Yalova Üniversitesi

Bilgisayar Mühendisliği

Makine Öğrenmesi Yöntemleri le Oyun Öneri Tahmini Sınıflandırması

Anıl Taha TOMAK[1, *]
Mehmet Ali YILDIZ[2, *]

[1]Yalova Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 77100, YALOVA

[2]Yalova Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 77100, YALOVA

Makale Bilgisi

Sunum: 20/05/2025

Anahtar Kelimeler

Makine Öğrenmesi,
Genetik Algoritma,
Öznitelik,
Sınıflandırma,
PCA,
Çapraz
Doğrulama

Öz

Bu çalışma, dijital oyunların kullanıcılar tarafından tavsiye edilip edilmediğini öngörmeye yönelik bir makine öğrenmesi tabanlı sınıflandırma sistemi geliştirmeyi amaçlamaktadır. Proje kapsamında, sekiz farklı sınıflandırma algoritması kullanılarak model performansları detaylı bir şekilde karşılaştırılmıştır. Modeller; ham veri, genetik algoritma ile seçilmiş öznitelikler ve Temel Bileşenler Analizi (PCA) ile boyutu azaltılmış veriler kullanılarak eğitilmiştir. Model değerlendirmeleri 10 katlı çapraz doğrulama yöntemi ile gerçekleştirilmiştir. Elde edilen bulgular, uygulanan öznitelik seçimi ve boyut indirgeme yöntemlerinin sınıflandırma başarısı üzerindeki etkilerini net bir şekilde ortaya koymakta ve oyun öneri tahminine yönelik en etkili yöntem kombinasyonlarını belirlemeye yardımcı olmaktadır.

Classification of Game Recommendation Prediction Using Machine Learning Methods

Keywords

Machine Learning, Genetic Algorithm, Feature selection, Cross-Validation, Classification,PCA

Abstract

This study aims to develop a machine learning-based classification system for predicting whether digital games are recommended by users. Within the scope of the project, the performances of eight different classification algorithms were comparatively analyzed. Models were trained using raw data, features selected through a genetic algorithm, and dimensionally reduced data obtained via Principal Component Analysis (PCA). The evaluation of the models was conducted using 10-fold cross-validation. The results clearly demonstrate the impact of different feature selection and dimensionality reduction techniques on classification performance, and help identify the most effective method combinations for game recommendation prediction.

1. GİRİŞ (INTRODUCTION)

Sınıflandırma problemleri, makine öğrenmesi alanında en sık karşılaşılan ve birçok sektörde geniş uygulama alanı bulan temel problemlerden biridir. Özellikle kullanıcı davranışlarını analiz etmeye yönelik çalışmalarda, sınıflandırma algoritmaları karar destek sistemlerinin temel yapı taşlarını oluşturmaktadır. Günümüzde dijital oyun sektörü, kullanıcı tercihlerini anlamaya yönelik veri odaklı yaklaşımlara giderek daha fazla yönelmekte; bu bağlamda, oyunların kullanıcılar tarafından önerilip önerilmediğini tahmin edebilen sistemler hem kullanıcı deneyimini artırmakta hem de oyun geliştiriciler için önemli geri bildirimler sağlamaktadır.
Bu çalışmada, Kaggle platformu üzerinden elde edilen Steam oyun yorum verileri kullanılarak, oyunların önerilme durumunu tahmin etmeye yönelik bir sınıflandırma modeli geliştirilmiştir. Farklı veri ön işleme teknikleri kullanılarak oluşturulan veri setleri üzerinde, sekiz farklı sınıflandırma algoritmasıyla performans değerlendirmeleri yapılmıştır. Bu kapsamda; ham veriler, Genetik Algoritma (GA) ile öznitelik seçimi yapılmış veriler ve Temel Bileşenler Analizi (PCA**)** ile boyutu azaltılmış veriler kullanılarak modeller oluşturulmuş ve karşılaştırılmıştır.

Geliştirilen modellerin değerlendirilmesi, 10 katlı çapraz doğrulama yöntemi ile gerçekleştirilmiş; elde edilen sonuçlar, farklı veri işleme yöntemlerinin sınıflandırma başarısına olan etkilerini ortaya koymuştur. Bu çalışma, oyun öneri tahminine yönelik etkili ve optimize edilmiş makine öğrenmesi temelli çözümler sunmayı hedeflemekte ve sektörel veri analitiği uygulamalarına katkı sağlamaktadır.

2. YÖNTEMLER (METHODS)

1.1 Veri Kümesi ve Ön İşleme (Dataset and Preprocessing)

Bu çalışmada kullanılan veri kümesi, Kaggle platformunda yayımlanan ve Steam dijital oyun platformuna ait kullanıcı değerlendirmelerini içeren Steam Reviews Dataset 2021 veri setidir [1]. Ham veri kümesi toplamda 16.657.838 kullanıcı yorumundan oluşmaktadır. Her bir kayıt, bir oyuna ait kullanıcı incelemesini temsil etmekte ve ilgili oyunun ismi, kullanıcı bilgileri, yorum içeriği, önerilme durumu (recommended), yorumun yazıldığı dil, zaman damgaları gibi birçok öznitelik içermektedir.

Çalışma kapsamında yalnızca İngilizce dilinde yazılmış yorumlar dikkate alınarak, veri kümesi yaklaşık 7 milyon satıra düşürülmüştür. Ardından, metin içerikli review sütununda kapsamlı bir temizlik süreci uygulanmıştır. Bu süreçte; yalnızca sembollerden oluşan, üç kelimeden az içeriğe sahip olan ya da genel olarak anlam ifade etmeyen yorumlar veri setinden çıkarılmıştır.

Modelin doğruluğunu ve verimliliğini artırmak adına, sınıflandırma açısından anlamlı bilgi içermeyen veya kimlik belirleyici nitelikte olan "app_id", "app_name", "review_id", "language", "timestamp_created", "timestamp_updated", "author.steamid" ve "author.last_played" sütunları veri setinden çıkarılmıştır.

Ön işleme sonrası veri setine ait öznitelikler Tablo 1'de görülmektedir.

Tablo 1. Veri seti öznitelikleri ve açıklamaları

Öznitelik	**Açıklama**
review	Kullanıcının oyun incelemesi
recommended	Kullanıcıların oyunu önerip önermediği
votes_helpful	Yararlı oyların sayısı
votes_funny	Komik oyların sayısı
weighted_vote_score	Yararlı oy sayısına göre puanlama
comment_count	İncelemeye yapılan yorum sayısı
steam_purchase	İnceleme yazarının uygulamayı steamden alıp almadığı
received_for_free	İnceleme yazarının uygulamayı ücretsiz alıp almadığı
written_during_early_access	İncelemenin erken erişim sırasında yazılıp yazılmadığı
author.num_games_owned	İnceleme yazarının sahip olduğu oyun sayısı
author.num_reviews	Yazarın şu ana kadar yaptığı inceleme sayısı

author.playtime_forever	Yazarın incelenen oyundaki toplam oynama süresi
author.playtime_last_two_weeks	Yazarın incelenen oyundaki son iki haftalık oynama süresi
author.playtime_at_review	Yazarın incelenen oyunun incelenme anındaki oynama süresi

Yapılan işlemler sonucunda elde edilen veri sayısı hâlâ oldukça büyük olduğundan, işlenebilir bir örneklem oluşturmak amacıyla 30.000 veri örneği seçilmiştir. Bu örneklem, recommended değişkenindeki sınıfların dengeli (önerilen ve önerilmeyen yorumlar eşit sayıda) olmasına özen gösterilerek oluşturulmuştur.

Veri setinde yer alan review sütunu doğal dil içeriği barındırdığından, bu metinler TF-IDF (Term Frequency–Inverse Document Frequency) yöntemi ile vektörleştirilerek makine öğrenmesi algoritmalarında kullanılabilecek sayısal forma dönüştürülmüştür. Ayrıca veri setinde bulunan boolean türündeki öznitelikler, etiket kodlama (label encoding) yöntemi ile sayısal değerlere çevrilmiştir.

Tüm bu ön işleme adımlarının ardından, dengeli sınıf dağılımına sahip, temizlenmiş ve modellemeye uygun bir veri seti elde edilmiştir.

1.1 Sınıflandırma Yöntemleri (Classification Methods)

Bu çalışmada, kullanıcı yorumlarına dayanarak oyunların önerilip önerilmediğini tahmin etmeye yönelik bir sınıflandırma problemi ele alınmıştır. Bu amaçla, sekiz farklı makine öğrenmesi algoritması kullanılmış ve bu algoritmaların çeşitli veri işleme yöntemleriyle olan performansları karşılaştırılmıştır [2].

2.2.1. Naive Bayes

Naive Bayes algoritması, Bayes teoremi temelli çalışan ve öznitelikler arasında koşulsal bağımsızlık varsayımına dayanan basit ama güçlü bir sınıflandırma yöntemidir. Metin madenciliğinde sıklıkla tercih edilir. Bu projede, TF-IDF ile sayısallaştırılmış yorum verileri üzerinde düşük hesaplama maliyeti ve yüksek hız avantajı sayesinde başarılı sonuçlar elde etmiştir.

