T.C.
İSTANBUL AYDIN ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

GÖZETİMLİ ÖĞRENME YÖNTEMLERİNİN KULLANIMI İLE FİLMLERİN IMDB PUANLAMA SİSTEMİNE GÖRE DERECELENDİRİLMESİNİ YAPAN MODELİN OLUŞTURULMASI

YÜKSEK LİSANS TEZİ

Ogtay SAFARALİYEV

Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Programı

EKİM, 2022

T.C.
İSTANBUL AYDIN ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

GÖZETİMLİ ÖĞRENME YÖNTEMLERİNİN KULLANIMI İLE FİLMLERİN IMDB PUANLAMA SİSTEMİNE GÖRE DERECELENDİRİLMESİNİ YAPAN MODELİN OLUŞTURULMASI

YÜKSEK LİSANS TEZİ

Ogtay SAFARALİYEV
(Y1913.010007)

Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Programı

Tez Danışmanı: Dr. Öğr. Üyesi Peri GÜNEŞ

EKİM, 2022

ONAY FORMU

ONUR SÖZÜ

Yüksek Lisans tezi olarak sunduğum "GÖZETİMLİ ÖĞRENME YÖNTEMLERİNİN KULLANIMI İLE FİLMLERİN IMDB PUANLAMA SİSTEMİNE GÖRE DERECELENDİRİLMESİNİ YAPAN MODELİN OLUŞTURULMASI" adlı çalışmanın, tezin proje safhasından sonuçlanmasına kadarki bütün süreçlerde bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurulmaksızın yazıldığını ve yararlandığım eserlerin Kaynakça 'da gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve onurumla beyan ederim. (21/07/2022)

Ogtay SAFARALİYEV

ÖNSÖZ

Bu çalışmada, Gözetimli öğrenme yöntemlerinin kullanımı ile filmlerin IMDB puanlama sistemine göre derecelendirilmesini yapan modelin oluşturmak amaçlanmıştır. Prodüksiyon şirketlerine, yatırımcılara karar verme aşamasında başarı kriteri olarak görülen İMDB puanını tahmin etmek önemli ve kompleks sorun olmuştur. Bu yazıda genetik algoritma, karar ağaçları, SVM, KNN gibi yöntemler kullanılarak bu soruna bakılmıştır.

Bu bilimsel çalışmamın gerçekleştirilmesinde değerli bilgilerini benimle paylaşan, saygıdeğer Dr. Öğr. Üyesi Peri GÜNEŞ hocama teşekkür ederim.

Ekim, 2022 Ogtay SAFARALİYEV

GÖZETİMLİ ÖĞRENME YÖNTEMLERİNİN KULLANIMI İLE FİLMLERİN IMDB PUANLAMA SİSTEMİNE GÖRE DERECELENDİRİLMESİNİ YAPAN MODELİN OLUŞTURULMASI

ÖZET

Film prodüksiyonu en pahalı yatırım alanlarından biridir ve büyük miktarda finansal zarara veya kâr neden olabilir. Bu tür yatırımlardan kaynaklanabilecek büyük kayıplar göz önüne alındığında, yatırımcıların ve yapım şirketlerinin belirli bir filme yatırım yapıp yapmama kararı vermesi kritik önem taşıyor. Bu nedenle, yatırımcı şirketlere karar verme süreçlerinde yardımcı olacak bir model oluşturmak çok tasarrufludur. Konu üzerinden bu zamana kadar farklı yöntemlerin uygulandığı çalışmalara yazıda yer verilmiştir. Google trendler, veri madeciliği gibi tekniklerin olduğu makalelerin içeriği daha kapsamlı şekilde incelenmiştir. Makine öğrenmesi, çeşitli alanlarda karar verme modellerinde sık kullanılmaktadır. Bundan başka öneri sistemlerini oluşturmadaki etkinliğini farklı araştırmalarla kanıtlamıştır. Başarı tahmin etme alanında bu tarz metodların kullanımı için veri toplama, özellik seçimi, veri ön işleme ve veri temizleme gibi işlemler yapılmıştır. Bu çalışmada, bir filmin yapım öncesi başarısının bir göstergesi olan IMDB (İnternet Film Veritabanı) derecelendirmesini kullanarak tahmin etmek için bir Genetik algoritma ile birlikte birkaç Makine öğrenmesi tekniğini (K-Nearest Neighbors, Support Vectors Machine, Decision Tree (C5)) kullanılmıştır. Genetik algoritma diğer algoritmalarla eğitim verilerinin doğruluğu (97,2% oranıyla) ve test verilerinin doğruluğu (90,5%) üzerinden karşılaştırıldıkta daha başarılı performans sergiliyor. Kullanılan KNN ise bu teknikler arasında en başarısız yöntem olmuştur. Filmin başarısı arttıkça gişe hasılatı da artıyor. Böylece gişe hasılatı ile film başarısı arasında bir ilişki olduğu varsayabiliriz. Ancak bu tahmin çok zor, çünkü bir filmi başarılı kılan faktörleri ayırt etmek zor ve insanların görüşlerini tespit etmek çok sübjektif bir konudur. Çalışmanın son bölümünde bu ilişki grafikle gösterilmektedir. Sonuçlar, makine öğreniminin bu alanda faydalı olduğunu ve nispeten iyi performansa sahip başarı tahmin modelleri oluşturmak için GA'nın kullanılabileceğini göstermektedir. Ve son

olarak tezin başka buna benzer çalışmalarda ne tarz katkı sağlayabileceğine değinilmiştir.

Anahtar kelimeler: IMDB, Genetik Algoritmalar, C5, Karar Ağacı, SVM.

CREATING A MODEL THAT GRADES MOVIES ACCORDING TO THE IMDB SCORING SYSTEM USING SUPERVISED LEARNING METHODS

ABSTRACT

Film production is one of the most expensive investment areas and can result in substantial financial loss or profit. Given the huge losses that can result from such investments, it is critical for investors and production companies to decide whether to invest in a particular film. Therefore, it is very cost-effective to create a model to assist investor companies in their decision-making processes. Studies on the subject in which different methods have been applied so far are included in the article. The content of the articles with techniques such as Google trends, data mining has been examined more comprehensively. Machine learning is frequently used in decision making models in various fields. Moreover, it has proven its effectiveness in creating recommendation systems with different studies. For the use of such methods in the field of success estimation, data collection, feature selection, data preprocessing and data cleaning have been performed. In this study, several Machine Learning methods (K-Nearest Neighbors, Support Vectors Machine, Decision Tree (C5)) along with a (GA) Genetic Algorithm are used to predict a movie using IMDB (Internet Movie Database) rating, which is an indicator of pre-production success. The genetic algorithm performs better when compared with other algorithms on the accuracy of the training data (97.2%) and the accuracy of the test data (90.5%). The KNN used was the most unsuccessful among these techniques. As the success of the movie increases, so does the box office revenue. Thus, we can assume that there is a relationship between the box office revenue and the success of the movie. However, this estimate is very difficult, because it is difficult to discern the factors that make a film successful, and it is very subjective to determine people's opinions. In the last part of the study, this relationship is shown graphically. The results show that machine learning is useful in this area and GA can be used to build success prediction models with relatively good performance. And finally, it has been mentioned how the thesis can contribute to other similar studies.

Keywords: IMDB, Genetic Algorithms, C5, Decision Tree, SVM.

İÇİNDEKİLER

KISALTMALAR LİSTESİ

C4.5 : Karar ağacı

GA : Genetik algoritma

İMDB : İnternet Film Veritabanı

KNN : K-En Yakın Komşu

ML : Makine öğrenmesi

SL : Denetimli Öğrenme

SVM : Destek Vektör Makinesi

UL : Denetimsiz öğrenme

Weka : Waikato Environment for Knowledge Analysis

ÇİZELGELER LİSTESİ

ŞEKİLLER LİSTESİ

I. GİRİŞ

Sinema sektörü olarak da bilinen film sektörü, eğlence alanında hakim bir sektör haline gelmiştir. Bu nedenle, belirli bir şehirde aynı anda birden fazla film sinemalarda olabilir. Ancak insanların hayatlarında yoğun, yoğun programları vardır. Bu yüzden her filmi izlemek için yeterli zamanları yok. Ayrıca, yüksek yaşam maliyeti nedeniyle insanlar sinemalarda tüm filmleri izleyemezler. Bu yüzden insanlar her zaman sinemada izlemeye değer daha iyi filmler bulmaya çalışırlar. Bu nedenle, diğer insanlardan öneriler ve spoiler içermeyen incelemeler aramaya çalışırlar (Tuna, 2013: 8).

İnternet Film Veritabanına (IMDB) göre, her yıl dünya çapında 20 binden fazla film üretiliyor, dağıtılıyor ve izleniyor. Bu rakamlar, bu endüstrinin pazar üzerindeki büyük etkisini göstermektedir. Filmlerin kültürel ve sosyolojik etkisinin yanı sıra, BoxOfficeMojo'ya göre sinema filmi, her yıl ortalama 10 milyar dolarlık büyüme ile günümüzde iş piyasasının önemli bir bölümünü işgal ediyor (Im ve Nguyen, 2011; Maharshi, vd. 2017: 631).

Film endüstrisinin günümüzde önemi, sadece üretilen paraya değil, aynı zamanda bu sektörle ilgilenen çok sayıda insana da bağlıdır. Bu faktörler, üretim şirketlerini kendi ülkelerindeki büyük ekonomik oyuncular ve genel olarak uluslararası ekonomideki önemli kuruluşlar arasında yapmaktadır (Sangkil, vd. 2010).

Basitçe söylemek gerekirse, bir filmi benimsemek, şirketin bu film başarısına bahse girdiği anlamına gelir. İzleyiciye göre başarılı bir film internette, örneğin IMDB'de yüksek reytinge sahip bir filmdir. Bu derecelendirmeler genellikle bir kişinin sonunda filmi izleyip izlemeyeceğini belirleyen ana faktördür. Bu bakış açısından, veri analistleri bir filmin başarısını belirlemek için kullanılabilecek faktörleri inceler. Bunlar çoğunlukla önceden yayınlanmış faktörlerdir. Filmin başarısı, reytingi veya gişe başarısı (Wang, vd. 2020: 32; Adéla, 2017: 16) ile temsil edilir.

A. Amaç

Film zevki çok karmaşık ve kişisel düzeyde sürekli değişiyor, ancak genel düzeyde daha tutarlı ve daha az değişken. Bu, Makine Öğreniminin yardımıyla bu tutarlılığı inceleme ve tahminler oluşturma fırsatını açar. Bu araştırmanın temel amacı, film reytinglerini daha vizyona girmeden tahmin edebilecek bir model geliştirmektir. IMDB'de filmlerle ilgili yönetmen, yazar, oyuncu vb. gibi birçok veri bulunmaktadır (Krishma ve Amit, 2019: 696). Bu doğrudan faktörlerin dışında, farklı filmler arasındaki ilişkiler (örn. yönetmenin önceki filmleri, oyuncuların/oyuncuların önceki filmleri, aynı filmin önceki bölümleri) film vb.) bir filmin reytingini dolaylı olarak da etkileyebilir. Dolayısıyla bu projenin hedefi, önemli özellikleri belirlemek için özellik seçimi ve özellik çıkarma tekniklerini kullanmak ve yeni filmleri 10 üzerinden derecelendirme kategorilerine ayırabilen bir sınıflandırıcı geliştirmektir.

Bu projenin kapsamı şunları içerir:

- Veri kümesini temizlemek ve geliştirmek
- Sınıflandırma algoritmalarını karşılaştırma
- Önemli özellikleri belirlemek için özellikleri seçmek ve çıkarmak için verileri analiz etme

B. Tanımlar ve Kısaltmalar

1. Makine Öğrenmesi Sınıflandırma Yöntemleri

Makine öğrenimi, eğitim verileri olarak da adlandırılan mevcut verileri kullanarak belirli bir görevdeki performansı artırmaya yönelik bir bilgisayar bilimi alanıdır. Amaç, bu geçmiş deneyimlerden çıkarımlar yapmak için matematiksel modeller oluşturmaktır. Makine öğrenimi, bu modelleri oluşturmak için istatistik yöntemlerini yoğun olarak kullanır. Başka bir deyişle, makine öğrenmesi, eğitim verilerini kullanarak bu matematiksel modellerin parametrelerinin optimizasyon işlemlerini içerir. Çoğu zaman makine öğrenimi veri madenciliği ile ilişkilidir. Veri madenciliği, büyük veri kümeleri üzerinde makine öğrenmesi yöntemlerinin uygulanmasıdır. Veri madenciliğinin finans sektöründeki sahtekarlıkların tespit edilmesinden emlakçılar için ev fiyatlarının tahmin edilmesine kadar geniş bir uygulama alanı vardır; tıbbi araştırmalardan otonom araçlara kadar. Film endüstrisi için film başarısını tahmin etmek de bu çeşitli alanlardan biridir.

Bir Machine Learning modeli oluşturmak için öncelikle modelden ne beklendiği veya hangi problemi çözmeye çalıştığı belirlenmelidir. Modeli eğitmek için kullanılan verileri anlamak, hedefi bilmek için çok önemlidir. Beş tür sorun genellikle bu gruplardan birine girer:

1. Sınıflandırma Problemi: Çıktının sınırlı sayıda grup veya bazen bir sayı olarak sınıflandırılması gerektiğinde.

2. Anormallik Tespit Problemi: Model, bir tür fenomeni veya sayıyı izler, daha sonra anormallikleri tespit etmek için kalıpları öğrenir.

3. Regresyon Problemi: Çıktı sayısal ve süreklidir, çoğu zaman trend grafiklerinde temsil edilir, amaçları genellikle azalan getirilerden kaçınmak veya karı artırmaktır.

4. Kümeleme Problemi: Bir sınıflandırma problemine benzer ancak kümeler oluşturmaya çalışmak için örüntüler aradığı bir denetimsiz öğrenme şeklidir. Yeni veriler yapı kümelerine gider.

5. Pekiştirme Problemi: Kararların daha önceki deneyimlere dayalı olarak verilmesi gerektiğinde, genellikle bir ortamda öğrenilir. Doğru kararlar için "ödüllendirilmek" ve bazen yanlış kararlar için "cezalandırılmak" için doğru kararların ne olduğunu bilmek için deneme yanılmaya bağlıdır.

Makine Öğrenimi algoritmaları iki ana kategoriye ayrılmaktadır:

a. Denetimli Öğrenme: Makine Öğrenimi algoritmalarının büyük çoğunluğunu içerir. Denetlenen olarak adlandırılırlar çünkü çözmeye çalıştıkları problemin çözümünü kısmen gerektirirler. Denetimli Öğrenmede (SL) x girdi değişkenlerini ve Y çıktı değişkenini şu şekilde bulabiliriz:

$$Y = f(x) \text{ (Denklem 1)}$$

SL'deki amaç, f eşleme fonksiyonunu o kadar iyi tahmin etmektir ki, yeni veriler x ile istendiğinde, algoritma, karşılık gelen Y çıktısını mükemmel bir şekilde tahmin edebilir. Denetimli Öğrenme algoritmaları ve iki ana kategoride gruplandırılmıştır:

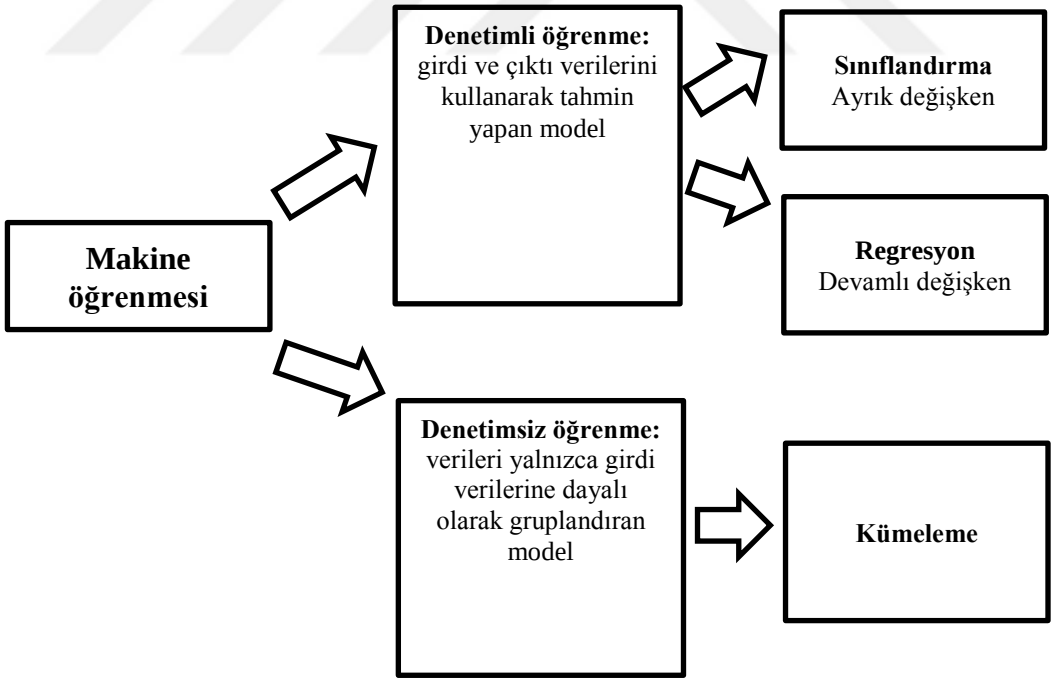
- Sınıflandırma algoritmaları: çıktı değişkeni, \kedi"/\köpek" veya \hastalık"/\hastalık yok" gibi bir kategoridir.

- Regresyon algoritmaları: çıktı değişkeni, \dolar" veya \ağırlık" gibi gerçek bir değerdir

b. Denetimsiz Öğrenme: Yalnızca x girdi verilerinin mevcut olduğu ve karşılık gelen Y çıktı değişkenlerinin olmadığı sorunları ifade eder. Denetimsiz Öğrenmenin (UL) amacı, veriler hakkında daha fazla bilgi edinmek için temel yapıyı veya verilerdeki dağılımı modellemektir. UL'ye bir örnek, hayvan resimlerinden oluşan bir veri tabanı verildiğinde, kaç farklı hayvan olduğunu tanımlayan ve her hayvanın tüm resimlerinden oluşan gruplar oluşturan bir algoritma olabilir. Bu durumda, algoritma tüm köpek resimlerini bir araya getirecek, ancak o gruba hangi etiketi koyacağına dair hiçbir fikri olmayacaktı. Başka bir deyişle, algoritma hangi hayvan olduğunu bilemezdi.

c. Takviyeli Öğrenme: Bu tür ML'de yazılım, kazandığı için ödül ve kaybettiği için ceza alır. Reinforcement Learning (RL) algoritmaları tipik olarak oyunlara ve sahtecilik dedektörlerine uygulanır.

