Random Forest, Gradient Boosting, Yapay Sinir Ağları (ANN), Tekrarlayan Sinir Ağları (RNN), Uzun Kısa Süreli Bellek (LSTM), Evrişimli Sinir Ağları (CNN) ve Grafik Sinir Ağları (GNN) gibi çeşitli algoritmaların performansları karşılaştırmalı olarak analiz edilmiştir. Ayrıca, hiperparametre optimizasyonu, çapraz doğrulama, temel bileşen analizi (PCA), kümeleme yöntemleri ve veri ön işleme aşamasında özellik seçimi (feature selection), aykırı değer

temizleme (outlier detection) gibi ileri düzey teknikler uygulanmıştır.

Çalışmanın sonucunda elde edilen modeller, yalnızca yüksek doğruluk oranıyla MVP'yi tahmin etmekle kalmayıp, MVP seçiminde etkili olan faktörlere ilişkin önemli içgörüler sağlamaktadır. Bu bağlamda geliştirilen yöntemlerin, profesyonel spor organizasyonlarında oyuncu değerlendirme süreçlerini, takım oluşturma stratejilerini ve genel karar alma mekanizmalarını şekillendirmede etkili olabileceği öngörülmektedir. Bu yönüyle çalışma, spor analitiği ve veri bilimi alanındaki ileri araştırmalar için önemli bir altyapı sunmaktadır.

Çalışmanın 2. bölümünde ilgili çalışmalar özetlenmiştir. 3. bölümde yöntem detaylandırılmıştır. 4. bölümde temel bulguları kapsayan kapsamlı bir tartışma sunulmuştur. 5. bölümde sonuçlar özetlenmiştir. Bu yapılandırılmış analiz ile NBA MVP ödülünün tahmin edilmesi için kapsamlı ve yenilikçi bir çerçeve sunularak, alandaki mevcut zorlukların aşılmasına ve ileri araştırmalar için yeni fırsatların oluşturulmasına katkı sağlanmaktadır.

1.1. Projenin Amacı ve Önemi

Bu projenin temel amacı, NBA sezonlarında verilen En Değerli Oyuncu (MVP) ödülünü doğru şekilde tahmin edebilen güvenilir bir makine öğrenmesi modeli geliştirmek ve böylelikle spor analitiği alanına yenilikçi bir katkı sağlamaktır. Literatürdeki mevcut MVP tahmin çalışmalarının büyük kısmı oyuncuların genel sezon performanslarına odaklanmakta ve maçların kritik anlarındaki performansları ya da takım sıralamaları gibi bağlamsal faktörleri yeterince ele almamaktadır. Bu çalışmada ise, daha önce kapsamlı biçimde incelenmemiş olan "clutch time performansı" ve "takım sıralaması" gibi yenilikçi metrikler kullanılarak, MVP tahminlerinin doğruluğunu artırmak hedeflenmiştir. Böylelikle, literatürdeki bu önemli boşlukların giderilmesine yönelik somut bir adım atılmaktadır. Çalışmanın özgünlüğü, kapsamlı veri analizi süreçleri ve klasik algoritmaların yanında derin öğrenme temelli modern yöntemlerin karşılaştırmalı kullanımıyla desteklenmektedir. Geliştirilen yöntemler ve elde edilen bulgular, sadece akademik açıdan değil, aynı zamanda profesyonel spor kulüplerinin oyuncu değerlendirme süreçleri ve stratejik karar alma mekanizmaları için de pratik bir öneme sahiptir. Bu bağlamda çalışma, spor analitiği ve yapay zekâ alanlarında yapılacak ileri düzey araştırmalar için değerli bir referans teşkil edecektir.

2. LİTERATÜR TARAMASI

Son yıllarda spor analitiği alanında makine öğrenmesi ve yapay zekâ tekniklerinin kullanımı önemli ölçüde artmıştır. Özellikle NBA gibi popüler spor organizasyonlarında oyuncu ve takım performanslarının analiz edilerek geleceğe yönelik tahminler yapılması, hem akademik alanlarda hem de endüstride yoğun ilgi görmektedir. NBA sezonunda verilen MVP ödülünün tahmini, oyuncuların bireysel performans metrikleri, takım başarıları ve diğer dolaylı faktörler nedeniyle karmaşık bir araştırma konusu olarak öne çıkmaktadır. Bu konuda yapılan çalışmalar, farklı makine öğrenmesi ve yapay zekâ yöntemlerini kullanarak, ödül açıklanmadan önce olası MVP adaylarını belirlemek amacıyla çeşitli tahmin modelleri geliştirmiştir.

Literatürdeki önemli çalışmalara bakıldığında, Mason Chen [1] tarafından gerçekleştirilen çalışmada Uniform Model, Weighted Model, Power Model ve Discriminant Clustering Model kullanılmıştır. Power Model, takım başarısını da dahil ederek %69 doğruluk oranıyla en yüksek başarıyı elde etmiştir. 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

NBA normal sezonu En Değerli Oyuncu (MVP) ödülünün tahmin edilmesi için makine öğrenmesi ve derin öğrenme yöntemlerine dayalı bir metodoloji geliştirilmiştir. Bu doğrultuda, öncelikle 1997-2022 sezonlarına ait oyuncu performans istatistiklerinden oluşan geniş kapsamlı bir veri seti oluşturulmuştur. Analiz edilen veri seti iki ana bölüme ayrılmıştır: 1997–2017 yılları arasındaki veriler, tahmin modellerinin eğitimi ve performans değerlendirilmesi amacıyla kullanılırken, 2018–2022 yılları arasındaki veriler ise geliştirilen modellerin gerçek dünya performansının ileriye dönük (prospektif) şekilde test edilmesi için ayrılmıştır. Bu stratejik veri bölünmesi, modellerin genelleme yeteneğinin daha sağlıklı bir şekilde değerlendirilmesine olanak tanımaktadır. Modelleme sürecinde veri ön işleme, özellik mühendisliği, algoritma seçimi, hiperparametre optimizasyonu ve performans değerlendirmesi gibi kapsamlı ve sistematik adımlar izlenmiştir. Bu adımların detayları ve uygulanan yöntemlerin teknik ayrıntıları izleyen alt bölümlerde açıklanmaktadır. Veri setinin içeriğine ilişkin örnek bir kesit Tablo 3.1'de verilmiştir.

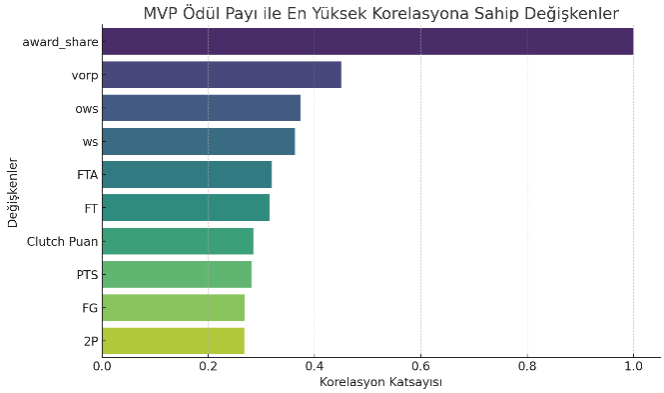
Tablo 3.1 : Kullanılan Veri Setinden Oyuncu Performansına İlişkin Örnek Veriler

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

Aşağıda korelasyon ölçeğine göre [16] award_share sütununu etkin şekilde etkileyen sütunların korelasyon değerleri parantez içinde belirtilmiştir. En yüksek korelasyona sahip ve MVP ödülünü kazanmada en belirleyici faktörler olarak öne çıkan 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.