2.2.2 k-Nearest Neighbors (kNN)

kNN algoritması, sınıflandırılacak verinin en yakın komşularına bakarak karar verir. Öklidyen mesafe gibi ölçütlerle benzerlik hesaplanır. Yüksek boyutlu metin vektörlerinde hesaplama maliyeti artmakla birlikte, belirli öznitelik seçimi ve boyut indirgeme adımlarından sonra anlamlı sonuçlar verebilmiştir.

2.2.3. Karar Ağaçları (Decision Tree)

Karar ağaçları, veriyi ardışık karar kuralları ile sınıflandıran açıklanabilir algoritmalardır. Yorumların sahip olduğu çeşitli öznitelikler (örneğin, oyun süresi, inceleme puanı gibi) üzerinden sezgisel olarak takip edilebilir sınıflandırma yolları üretmiştir. Basitliği nedeniyle hızlı ve anlaşılabilir bir modelleme sağlamıştır.

2.2.4. Linear Support Vector Machine (Linear SVM)

SVM algoritması, verileri en iyi ayıran hiper düzlemi bulmayı hedefler. Özellikle yüksek boyutlu TF-IDF vektör uzayında etkili karar sınırları oluşturabilen Linear SVM, metin tabanlı sınıflandırma için güçlü bir seçenek olmuştur. Projede, doğruluk oranları bakımından en iyi performans gösteren modellerden biri olmuştur.

2.2.5. Lojistik Regresyon (Logistic Regression)

Logistic Regression, veriyi olasılıksal bir yaklaşımla iki sınıfa ayırır. Doğrusal ayrılabilirlik varsayımı ile çalışan bu algoritma, TF-IDF ile sayısallaştırılmış verilerde basitliği ve hız avantajı ile öne çıkmıştır. Özellikle temel model olarak karşılaştırma için referans noktası görevi görmüştür.

2.2.6. Random Forest

Random Forest, birçok karar ağacından oluşan bir topluluk yöntemidir. Rastgele öznitelik alt kümeleri üzerinden eğitilen ağaçlar sayesinde aşırı öğrenme (overfitting) riski azaltılmıştır. Özellikle öznitelik seçimi (genetik algoritma) sonrası yüksek kararlılık ve genel doğruluk oranları sağlamıştır.

2.2.7 Multilayer Perceptron (MLP)

MLP, yapay sinir ağlarının en temel formudur ve birden fazla gizli katman içerir. Karmaşık örüntüleri öğrenme kapasitesi sayesinde, metin verilerinden çıkarılan soyut ilişkileri modelleyebilmiştir. Bu çalışmada, diğer algoritmalara göre daha fazla eğitim süresi gerektirse de tatmin edici doğruluk değerleri sunmuştur.

2.2.8 Extreme Gradient Boosting (XGBoost)

XGBoost, zayıf sınıflandırıcıları ardışık olarak güçlendiren ve her yeni modelle hataları azaltmaya çalışan güçlü bir topluluk algoritmasıdır. Özellikle yüksek boyutlu ve gürültülü veri üzerinde sağladığı yüksek başarı ile projenin en etkili modellerinden biri olmuştur. Eğitim süresi diğer modellere göre daha uzun olsa da performans avantajı dikkat çekicidir.

2.3 Özellik Seçimi (Feature Selection)

Makine öğrenmesi uygulamalarında yüksek boyutlu veri setleri, modellerin hem öğrenme süresini hem de genelleme yeteneğini olumsuz etkileyebilir. Bu nedenle öznitelik seçimi (feature selection), sınıflandırma performansını artırmak, model karmaşıklığını azaltmak ve yorumlanabilirliği geliştirmek için yaygın olarak kullanılan bir ön işleme tekniğidir.

Öznitelik seçimi; veri kümesindeki tüm öznitelikler arasından, probleme en çok katkı sağlayan alt kümenin belirlenmesini amaçlar. Bu işlem, bazı özelliklerin sınıflandırma için önemsiz veya redundant (tekrarlı) olduğunu varsayarak yalnızca en bilgilendirici öznitelikleri seçer.

Bu çalışmada, öznitelik seçimi problemi, sezgisel ve optimizasyon odaklı bir yaklaşım olan genetik algoritma (GA) ile çözülmüştür. Genetik algoritmalar, doğal seçilim mekanizmasından ilham alan evrimsel algoritmalardır. Çözüm adayları, popülasyonlar halinde değerlendirilir; daha iyi performans gösteren bireyler çaprazlama, mutasyon gibi işlemlerle yeni nesiller oluşturur. Bu süreç sonunda, öznitelik seçiminde optimuma yakın alt kümeler bulunabilir.

2.3.1 Genetik Algoritma Kullanılarak Özellik Seçimi (Feature Selection Using GA)

Bu çalışmada kullanılan genetik algoritma sürecinde:

- Başlangıç Popülasyonu: GA başlangıç popülasyonu, öznitelik seçim maskeleri olarak oluşturulmuş bireylerden oluşmaktadır. Her birey, veri kümesindeki özniteliklerin seçilip seçilmediğini belirten bir dizi 0 ve 1 değerinden oluşmaktadır.
- Fitness Fonksiyonu: Fitness değeri, seçilen öznitelikler kullanılarak Random Forest sınıflandırıcısı ile hesaplanmaktadır. Model, 3 katlı Stratified K-Fold çapraz doğrulama ile eğitilmekte ve her bireyin fitness değeri doğruluk oranına (accuracy) göre belirlenmektedir.
- Ebeveyn Seçimi: Ebeveyn seçimi sürecinde elitizm stratejisi uygulanmış ve en yüksek fitness değerine sahip bireyler korunmuştur. Diğer ebeveynler ise rulet tekeri yöntemi ile belirlenmiştir.
- Çaprazlama ve Mutasyon: Çocuk bireylerin oluşturulması için tek noktalı çaprazlama ve kontrollü mutasyon stratejileri uygulanmıştır. Mutasyon olasılığı %1 olarak belirlenmiştir.
- Döngü Sayısı: GA süreci toplam 20 nesil boyunca tekrarlanmıştır. Bu süreçte her nesilde fitness değeri en yüksek bireyler seçilmiş ve en iyi birey Hall of Fame (HoF) havuzuna eklenmiştir.

  Genetik algoritma sonucunda seçilen öznitelik sayısı 7'dir. Bu öznitelikler:

- votes_helpful
- comment_count
- author.num_games_owned
- author.playtime_forever
- author.playtime_last_two_weeks
- author.playtime_at_review
- recieved_for_free

Bu işlemler, genetik algoritmanın yalnızca sınıflandırma doğruluğunu artırmakla kalmayıp, aynı zamanda modelin daha az ancak daha anlamlı öznitelikler ile daha verimli çalışmasını sağladığını göstermektedir.

2.4 PCA ile Özellik Çıkarımı (Feature Reduction Using PCA)

Bu çalışmada, veri setinin boyutunu azaltarak daha etkili ve daha hızlı bir sınıflandırma performansı elde etmek amacıyla Ana Bileşen Analizi (Principal Component Analysis- PCA) yöntemi kullanılmıştır. PCA, yüksek boyutlu verilerdeki değişkenliği en iyi temsil eden yeni bileşenler oluşturarak, orijinal özelliklerin doğrusal kombinasyonlarından meydana gelen bir alt uzayda veriyi yeniden ifade eder. Bu sayede, bilgi kaybını en aza indirerek modelin hem eğitim süresini azaltması hem de aşırı öğrenmeyi (overfitting) engellemesi hedeflenir.

PCA ile boyut indirgeme sonrasında elde edilen veriler, aynı sınıflandırma algoritmalarıyla yeniden eğitilmiş ve performans sonuçları analiz edilmiştir. Elde edilen sonuçlar, özellikle yüksek boyutlu veri setlerinde PCA'nın sınıflandırma başarısını önemli ölçüde etkilemeden işlem süresini düşürdüğünü ve model karmaşıklığını azalttığını göstermiştir.

3. BULGULAR VE TARTIŞMA (FINDINGS AND DISCUSSION)

Bu çalışmada, oyuncuların yaptığı incelemelere göre oyunların önerilip önerilmediğini sınıflandırmak amacıyla 8 farklı sınıflandırma modeli geliştirilmiştir. Bu modeller şunlardır: Naive Bayes, K-Nearest Neighbors (KNN), Decision Tree, Linear SVM, Logistic Regression, Random Forest, Multilayer Perceptron (MLP) ve XGBoost.

Veri boyutunun çok büyük olmamasıyla birlikte model doğrulama ve performans ölçümünde daha güvenilir sonuçlar elde edebilmek amacıyla 10 katlı çapraz doğrulama (10-fold cross-validation) yöntemi uygulanmıştır. Bu yöntem, veri kümesini 10 eşit parçaya böler; her seferinde bir parçayı test kümesi olarak ayırıp kalan 9 parça ile modeli eğitir. Bu işlem 10 kez tekrarlanır, her tekrarda test kümesi olarak farklı veri kümesi seçilir, her veri noktası hem eğitim hem de test sürecine dahil olmuş olur.

