BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ TEZLİ YÜKSEK LİSANS PROGRAMI

İSTATİSTİKSEL ANALİZ YÖNTEMLERİ VE MAKİNE ÖĞRENME YÖNTEMLERİ İLE FİLM BAŞARI TAHMİNİ

HAZIRLAYAN

BUGAY SARIKAYA

YÜKSEK LİSANS TEZİ

ANKARA - 2021

BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ TEZLİ YÜKSEK LİSANS PROGRAMI

İSTATİSTİKSEL ANALİZ YÖNTEMLERİ VE MAKİNE ÖĞRENME YÖNTEMLERİ İLE FİLM BAŞARI TAHMİNİ

HAZIRLAYAN

BUGAY SARIKAYA

YÜKSEK LİSANS TEZİ

TEZ DANIŞMANI

DR. ÖĞR. ÜYESİ DUYGU DEDE ŞENER

ANKARA – 2021

BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Bilgisayar Mühendisliği Anabilim Dalı Bilgisayar Mühendisliği Tezli Yüksek Lisans Programı çerçevesinde Bugay Sarıkaya tarafından hazırlanan bu çalışma, aşağıdaki jüri tarafından Yüksek Lisans Tezi olarak kabul edilmiştir.

Tez Savunma Tarihi: 11 / 08 / 2021

Tez Adı: İstatistiksel Analiz Yöntemleri ve Makine Öğrenme Yöntemleri ile Film Başarı Tahmini

Tez Jüri Üyeleri	**İmza**
Dr. Öğr. Üyesi Damla TOPALLI, Atılım Üniversitesi	……..
Dr. Öğr. Üyesi Duygu DEDE ŞENER, Başkent Üniversitesi	……..
Dr. Öğr. Üyesi Tülin ERÇELEBİ AYYILDIZ, Başkent Üniversitesi	……..

ONAY

Prof. Dr. Ömer Faruk ELALDI
Fen Bilimleri Enstitüsü Müdürü
Tarih : … / … / ….…….

BAŞKENT ÜNİVERSİTESİ

FEN BİLİMLER ENSTİTÜSÜ

YÜKSEK LİSANS ÇALIŞMASI ORİJİNALLİK RAPORU

Tarih: … / … / 20…

Öğrencinin Adı, Soyadı : Bugay Sarıkaya

Öğrencinin Numarası : 21920399

Anabilim Dalı : Bilgisayar Mühendisliği

Programı : Bilgisayar Mühendisliği Tezli Yüksek Lisans

Danışmanın Unvanı/Adı, Soyadı : Dr. Öğr. Üyesi Duygu Dede Şener

Tez Başlığı : İstatistiksel Analiz Yöntemleri ve Makine Öğrenme Yöntemleri ile Film Başarı Tahmini

Yukarıda başlığı belirtilen Yüksek Lisans tez çalışmamın; Giriş, Ana Bölümler ve Sonuç Bölümünden oluşan, toplam 32 sayfalık kısmına ilişkin, 24 / 08 / 2021 tarihinde tez danışmanım tarafından Turnitin adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı %8' dir. Uygulanan filtrelemeler:

1. Kaynakça hariç
2. Alıntılar hariç
3. Beş (5) kelimeden daha az örtüşme içeren metin kısımları hariç

"Başkent Üniversitesi Enstitüleri Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Usul ve Esaslarını" inceledim ve bu uygulama esaslarında belirtilen azami benzerlik oranlarına tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Öğrenci İmzası:…….

ONAY

Tarih: … / … / 20…

Dr. Öğr. Üyesi Duygu DEDE ŞENER

TEŞEKKÜR

Sayın Dr. Öğr. Üyesi Duygu Dede Şener'e (tez danışmanı), çalışmanın sonuca ulaştırılmasında ve karşılaşılan güçlüklerin aşılmasında her zaman yardımcı ve yol gösterici olduğu için…

ÖZET

Bugay SARIKAYA

İSTATİSTİKSEL ANALİZ YÖNTEMLERİ VE MAKİNE ÖĞRENME YÖNTEMLERİ İLE FİLM BAŞARI TAHMİNİ

Başkent Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

2021

Film endüstrisinde başarılı bir sinema filmi çekmek için büyük yatırımlar yapılmaktadır. Ancak, büyük yatırımlara rağmen, beklenildiği gibi başarılı olamayan bazı film örnekleri mevcuttur. Bu nedenle, bir filmin başarısını büyük ölçekte tahmin etmek, film yapımcıları için film vizyona girmeden önce çok önemlidir. Bu çalışmada, yapımcılara film yatırımı konusunda bir öngörü sağlamak için sınıflandırmaya dayalı bir tahmin modeli geliştirilmesi amaçlanmıştır. Bir filmin başarısını tahmin etmek için önerilen modelde farklı istatistiksel analiz ve makine öğrenmesi yaklaşımları kullanılmıştır. Esas olarak, hangi film özelliğinin filmin başarısı ile yüksek oranda ilişkili olduğunu ve film başarısını tahmin etmede hangi makine öğrenme tekniğinin daha iyi olduğunu tespit etmeye odaklanılmıştır. Bunun için öncelikle ki-kare (chi-square) analizi ve varyans analizi testi kullanılarak istatistiksel bir analiz yapılmıştır. Ardından sınıflandırma yöntemlerinden Rastgele Orman (Random Forest), Destek Vektör Makinesi (Support Vector Machine) ve Yapay Sinir Ağı (Artificial Neural Network), regresyon yöntemlerinden, Çoklu Doğrusal Regresyon (Multi Linear Regression), Destek Vektör Regresyon (Support Vector Regression), Rastgele Orman Regresyon (Random Forest Regression) gibi farklı makine öğrenme teknikleri kullanılarak karşılaştırmalı bir analiz yapılmıştır. Deneysel sonuçlar, bir filmin başarısının en önemli belirleyicilerinin "oyOrtalamasi", "oySayisi", "gelir" ve "butce" olduğunu göstermektedir. Bunun yanı sıra Rastgele Orman, diğer makine öğrenme yöntemleri arasında film başarısını tahmin etmede %96 doğruluk (accuracy) oranıyla başarılı olmuştur. Destek Vektör Regresyon, diğer regresyon yöntemleri arasında film başarı tahmin etmede 1.77 Kök Ortalama Kare Hatası (Root Mean Square Error, RMSE) değeri ile başarılı olmuştur.

ANAHTAR KELİMELER: Film başarı tahmini, Film başarı sınıflandırması, Makine Öğrenimi, İstatistiksel Analiz

ABSTRACT

Bugay SARIKAYA

MOVIE SUCCESS PREDICTION WITH STATISTICAL ANALYSIS TECHNIQUES AND MACHINE LEARNING METHODS

Baskent University Institute of Science and Engineering

Department of Computer Engineering

2021

In the movie industry, huge investments have been made to shoot a successful motion picture. However, despite large investments, there are some movie examples that cannot be successful as expected. Therefore, predicting the success of a movie is so important on a large scale for the movie producers before releasing the movie. In this study, a classification-based prediction model is aimed to develop for providing a foresight to the producers about investing on a movie. Different statistical analysis and machine learning approaches were used in the proposed model for predicting success of a movie. We mainly focus on detecting which movie attribute is highly correlated with the success of the movie and which machine learning technique is better at predicting the movie success. To do so, firstly a statistical analysis was conducted by using chi-square analysis and analysis of variance test. Then a comparative analysis was performed by using different machine learning techniques including Random Forest, Support Vector Machine and Artificial Neural Network, and Multiple Linear Regression, Support Vector Regression and Random Forest Regression methods as regression methods. The experimental results indicate that the most important predictors of a movie's success are "voteAverage", "voteCount", "revenue" and "budget. In addition to this, Random Forest has become successful by the accuracy of 96% in predicting movie success among other machine learning methods, and Support Vector Regression has become successful by the Root Mean Square Error (RMSE) 1.77 in predicting movie success among other regression methods.

KEYWORDS: Movie success prediction, Movie success classification, Machine Learning, Statistical Analysis

İÇİNDEKİLER

TABLOLAR LİSTESİ

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

SİMGELER VE KISALTMALAR LİSTESİ

AI	Artificial Intelligence
FN	False Negative
FP	False Positive
IMDB	The Internet Movie Database
OMDB	The Open Movie Database
RF	Random Forest
RMSE	Root Mean Square Error
TMDB	The Movie Database
TN	True Negative
TP	True Positive
YSA	Yapay Sinir Ağı

1. GİRİŞ VE LİTERATÜR ÇALIŞMASI

Film sektöründe başarıyı ölçmek için genelde filmlerin IMDB (Internet Movie Database) puanları ya da gişe hasılatları değerlendirilir. Yüksek bir IMDB puanı ve gişe hasılatı alabilecek bir sinema filmi çekebilmek, yatırımcılara ciddi büyüklükte bir maliyete mal olmaktadır. Yüksek bütçelere, detaylı çalışmalara rağmen hedeflenen başarıya ulaşamayan birçok sinema filmi mevcuttur.

1.1. Çalışmanın Motivasyonu, Amacı ve Tanımı

Film endüstrisi, tüm dünyada genişleyen endüstrilerden biridir. Bu endüstrinin hızlı büyümesi ve ekonomik etkisi ile birlikte birçok araştırmacı film endüstrisi üzerinde çalışmaktadır. Özellikle bir filmin başarısı üzerinde etkisi olan faktörleri araştırmak için öngörücü modeller yürütmek son on yılda popüler bir araştırma alanı haline gelmiştir. Bir sinema filminin başarısının filmin gişe, bütçe, gelir ve popülerlik düzeyi gibi birçok özelliğine bağlı olduğu bilinen bir gerçektir. Bir filmin başarısında etkisi olan yönetmen, oyuncular gibi başka önemli faktörler olsa da her film her seferinde beklenen gişeyi veya başarıyı elde edemez. Bu nedenle, yapımcıların bir filmin başarısını vizyona girmeden önce tahmin etmeleri temel bir ihtiyaç haline gelmiştir. Son zamanlarda, farklı makine öğrenmesi yaklaşımlarını kullanarak film başarısını tahmin etmeye odaklanan bazı çalışmalar bulunmaktadır. Filmin başarı ihtimalini yükseltmek için doğru bileşenler bir araya getirilmeli ve doğru yatırımlar yapılmalıdır.

1.2. Önceki Çalışmalar

Ahmad J. ve arkadaşları [1] tarafından, veri madenciliğini kullanarak film başarı tahmini üzerine çalışma yapılmıştır. Filmin ismi, çıkış yılı, türü, direktörü, müzik direktörleri, yapımcılarını kullanarak bir veri kümesi oluşturulmuştur. X2 analizini kullanarak çeşitli özellikler arasındaki korelasyonları bulmayı içeren film başarısını tahmin etmek için bir matematiksel model önerilmiştir. Çalışmada simülasyon verileri kullanılmıştır ve sadece Bollywood filmlerinde test edilmiştir ve ayrıca aktörlerin ve film türlerinin filmin başarısını etkilediği sonucuna varılmıştır.

Ping-Yu Hsu ve arkadaşları [2] tarafından, IMDB nitelikleri ile film kullanıcı puanlarını tahmin etme üzerine çalışma yapılmıştır. IMDB öznitelikleri ile kullanıcı derecelendirmelerini tahmin etmek için özel bir model geliştirilmiştir. Kullanılan veri setinde 32968 film bulunmakta olup 31506 film eğitim verisi ve diğerleri test verisi olarak kullanılmıştır. Doğrusal Kombinasyon (Linear Combination), Çoklu Doğrusal Regresyon (Linear Regression) ve sinir ağı yöntemleri (Neural Network) ve X2 analizi yöntemleri kullanılmıştır. Yazar, oyuncu, yönetmen gibi bazı özelliklerin kullanıcı puanlarını derinden etkilediğini belirtilmiştir.

Eker ve arkadaşları [3] tarafından, makine öğrenmesi ile film başarı tahmini üzerine bir çalışma yapılmıştır. Özelliklerin film sınıflandırması üzerindeki etkisi ölçülmeye çalışılmıştır. Çalışmalarında karar ağaçları, K-NN, Rastgele Ağaç (Random Forest), c4.5, c5.0 ve Boosting algoritmaları kullanılmıştır. IMDB web sitesi ve Facebook web sitesinden alınan veri seti üzerinde farklı makine öğrenmesi algoritmaları kullanılarak karşılaştırmalı bir çalışma yapılmıştır. Veri kümeleri IMDB web sitesi üzerinden alınmıştır. Veri setindeki toplam 5043 filmin 28 farklı değişkeni ile üzerinde çalışılmıştır. Bütçe, filmin IMDB linki, en boy oranı, içerik değerlendirmesi, ülke, dil, film çıkış tarihi, film türü, renk, posterdeki yüz sayısı, oyuncuların toplam Facebook beğeni sayısı, film için oy kullananların sayısı, film için IMDB' deki eleştirel yorum sayısı, filme oy verenlerin sayısı, filmin ismi, dakika bazında film süresi, yapımcı ismi, yapımcının Facebook' ta ki beğeni sayısı gibi özellikler üzerinde çalışmalar yapılmıştır. IMDB puanında kullanıcı oylarının en önemli faktör olduğu, filmin üretildiği ülkenin ise IMDB puanını belirlemede en az önemli faktör olduğunu gözlemlenmiştir.