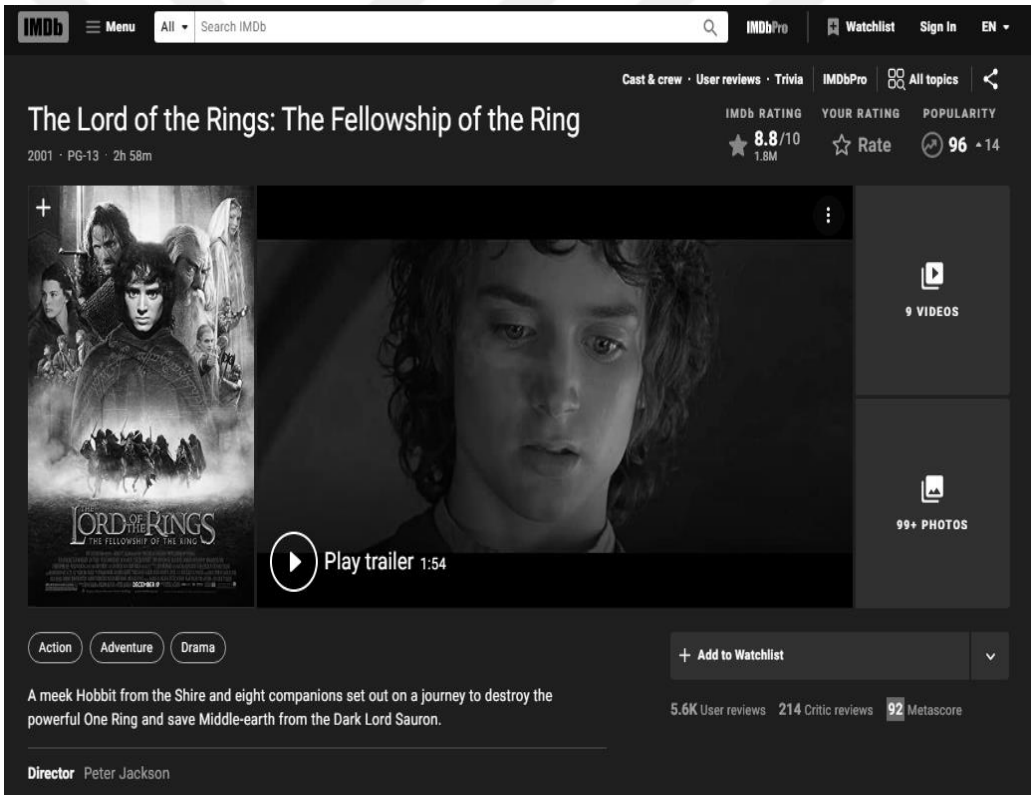
Şekil 1'de, iki ana ML algoritması alt türünü gösteren bir şema gösterilmektedir:



Şekil 1 Makine öğrenmesinin türleri

2. IMDB (Internet Movie Database)

İnternet Film Veritabanı (IMDB) 1990 yılında piyasaya sürüldü ve filmler, televizyon programları, video oyunları, oyuncu ve film ekibi biyografileri, film incelemeleri ve olay özeti hakkında bilgi içeren en büyük çevrimiçi veritabanlarından biridir (Jaiswal, vd. 2020; Wernard ve Sander, 2015: 123). IMDB'nin veri tabanında 4,2 milyon unvan ve 7,8 milyon kişilik kaydı ile birlikte 75 milyon kayıtlı kullanıcısı bulunmaktadır (Basuroy, vd. 2003: 109). Amazon Inc'in bir yan kuruluşu olarak IMDB, gelirini reklamlar, ortaklıklar ve lisanslama yoluyla elde eder. Kayıtlı kullanıcılar, platforma gönüllü olarak içerik katkıda bulunabilir. Ancak, gönderilen veriler platformda gösterilmeden önce bir dizi tutarlılık kontrolünden geçer ve katkıda bulunanlar verileri anında ekleyemez, silemez veya değiştiremez, bu da onu Wikipedia.com gibi diğer kullanıcı katkılı web sitelerinden farklı kılar.



Şekil 2 IMDB Ana Sayfası

İnternet Film Veritabanı (IMDB), filmler (uzun ve kısa), TV dizileri, video oyunları vb. ile ilgili bilgileri aramak için en büyük çevrimiçi veritabanlarından biridir (Simonoff ve Sparrow, 2000). Ayrıca her filmin oyuncu kadrosu, ilgili kişilerin biyografisi, eğlence endüstrisi, film incelemeleri, film hikayesi vb.hakkında bilgi içerir (Chakraborty, vd. 2019; Lopamudra, vd. 2020: 7). İçeriğin çoğu kullanıcı

tarafından oluşturulur ve kullanıcılar filmleri derecelendirerek, incelemeler ve filmlerle ilgili diğer faydalı bilgileri sağlayarak görüşlerini dile getirmeye teşvik edilir.

IMDB veri seti, daha fazla çalışma için ilgi çekici bir unsur getiriyor çünkü insanların çoğu film endüstrisinin ve ilgili aktörlerin farkındadır ve insanlara film veri setinin bir görselleştirmesi sunulduğunda, çoğu onlara bakmaya çalışacaktır. En sevdiğiniz filmler ve aktörler ve aktörler arasındaki karmaşık ilişkileri keşfedilmektedir. IMDB veri seti çok temiz ve yapılandırılmış bir formatta saklanır ve her bir film ve dökümü hakkında çok zengin bilgiler içerir, böylece geniş bir veri analizi yelpazesine izin verir (Gaurvi ve Anirudh, 2015: 5).

C. Çalışmaya Genel Bakış

Bu tezin geri kalanı şu şekilde organize edilmiştir: Bölüm 2'de, filmin reyting tahmini ile ilgili önceki çalışmaları inceliyoruz. Bölüm 3'te verileri, veri madenciliğini ve işlenmesi hakkında yazı bulunmaktadır. Ek olarak bu çalışmada kullanılacak çeşitli makine öğrenimi tekniklerini tanıtılmaktadır. Bölüm 4'te uygulanan yöntemlerin analizi yapılıyor. Deney ve sonuçlar de bu bölümde sunulmaktadır. Son olarak, Bölüm 5'te sonuca varıyoruz ve gelecekteki çalışmalar için kapıyı açıyoruz.

II. LİTERATÜR ARAŞTIRMASI

A. IMDB Verilerine Dayalı Film Başarısını Tahmin Etme

Araştırmalarında Nithin V.R., Pranav M., Sarath Babu P.B. ve Lijiya A., IMDB verilerine dayanarak film başarısını ve IMDB puanını tahmin etmek için bir model geliştirdi. 2000'den 2012'ye kadar Amerika Birleşik Devletleri'nde ve İngilizce olarak vizyona giren filmler için tahmin yaptılar. Başlangıçta IMDB'den veri seti aldılar. Gişe detayları hakkında bilgisi olmayan filmleri kaldırdılar. Wikipedia ve Rotten Tomatoes'dan veri setinin eksik veri alanlarını doldurdular. Veri seti hem nominal hem de sayısal özniteliklerin bir koleksiyonu olduğundan, regresyon işlemi için karşılık gelen nominal değerleri sayısal değerlere dönüştürmek zorunda kaldılar. Göz önünde bulundurulan tüm farklı özellikler ile film geliri arasındaki korelasyonu aldılar. Fazlalık ve alakasız özniteliklerden kaçınmak için, öznitelikler arasındaki korelasyonu kendileri aldılar. En iyi özellik alt kümesini seçerken, açgözlü geriye dönük prosedürü kullanmışlardır.

Geliri tahmin etmek için Lineer Regresyon modeli, Lojistik Regresyon modeli, Destek Vektör Makinesi Regresyon modeli (SVM modeli) kullanmışlar ve her bir yöntemin çıktısını karşılaştırmışlardır. Doğrusal regresyon modelinde, standart en küçük kareler doğrusal regresyonunu kullanmışlardır. Lojistik regresyon modeli için, regresyon problemini bir sınıflandırma problemine değiştirmek zorunda kaldılar. Böylece gelirleri bölümlere ayırdılar ve her bölüme film bırakarak bir histogram oluşturdular. Karşılaştırma sonucunda, lineer regresyon modelinin en doğru model olduğunu ve bunun yaklaşık %51 doğru olduğunu, lojistik regresyon modelinin %42.2 ve SVM modelinin %39 doğruluğa sahip olduğunu saptamışlardır. Araştırmanın başında veri setinden 20 özellik belirlemelerine rağmen bütçe, yönetmen, yazar vb. sadece 7 özellik en önemli özellikler olarak tespit etmişlerdir.

Ancak, araştırma sonuçlarına göre, tüm filmlerin başarı yüzdesi endüstriyel kullanım için iyi görünmüyordu ve eğitim setleri (yani 1050 film) daha büyük olsaydı ve sosyal ağ verileri gibi ek özellikleri göz önünde bulundurduklarında

inanıyorlardı. Haber analizi, modelin performansını iyileştirebilirler. Araştırmalarında göz ardı ettikleri bir diğer önemli gerçek ise filmlerin kendi aralarındaki ilişkidir. Örneğin, belirli bir filmin yönetmeni başka popüler filmler yönetti mi? Vs. Doğruluğunu daha da artırabileceklerini düşünselerdi (Nithin ve Lijiya, 2014: 367).

B. Google Trendler Kullanılarak IMDB Film Puanını Tahmin Etme

Deniz Demir, Olga Kapralova ve Hongze Lai, araştırmalarında, filmle ilgili bilgiler için Google arama frekanslarını kullanarak IMDB film derecelendirmelerini tahmin etmek için bir model geliştirdiler (Burgos, vd. 2015). Modeli eğitmek için girdi olarak IMDB veri setini ve Google arama frekanslarını kullandılar. Bunun için Amerika'da 3 yıl arasında 400 film kullanmışlar. Filmlerin %50'si IMDB puanında iyi (derecelendirme 6'dan büyük), diğerleri IMDB puanında kötü (derecelendirme 6 veya altı). Her film için film adı, yönetmen, film oyuncuları ve çıkış tarihi topladılar. Film popülerliğini tahmin etmek için iki farklı yaklaşım kullandılar. Bir yaklaşım, Google Trendler ve Google AdWords istatistiklerini birleştirmek, diğeri ise yalnızca Google Trendler istatistiklerini kullanmaktı.

İlk yaklaşımda, lojistik sınıflandırıcı, DVM modeli ve çok katmanlı algılayıcının performansını karşılaştırmışlardır. Ancak, bu yöntemlerin hiçbiri çıktıda %55'ten fazla doğruluk vermedi. Dolayısıyla sonuç, önemli bir çıktı vermeyen, adil bir yazı turasına benzer şekilde uygulanan tüm bu yöntemlerdi. İkinci yaklaşımda ise yalnızca Google Trendler verilerini kullanmışlardır. Ardından, ilk yaklaşımda kullanılan aynı 3 modeli test ettiler. İlginç bir şekilde, SVM modelinin %72 doğruluğa sahip olduğu ve diğer ikisinin yaklaşık %60 doğruluğa sahip olduğu ikinci yaklaşımda doğruluklar biraz daha yüksek olmuştur (Burgos, vd. 2015).

Bulgularına göre, genel olarak, belirli bir film için Google arama sorgularının sayısı, film vizyona girmeden bir hafta önce artmaya başlıyor ve çıkış tarihi civarında en yüksek seviyeye ulaşıyor. Ardından, yaklaşık 4 ay sonra arama eğilimi ortadan kalkar. Ayrıca, yayın sonrası uzun vadeli Google arama etkinliğinin, yayın öncesi ve kısa vadeli yayın sonrası arama etkinliğinden daha yüksek tahmin yeteneğine sahip olduğunu gözlemlediler. Bu, daha iyi bir tahmin için film vizyona girdikten sonra en az 3-4 ay beklemeleri gerektiği anlamına geliyor. Ancak o zaman, bir tahminin değeri çok değerli olmayabilir (Demir, vd. 2012).

C. Veri Madenciliği Teknikleriyle "Hype" Analizine Dayalı Olarak Filmlerin Gişe Başarısının Tahmin Edilmesi

A. Sivasantoshreddy'de P. Kasat ve A. Jain, yutturmaca analizini kullanarak bir filmin gişe açılış tahminini tahmin etmeye çalıştı. Esas olarak bu makale, hype analizi için twitter verilerine odaklanmaktadır. Hype analizinin arkasındaki ana mantık, bir filmin başarısı, büyük ölçüde açılış hafta sonu gelirine ve vizyona girmeden önce insanlar arasında ne kadar yutturmaca aldığına bağlıdır. İlk başta web tarayıcısını kullanarak bir filme ait tweet sayısını buldular. Bu tweetler saat bazında toplanmaktadır. Hype ölçümü için üç faktör vardır. Birinci faktör, "Saniyedeki ilgili tweet sayısı"nı hesaplamaktır. İkinci faktör, "Tweetleri gönderen farklı kullanıcıların sayısını bulun". Üçüncü faktör "Bir tweetin erişimini hesapla". Burada bir tweetin erişimi, bazı farklı kişilerin tweet'lerinin farklı değere sahip olduğu anlamına gelir. Tanınmış bir aktör veya yönetmenin bir film için olumlu bir tweet atmasının, ortalama bir kişi tarafından atılan bir tweet'ten daha değerli olduğunu varsayalım. Bir tweetin erişimini hesaplamak için belirli bir kullanıcının takipçisini hesaplamışlar. Saniyedeki ilgili tweet sayısını hesapladılar, İkinci faktör "Her film için bu üç faktörün ortalama değerini alarak bir tweetin erişimini hype faktörü olarak bulun ve hesaplayın. Hype faktörüne, filmin gösterime gireceği ekran sayısına ve tüm biletlerin gösteri başına ekran başına ortalama fiyatına dayalı analizleri. Toplam model çok basit hesaplamalar ve sadece bir filmle ilgili tweet sayısını saydılar, ancak tweet'in olumlu mu yoksa olumsuz mu olduğunu bilmek için herhangi bir dil işleme kullanmıyorlar. Bir gişe filminin sinemalarda gösterime girmesinden önce finansal başarısının tahmininde bir sinir ağı kullanılmıştı. Bu tahmin 9 sınıfa ayrılmış bir sınıflandırma problemine dönüştürülmüştür. Model çok az özellik ile temsil edildi (Ajay, vd. 2012).

D. "Collaborative" Filtreleme ve Bulanık Sisteme Dayalı Film Başarısının Tahmini İçin Yapılan Kullanıcı Eğilimli Analiz

Tomar ve Verma 2015'deki çalışmalarında, ortak bir filtreleme sistemi kullanarak kullanıcıların film derecelendirmelerini tahmin eder ve doğrular. Kullanılan tahmin stratejisi, bir kullanıcının güvenilirliğini belirlemek için önceki derecelendirme denemelerini kullanmaktan oluşur. Bu, film derecelendirmesinin doğruluğunu sağlamak için yapılır. Bir sonraki adım, bilinmeyen kullanıcıları,

bilinen kullanıcılarla kişilik benzerliklerini analiz etmek için değerlendirmektir (işbirlikçi filtreleme). Bu değerlendirme, kullanıcıların film türü ilgilerine göre gruplanmasına yol açar. Daha sonra yazarlar, bir grup benzer kullanıcıyı kullanarak bir kullanıcının derecelendirmesini tahmin etmek için bir benzerlik hesaplaması yürütür. Son olarak, öneri olmadığında işbirlikçi filtrelemenin sonuçlarını iyileştirmek için bulanık bir çıkarım süreci kullanırlar. Yazarlar, günde ortalama film sayısından daha fazla puan alan veya toplam kullanıcı puanlarının yüzde yirmi beşinden fazlasını puanlayan kullanıcıları filtrelemeye çalışır. Yazarlar benzerlik indeksini ve tahminini hesaplar. İşbirlikçi filtrelemeyi tek başına kullanırlar ve ardından onu bulanık sistemle birleştirirler. Sonuçlar, işbirlikçi süzme ve bulanık sistem bir arada kullanıldığında, tek başına işbirlikçi süzme kullanımına kıyasla, tahmin karekök hatasında ve ortalama güvenilirlikte gözle görülür bir gelişme olduğunu göstermektedir (Rachit ve Cherag, 2015: 478).

E. Film Başarısını Tahmin Eden Diğer Çalışmalar

Başka bir bakış açısıyla, (Oghina, vd. 2012) 'deki çalışma, iki tür özelliği kullanarak bir kullanıcı film derecelendirmesini tahmin etmeye çalışır: tweet'lerden ve yorumlardan yüzeysel ve metinsel özellikler. Yüzey özellikleri, bir film etrafındaki çevrimiçi etkinliklerin sayısını veya hacmini (niceliksel) içerir ve metinsel özellikler, bu çevrimiçi etkinliklerin anlamına (niteliksel) atıfta bulunur. Derecelendirmeyi tahmin etmek için yazarlar doğrusal bir regresyon modeli uygular. Farklı özellik kombinasyonlarını kullanarak bu modeli birden çok kez çalıştırırlar ve her kombinasyondan elde edilen hatayı karşılaştırırlar. En iyi sonuç, tweeter özelliklerini Facebook'taki beğeni/beğenmeme özellikleriyle birleştirmek. Bu kombinasyon, %42 ortalama mutlak hata ve %52 kök kare ortalama hata verir (Oghina, vd. 2012).

(Parimi ve Caragea, 2013)'de amaç, filmleri gelirlerine göre sınıflandırdıktan sonra vizyon öncesi gişe başarısını tahmin etmektir. Veriler, aktörler, yönetmenler, çalışma zamanı, tür çıkış tarihi, devam filmi bilgileri ve bütçe tarafından açıklanan filmlerden oluşur. Veriler BOXOfficeMojo'dan toplanır. Yazarlar, her filmin özelliklerini iki grupta sınıflandırır: içerik özellikleri ve bağlantı özellikleri. Bağlantı özellikleri, özellikler arasındaki bağımlılıkları yakalar ve yönlendirilmiş ağırlıklı grafik oluşturmak için kullanılır. Grafikte, ortak oyuncuları, yönetmenleri veya aynı

türe sahip olmaları veya aynı zamanlarda gösterime girmeleri durumunda iki film birbirine bağlıdır. Ağırlıklandırma şeması, kenarlardaki olumsuz etkileri yakalayan radikal temel işlevi kullanır. Yönlendirilmiş çizgedeki sınıflandırma, transdüktif algoritmaya dayalıdır. Son olarak, yazarlar sınıflandırmayı matris çarpanlarına ayırmayı kullanarak yaparlar. Elde edilen en yüksek doğruluk, lojistik regresyon kullanılarak %33.56 idi (Parimi ve Caragea, 2013: 575).

(Borga ve Robert, 2012: 36)'deki çalışma, yapımcı veya stüdyonun karar verme sürecini kolaylaştırmak için yirmi dokuz değişkenin Amerika Birleşik Devletleri gişe geliri üzerindeki etkisini ve ayrıca gelir-bütçe oranını incelemektedir. Toplanan veriler 2010'un en çok hasılat yapan filmlerini içeriyor. Yazarlar, ABD gişesinin önemli tahmincileri olabilecek özellikleri belirlemek için bir korelasyon testi uyguluyorlar. Bu test için kullanılan teknik, sıradan en küçük kareler yaklaşımı ve aşamalı prosedürdür. Sonuçlar, bütçe, devam filmi, animasyon, spor, eğitim ve yıldız gücünün önemli belirleyiciler olduğunu göstermektedir. Ayrıca deney, R-reytingi2 ve Drama'nın gişe üzerinde olumsuz bir etkisi olduğunu gösteriyor. Bir diğer önemli özellik, 6.517'lik ortalama gelir-bütçe oranıyla sonuçlanan Korku türüdür (Borga ve Robert, 2012: 41).

(Shraddha, vd. 2015) yazarları, daha iyi tahmin doğruluğu elde etmek için bir filmin hem klasik hem de sosyal özelliklerini birleştirir. Veri seti 20 farklı film için özellikler içermektedir. Yazarlar bu eserdeki filmleri manuel olarak başarılı veya başarısız olarak sınıflandırır. WEKA'da (http://www.cs.waikato.ac.nz/ml/weka/) J48'i kullanarak, test doğruluğu %66'ya eşit olan karar ağaçları elde ettiler (Shraddha, vd. 2015: 299). Bu çalışmada, topluluk yöntemleri kullanılarak karar ağaçları, k-en yakın komşular (k-NN) ve doğrusal regresyon önerilmiş ve karar ağaçlarının rastgele ormanlara, torbalama ve artırmaya dayalı tahmin performansları k-NN ve k-NN ile karşılaştırılmıştır. 1439 film örneğini kullanarak torbalama ve artırmaya dayalı doğrusal regresyon. Bu, topluluk yöntemlerini kullanan karar ağaçlarının, topluluk yöntemlerinin k-NN ve doğrusal regresyon analizinden daha iyi uygulama etkinliğini sağladığını göstermektedir (Lee, vd. 2020).

Çevrimiçi etkinliklerin genel olarak tahmin üzerindeki büyük etkisi göz önüne alındığında, (Yasseri ve Kertész, 2013) yazarları, çevrimiçi kullanıcıların toplu etkinlik verilerine dayanarak filmlerin finansal başarısı için tahmine dayalı bir model oluştururlar. Bir filmin popülaritesi, editörlerin ve izleyicilerin aktivite düzeyini ve

filmin Wikipedia'ya karşılık gelen girişini ölçerek ve analiz ederek önceden tahmin edilebilir. Kullanılan veriler, ilgili Wikipedia sayfalarıyla birlikte ABD filmlerini içerir. Eksiksiz veri seti, finansal verilerin yanı sıra Wikipedia faaliyet kayıtlarını da içerir. Modeldeki ilk adım, popülerliği dört farklı etkinlik ölçüsüne dayalı olarak tahmin etmektir: kullanıcı sayısı, görüntüleme sayısı, yapılan düzenleme sayısı ve son olarak düzenleme dizisinin işbirlikçi titizliği (bu, sayılan düzenleme sayısıyla aynıdır). bir kullanıcı tarafından yapılan tüm düzenlemeler tek bir düzenleme olarak). Bu metrikleri kullanarak, yazarlar Pearson korelasyon katsayısını hesaplar ve sonuçları çok değişkenli doğrusal regresyon modelini kullanarak sınıf etiketini tahmin etmek için kullanır. Elde edilen belirleme katsayısı R2 %77'ye eşittir (Yasseri ve Kertész, 2013).