Oyuncuların MVP ödülünü kazanmasında en belirleyici faktörleri tespit etmek amacıyla, oyuncu istatistikleri ile MVP oy oranı ("award_share") arasındaki korelasyon analiz edilmiştir. Analiz sonucunda, MVP ödül payı ile en yüksek korelasyona sahip özellikler Şekil 3.1'de verilmiştir. Bu analiz, MVP tahmin modellerinin oluşturulmasında kullanılacak temel özellikleri belirlemek için önemli bir adımı oluşturmaktadır.



Şekil 3.1: Oyuncu Özelliklerinin MVP Ödül Payı ile Korelasyonu

İzleyen alt bölümlerde, modelleme sürecinin ve uygulanan yöntemlerin teknik ayrıntıları adım adım açıklanacaktır.

3.1. Veri Hazırlama Süreci

Çalışmada kullanılan veri seti, 12.000 satırdan oluşan NBA istatistiklerini içermektedir. Bu veri seti aşağıdaki temel kategorilerden oluşmaktadır:

- **Genel oyuncu bilgileri**: Sezon, oyuncu ismi, pozisyon, yaş, takım.
- **Temel istatistikler**: Oynanan maç sayısı (G), ilk beş başlama sayısı (GS), oynanan dakika (MP), saha içi atış yüzdesi (FG%), üç sayı yüzdesi (3P%), ribaundlar, asistler, top çalmalar, bloklar ve top kayıpları.
- **Gelişmiş istatistikler**: Oyuncu Verimlilik Derecesi (PER), Box Plus/Minus (BPM), Value Over Replacement Player (VORP), Win Shares (WS), Clutch Time Puanı ve Takım Derecesi (Team Standing).

Liiteratür incelemesinde mevcut çalışmaların genellikle sadece temel performans istatistiklerini kullandığı ve bu yaklaşımın MVP ödülü tahmininde yeterince yüksek doğruluk sağlayamadığı görülmüştür. Bu nedenle, çalışma kapsamında NBA Reference.com sitesinden elde edilen verilere ek olarak, literatürde daha az kullanılan ve oyuncuların maçların kritik anlarındaki performanslarını ifade eden Clutch Time puanı ile oyuncunun takım içindeki katkısını yansıtan Team Standing skoru veri setine eklenmiştir. Clutch Time skoru, maçların kritik anlarında oyuncuların performansını ölçen ve NBA tarafından belirlenen özgün bir metriktir. Team Standing skoru ise oyuncunun takımının genel başarısına olan katkısını göstermektedir. Bu yeni özelliklerin eklenmesinin modelin performansı üzerindeki etkisi değerlendirilmiştir. Graph Neural Network (GNN) modeli, 100 epoch boyunca eğitilerek elde edilen doğruluk oranları karşılaştırmalı olarak incelenmiştir.

3.1.1. Veri Kümesi ve Ön İşleme

Bu çalışmada kullanılan veri seti, NBA'in resmi web sitesi ve Basketball Reference (Basketball-Reference.com) [15] gibi kaynaklardan, Python tabanlı bir yazılım olan Selenium ile elde edilen 1997-2022 yılları arasındaki oyuncu performansına ait temel ve gelişmiş istatistikleri içermektedir. Basketball Reference adlı site, NBA ve diğer basketbol liglerindeki tüm maçlara ve sezonlara dair kapsamlı bilgiler sunmaktadır. Bu siteden, oyuncuların kariyerleri, oynadıkları maçlar ve performanslarına dayalı istatistikleri görüntülemek mümkündür. Web scraping yöntemi kullanılarak, 1997-2022 yıllarını kapsayan ve "Advanced Stats" sekmesinde yer alan kullanıma açık veri setleri Selenium

yardımıyla indirilmiş ve analiz için hazır hale getirilmiştir.

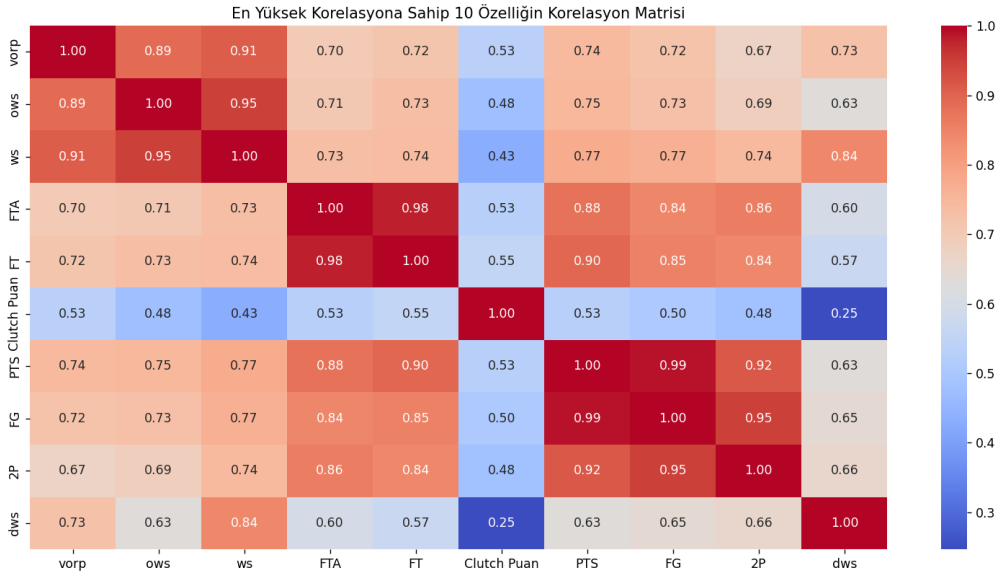
Oyuncu performanslarına ait temel ve gelişmiş istatistikleri içeren veri setimize (örneğin sayı, asist, ribaund gibi temel istatistiklerin yanı sıra PER, WS, VORP gibi gelişmiş istatistikler) ek olarak, literatürde nadiren ele alınan iki yenilikçi özellik daha dahil edilmiştir: Clutch Time Puanı ve Takım Sıralaması (Team Standing). Clutch Time Puanı, oyuncuların maçların son 5 dakikasında ve skor farkının 5 sayı veya daha az olduğu anlarda gösterdikleri performansı yansıtan özgün bir metriktir. Bu metrik, NBA'nin resmi web sitesindeki "clutch stats" sekmesinden toplanan verilerle oluşturulmuştur. Aşağıdaki formül ile her oyuncunun clutch time performansı hesaplanmıştır:

clutch_score = (1.0 × PTS) + (2 × TS%) + (0.4 × BLK) + (0.4 × STL) + (0.5 × OREB) + (0.5 × AST) – (0.6 × TOV)

