Machine Learning ile NBA 2025 MVP Tahmini

Yağmur Kalyoncu   
*Bilgisayar Mühendisliği*  
 *Ankara Üniversitesi* Ankara, Türkiye  
kalyoncuyagmur67@gmail.com

Halime Şahnur Çöğür  
 *Bilgisayar Mühendisliği*  
 *Ankara Üniversitesi* Ankara, Türkiye  
 sahnurcogur98@gmail.com

*Özet*— Bu projede NBA oyuncularının sezonluk performans istatistiklerine dayanarak en değerli oyuncu (MVP) ödülünü kimin kazanacağını tahmin eden bir makine öğrenimi modeli geliştirilmiştir. Bu amaçla Random Forest, XGBoost ve Logistic Regression kullanılarak model eğitilmiş ve modellerin doğruluk oranı karşılaştırılmıştır.

Keywords— makine öğrenmesi, MVP tahmini, Random Forest Classifier, XGBoost, Logistic Regression.

# GİRİŞ

Basketbol dünyasında en öne çıkan ödüllerden biri olan NBA MVP ödülü öznel unsurlar taşımasına rağmen veri yapısı açısından çok zengin bir yapıya sahiptir. Bu nedenle makine öğrenmesi uygulamaları için uygundur. Bu proje 2004-2024 NBA sezonları arasındaki oyuncu istatistiklerini kullanarak 2025 MVP ödülünü tahmin etmeye çalışmaktadır. Basketbol referans.com veri kaynaklarından elde edilen verilerle Random Forest Classifier, XGBoost, Logistic Regression gibi algoritmaların tahmin başarımı karşılaştırılmıştır. Ayrıca her sezonun gerçek en değerli oyuncusunun bilindiği bir veri setiyle eğitilen modeller oyuncu-performans metrikleriyle MVP seçimi arasındaki ilişkileri de yakalamayı hedeflemektedir.

# YÖNTEM

## Veri Toplama

Projede kullanılan oyuncu istatistikleri [Basketball-Reference](https://www.basketball-reference.com/) web sitesinden 2004-20024 yıllarını kapsayacak şekilde web scraping yöntemiyle toplanmıştır. Her oyuncu için ; oyuncu adı, pozisyon, takım, maç sayısı (G), dakika (MP), sayı (PTS), asist (AST), ribaund (TRB), şut yüzdesi (FG%), yaş verileri elde edilmiş ve her sezonun MVP’si etiketlenerek veri kümesine “MVP” sütunu 1/0 olarak eklenmiştir.

## Veri Ön İşleme

İlk önce CSV formatındaki veri dosyaları pandas kütüphanesi kullanılarak veri çerçevesine (DataFrame) yüklenmiştir. Modelin eğitilmesinde kullanılacak olan **'PTS', 'AST', 'TRB', 'FG%', 'MP'** sütunları veri setinden çekilmiş ve hedef değişken olarak yalnızca **'MVP'** etiketi korunmuştur.

Eksik veri problemiyle başa çıkmak için **SimpleImputer** sınıfı kullanılmış ve tüm eksik değerler ilgili sütunların ortalama değeri ile doldurulmuştur. Bu yöntem verideki boşlukların modele olumsuz etkisini azaltarak daha istikrarlı bir öğrenme süreci sağlar.

Sonraki adımda tüm özellikler (feature'lar) **StandardScaler** kullanılarak normalize edilmiştir. Bu işlem her sütunun ortalamasını 0, standart sapmasını 1 yaparak verileri normalize eder. Bu işlem, farklı ölçekteki değişkenlerin aynı ölçekte temsil edilmesini sağlar ve modellerin daha verimli çalışmasına yardımcı olur.

## Modelleme

MVP tahmini için üç farklı model kullanıldı:

* **Random Forest**: Karar ağaçları tabanlı bu topluluk yöntemi, değişkenler arası ilişkileri iyi modelleyebilmesi nedeniyle tercih edildi.
* **XGBoost**: Özellikle sınıflandırma problemlerinde yüksek doğruluk sağlayan bu boosting algoritması, model karşılaştırmasında önemli bir yer tuttu.
* **Lojistik Regresyon**: Temel bir doğrusal sınıflandırma yöntemi olarak kullanıldı.

Performans değerlendirmesi için öncelikli metrik olarak **precision** kullanılmıştır. Precision metriğinin tercih edilmesinin nedeni, MVP tahmininde yanlış pozitiflerin (yani MVP olmayan bir oyuncunun MVP olarak tahmin edilmesinin) minimize edilmesinin önem taşımasıdır. Bu sayede model daha güvenilir ve tutarlı MVP adayları ortaya koyabilmiştir.