Daha az kişi, özellikleri olarak kullanılan bazı önceden yayınlanmış verilerle bir tahmin modeli yapmıştı. Çoğu durumda, insanlar çok az özelliği dikkate aldı. Sonuç olarak, modelleri kötü çalışıyor. Ancak, bir filmin başarısının büyük ölçüde bağlı olduğu izleyicilerin katılımını göz ardı ettiler. Her ne kadar az sayıda insan duygu analizi için NLP'nin birçok uygulamasını benimsiyor (Burgos, vd. 2015) ve test alanları için film incelemeleri topladı. Ancak tahminin doğruluğu, test alanının ne kadar büyük olduğuna bağlıdır. Küçük bir alan, ölçüm için iyi bir fikir değildir. Yine çoğu eleştirmenlerin eleştirilerini dikkate almadı (Pimwadee ve Lina, 2005). Bu çalışma, sinir ağları, regresyon ve karar ağaçları gibi çeşitli Veri Madenciliği tekniklerini kullanan bir tahmine dayalı model inşa ederek bir filmin kârını tahmin etmeyi amaçlamaktadır. Model, gişe geliri tahmininin elde edilmesini sağlayacaktır. En doğru tahminleri belirlemek için iki ölçüm kullanıldı: kategorik modeller için yanlış sınıflandırma hatası ve sürekli model için ortalama kare hatası. Bu çalışmada, çok katmanlı algılayıcı kullanılarak en iyi tahmin sonuçları elde edilmiştir. Bağımlı değişken arasındaki temsili ayrımla ilgili olarak, çok sınıflı model, tahmin edilecek sınıf sayısının artmasıyla açıklanan diğerlerine göre çok daha yüksek bir hata oranı sunar (Galvão ve Henriques, 2018).

(Saranya ve Hussain, 2015) Naive Bayes ağacını kullanarak belirli bir kullanıcının bir film derecelendirmesini tahmin eder. Veriler Movie Lens'ten (https://movielens.org/) toplanır. Her kullanıcı için bilgiler, ad, yaş, cinsiyet, film türü ve bulanık kuralı izleyen soruları içerir. Tahmin amacıyla, yazarlar iki endeks hesaplar: iki kullanıcı arasındaki benzerlik düzeyi olan kosinüs benzerliği ve belirli

bir kullanıcı tarafından bir filmin derecelendirmesinin tahmini olan tahmin edilen derecelendirme. Naive Bayes yaklaşımı, filmleri kullanıcı bilgilerine göre ayırır, ardından kullanıcıları ilgi alanlarına göre ayırır. Yazarlar, belirli bir kullanıcıya, tahmin edilen derecelendirmelerine göre bir dizi film öneren bir öneri sistemi oluşturur. (Saranya ve Hussain, 2015: 5858)

(Anand, vd. 2015) yazarları, hem klasik hem de sosyal faktörlerin entegrasyonunun ve klasik faktörler arasındaki karşılıklı ilişkinin yüksek bir doğruluğa yol açtığını öne sürmektedir. Kullanılan veriler, filmin analizi için göz önünde bulundurulan klasik faktörleri (oyuncular, yapımcılar, tür vb.) içerir. Bunlar IMDB ve benzeri sitelerden alınmıştır. Ayrıca, veriler şunları içeren sosyal faktörleri içerir: 1. Twitter'da atılan tweetlerin duygu analizi, 2. Film öncesi videolarda (film fragmanları) YouTube izlenme isabetleri, 3. Film çıkış tarihi olarak izlenmelerin artış oranı yaklaşımlar. Çok değişkenli bir doğrusal regresyon modeli sunulur ve bu, %70,57'lik bir R2 ile sonuçlanır (Anand, vd. 2015).

(Lash ve Zhao, 2016) 'daki yazarlar, film başarısı için oyuncu kadrosuna, içerik ve yayın zamanına dayalı olarak tahmine dayalı bir model oluşturuyor. BoxOfficeMojo'dan toplanan veriler, farklı film bilgilerini içerir. Yazarlar nitelikleri türlerine göre ayırdılar: derecelendirmeyi, türü ve konuyu içeren “hangi nitelikler”. Yıldız gücünü ve işbirlikçi ağı içeren “kim nitelikleri”. Yayınlanan verilerin ortalama yıllık kârını, türe göre yüzde olarak yıllık kârlılığı ve AWPG adlı bir endeksi içeren “ne zaman öznitelikleri”. Yazarlar, karlılıklarını tahmin etmek amacıyla filmleri sınıflandırmak için %77,1 doğruluk elde eden bir lojistik doğrusal regresyon kullanır (Lash ve Zhao, 2016: 889).

Asad, vd. 2012'deki çalışmada, WEKA'da C4.5 ve PART kullanılarak önceden yayınlanmış filmlerin popülerliğine ilişkin bir sınıflandırma şeması önermektedir. Filmlerle ilgili veriler IMDB'den ve finansal veriler BoxOfficeMojo'dan toplanır. Yazarlar, filmlerin derecelendirmesini mükemmelden berbata kadar dört kategoride gruplandırdı. İki sınıflandırıcı yaklaşık olarak aynı sonuçları gösterdi (%1'den az varyasyonla). Derecelendirmeyi etkileyen faktörleri anlamak için yazarlar iki veri seti oluşturur. İlk veri seti filme ait yıl, oyuncu kadrosu gibi bilgileri, ikinci veri seti ise finansal bilgileri içermektedir. Sonuçlar, filmin içerik bilgilerinin derecelendirmeyi finansal bilgilerden daha fazla etkilediğini gösteriyor. İçerik bilgilerine göre film reytingi tahmininde doğruluk oranı finansal bilgilerden elde

edildiğinde %56 iken %77'dir (Khalid, vd. 2012: 751). Asur ve Huberman 2010'da sosyal medya verilerinin gişe tahmin performansını etkileyebileceğini göstermiştir. Çalışmalarında, insanların hakkında olumlu tweetler paylaştığı filmlerin daha fazla izleyici kitlesine sahip olacağı varsayılmıştır. Bu varsayım modellemesini Lineer Regresyon yaklaşımı ile ispatlamışlardır (Sitaram ve Bernardo, 2010).

Eliashberg ve arkadaşlarının çalışmasıyla 2014'de yalnızca filmin senaryosunu ve tahmini bütçesini kullanarak çok farklı bir yaklaşım getirildi. Çalışmalarında, çekirdek tabanlı yaklaşıma dayalı olarak senaryolardan semantik kullanarak film yapımının onaylanması anlamına gelen yeşil ışıklandırma sürecine yardımcı olmayı amaçladılar (Jehoshua, vd. 2014: 2641). Bahsedilen Wikipedia çalışmasına dayalı çok benzer bir çalışma da filmlerin vizyona gireceği ülkeler farkıyla ilk hafta sonu gişe değerlerini tahmin etmeyi amaçladı. Bu çalışmada ABD'nin yanı sıra İngiltere, Avustralya, Japonya ve Almanya gibi farklı ülkelerden film verileri toplanmıştır. Aynı tahmin değişkenlerini ve belirleme katsayısı olan aynı değerlendirme ölçütünü kullandılar. Ancak ABD dışındaki ülkeler için daha düşük performans sonuçları aldılar. ABD'nin aksine Wikipedia sayfalarının daha düşük web sitesi trafiğinin bu daha düşük sonuçlara yol açabileceğini varsaydılar. Bu çalışma, sadece ABD'den toplanan filmlerin benzer çalışmalar için yeterli olabileceğini göstermiştir (Brian de ve Ryan, 2014).

Çalışmaların çoğu tek bir sosyal medya kaynağından yararlanılırken, Apala ve arkadaşları 2013'de birden fazla sosyal medya kaynağını bir araya getirme fikrini ortaya attı. Hem tür hem de yönetmen gibi film meta verilerini ve kullanıcıların YouTube'un filmin resmi fragmanı hakkındaki yorumlarını kullandılar. Sharda ve Delen gibi, Apala ve arkadaşları gişe performansını Hit, Neutral ve Flop olmak üzere üç kategoriye ayırarak bu regresyon problemini bir sınıflandırma problemine dönüştürdü. Ancak 35 filmden oluşan yetersiz veri nedeniyle sonuçları çok düşüktü (Apala, vd. 2013: 1213).

(Delen ve Sharda, 2009: 33) 'te, Lydia (haber verilerini toplamak ve analiz etmek için yüksek hızlı metin işleme sistemi) tarafından üretilen nicel haber verilerinin yer aldığı Haber analizi yoluyla film brüt tahmini iyileştirilmeye çalışıldı. İki farklı model (regresyon ve k-en yakın komşu modelleri) içeriyordu. Ama sadece yüksek bütçeli filmleri düşündüler. İsim olarak ortak bir kelime kullanıldığında model başarısız oldu ve bir film hakkında haber olup olmadığını tahmin edemedi

(Wenbin ve Steven, 2009). Araştırmalar, hem olumlu hem de olumsuz çevrimiçi incelemelerin haftalık gişe sonuçlarıyla ilişkili olduğunu, dolayısıyla gişe geliri tahmin edilirken dikkate alınabileceğini göstermiştir (Basuroy, vd. 2003: 109). Rolleri farklı, hem profesyonel inceleme içerikleri hem de kullanıcı tarafından oluşturulan incelemeler gişe geliri ile ilişkilidir (Basuroy, vd. 2003: 109).

Bu yazıda, Amerika Birleşik Devletleri'nde üretilen 100'den fazla film hakkında tarihsel verileri kullanan bir araştırmayı rapor ediyoruz. Bu verilerden karar ağaçları çıkarılarak bir filmin daha yapım aşamasından önce karlı olup olmayacağının tahmin edilmesi sağlanmıştır. Sonuçlar, ortaya çıkan modelin bir filmin karlı olup olmayacağını %70'in üzerinde bir doğrulukla tahmin ettiğini ve bu modelin film yapımcıları için bir karar destek aracı olarak kullanılabileceğini göstermektedir. Burada sunulan yaklaşım, müzik veya video oyunu endüstrileri gibi eğlence sektörünün diğer dallarına da aynı şekilde uygulanabilir (Burgos, vd. 2015). Bu raporda amaç, IMDB'den film verilerini ve YouTube ve Wikipedia'daki sosyal medya verilerini toplamak ve küçük veri kümeleriyle yüksek doğruluklarıyla bilinen iki makine öğrenimi algoritmasının (Random Forest ve XGBoost) performansını karşılaştırmaktır, ancak büyük özelliklere sahiptir. Ayarlamak (Mhowwala, vd. 2020: 383).

Bununla birlikte, IMDb film önerileri, bağlamsal uyarlanabilirliklerini doğal olarak sınırlayan etkileşim veya kullanıcı kontrolü için herhangi bir araç sunmaz. Bu çalışmada, etkileşimli film önerileri sunan ve kullanıcı derecelendirme verileri için IMDb web sitesini entegre eden MovieBrain adlı Google Chrome uzantımızı açıklıyoruz. Kullanıcıların öneri sürecini ve sonuçlarını manuel olarak yeniden ayarlamasına olanak tanıyan dinamik ayarlar ve tür filtreleri mevcuttur (Dooms, vd. 2014).

Bu yazıda, öncelikle bu etkili faktörlerin stratejik bir araştırmasını sunuyoruz. Ardından, büyük veriyi kullanarak bu faktörleri modelleyerek bir filmin BOR'unu tahmin etmek için yeni bir çerçeve ortaya koyduk. Spesifik olarak çerçeve, bir dizi özellik öğrenme modeli ile bir tahmin ve sıralama modelinden oluşur. Özellikle, öğrenme özellikleri için tasarlanmış iki model vardır: a. aktörlerin, yönetmenlerin ve şirketlerin gizli temsillerini eşzamanlı olarak öğrenmek için işbirliği ilişkilerini toplu olarak yakalayabilen yeni bir dinamik heterojen ağ yerleştirme modeli; b. Fragmanlardan film kalitesinin üst düzey temsillerini ortaya çıkarmak için

tasarlanmış derin sinir ağı tabanlı bir model. Öğrenilen özelliklere dayanarak, BOR tahmin sonuçlarını elde etmek için karşılıklı olarak geliştirilmiş bir tahmin ve sıralama modeli geliştiriyoruz. Son olarak, çerçeveyi Çin film pazarına uyguluyoruz ve gerçek dünya verilerini kullanarak kapsamlı bir performans değerlendirmesi yapıyoruz (Wang, vd. 2020: 32).

Bu çalışmanın amacı, doğal dil işleme teknikleri, metin madenciliği ve derin öğrenme sinir ağlarını kullanarak, yalnızca film senaryosundaki bilgilere dayanarak film gişe performansını tahmin etmek için tahmine dayalı bir model geliştirmektir. Bu yaklaşım, Brezilya Film Ajansı'nın seçim sürecine özel olarak odaklanarak, yatırımcının projenin önceki adımlarında karar verme sürecini optimize etmeyi amaçlar (Corrêa de Sá, 2020: 52). Bu çalışmada, bir filmin vizyona girmeden önce gişedeki finansal performansını tahmin etmede sinir ağlarının kullanımı araştırılmaktadır. Modelimizde, tahmin problemi, gişe hasılatlarının nokta tahminini tahmin etmek yerine bir sınıflandırma problemine dönüştürülür, gişe hasılatlarına dayalı bir film dokuz kategoriden birinde sınıflandırılır, bir 'flop' ile Bir 'gişe rekorları kıran' Sinir ağımızın son literatürde önerilen modellerin yanı sıra 10 katlı çapraz doğrulama metodolojisini kullanan diğer istatistiksel tekniklerle karşılaştırılması, sinir ağlarının bu ortamda çok daha iyi bir tahmin işi yaptığını gösteriyor (Sharda ve Delen, 2006: 246). Tahmin modellerimizde, gişe gelirlerinin nokta tahminini tahmin etmek yerine tahmin problemini bir sınıflandırma problemine dönüştürdük; Bir filmi (gişe hasılatlarına dayanarak) "flop"tan "gişe rekorları kıran (blockbuster)"a kadar dokuz kategoride sınıflandırdık. Burada, bireysel modelleri örneklerinkilerle karşılaştırdığımız heyecan verici tahmin sonuçlarımızı sunuyoruz (Delen ve Sharda, 2009: 34).

III. METODOLOJİ

A. Veri Toplama

Veri setini oluşturmak için dizi ve film içeren IMDB çevrimiçi veritabanı kullanılmaktadır. Veri hazırlamak için kullanılacak dosyalar (http://www.imdb.com/interfaces) adresiden alınmıştır. Filmlerin yönetmeni, bütçe, film süresi gibi kriterleri 43 tane dosya halinde Çizelge 1'de tanımları ile birlikte gösterilmektedir.

Çizelge 1 IMDB veri dosyalarının isimleri ve kısa açıklaması

Dosya isimleri	**Kısa tanım**
actors.list.gz	Her aktör için bulunan film listesi
actresses.list.gz	Her aktris için bulunan film listesi
aka-names.list.gz	Aktör\aktrislerin diğer bilinen adları listesi
aka-titles.list.gz	Film başlıklarının diğer bilinen isimleri listesi
alternate-versions.list.gz	Her gösteri için geçmiş düzenleme
biographies.list.gz	Her aktörün/aktrisin biyografisi
business.list.gz	Filmlerin gişe başarısı ve bütçeleri
certificates.list.gz	Her film için yaş sınırı
cinematographers.list.gz	Her görüntü yönetmeni için film listesi
complete-cast.list.gz	Oyuncu listesi tamamlanan filmler
complete-crew.list.gz	Yapım ekibi listesi tamamlanan filmler
composers.list.gz	Her bestecisinin bulunduğu filmler listesi
costume-designers.list.gz	Kostüm tasarımcılarının listesi
countries.list.gz	Filmlerin yapıldığı ülkeler listesi
directors.list.gz	Film yönetmenlerinin bulunduğu liste
distributors.list.gz	Film dağıtımcılarının bulunduğu liste
editors.list.gz	Film kostüm tasarımcıları listesi
filesizes	IMDB arayüzündeki her dosyanın boyutu
genres.list.gz	Film türlerinin listesi
german-aka-titles.list.gz	Başlıkların Almanca tercümesi
goofs.list.gz	Her filmde saçmalıklar
italian-aka-titles.list.gz	Başlıkların İtalyanca tercümesi
keywords.list.gz	Anahtar kelimelerin bulunduğu liste
language.list.gz	Filmin hangi dilde olduğunu içinde bulunduran liste
literature.list.gz	Filmler hakkında bazı literatürler
locations.list.gz	Her filmin çekildiği yer
miscellaneous-companies.list.gz	Her film için çeşitli şirketlerin listesi
movie-links.list.gz	Filmlerin IMDB linklerinin bulunduğu liste
movies.list.gz	Filmler ve yıllarının belirtildiği liste
mpaa-ratings-reasons.list.gz	Derecelendirilen her film için MPAA derecesinin nedeni
plot.list.gz	Film konularının listesi
producers.list.gz	Film yapımcılarının listesi
production-companies.list.gz	Film üretim şirketleri listesi
production-designers.list.gz	Üretim şirketlerinin listesi

Çizelge 1 IMDB veri dosyalarının isimleri ve kısa açıklaması Devam

quotes.list.gz	Filmlerde kullanılan alıntılar listesi
ratings.list.gz	Filmlerin IMDB puanlarını içeren liste
release-dates.list.gz	Yayınlanma tarihleri
running-times.list.gz	Film süreleri listesi
sound-mix.list.gz	Her film için ses miksi yapan şirketler
special-effects-companies.list.gz	Film özel efekt şirketlerinin listesi
technical.list.gz	Üretim zamanı kullanılan aletler hakkında teknik bilgiler
trivia.list.gz	Triviaların bulunduğu filmler listesi
writers.list.gz	Yazarların bulunduğu filmler listesi

B. Özellik Seçimi

İlk olarak modelin geliştirilmesinde etkisi olmayacağı düşünülen özellikler veri setinden çıkarılıyor. Örneğin aktör biografisi, farklı film şirketleri listesi gibi özellikler başarı faktörünü etkilemediği için kullanılmayacaktır. Çizelge 2. de kullanılmayan dosyalar gösterilmektedir.

Çizelge 2 Veri kümesinden kullanılmayan dosya isimleri

Dosya isimleri
aka-names.list.gz
alternate-versions.list.gz
filesizes.gz
german-aka-titles.list.gz
italian-aka-titles.list.gz
miscellaneous-companies.list.gz
biographies.list.gz
complete-cast.list.gz
complete-crew.list.gz
movie-links.list.gz
aka-titles.list.gz

Bu çalışmanın temel amacı, yatırımcılara böyle bir filme yatırım yapıp yapmama konusunda tavsiyelerde bulunmak için bir karar verme tekniği vermek adına, bir filmin başarısını daha prodüksiyondan önce tahmin etmek olduğundan, başarı göstergesi olarak seçtiğimiz tüm özellikler ön seçimdir. üretim bilgileri. Bu nedenle, Çizelge 3'te gösterilen tüm dosyaları da kaldırdık. Bu dosyalar, bir film prodüksiyonunda izleyiciler tarafından bulunan hatalar, sertifikalar ve bir filmin

yapımından sonra kazandığı MPAA derecelendirmeleri (Mahesh, vd. 2010: 294), bir film gösterimi gibi üretim sonrası bilgileri içerir. Örneğin tarih ve önemsiz şeyler, vb.

Çizelge 3 Filmin yapım sonrası bilgilerini içeren dosyalar

Dosya isimleri
trivia.list.gz
technical.list.gz
quotes.list.gz
mpaa-ratings-reasons.list.gz
release-dates.list.gz
goofs.list.gz
certificates.list.gz

Veritabanı, her film hikayesinin bir açıklamasını saklayan bir "Plot" dosyası içerir. Ayrıca, içinden çıkarabildiğimiz tek ilginç bilgi anahtar kelimeler olduğu için bu dosyayı da eledik. Bunlar "Anahtar kelimeler" dosyasında bulunabilir. Veri kümesinde çok büyük sayıda film bulunduğundan bazı yapım türleri çıkartılmıştır. Dolasıyla ABD'de üretilmeyen tüm yapımlar dosyalardan çıkartılmıştır.