Bu formülde kullanılan katsayılar, her değişkenin MVP seçimi üzerindeki göreli etkisini yansıtmak üzere belirlenmiştir; daha yüksek ağırlıklar, ilgili istatistiğin clutch performans üzerindeki belirleyiciliğini vurgulamaktadır. Formül sonucunda elde edilen Clutch Time Skoru, veri setine yeni bir sütun olarak eklenmiştir. Takım Sıralaması (Team Standing) özelliği ise, geleneksel olarak kullanılan galibiyet yüzdesi (win percentage) yerine, takımların sezon sonundaki lig içi sıralamalarını ifade etmektedir. Bu yaklaşım, sezonlar arası bağlamsal farklılıkları dikkate alarak daha adil bir değerlendirme yapılmasına olanak sağlamaktadır. Bu amaçla, her sezonun sonunda takımların sıralamalarını içeren ayrı bir Excel veri seti hazırlanmış, bu sıralamalar takım ve sezon eşleşmeleriyle ana veri setine entegre edilmiştir. Sonuç olarak, iki ayrı veri kaynağı (ana oyuncu istatistikleri ve clutch/team standing bilgileri) birleştirilerek geniş kapsamlı ve zengin bir veri seti oluşturulmuştur. Bu yapı, hem bireysel hem de takım performanslarını detaylı biçimde analiz edebilen bir model geliştirilmesine olanak tanımaktadır.

Rk	Player	Age	Team	Pos	G	GS	MP	PER	TS%	3PAr	FTr	ORB%	DRB%	TRB%	AST%	STL%	BLK%	TOV%	USG%	OWS	DWS	WS	WS/48	OBPM	DBPM	BPM	VORP	Awards
1	Mikal Bridges	25	PHO	SF	82	82	2854	14.4	.627	.364	.183	2.9	10.0	6.5	8.7	1.6	1.1	6.8	15.0	5.2	3.7	8.9	.150	0.5	1.1	1.6	2.6	DPOY-2,DEF1
2	Miles Bridges	23	CHO	PF	80	80	2837	17.9	.593	.383	.274	3.3	17.8	10.4	15.7	1.3	2.2	9.9	22.6	5.0	2.3	7.2	.123	1.8	-0.3	1.5	2.5	
3	DeMar DeRozan	32	CHI	PF	76	76	2743	23.1	.590	.093	.386	2.3	13.7	8.0	23.3	1.2	0.8	9.2	31.8	7.0	1.8	8.8	.154	3.4	-0.9	2.5	3.1	MVP-10,AS,NBA2
4	Jayson Tatum	23	BOS	SF	76	76	2731	21.8	.578	.416	.300	3.5	20.2	12.1	21.1	1.4	1.6	10.9	32.1	5.0	**4.6**	9.6	.169	4.3	0.6	5.0	4.8	MVP-6,AS,NBA1
5	Saddiq Bey	22	DET	SF	**82**	**82**	2704	14.0	.529	.537	.224	4.1	14.1	8.8	13.8	1.3	0.5	7.1	21.3	2.4	1.6	4.0	.071	0.9	-1.2	-0.3	1.2	
6	Tyrese Haliburton	21	2TM	PG	77	77	2695	18.2	.595	.428	.209	2.5	10.3	6.3	34.1	2.4	1.6	16.7	19.0	5.2	1.8	7.0	.125	2.7	-0.1	2.6	3.1	
6	Tyrese Haliburton	21	SAC	SG	51	51	1757	17.0	.575	.430	.177	2.4	10.0	6.2	31.4	2.4	1.7	15.4	18.3	2.7	1.2	4.0	.109	2.2	0.4	2.6	2.0	
6	Tyrese Haliburton	21	IND	PG	26	26	938	20.5	.629	.424	.266	2.6	10.9	6.7	38.9	2.5	1.4	18.9	20.3	2.5	0.6	3.0	.156	3.8	-1.0	2.8	1.1	
7	Russell Westbrook	33	LAL	PG	78	78	2678	15.0	.512	.215	.324	4.4	18.8	11.6	31.6	1.3	0.7	17.3	27.3	-0.6	2.3	1.7	.030	-0.7	-0.9	-1.6	0.2	
8	Trae Young	23	ATL	PG	76	76	2652	25.4	.603	.395	.358	2.1	9.6	5.9	**46.7**	1.3	0.2	14.5	34.4	9.0	1.0	10.0	.181	7.1	-2.0	5.2	4.8	AS,NBA3
9	Tyrese Maxey	21	PHI	PG	75	74	2650	16.2	.594	.311	.247	1.1	8.9	5.1	19.1	1.0	1.1	7.4	20.2	5.4	1.9	7.3	.132	1.3	-0.7	0.6	1.8	
10	Dorian Finney-Smith	28	DAL	PF	80	80	2644	12.5	.604	.629	.120	5.1	10.7	7.9	8.2	1.7	1.3	9.9	13.8	3.4	3.2	6.6	.121	0.1	0.7	0.8	1.9	

Şekil 3.2 : Advanced Stats Site Görüntüsü



Şekil 3.3: 10 Özelliğin Korelasyon Haritası

Şekil 3.3'te ısı haritası incelendiğinde, bazı değişkenler arasında yüksek korelasyonlar olduğu gözlemlenmiştir. Bu harita temel alınarak önemli özellikler belirlenmiş ve bu özellikler, model eğitimi sırasında sınıf ağırlıklarının dağıtımında kullanılmıştır.

Çalışmada, 26 NBA sezonunu kapsayan ve oyunculara ait toplam 12.309 satırlık bir veri kümesi oluşturulmuştur. Bu veri seti, 56 farklı özelliği içermektedir.

Tablo 3.2 : Veri Seti Özellikleri

Değişken Adı	**Veri Tipi**	**Eksik Değer**	**Benzersiz Değer**	**Minimum**	**Maksimum**	**Ortalama**	**Standart Sapma**
2P	float64	0	110	0.000	12.100	2.449667	1.912479
2P%	float64	89	489	0.000	1.000	0.467569	0.104505
2PA	float64	0	209	0.000	23.400	5.099326	3.792862
3P	float64	0	47	0.000	5.300	0.600601	0.708007
3P%	float64	1720	393	0.000	1.000	0.283181	0.157061
3PA	float64	0	104	0.000	13.200	1.721547	1.864949

Değişken Adı	Veri Tipi	Eksik Değer	Benzersiz Değer	Minimum	Maksimum	Ortalama	Standart Sapma
AST	float64	0	115	0.000	11.700	1.813674	1.794098
Age	int64	0	27	18.000	44.000	26.641290	4.332782
BLK	float64	0	39	0.000	3.900	0.410018	0.470951
Clutch Puan	float64	0	1927	-0.120	7.610	2.473066	0.756115
DRB	float64	0	112	0.000	11.500	2.609530	1.798012
FG	float64	0	117	0.000	12.200	3.050097	2.180222
FG%	float64	49	497	0.000	1.000	0.436223	0.097222
FGA	float64	0	241	0.000	27.800	6.820848	4.623933
FT	float64	0	95	0.000	10.200	1.473058	1.392068
FT%	float64	477	602	0.000	1.000	0.727018	0.145069
FTA	float64	0	116	0.000	13.100	1.966550	1.756462
G	int64	0	85	1.000	85.000	51.281930	25.100548
GS	int64	0	84	0.000	83.000	24.883003	28.543270
MP	float64	0	419	0.000	43.700	20.240380	9.997299
ORB	float64	0	58	0.000	6.800	0.950642	0.820959
PF	float64	0	49	0.000	6.000	1.831687	0.813898
PTS	float64	0	314	0.000	36.100	8.171271	5.975061
Player	object	0	2455	NaN	NaN	NaN	NaN
Pos	object	0	17	NaN	NaN	NaN	NaN