3.1 Karmaşıklık Matrisi (Confusion Matrix)

Sınıflandırma modellerinin başarımını değerlendirmek amacıyla karmaşıklık matrisi kullanılmıştır. Bu matris, modelin gerçek ve tahmin edilen sınıflar arasındaki ilişkiyi tablo şeklinde sunar. Matris dört temel bileşenden oluşur: doğru pozitif (TP), doğru negatif (TN), yanlış pozitif (FP) ve yanlış negatif (FN). Bu değerler üzerinden doğruluk, hassasiyet, duyarlılık, özgüllük, F1 skoru ve MCC gibi performans metrikleri hesaplanarak modelin sınıflandırma yeteneği detaylı şekilde analiz edilmiştir.

3.2 Performans Metrikleri (Performance Metrics)

Model değerlendirmelerinde aşağıdaki altı metrik kullanılmıştır:

Accuracy (Doğruluk):

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Tüm doğru sınıflandırmaların toplam sınıflandırma sayısına oranıdır.

Precision (Kesinlik):

$$\frac{TP}{TP + FP}$$

Pozitif tahmin edilen örneklerin ne kadarının gerçekten pozitif olduğunu ölçer.

Recall (Duyarlılık / Sensitivity):

$$\frac{TP}{TP + FN}$$

Gerçek pozitiflerin ne kadar doğru tespit edildiğini gösterir.

Specificity (Özgüllük):

$$\frac{TN}{TN + FP}$$

Gerçek negatiflerin ne kadarının doğru tahmin edildiğini gösterir.

F1-Score:

$$2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Precision ve recall'un dengeli bir ortalamasıdır.

Matthews Correlation Coefficient (MCC):

$$\frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

-1 ile 1 arasında değer alır ve sınıflandırma başarısını daha dengeli yansıtır.

Bu metriklerin her biri, sadece doğruluğa odaklanmak yerine, sınıflandırma kalitesini çok yönlü analiz etmeyi sağlar [3].

3.3 Ham Veri ile Sonuçlar (Results with Raw Data)

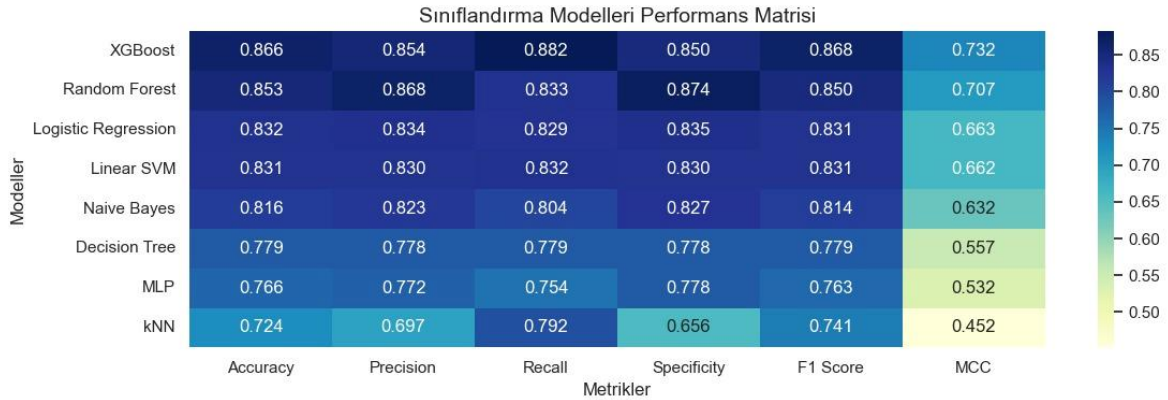
Bu bölümde, sınıflandırma modellerinin ham veriler üzerinde gerçekleştirdiği performans değerlendirmesi sunulmaktadır. Bu süreçte, herhangi bir boyut indirgeme veya öznitelik seçimi uygulanmamış olup, modellerin doğrudan ham verilerle elde ettiği sonuçlar analiz edilmiştir. Performans değerlendirmesi, 10 katlı çapraz doğrulama (cross-validation) yöntemi kullanılarak gerçekleştirilmiş. Doğruluk (accuracy), kesinlik (precision), duyarlılık (recall), F1 skoru, özgüllük (specificity) ve Matthew's korelasyon katsayısı (MCC) olmak üzere altı temel metrik üzerinden ölçülmüştür.

Ham veriler üzerinde gerçekleştirilen analizde, XGBoost modeli %86,6 doğruluk oranı ile en başarılı sonuçları elde etmiştir. Random Forest modeli %85,3 doğruluk oranı ile ikinci sırada yer alırken, Logistic Regression modeli %83,2 doğruluk oranı ile üçüncü sırada yer almıştır. F1 skoru açısından da en yüksek başarıyı XGBoost modeli göstermiştir. Bu sonuçlar, ham veriler üzerinde uygulandığında XGBoost modelinin diğer modellere kıyasla daha tutarlı ve etkili bir performans sergilediğini ortaya koymaktadır.

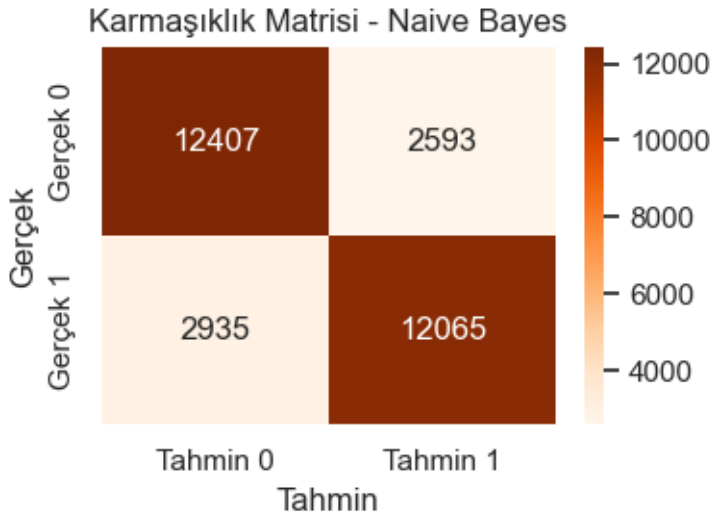
En düşük performansı sergileyen model ise %72,4 doğruluk oranı ile kNN modeli olmuştur. Bu model, doğruluk oranının yanı sıra duyarlılık dışında tüm metriklerde en düşük değerlere sahiptir. Duyarlılık değerinin yüksek olmasına karşın, özgüllük değerinin düşük olması modelin pozitif örnekleri yakalarken negatif örnekleri kaçırdığını gösterir, bu da modelin tutarsız yapısını ortaya koymaktadır. Bu durumu destekleyen bir diğer gösterge ise düşük MCC değeridir; bu durum, modelin sınıflandırmada dengesiz bir performans sergilediğine işaret etmektedir.

MLP ve Decision Tree modelleri çok yüksek doğruluk oranları vermeseler de duyarlılık ve özgüllük değerlerinin birbirine yakın olması ve MCC değerinin de daha yüksek olması, her iki modeli de kNN modeline kıyasla daha tutarlı yapmaktadır.

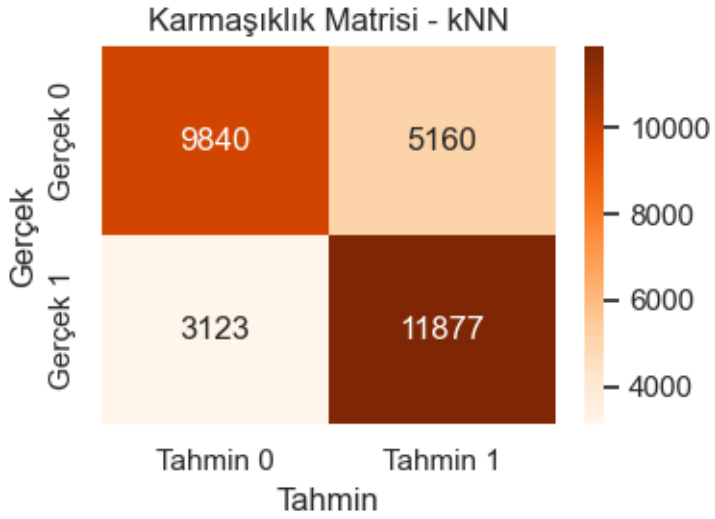
Tablo 2. Ham veriler kullanılarak elde edilen sınıflandırma performans matrisi