C. Veri Ön İşleme ve Veri Temizleme

Dosyalar arasındaki ve aynı dosyadaki tutarsız biçim ve kalıplar nedeniyle, her dosyayı temizlemek ve yeniden biçimlendirmek için bir Java uygulaması geliştirdik. Temizlediğimiz ilk dosya filmlerin listesi "movies.list". Tüm TV dizileri, ABD dışı filmler ve kısa filmlere ek olarak, yılı bilinmeyen herhangi bir filmi sildik (gösterim süresi 60 dakikadan az). "İçinde Gölgeler 1991" ve "Gölgeler İçinde 2015" gibi aynı isme sahip, ancak yapım yılları farklı olan filmleri ayırt etmek için yılı film başlığının bir parçası olarak ekledik.

Her film için türlerin listesini "genres.list"ten çıkardık. Bir sonraki adım, tüm şovları, haberleri ve müzikalleri silmekti. Bu ayıklanan bilgileri yeni bir dosyada sakladık. Çizelge 4, yeni tür veri dosyasının bir alt kümesini göstermektedir. Bir filmin birden fazla türü varsa, türlerin hepsinin listelendiğini ve virgülle ayrıldığını unutmayın.

Çizelge 4 Türleriyle birlikte filmlerin listesi

Film ismi	Filmin türü\türleri
The Elegant Clockwork of the Universe (2013)	Dram, Gizem, Bilimkurgu
The Elektra/Vampyr Variations (2009)	Komedi, Fantezi, Korku
The Elementary Sherlock Holmes (2009)	Biyografi, Belgesel
The Elements Club: Lord of Flawless Strength (2014)	Romantik
The Elements Club: Unity Match (2015)	Romantik
The Elements of Me (2016)	Reality-TV
The Elements of Me; Introducing Chosen Wilkins (2016)	Komedi
The Elephant King (2006)	Dram, Romance

Çizelge 5, yeniden biçimlendirildikten ve IMDB film kimliği gibi tüm istenmeyen bilgiler hariç tutulduktan sonra derecelendirme dosyası "ratings.list"in bir kısmını göstermektedir.

Çizelge 5 IMDB puanına sahip filmlerin listesi (10'un üzerinde puan)

Title	Rating (/10)
American Pie (1999)	7.0
American Pie 2 (2001)	6.4
American Pie Presents Band Camp (2005)	5.0
American Pie Presents Beta House (2007)	5.3
American Pie Presents The Naked Mile (2006)	5.1
American Pie Revealed (2003)	6.8

Derecelendirme, üretim yılı, tür (N.A., vd. 2013: 52) ve çalışma süresini içeren sonuçtaki dosya grubunu tek bir dosyada birleştirdik. Çizelge 6, bu dosyadan bir alıntıyı göstermektedir.

Çizelge 6 Yapım yılı, türü, derecelendirmesi ve yayınlanma süresiyle birlikte filmlerin listesi

Filmin adı	Yılı	Filmin türü	10 üzerinden puanı	Film süresi (dak.)
Painted Faces	1929	Suç, Gizem	5.5	74
I've Got Your Number	1934	Komedi, Romantik	6.6	69
Bambi	1942	Animasyon, Dram, Aile	7.4	70
The Strongest Man in the World	1975	Komedi, Aile, Fantezi	5.9	92
Blue Ecstasy in New York	1980	Yetişkin, Dram	6.0	90

Çizelge 6 Yapım yılı, türü, derecelendirmesi ve yayınlanma süresiyle birlikte filmlerin listesi Devam

Cheetah	1989	Macera, Aile	6.1	83
Adventures of Buttwoman	1991	Yetişkin	6.8	90
No Escape No Return	1993	Aksiyon	3.8	91
Playback	1997	Yetişkin	4.3	120
Love American Style	1999	Komedi, Romantik	6.9	60
The Tooth Fairy	2006	Korku, Gerilim	4.6	89

Bir sonraki adımda aktör, yönetmen, senarist, film yapım şirketi ve diğer kriterleri veri setine ekledim. Ardından kalan özellikleri (aktörler, yapım şirketleri, yönetmenler, yazarlar vb.) ilişkili dosyalarından çıkardık ve bunları son veri kümesine ekledik. Farklı dosyalar arasındaki bağlantı, filmin adıydı. Aktörler/aktrisler dosyaları göz önüne alındığında, her film için birincil oyuncuları/aktrisleri ondan çıkarmamız imkansızdı. Dosyalar, mevcut tüm aktörleri/aktrisleri (IMDB veritabanından) alfabetik sırayla listeler, ardından her birinin rol aldığı filmlerin listesi gelir. Bunun için "OMDBAPI" adlı ücretsiz bir genel API kullandık (http:/ /www.omdbapi.com/). Bu API, birinin bir film adı ve üretim yılını içeren bir veri isteği göndermesine izin verir. Json yanıtı, birincil aktörleri/aktrisleri içerir. İşlemin bir örneği Şekil 3'de gösterilmektedir.

{"Title":"Prisoners","Year":"2013","Rated":"R","Released":"20 Sep 2013","Runtime":"153 min","Genre":"Crime, Drama, Mystery","Director":"Denis Villeneuve","Writer":"Aaron Guzikowski","Actors":"Hugh Jackman, Jake Gyllenhaal, Viola Davis","Plot":"When Keller Dover's daughter and her friend go missing, he takes matters into his own hands as the police pursue multiple leads and the pressure mounts.","Language":"English","Country":"United States","Awards":"Nominated for 1 Oscar. 10 wins & 38 nominations total","Poster":"https://m.media-amazon.com/images/M/MV5BMTg0NTIzMjQ1NV5BMl5BanBnXkFtZTcwNDc3MzM5OQ@@._V1_SX300.jpg","Ratings":[{"Source":"Internet Movie Database","Value":"8.1/10"},{"Source":"Rotten Tomatoes","Value":"81%"},{"Source":"Metacritic","Value":"70/100"}],"Metascore":"70","imdbRating":"8.1","imdbVotes":"685,917","imdbID":"tt1392214","Type":"movie","DVD":"17 Dec 2013","BoxOffice":"$61,002,302","Production":"N/A","Website":"N/A","Response":"True"}

Şekil 3 Özelliklerin gösterimine örnek

Bu kaynak vasıtasıyla belirtilen özelliklerden yalnızca oyuncular kısmını (örnek: Hugh Jackman, Jake Gyllenhaal, Viola Davis) veri setinde kullanacığım. Bu işlemi veri kümesinin oyuncular özelliğinin ana aktörlerden oluşması için yaptım. Veri kriterlerinin geri kalanı IMDB dosyalarından önceki aşamalarda alınmıştır. İstekleri "OMDAPI" ye göndermek ve her film için oyuncuları/aktrisleri çıkarmak ve sonuçları kaydetmek için bir PHP betiği geliştirdik.

Yukarıda belirtilen tüm veri azaltma, işleme ve çıkarma işlemleriyle, aşağıdaki özellikleri içeren son veri setimizi oluşturduk: Yıl, derecelendirme, oy sayısı

(derecelendirme dosyasından), çalışma süresi, tür, distribütörler, yapım şirketleri, özel efekt şirketler, ses karışımı anahtar kelimeler, oyuncular, aktrisler, görüntü yönetmenleri, besteciler, kostüm tasarımcıları, yönetmenler, editörler, yapımcılar, yapım tasarımcıları ve yazarlar. Tüm eksik değerler NA ile değiştirildi. Belirli bir film için bir özelliğin birden fazla örneği varsa, bu örnekleri virgülle ayırdığımızı unutmayın (örneğin, aynı film için birden fazla yapım şirketi olduğunda).

Bazı metin tipi özelliklerimiz olduğu göz önüne alındığında, bu özelliklerin her değerini ortalama derecelendirmesiyle değiştirdik. Örneğin, bir X oyuncusu için, X'in rol aldığı tüm filmleri çıkardık, sonra bu filmlerin ortalama puanını hesapladık. Veri setinde X'i bu ortalama puanla değiştirdik. Bu yaklaşım, tek bir filmde az sayıda kullanıcı oy oyu ve yüksek reyting alan faktörlere (oyuncular, yönetmenler vb.) karşı haksız bir önyargı verebilir. Bu yanlılığı önlemek için, (Saraee, White ve Eccleston, 2004) 'de yapılana benzer şekilde, oy sayısı binden az olan tüm filmleri veri setimizden çıkardık.

Metinden sayısala dönüştürme işlemini gerçekleştirdikten sonra ve öznitelikler aynı anda birden fazla değere sahip olabileceğinden (aynı film için birden fazla oyuncu, yönetmen, yazar vb.), tüm bu değerleri ortalamalarıyla değiştirdik.

Bir sonraki adım olarak, sürekli değerleri tahmin etmek yerine sınıflandırma problemine sahip olmak için derecelendirmeyi en yakın tam sayıya yuvarladık.

1. Filmin Türü

Bazı tür türleri, bazı araştırmacılar tarafından önemli belirleyiciler olarak tanımlanmıştır. Litman'a göre, bilimkurgu türü, filmin finansal başarısı ile pozitif olarak ilişkilidir. Topf 2010'da komedi ve aksiyon buluyor, Terry ve diğerleri. Aynı yılda aksiyon ve çocuklar ile Hasbrouck ve Deniz animasyonlarının tiyatro gelirleri açısından önemli olduğunu belirtmişlerdir. İtalya'da çizgi roman türü, Fransa'da, komedi ya da romantik komedi ve Polonya'da belgesel ya da komedi. Sharda ve Delen (2006) ve Karniouchina ve arkadaşları korku filmlerini daha düşük gelirlerle ilişkilendiriyor; ancak, önemli ölçüde daha düşük bütçeleri nedeniyle, korku filmleri yatırım getirisi ile pozitif olarak ilişkilidir. Hasbrouck ve Deniz bunu korku filmlerinde genellikle yıldız olmaması ve bunun da maliyetleri önemli ölçüde artırması ile açıklamaktadır. Aksine, batı ve bilim kurgu filmleri, şişirilmiş bütçeleri nedeniyle önemli ölçüde daha düşük yatırım getirisi sağlar (Sharda ve Delen, 2006).

Türlerin etkisine ilişkin sonuçlar genellikle çok tutarlı değildir. Örneğin, Karniouchina ve arkadaşları film türlerinin daha önce önerilenden daha önemli olduğunu buldu. Buna karşılık, Pangaker ve Smith dramanın küresel gişe hasılatı ile yalnızca negatif ilişkili olduğunu bulmuş ve çağdaş küresel film pazarında türün artık bu kadar önemli bir rol oynamadığı sonucuna varmıştır.

2. Yayınlanma Tarihi

Yayın tarihi, yapımcının sinema gelirlerini etkileyebilecek iki temel olguyu dikkate almasını gerektiren önemli bir stratejik karardır - sinemaya katılım ve rekabet. Yazarlar, hipotezlerini belirli aylarda sinemaseverlerin daha sık sinemaya gittiği mantığına dayandırarak, genellikle Noel öncesi ve yaz çıkışlarına odaklandılar. Artan bu talep, yapımcıları filmlerini bu tarihlerde vizyona sokmaya teşvik etmekte ve bu da rekabeti artırmaktadır. Çıkış tarihinin bir filmin finansal başarısı üzerindeki nihai etkisi, yıl boyunca bu iki faktör arasındaki dengenin gelişimine bağlıdır.

Litman Noel çıkışını tiyatro başarısının olumlu bir etkisi olarak tanımladı, ancak yaz gösterimi ile ikincisi arasında hiçbir ilişki bulamadı. Thurau ve arkadaşları yaz gösteriminin açılış gişe sonuçları üzerinde olumlu bir etkisi olduğunu keşfettiler, ancak uzun vadeli gişe üzerindeki olumlu etki, bir yaz gösteriminin filmin ödül alma şansını azalttığı gerçeğiyle azaldı. Polonya'da Nisan, Temmuz ve Ağustos aylarında gösterime giren filmler diğerlerinden önemli ölçüde daha kötü performans gösterdi. Pangaker ve Smith tatillerde filmin çıkış tarihinin önemli bir etkisi bulamamışlardır.

3. Erkek ve Kadın Oyuncular

Litman'ın orijinal çalışması, yıldızların varlığı ile film başarısı arasında herhangi bir ilişki doğrulamadı. Litman, bulgularından, süper yıldızların varlığının, yalnızca iyi performanslara katkıda bulundukları ve filmin kalitesini artırdıkları ölçüde alakalı olduğu sonucuna varıyor. Thurau ve arkadaşları, yıldız gücünün hem uzun hem de kısa vadeli gişe üzerinde biraz olumsuz dolaylı etkisi bile buldu. Bununla birlikte, son çalışmalar tam tersi sonuçlara işaret etmektedir.

Kim'e göre, bir filmdeki ana aktör veya aktris, büyük olasılıkla bu filmin gelir elde etmede başarılı olup olmadığını belirleyecektir. Bu, bir yönetmeni işe alırken de aynı olur. De Vany ve Walls, süperstar varlığının olasılık kütlesini daha yüksek

sonuçlara kaydırdığını bulmuşlardır. Karniouchina, film vızıltılarının ve yıldız vızıltılarının açılış hafta sonu ve uzun vadeli gişe gelirleri üzerindeki etkisini inceleyerek, değişkenleri "bir unvan veya projede yer alan bir top yıldızla ilgili genel heyecan ve beklenti olarak tanımladı. arama tabanlı önlemlerden". Sonuçlar, sinemadaki ilginin sinema gösterimi boyunca gişe gelirini artırmada etkili olduğunu gösteriyor. Star vızıltı açılış haftası gişe gelirlerini artırıyor; ancak, altta yatan film izleyicilerde yankı uyandırmazsa, sonraki haftalarda gelir üzerinde olumsuz bir etkisi olabilir.

Ravid yıldızların marjinal değerlerini yakaladığını ve bu nedenle filmin RoI'sine katkıda bulunmadığını varsayarak "rant yakalama hipotezini" test etti. Bulguları büyük ölçüde hipotezi desteklemektedir: yıldızları kullanan filmler genellikle daha pahalıdır, daha yüksek gelire sahiptir, ancak yatırım getirisi daha düşüktür. Ravid bütçeyi gelir modeline dahil ettiğinde, değişken bütçe tüm önemi aldı - yüksek bütçe, harcama kaynağından bağımsız olarak yüksek gelirlere işaret ediyor. Benzer bir sonuca Liu ve diğerleri tarafından da ulaşılmıştır. Bulguları yıldızların finansman çekme eğiliminde olduğunu doğrulamaktadır. Filme bir yıldızın dahil edilmesi genellikle daha yüksek bir bütçe anlamına gelir; bu nedenle, asıl etkisi daha yüksek bütçe yoluyladır.

De Vany ve Walls sinema filmi kârının sonlu ortalama ve sonsuz varyans ile istikrarlı bir Pareto dağılımına sahip olduğunu bildirmiştir. Ravid'in rant yakalama hipotezine paralel olarak, süperstar filmlerin ortalama, beklenen ve en olası kârı arasındaki farktan kaynaklanan çarpık şekli, "süper yıldızın laneti" olarak adlandırdıkları şeyi açıklar. Dağıtımın çarpıklığı, beklenen değerin en olası sonuçtan önemli ölçüde daha büyük olmasına neden olur, bu nedenle bir stüdyo süper yıldıza kâra beklenen katkısını öderse, film büyük olasılıkla para kaybedecektir.

4. Film Süresi

Kim 2013'te içeriğin yanı sıra, filmin uzunluğunun da teatral başarı üzerinde bir etkisi olduğunu bulmuştur – film ne kadar uzunsa, o kadar çok izleyici çeker.

5. Önemli Dağıtımcı Şirketler

Büyük stüdyoların/distribütörlerin 20 tercihli tiyatrolara daha iyi erişime ve daha kapsamlı dağıtım bağlantılarına sahip olması beklenebilir. Bu akıl yürütme,

hepsi de ana dallardan biri tarafından filmin teatral başarısı üzerinde önemli olumlu bir etki bulan Litman, Bagella ve Becchetti veya Pangaker ve Smith'in bulgularıyla uyumludur. Buna ek olarak, Chang ve Ki 2005'te filmin uzunluğunun çoğunlukla dağıtımcıya bağlı olduğunu buldu: büyük dağıtımcılar tarafından yayınlanan filmler sinemalarda daha uzun süre gösterildi.

İlginç bir şekilde, Karniouchina ve diğerlerinin 2010'da yaptıkları analizi, büyük yapımcıların ve dağıtımcıların, bütçeler ve ekranlar kontrol edildiğinde daha düşük gelir ve karlarla ilişkili olduğunu, yani daha yüksek bütçeler kullanma eğiliminde olmalarına rağmen, daha başarılı filmler üretmediklerini göstermektedir.

Büyük stüdyoların etkisine odaklanan literatür oldukça azdır. Aynı şey ortak yapımlar için de geçerli. Ancak Bozdoğan'ın 2016'da çalışmasından bahsetmeye değer – onun bulgularına göre ortak yapımlar Fransız sinemalarında tiyatro gelirleri açısından daha kötü performans gösteriyor.

D. Veri Seti

Bölüm 3'te daha önce bahsedildiği gibi, temizlenmiş veri seti 18 öznitelikten (Çizelge 7) ve filmin IMDB derecelendirmesini temsil eden 8 farklı sınıf etiketinden oluşur (2'den 9'a kadar). Nihai veri seti 4.883 filmden oluşuyor.

Çizelge 7 Nihai veri kümesinin nitelikleri ve türleri

Özellik Adı	Özellik Türü
Erkek oyuncular	Devamlı
Görüntü yönetmenleri	Devamlı
Besteciler	Devamlı
Kostüm tasarımcıları	Devamlı
Yönetmenler	Devamlı
Film Distribütörleri	Devamlı
Editörler	Devamlı
Tür	Devamlı
Anahtar kelimeler	Devamlı
Yapımcılar	Devamlı
Üretim şirketleri	Devamlı
Üretim tasarımcıları	Devamlı
Filmin süresi	Ayrık
Ses-miks	Devamlı
Özel efekt şirketleri	Devamlı
Yazarlar	Devamlı
Yıl	Ayrık
Kadın oyuncular	Devamlı

Verilerdeki eksik değerlerle başa çıkmak için, her durumda Ci, Ai özniteliğinin her eksik değerini, Ci ile aynı sınıf etiketine sahip tüm durumların Ai'nin ortalama değeri ile değiştiririz. Örneğin, 5 dereceli bir M filmimiz olduğunu ve editör özniteliğinin eksik olduğunu varsayalım. Derecelendirmesi 5 olan tüm filmlerin tüm editör değerlerinin ortalamasını alıyoruz. Bu ortalama, editörü için bir değer olarak M'ye atanır.

E. K-Nearest Neighbors Algoritması

K-En Yakın Komşular (KNN) örnek tabanlı öğrenme özeliğine sahip makine öğrenmesi algortimasıdr. KNN tekniğinin temel çalışma prensibi daha önceki benzer durumlara dayanan yeni aşamanı sınıflandırmaktır.

KNN, yeni bir örneğin benzer durumlarını bu örneğin komşuları olarak tanımlar. Bu teknikte benzerliği tanımlamak için birçok yaklaşım kullanılmaktadır. Durumlar arasındaki Öklid mesafesi böyle bir yaklaşımdır. A ve B durumları arasındaki Öklid Uzaklığı aşağıdaki formülle hesaplanır:

$$E(A, B) = \sum_{i=0}^{n} \sqrt{(a_i - b_i)^2} \quad \text{(Denklem 2)}$$

Burada a_i ve b_i, sırasıyla A ve B'nin i'nci niteliğinin değerini temsil eder ve n, niteliklerin toplam sayısıdır. Dolayısıyla, A ve B durumları, karşılık gelen niteliklerinde az da olsa farklılık gösteriyorsa, komşu olarak kabul edilir.