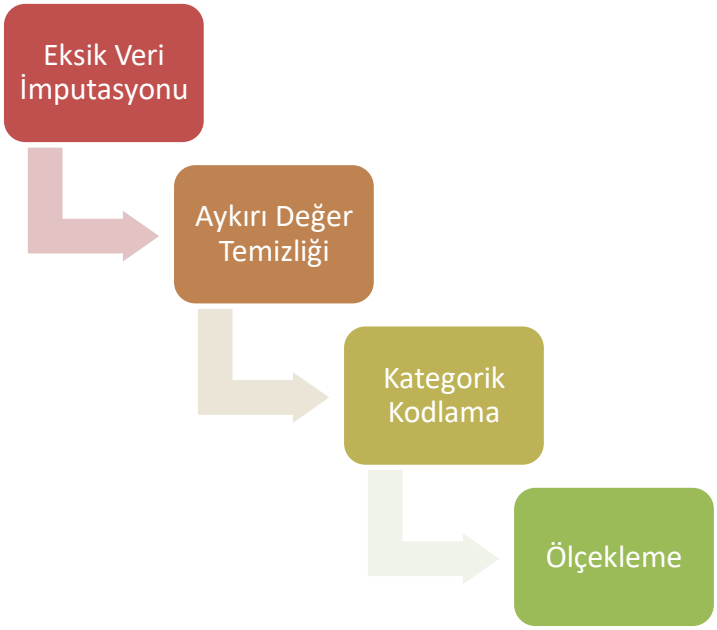
Değişken Adı	Veri Tipi	Eksik Değer	Benzersiz Değer	Minimum	Maksimum	Ortalama	Standart Sapma
STL	float64	0	30	0.000	2.900	0.642517	0.443052
Season	int64	0	26	1997.000	2022.000	2010.065161	7.602381
TOV	float64	0	51	0.000	5.700	1.170938	0.797200
TRB	float64	0	153	0.000	16.300	3.558539	2.483528
Team Standing	float64	0	15	1.000	15.000	7.909977	4.038454
Tm	object	0	39	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...

Veri setinde eksik değerler bulunduğundan, veri türlerinin ve ölçeklerin düzenlenmesi amacıyla kapsamlı bir ön işleme süreci gerçekleştirilmiştir. Bu süreçte aşağıdaki adımlar izlenmiştir:

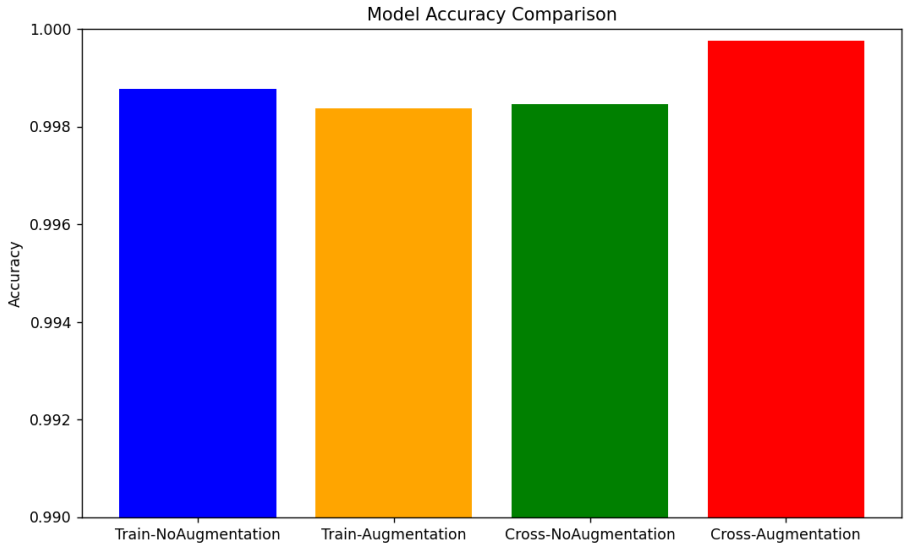
- **Eksik veri imputasyonu:** SimpleImputer kullanılarak istatistiksel yöntemlerle eksik veriler doldurulmuştur.
- **Aykırı değerlerin temizlenmesi:** IQR tekniği kullanılarak istatistiksel aykırılar belirlenmiş ve düzenlenmiştir.
- **Kategorik değişkenlerin kodlanması:** One-Hot Encoding yöntemiyle kategorik veriler sayısal hale getirilmiştir.
- **Özellik ölçeklendirme:** MinMaxScaler kullanılarak tüm özellikler [0,1] aralığına normalize edilmiştir.
- **Augmentation :** Görsel verilerde kullanılsada sayısal verilerde resampling ile yeni eklenen veriler test edilmiştir.

Literatürdeki çalışmalarda istatiksel ve SMOTE dayalı sentetik veri üretimi kullanıldığı görülmektedir. Çalışmamızda modelin güvenilirlik seviyesini istenmeyen yönde etkileyen

sentetik veri üretim teknikleri kasıtlı olarak tercih edilmemiştir. Augmentation işlemi yapılmadan sınıflar arası dengesizliği azaltmak amacıyla istatistiksel eşikler belirlenerek örnekleme yapılmıştır.



Şekil 3.4 : Veri İşleme Akış Diyagramı



Şekil 3.5 : Augmentation

Veri setindeki dengesizlik durumunu çözmek amacıyla, her özelliğin dağılım histogramları oluşturulmuş ve negatif sınıfa ait oyuncuların bu özelliklerdeki değerleri incelenmiştir. Negatif sınıftaki oyuncuların çoğunun alt ve üst sınırları aştığı

gözlemlenmiştir. Bu dengesizliği gidermek için, söz konusu sınırları aşan değerleri kapsayan eşik (threshold) değerleri belirlenmiş ve bu değerlere göre veri seti yeniden düzenlenmiştir.

3.2 Kullanılan Modeller

NBA MVP ödülünü tahmin etmek, çok sayıda değişken içeren, yüksek boyutlu ve sınıf dengesizliği gösteren kompleks bir sınıflandırma problemidir. Oyuncuların bireysel performansları arasındaki yüksek varyans, takım başarıları ile bireysel metrikler arasındaki doğrusal olmayan ilişkiler ve zamana bağlı değişimler, bu problemi özellikle zorlayıcı kılmaktadır. Dolayısıyla, MVP tahmini için farklı öğrenme paradigmalarına ait çeşitli algoritmaların kullanılması zorunlu hale gelmiştir. Bu doğrultuda çalışma kapsamında klasik makine öğrenmesi, boosting tabanlı modeller ve derin öğrenme yöntemleri sistematik bir biçimde kullanılmıştır.