Saraee ve arkadaşları [4] tarafından, film puanlarının analizi ve tahmini için bir veri madenciliği yaklaşımı üzerine çalışmalar yapılmıştır. Çalışmalarında 390.000'den fazla film, televizyon dizisi ve video oyunu için ücretsiz, kullanıcı tarafından yönetilen, çevrimiçi bir kaynak olan ve başlık, tür, gişe rekorları kıran gibi bilgileri içeren IMDB' den alınmıştır. Büyük bütçeli filmlerin düşük bütçeli filmlerden daha popüler olup olmadığını keşfetmek, "altın çağ" kanıtlanabilir ve belirli bir aktörün veya aktrisin bir filmin başarılı olmasına yardımcı olup olmayacağı ile ilgili çalışmalar yapılmıştır. Kaynak verilerin formatı nedeniyle IMDB' de veri madenciliği yapmanın zor olduğu gözlemlenmiştir. Filmde yer alan yönetmen ve aktörlerin/aktrislerin film başarısında çok etkili olduğu gözlemlenmiştir.

Lash ve Zhao [5] tarafından film yatırımlarıyla ilgili kararları tahmin etmenin bir yolu önerilmiştir. Film prodüksiyonunun ilk aşamalarında film yatırım kararlarını desteklemek

için film karlılığı tahmin edilmeye çalışılmıştır. Çeşitli kaynaklardan gelen verilerden yararlanarak ve sosyal ağ analizi ve metin madenciliği teknikleri kullanarak "kim" oyuncu kadrosunda, bir filmin "ne" hakkında olduğu, bir filmin "ne zaman" olacağı gibi çeşitli türde özellikleri çıkaran bir sistem önerilmiştir. Bu çalışma, tarihsel bir veri kullanarak film yapımında erken bir yatırım kararı alınmasına yardımcı olmuştur. Bu çalışmada, kar esas olarak gişe geliri üzerinden hesaplanmıştır. Bununla birlikte, birçok film için satılık eşyalar gibi başka gelir kaynakları da gözlemlenmiştir.

Kyuhan Lee ve arkadaşları [6] tarafından, makine öğrenimi teknikleriyle film başarısını tahmin etme ve doğruluk oranlarını arttırma yolları üzerinde çalışma yapılmıştır. Tahmin modelinin performansını iyileştirmek için çoklu yaklaşımları incelenmiştir. Transmedya hikaye anlatımı teorisinden türetilen yeni bir özellik geliştirilmiş ve eklenmiştir. Gişe performansını tahmin etme araştırmasında nadiren benimsenen bir topluluk yaklaşımı kullanılmıştır. Sonuç olarak, önerilen model olan Cinema Ensemble Model (CEM), mevcut makine öğrenimi algoritmalarını kullanan geçmiş çalışmalardan tahmin modellerinden daha iyi performans gösterdiği gözlemlenmiştir.

Hemraj Verma ve Garima Verma [7] tarafından, çeşitli makine öğrenme tekniklerini kullanarak tahmin modellerinin karşılaştırmalı bir analizi yapılmıştır. Modeller, bir filmin vizyona girmeden önce başarılımı yoksa başarısızmı olacağını tahmin etmek için kullanılmıştır. Karşılaştırmalar için Karar Ağacı, Rastgele Orman, Destek Vektör Makinesi, Lojistik Regresyon, uyarlanabilir ağaç güçlendirme ve Yapay Sinir Ağı algoritmaları kullanılmıştır. Modellerde kullanılan ana faktörlerin, başrol oyuncusunun reytingleri, bir filmin IMDB sıralaması, filmin müzik sıralaması ve bir filmin vizyona girmesi planlanan toplam ekran sayısı olduğu gözlemlenmiştir. Çoğu modelin sonuçları %80 - %90 arasında bir değer elde edilmiştir. Rastgele orman ve Lojistik Regresyon teknikleri ile %92 doğruluk oranı elde edilmiştir. Sonuçlara göre bir filmin başarısının en önemli belirleyicileri olarak müzik reytingi, IMDB reytingi ve ekran sayısı olduğu gözlemlenmiştir.

Wenbin Zhang and Steven Skiena [8] tarafından, haber analizi yoluyla film brüt tahminini iyileştirme yöntemleri üzerinde çalışmalar yapılmıştır. Geleneksel film brüt tahminleri için, IMDB den alınan sayısal ve kategorik film verileri kullanılmıştır. İnsanların film brütlerini tahmin etmelerine yardımcı olmak için büyük ölçekli haber analiz sistemleri Lydia tarafından oluşturulan nicel haber verileri kullanılmıştır. İki farklı modeli (regresyon ve k-nn modelleri) analiz ederek, yalnızca haber verilerini kullanan modellerin IMDB verilerini kullananlara benzer performans gösterebildikleri gözlemlenmiştir. Ayrıca IMDB

verileri ile haber verilerinin birleşimini kullanarak daha iyi performans elde edilebileceği belirtilmiştir.

Subramaniyaswamy ve arkadaşları [9] tarafından, çoklu regresyon ve destek vektör makineleri kullanarak film gişe başarısı tahmin edilmeye çalışılmıştır. Veri kümeleri, 2016 yılında yayınlanan filmler için BoxOfficeMojo ve Wikipedia' dan alınan bilgilerle oluşturulmuştur. Fragman görüntüleri YouTube' dan alınmıştır. Açılış tarihi, film ismi, bütçe, yurtiçi brüt, uluslararası brüt, toplam brüt, trailer görünümleri, Wikipedia görünümleri, stüdyo, tür vb. verileri üzerinde çalışılmıştır. Tekrar eden film verileri ve düşük bütçeli filmleri veri setlerinden çıkartılmış ve elde ettikleri toplam 138 film kalmıştır. Bu filmlerin çoğunluğu büyük stüdyolar tarafından yayınlanmıştır. Sonuç olarak SVM yöntemi ile yapılan denemelerde %56.52' lik bir doğruluk oranı elde edilirken, çoklu regresyon sonucunda ise 0.88 olan R değeri elde edilmiştir.

Anand Bhave ve arkadaşları [10] tarafından, film başarısını öngörmede farklı faktörlerin rolleri hakkında çalışma yapılmıştır. Çalışmaları hem klasik hem de sosyal faktörlerin (beklenti ve kullanıcı geri bildirimi) entegrasyonunun ve klasik faktörler arasındaki karşılıklı ilişkinin incelenmesinin daha fazla doğruluğa yol açacağını ileri sürmüştür. Vizyona girmeyen filmlerin genel başarısını, klasik özelliklerin yanı sıra sosyal medya kanalları aracılığıyla kullanıcı beklentisi veya geri bildirimleri de dikkate alınarak doğru bir şekilde tahmin edilebileceği gösterilmiştir. IMDB derecelendirmesi, YouTube görüntüleme sayısı ve bir filmin gösterime girdiği salon sayısı gibi özellikleri göz önüne alınmış, çok değişkenli doğrusal regresyon kullanarak R-kare değeri 0.70 bulunmuştur.

Rijul ve arkadaşları [19] tarafından, IMDB' den bir veri kümesi oluşturulmuştur. IMDB verileri üzerinde detaylı bir analiz verilmiş ve filmlerin IMDB puanını tahmin etmeye çalışılmıştır. Çalışmalarında IMDB puanı, yönetmen, brüt, bütçe vb. kategorik ve sayısal öznitelikler kullanılmıştır. Deneysel çalışmaları sonucunda film puanı tahmininde en başarılı olan yöntem olarak Rastgele Orman algoritması gözlemlenmiştir.

2. YÖNTEMLER

Bu bölümde önerilen modelde kullanılan istatistiksel analiz teknikleri ve makine öğrenmesi teknikleri anlatılmaktadır.

2.1. Veri Ön İşleme Yöntemleri

2.1.1. Bir sıcak kodlama (One hot encoding)

Makine öğrenme yöntemleri kategorik veriler üzerinde çalışma yürütememektedir. Bu nedenden dolayı çalışmalarda kullanılacak verilerin sayısal verilere dönüştürülmesi gerekmektedir. Bir sıcak kodlama, kategorik değişkenleri ikili (binary) olarak kullanılması anlamına gelmektedir. Bu yöntemin kullanımında kategorik değerlere sahip verilerin tamsayı değerlerine eşlenme işlemi yapılmaktadır [29]. Daha sonra, her bir tamsayı değeri, 1 ile işaretlenmiş tamsayı indeksi dışındaki tüm değerleri sıfır olan bir ikili vektör olarak temsil edilir.

2.1.2. Özellik Ölçeklendirme (Feature Scaling)

Özelliklerin taban ve tavan değerlerinin arasındaki farkın çok büyük olması makine öğrenmesi algoritma sonuçlarını doğruluktan uzaklaştıracaktır. "Giriş değerlerinin her birini kabaca aynı aralıkta tutarak eğim açılımı hızlandırabilir. Bunu önlemenin yolu, giriş değişkenlerinin aralıklarını değiştirmektir ve böylece bunların hepsi aynı olur" [30]. İdeal formülazyon:

$$-1 \leq x(i) \ \leq veya - 0.5 \ \leq x(i) \leq 0.5 \qquad (1)$$

Bunu sağlamak için özellik ölçekleme (feature scaling) yöntemi kullanılır. Özellik ölçeklendirme, giriş değerlerini giriş değişkeninin maksimum değer eksi en düşük değerinden bölerek yeni bir aralığına dönüştürür. Girdi değişkeni sıfırdan yeni bir ortalama değer ile sonuçlanan değişkendir. Özellik ölçeklendirme aşağıdaki denklemle ifade edilir.

$$x_i = \frac{x_i - \mu_i}{s_i} \qquad (2)$$

2.2. İstatistiksel Yöntemler

2.2.1. Korelasyon Matrisi

Korelasyon matrisi [28], farklı değişkenler için korelasyon katsayılarını gösteren bir tablodur. Matris, bir tablodaki tüm olası değer çiftleri arasındaki korelasyonu gösterir. Büyük bir veri kümesini özetlemek ve verilen verilerdeki kalıpları belirlemek ve görselleştirmek için güçlü bir araçtır. Bir korelasyon matrisi, değişkenleri gösteren satır ve sütunlardan oluşur. Tablodaki her hücre korelasyon katsayısını içerir.

Ek olarak, korelasyon matrisi, diğer istatistiksel analiz türleri ile birlikte sıklıkla kullanılır. Örneğin, Çoklu Doğrusal Regresyon modellerinin analizinde yardımcı olabilir. Çoklu doğrusal regresyonda, korelasyon matrisi, bir modeldeki bağımsız değişkenler arasındaki korelasyon katsayılarını belirler.

Çalışmamızda her bir öznitelik ile hedef öznitelik olan film başarısı arasındaki korelasyonu elde etmek için öznitelik korelasyon matrisi kullanılmıştır. Verileri özetlemenin ve veri sahibini, veri analizinin daha verimli bir şekilde gerçekleştirilebileceği şekilde yüksek düzeyde ilişkili niteliklere daha fazla odaklanması için yönlendirmenin yaygın bir yoludur. Çalışmamızda Pearson korelasyon katsayısı kullanılmıştır. Puan 0 ile 1 arasında değişmektedir, 1'e yakın değerler yüksek korelasyonu, 0'a yakın değerler ise düşük korelasyonu temsil etmektedir.

2.2.2. Ki-Kare Testi

Ki-kare ($\chi 2$) testi [11] bir hipotez test yöntemidir. Ki-kare testi, test istatistiği, sıfır hipotezi altında ki-kare dağıtıldığında, özellikle Pearson' ın ki-kare testi ve bunun varyantlarında gerçekleştirilmesi geçerli olan istatistiksel bir testtir. Pearson' ın ki-kare testi, bir beklenmedik durum tablosunun bir veya daha fazla kategorisinde beklenen frekanslar ile gözlemlenen frekanslar arasında istatistiksel olarak anlamlı bir fark olup olmadığını belirlemek için kullanılır. Bu testin standart uygulamalarında, gözlemler birbirini dışlayan sınıflara sınıflandırılır. Popülasyondaki sınıflar arasında fark olmadığına dair sıfır hipotez doğruysa, gözlemlerden hesaplanan test istatistiği $\chi 2$ frekans dağılımını takip eder. Testin amacı, sıfır hipotezinin doğru olduğunu varsayarak gözlemlenen frekansların ne kadar muhtemel olduğunu değerlendirmektir. $\chi 2$ dağılımını izleyen test istatistikleri, gözlemler bağımsız olduğunda ortaya çıkar. Çiftlerin gözlemlerine dayalı olarak bir rasgele değişken çiftinin bağımsızlığının sıfır hipotezini test etmek için $\chi 2$ testleri de vardır.

Belirtilen sıfır hipotezin doğru olup olmadığını tespit etmek için gözlemlenen değerleri beklenen değerlerle karşılaştırmak için kullanılır. Sıfır hipotez, karşılaştırılan veriler arasında hiçbir fark olmadığını belirtir. Bu test için, tanımlanan anlamlılık düzeyinden (0.05) küçük veya buna eşit bir p değeri, gözlemlenen dağılımın beklenen dağılımla aynı olmadığı sonucuna varmak için güçlü bir kanıt olduğunu gösterir. Ayrıca, ki-kare istatistiğinin hesaplanmasında kullanılan veriler rastgele, ham, birbirini dışlayan, bağımsız değişkenlerden alınmış ve yeterince büyük bir örneklemden alınmış olmalıdır. Ki-kare testi aşağıdaki denklemle ifade edilir.

$$x_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (3)$$

2.2.3. Varyans Analizi (ANOVA)

Varyans Analizi (ANOVA) [12], istatistik bilim dalında, grup ortalamaları ve bunlara bağlı olan işlemleri analiz etmek için kullanılan bir istatistiksel modeller koleksiyonudur. Varyans Analizi kullanılmaktayken belirlenmiş bir değişkenin gözlemlenen varyansı farklı değişim kaynaklarına dayandırılabilen varyans bileşenine ayrılır. "Varyans Analizi" birkaç grubun ortalamalarının birbirine eşit mi eşit değil mi olduğunu sınamak için bir çıkarımsal istatistik sınaması olur ve bu sınama iki-grup için yapılan t-test sınamasını çoklu-gruplar için genelleştirir. Eğer, çoklu değişkenli analiz için birbiri arkasından çoklu iki-örneklemli-t-sınaması yapmak istenirse bunun I. tip hata yapma olasılığını artırma sonucu doğurduğu aşikardır.