Yeni bir N örneğini sınıflandırmak için algoritma, N ile eğitim setindeki her bir durum arasındaki yukarıda gösterildiği gibi Öklid mesafesini hesaplayarak çalışır. Komşular Kümesinden en yakın K örnekleri seçilir ve N bu örneklere göre sınıflandırılır (Musa, 2013: 52). Sınıflandırma ayrıysa, N, Komşular Kümesi'nde gösterilen çoğunluk sınıfı etiketine atanır. Sınıflandırma sürekli ise Komşu Kümedeki sınıf etiketlerinin ortalaması hesaplanır ve sonuç N'ye atanır. Şekil 4, KNN'nin sözde kodunu gösterir.

KNN (Training Set, Testing Instances):

```
Foreach Testing_Case t:
K_neighbors = t.Compute_Ecludian(Training_Set,K);
If(Discrete):
t.Classlabel = K_neighbors.get_most_existing_Classlabel;
if(Continuous):
t.ClassLabel = K_neighbors.getAverage();
```

Şekil 4 KNN Algoritması

Örneğin, Çizelge 8'de gösterilen filmlerimiz olduğunu varsayalım; burada başlık filmin başlığını, çalışma süresi toplam yayınlanma süresini dakika olarak, tür derecelendirmesi kendi türünün ortalama derecelendirmesini ve son olarak da filmin puanını temsil eder.

Çizelge 8 Veri seti ve bir filmin puanının KNN ile tahmini

ID	Filmin original adı	Film süresi	Film türünün puanı	Puan
B_1	Taking the Turn (2006)	134	5.35	6.9
B_2	Craig Shoemaker: Daditude (2012)	82	6.98	7.4
B_3	Angel of Nanjing (2015)	70	7.21	7.9
B_4	One Square Mile (2014)	96	6.28	6.3
B_5	Who Else to Blame? (2011)	97	4.80	3.7
B_6	Dead Last (2001)	60	6.33	7.8
B_7	Batman v Superman: Dawn of Justice (2016)	151	5.75	?

KNN kullanarak B_7'in derecesini tahmin etmek için, ilk görev Çizelge 8'de gösterilen B_7 ve M_{id} (id=1…6) arasındaki Öklid mesafesini hesaplamaktan ibarettir.

$$E(B_7,B_{id})= \sqrt{(\text{film\_süresi}(B_7) - \text{film\_süresi}(B_{id}))^2} + \sqrt{(\text{tür}(B_7) - (\text{tür}(B_{id}))^2}$$

$E(B_7,B_1) = \sqrt{(151 - 134)^2} + \sqrt{(5.75 - 5.35)^2} = 17.4$

$E(B_7,B_2) = \sqrt{(151 - 82)^2} + \sqrt{(5.75 - 6.98)^2} = 70.23$

$E(B_7,B_3) = \sqrt{(151 - 70)^2} + \sqrt{(5.75 - 7.2)^2} = 82.45$

$E(B_7,B_4) = \sqrt{(151 - 96)^2} + \sqrt{(5.75 - 6.28)^2} = 55.53$

$E(B_7,B_5) = \sqrt{(151 - 97)^2} + \sqrt{(5.75 - 4.8)^2} = 54.95$

$E(B_7,M_6) = \sqrt{(151 - 60)^2} + \sqrt{(5.75 - 6.33)^2} = 91.58$

Yukarıdaki sonuçlar göz önüne alındığında, 1NN kullanılarak, B_7'in en yakın komşusu B_1'dir, dolayısıyla B_7'in tahmin edilen derecesi 6.9'dur.

K = 2 kullanıldığında, B_7'in en yakın iki komşusu B_1 ve B_5'dır ve dolayısıyla sınıflandırma şöyle olur:

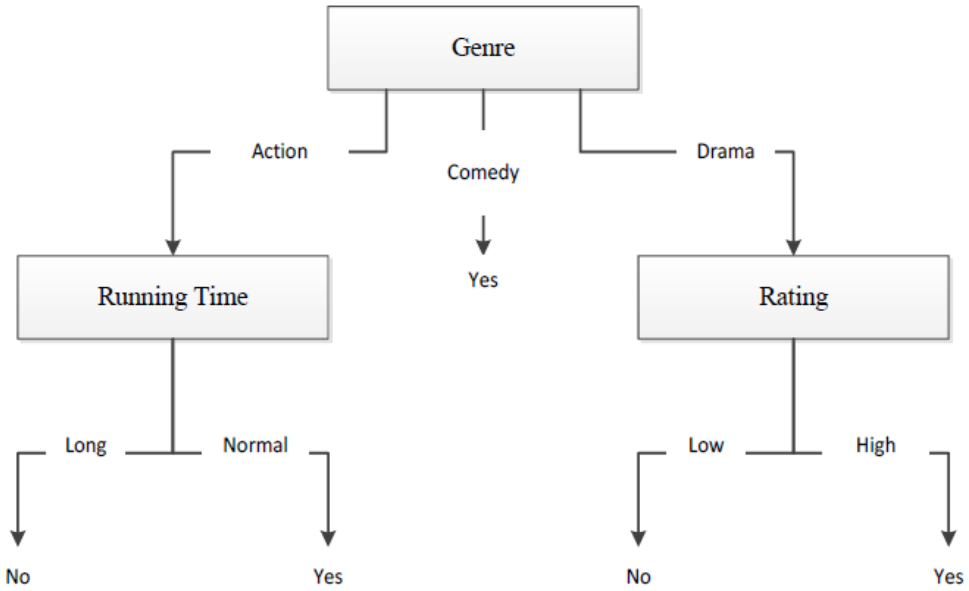
2NN = $(R_1+R_5)/2$ = 5,3 burada R_1 ve R_5, sırasıyla B_1 ve B_5 değerleridir

F. Decision Tree Algoritması

Karar ağacı öğrenme, tümevarımsal çıkarım algoritmalarına dayalı bir makine öğrenme tekniğidir (Meenakshi, vd. 2018).

Birkaç algoritma, sınıflandırma amacıyla karar ağacı öğrenmesini kullanır. Bu bölümde karar ağaçlarının ne olduğunu açıklayarak başlıyoruz, ardından ID3'ü tanımlıyoruz. Çalışmamızda kullandığımız karar ağacı öğrenme algoritması.

Karar ağacı: Bir dizi örnek verildiğinde, bir karar ağacı her bir örneği belirli kriterlere göre sınıflandırır. Bu yöntem, ayrık değerli hedef fonksiyonları tahmin etmek veya tahmin etmek için kullanılır ve bu kapsamda her bir karar ağacı öğrenilen bir fonksiyonu temsil eder. Bir karar ağacı, her biri bir teste (duruma) karşılık gelen bir dizi düğümden oluşur. N_i düğümündeki test başarılı olursa, bir alt düğümdeki test gerçekleştirilir. Bu bir yaprağa kadar devam eder. Karar ağaçlarındaki yapraklar sınıf etiketini temsil eder (Abdoulaziz, 2019: 46). Şekil 5'de bir karar ağacı örneği gösterilmektedir; burada sorun, niteliklere, türe, derecelendirmeye ve yayınlanma süresine dayalı olarak bir filmin izlenip izlenmeyeceğine karar vermektir. Bir karar ağacından hipotezler çıkarabiliriz. Bu örnekte bir hipotez şudur: "Türü "aksiyon" ise ve "normal" bir çalışma süresine sahipse bir film izleyebilirim".



Şekil 5 Tür, film süresi ve Derecelendirmeye göre bir filmin izlenip izlenmeyeceğine karar vermek için karar ağacı.

Yukarıda bahsedilen karar ağacında, ayrık değerli niteliklere sahibiz (örneğin, Derecelendirme: düşük, yüksek), ancak çoğu durumda, nitelikler sürekli değerlere

sahiptir. Bir özniteliğin sahip olabileceği çok sayıda farklı sayısal değer göz önüne alındığında, tüm bu değerleri tek tek bir karar ağacına koymak imkansızdır.

Karar ağaçları çok büyük ve karmaşık hale gelebilir ve bu da insan uzmanlar tarafından yorumlanmalarını zorlaştırır. Bunun için bazı alanlarda onları kural kümelerine dönüştürmeye ihtiyaç vardır. Kurallardan oluşan bir kural seti. Kural, niteliklerdeki koşulların ve bir sınıf etiketinin birleşimidir. Örneğin, Şekil 6, Şekil 5'deki karar ağacından çıkarılan üç kuralı göstermektedir. Kökten yaprak düğüme giden bir yoldaki tüm koşulların tek bir bağlantıda birleştirilmesiyle bir kural oluşturulur.

Kural 1:	**Kural 2:**	**Kural 3:**
If Tür = Aksiyon	If tür = aksiyon	If tür = komedi
Film süresi = uzun	Film süresi = Normal	Izlenmeli = evet
İzlenmeli = hayır	Izlenmeli = evet	

Şekil 6 Bir karar ağacından çıkarılan bir kural seti

Kural kümeleri, bazı koşullar ortadan kaldırılarak budanabilir. İki budama tekniği vardır: Yalnızca kuralın doğruluğunu olumsuz yönde etkilemiyorsa (verileri doğru bir şekilde sınıflandırma potansiyeli) bir koşulun kaldırılmasından oluşan "Post-pruning". Başka bir budama türü, ağacı bir kural kümesine dönüştürmeden hemen önce budayan "Error-reduced Pruning" olarak adlandırılır. "Error-reduced Pruning", ağaçtaki bir düğümü, alt ağacının en popüler sınıf etiketiyle değiştirir. Bu değişikliği yalnızca ağacın doğruluğu etkilenmezse koruruz. Örneğin, Şekil 5'deki Çalışma Süresi düğümünü yalnızca ağacın doğruluğu bozulmazsa "Hayır" ile değiştiririz.

ID3: ID3, vakaları sınıflandırmak için karar ağaçlarını kullanan makine öğreniminde en yaygın kullanılan tekniklerden biridir. ID3'teki ilk adım, en iyi sınıflandırıcı niteliğini seçmektir. İkincisi, ağacı bu özniteliğe böldükten sonra ne kadar bilgi kazanıldığı ile tanımlanır (Quinlan J. R., 1993).

Bunun için ID3, entropi adı verilen bilgi teorisinde iyi bilinen bir konsepte dayanır. Entropi, bir dizi verinin saflığını veya saflığını karakterize eder. Makine öğrenimi ve tahmin alanında, entropi, hedef işlevine dayalı olarak verilerin ne kadar iyi tahmin edilemez olduğunu ifade eder. İlgi alanımız, verilerin entropisini azaltmaktır (Shraddha, vd. 2015: 298).

Bir veri seti, S ve bir ikili hedef fonksiyonu (+, -) verildiğinde, S'nin entropisi şöyledir:

$$\text{Entropy}(S) = -P_{(+)} \lg P_{(+)} - P_{(-)} \lg P_{(-)} \text{ (Denklem 3)}$$

Burada P (+) ve P (-) sırasıyla pozitif ve negatif değerlerin oranıdır. (lg, log2'dir)

Denklemden çıkarabiliriz. 1, tüm durumlar aynı sınıf etiketine sahip olduğunda (bu durumda, P+ veya P- eşittir 1) entropinin değeri 0'dır, bu da S'nin kolayca tahmin edilebileceği anlamına gelir. Bu denklem, veri setinde 2'den fazla sınıf etiketi olduğunda kolayca genelleştirilebilir:

$$\text{Entropy}(S) = \Sigma_{c \in ClassLabels} -P_c \lg P_c \text{ (Denklem 4)}$$

ID3'ün amacı, sınıflandırma sürecini kolaylaştırmak için mümkün olan en düşük entropiye sahip alt kümelerle sonuçlanan özniteliği seçmektir. yani alt kümelerin her biri mümkün olduğunca homojen olmalıdır. Bu amaçla, ID3, her bir öznitelik için, veri kümesi bu özniteliğe dayalı olarak bölünürse beklenen entropi azalmasını hesaplar. Bu beklenen azalma, A niteliğinin Bilgi Kazanımı olarak adlandırılır.

A'nın bir S kümesine göre bilgi kazancı aşağıdaki fonksiyon tarafından verilir:

$$\text{Info}(S, A) = -\Sigma_{a \in A} \frac{|Sa|}{|S|} Entropy(S_a) \text{ (Denklem 5)}$$

S'nin A'ya göre bölünmesinden elde edilen kazanç daha sonra aşağıdaki gibi hesaplanır:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \text{Info}(S, A) \text{ (Denklem 6)}$$

En yüksek kazanıma sahip öznitelik daha sonra karar ağacındaki bir düğümden seçilir. Düğüm, A ve bir değerden oluşan bir koşulu kodlar. İşlem, tüm nitelikler tamamen tükenene kadar ağaçtaki her düğüm için tekrarlanır.

Örneklemek için, Çizelge 9'da gösterilen veri setini düşünelim; her bir vaka kendi türüne (Komedi, Drama, Aksiyon), yayınlanma süresine (Kısa, Uzun, Ortalama), bütçesine (Düşük, Yüksek) ve filme uygun olup olmadığına göre karakterize edilen bir filmi temsil eder. başarılıydı ya da değildi (1 = başarı, 0 = başarısızlık). ID3'ün ilk adımı, tüm veri kümesinin entropisini hesaplamaktır.

$\text{Entropy}(S) = -3/6 \log_2 3/6 - 3/6 \log_2 3/6 = 1.$

Bu durumda, veri seti sınıf etiketine göre eşit olarak bölünür. 1'in entropisi, bir durumu (0 veya 1'e) sınıflandırmak için 1 bitin gerekli olduğunu gösterir.

Çizelge 9 Üç öznitelikle tanımlanan bir film veri seti: Tür, Film süresi, Bütçe ve Başarıyı gösteren sınıf etiketi

ID	**Film türü**	**Film süresi**	**Bütçe**	**Başarısı**
1	Komedi	Kısa	Düşük	1
2	Dram	Kısa	Düşük	1
3	Aksiyon	Uzun	Yüksek	1
4	Aksiyon	Uzun	Düşük	0
5	Aksiyon	Ortalama	Düşük	0
6	Dram	Ortalama	Yüksek	0

Sonraki adım, her bir öznitelik için bilgi kazancını hesaplamaktır. Tür özniteliği ile başlayarak, veri seti Çizelge 10'de gösterilen 3 alt gruba ayrılabilir.

Entropy (*Komedi*) = -1/1 $\log_2$ 1/1 – 0/1 $\log_2$ 0/1 = 0

Entropy (*Dram*) = -1/2 $\log_2$ 1/2 – 1/2 $\log_2$ 1/2 = 1

Entropy (*Aksiyon*) = -1/3 $\log_2$ 1/3 – 2/3 $\log_2$ 2/3 = 0.918

Info(S, Tür) = $-|S_{komedi}|/|S|\ Entropy(S_{komedi}) - |S_{dram}|/|S|\ Entropy(S_{dram})$

$-|S_{aksiyon}|/|S|\ Entropy(S_{aksiyon}) = -\frac{1}{6}*0 - \frac{2}{6}*1 - \frac{3}{6}*0.918$ = -0.792

Gain(S, Genre) = 1+ (– 0.792) = 0.208

Çizelge 10 Tür kriterine göre bölünen alt veri kümeleri

Alt kümeler				
1	*Komedi*	Kısa	Düşük	1
2	*Aksiyon*	Uzun	Yüksek	1
3	*Aksiyon*	Uzun	Düşük	0
4	*Aksiyon*	Ortalama	Düşük	0
5	*Dram*	Kısa	Düşük	1
6	*Dram*	Ortalama	Yüksek	0

Film süresine gelince, alt kümeler Çizelge 11'da gösterilmektedir.

Entropy (*Kısa*) = -2/2 $\log_2$ 2/2 – 0/2 $\log_2$ 0/2 = 0

Entropy (*Ortalama*) = -0/2 $\log_2$ 0/2 – 2/2 $\log_2$ 2/2 = 0

Entropy (*Uzun*) = -1/2 $\log_2$ 1/2 – 1/2 $\log_2$ 1/2 = 1

Çizelge 11 Film süresine göre bölünen alt veri kümeleri

Alt kümeler				
1	Komedi	*Kısa*	Düşük	1
2	Dram	*Kısa*	Düşük	1
3	Aksiyon	*Uzun*	Yüksek	1
4	Aksiyon	*Uzun*	Düşük	0
5	Aksiyon	*Ortalama*	Düşük	0
6	Dram	*Ortalama*	Yüksek	0

Info(S, Filmin süresi) = $-|S_{\text{uzun}}|/|S|\ Entropy(S_{\text{uzun}})-|S_{\text{kısa}}|/|S|$

$Entropy(S_{\text{kısa}})-|S_{\text{ortalama}}|/|S|\ Entropy(S_{\text{ortalama}}) = -\frac{2}{6}*1-\frac{2}{6}*0-\frac{2}{6}*0$ = -0.333

Gain(S, Filmin süresi) = 1 + (– 0.333) = 0.667

Bütçe özniteliğine gelince, alt kümeler Çizelge 12'da gösterilmektedir.

Entropy (*Düşük*) = -2/4 $\log_2$ 2/4 – 2/4 $\log_2$ 2/4 = 1

Entropy (*Yüksek*) = -1/2 $\log_2$ 1/2 – 1/2 $\log_2$ 1/2 = 1

Çizelge 12 Bütçeye göre bölünen alt veri kümeleri

Alt kümeler				
1	Komedi	Kısa	*Düşük*	1
2	Dram	Kısa	*Düşük*	1
3	Aksiyon	Uzun	*Düşük*	0
4	Aksiyon	Ortalama	*Düşük*	0
5	Aksiyon	Uzun	*Yüksek*	1
6	Dram	Ortalama	*Yüksek*	0

Info(S, Bütçe) = $-|S_{\text{yüksek}}|/|S|\ Entropy(S_{\text{yüksek}})-|S_{\text{düşük}}|/|S|\ Entropy(S_{\text{düşük}}) = -\frac{4}{6}*1-\frac{2}{6}*1$ = -1

Gain (S, Bütçe) = 1 – 1 = 0

Yukarıda hesaplanan sonuçlar göz önüne alındığında, en yüksek kazanıma sahip nitelik Çalışma süresidir. Bu öznitelik daha sonra C5 tarafından ağacın kökü olacak şekilde C5 tarafından seçilir. Koşullar, bu özelliğin olası her değerine karşılık gelir.

Ardından, her adımda (her alt düğüm için), ağaç tamamen oluşturulana kadar yukarıdaki işlem alt kümelerde tekrarlanır. Çalışma süresinin seçilmesi Çizelge 11'de gösterildiği gibi üç alt küme düğümü (Kısa, Uzun ve Ortalama) ile sonuçlanır.

Çizelge 11'in Alt Küme 2'yi (S_2) göz önünde bulunduralım (Çalışma Süresi için =" Uzun").

Entropy (S_2) = -1/2 lg 1/2 – 1/2 lg 1/2 = 1

S_2'yi Bütçeye göre bölüyorsanız:

Entropy (*Düşük*) = -1/1 $\log_2$ 1/1 = 0

Entropy (*Yüksek*) = -1/1 $\log_2$ 1/1 = 0

Info (S_2, Bütçe) = $-|S_{2\text{yüksek}}|/|S_2|\ Entropy(S_{2\text{yüksek}}) - |S_{2\text{düşük}}|/|S_2|$

$Entropy(S_{2\text{düşük}}) = -\frac{1}{2}*0 - \frac{1}{2}*0 = 0$

Gain (S_2, Bütçe) = 1

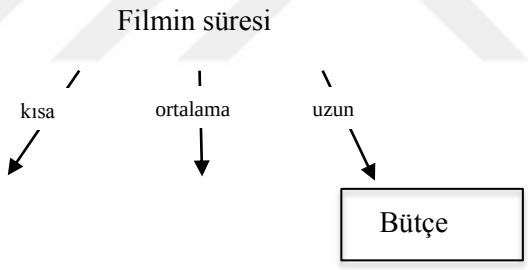
S_2'yi Türe göre bölüyorsanız:

Entropy (*Aksiyon*) = -1/2 lg 1/2 -1/2 lg 1/2 = 1

Info (S_2, Tür) = $-|S_{2aksiyon}|/|S_2|\ Entropy(S_{2aksiyon}) = -\frac{2}{2} * 1 = -1$

Gain (S_2, Tür) = 0

Şimdiye kadar inşa edilen ağaç Şekil 7'te gösterilmektedir.