## 2025 MVP Tahmini

Modelleme sürecinde elde edilen en başarılı model olan Random Forest, 2025 sezonu için MVP adaylarını tahmin etmek amacıyla kullanılmıştır. Öncelikle, 2005-2024 yıllarına ait oyuncu istatistikleri ile model yeniden eğitilmiş, ardından 2025 sezonuna ait oyuncu verileri üzerinde tahminler gerçekleştirilmiştir. Model, her oyuncunun MVP olma olasılığını hesaplayarak, en yüksek olasılığa sahip ilk beş oyuncuyu belirlemiştir. Bu yaklaşım, oyuncuların sadece geçmiş performanslarına değil, aynı zamanda sezondaki güncel istatistiklerine de dayanarak tahmin yapılmasını sağlamıştır. Modelin verdiği olasılık skorları, oyuncuların MVP olma potansiyelleri hakkında nicel bir fikir sunmaktadır. Bu sayede, hem favori adaylar hem de sürpriz potansiyeli taşıyan isimler daha objektif şekilde ortaya konabilmiştir.

# SONUÇLAR

Fig. 1’de de görüldüğü gibi **Random Forest** ve **XGBoost** modelleri yüksek precision skorları ile MVP tahmininde oldukça başarılıdır. Özellikle Random Forest modeli diğer modellere göre daha iyi performans sergilemiştir. Bu, karar ağaçlarının karmaşık veri yapıları ve değişkenler arası ilişkileri etkili biçimde modelleyebilmesi ile açıklanabilir. Lojistik Regresyon modelinin performansının daha düşük kalması MVP tahmininin doğrusal olmayan karmaşık özelliklere dayalı olmasından kaynaklanabilir. Bu model sadece lineer ilişkileri yakalayabildiği için, spor performans verilerindeki etkileşimleri yeterince yansıtamamıştır.

## Model Karşılaştırması

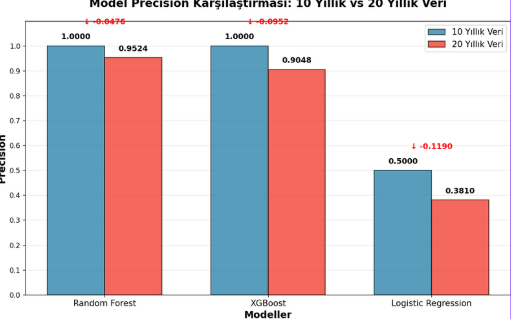
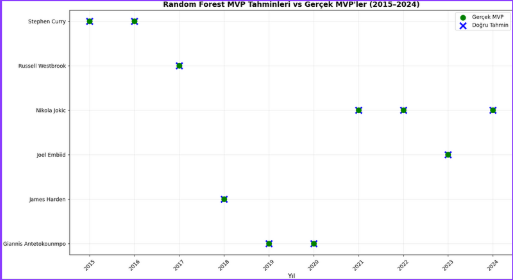


Fig. 1



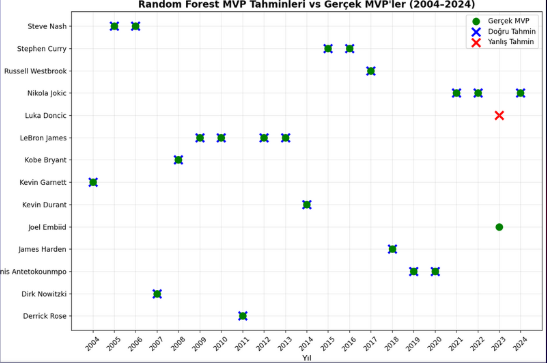
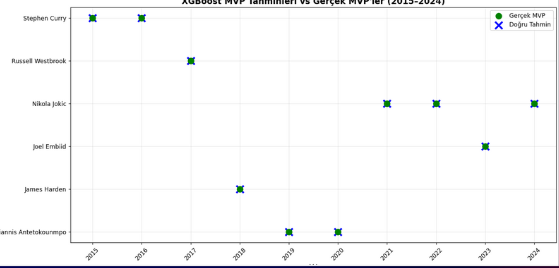


Fig. 2.

* (Fig. 2.)Bu modelde 10 yıllık ve 20 yıllık doğruluk oranları karşılaştırıldığında doğruluğun düştüğünü görüyoruz. Bu durumu "makine öğrenmesinde daha fazla verinin her zaman daha iyi sonuç vermeyeceği" gerçeğiyle açıklayabiliriz.