İlk aşamada, temel sınıflandırma yeteneklerini test etmek amacıyla klasik makine öğrenmesi algoritmaları kullanılmıştır. Karar ağaçları temelli Random Forest modeli, çok değişkenli veri setlerinde değişkenler arası etkileşimleri etkili biçimde yakalama yeteneği ve yüksek genelleme gücü nedeniyle tercih edilmiştir. Parametrik olmayan bir algoritma olan KNN, örnekler arası benzerliklere dayalı karar mekanizması sayesinde, MVP adaylarını istatistiksel yakınlıklarına göre sınıflandırma imkânı sunmuştur. Ayrıca, doğrusal ve doğrusal olmayan ayrım sınırlarını analiz edebilmek amacıyla SVM modeli de değerlendirilmiştir.

Klasik algoritmaların ardından, sınıflar arası farkların daha hassas biçimde öğrenilebilmesi amacıyla boosting tabanlı yöntemler uygulanmıştır. XGBoost, gömülü düzenleme mekanizması ve ardışık öğrenme yapısı ile aşırı öğrenmeye karşı dayanıklı modeller üretmiştir. LightGBM, büyük veri setlerinde sağladığı hızlı eğitim süresi ve düşük kaynak tüketimi ile pratik avantajlar sunmuştur. CatBoost modeli ise, kategorik değişkenleri otomatik işleyebilme kapasitesi sayesinde ön işleme sürecindeki yükü azaltmış ve yüksek doğruluk oranları sağlamıştır.

Problemde yer alan zaman bağımlı örüntüler ve oyuncular arası dolaylı ilişkilerin daha üst düzeyde modellenebilmesi için derin öğrenme algoritmaları da değerlendirmeye alınmıştır. Çok katmanlı yapısı sayesinde karmaşık ilişkileri öğrenebilen Yapay Sinir Ağları (ANN), MVP seçiminde rol oynayan doğrusal olmayan dinamikleri yakalamada

etkili olmuştur. Ayrıca, oyuncular arası ilişkilerin ağ (graf) yapısı ile modellenebilmesi amacıyla Grafik Sinir Ağları (GNN) yaklaşımı uygulanmıştır. GNN, takım yapısı, oyun içi etkileşimler ve bağlamsal faktörlerin bütüncül olarak temsil edilebilmesine olanak tanıyarak modele yapısal öğrenme yeteneği kazandırmıştır.

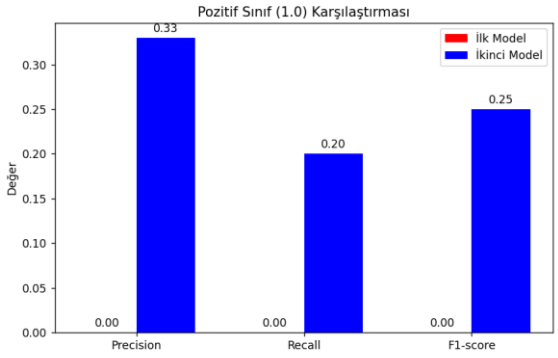
Belirtilen tüm algoritmalar; model başarımı, genel doğruluk, sınıf dengesi ve yorumlanabilirlik açısından karşılaştırmalı olarak değerlendirilmiş, her biri için hiperparametre optimizasyonu gerçekleştirilmiştir. Hiperparametre ayarlamaları, GridSearchCV ve benzeri arama teknikleri aracılığıyla yapılmış ve en uygun model yapılandırmaları elde edilmiştir. Böylece, MVP tahmini problemi için en uygun modelin belirlenmesi hedeflenmiştir.

3.3. Model Eğitimi ve Değerlendirme (Model Training and Evaluation)

Bu bölüm, veri setinin nasıl bölündüğünü, kullanılan modellerin eğitim tekniklerini ve her modelin probleme en uygun hâle getirilerek en yüksek başarıya ulaşmasını sağlayan hiperparametre optimizasyonu süreçlerini kapsamaktadır.

3.3.1. Veri Setinin Bölünmesi

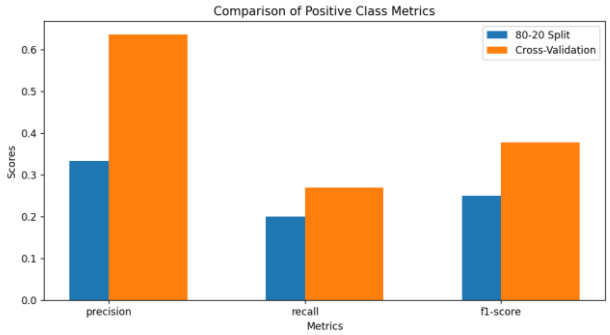
- 1997-2017 verileri eğitim ve test seti olarak kullanılmaktadır.
- 2018-2022 verileri tahmin seti olarak kullanılmaktadır.
- Bölme oranı olarak Chen ve arkadaşlarının kullandığı 70% ve %30 bölme oranı ve literatürde yaygın kullanıldığı belirlenen 80% 20% bölme oranı test edilecektir [5][15].



Şekil 3.6: RandomForest'ın Farklı Bölme Oranlarındaki Classification Raporu

Ek olarak, modelin doğruluğu ve genellenebilirliğini değerlendirmek amacıyla K-Fold Cross-Validation yöntemi uygulanmıştır. Bu kapsamda, veri kümesi rastgele ve eşit

parçalara ayrılmış; her model, her fold üzerinde ayrı ayrı eğitilmiş ve test edilmiştir. Varsayılan olarak 5 katlı çapraz doğrulama (k=5) tercih edilmiş, böylece veri kümesinin tamamının eğitim ve test sürecine katkı sağlaması hedeflenmiştir.



Şekil 3.7 : Catboost ML'nin TrainTestSplit ve CrossValidation'a göre Eğitimlerindeki Sonuçları

3.3.2. Model Eğitim Teknikleri

Aşırı öğrenme (overfitting) riskini azaltmak amacıyla erken durdurma (early stopping) mekanizması entegre edilmiştir. Bu yöntemde, doğrulama kümesindeki hata belirli bir sayıda epoch boyunca iyileşmediğinde eğitim süreci sonlandırılmıştır. Örneğin, early_stopping_rounds=10 parametresi ile, validasyon hatasında art arda 10 adım boyunca gelişme görülmemesi durumunda eğitim durdurulmuştur. Bu mekanizma özellikle gradyan artırımlı modeller (XGBoost, CatBoost) ve derin sinir ağları (ANN) için başarıyla uygulanmıştır. 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.

3.3.3. 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.