Şekil 7 Çizelge 2'de sunulan veri seti üzerinde C5'in iki adımından sonra oluşturulan kısmi karar ağacı

ID3'ün en son devamı olan C5, algoritmayı bir winnowing özelliği ile çalıştırma imkanı sağlar. C5'teki harmanlama tekniği, ağacı oluşturmadan önce kullanılan özniteliklerin sayısını azaltmaya çalışır. Bu teknik, verileri rastgele iki yarıya bölerek çalışır. İlk yarı, C5 normal sürecini kullanarak bir ağaç oluşturmak için kullanılır. Diğer yarısı ise ağacı budamak için kullanılır. Budanan nitelikler alakasız olarak kabul edilir ve kazanılmış nitelikler olarak adlandırılır. C5'in son ağacı, kazanılmış öznitelikler olmadan ilk veri seti kullanılarak oluşturulur.

Sürekli nitelikler için değerler, her bir nitelik için kesme noktaları (eşikler) hesaplanarak aralıklara dönüştürülür. Bu kesme noktaları, test düğümlerini

oluşturmak için kullanılır. Hesaplama eşikleri için sözde kod Şekil 8'de gösterilmiştir.

```
Cutpoints (Data):
Foreach attribute A in Data:
Order(Data) based on A;
Foreach case Ci in Data:
if Ci.ClassLabel ≠ Ci+1.ClassLabel:
A.ThreshHold[i] = (Ci.A + Ci+1.A)/2;
```

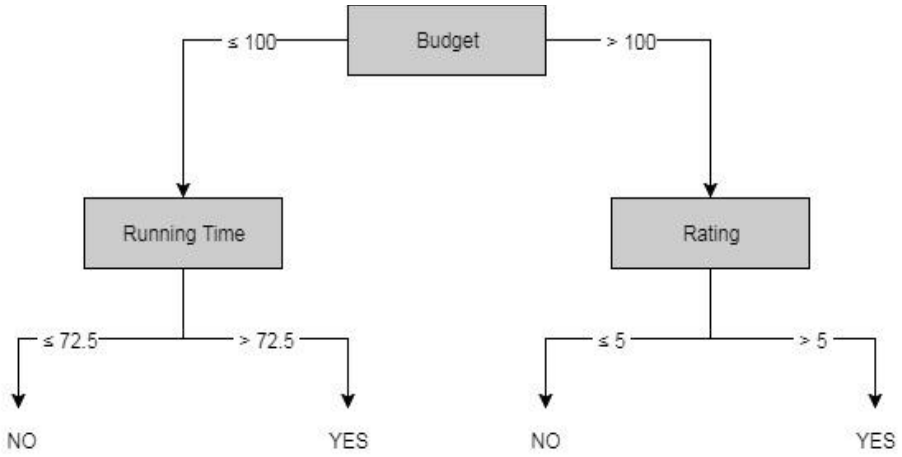
Şekil 8 Bir veri kümesi "Veri"den eşiklerin nasıl çıkarılacağını gösteren algoritma

Örneğin, Çizelge 13'te gösterilen veri kümesinin Film Süresi özniteliğine göre sıralandığını varsayalım. Bu sette, Film süresi özniteliği için üç farklı kesme noktası değeri vardır. Bunlar, sınıf etiketinin değiştiği ardışık iki öznitelik değerinin ortalaması olarak hesaplanır. Kesim noktası 1 (70+75)/2 = 72,5'tir. Kesim noktası 2 (85+100)/2 = 92,5'tir. Kesim noktası 3 (100+150)/2 = 125'tir.

Çizelge 13 Bir "Filmin süresi" özniteliği ve filmin izlenip izlenmeyeceğini gösteren bir sınıf etiketi içeren veri seti

Film süresi	Izlenmeli
50.2	1
70	1
75	0
85	0
100	1
150	0

Sürekli değerli nitelikler kullanılarak oluşturulan bir karar ağacı örneği Şekil 9'da gösterilmektedir.

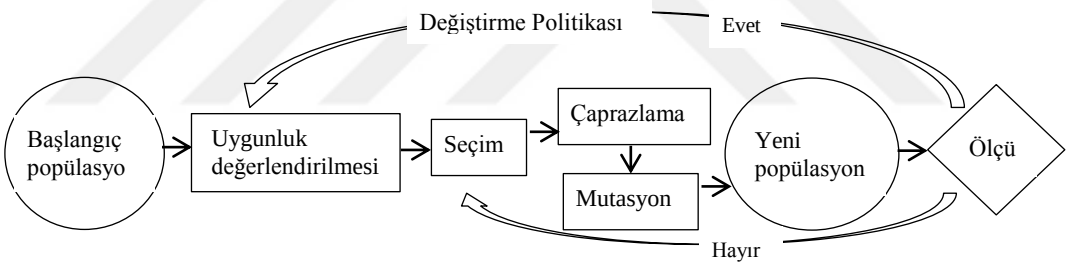


Şekil 9 Tür, Film süresi ve Derecelendirmeye dayalı bir film izleyip izlememe karar ağacı sayısal veriler olarak.

G. Genetik Algoritma

Genetik Algoritma meta-sezgisel özellikli, yaygın kullanılan yöntemdir. Holland tarafından 1960'larda keşf edilen bu teknik doğal adaptasyon sürecini irdelemek için kullanılmıştır. Genetik algoritmada problemin çözümleri arasından en uygununu bulmak için doğanın seçilim, güçlü olanın hayatta kalma gibi özellikleri kullanılmaktadır.

GA'nın ilk adımında kromozomlar populasyonu oluşturuluyor. Buradaki kromozomların her biri olası çözümü simgelemektedir ve çözümün yeterliliğini tanımlayan uygunluk değerini kullanmaktadır. Kromozomun hayatta kalma olasılığı ona karşılık gelen uygunluk değerine bağlıdır. GA, mevcut popülasyondan uygunluklarına göre kromozomları seçer, ardından yeni kromozomlar oluşturmak için belirli genetik işlemleri gerçekleştirir. İstenen bir çözüme ulaşılana veya başka bir durdurma kriterine ulaşılana kadar süreç devam eder. Normalde, böyle bir kriter probleme bağlıdır. Şekil 10, bir GA sürecini tanımlayan bir akış şemasını göstermektedir.



Şekil 10 GA şeması

1. Kodlama Şeması

Biyolojide kromozom, her biri belirli bir özelliği (örneğin saç rengi) temsil eden bir gen dizisidir. GA'daki en önemli kısım, GA'nın performansını etkileyen kodlama şeması ve optimal çözümü ne kadar hızlı bulabileceğidir. Kodlama şeması, kromozomların temsiline karşılık gelir.

Başlangıçta, kromozomların temsili, her bir genin 0 veya 1 değerine sahip olduğu ikili kodlamaya dayanıyordu. Çizelge 14, ikili kodlama kullanılarak temsil edilen üç kromozomu göstermektedir.

Çizelge 14 İkili sayı kodlama ile kromozom gösterimi

Kromozom ID	**Kromozom gösterimi**
C1	1 0 0 0 1 0 0 1
C2	1 1 1 1 1 1 0 0
C3	1 0 0 0 0 0 0 1

Genetik Algoritmalar üç temel öğe içerir: popülasyon, uygunluk fonksiyonu ve GA operatörleri.

Kromozom popülasyonu, GA'nın arama uzayını temsil eder. Bu, rastgele veya başka bir algoritma (örneğin açgözlü yaklaşımlar veya diğer buluşsal yöntemler) kullanılarak oluşturulabilir.Uygunluk fonksiyonu, bir çözümün diğerlerine kıyasla ne kadar iyi olduğunu tanımlar ve eldeki probleme göre formüle edilir.

GA operatörleri, GA'nın mevcut kromozomlardan yeni kromozomlar oluşturmasına izin verir.

2. Genetik Operatörler

Tipik olarak, GA iki farklı genetik işleme dayanır: "Crossover" ve "Mutasyon". İlki, yenilerini oluşturmak için genleri değiştirerek kromozomların birleştirilmesinden oluşur. Sonuncusu belirli bir kromozomu değiştirir.

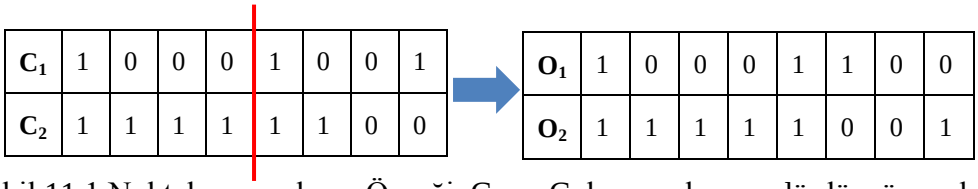
3. Çaprazlama

Çaprazlama, yeni kromozomlar üreten adım olduğu için üreme sürecinin özüdür. Üremede çaprazlama, iki kromozomun genlerini değiştirerek iki yeni yavru oluşturmak için yeniden birleştirildiği süreçtir. Her yeni yaratılan yavru, her iki ebeveynden gelen genlerin bir kombinasyonunu içerir. Çaprazlama belirli bir olasılıkla gerçekleşir. Çaprazlamanın amacı, popülasyona yayılmış iyi özellikleri tek bir kromozomda birleştirmektir.

Çaprazlamanın en popüler sürümleri şunlardır: 1-Noktalı Çaprazlama, N-Noktalı Çaprazlama ve Tekdüzenli Çaprazlama.

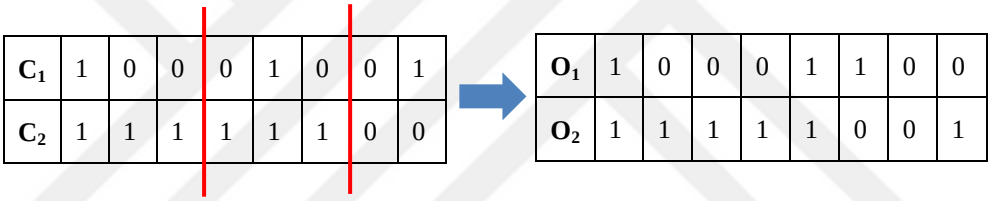
1-Noktalı çaprazlama yöntemi: 1-Point Crossover, bu operatörün ilk ve en yaygın kullanılan versiyonudur. Her iki ebeveyni de aynı yerde kesmekten ibarettir.

Daha sonra, birinci ebeveynin birinci kısmı ile ikinci ebeveynin ikinci kısmı birleştirilerek birinci yavru oluşturulur. İkincisi, ikinci ebeveynin ilk kısmı ile birinci ebeveynin ikinci kısmı birleştirilerek oluşturulur. Şekil 11, Çizelge 14'dan C1 ve C2 kromozomları arasındaki geçişi göstermektedir.



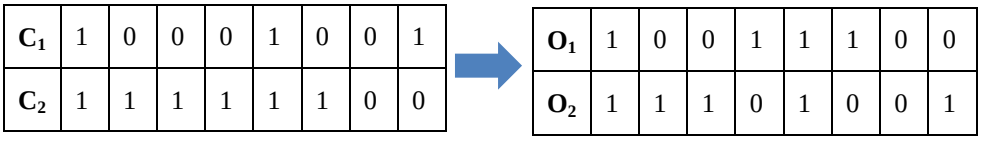
Şekil 11 1 Noktalı çaprazlama Örneği. C_1 ve C_2 kromozlarının dördüncü genden sonrası kesilmiştir.

N-Noktası çaprazlama yöntemi: Çaprazlamanın bu versiyonunda kromozomlar birden fazla yerde (N yerde) kesilir. Şekil 12, C_1 ve C_2'nin 2 Noktalı Geçişini göstermektedir. N-Point Crossover, ebeveynlerden gelen birden fazla parçanın birleştirilmesinden oluştuğu için oluşturulan çocuklarda daha fazla değişikliğe izin verir.



Şekil 12 2 Noktalı Çaprazlama örneği. C_1 ve C_2'nin üçüncü ve altıncı genden sonra kromozomlar kesilmiştir.

Üniforma çaprazlama yöntemi: Bu çaprazlama biçiminde, bir çocuk belirli bir geni ebeveynlerden birinden rastgele miras alır ve diğer çocuk, diğer ebeveynden karşılık gelen geni alır. Şekil 13, bu tür çapraz geçişin bir resmini göstermektedir.



Şekil 13 O_1'in C_1'den 1,2,3,5 ve 7. genlerini ve C_2'den 4, 6 ve 8. genlerini miras aldığı Uniform Çaprazlama örneği. O_2, C_1'den 4, 6 ve 8. genlerini ve C_2'den 1,2,3,5 ve 7. genlerini miras alır.

4. Mutasyon

Çaprazlamanın bir sınırlaması, tamamen yeni bir çözüm yaratmaması, ancak öncekileri birleştirmesidir. Bu gerçek bizi nihayetinde yerel optimal çözümlere götürebilir. Örneğin, yalnızca "1" ile başlayan kromozomları içeren bir popülasyon verildiğinde. Tek başına çaprazlama, yalnızca “1” ile başlayan kromozomlar

oluşturur. Optimal çözüm "0" ile başlarsa, GA onu asla bulamaz. Bu sorunun üstesinden gelmek için mutasyon kullanılabilir. Çaprazlamaya benzer şekilde, birkaç farklı mutasyon operatörü vardır: ikili gösterim için mutasyon, gerçek değerli gösterim için mutasyon, takas yoluyla mutasyon ve ters çevirme mutasyonu.

İkili gösterimde mutasyon, genin değerini 0'dan 1'e çevirmekten veya tam tersini içerir.

Gerçek değerli temsil için mutasyon, gen değerinin olası gen değerleri kümesinden rastgele seçilen bir değerle değiştirilmesinden oluşur (Şekil 14).

Şekil 14 Gerçek değerli temsil için mutasyon. Gen 7 mutasyona uğramıştır.

Değiştirme yoluyla mutasyon (Şekil 15), rastgele iki genin seçilmesi ve değerlerinin değiştirilmesinden oluşur. İnversiyon Mutasyonu (Şekil 16), bir gen segmentini tersine çevirmekten oluşur.

Şekil 15 Değiştirme yoluyla mutasyon. Gen 2 ve Gen 5 değiştirilir

Şekil 16 İnversiyon Mutasyonu. Kalın yazılan dize ters çevrilir

5. Seçim Prosedürü

Uygun kromozomların hayatta kalma şansının daha yüksek olması gerektiğinden, seçim prosedürü bir kromozomun uygunluğuna dayanır. Denk. 26, kromozom i için seçim olasılığını gösterir.

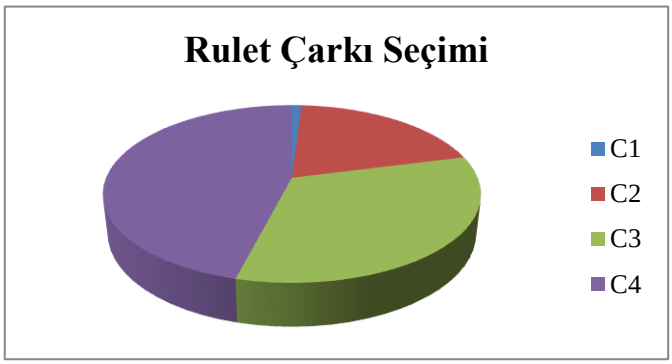
$$S_i = \frac{f_i}{\Sigma_{j \epsilon P} f_j} \quad \text{(Denklem 7)}$$

Burada f_i, i kromozomunun uygunluk değeridir ve P, kromozom popülasyonunun tamamıdır.

Bu prosedürün en yaygın kullanılan versiyonları “Rulet Çarkı Seçimi”, “K-Turnuva Seçimi” ve “Sıra Seçimi”dir.

Rulet Çarkı Seçimi: Bu seçim tekniğini göstermek için kromozomlar arasında bölünmüş bir rulet çarkı hayal edebiliriz. Her kromozoma, çarkın uygunluğuyla

orantılı bir parçası verilir. Bir zar atılır. Zarı alan kromozom daha sonra seçilir. Şekil 17 rulet çarkını ve Çizelge 15'de gösterilen kromozomlar arasındaki dağılımını göstermektedir. Açıktır ki, kromozom ne kadar uygun olursa, seçim olasılığı o kadar yüksek olur.



Şekil 17 Rulet Çarkı Örneği

Çizelge 15 Uygunluk değerleriyle birlikte kromozomlar

Kromozom	**Uygunluk değeri**
C1	2
C2	30
C3	50
C4	70

Sıra Seçimi: Rulet tekerleği seçimi kullanılarak, çok düşük uygunluğa sahip kromozomlar, eğer varsa, çok az seçilme şansı elde eder. Şekil 17'de gösterildiği gibi, C1 seçilme şansı sadece %1'dir. Bunun üstesinden gelmek için, sıra seçimi tüm kromozomları uygunluk değerlerine göre sıralar (en uygun kromozom en yüksek sırayı alır) ve daha sonra inceliklerden ziyade sıralara rulet çarkı uygulanır. Çizelge 16, Çizelge 15'deki kromozomları göreli sıralarıyla gösterir. Şekil 18 ise yeni rulet çarkı yerleşimini göstermektedir. Açıkçası, minimum seçilme şansı olan C1 şimdi daha iyi oluyor.

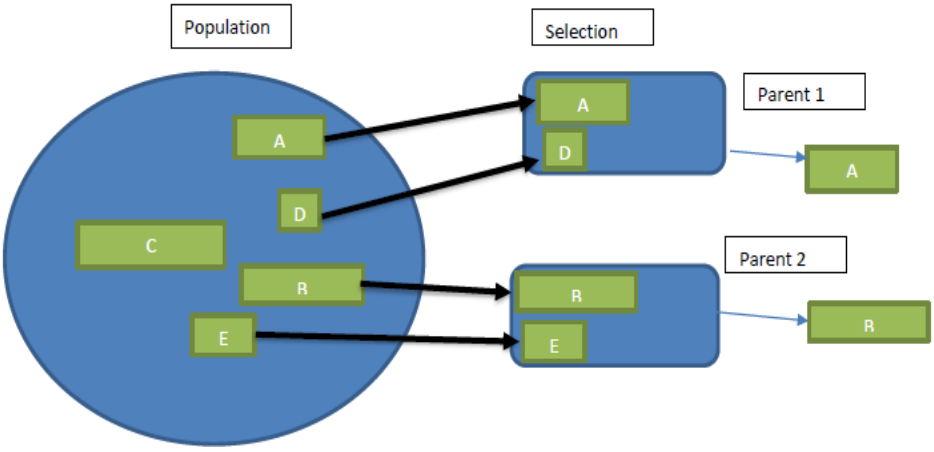
Çizelge 16 Dereceli Kromozomlar

Kromozom	**Uygunluk değeri**	**Sıra #**
C1	2	1
C2	30	2
C3	50	3

C4	70	4

Şekil 18 Kromozomlara sıralarına göre ayrılmış rulet çarkı

K-Turnuva Seçimi: K-Turnuva Seçimi, popülasyondan rastgele K kromozomlarının seçilmesi, en uygun olanın ilk ebeveyn olarak seçilmesi, diğer K-1'in popülasyona geri döndürülmesi ve diğer ebeveyni seçmek için işlemin tekrarlanmasından oluşur (Şekil 19).



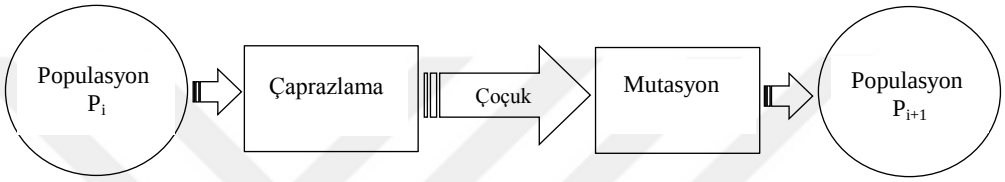
Şekil 19 2-Turnuva Seçimi Çizimi

K-Turnuva seçiminin bir avantajı, kromozomların popülasyonun tamamında değil, her bir alt kümede birbirleriyle rekabet etmesidir. Bu, düşük uygunluğa sahip kromozomların seçilme şansının daha yüksek olmasına neden olur. Örneğin, Şekil 19'da A kromozomu, yalnızca daha düşük bir uygunluğa sahip olan D ile rekabet etmektedir. Diğer seçim türlerinde A, her ikisi de daha yüksek uygunluk değerlerine sahip olan C ve B kromozomları dahil tüm popülasyonla rekabet etmek zorundaydı. Ancak bir dezavantajı, popülasyonda en düşük uygunluk değerine sahip kromozomun hiçbir zaman seçilememesidir. Bunun nedeni, bu kromozomun her zaman daha yüksek uygunluğa sahip diğerleriyle karşılaştırılmasıdır.