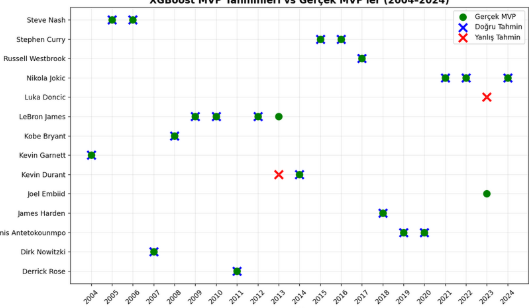
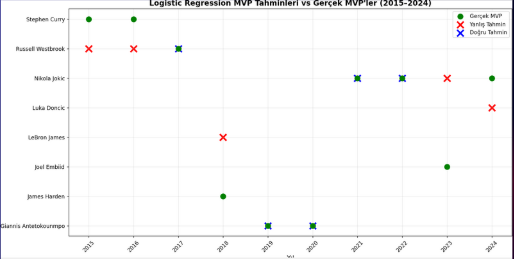


Fig. 3.

* (Fig. 3.)10 yıllık ve 20 yıllık doğruluk tahmini sonuçlarını karşılaştırdığımızda yine doğruluk oranının düştüğünü görüyoruz. Bu durumu "makine öğrenmesinde daha fazla verinin her zaman daha iyi sonuç vermeyeceği" gerçeğiyle açıklayabiliriz.



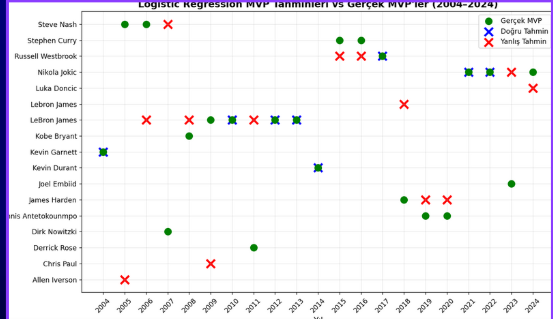


Fig. 4.

* (Fig. 4.)Basit ve yorumlanabilir bir model olmasına rağmen hem 10 hem de 20 yıllık veri setlerinde precision açısından en düşük sonucu vermiştir. Bu durum modelin doğrusal sınırlamalara sahip olması ve sınıflar arasındaki dengesizliği yeterince telafi edememesinden kaynaklanmaktadır.

## Yapılan Analizler

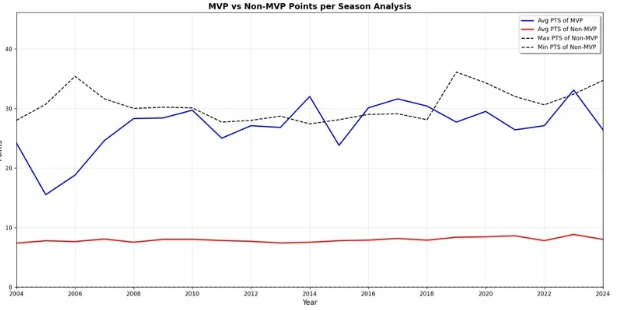
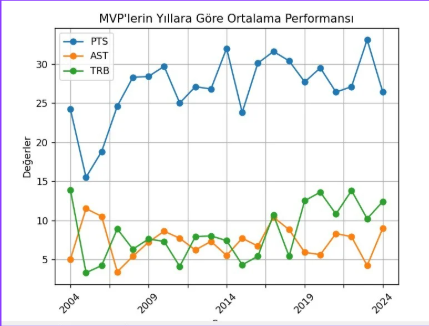
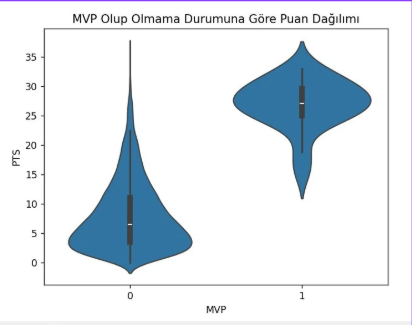