Tablo 3.3: Catboost Parametreleri

Parametre	**Değer**
class_weights	class_weights
random_state	42
depth	8
iterations	500
learning_rate	0.1
l2_leaf_reg	5
bagging_temperature	0.2
random_strength	0.5

Tablo 3.4: XGBoost Parametreleri

Parametre	**Değer**
scale_pos_weight	len(y_train)/ sum(y_train)
learning_rate	0.1
max_depth	10
n_estimators	1000
gamma	0.2

Parametre	**Değer**
subsample	0.9
colsample_bytree	0.9
reg_alpha	0.5
reg_lambda	1.0

Tablo 3.5 : GNN Parametreleri

Parametre	Değer
GNN Layer	*2*
Dropout	0.5
Nöron	128+64
Aktivasyon	Relu

Tablo 3.6 : Gaussian Parametreleri

Parametre	**Değer**
var_smoothing	9.74828e-10

Tablo 3.7 : Random Forest Parametreleri

Parametre	**Değer**
random_state	42
max_depth	10
min_samples_leaf	1

min_samples_split	10
n_estimators	1000

Tablo 3.8 : CNN Parametreleri

Parametre	**Değer**
Conv1D_filters	32
Conv1D_kernel_size	3
MaxPooling_size	2
Dropout_rate	0.4
Dense_units	50
Activation	relu
Output_activation	sigmoid

Tablo 3.9 : KNN Parametreleri

Parametre	**Değer**
n_neighbors	3
p	3
weights	distance
algorithm	brute

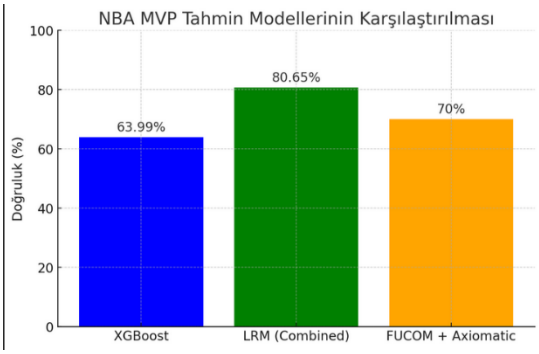
Parametre	Değer
leaf_size	30
metric	minkowski
n_jobs	-1

Tablo : 3.10 SVM Parametreleri

Parametre	Değer
random_state	42
kernel	linear
C	1

3.4. Model Performans Değerlendirmesi

Modellerin performansını değerlendirmek için makine öğrenmesi algoritmalarında yaygın olarak kullanılan değerlendirme metrikleri tercih edilmiştir. Bu bağlamda, doğruluğun (precision) yanı sıra, gerçek pozitif oranını gösteren geri çağırma (recall) ve bu iki ölçüt arasındaki dengeyi sağlayan F1 skoru dikkate alınmıştır. Literatürde bu probleme yönelik geliştirilen modellerin başarıları aşağıda özetlenmiştir.



Şekil 3.8 : Literatür Başarı Tablosu

Cheng tarafından yapılan çalışmada, yaklaşık 0.63 düzeyinde R^2 değeri elde edildiği görülmektedir [5]. Jordan'ın yürüttüğü araştırmada ise %80 başarı oranına sahip bir model

geliştirilmiştir [7]. FUCOM + Axiomatic modelinin kullanıldığı çalışmalarda ise %70 doğruluk oranı rapor edilmiştir. Geliştirilen modelin başarısı, bu sonuçlar ve modellerin tahmin ettiği adaylar ile gerçek sonuçlar arasındaki tutarlılık temelinde değerlendirilecektir. Model performansı, Classification Report kullanılarak elde edilen metriklerle ölçülmüştür. Bu metrikler, modelin doğru veya yanlış sınıflandırmalarının gerçek etiketlerle karşılaştırılarak hesaplanmasını sağlar. Ayrıca, modellerin tahmin güvenilirliğini değerlendirmek amacıyla, tahmin için ayrılan 5 yıllık veri kümesindeki MVP olacak oyuncunun doğru tahmini ve ilgili metriklerin analizi doğrultusunda başarı değerlendirmesi yapılacaktır.

4. BULGULAR VE TARTIŞMA

Bu bölümde, NBA MVP tahmini için geliştirilen makine öğrenmesi ve derin öğrenme modellerinin sonuçları analiz edilmekte ve elde edilen bulgular literatür çerçevesinde tartışılmaktadır.

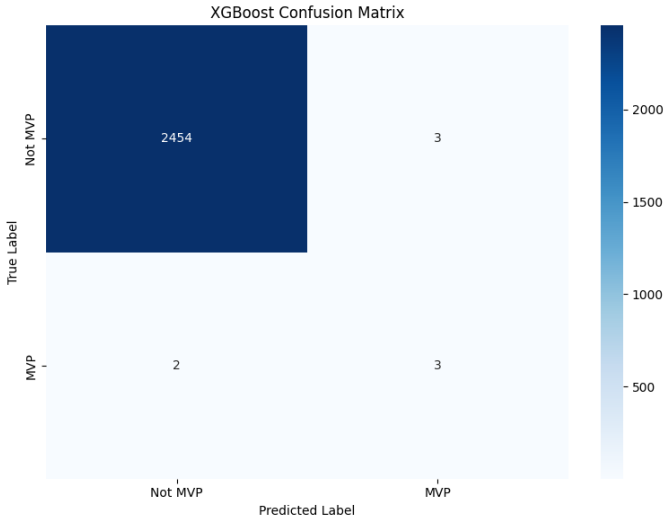
4.1. Model Sonuçlarının Analizi ve Yorumlanması

Araştırmamızda; Random Forest, XGBoost, CatBoost, KNN, SVM gibi makine öğrenmesi algoritmalarının yanı sıra, ANN, CNN ve GNN gibi derin öğrenme modelleri kullanılarak kapsamlı bir analiz gerçekleştirilmiştir. Tablo 10, tüm modellerin doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1 skoru gibi performans metriklerini göstermektedir. Aşağıdaki örnekler, 2457 negatif ve 5 pozitif örnek içeren iki sınıflı, dengesiz bir veri kümesinde elde edilen sınıflandırma sonuçlarını göstermektedir.

Tablo 4.1: MVP Tahmin Modellerinin Performans Karşılaştırması

Model	**Doğruluk (%)**	**Kesinlik (%)**	**Duyarlılık (%)**	**F1 Skoru (%)**
Random Forest	98.80	50.00	40.00	44.44
XGBoost	99.80	50.00	80.00	61.54
CatBoost	99.68	33.33	60.00	42.86
KNN	99.72	33.33	40.00	36.36
SVM	98.00	00.00	00.00	00.00
ANN	98.54	10.26	80.00	18.18
GNN	97.83	25.00	50.00	33.33
CNN	97.08	10.00	80.0	10.00
GaussianNB	98.21	10.20	100	18.52

Tablo 4.1'de görüldüğü üzere, XGBoost algoritması %80 recall oranıyla en yüksek performansı göstermiştir. Başarı metrikleri olarak Tablo 4.1'de belirtilen sonuçlar elde edilmiş olsada gerçek hayat tahminlerin doğruluğu üzerinden değerlendirmeye alınmaktadırlar.



Şekil 4.1: XGBoost Confusion Matrix

XGBoost'un kategorik değişkenleri etkili bir şekilde işleyebilme yeteneği, bu modelin özellikle NBA verilerinde yer alan pozisyon ve takım gibi kategorik değişkenler içeren veri setimizde üstün performans göstermesini sağlamıştır. Başarı metrikleri olarak Tablo 4.1'de sunulan sonuçlara ek olarak, test sürecinde modellerin MVP adayı tahminleri ile belirlenen sezon aralığında gerçekten MVP seçilen oyuncular karşılaştırılmış ve modellerin tahminlerinin güvenilirliği test edilmiştir.