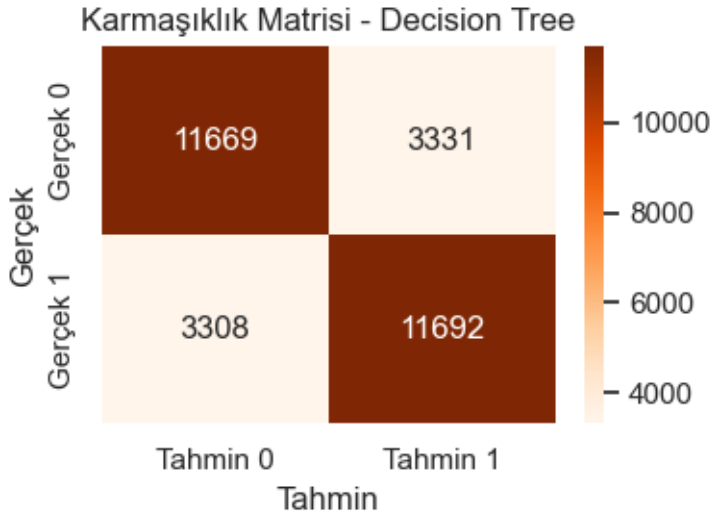
Tablo 3. Naive Bayes karmaşıklık matrisi



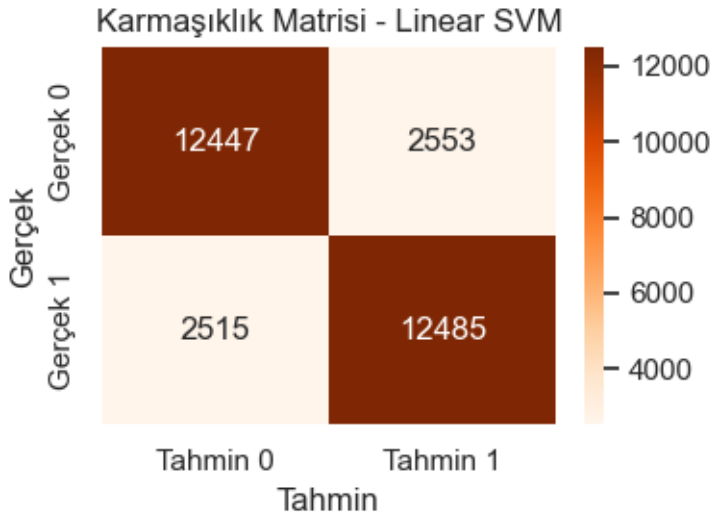
Tablo 4. kNN karmaşıklık matrisi



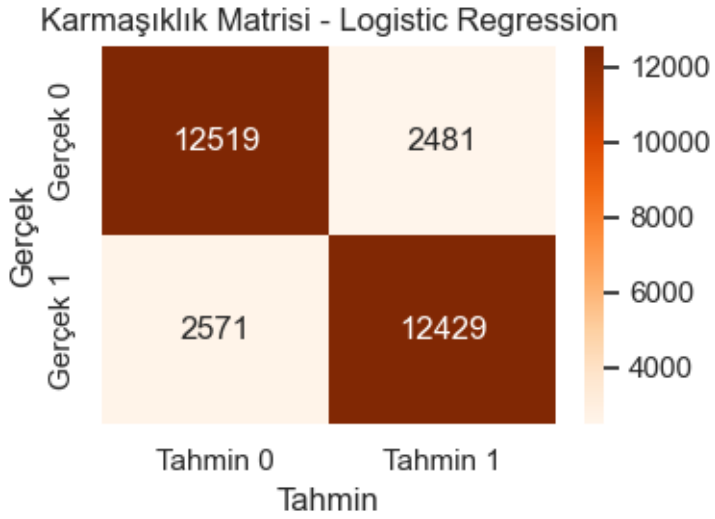
Tablo 5. Decision Tree karmaşıklık matrisi



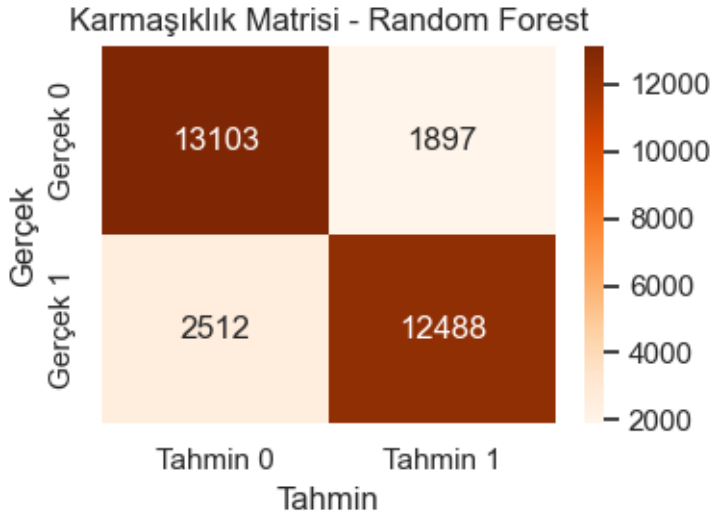
Tablo 6. Linear SVM karmaşıklık matrisi



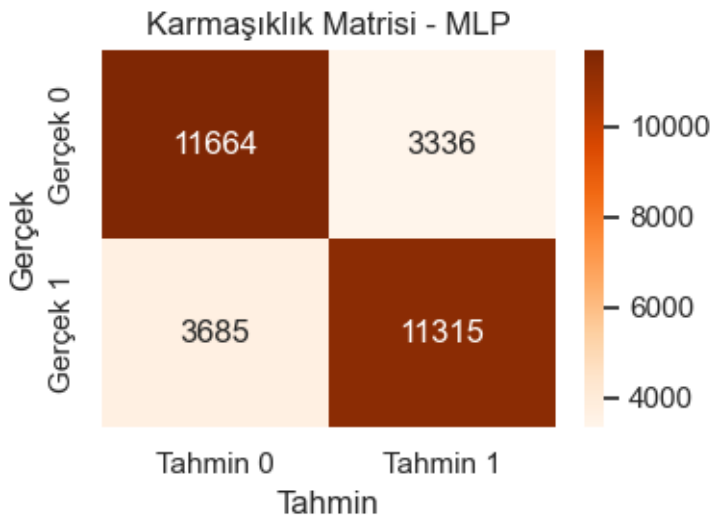
Tablo 7. Logistic Regression karmaşıklık matrisi



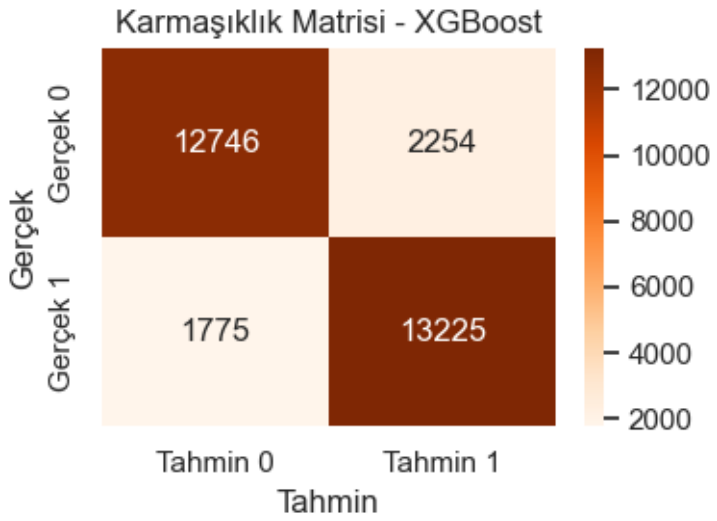
Tablo 8. Random Forest karmaşıklık matrisi