ANOVA analizi için öncelikle hipotezlerin oluşturulması gerekmektedir. Sıfır hipotezi, ilişkinin yokluğunu ifade eder. Sıfır hipotezi reddedilirse ilişkinin/farkın varlığının olduğu ifade edilir. ANOVA analizi sonuçlarına göre F istatistiği bulunur. P değeri, F istatistiği ve F dağılımı kullanılarak bulunur. Eğer p değeri 0.05’ ten küçükse tüm ortalamaların eşit olduğu sıfır hipotezi reddedilir [33].

Varyans Analizi [12], bir veri setinde bulunan gözlemlenen toplu değişkenliği iki bölüme ayıran istatistikte kullanılan bir analiz aracıdır: sistematik faktörler ve rastgele faktörler. Sistematik faktörlerin, verilen veri seti üzerinde istatistiksel bir etkisi varken, rastgele faktörlerin yoktur. Analistler, bir regresyon çalışmasında bağımsız değişkenlerin bağımlı değişken üzerindeki etkisini belirlemek için ANOVA testini kullanır. Çalışmamızda iki bağımsız değişkenin dikkate alındığı üç grup verinin ortalamaları arasında karşılaştırma

yapmak istediğimiz için iki yönlü ANOVA kullanılmıştır. Dikkate alınan değişkenler film başarısı ve geri kalan niteliklerdir. Ayrıca, birden çok bağımlı değişkeni aynı anda değerlendirerek ANOVA analizinin yeteneklerini genişletmek için Çok Değişkenli ANOVA (MANOVA) kullanıldı.

2.3. Sınıflandırma Yöntemleri

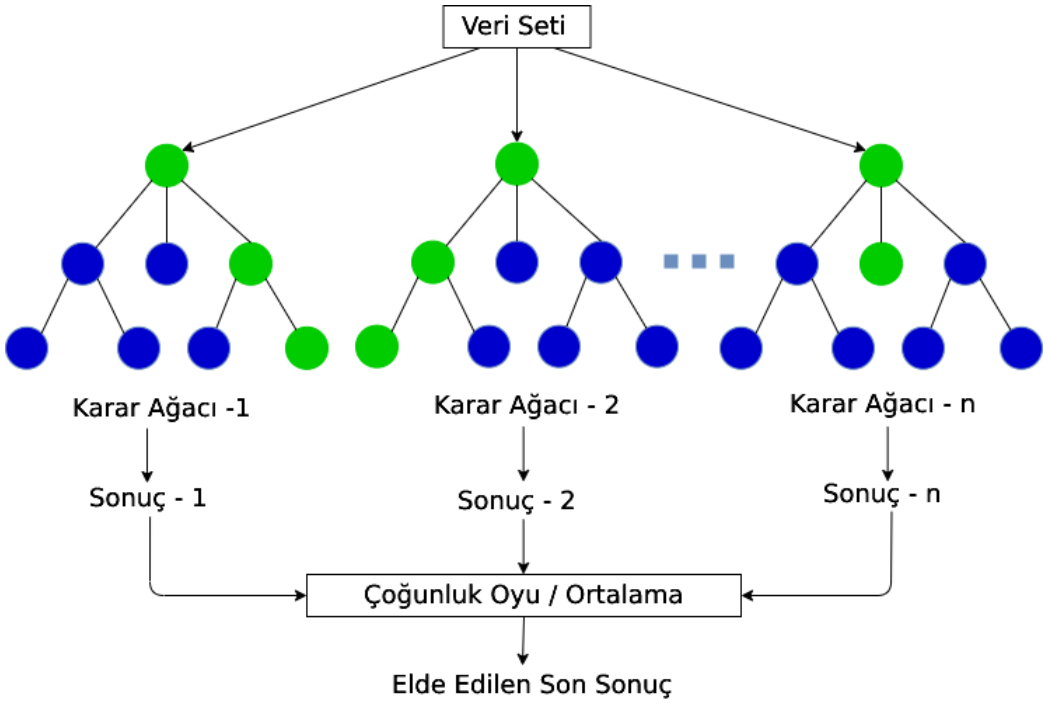
2.3.1. Rastgele Orman (Random Forest)

Rastgele orman, regresyon ve sınıflandırma sorunlarını çözmek için kullanılan bir makine öğrenimi tekniğidir. Karmaşık sorunlara çözümler sağlamak için birçok sınıflandırıcıyı birleştiren bir teknik olan topluluk öğrenmeyi kullanır. Rastgele bir orman algoritması birçok Karar Ağacından oluşur. Rastgele orman algoritması tarafından oluşturulan 'orman', torbalama veya önyükleme toplama yoluyla eğitilir. Torbalama, makine öğrenimi algoritmalarının doğruluğunu artıran bir topluluk meta algoritmasıdır. Rastgele orman algoritması, karar ağaçlarının tahminlerine dayanarak sonucu belirler. Çeşitli ağaçların çıktısının ortalamasını veya ortalamasını alarak tahmin eder. Ağaç sayısının artması sonucun kesinliğini arttırır. Rastgele bir orman, bir Karar Ağacı algoritmasının sınırlamalarını ortadan kaldırır. Veri kümelerinin fazla takılmasını azaltır ve hassasiyeti artırır. Paketlerde çok fazla konfigürasyon gerektirmeden tahminler üretir.

Karar ağaçları, rastgele bir orman algoritmasının yapı taşlarıdır. Karar ağacı, ağaç benzeri bir yapı oluşturan bir karar destek tekniğidir. Karar ağaçlarına genel bir bakış, Rastgele Orman algoritmalarının nasıl çalıştığını anlamamıza yardımcı olacaktır. Bir Karar Ağacı üç bileşenden oluşur: karar düğümleri, yaprak düğümler ve kök düğüm. Bir Karar Ağacı algoritması, bir eğitim veri setini diğer dallara ayrılan dallara böler. Bu dizi, bir yaprak düğüm elde edilene kadar devam eder. Yaprak düğüm daha fazla ayrılamaz. Karar ağacındaki düğümler, sonucu tahmin etmek için kullanılan nitelikleri temsil eder. Karar düğümleri, yapraklara bir bağlantı sağlar.

Rastgele ormanlar veya rastgele karar ağaçları [13], her ağacın bağımsız olarak örneklenen rastgele bir vektörün değerlerine bağlı olduğu ve ormandaki tüm ağaçlar için aynı dağılıma sahip olduğu ağaç tahmincilerinin birleşimidir. Çok sayıda karar ağacının oluşturulmasını gerektiren sınıflandırma, regresyon ve diğer görevler için kullanılan bir topluluk öğrenme yöntemidir. Sınıflandırma görevi için, oluşturulan her ağaç, bir çıktı sınıfı olarak bir sınıfı tahmin eder ve nihai karar, ağaçlardan seçilen çoğunlukla tahmin edilen

sınıflar tarafından verilir. Rastgele ormanlar ve varyantları kara kutu modelleri olarak adlandırılır ve biyoinformatik, finans ve sağlık sistemleri gibi çeşitli araştırma alanlarında uygulanmıştır.

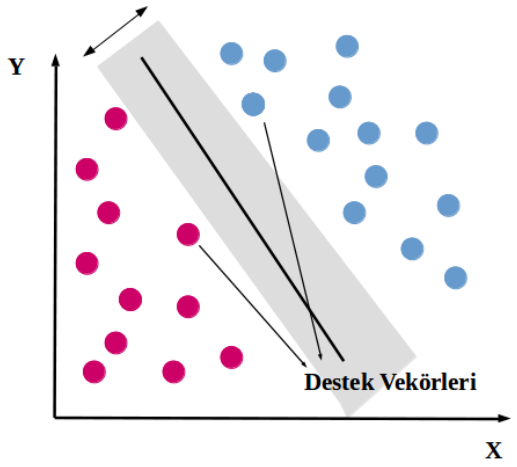


Şekil 1. Karar ağacı şeması [24]

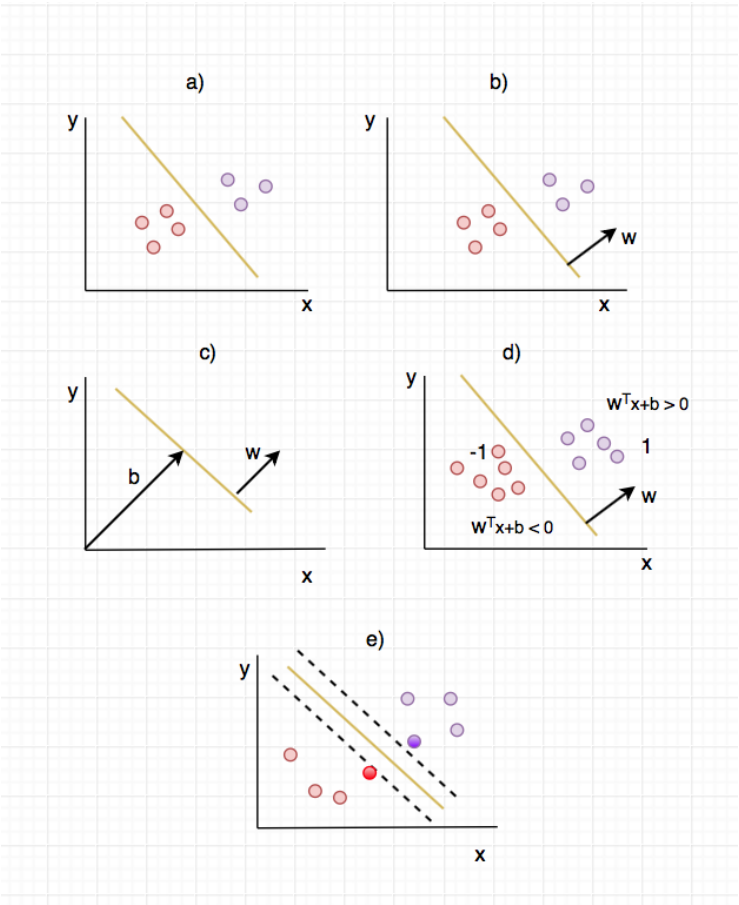
2.3.2. Destek Vektör Makinesi (SVM, Support Vector Machine)

SVM (Support Vector Machine) [14], veri sınıflandırma için kullanılan bir makine öğrenme algoritmasıdır. SVM' nin ana hedefi, bir veri kümesini birden çok kez iki farklı sınıfa en iyi şekilde bölen bir hiperdüzlem bulmaktır (sınıf sayısını eşleştirmek için gerektiği kadar). Destek vektörleri, hiperdüzlemin en yakın veri noktalarıdır, bu nokta hiperdüzlemin tanımlanmasına yardımcı olur, böylece tüm hesaplamalar bu noktalar üzerinden yapılır. Bu hiperdüzlem, iki sınıfı birbirinden ayıran bir kenar boşluğu alanı yaratır. Buradaki hata fonksiyonu, hata azaldıkça kenar boşluğu büyüyecek şekilde tasarlanmıştır. Açıkça bölünen bir hiperdüzlem yoksa, tüm özellik uzayı yeni bir yüksek boyutlu özellik uzayına dönüştürülür. Bu çekirdeklenme olarak bilinir. SVM' ler, küçük ila orta örnek boyutuna sahip temiz veri kümelerinde doğru sonuçlar üretir. Ancak daha büyük veri kümeleriyle uğraşırken, hesaplama maliyetleri çok fazla olabilir ve ayrıca büyük veri kümelerinin

gürültülü doğasına karşı oldukça hassastır. Şekil 2 ve Şekil 3'te iki sınıfı en iyi ayıran destek vektörleri gösterilmiştir.



