Filmlerin Derecelerinin Veri Madenciliği Teknikleri ile Tahmin Etme

Fatma Betül Yazıcı*
*Department of Computer Engineering
İstanbul Technical University, Turkey, Email: betulyazicii@gmail.com

***Abstract*—Filmlerin puanları, filmi izleyenlerin ne kadar beğendiği bilgisini verir. Yeni bir film vizyona girdiği zaman filmin puanı bilinmemektedir. İzleyiciler açısından filmlerin puanlarını önceden bilmek filmleri seyredip seyretmeme kararını etkiler. Sinema şirketleri açısından filmlerin puan bilgisinin önceden tahmin edilmesi filmler için kaç sinema salonunda, kaç seans, hangi gün ve saatlerde filmlerin gösterilmesi kararını etkiler. Bu çalışma ile IMDB (Internet Movie Database) verileri kullanılarak sınıflandırma teknikleri ile filmlerin gelecekteki puanlarının tahmin edilmeye çalışılmıştır. Farklı sınıflandırma modelleri ve sınıflandırma değerlendirme ölçütleri ile sonuçlar karşılaştırmalı bir şekilde gösterilmektedir.**

***Index Terms*—Movie Rating Prediction, Machine Learning, Classification, Feature Selection, Outlier Detection**

I. Giriş

Film, insanların hayatında gittikçe daha önemli bir eğlence haline geliyor. İnsanların film seçimleri ve neyi sevebileceğinin tahmin edilmesi ilginç bir konudur. Bu çalışma ile bir film derecelendirme tahmin sistemi kuruyoruz. Filmlerin kullanıcılar tarafından puanlanmış veri kümesini kullanarak puanlanmamış bir film geldiğinde derecesinin ne olabileceğini tahmin etmeye çalışıyoruz. Bu tahmin sonucu ileride film önerisi olarak da sunulabilir. Yüksek puan olarak tahmin edilen filmleri kullanıcılara önerilebilir.

Filmlerin başarısının tahmin edilmesi üzerine birçok çalışma mevcuttur. Öneri sistemleri için filmlerin puanlarının tahmin edilmesi üzerine yapılan çalışmalar filmlerin süresi, türü gibi filmlerin niteliklerine bakmadan, kullanıcıların benzerliklerine bakarak filmlerin puanlarını tahmin etmektedir. Bir kullanıcının bir filme verdiği puanın, benzer kullanıcılar tarafından da aynı verileceği fikrine dayanmaktadır. Bu çalışmada filmlerin nitekiklerinden yola çıkarak puanlarının tahmin edilmesi üzerine bir çalışmadır.

II. Literatür İncelemesi

Filmlerin derecelerinin belirlenmesi konusunda yapılmış birçok çalışma bulunmaktadır. Birçok araştırmacı makine öğrenmesi teknikleri kullanarak filmlerin derecelerini tahmin etmek için çalışmalar yapmıştır. Gaenssle ve Budzinski yaptıkları bir çalışmada 2002'den 2014 yılına kadar Rusya'da oynayan uluslararası filmlerin başarısının ölçülmesi üzerine bir çalışma yapmıştır. Çalışmanın sonucuna göre bütçenin filmin başarısının tahminlenmesinde yüksek bir etkiye sahip olduğu görülmüştür. [1]

Warda ve Zakia isimli araştırmacıların yaptığı çalışmada Wikipedia ve IMDB'den filmlerin bilgilerini çekerek Bagging, Random Forest, J48, IBK ve Naive Bayes sınıflandırma yöntemlerini kullanmışlardır. Dengesiz veri setlerini önce resampling sonra SMOTE algoritmalarını kullanarak %57.9 oranındaki sınıflandırma başarısını %99.23 oranına çıkarmışlardır. [2]

III. Metodoloji

Önerilen metodoloji 4 ana başlık altında incelenmektedir. Veri hazırlama, veri önişleme, özniteliklerin seçilmesi ve sınıflandırma algoritmalarının uygulanmasıdır.