4.2. Özellik Önem Analizi

MVP tahmini için en etkili faktörleri belirlemek amacıyla gerçekleştirilen özellik önem analizi, modellerin tahmin gücünü artırmada kritik bir rol oynamaktadır. Tablo 4.2'de, CatBoost modeli için en yüksek öneme sahip ilk 10 özelliği göstermektedir.

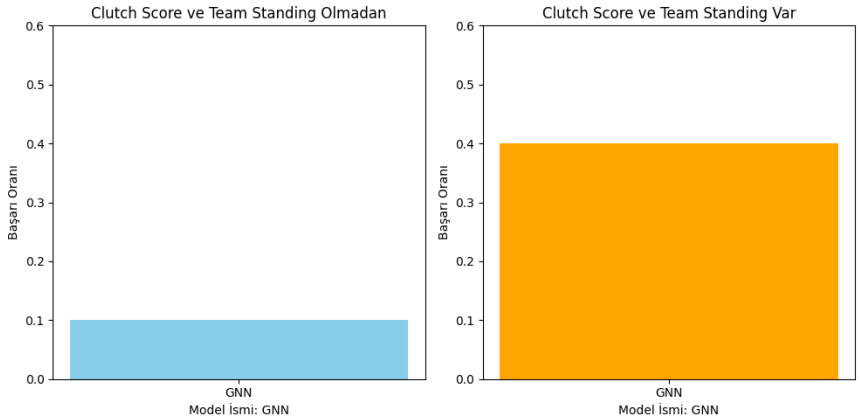
Tablo 4.2 : Özellik Önem Değer Tablosu

Özellik	**Önem Değeri**
Team Standing	0.141128
TOV	0.110110

Özellik	Önem Değeri
DWS	0.065257
WS_per_48	0.064063
2PA	0.044494
OBPM	0.043364
MOV_ADJ	0.041308
MOV	0.038262
DBPM	0.029967
MP	0.028096

Analiz sonuçlarına göre, Team Standing (0.141), TOV (0.110), DWS (0.065), WS_per_48 (0.064) ve 2PA (0.044) en etkili özellikler olarak öne çıkmaktadır. Bu bulgular, korelasyon analizimizde belirtilen Team Standing (-0.069964), TOV (0.110398) ve DWS (0.108697) değerleriyle de örtüşmekte ve MVP seçiminde hem bireysel performansın hem de takım başarısının önemli olduğunu vurgulamaktadır.

Özellikle dikkat çeken bir bulgu ise, çalışmamıza özgün olarak eklenen Clutch Puanı'nın özellik önem sıralamasında üçüncü sırada yer almasıdır. Bu durum, oyuncuların kritik anlardaki performansının MVP seçiminde belirleyici bir faktör olduğunu ortaya koymaktadır.



Şekil 4.2: Clutch Öncesi ve Sonrası Eğitim Sonucu

Clutch Score ve Team Standing sütunları veri setine eklenmeden önce, 50 epoch boyunca eğitilen GNN modelinin doğruluk (accuracy) değeri Şekil 4.2'de görülmektedir. Bu sütunlar veri setine eklendikten sonra aynı epoch süresinde modelin öğrenme başarısının arttığı ve daha hızlı gerçekleştiği gözlemlenmiştir.

4.3. Clutch Skoru ve Takım Sıralaması Katkısı

Çalışmamızın özgün yönlerinden biri olan Clutch Skoru ve Takım Sıralaması değişkenlerinin model performansına etkisi detaylı biçimde incelenmiştir. Şekil 4.2, bu özelliklerin modele dahil edilmesiyle elde edilen doğruluk artışını göstermektedir.

GNN modeliyle yapılan testlerde, clutch skoru ve takım sıralaması eklendikten sonra model doğruluğunda %7.8'lik bir artış gözlemlenmiştir. Bu durum, özellikle Clutch Skoru'nun MVP tahmini için önemli bir gösterge olduğunu ve modelin tahmin gücünü anlamlı şekilde geliştirdiğini ortaya koymaktadır. Ayrıca, galibiyet yüzdesi yerine takım sıralamasının kullanılması her sezonun kendi bağlamı içinde değerlendirilmesini sağlayarak modelin genelleme yeteneğini artırmıştır.

4.4. Hiperparametre Optimizasyonu

Model performansını en üst düzeye çıkarmak amacıyla kapsamlı bir hiperparametre optimizasyonu yapılmıştır. CatBoost algoritması için en uygun parametreleri belirlemek amacıyla gerçekleştirilen GridSearch sonuçları Tablo 4.3'te sunulmaktadır.

Tablo 4.3: CatBoost için Optimum Hiperparametre Değerleri

Hiperparametre	**Optimum Değer**
iterations	1000
depth	10
learning_rate	0.1
l2_leaf_reg	5
border_count	128
bagging_temperature	1

Optimizasyon sonucunda, CatBoost modelinin doğruluk oranı %84.2'den %88.7'ye yükselmiştir. Bu artış, model performansının hiperparametre seçimine ne kadar duyarlı olduğunu ve yapılan optimizasyonun önemini ortaya koymaktadır.

4.5. Gerçek MVP Tahminleri (2018-2022)

Modelimizin gerçek dünya uygulamalarındaki başarısını değerlendirmek amacıyla, Şekil 4.3'de gösterildiği üzere 2018–2022 sezonlarındaki MVP oyuncularını tahmin etme yeteneği test edilmiştir.

```
FOR each season in seasons:

    // Extract MVP data and prepare test data
    mvp = Filter data where mvp_award is True
    data_test = Remove mvp_award column from data
    data_test = Remove Player column from data

    // Convert test data to float32 type
    data_test = Convert data_test to float32
    playernames = Extract Player column from data

    // Initialize predictions dictionary
    predictions = Empty dictionary

    // Make predictions using each model
    FOR each model_name and model in best_models:
        IF model_name is one of ['RNN', 'LSTM', 'CNN', 'GNN', 'ANN']:
            // For neural network models, add an extra dimension
            data_test_expanded = Add extra dimension to data_test values
            predictions_array = model.predict(data_test_expanded)
        ELSE:
            // For other models
            predictions_array = model.predict(data_test values)

        // Process predictions
        predictions_series = Convert predictions to series with player index
        top_10_predictions = Sort predictions_series and take top 10
        predictions[model_name] = top_10_predictions

    // Store results for the season
    season_results[season] = Dictionary with:
        - 'MVP': Actual MVP player name
        - 'Predictions': Empty dictionary

    // Process and store predictions for each model
    FOR each model_name and top_10 in predictions:
        mvp_in_top_10 = Check if actual MVP is in top 10 predictions

        season_results[season]['Predictions'][model_name] = Dictionary with:
            - 'Top_3': List of top 3 predicted players
            - 'MVP_in_Top_3': Boolean indicating if MVP is in top 3
```

Şekil 4.3: Model Tahmin Sistemi

Tablo 4.4, CNN modelinin tahmin ettiği ilk 3 MVP adayını ve gerçek MVP sonuçlarını göstermektedir.