Şekil 2. Destek vektörleri [25]



Şekil 3. Doğrusal destek vektörleri [26]

Destek vektör makinesi algoritmaları kullanım alanları:

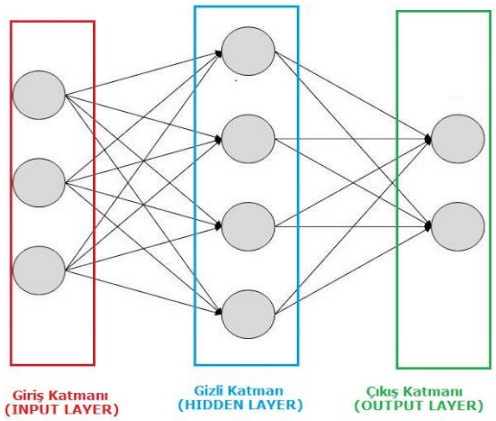
- Hem standart endüktif hem de iletken ayarlarda etiketli eğitim örneklerine olan ihtiyacı önemli ölçüde azaltabileceğinden, metin ve köprü metni sınıflandırmada yardımcı olur.
- "Görüntülerin sınıflandırılması da SVM' leri kullanarak yapılabilir. Deneysel sonuçlar, SVM' lerin geleneksel arama sorgulama şemalarına göre sadece üç veya dört arama sonuçları geri bildiriminden sonra önemli ölçüde daha yüksek bir arama doğruluğu sağladığını göstermektedir. Vapnik tarafından öne sürülen ayrıcalıklı yaklaşımı kullanan SVM modifiye edilmiş bir sürümünü kullananlar da dahil olmak üzere, bu, görüntü bölümlendirme sistemleri için de geçerlidir." [14]
- "SVM algoritması biyolojik ve diğer bilim dallarında yaygın olarak uygulanmıştır. Doğru sınıflandırılmış bileşiklerin %90'ına kadar proteinleri sınıflandırmak için kullanılmıştır. SVM modellerinin yorumlanması için bir mekanizma olarak SVM ağırlıklarına dayanan permütasyon testleri önerilmiştir. Destek vektör makinesi ağırlıkları geçmişte SVM modellerini yorumlamak için kullanılmıştır." [14]

2.3.3. Yapay Sinir Ağı (YSA, Artificial Neural Network)

Yapay sinir ağı (YSA) [15], insan beyninin bilgileri analiz etme ve işleme yönteminin benzerinin simüle edilmesi için tasarlanmış bir bilgi işlem sisteminin parçası olarak düşünülebilir. YSA, yapay zekanın (AI) temeli olarak görülebilir ve insan veya istatistik standartlarına göre zor olan sorunların çözümünde kullanılır. Yapay sinir ağları, fazla sayıya sahip veri setlerinde daha iyi sonuçlar üretmelerini sağlayacak kendi kendine öğrenme yeteneklerine sahiptir. Yapay sinir ağları, bir ağ gibi birbirine bağlı nöron düğümleri ile insan beynine benzer bir şekilde inşa edilmiştir. İnsan beyninde nöron adı verilen yüz milyarlarca hücre vardır. Her nöron, bilgiyi beyne doğru (girdiler) ve beyinden uzağa (çıkışlar) taşıyarak bilgiyi işlemekten sorumlu olan bir hücre gövdesinden oluşur. Yapay sinir ağlarında, öğrenme, hatırlama ve genelleme kabiliyetleri beyindeki biyolojik sinir ağlarının yapısına benzer bir şekilde yönetilir. Örnekler kullanılarak öğrenme işlemi gerçekleştirilir. Öğrenme

işlemi gerçekleştirilirken, giriş çıkış bilgileri ile kurallar belirlenir. YSA, tahmin, kontrol, teşhis, sınıflandırma, veri ilişkilendirme, veri filtreleme ve yorumlama gibi alanlarda aktif olarak kullanılmaktadır. Yapay sinir ağlarının ve ilgili problemlerin özelliklerinin karşılaştırılması ile problemler için en uygun doğru ağın seçimi belirlenir.

Yapay sinir ağları, yapay sinir hücrelerinin birbirine bağlanmasıyla oluşan yapılardır. Yapay sinir ağları, Giriş Katmanı, Ara (Gizli) Katmanlar ve Çıkış Katmanı olmak üzere üç ana katmanda incelenir. İlk olarak bilgiler yapar sinir ağına girdi katmanından iletilir ve ara katmanlarda işlenerek, devamında çıktı katmanına iletilirler. Ağın girdiler için en doğru çıktıları üretebilmesi için ağ ağırlıklarının doğru değerlerinin olması gerekmektedir.



Şekil 4. Yapay sinir ağları katmanları [27]

YSA, eğer tek katmandan oluşuyor ise Tek Katmanlı Sinir Ağı (Single Layer Artificial Neural Network), eğer fazla sayıda nörondan ve gizli katmandan oluşuyorsa Çok Katmanlı Sinir Ağı (Multilayer Artificial Neural Network) denir.

2.4. Regresyon Yöntemleri

2.4.1. Çoklu Doğrusal Regresyon (Multiple Linear Regression)

Regresyon modelleri, gözlemlenen verilere bir çizgi uydurarak değişkenler arasındaki ilişkileri tanımlamak için kullanılır. Regresyon, bağımsız değişken(ler) değiştikçe bağımlı değişkenin nasıl değiştiğini tahmin etmenizi sağlar.

Regresyon analizi, sebep sonuç ilişkisi olan değişkenler arasındaki ilişkiyi tahmin etmeye yönelik istatistiksel bir tekniktir. Tek değişkenli regresyonun ana odak noktası, bağımlı bir değişken ile bir bağımsız değişken arasındaki ilişkiyi analiz etmek ve bağımlı-bağımsız değişken arasındaki doğrusal ilişki denklemini formüle etmektir. Bir bağımlı değişkenli ve birden fazla bağımsız değişkenli regresyon modellerine Çoklu Doğrusal Regresyon denir. Çoklu doğrusal regresyon aşağıdaki denklemle ifade edilir. "ε-SVM regresyonunda, eğitim verisi seti, tahmin değişkenlerini ve gözlemlenen yanıt değerlerini içerir. Amaç, her x eğitim noktası için y_n'den ε'den büyük olmayan bir değerle sapan ve aynı zamanda mümkün olduğu kadar düz olan bir f(x) fonksiyonu bulmaktır" [31].

$$y \leq \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \in \qquad (4)$$

2.4.2. Destek Vektör Makinesi Doğrusal Regresyon (Support Vector Machine Linear Regression)

Sınıflandırma ve regresyon analizi için verileri analiz eden ilişkili öğrenme algoritmalarına sahip Denetimli Makine Öğrenimi Modelleri, Destek Vektör Regresyonu olarak bilinir. SVR, Destek Vektör Makinesi veya SVM konseptine dayalı olarak oluşturulmuştur. Verilerin doğrusal olarak ayrılamadığı durumlarda sınıflandırma problemlerinde veya sınıf atamalarında kullanılabilecek popüler Makine Öğrenimi modellerinden biridir. SVM regresyonu, çekirdek işlevlerine dayandığı için parametrik olmayan bir teknik olarak kabul edilir.

2.4.3. Rastgele Orman Regresyonu (Random Forest Regression)

Karar Ağacı yapı olarak basitçe anlaşılır ve yorumlanabilir bir algoritmadır ve modelin özellikleri öğrenmesi için tek bir ağaç yeterli olmayabilir. Random Forest ise karar vermek için birden fazla karar ağacının nitelik özelliklerini kullanan "Ağaç" tabanlı bir algoritmadır. Karar Ağacı algoritmasının, basit yapısından dolayı aşırı uydurmaya neden olduğu zamanlar olmaktadır ve bu büyük bir dezavantajdır. Karar Ağacı Regresyonu yerine Rastgele Orman Regresyonu uygulanarak bu sorun sınırlandırılabilir. Random Forest algoritması diğer regresyon modelleri ile karşılaştırıldığında daha performanslı ve güvenilirdir [32].

Regresyon görevleri için, tek tek ağaçların ortalama veya ortalama tahmini döndürülür. Rastgele karar ormanları, karar ağaçlarının eğitim setlerine fazla uyma

alışkanlığını düzeltir. Rastgele ormanlar genellikle karar ağaçlarından daha iyi performans gösterir, ancak doğrulukları gradyan destekli ağaçlardan daha düşüktür. Ancak, veri özellikleri performanslarını etkileyebilir.

2.5. Değerlendirme Yöntemleri

2.5.1. Karışıklık Matrisi (Confusion Matrix)

Karışıklık matrisi, tahmin modelinin performansını görselleştirmenin tablo şeklinde bir yoludur. Karışıklık matrisindeki her giriş, sınıfları doğru veya yanlış sınıflandırdığı model tarafından yapılan tahminlerin sayısını gösterir. Karışıklık matrisi çoğu zaman bir ikili sınıflandırma problemi için açıklanır.

- Gerçek Pozitifler (True Positive, TP): Sınıflandırıcının pozitif sınıfı pozitif olarak doğru tahmin ettiği tahminlerin sayısını ifade eder.
- Gerçek Negatifler (True Negative, TN): Sınıflandırıcının negatif sınıfı negatif olarak doğru tahmin ettiği tahminlerin sayısını ifade eder.
- Yanlış Pozitifler (False Positive, FP): Sınıflandırıcının negatif sınıfı pozitif olarak yanlış tahmin ettiği tahminlerin sayısını ifade eder.
- Yanlış Negatifler (False Negative, FN): Sınıflandırıcının pozitif sınıfı negatif olarak yanlış tahmin ettiği tahminlerin sayısını ifade eder.

Makine öğrenimi modeli için değerlendirme kriterleri olarak karışıklık matrisi kullanılabilir. Model için etkili performans ölçütleri sunar. Çalışmada başarısız, orta başarılı ve başarılı sınıfları olduğu için çoklu sınıf sınıflandırması kullanılmıştır.

$$TOPLAM = TP + TN + FP + FN \quad (5)$$

$$GERÇEK\ POZİTİFLER = TP + TN \quad (6)$$

$$GERÇEK\ NEGATİFLER = TN + FN \quad (7)$$

2.5.2. Doğruluk Oranı (Accuracy Rate)

Genel olarak, sınıflayıcının ne sıklıkta doğru tahmin ettiğinin bir ölçüsüdür. Bir makine öğrenimi sınıflandırma algoritmasının doğruluğu, algoritmanın bir veri noktasını ne sıklıkla doğru şekilde sınıflandırdığını ölçmenin bir yoludur. Doğruluk, tüm veri noktalarından doğru tahmin edilen veri noktalarının sayısıdır. Daha resmi olarak, gerçek

pozitiflerin ve gerçek negatiflerin sayısının gerçek pozitiflerin, gerçek negatiflerin, yanlış pozitiflerin ve yanlış negatiflerin sayısına bölünmesiyle tanımlanır. Gerçek pozitif veya gerçek negatif, algoritmanın sırasıyla doğru veya yanlış olarak doğru bir şekilde sınıflandırdığı bir veri noktasıdır. Yanlış pozitif veya yanlış negatif ise, algoritmanın yanlış sınıflandırdığı bir veri noktasıdır.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

2.5.3. Çapraz Doğrulama (Cross-Validation)

Çapraz doğrulama, makine öğrenimi modellerinin becerisini tahmin etmek için kullanılan istatistiksel bir yöntem, sınırlı bir veri örneğinde makine öğrenimi modellerini değerlendirmek için kullanılan bir yeniden örnekleme prosedürüdür. Prosedür, belirli bir veri örneğinin bölüneceği grup sayısını ifade eden k adında tek bir parametreye sahiptir. Bu nedenle, prosedür genellikle k-katlı çapraz doğrulama olarak adlandırılır. k için özel bir değer seçildiğinde, modele referansta k yerine kullanılabilir, örneğin k 10 olarak seçildiğinde 10 kat çapraz doğrulama uygulanacağı anlamına gelmektedir. Anlaşılması kolay ve genellikle basit bir eğitim/test ayrımı gibi diğer yöntemlere göre model becerisinin daha az yanlı veya daha az iyimser bir tahminle sonuçlandığı için popüler bir yöntemdir.

2.5.4. Kök Ortalama Kare Hatası (RMSE, Root Mean Square Error)

Kök Ortalama Kare Hatası (RMSE), artıkların (tahmin hataları) standart sapmasıdır. Artıklar, regresyon çizgisi veri noktalarından ne kadar uzakta olduğunun bir ölçüsüdür. RMSE, bu artıkların ne kadar yayıldığının bir ölçüsüdür. Başka bir deyişle, verilerin en uygun çizgi etrafında ne kadar yoğun olduğunu söyler. Ortalama karekök hata, deneysel sonuçları doğrulamak için klimatoloji, tahmin ve regresyon analizinde yaygın olarak kullanılır. RMSE aşağıdaki denklemle ifade edilir [34].

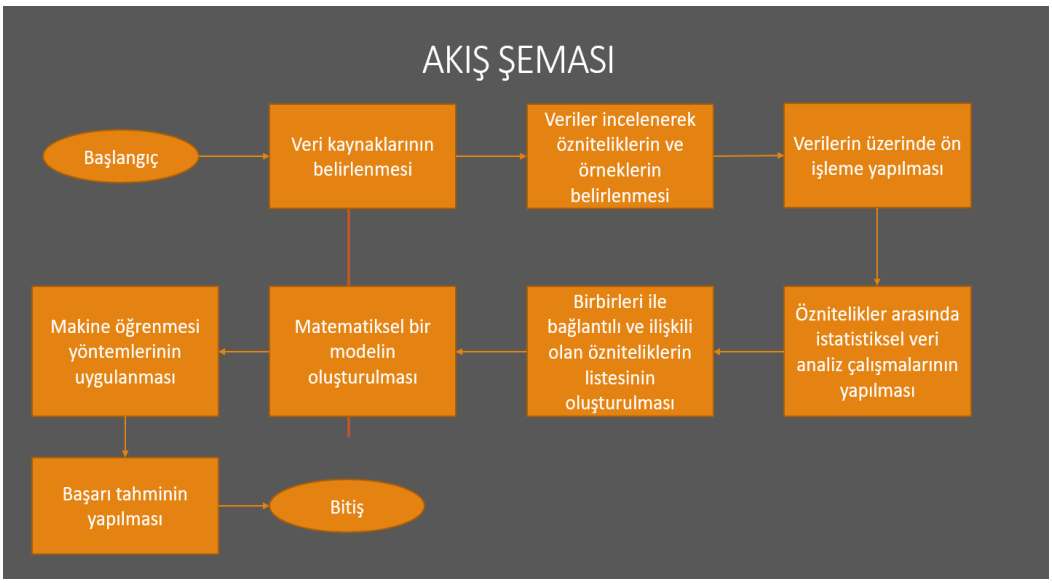
$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}} \quad (9)$$

3. YAPILAN ÇALIŞMALAR

3.1. Veri Setinin Oluşturulması, Ön İşleme Aşamaları

Bu çalışmada, TMDB (The Movie Database) [16] ve OMDB (The Open Movie Database) [17]'den 2000' den 2020' ye kadar yayınlanan 60000 filmden oluşan bir veri seti toplanmıştır. Film verileri üzerinde yapılan temizleme işlemleri (kirli ve boş veriler içeren filmler) sonucunda 4899 film, veri kümesi olarak kullanılmaya karar verilmiştir. Veri seti tarih, tür, dil, sezon gibi çeşitli özniteliklerden oluşmaktadır, özelliklerin IMDB derecelendirmesi kategorik, gişe, bütçe, IMDB oyları, popülerlik, gelir, çalışma süresi özellikleri sayısal değerlerdir. Buna ek olarak, veri setinin örnek dağılımı başarılı sınıf için 1215 örnek, ortalama başarılı için 2663 örnek ve başarısız sınıf için 1021 örnek bulunmaktadır.