Fig. 5. Yıllara Göre MVP vs non-MVP Puanları

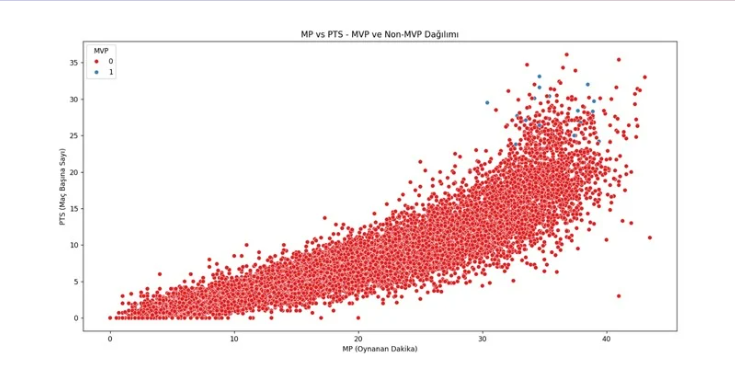
* Fig. 5. grafiği MVP olan oyuncuların maç başına sayı üretiminin genellikle daha yüksek olduğunu, ancak sadece sayı değil, birçok farklı faktörün MVP seçiminde etkili olduğunu göstermektedir.



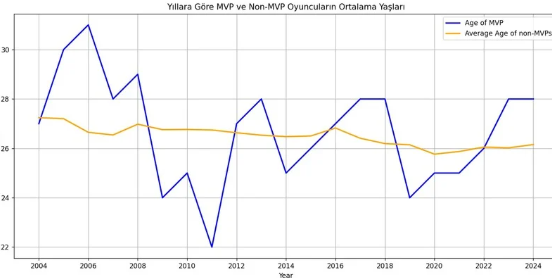
* Yorum: PTS değeri AST ve TRB değerlerinden çok daha büyük. Bu MVP bulunurken PTS özelliğinin daha baskın olduğunu gösterir.



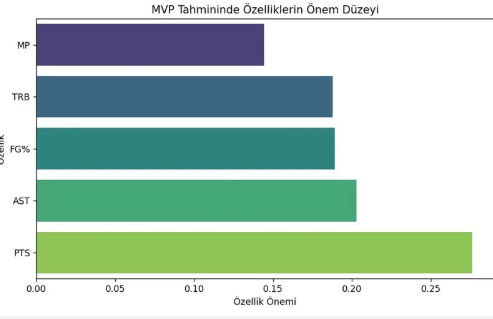
* Yorum: MVP olmayan oyuncuların PTS değerleri daha düşük sayılarda yoğunlaşmışken MVP oyuncuların PTS değerleri daha yüksek sayılarda yoğunlaşmıştır.



* Yorum: Daha fazla dakika oynayan oyuncuların genellikle PTS skoru daha yüksektir.



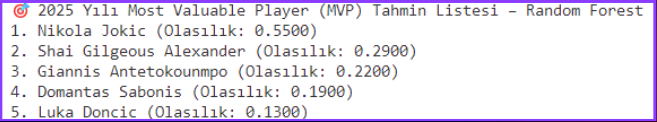
* Yorum: Bu grafik MVP oyuncularının yaş açısından her sezon farklı bir profil çizebildiğini ortaya koyuyor. MVP olmanın yaştan çok performansa ve etkiye dayalı olduğu net şekilde görülüyor.



Bu grafikte MVP hesaplarken özelliklerin ne kadar önemli olduğunu görüyoruz.

## 2025 MVP Tahmin Sonuçları

Model tarafından **2025 sezonu için** tahmin edilen potansiyel MVP adayları, geçmiş MVP profilleriyle büyük ölçüde uyumludur. Ancak, bazı beklenmedik isimlerin listede yer alması, modelin henüz sezon içi form değişikliklerini veya taktiksel değişiklikleri tam anlamıyla yakalayamamasından kaynaklanabilir. Bu durum, veri setinin sınırlamaları ve modelin sadece sayısal istatistiklere dayanmasından ötürü beklenen bir sonuçtur.



# SONUÇ

Bu projede, geçmiş sezon verileri kullanılarak NBA MVP’si tahmini yapılmıştır. Uygulanan makine öğrenmesi modelleri arasında en başarılı sonuç Random Forest ile elde edilmiştir. Sonuçlar, istatistiksel verilerin oyuncu başarısını yordamada etkili olduğunu göstermiştir

##### REFERANSLAR

1. basketball-reference.com. “NBA MVP & ABA Most Valuable Player Award Winners” basketball-reference.com, <https://www.basketball-reference.com/awards/mvp.html>
2. basketball-reference.com. “2023-24 NBA Player Stats: Per Game” basketball-reference.com, https://www.basketball-reference.com/leagues/NBA\_2024\_per\_game.html