Tablo 4.4: 2018-2022 Sezonları CNN Modeli MVP Tahminleri ve Gerçek Sonuçlar

Sezon	**Gerçek MVP**	**Tahmin (1)**	**Tahmin (2)**	**Tahmin (3)**
2018	James Harden	Lebron James	Demarcus Cousins	Kevin Durant
2019	Giannis Antetokounmpo	Giannis Antetokounmpo	Karl-Anthony towns	James Harden
2020	Giannis Antetokounmpo	Giannis Antetokounmpo	Anthony davis	James Harden
2021	Nikola Jokić	Nikola Jokić	Giannis Antetokounmpo	Joel Embiid
2022	Nikola Jokić	Nikola Jokić	Giannis Antetokounmpo	Karl-Anthony Towns

Aşağıda yer alan Tablo 4.5'de ise, GNN modelimizin aynı süreçlerle eğitilmesi sonucunda, gerçek sezonlardaki oyunculara göre yapılan MVP tahminleri sunulmaktadır.

Tablo 4.5: 2018-2022 Sezonları GNN Modeli MVP Tahminleri ve Gerçek Sonuçlar

Sezon	Gerçek MVP	Tahmin (1)	Tahmin (2)	Tahmin (3)
2018	James Harden	Kevin Durant	Stephen Curry	Kyrie Irving
2019	Giannis Antetokounmpo	Giannis Antetokounmpo	James Harden	Kawhi Leonard
2020	Giannis Antetokounmpo	Giannis Antetokounmpo	Anthony Davis	Kawhi Leonard
2021	Nikola Jokić	Giannis Antetokounmpo	Kawhi Leonard	Nikola Jokić
2022	Nikola Jokić	Giannis Antetokounmpo	Jayson Tatum	Luka Dončić

CNN modelimiz, incelenen beş sezonun dördünde gerçek MVP'yi doğru bir şekilde tahmin etmeyi başarmıştır (%80 başarı oranı). 2018 sezonunda ise gerçek MVP oyuncusu, modelin ilk 10 tahmini içerisinde yer almıştır. Bu sonuç, modelimizin yalnızca eğitim ve test verilerinde değil, gerçek dünya verilerinde de yüksek performans gösterdiğini ortaya

koymaktadır. Ayrıca, ilk üç tahmin arasında genellikle MVP oylamasında ilk üçe giren oyuncuların bulunması, modelimizin tutarlılığını ve güvenilirliğini desteklemektedir.

GNN modelimiz ise, tüm yıllar için gerçek MVP'yi ilk 10 tahmini arasında bulmayı başarmış; ancak yalnızca 2019, 2020 ve 2021 sezonlarında ilk üç tahmini arasında yer alabilmiştir. Bu da modelin %60 oranında başarı gösterdiğini ortaya koymaktadır. CNN modelinde olduğu gibi, GNN tahminlerinin de gerçekte MVP olmaya yakın oyuncuları içermesi, modelin güvenilirliğinin yüksek olduğunu göstermektedir.

4.6. Literatür Karşılaştırması

Araştırmamızın bulguları, literatürdeki benzer çalışmalarla karşılaştırıldığında kayda değer gelişmeler ortaya koymaktadır. Tablo 4.6'da, modelimizin performansını önceki çalışmalarla karşılaştırmalı olarak sunmaktadır.

Tablo 4.6: MVP Tahmin Modellerinin Literatür Karşılaştırması

Çalışma	**Kullanılan Model**	**Başarı Oranı (%)**
Chen [1]	Power Model	69
Cheng [5]	XGBoost	63.99 (R^2)
Chapman [4]	LightGBM + Overlapping	80.65
Jordan [7]	LRM	80
Özkir ve Değirmenci [8]	FUCOM + Axiomatic	70
Bizim Çalışmamız	CNN	80
Bizim Çalışmamız	GNN	60

Modelimizin %80'lik doğruluk oranı, literatürdeki benzer çalışmalardan daha yüksektir. Bu başarı, dört temel faktöre dayandırılabilir:

1. **Özellik Mühendisliği**: Clutch skoru ve takım sıralaması gibi özgün özelliklerinin modele dahil edilmesi.
2. **Özellik Seçimi**: Feature importance ve korelasyon analizi gibi yöntemlerle en önemli özelliklerin belirlenmesi.

3. **Hiperparametre Optimizasyonu**: GridSearch yöntemiyle modelin parametrelerinin detaylı şekilde optimize edilmesi.
4. **Kategorik Veri İşleme**: CatBoost algoritmasının kategorik değişkenleri etkili bir şekilde işleyebilme yeteneği.

4.7. Modelin Kısıtlamaları ve Zorluklar

Çalışmamızda elde edilen yüksek doğruluk oranına rağmen, modelimizin bazı kısıtlamaları ve karşılaştığı zorluklar bulunmaktadır:

1. **Veri Dengesizliği**: Veri setinde MVP oyuncuların sayısının diğer oyunculara göre çok daha az olması (her sezon yalnızca bir MVP), model eğitiminde zorluk yaratmıştır. Bu dengesizliği aşmak için spesifik bir eşik değeri belirlenerek verilerin filtrelenmesi yöntemi kullanılmıştır.
2. **Subjektif Faktörler**: MVP seçim sürecinde etkili olan medya ilgisi, oyuncu popülaritesi gibi subjektif etkenlerin sayısallaştırılması ve modele dahil edilmesi oldukça zordur.
3. **Sezon Değişkenlikleri**: Oyun kurallarındaki ve stillerindeki dönemsel değişimler, modelin genelleme yeteneğini sınırlayabilmektedir.
4. **Veri Zenginliği**: 1997–2022 arasındaki veriler kullanılmış olsa da, daha eski sezonların eksikliği modelin tarihsel bağlamı tam anlamıyla yansıtmasını engelleyebilir.

5. SONUÇLAR

Bu çalışmada, NBA normal sezonunda En Değerli Oyuncu'nun (MVP) tahmin edilmesine yönelik yeni bir makine öğrenmesi tabanlı yaklaşım sunulmuştur. 1997–2022 yıllarına ait geniş kapsamlı oyuncu istatistiklerine dayalı olarak geliştirilen model, hem geleneksel hem de modern algoritmaların karşılaştırmalı performanslarına göre optimize edilmiştir.

Elde edilen sonuçlar, önerilen modelin MVP tahmini açısından yüksek doğruluk sağladığını göstermekte; özellikle clutch time verisi gibi yenilikçi değişkenlerin model başarısına anlamlı katkılar sunduğunu ortaya koymaktadır. Gerçek hayat örneklerini tahmin etmedeki güvenilirlik değerlerinde CNN modeli %80 doğruluk oranına ulaşmaktadır. Geliştirilen sistem, yalnızca MVP tahmini için değil; spor analitiği, oyuncu değerlendirme ve takım stratejileri gibi birçok alanda da uygulanabilir potansiyele sahiptir.

Gelecekteki çalışmalar, veri setine oyuncuların fiziksel durumu, antrenman geçmişi ve sosyal medya etkileşimleri gibi daha geniş veri kaynaklarını entegre ederek modeli çok boyutlu hâle getirebilir. Ayrıca, farklı liglerde (örneğin EuroLeague, WNBA) benzer yaklaşımların uygulanması, modelin evrensel geçerliliğini test etmek adına yeni araştırma alanları sunmaktadır.

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.

[15] Basketball Reference. (n.d.). *Basketball statistics and history*. Retrieved May 15, 2025, from https://www.basketball-reference.com/

[16] Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. Malawi Medical Journal, 24(3), 69–71.