OMDB ve TMDB veritabanlarından film verilerinin çekilmesi ve verilerin üzerinde yapılan ön işleme adımları için bir NodeJS uygulaması yazıldı. JavaScript dili kullanılarak düzenlemeler yapıldı. JavaScript metotlarıyla ilgili ön işleme aşamaları rahat bir şekilde uygulanmıştır. Veri kaynaklarına günlük belirli bir sayıda istek atılabildiği için veri setinin oluşturulması uzun süre almıştır. Film verileri, aralarında herhangi bir ilişki barındırmadığından NoSQL bir veritabanı olan MongoDB' de tutuldu.



Şekil 5. Çalışmanın akış şeması

3.1.1. Veri Ön İşleme Adımları

Veri setinde her film, filmin başarısını temsil eden bir IMDB derecesine sahiptir. İki sınıflı bir ikili sınıflandırma probleminden farklı olarak, çok sınıflı sınıflandırma, film başarı bilgisini daha geniş bir şekilde kapsayacaktır. Problemi çok sınıflı bir sınıflandırma problemine dönüştürmek için yapay bir sınıf inşası gerçekleştirilmiştir. Bu nedenle, her film kendi IMDB puanına göre kategorize edilir, örneğin puanları [7-10] aralığında olan filmler "başarılı" sınıfına, [5-6.99] aralığındaki puana sahip filmler "ortalama başarılı" sınıfı, [0-4.99] aralığındaki puana sahip filmler "başarısız" sınıfına atanmıştır. Bu şekilde problemimiz çok sınıflı bir probleme dönüştürülmüştür.

Ayrıca, sınıflandırma algoritmaları uygulanmadan önce eksik değer giderme (missing value removal) ve kategorik özellikleri sayısal özelliklere dönüştürme işlemleri yapılmıştır. Kategorik değerleri sayısal değerlere dönüştürmek için bir sıcak kodlama (one hot encoding) uygulandı. Bu şekilde makine öğrenmesi algoritmalarına sağlanacak genelleştirilmiş bir form elde edilebilir. Son olarak, gişe, gelir ve oy sayıları gibi verilerin bağımsız niteliklerinin aralığını normalleştirmek için bazı niteliklere özellik ölçekleme yaklaşımı uygulandı.

genre_action	genre_adventure	genre_animation	genre_comedy	genre_crime	genre_documentary
0	0	0	0	0	0
0	0	0	0	0	0
0	1	0	0	0	0
1	1	0	0	0	0
0	0	0	1	0	0
0	0	1	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
1	1	0	0	0	0
1	0	0	0	0	0
0	0	0	0	0	1
0	0	0	0	0	0
1	0	0	0	0	0
0	0	0	1	0	0
1	0	0	0	0	0
1	0	0	0	0	0

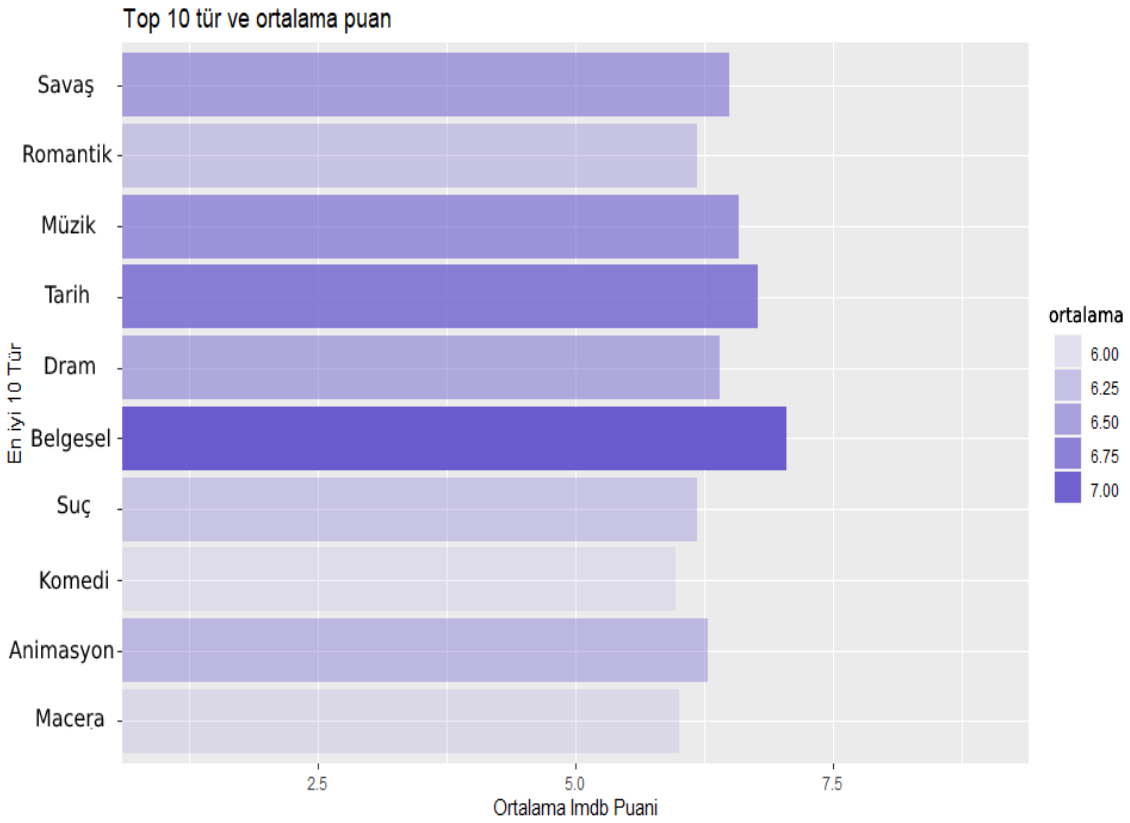
Şekil 6. Film türleri için bir sıcak kodlama yaklaşımı

3.2. Deneysel Sonuçlar

Yapılan çalışmalar sonucunda elde edilen deneysel sonuçlar istatistiksel ve algoritmik olarak ikiye ayrılmaktadır.

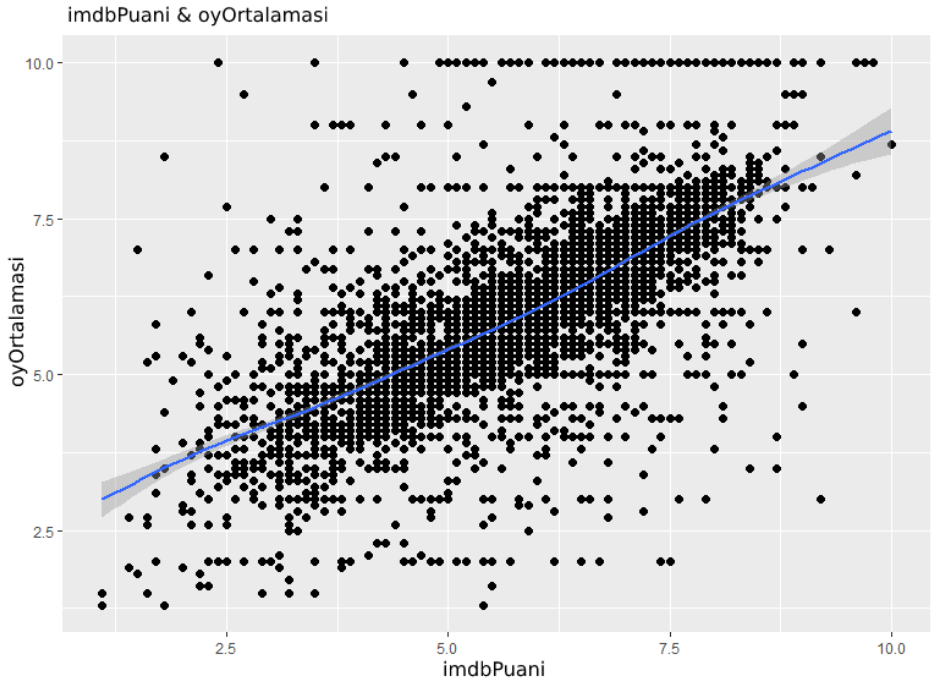
3.2.1. İstatistiksel Sonuçlar

Şekil 7' de en iyi IMDB puan ortalamasına sahip ilk 10 film türleri verilmiştir. En başarılı filmlerin belgesel, tarih ve müzikal film türlerindeki filmlerin olduğu gözlemlenmiştir.

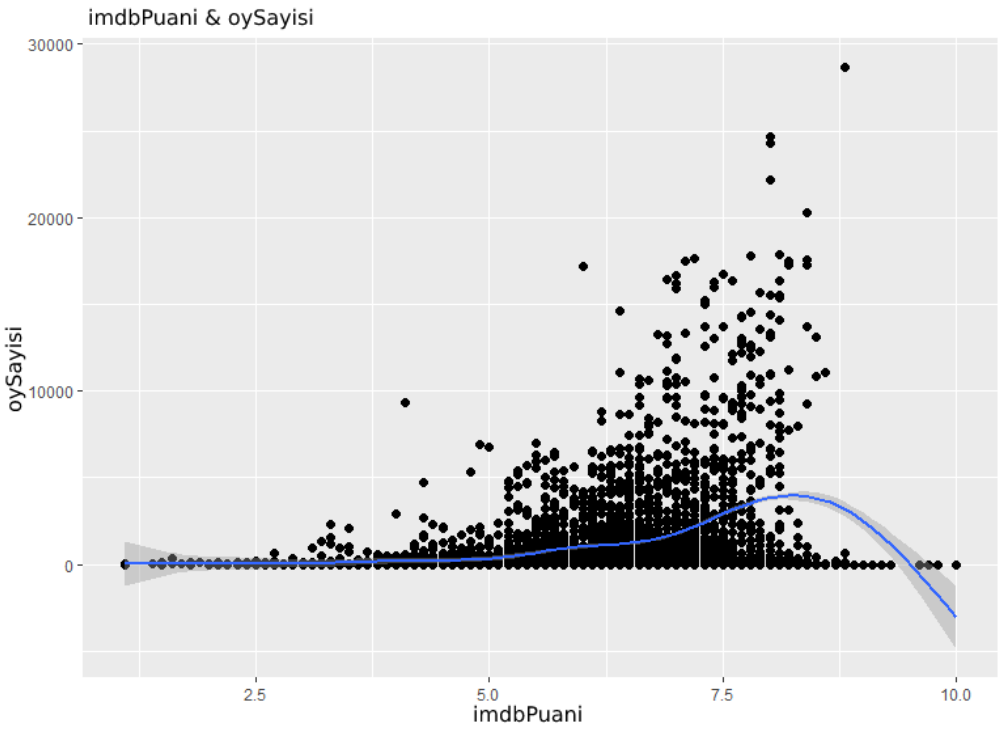


Şekil 7. IMDB puanına göre ilk 10 film türü

Şekil 8’ de IMDB puanı ve sitedeki kullanıcı oyu öznitelikleri kendi aralarında ki dağılımın standart normal dağılım olduğu görülmüştür. Şekil 9’ da ise IMDB puanı ve oy sayısı arasında ki dağılımın negatif çarpık olduğu görülmüştür.