Tablo 9. MLP karmaşıklık matrisi



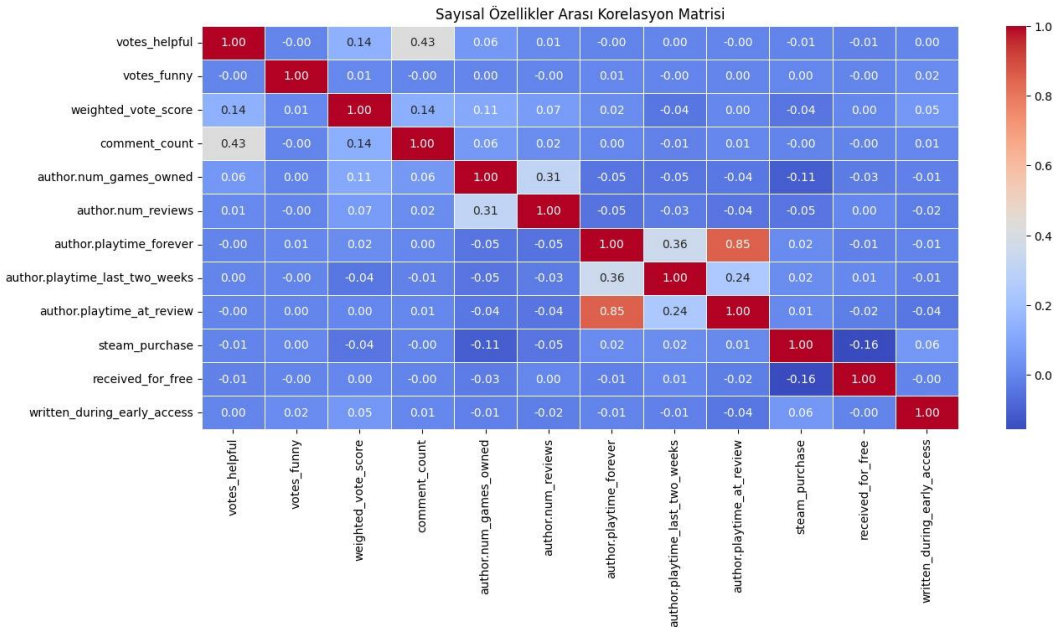
Tablo 10. XGBoost karmaşıklık matrisi



Korelasyon matrisi, bir veri setindeki değişkenler arasındaki ilişkilerin derecesini ve yönünü belirlemek amacıyla kullanılan simetrik bir tablodur. Matrisin her bir hücresinde, iki değişken arasındaki korelasyon katsayısı yer alır ve bu değer -1 ile 1 arasında değişir. Pozitif değerler değişkenlerin birlikte artma eğiliminde olduğunu, negatif değerler ise bir değişken artarken diğerinin azaldığını gösterir. Sıfır değeri ise iki değişken arasında herhangi bir ilişki olmadığını ifade eder.

Korelasyon katsayısının büyüklüğü ilişkinin kuvvetini belirtirken, işareti ilişkinin yönünü belirler. Örneğin, 0.8 gibi pozitif bir değer güçlü bir pozitif ilişkiyi ifade ederken, -0.7 gibi bir değer güçlü bir negatif ilişkiyi işaret eder. Bu nedenle, korelasyon matrisi veri analizinde değişkenler arasındaki ilişkileri görselleştirmek ve daha derinlemesine analizler gerçekleştirmek için etkili bir araçtır.

Tablo 11. Sayısal özellikler arası korelasyon matrisi



3.4 Genetik Algoritma ile Seçilen Özellikler Üzerinden Sonuçlar

Bu bölümde, genetik algoritmaların (GA) öznitelik seçimi sürecindeki rolü ele alınmıştır. GA, biyolojik evrim prensiplerinden ilham alan ve çözüm uzayında en uygun öznitelik altkümelerini belirlemeyi hedefleyen bir arama ve optimizasyon yöntemidir. Bu yaklaşım, yüksek boyutlu veri kümelerinde sınıflandırma performansını artırmak amacıyla bilgi açısından en etkili öznitelikleri seçerek modellerin genel doğruluk oranını optimize etmeyi amaçlar.

Değerlendirme, 10 katlı cross-validation yöntemiyle gerçekleştirilmiş ve her model için accuracy, precision, recall, F1 score, specificity ve Matthews Correlation Coefficient (MCC) metriklerinin değerleri hesaplanmıştır.

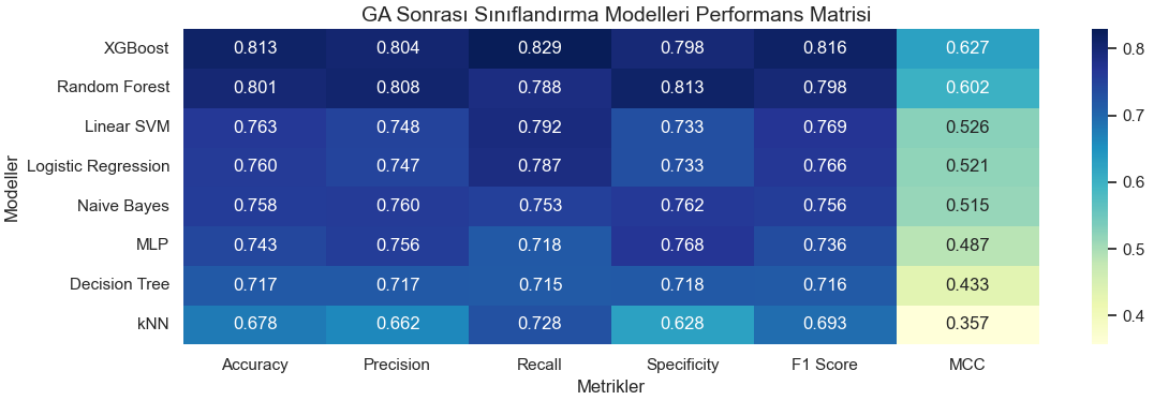
Genetik algoritma ile seçilen özelliklerin kullanımı sonucunda, tüm modellerin doğruluk oranı ve diğer performans metriklerinde belirgin bir düşüş gözlemlenmiştir. XGBoost modeli %81,3 doğruluk oranı ile en yüksek performansa sahip model olmasına karşın, ham veriler ile elde edilen doğruluk oranına kıyasla daha düşük bir sonuç ortaya çıkmıştır. Bunun yanı sıra, yalnızca doğruluk oranında değil, diğer tüm metriklerde de kayda değer bir performans azalması tespit edilmiştir.

Genetik algoritma ile yapılan değerlendirmelerde, ham verilerle elde edilen sonuçlarda olduğu gibi Random Forest modeli yine en yüksek ikinci doğruluk oranını sağlamıştır. Bununla birlikte, %76,3 doğruluk oranı ile Linear SVM modeli, genetik algoritma uygulanması sonrasında üçüncü en iyi performansı gösteren model olarak öne çıkmıştır.

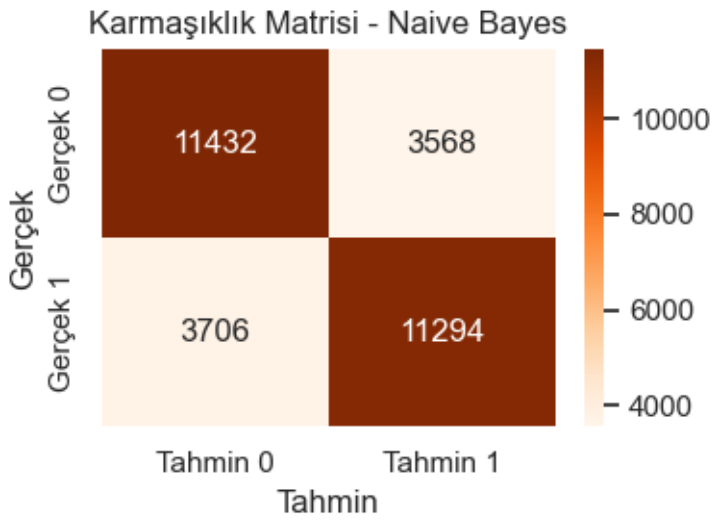
Genetik algoritma ile yapılan değerlendirmelerde en düşük doğruluk oranı %67,8 ile kNN modelinde elde edilmiştir. Ham verilerle yapılan değerlendirmelerde de düşük performans sergileyen kNN modeli, genetik algoritma uygulanması sonrasında daha da kötü bir sonuç ortaya koymuştur. Sadece doğruluk oranında değil, duyarlılık, özgüllük ve MCC değerlerinde de belirgin bir düşüş gözlemlenmiştir. Özellikle duyarlılık değerinin yüksek olmasına karşın özgüllük değerinin düşük olması ve MCC değerinin 0.357 gibi oldukça düşük bir seviyede kalması, kNN modelinin genetik algoritma ile tutarsız ve zayıf bir performans sergilediğini göstermektedir.

Ham veriye kıyasla, genetik algoritma ile seçilen özniteliklerin kullanılması durumunda tüm metriklerde belirgin düşüşler yaşanmıştır. Bu durum, özellik seçiminin başarısız bir şekilde gerçekleştirildiğini ya da özellik seçiminin sınıflar arası ayrımı zorlaştırarak modelin performansını olumsuz etkilediğini göstermektedir.