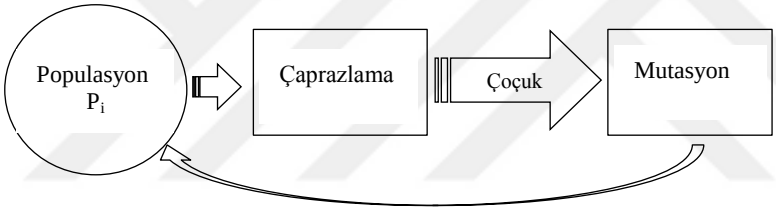
Değiştirme Politikası: Değiştirme politikası, mevcut nesilden hangi kromozomların değiştirileceğini tanımlar. "Generational GA" ve "Steady-State GA" iki farklı değiştirme politikasıdır.

Generational GA (Şekil 20), Holland tarafından tanıtılan ilk politikadır. Her nesilde, yeni oluşturulan yavrulardan oluşan, aynı büyüklükte tamamen yeni bir popülasyon yaratılmasından oluşur.

Kararlı Durum GA (Şekil 21), aynı popülasyonu tutmaktan, ancak eski, genellikle zayıf kromozomları yeni yavrularla değiştirmekten oluşur. Bu prosedür doğada meydana gelene daha yakındır.



Şekil 20 Kuşak GA'sının genel süreci (Generational GA)



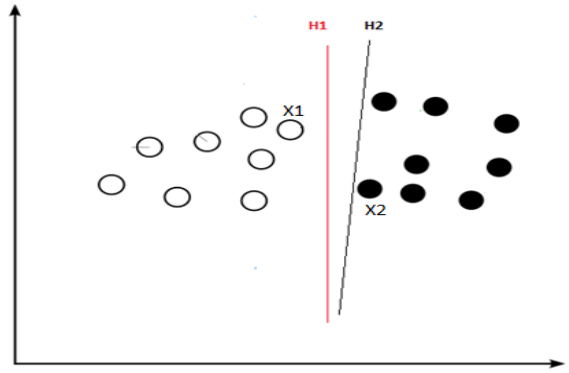
Şekil 21 Kararlı durum GA'sının genel süreci (Steady-State GA)

Steady-State GA'nın bir avantajı, yeni yavruların ebeveynleri ve diğer kromozomlarla yaratıldıkları anda rekabet etmesine izin vermesidir. Oysa kuşaksal GA, her kuşakta daha fazla çeşitlilik sağlar ve bu nedenle optimal çözüme daha hızlı yakınsamaya eğilimlidir. *Elitizm*, elit kromozomların nesilden nesile kopyalanması işlemidir. Prosedürün amacı, şimdiye kadar bulunan en iyi çözümlerin kaybolmamasını sağlamaktır.

H. Destek Vektör Makinesi

Destek Vektörleri Makinesi (SVM), 60'ların başında Vladimir Vapnik'in doktora tezinde ilk kez tanıtılan denetimli bir öğrenme tekniğidir. Mevcut versiyon Cortes ve Vapnik tarafından 1995'de icat edilmiştir. Durumlar veri noktalarıdır ve sınıflandırma etiketleri ikilidir. SVM'nin amacı, bu veri noktalarını etiketlerine göre bölmek için hiper düzlemler bulmak ve oluşturmaktır. Şekil 22, her rengin bir sınıf

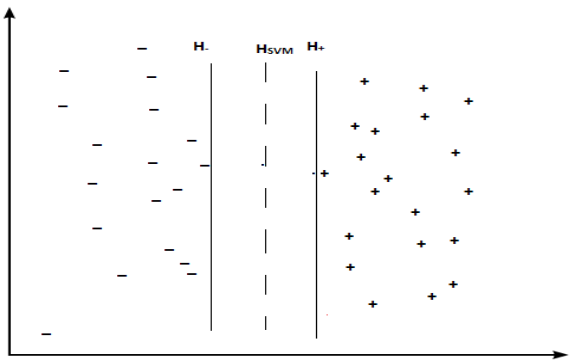
etiketini gösterdiği renklerine göre veri noktalarını doğru ve kesin olarak bölen iki hiper düzlemi (H_1 ve H_2) göstermektedir.



Şekil 22 İki hiperdüzlemle ayrılmış veri noktaları. Renkler sınıf etiketlerini gösterir.

Birkaç hiperdüzlem bu veri noktalarını bölebildiğinden, zorluk hangi hiperdüzlemin seçileceğini bilmektir. SVM, en büyük işlevsel marjı olan hiper düzlemi arar. İşlevsel marj, hiperdüzlem ile en yakın veri noktası arasındaki mesafe olarak tanımlanır (sınıf etiketine bakılmaksızın). Bu nedenle, Şekil 22'de, H_1 ile en yakın veri noktası X_1 arasındaki mesafe, H_2 ile en yakın veri noktası X_2 arasındaki mesafeden daha büyük olduğu için, SVM, H_2'ye göre H_1'i tercih eder.

Her biri pozitif (+) veya negatif (-) olarak etiketlenmiş n vakamız olduğunu varsayalım. SVM, sırasıyla pozitif veri noktalarına ve negatif veri noktalarına mümkün olan en yakın hiperdüzlemler olan iki hiperdüzlem H+ ve H-'yi arar (Şekil 23). H_+ ve H_- ile sınırlanan alana "sokak" denir. Bu veri noktalarını ayıran herhangi bir hiperdüzlem (istenen hiperdüzlem H_{SVM} dahil) sokak bölgesi içinde yer alacaktır ve H_{SVM}'nin tüm veri noktalarına en büyük mesafeye sahip olması, yani sokağın ortasında olması arzu edilir. Bu nedenle SVM, H_+ ve H_- (cadde genişliği) arasındaki mesafeyi maksimize etmeyi amaçlar.



Şekil 23 SVM tarafından hesaplandığı şekliyle H_+, H_- ve H_{SVM}

Aşağıda H_{SVM}, H_+ ve H_- hiperdüzlemlerinin denklemlerini gösterilmiştir, burada x bir durumu temsil eden vektör, w hiperdüzlemlere normal olan bir vektör ve b bir sabittir.

$$H_{SVM}: w*x + b = 0 \text{ (Denklem 8)}$$

$$H_+: w*x + b = +1 \text{ (Denklem 9)}$$

$$H_-: w*x + b = -1 \text{ (Denklem 10)}$$

Bilinmeyen bir u durumu verildiğinde, H_{SVM}'nin sağındaysa ve aşağıdaki karar kuralını karşılıyorsa pozitif olduğu söylenir:

$$w*u + b \geq 0 \text{ (Denklem 11)}$$

Problem, tüm x_+ pozitif veri noktaları için aşağıdakiler geçerli olacak şekilde optimal w vektörünü bulmaktan ibarettir:

$$w*x_+ + b \geq +1 \text{ (Denklem 12)}$$

Ve tüm negatif veri noktaları için x-:

$$w*x_- + b \leq -1 \text{ (Denklem 13)}$$

Yukarıdaki iki kısıtlamayı tek bir denklemde birleştirmek için yeni bir yi değişkeni kullanılır. Pozitif durumlar için +1 ve negatif olanlar için -1 değerine sahiptir. Yukarıdaki iki denklemi ile çarparak şunu elde ederiz:

$$y_i * (w*x^i + b) - 1 \geq 0 \text{ (Denklem 14)}$$

Hiperdüzleme ait olan herhangi bir xi için aşağıdaki Hiperdüzlem Nokta Kuralı geçerlidir:

$$y_i * (w*x_i + b) - 1 = 0 \text{ (Denklem 15)}$$

Caddenin genişliğini maksimize etmek için bu genişlik için bir denkleme ihtiyacımız var. Sırasıyla H_+ ve H_- üzerine yerleştirilmiş x_1 ve x_2 olmak üzere iki noktamız olduğunu varsayalım. Bu iki veri noktasının denklemleri sırasıyla:

$$w*x_1 + b = +1 \text{ (Denklem 16)}$$

$$w*x_2 + b = -1 \text{ (Denklem 17)}$$

Yukarıdaki denklemleri çıkarma ile ve w'nin büyüklüğüne bölerek aşağıdaki genişlik denklemini elde ederiz:

$$\frac{w * (x_1 - x_2)}{|w|} = \frac{2}{|w|}$$ (Denklem 18)

Amaç, $\frac{2}{|w|}$'ye eşit olan genişliği maksimize etmektir, bu nedenle, sorun w'yi en aza indirgemektir, bu da aşağıdakilere yol açar:

$$MAX\ (2/|w|) \equiv MIN\ (|w|) \equiv MIN\ (½\ |w|^2)$$ (Denklem 19)

Küçültmek daha kolay olduğu için MIN (|w|) yerine MIN (½ |w|2) kullanılır.

Şimdi sorun, bir kısıtlama altında bir ifadeyi en aza indirmektir (Hiperdüzlem Noktaları Kuralı. SVM, bir optimizasyon tekniği olarak Lagrange çarpanlarını kullanarak bu sorunu çözmektedir.

Lagrange çarpanları, ilk ifadeyi ve verilen kısıtlamaları tek bir ifadede birleştirerek farklı kısıtlamalara tabi belirli bir ifadeyi en üst düzeye çıkarmayı veya en aza indirmeyi amaçlar (Hand ve Finch, 1998). Bu durumda, amaç fonksiyonundan (½ $|w|^2$) bir sabit, α_i ile çarpılan tüm kısıtlamaları (Köprü Noktaları Kuralı) çıkararak ana denklemi türetir.

$$L = ½\ |w|^2 - \Sigma\ \{\alpha_i * [y_i * (w * x_i + b) - 1]\}$$ (Denklem 20)

Burada α_i kısıt çarpanı olarak adlandırılır.

w için minimum değeri bulmak, Denklemin türevinin bulunduğu değeri bulmaktan ibarettir.

$$\frac{dL}{dw} = w - \Sigma\ (\alpha_i * y_i * x_i) = 0 \rightarrow w = \Sigma\ (\alpha_i * y_i * x_i)$$ (Denklem 21)

$$\frac{dL}{db} = - \Sigma\ (\alpha_i * y_i) = 0 \rightarrow \Sigma\ (\alpha_i * y_i) = 0$$ (Denklem 22)

w'nin Denklem'deki değeriyle değiştirilmesi. SVM'nin ilk karar kuralında yukarıdaki ilk denklem, şunu elde ederiz:

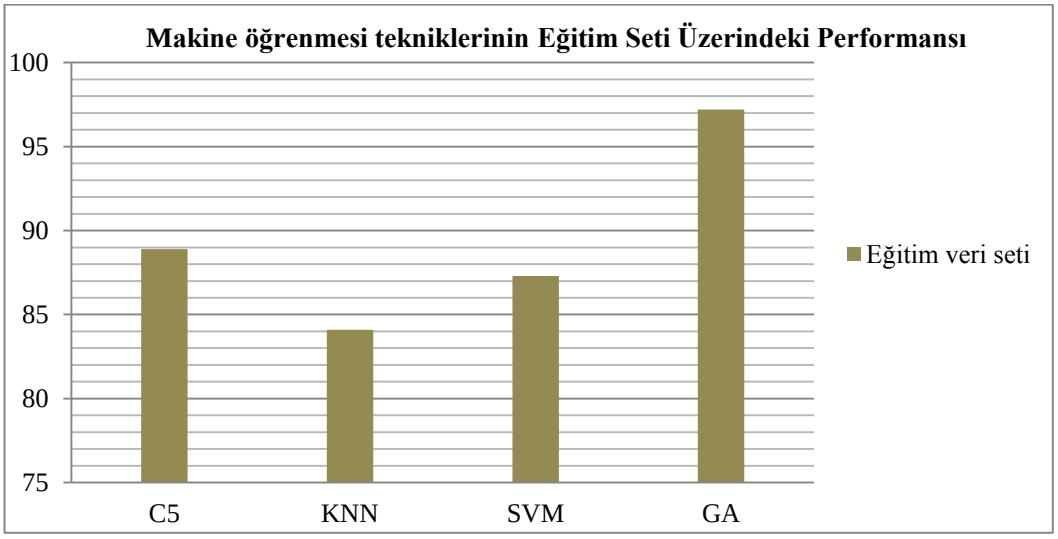
$\sum_{i=0}^{n}(\alpha_i * y_i * x_i) * x + b \geq 0$ ise x pozitiftir, aksi halde negatiftir. Denk. Bu son denklem yeni bilinmeyen durumları sınıflandırmak için kullanılan karar kuralıdır.

IV. BULGULAR

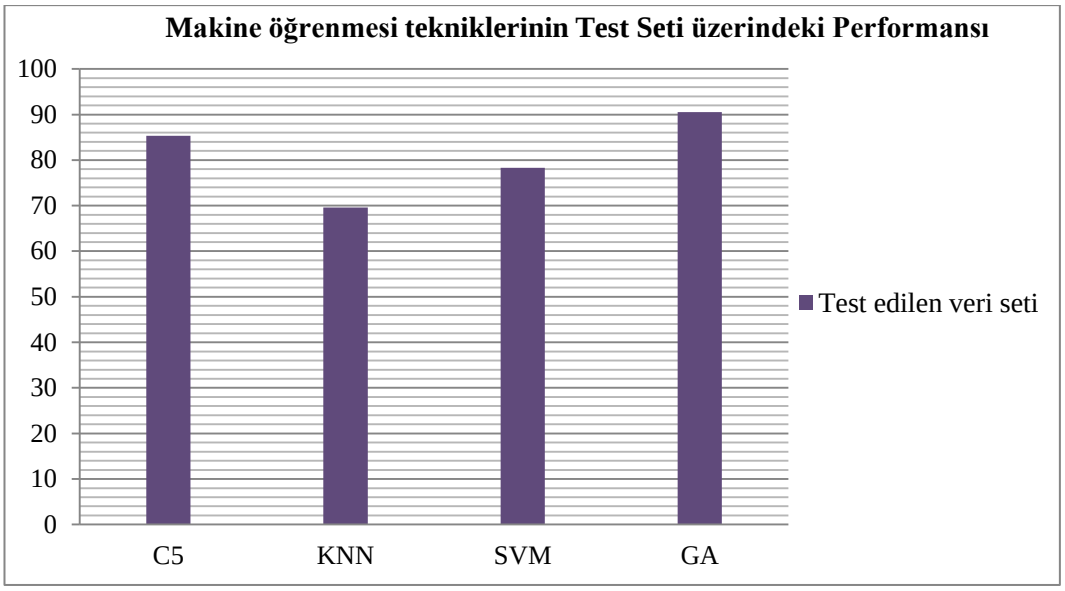
Çalışmada 10 katlı çapraz doğrulama tekniği kullanılmaktadır. Başlangıç veri kümesi rastgele şekilde yaklaşık olarak eşit oranda 10 farklı katmana bölünmüştür. Her kat Fi için algoritmalarımızı diğer 9 kat Fj'nin (her j ≠ i için) kombinasyonu üzerinde eğitiyoruz ve Fi üzerinde test ediyoruz. Her i için eğitim doğruluğunu ve test doğruluğunu hesaplıyoruz. Hem eğitim hem de test verilerinde ortalama doğruluğu standart sapma ile birlikte rapor ederiz (Çizelge 17). Ayrıca rastgele ve çoğunluk sınıflandırıcılarının sonuçlarını da dahil ediyoruz. Ortalama test ve eğitim doğruluğu açısından elde edilen en yüksek sonuçları kalın harflerle ve rastgele veya ana sınıflandırıcı dışındaki en düşük sonuçları italik olarak vurgularız. Şekil 24 ve Şekil 25, grafiklerde aynı sonuçları göstermektedir.

Çizelge 17 Çalışmadaki algoritmaların eğitim ve test verileri üzerindeki performansı

Yöntemler	**Eğitim verilerinin doğruluğu**	**Test verilerinin doğruluğu**
C5	88.9 (1)	85.3 (1.8)
KNN	*84.1 (0.2)*	*69.6 (1.5)*
SVM	87.3 (0.2)	78.3(1)
GA	97.2 (0.6)	90.5 (0.9)



Şekil 24 Algoritmaların eğitim seti üzerindeki performansı



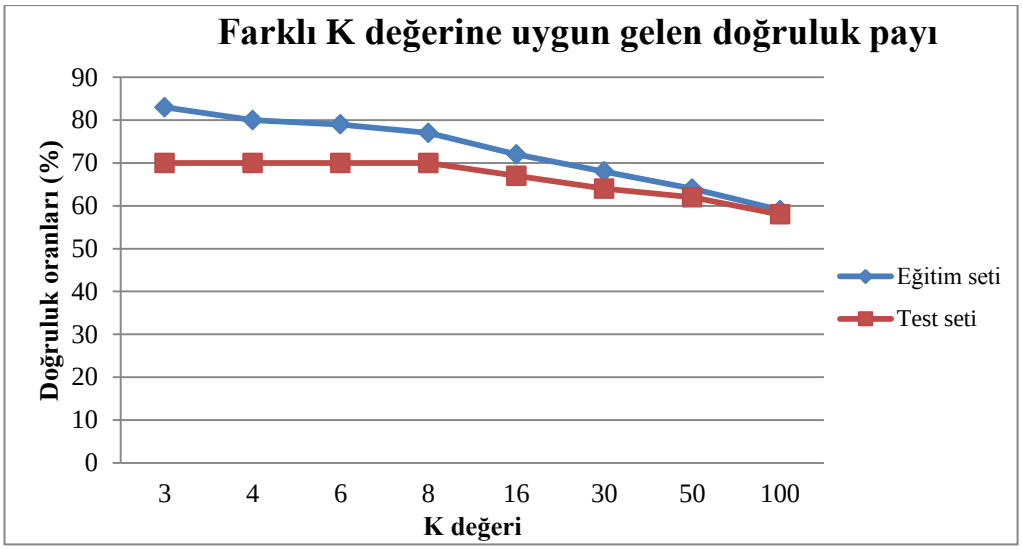
Şekil 25 ML tekniklerinin test setindeki performansı

ID3: C5'i (ID3'ün en son torunu, Rulequest'ten (www.rulequest.com) kaynak kodu olarak edinilebilir) winnowing ile 10 kat üzerinde çalıştırdık. Çizelge 17'deki sonuçlar, fazla veri uydurma olmadığını ve düşük bir standart sapma ile sonuçlandığından tekniğin kararlı olduğunu (farklı 10 kat üzerinde) göstermektedir. Ayrıca, GA'mızın ardından en iyi performansı C5'in gösterdiğini fark ettik.

Destek Vektör Makineleri: SVM performansı hiçbir parametreden etkilenmez. K-NN (Scikit-Learn) ile aynı kütüphaneyi kullandık. Standart sapmanın düşük değeri ile gösterildiği gibi, fazla veri uydurma yoktur ve teknik stabildir (farklı 10 kat üzerinde).

K-En Yakın Komşu: KNN'nin varsayılan benzerlik ölçümünü kullandık - Öklid Mesafesi. K, KNN kullanırken dikkate alınması gereken ana parametredir. K'yi arttırırken, yeni bir tane sınıflandırmak için kullanılan durum sayısını artırıyoruz. Sınıflandırma, daha küçük bir K kullanımına göre daha fazla sayıda benzer duruma dayanacağından, bu bir avantaj gibi görünebilir. Ancak bu avantaj, verilere ve benzer durumların birbirleri üzerindeki etkisinin ne kadar büyük olduğuna bağlıdır. Örneğin, artan K, gürültülü verilerle uğraşırken KNN'nin performansını bozabilir, çünkü K ne kadar büyükse, komşuluğa o kadar fazla gürültü dahildir.

Makine öğrenimi için Scikit-Learn 0.18.1 sürümünü (http://scikit-learn.org/) - bir Python kitaplığı - kullandık. KNN'yi her seferinde farklı bir K değeriyle birden çok kez çalıştırırız. KNN'nin K'ye göre davranışı Şekil 26'de gösterilmektedir. En iyi sonuçları verdiği için K'yi 3 olarak seçiyoruz.



Şekil 26 Kullanılan k ile ilgili olarak KNN doğruluğunun test ve eğitim verileri üzerindeki değişimi.