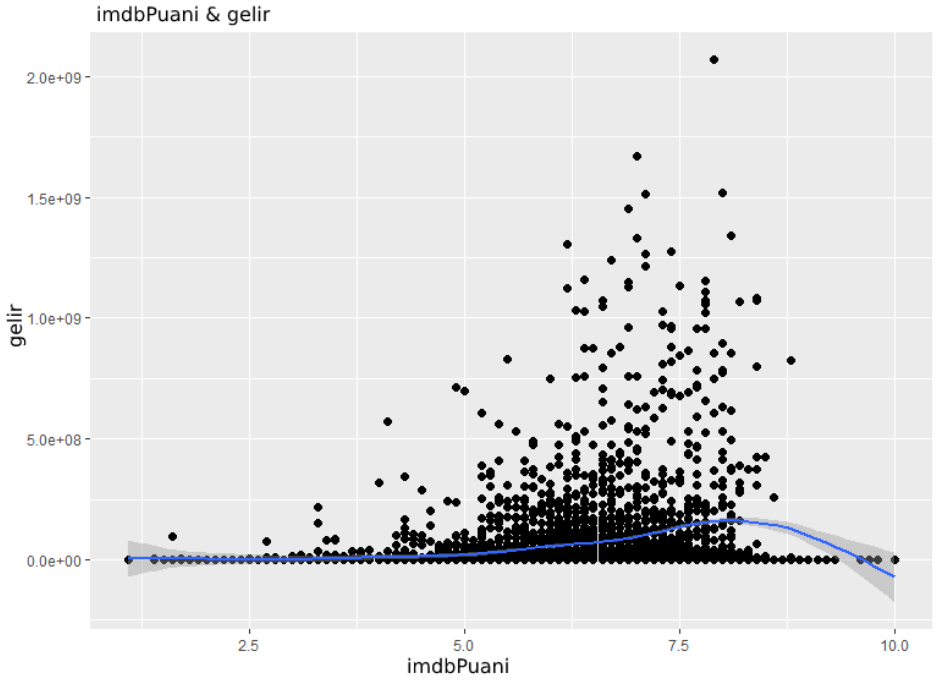


Şekil 8. imdbPuani & oyOrtalamasi dağılım grafiği

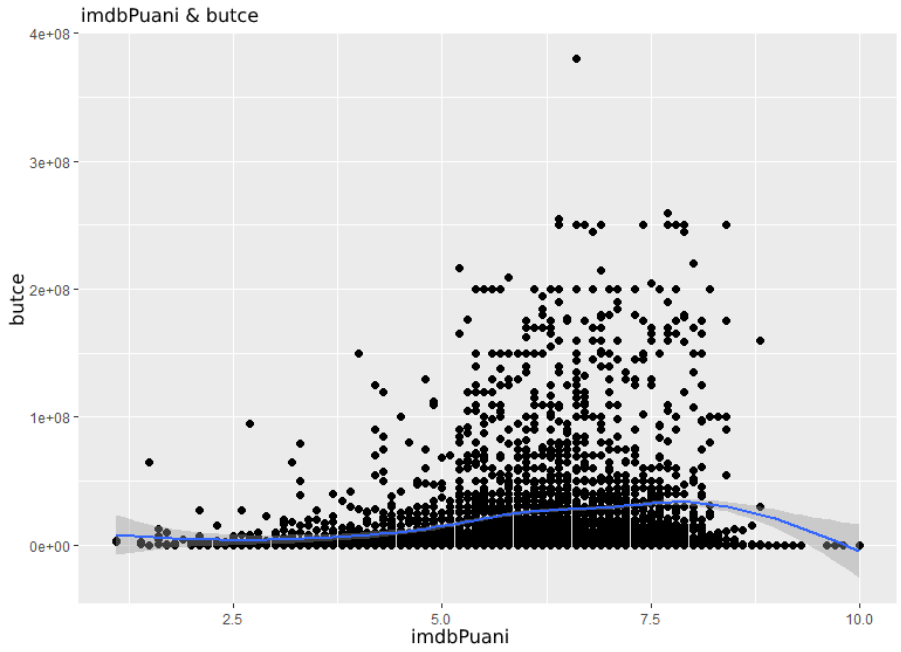


Şekil 9. imdbPuani & oySayisi dağılım grafiği

Şekil 10 ve Şekil 11’ de IMDB puanı ve gelir, IMDB puanı ve bütçe arasındaki dağılımın negatif çarpık olduğu görülmüştür.

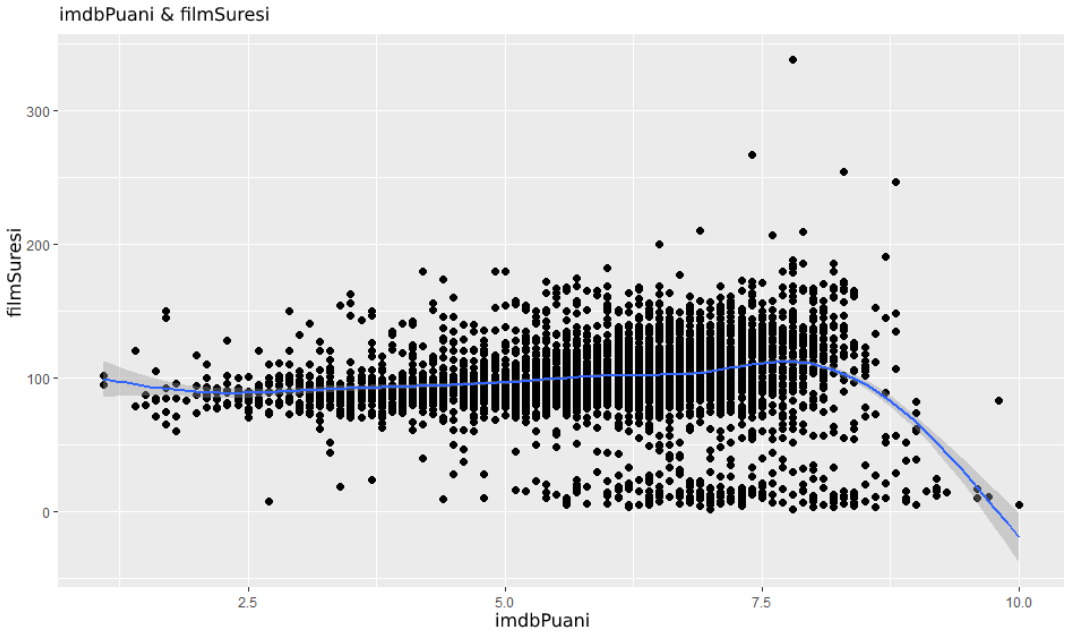


Şekil 10. imdbPuani & gelir dağılım grafiği

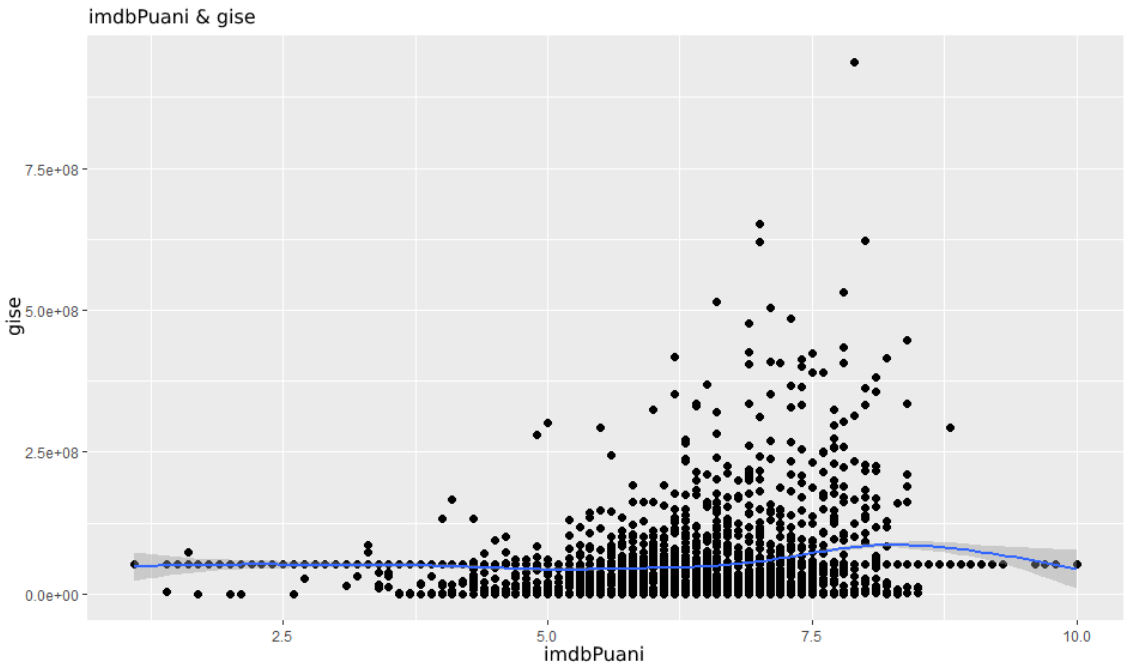


Şekil 11. imdbPuani & butce dağılım grafiği

Şekil 12 ve Şekil 13’ te de IMDB puanı ve film süresi, IMDB puanı ve gişe arasındaki dağılımın negatif çarpık olduğu görülmüştür.

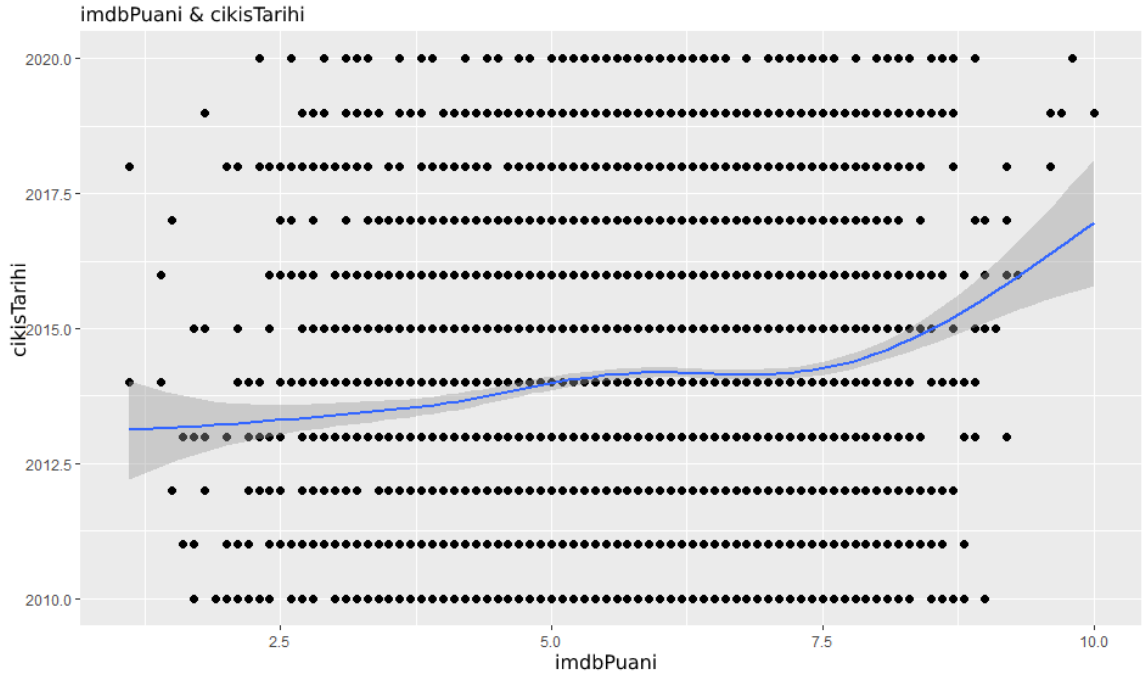


Şekil 12. imdbPuani & filmSuresi dağılım grafiği

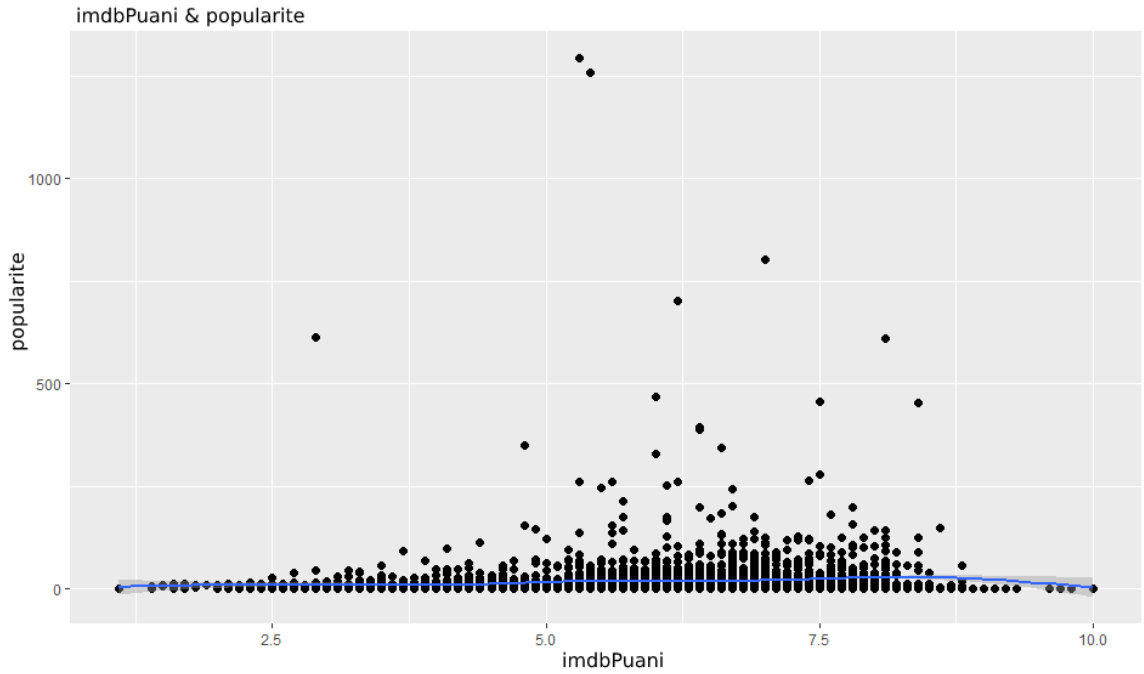


Şekil 13. imdbPuani & gise dağılım grafiği

Şekil 14’ te IMDB puanı ve çıkış yılı arasında pozitif çarpık bir dağılım görülmüştür. Şekil 15’ te IMDB puanı ve film popülerliği (kullanıcı yorumları, takibi) arasında ki dağılımın ikili mod dağılımı olduğu görülmüştür.