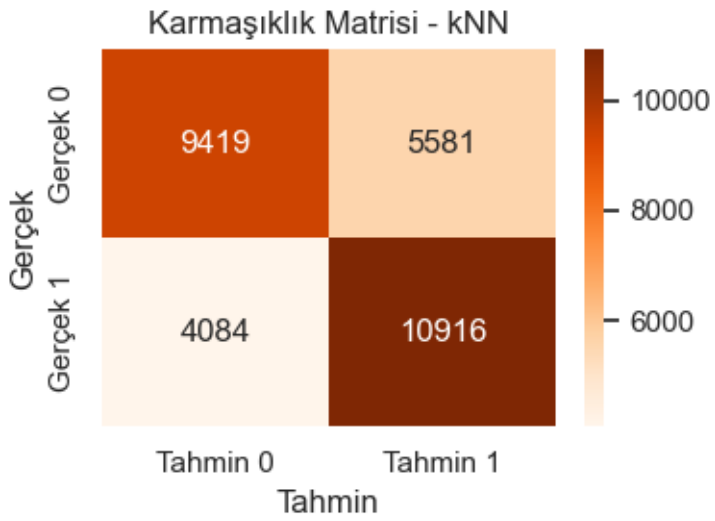
Tablo 12. Genetik algoritma ile seçilen özniteliklerle elde edilen sınıflandırma performans matrisi



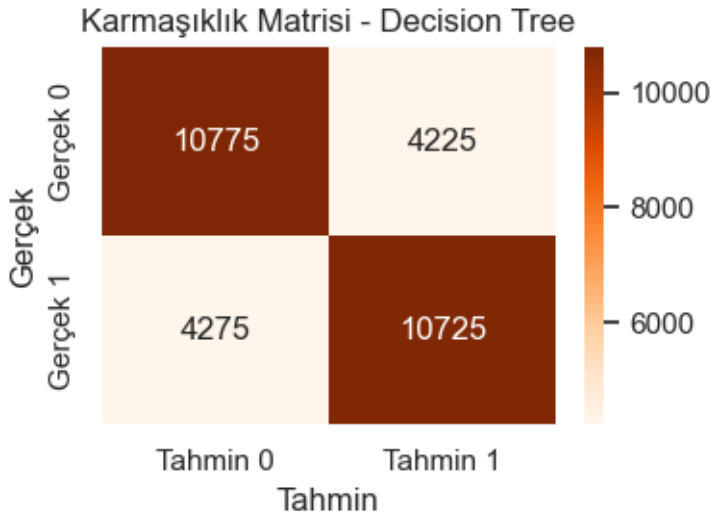
Tablo 13. Naive Bayes karmaşıklık matrisi



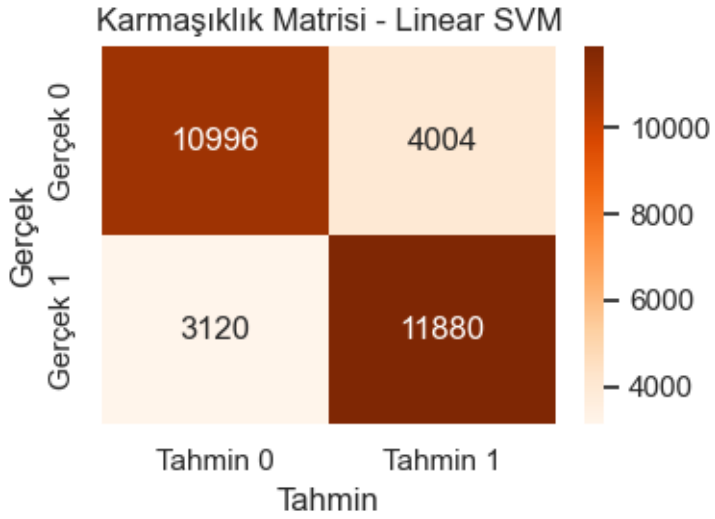
Tablo 14. kNN karmaşıklık matrisi



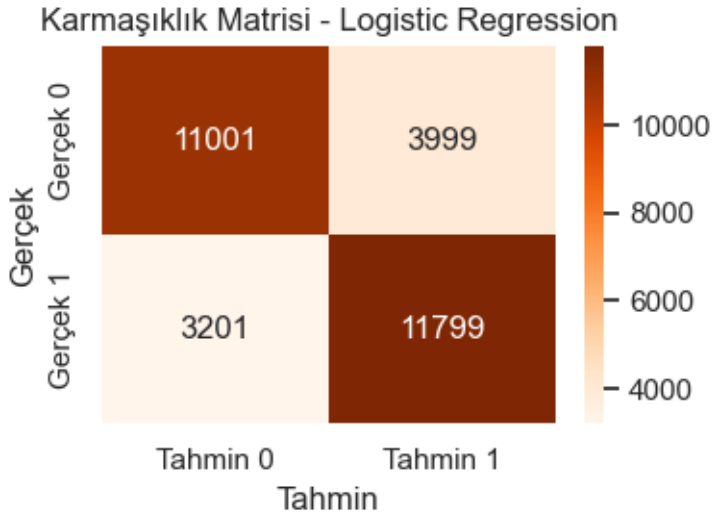
Tablo 15. Decision Tree karmaşıklık matrisi



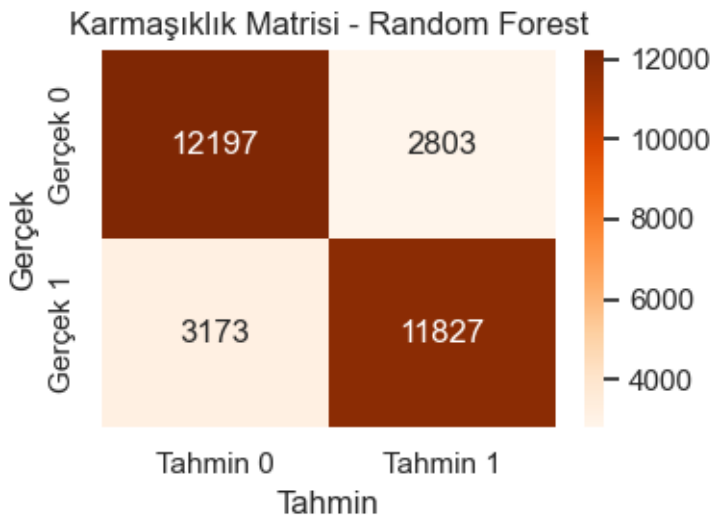
Tablo 16. Linear SVM karmaşıklık matrisi



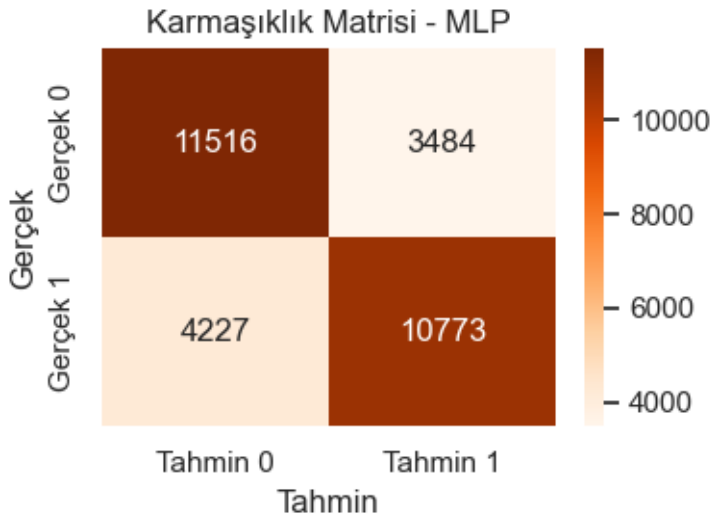
Tablo 17. Logistic Regression karmaşıklık matrisi



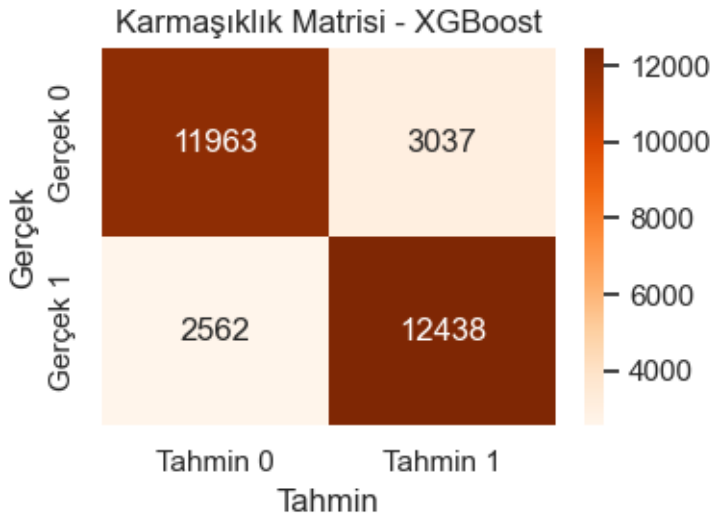
Tablo 18. Random Forest karmaşıklık matrisi



Tablo 19. MLP karmaşıklık matrisi



Tablo 20. XGBoost karmaşıklık matrisi



3.5 PCA ile Özellik Çıkarımı Sonuçları (Feature Reduction Results with PCA)

Bu çalışmada, özellik boyutunun azaltılmasına yönelik olarak Principal Component Analysis (PCA) yöntemi

uygulanmış ve elde edilen öznitelikler doğrultusunda sınıflandırma performansı analiz edilmiştir. PCA, doğrusal bir boyut indirgeme tekniği olup, yüksek boyutlu veri kümelerinde yer alan korelasyonlu öznitelikleri ortogonal bileşenler aracılığıyla yeniden yapılandırarak veri setindeki varyansın büyük kısmını daha az sayıda bileşende toplamayı hedeflemektedir. Bu sayede hem hesaplama maliyetinin düşürülmesi hem de modellerin aşırı öğrenmeye karşı daha dirençli hâle getirilmesi amaçlanmıştır.