Şekil 26, K arttıkça KNN performansındaki düşüşü göstermektedir. Bunun nedeni verilerdeki gürültüdür. Standart sapmanın düşük değeri göz önüne alındığında, fazla veri uyumu yoktur ve teknik stabildir (farklı 10 kat üzerinde).

Genetik Algoritma: C5 kullanarak, eğitim kıvrımlarını kullanarak 50 farklı kural kümesi oluşturuyoruz. Örneklenen GA'yı, ilk popülasyon olarak oluşturulan 50 kural kümesini kullanarak aynı veriler üzerinde eğitiriz, ardından GA tarafından üretilen son kural kümesini test katında test ederiz. GA'daki rastgele eleman nedeniyle bu deneyi 30 kez tekrarladık.

A. Parametrelerin Etkisi

GA performansı 5 parametreden yüksek oranda etkilenir: K-turnuva seçiminde kullanılan K, çocuğu mutasyona uğratma olasılığı (μ1), en iyi ebeveyni mutasyona uğratma olasılığı (μ2), çaprazlama tabanlı kural seçme olasılığı güven (Pgüven), seçkinlik yüzdesi (Elit_yüzdesi) ve platonun maksimum uzunluğu (MAXPLAT) üzerine. GA'yı parametrelerin farklı değerleriyle çalıştırıyoruz. Çizelge 18'de, test verilerinde en iyi sonuç verenler rapor edilmiştir.

Çizelge 18 GA parametrelerinin en iyi değerleri

Parametre	**Değer**
K	3
μ1	0.05

Çizelge 18 GA parametrelerinin en iyi değerleri Devam

μ2	0.1
Pgüven	0.7
Elit_yüzdesi	20%
MAXPLAT	5
Populasyon ölçüsü	50

B. Sonuçların Tartışılması

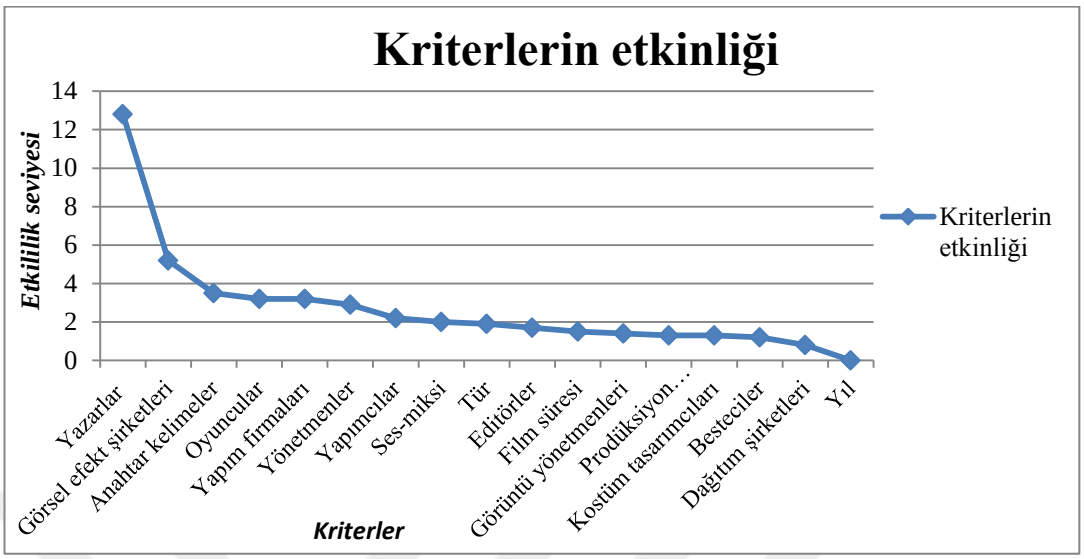
Çizelge 17, Şekil 24 ve Şekil 25'de gösterildiği gibi, GA, %90,5'lik bir test doğruluğu ve ardından C5 (%85,3) elde eden test verilerinde diğer tüm makine öğrenimi tekniklerinden daha iyi performans gösterir. Bu, karar kurallarının bu belirli verilerden tahminde iyi olduğunu gösterir. GA'nın aksine, K-NN en kötü sonuçları gösterir. K-NN tarafından kullanılan varsayılan mesafe ölçüsü olan durumlar arasındaki Öklid mesafesinin, problem için benzer örnekleri tanımlamak için uygun bir yaklaşım olmadığını çıkarmak mümkündür.

C. Öznitelik Etkisi

C5 tarafından sağlanan harmanlama tekniği, sınıf etiketi ile öznitelik korelasyonunu analiz etmek için bir yöntem olarak kullanılabilir. Bu teknikle C5, karar ağacından çıkarılan öznitelikleri, yani sınıflandırma sürecinde etkisiz olan öznitelikleri listeler. Ancak, bu tür niteliklerin GA'mıza dahil edilmesinin daha iyi sonuçlar verdiğini fark ettik. Nitekim, GA'mızı bu niteliklerle ve bunlar olmadan çalıştırdık. İkinci durumda, doğruluk bozuldu. Niteliklerin etkinliğini değerlendirmek için, GA tarafından üretilen son kural setini inceliyoruz ve kullanılmayan nitelikleri not ediyoruz. Ayrıca, doğruluğu bozmadan sınıflandırıcıdan hangi özniteliklerin çıkarılabileceğini araştırıyoruz. Bunun için ve her öznitelik için, ona karşılık gelen tüm koşulları GA optimal kural kümesinden kaldırırız ve eğer bu, performansı bozmazsa (eğitim doğruluğu açısından), niteliği sınıflandırıcıdan kaldırırız.

Ayrıca, bu özelliği kaldırırken doğruluktaki düşüşü hesaplayarak her bir özelliğin etkililik seviyesini ölçüyoruz. Örneğin, ilk doğruluğun %95 olduğunu ve A özniteliğini çıkardıktan sonra %92'lik bir doğruluk elde ettiğimizi varsayalım. Böylece A'nın etkililik seviyesi 3'tür. Dolayısıyla, 0 veya daha az etkililik seviyesine

sahip bir nitelik, performansı etkilemeden çalışmadan çıkarılabilir. Tüm niteliklerin etkinlik düzeyi Şekil 27'de gösterilmektedir.

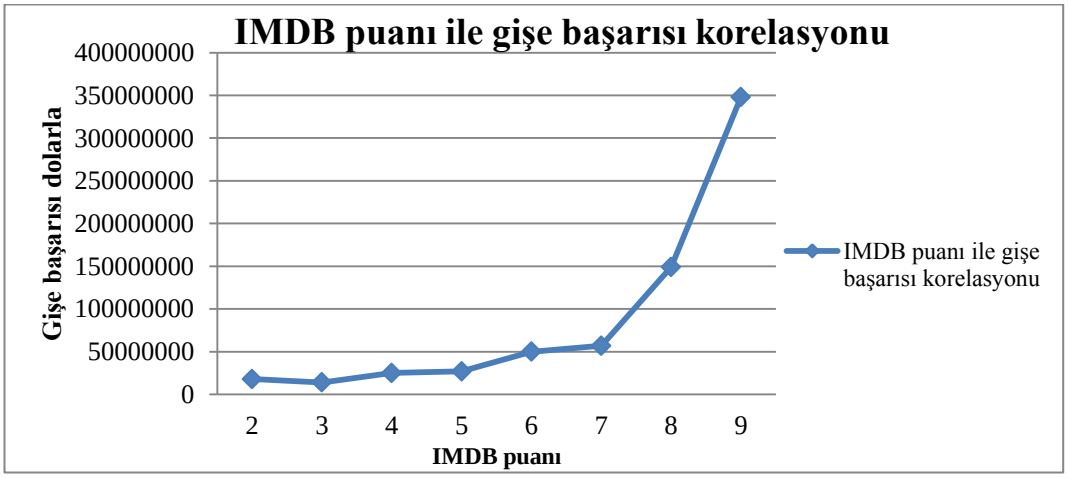


Şekil 27 Her özelliğin hesaplanan etkinlik düzeyi

Şekil 27'de görüldüğü gibi, etkinlik değeri 0 olduğu için modelimizin performansını etkilemeden "Yıl" özniteliğini güvenle kaldırabiliriz.

D. IMDB Derecelendirmesi ve Gişe Başarısı

Bölüm 1'de bahsedildiği gibi, yatırımcılar ve prodüksiyon şirketleri, filmin reytingi kadar kârıyla da ilgilenirler. Şekil 28'de, gişe büyümesine karşı IMDB notunu çiziyoruz. Grafik, bir filmin gişe artışının IMDB puanıyla yüksek oranda ilişkili olduğunu gösteriyor. Bu nedenle, ikincisi öncekinin bir göstergesi olarak kullanılabilir.



Şekil 28 IMDB puanı ile gişe büyümesi arasındaki korelasyon

V. SONUÇ VE GELECEK ÇALIŞMALAR

Film prodüksiyonu, büyük kâr veya zararla sonuçlanabilecek güvencesiz bir yatırım alanıdır. Filmlerin yapımlarından önceki başarısını nispeten yüksek bir hassasiyetle tahmin eden pratik bir model geliştirmek, yatırımcıları ve birçok şirketi zarar ve iflastan kurtarabilir. Bu tezde, bu problemi makine öğrenme tekniklerini kullanarak ele alıyoruz: Genetik Algoritmalar, C5, Destek Vektörleri Makinesi, K-En Yakın Komşular, tüm teknikleri IMDB'den toplanan filmlerde test ettik. Veri seti 4883 filmden oluşmaktadır. Tüm teknikler, bu problemde makine öğreniminin doğruluğunu gösteren umut verici sonuçlar verdi. C5 tarafından üretilen kural kümelerini optimize eden kendi genetik algoritmamızı geliştirdik. Uygulanan GA'mız, görünmeyen filmlerde yaklaşık %90.5'lik bir tahmin doğruluğu ile en yüksek sonuçları elde etti (1998'den 1999'a kadar olan bir doğrulama film setinde %88.16).

Yaklaşımımızın avantajı, yatırımcıların ve yapım şirketlerinin bir filmin başarı şansını artıran bazı film özelliklerine karar vermelerine yardımcı olan, insan tarafından okunabilir bir biçimde bir dizi kural sağlamasıdır. Yaklaşımımızın, sırasıyla “Final Fantasy: The Spirits Within (2001)” ve “Titan A.E (2000)” filmlerinde FOX stüdyosu ve Square Picture'da olduğu gibi yapım şirketlerinin büyük kayıplara uğramasını engelleyebileceğine inanıyoruz.

Deneylerimiz, bazı özelliklerin filmlerin başarısını daha fazla etkilediğini gösterdi, tam olarak, Yazarlar en çok Özel Efekt Şirketleri, Anahtar Kelimeler, Aktörler vb. film gişe büyümesi ile korelasyon. Gelecekteki çalışmalarda, başarı ölçüsü olarak gişe artışını kullanırken diğer meta-sezgisel yöntemlerin performansını test etmeyi planlıyoruz. Ayrıca, TV dizilerinde, belgesellerde ve şovlarda kullanmak için modelimizin alan uyarlamasını yapmak ilginç bir fikir olabilir.

VI. KAYNAKÇA

MAKALELER

AJAY S. S. R., PRATİK K, ABHİYASH J., (2012). "Box-Office Opening Prediction of Movies Based on Hype Analysis through Data Mining" **International Journal of Computer Applications**, cilt 56, sayı 1.

ANAND B., HİMANSHU K., VİNAY B., PRANALİ K., (2015). **International Conference on Pervasive Computing**, "Role of Different Factors in Predicting Movie Success".

APALA, K. R., JOSE, M., MOTNAM, S., CHAN, C.-C., LISZKA, K. J., GREGORIO, F., (2013). "Prediction of Movies Box Office Performance Using Social Media", **International Conference on Advances in Social Networks Analysis and Mining (ASONAM)**, ss.1209-1214.

BASUROY, S., CHATTERJEE, S., RAVİD, S. A., (2003). "How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets. Journal of Marketing", ss.103–117.

BORGA D., ROBERT B. H., (2012). "When to Greenlight: Examining the Pre-Release Factors That Determine Future Box Office Success of a Movie in the United States", **International Journal of Economics and Management Sciences**, cilt 2, sayı 3. ss.35-42.

BRİAN DE S., RYAN C., (2014). "Prediction of Foreign Box Office Revenues Based on Wikipedia Page Activity".

BURGOS, M. C., CAMPANARİO, M. L., LARA, J. A., LİZCANO, D., (2015). "Using Decision Trees to Characterize and Predict Movie Profitability on the US Market", **Proceedings of the International MultiConference of Engineers and Computer Scientists**.

CHAKRABORTY, P., RAHMAN, M. Z., RAHMAN, S., (2019). "Movie Success Prediction Using Historical and Current Data Mining", **International Journal of Computer Applications**, cilt 178, sayı 47.

DELEN, D., SHARDA, R., (2009). "Predicting the Financial Success of Hollywood Movies Using an Information Fusion Approach", **Endüstri Mühendisliği Dergisi**, cilt 21, sayı 1, ss.30-37.

DEMIR, D., KAPRALOVA, O., LAI, H., (2012). "Predicting IMDB Movie Ratings Using Google Trends".

DOOMS, S., PESSEMIER, T. D., MARTENS, L., (2014). "Improving IMDb Movie Recommendations with Interactive Settings and Filters.".

GALVÃO, M., HENRIQUES, R., (2018). "Forecasting Movie Box Office Profitability", **Journal of Information Systems Engineering & Management**, cilt 3, sayı 3.

GAURVI L., ANIRUDH P., (2015). "Predicting Movie Ratings at IMDB", **Department of Computer Science & Engineering and Information Technology**, Jaypee University of Information Technology.

IM, D., NGUYEN, M. T., (2011). "Predicting Box-Office Success of Movies in the U.S. Market".

JAISWAL, M., PRASAD, A., SRIVASTAVA, A., SIDDQUI, T. J., (2020). "Prediction and Analysis of Movie Success with Machine Learning Approach".

JEHOSHUA E., SAM K. H., Z. JOHN Z., (2014). "Assessing Box Office Performance Using Movie Scripts: A Kernel-Based Approach" **Transactions on Knowledge and Data Engıneering**, cilt 26, sayı 11, ss.2639-2648.

KHALID I. A., TANVIR A., MD. SAIEDUR R., (2012). "Movie Popularity Classification Based on Inherent Movie Attributes Using C4.5, PART and Correlation Coefficient", **International Conference on Informatics, Electronics & Vision**, ss.747-752.

KRISHMA, AMIT C., (2019), "Does Sentiment Effect the Gross Income of a Movie?", **Journal of Emerging Technologies and Innovative Research**, cilt 6, sayı 5, ss.693-701.

LASH, M. T., ZHAO, K., (2016). "Early Predictions of Movie Success: The Who, What, and When of Profitability.", **Journal of Management Information Systems**, cilt 33, sayı 3, ss.874–903.

LEE, S., KC, B., CHOEH, J. Y., (2020). "Comparing Performance of Ensemble Methods in Predicting Movie Box Office Revenue", Heliyon.

LOPAMUDRA P., LY (HARRIET) B., RISHI M., (2020). "Predicting Box Office Success: Do Critical Reviews Really Matter?".

MAHARSHI V., KEVAL S., NITIN U., (2017). "Is Movie a Box Office Success? Analyzing Search Queries to Predict Commercial Success", **International Journal of Control Theory and Applications**, cilt 9, sayı 41, ss.629-640.

MAHESH J., DIPANJAN D., KEVIN G. N. A. S., (2010). "Movie Reviews and Revenues: An Experiment in Text Regression" ss. 293–296.

MEENAKSHI, K., MARAGATHAM, G., AGARWAL, N., GHOSH, I., (2018). "A Data Mining Technique for Analyzing and Predicting the Success of Movie", **National Conference on Mathematical Techniques and Its Applications** (NCMTA 18).

MHOWWALA, Z., SULTHANA, A. R., SHETTY, S. D., (2020). "Movie Rating Prediction Using Ensemble Learning Algorithms", **International Journal of Advanced Computer Science and Applications** (IJACSA), ss.383-388.

N.A. P., E.V.D.M. S., (2013). "The Determinants of Box Office Performance in the Film Industry Revisited", **South African Journal of Business Management**, cilt 44, sayı 3, ss.47-58.

NİTHİN VR., LİJİYA, A., (2014). "Predicting Movie Success Based on IMDB Data", **International Journal of Data Mining Techniques and Applications**, ss. 365-368.

OGHİNA, A., BREUSS, M., TSAGKİAS, M., RİJKE, M., (2012). "Predicting IMDB Movie Ratings Using Social Media."

PARİMİ, R., CARAGEA, D., (2013). "Pre-release Box-Office Success Prediction for Motion Pictures", (MLDM), ss.571-585.

PİMWADEE C., LİNA Z., (2005). "Movie Review Mining: A Comparison between Supervised and Unsupervised Classification Approaches".

RACHİT T., CHERAG V., (2015). "User Propensity Analysis for Movie Prediction Rating Based on Collaborative Filtering and Fuzzy System", **International**

Journal of Innovative Science, Engineering & Technology, cilt 2, sayı 9, ss.471-479.

SANGKİL, M., PAUL, K. B., DAWN, I., (2010). "Dynamic Effects Among Movie Ratings, Movie Revenues, and Viewer Satisfaction".

SARANYA, A., HUSSAİN, A., (2015). "User Genre Movie Recommendation System Using NB Tree", **International Journal of Innovative Research in Science, Engineering and Technology**, cilt 4, sayı 7, ss.5854-5859.

SHARDA, R., DELEN, D., (2006). "Predicting Box-Office Success of Motion Pictures with Neural Networks", **Expert Systems with Applications**, ss.243–254.

SHRADDHA M., HİTARTHİ B., DARSHANA D., (2015). "A Compendium for Prediction of Success of a Movie Based Upon Different Factors", **International Journal of Advanced Research in Computer and Communication Engineering**, cilt 4, sayı 12, ss.297-300.

SİMONOFF, J. S., SPARROW, I. R., (2000). "Predicting Movie Grosses: Winners and Losers, Blockbusters and Sleepers", Chance, cilt 13, sayı 3.

SİTARAM A., BERNARDO A. H., (2010). "Predicting the Future With Social Media", **International Conference on Pervasive Computing.**

WANG, Z., ZHANG, J., Jİ, S., MENG, C., Lİ, T., ZHENG, Y., (2020). "Predicting and Ranking Box Office Revenue of Movies Based on Big Data. Information Fusion", ss.25–40.

WENBİN Z., STEVEN S., (2009). "Improving Movie Gross Prediction Through News Analysis".

WERNARD S., SANDER W., (2015). "Predicting Ratings for New Movie Releases From Twitter Content", ss.122-126.

YASSERİ, T., KERTÉSZ, J., (2013). "Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data".

TEZLER

ABDOULAZİZ, A., (2019). "Doğal Dil İşleme ve Veri Madenciliği Kullanarak Tvitler Üzerinden Film Derecelendirilmesi", (Yüksek lisans tezi), Bilgisayar Mühendisliği Anabilim Dalı, Konya Teknik Üniversitesi.

ADÉLA D., (2017). "Determinants of the Box Office Success in the European Film Market" (Yüksek lisans tezi), Department of Finance, Masaryk University

CORRÊA DE SÁ, R. C., (2020). "Movie's Box Office Performance Prediction", (Yüksek lisans tezi), NOVA Information Management School, Specialization in Knowledge Management and Business Intelligence, Lizbon.

MUSA M., (2013). "Filmöneri sistemleri için hibrit bir yöntem geliştirilmesi", (Yüksek lisans tezi), Bilgisayar Mühendisliği Anabilim Dalı, Ege üniversitesi.

TUNA H., (2013). "Internet Based Movie Genre Suggestion Model Considering Demographical Information of Users", (Yüksek lisans tezi), The Department of Informatıon Systems, The Middle East Technical University.

ÖZGEÇMİŞ

Ad-Soyad: Ogtay SAFARALİYEV

ÖĞRENİM DURUMU

- **LİSANS:** (2019), Azerbaycan Devlet İktisat Üniversitesi İT ve Sistemleri Mühendisliği Bölüm Mezunu
- **YÜKSEK LISANS:** İstanbul Aydın Üniversitesi Bilgisayar Mühendisliği Tezli Yüksek Lisans