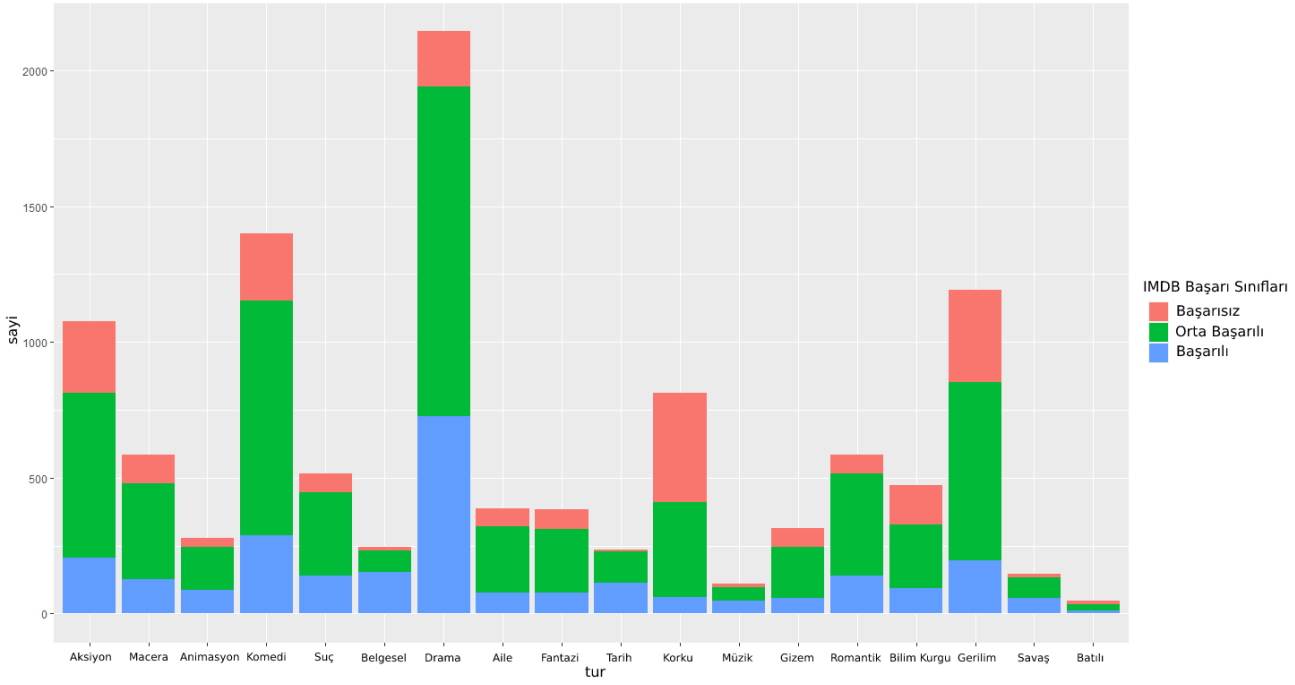


Şekil 14. imdbPuani & cikisTarihi dağılım grafiği



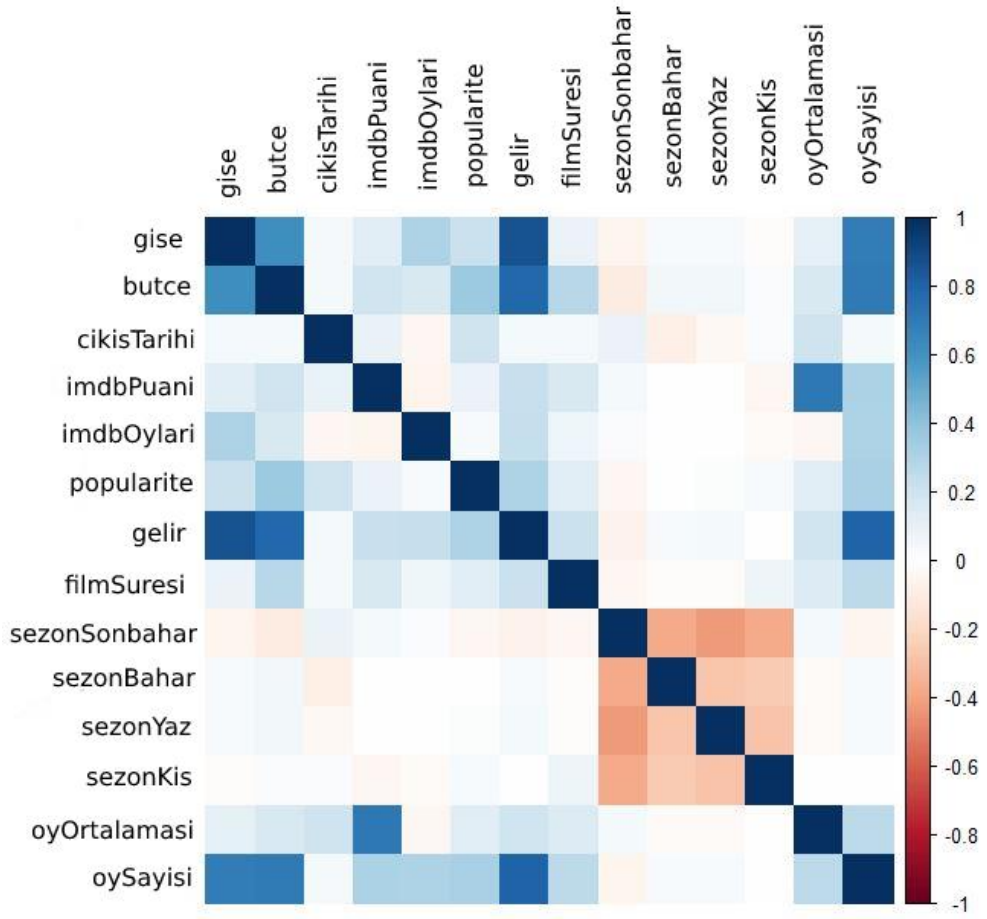
Şekil 15. imdbPuani & popularite dağılım grafiği

Niteliklere göre filmler arasındaki başarı dağılımı hakkında fikir edinmek için bazı dağılım grafikleri elde edildi. Şekil 16, tür özelliğine göre film başarı dağılımını vermektedir. Kırmızı, yeşil ve mavi gibi farklı renklerle verilen sınıf kategorileri sırasıyla başarısız, ortalama ve başarılı film sınıfını temsil etmektedir. Rakamlara göre en fazla başarılı film "drama" ve "komedi" filmlerinde izlenebilirken, "korku" ve "gerilim" filmlerinin çoğu başarısız kategorisinde yer alıyor.



Şekil 16. Filmin başarısına ve türlere göre filmlerin sayısı

Daha sonra Şekil 17' de verilen korelasyon matrisi elde edilmiştir. Matrise göre, "IMDB Puanı" adı olarak verilen film başarısı, "Oy Ortalaması", "Oy Sayısı", "Gelir" ve "Bütçe" öznitelikleri ile yüksek oranda ilişkilidir. Başarı ile yüksek düzeyde ilişkili nitelikler elde edildikten sonra, elde edilen bu niteliklere odaklanılarak aralarındaki ilişkiyi sonuçlandırmak için istatistiksel analizler yapılmıştır. Bunun için ki-kare ($\chi 2$) testi, ANOVA ve MANOVA testleri yapılarak film başarısı üzerinde en büyük etkiye sahip film özelliği araştırılmıştır.



Şekil 17. Film niteliklerinin korelasyon matrisi

Ki-kare ($\chi 2$) testi istatistiği ve karşılık gelen p değerleri Tablo 1' de verilmiştir. Elde edilen p değerlerine göre tamamı 0.05 anlamlılık düzeyinden daha küçük olan her bir öznitelik ile filmin başarısı (imdbPuani) arasındaki ilişkinin şu olduğu sonucuna varabiliriz. Elde edilen p değerlerine göre her bir özniteliğin film başarısı ile olan ilişkisi istatistiksel olarak anlamlıdır fakat sezonYaz özniteliği için elde edilen p değeri 0.05' den büyük olduğu için aynı durum bu öznitelik için söylenememektedir. Buna ek olarak tablodaki "Df" değeri "serbestlik derecelerini" temsil etmektedir.

Tablo 1. Ki-Kare test sonucu

Öznitelikler	**X-squared**	**Df**	**p-value**
oyOrtalamasi	9645.6	156	P < 0.001
oySayisi	8865.4	3222	P < 0.001
gelir	10081	4310	P < 0.001
butce	5728	2154	P < 0.001
filmSuresi	3230	372	P < 0.001
gise	9119.6	3978	P < 0.001
cikisTarihi	125.36	20	P < 0.001
popularite	19113	8300	P < 0.001
sezonKis	14.283	2	P < 0.001
sezonSonbahar	14.283	2	P < 0.001
sezonYaz	4.9467	2	0.0843
sezonBahar	17.391	2	P < 0.001
imdbOylari	4825.6	1512	P < 0.001

İki yönlü ANOVA sonuçları Tablo 2' de verilmiştir. Tablodaki öznitelikler, korelasyon puanlarına ve ki-kare istatistik sonuçlarına göre seçilmiştir. Sonuçlara göre başarılı, ortalama başarılı ve başarılı film gruplarının ortalamaları arasındaki fark istatistiksel olarak anlamlıdır, çünkü elde edilen tüm p değerleri 0.05 anlamlılık düzeyinden küçüktür. Tüm seçilen özniteliklerin olduğu sonucuna varılabilir. Bir filmin başarısını vurgulayan önemli faktörlerdir.

Tablo 2. İki yönlü ANOVA sonuçları

Two Way with IMDB Success Categories	**Df**	**Sum Sq**	**Mean Sq**	**F Value**	**Pr(>F)**
oyOrtalamasi	2	7449	3725	4062	P < 0.001
oySayisi	2	8.003e+09	4.001e+09	592.8	P < 0.001
gelir	2	1.532e+19	7.662e+18	259.1	P < 0.001
butce	2	6.593e+17	3.296e+17	192.1	P < 0.001
filmSuresi	2	224351	112175	163.1	P < 0.001
gise	2	9.171e+17	4.585e+17	129.5	P < 0.001

ANOVA sonuçlarının yanı sıra, birden fazla bağımlı değişkeni aynı anda değerlendirmek için MANOVA da yapılmıştır. "oyOrtalamasi" ve "oySayisi" nitelikleri, bu görev için seçilmiş niteliklerdir. Tablo 3'te Elde edilen sonuçlara göre, bu özelliklerin gruplandırılmasıyla grup ortalamaları arasındaki farkın istatistiksel olarak anlamlı olduğu sonucuna varılabilir.

Tablo 3. Çok değişkenli ANOVA (MANOVA) sonuçları

MANOVA with IMDB Success Categories	**Df**	**Pillai**	**approx F**	**num Df**	**den Df**	**Pr(>F)**
voteAverage / voteCount	2	0.45203	1596.1	4	21864	P < 0.001

3.2.2. Algoritmik Sonuçlar

Tablo 4 ve Tablo 5' te Rastgele Orman ve Destek Vektör Makinesi yöntemlerine göre karışıklık matrisleri verilmiştir. 0 başarısız, 1 orta başarılı, 2 başarılı sınıfı belirtmektedir.

Tablo 4. RF karışıklık matrisi

RF (Karışıklık Matrisi)	0	1	2
0	197	7	0
1	1	531	1
2	0	10	233

Tablo 5. SVM karışıklık matrisi

SVM (Karışıklık Matrisi)	0	1	2
0	204	0	0
1	0	532	1
2	0	2	241

Tablo 6 ve 7' de Rastgele Orman ve Destek Vektör Makinesi yöntemlerinden elde edilen sınıflandırma performans sonuçları belirtilmiştir. Her başarı sınıfının kendi ortalama doğruluk oranı ve bütün sınıfların hepsinin ortalama doğruluk oranı hesaplanmıştır.

Tablo 6. RF sınıflandırma performansları

RF	Sınıf-1 (Başarısız)	Sınıf-2 (Orta Başarılı)	Sınıf-3 (Başarılı)	Sınıf 1-2-3 (Ortalama)
Doğruluk (Accuracy)	0.96	0.98	0.94	0.96

Tablo 7. SVM sınıflandırma performansları

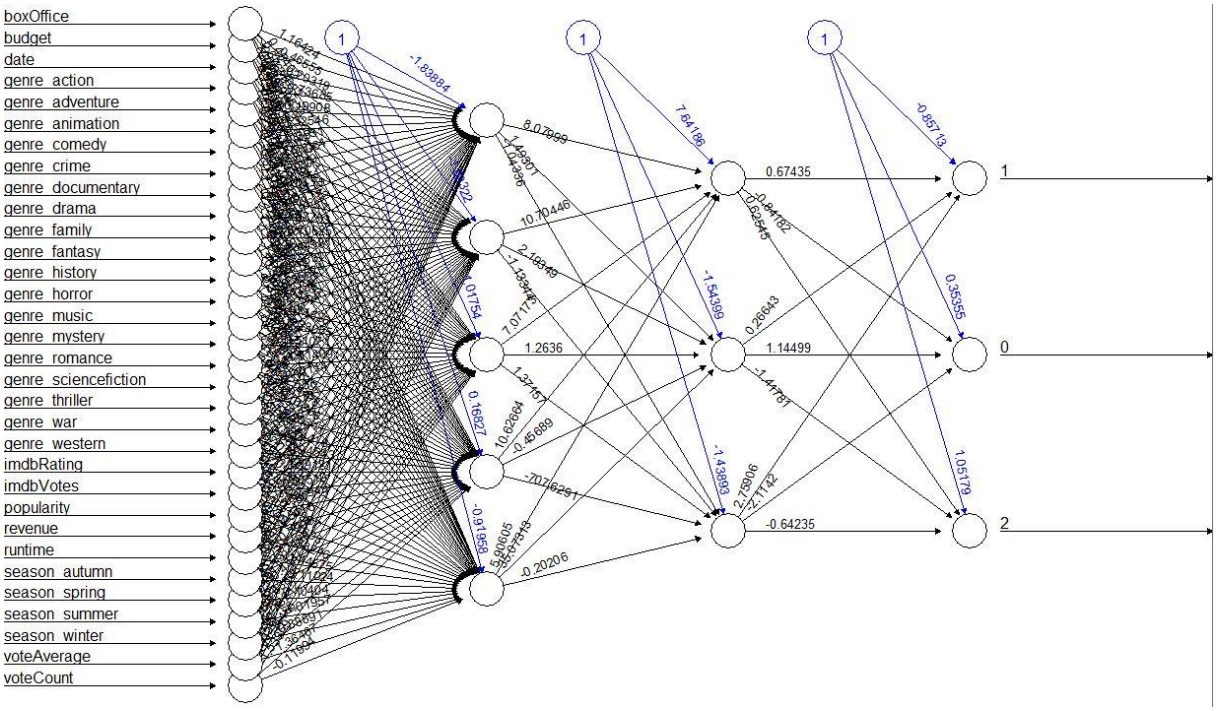
SVM	Sınıf-1 (Başarısız)	Sınıf-2 (Orta Başarılı)	Sınıf-3 (Başarılı)	Sınıf 1-2-3 (Ortalama)
Doğruluk (Accuracy)	0.87	0.88	0.90	0.89

Tablo 8’ den görüldüğü üzere kernel: radio, gamma: 0.1, cost: 60 değerleri ile algoritmayı uyguladığımızda en yüksek doğruluk değerine ulaşılmıştır. Gamma değeri 1 olduğu zaman, cost değerlerinde değişiklik yapılsa bile doğruluk değerinde bir değişiklik olmadığı görülmektedir. Gamma yükseldikçe doğruluk oranının düştüğü, cost yükseldikçe ise doğruluk oranının yükseldiği gözlenmiştir. Doğruluk oranının yükseltilmesi için SVM algoritmasında kullanılan parametrelere dikkat edilmesi gerekmektedir.