Uygulanan PCA sonrasında, elde edilen düşük boyutlu temsil kullanılarak sekiz farklı makine öğrenmesi algoritması üzerinde sınıflandırma işlemleri gerçekleştirilmiştir. Bu süreçte, diğer tüm deneysel aşamalarda olduğu gibi 10 katlı çapraz doğrulama yöntemi tercih edilmiştir. Modellerin başarıları doğruluk (accuracy), kesinlik (precision), duyarlılık (recall), F1 skoru, özgüllük (specificity) ve Matthew's korelasyon katsayısı (MCC) gibi yaygın performans metrikleri kullanılarak değerlendirilmiştir.

Elde edilen sonuçlara göre, PCA sonrasında %82,5 doğruluk oranı ile en yüksek performansı sergileyen model XGBoost olmuştur. Bu modeli %81,2 doğruluk oranı ile Linear Support Vector Machine (SVM) ve Logistic Regression algoritmaları takip etmiştir. Her ne kadar bu üç model, PCA sonrasında genel sınıflandırma başarılarını büyük ölçüde korumuş olsa da elde edilen sonuçlar boyut indirgeme işleminin modellerin performansı üzerinde az da olsa olumsuz bir etki yarattığını göstermektedir.

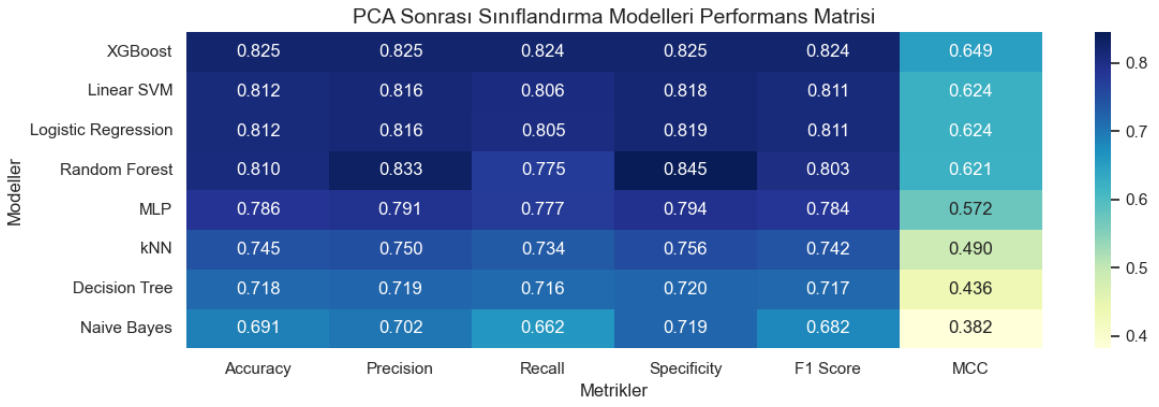
RandomForest algoritması her ne kadar doğruluk oranı bakımından dördüncü sırada yer alsa da %84,5 özgüllük (specificity) ve %83,3 precision değerleri ile diğer modellerden ayrışmaktadır. Bu durum, RandomForest'ın özellikle negatif sınıfların doğru şekilde sınıflandırılmasında başarılı bir yapı sergilediğini ortaya koymaktadır.

Öte yandan, Naive Bayes ve Decision Tree modellerinde belirgin performans düşüşleri gözlemlenmiştir. Naive Bayes, %69,1 doğruluk oranı ve 0.382 MCC değeri ile tüm modeller arasında en düşük performansı göstermiştir. Benzer şekilde, Decision Tree algoritması da düşük metrik değerleri ile öne çıkmıştır. Bu sonuçlar, PCA'nın bazı basit ya da varsayımsal modellere entegre edildiğinde sınıf ayrımını zorlaştırabileceğine işaret etmektedir.

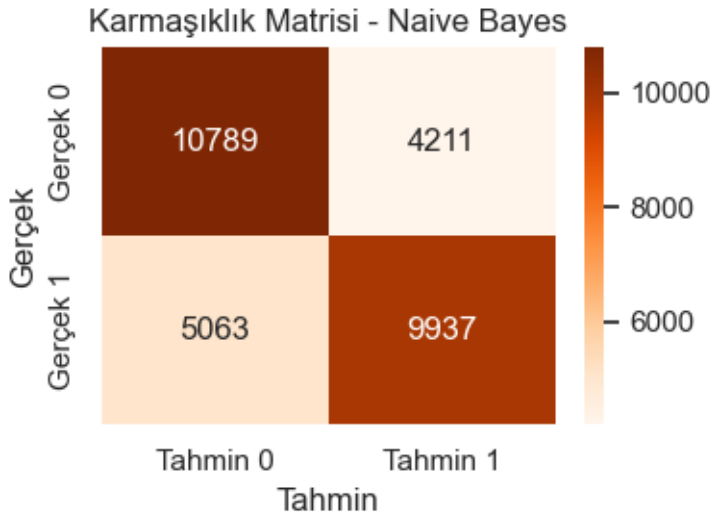
PCA uygulaması sonrasında, önceki modellerde en düşük performansı gösteren kNN modelinin doğruluk oranında artış gözlemlenmiş ve model, Decision Tree ve Naive Bayes algoritmalarına kıyasla daha iyi bir performans sergilemiştir. Ayrıca, duyarlılık ve özgüllük değerleri arasındaki farkın azalması, kNN modelini diğer yöntemlere göre daha tutarlı hale getirmiştir.

Genel olarak değerlendirildiğinde, PCA uygulaması ile elde edilen öznitelik temsilleri, genetik algoritma ile yapılan özellik seçimine kıyasla daha stabil ve güvenilir bir performans ortaya koymuştur. Özellikle karmaşık modellerin PCA sonrasında da yüksek sınıflandırma başarıları göstermesi, bu yöntemin veri setindeki bilgi içeriğini büyük ölçüde koruduğunu ve sınıflandırma sürecine katkı sağladığını göstermektedir.

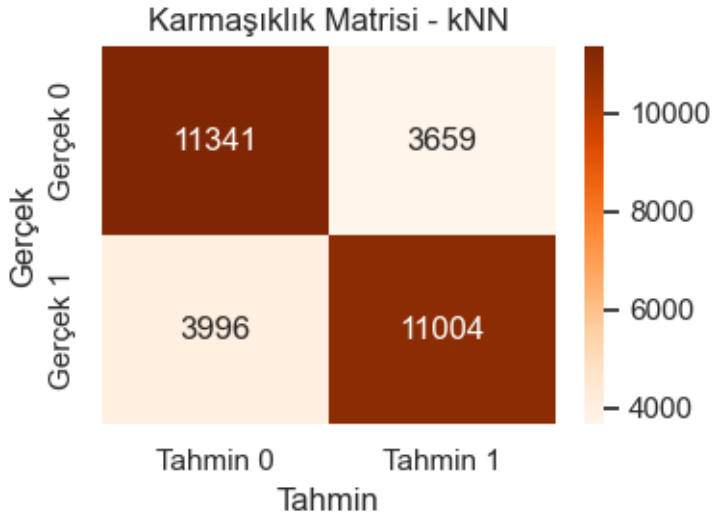
Tablo 21. PCA ile boyut indirgeme sonrası elde edilen sınıflandırma performans matrisi



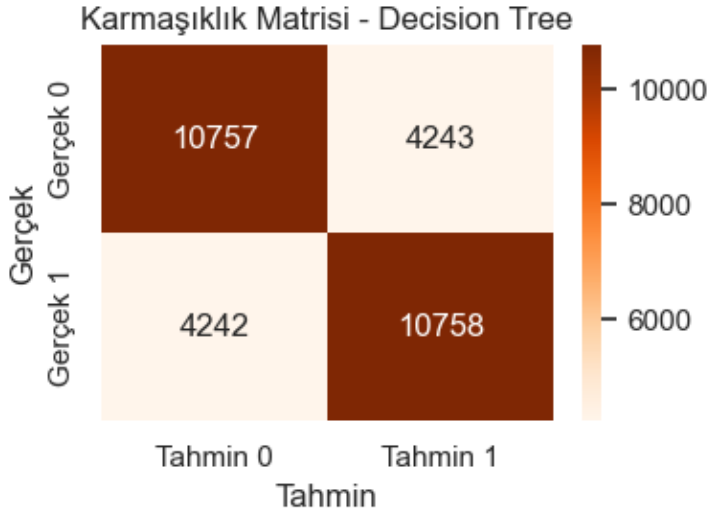
Tablo 22. Naive Bayes karmaşıklık matrisi