Tablo 8. SVM sınıflandırma performansları (kernel, gamma, cost değerleri)

kernel	**gamma**	**cost**	**accuracy (ortalama)**
radial	0.1	10	0.89
radial	0.1	20	0.89
radial	0.1	30	0.89
radial	0.1	40	0.89
radial	0.1	50	0.89
radial	0.1	60	0.89
radial	0.5	10	0.82
radial	0.5	20	0.82
radial	0.5	30	0.82
radial	0.5	40	0.82
radial	0.5	50	0.82
radial	0.5	60	0.82
radial	1	10	0.81
radial	1	20	0.81
radial	1	30	0.81
radial	1	40	0.81
radial	1	50	0.81
radial	1	60	0.81

YSA yöntemini uygulamak için Şekil 18' de ki gibi katmanlı bir yapı hazırlanmıştır. Siyah çizgiler her bir katman arasındaki bağlantıları ve her bağlantıdaki ağırlıkları, mavi çizgiler ise her adımda eklenen bias terimini göstermektedir. Önyargı, doğrusal bir modelin kesişimi olarak düşünülebilir. Tablo 10, her üç sınıfın ortalama doğruluk değerlerini vermektedir. Tablo göre Rastgele Orman yöntemi 0.96 doğruluk oranı ile film başarısını tahmin etmede en başarılı yöntem olmuştur. Ayrıca en düşük doğruluğun SVM yöntemine ait olduğu gözlemlenebilir.



Şekil 18. YSA modelinin her bağlantıdaki ağırlıklarla grafiksel gösterimi

Tablo 9. ANN sınıflandırma performansları

ANN	Sınıf-1 (Başarısız)	Sınıf-2 (Orta Başarılı)	Sınıf-3 (Başarılı)	Sınıf 1-2-3 (Ortalama)
Doğruluk (Accuracy)	0.92	0.96	0.92	0.94

Tablo 10. Makine öğrenmesi algoritmalarının sınıflandırma performansları

	Metot		
	RF	**SVM**	**ANN**
Doğruluk(Accuracy)	0.96	0.89	0.94

Veri seti ile ilgili bazı istatistiksel analiz gözlemleri yapıldıktan sonra üç farklı makine öğrenmesi algoritması kullanılarak film başarı tahmini yapılmıştır. Kullanılan yöntemlerin performansını karşılaştırmak için doğruluk metriği kullanılmıştır. Ayrıca her örneğin veya filmin önerilen modelin eğitiminde ve testinde yer alması sağlanarak daha doğru bir model elde etmek için 10 kat çapraz doğrulama tekniği kullanılmıştır. İkili sınıf problemini çok sınıflı probleme genişletmek için bire karşı hepsi yöntemi kullanıldı. Tüm deneyler RStudio üzerinde R dili kullanılarak yapıldı. Deneysel sonuçlar sonrasında elde edilen verilere göre film başarı tahmininde en etkili makine öğrenmesi algoritması olarak Rastgele Ağaç Sınıflandırma algoritması gözlenmiştir.

Tablo 11. Regresyon yöntemlerinin RMSE değerleri

Metot	**RMSE**
Multi Linear Regression	1.79
Support Vector Linear Regression	1.81
Random Forest Regression	1.77

Regresyon yöntemleri sonucunda elde edilen RMSE değerlerine bakarak Çoklu Doğrusal Regresyon, Destek Vektör Doğrusal Regresyon ve Rastgele Ağaç Regresyon değerlerinin birbirlerine çok yakın sonuçlar ürettiğini ve bunlardan en iyi yöntemin Rastgele Orman Regresyon olduğu gözlenmiştir.

4. TARTIŞMA

Film endüstrisinin hızlı büyümesiyle birlikte başarılı filmlerin piyasaya sürülmesi için büyük yatırımlar yapılmıştır. Film başarısının sadece bir veya iki kritere bağlı olmadığı, film başarısı ile ilgili tüm kriterlerin araştırılması için kapsamlı çalışmaların yapılması gerektiği de bilinen bir gerçektir. Bu nedenle film vizyona girmeden önce bu başarı kriterlerini bilmek, yapımcıların filmlere yatırım yapabilmesi için çok önemlidir.

Bu alanda yapılan çalışmalarda genelde simülasyon verileri kullanıldığı için, ne yazık ki yatırımcılara somut bir yönlendirme yapmanın güçlüğü fark edilmiştir. Bu çalışmada kullanılan tüm film verilerinin gerçek film verileri olması, üzerinde deneysel çalışmalar yapılan bütün niteliklerin istatistiksel analizler sonucunda değerlendirilip seçilmesi, yatırımcılar karşısında daha somut ve daha kararlı bir sonuç elde edilmesinde büyük rol oynamıştır.

Bu alanda yapılan diğer çalışmalarda kullanıldığı belirtilen film oyuncularının ve film yapımcılarının popülerliği gibi özelliklerin ilgili veri kaynaklarından elde edilemeyeceği gözlemlenmiştir. Çalışma doğruluğunu korumak için simülasyon veriler ile ekstra özellikler eklenmesi gibi yöntemlerden kaçınılmalıdır. Filmin posteri, videosu gibi özelliklerin ise film başarısı ile ilişki olmadığı gözlemlenmiş, bu alanda yapılacak çalışmalarda bu özelliklerin kullanılmasının çalışmaya bir katkı sağlamayacağı gözlemlenmiştir.

5. SONUÇ VE ÖNERİLER

Bu çalışmada, farklı istatistiksel analiz teknikleri ve makine öğrenmesi yaklaşımlarını kullanarak bir filmin başarısını tahmin etmek için bir model sunulmaktadır. Temel olarak, bir filmin hangi özelliğinin filmin başarısıyla yüksek oranda ilişkili olduğunu ve filmin başarısını tahmin etmede hangi makine öğrenme tekniğinin daha iyi olduğunu tespit etmeye odaklanılmıştır. Deneysel sonuçlara göre bir filmin başarısının en önemli destekçileri "oyOrtalaması", "oySayısı", "gelir" ve "bütçe" özellikleridir. Buna ek olarak Rastgele Orman algoritmasının, diğer makine öğrenmesi yöntemleri arasında film başarısını tahmin etmede en başarılı olduğu görülmüştür.

Tamamlayıcı bir araç olarak önerilen çalışma, film endüstrisindeki karar vericiler ve yapımcılar için pratik bir çıkarım sağlayabilir. Bir filme yatırım yapma konusunda daha vizyona girmeden karar verebilirler. Gelecekte, daha güvenilir ve sağlam tahmine dayalı model elde etmek için yeni filmler ve nitelikler eklenerek mevcut veri seti genişletilebilir.

Film başarı tahmini problemini çözmek için algoritmik çözümlerin yanında daha kuvvetli ve detaylı şekilde yapılacak istatistiksel analizlerin ciddi bir etkiye sahip olacağı öngörülmektedir. İki veya daha fazla niteliğin birbirleriyle etkileşimleri daha detaylı incelenmeli ve film başarısına katkıları sorgulanabilir.

KAYNAKLAR

[1] Ahmad, J., Duraisamy, P., Yousef, A., & Buckles, B., "Movie success prediction using data mining", IEEE, in 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-4). (2017, July).

[2] Hsu, P. Y., Shen, Y. H., & Xie, X. A., "Predicting movies user ratings with IMDB attributes", in International Conference on Rough Sets and Knowledge Technology (pp. 444-453). Springer, Cham,. (2014, October).

[3] Eker, A. G., Duru, N., Kat, O., & Ildırar, A., "Makine Öğrenmesi ile Film Başarı Tahmini", IEEE, in 2018 3rd International Conference on Computer Science and Engineering (UBMK) (pp. 610-614). (2018, September).

[4] Saraee, M. H., White, S., & Eccleston, J., "A data mining approach to analysis and prediction of movie ratings", Transactions of the Wessex Institute, 343-352. (2004).

[5] Lash, M. T., & Zhao, K., "Early predictions of movie success: The who, what, and when of profitability", Journal of Management Information Systems, 33(3), 874-903. (2016).

[6] Lee, K., Park, J., Kim, I., & Choi, Y., "Predicting movie success with machine learning techniques: ways to improve accuracy", Information Systems Frontiers, 20(3), 577-588. (2018).

[7] Verma, H., & Verma, G., "Prediction model for bollywood movie success: A comparative analysis of performance of supervised machine learning algorithms", The Review of Socionetwork Strategies, 14(1), 1-17. (2020).

[8] Zhang, W., & Skiena, S., "Improving movie gross prediction through news analysis", IEEE, in 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (Vol. 1, pp. 301-304), (2009, September).

[9] Subramaniyaswamy, V., Vaibhav, M. V., Prasad, R. V., & Logesh, R., IEEE, "Predicting movie box office success using multiple regression and SVM", in 2017 international conference on intelligent sustainable systems (ICISS) (pp. 182-186). (2017, December).

[10] Bhave, A., Kulkarni, H., Biramane, V., & Kosamkar, P., "Role of different factors in predicting movie success", IEEE, in 2015 International Conference on Pervasive Computing (ICPC) (pp. 1-4). (2015, January).

[11] Plackett, R. L., Karl, "Pearson and the chi-squared test", International Statistical Review/Revue Internationale de Statistique, 59-72. (1983).

[12] Scheffe, H., "The analysis of variance" (Vol. 72). John Wiley & Sons. (1999).

[13] Breiman, L., "Random forests", Machine learning, 45(1), 5-32. (2001).

[14] Noble, W. S., "What is a Support Vector Machine", Nature biotechnology, 24(12), 1565-1567. (2006).

[15] Wang, S. C., "Artificial neural network", In Interdisciplinary computing in java programming (pp. 81-100). Springer, Boston, MA. (2003).

[16] https://www.themoviedb.org/ Accessed: 20.04.2021

[17] https://www.omdbapi.com/ Accessed: 20.04.2021

[18] Fawcett, T., "An introduction to ROC analysis", Pattern recognition letters, 27(8), 861-874. (2006).

[19] Dhir, R., & Raj, A., "Movie success prediction using machine learning algorithms and their comparison", IEEE, In 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC) (pp. 385-390), (2018, December).

[20] Yim, J., & Hwang, B. Y., “Predicting movie success based on machine learning using twitter”, KIPS transactions on Software and Data Engineering, 3(7), 263-270. (2014).

[21] Quader, N., Gani, M. O., Chaki, D., & Ali, M. H., “A machine learning approach to predict movie box-office success”, IEEE, in 2017 20th International Conference of Computer and Information Technology (ICCIT) (pp. 1-7). (2017, December).

[22] Quader, N., Gani, M. O., & Chaki, D., “Performance evaluation of seven machine learning classification techniques for movie box office success prediction”, IEEE, in 2017 3rd International Conference on Electrical Information and Communication Technology (EICT) (pp. 1-6). (2017, December).

[23] Shankhdhar, A., Agrawal, V., & Rajpoot, V., “Analysing Movie Success Based on Machine Learning Algorithm”, In IOP Conference Series: Materials Science and Engineering (Vol. 1119, No. 1, p. 012008). IOP Publishing. (2021, March).

[24] “Decision tree vs random forest algorithm”, https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm / Accessed: 22.07.2021

[25] “Destek vektör makineleri”, https://veribilimcisi.com/2017/07/19/destek-vektor-makineleri-support-vector-machine / Accessed: 24.07.2021

[26] “Destek vektör makineleri”, https://medium.com/@k.ulgen90/ Makine Öğrenimi Bölüm-4 (Destek Vektör Makineleri) / Accessed: 24.07.2021

[27] “Yapay sinir ağları”, https://mesutpiskin.com/blog/yapay-sinir-agi-derin-ogrenme.html/yapay_sinir_agi/ Accessed: 25.07.2021

[28] “Korelasyon analizi”, https://www.veribilimiokulu.com/korelasyon-analizir-nedir / Accessed: 24.07.2021

[29] “One hot encoding”, https://womaneng.com/one-hot-encoding-nedir-nasil-yapilir / Accessed: 23.07.2021”

[30] “Feature scaling and normalization”, https://veribilimcisi.com/2017/07/18/ozellik-olcekleme-ve-normallestirme-nedir-feature-scaling-and-normalization / Accessed: 24.07.2021

[31] “Support vector machine regression”, https://www.mathworks.com/help/stats/understanding-support-vector-machine-regression.html / Accessed: 24.07.2021

[32] “Random forest regression”, https://towardsdatascience.com/machine-learning-basics-random-forest-regression-be3e1e3bb91a / Accessed: 25.07.2021

[33] “Introduction to ANOVA”, https://online.stat.psu.edu/stat500/book/export/html/479/ Accessed: 20.08.2021

[34] “Root Mean Square Error”, https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/ Accessed: 26.07.2021