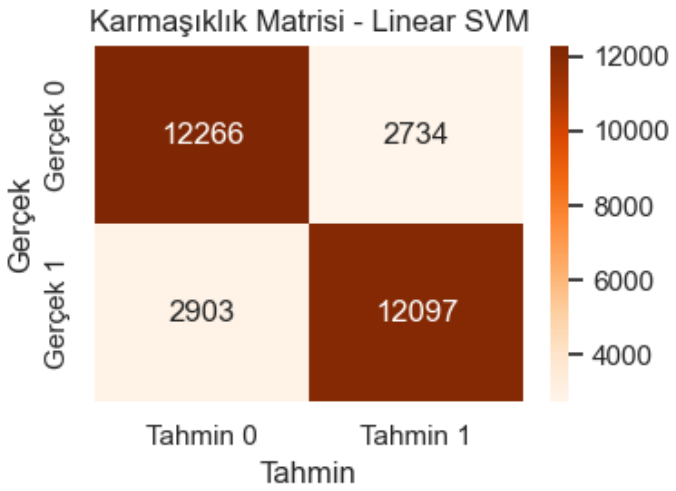
Tablo 23. kNN karmaşıklık matrisi



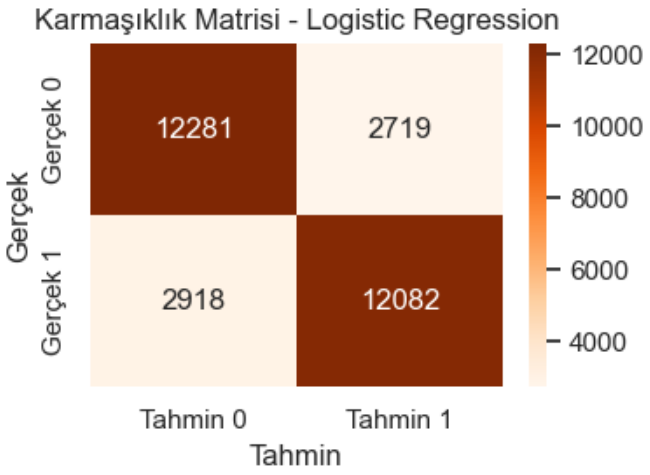
Tablo 24. Decision Tree karmaşıklık matrisi



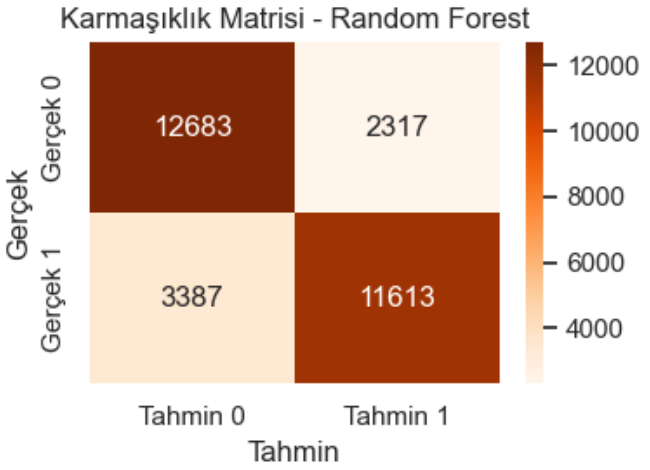
Tablo 25. Linear SVM karmaşıklık matrisi



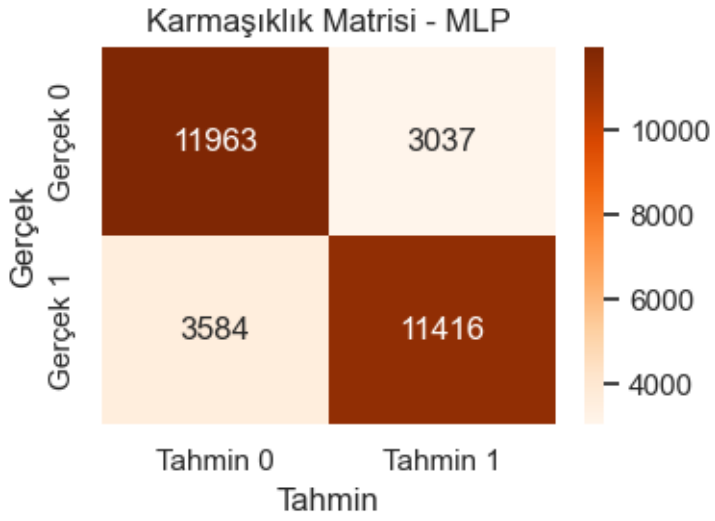
Tablo 26. Logistic Regression karmaşıklık matrisi



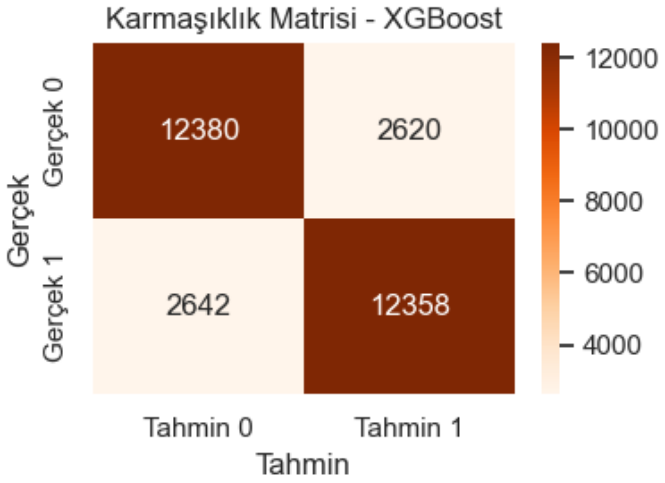
Tablo 27. Random Forest karmaşıklık matrisi



Tablo 28. MLP karmaşıklık matrisi



Tablo 29. XGBoost karmaşıklık matrisi



4 SONUÇLAR (CONCLUSIONS)

Bu çalışmada, dijital oyun platformlarındaki kullanıcı yorumlarını analiz ederek oyunların tavsiye edilip edilmediğini tahmin edebilen makine öğrenmesi tabanlı bir sınıflandırma sistemi geliştirilmiştir. Steam platformuna ait kullanıcı yorumlarından oluşturulan veri kümesi üzerinde; NaiveBayes, kNN, Decision Tree, Linear SVM, Logistic Regression, Random Forest, MLP ve XGBoost olmak üzere sekiz farklı sınıflandırma algoritması test edilmiştir.

Modelleme sürecinde, üç farklı veri kümesi yapılandırılmıştır: ham veriler, Genetik Algoritma (GA) ile seçilmiş özniteliklerden oluşan alt küme ve Ana Bileşenler Analizi (PCA) ile boyutu azaltılmış veri temsilleri. Tüm deneyler 10 katlı çapraz doğrulama yöntemi ile yürütülmüş ve her bir algoritma için doğruluk, kesinlik, duyarlılık, özgüllük, F1 skoru ve Matthews Korelasyon Katsayısı (MCC) gibi performans metrikleri hesaplanmıştır.

Ham verilerle yapılan sınıflandırma sonuçlarına göre, %86,6 doğruluk oranı ile en yüksek performansı XGBoost modeli göstermiştir. Genetik algoritma ile yapılan öznitelik seçimi sonrasında ise genel olarak tüm modellerde performans düşüşü yaşanmış ve bu durum özellikle kNN ve Naive Bayes gibi daha basit modellerde belirgin hale gelmiştir. PCA uygulaması ise bilgi kaybını minimumda tutarak, bazı modellerde ham veriye kıyasla olmasa da genetik algoritma ile yapılan sınıflandırma sonuçlarına göre daha istikrarlı sonuçlar elde edilmesini sağlamıştır. Her ne kadar çoğu model ham veriye kıyasla daha düşük performans göstermiş olsa da özellikle kNN modeli doğruluk oranı ve tutarlılık açısından olumlu yönde bir gelişme göstermiştir.

Genel değerlendirme sonucunda, PCA ile boyutu indirgenmiş veriler üzerinde çalışan XGBoost modeli hem yüksek doğruluk hem de dengeli metrik değerleriyle en güçlü model olarak öne çıkmıştır. Genetik algoritma ile seçilen öznitelikler sınıflandırma başarısını istenilen düzeyde iyileştirememiş, ancak Random Forest ve

Linear SVM gibi bazı modellerde makul seviyelerde sonuçlar elde edilmiştir.

Gelecek çalışmalar kapsamında, derin öğrenme tabanlı yöntemlerle model performansının artırılması ve kullanıcı yorumlarının duygu analizi gibi doğal dil işleme yöntemleriyle daha da zenginleştirilmesi planlanmaktadır. Ayrıca, öznitelik seçiminin açıklanabilir yapay zekâ yaklaşımları ile desteklenmesi, model çıktılarının yorumlanabilirliğini artırarak daha güçlü karar destek sistemlerinin geliştirilmesine katkı sağlayacaktır.

KAYNAKLAR (REFERENCES)

[1] https://www.kaggle.com/datasets/najzeko/steam-reviews-2021

[2] https://scikit-learn.org/stable/supervised_learning.html

[3] https://medium.com/@serapozden922/confusion-matrix-kar%C4%B1%C5%9F%C4%B1kl%C4%B1k-matrisi-62c43b8ad2b0