A. Veri Hazırlama

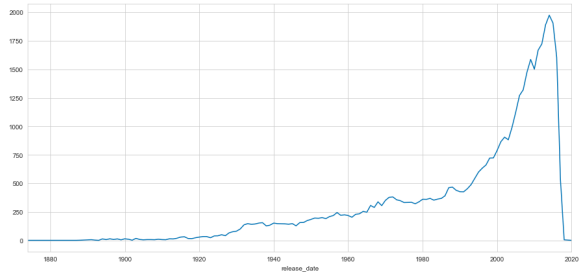
Bu çalışma için IMDB veri kümesi kullanılmıştır. Veri kümesi 45466 kayıt ve 12 nitelikten oluşmaktadır. Sınıf etiketi bulunmamaktadır. Niteliklere ait açıklamalar Tablo 1'de gösterilmektedir. [3]

TABLO I
Veri Kümesi Nitelik Bilgileri

Niteliğin Adı	***Niteliğin Açıklaması***
Title	Filmlerin orjinal dillerindeki ismidir
Release Date	Filmlerin yayım yılının bilgidir
Adult	Yetişkinlere uygun olup olmamasını belirten bir flagdir
Genre	Filmin türüne ait bilgileri içermektedir
Budget	Filmin bütçesinin ne olduğu bilgisini göstermektedir
Language	Filmin hangi dilde olduğu bilgisini içermektedir
Status	Filmin durumu hakkında bilgi verir
Revenue	Filmden elde edilen gelir bilgisidir
Runtime	Filmin süresinin bilgisidir
Vote Count	Filmdeki oyların kaç kişi tarafından verildiği bilgisini tutmaktadır
Popularity	Filmlerin popülerlik puan bilgisi
Vote Average	Filmlerin oylarının ortalaması

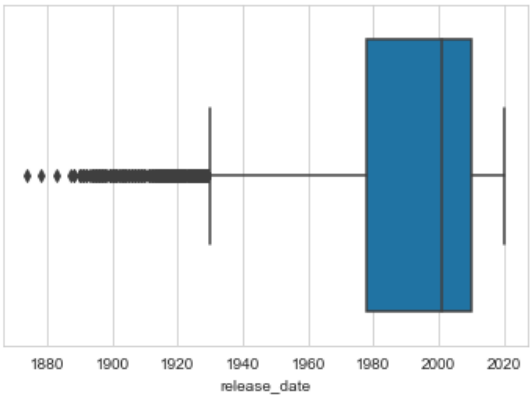
Veri kümesindeki nitelikler için önişlemede yapılan işlemlerin detayları aşağıdaki gibidir.

1) Filmlerin Yayım Yılı: Veri kümesinde 1874 yılından 2020 yılına kadar yayınlanmış filmleri içermektedir. Şekil 1'de filmlerin yıllara göre dağılımını göstermektedir.



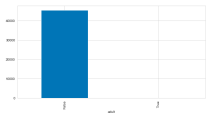
ŞEKİL 1. Filmlerin yıllara göre dağılımı

Yıl niteliğindeki aykırı değerleri belirlemek için kutu grafiği yöntemi kullanılmıştır. Şekil 2'de yıllara ait kutu grafiği yöntemi gösterilmektedir. Kutu grafiği yöntemi, veri çeyreklerini ve ortalamaları görüntüleyerek sayısal verilerin ve değişkenliğin görsel olarak dağılımını göstermek için kullanılmaktadır. Bu yöntem ile veri aykırı değerlerden temizlenmiştir.



ŞEKİL 2. Yıllara ait kutu grafiği

2) Filmlerin Yaş Sınırı Durumu: Filmlerin yetişkinlere uygun olup olmamasını belirten bir flagdir. False ve True şeklinde iki değer almaktadır. False değeri yetişkinlere uygun oldugunu, true değeri uygun olmadığını gösterir. 45465 kayıt false değerini, 1 kayıt ise true değeri içermektedir. Dengesiz veri kümesi olduğundan bu nitelik veri sınıflandırmada kullanılmamıştır. Şekil 3'de bu niteliğe ait farklı değerlerin veri kümesindeki sayılarını göstermektedir.



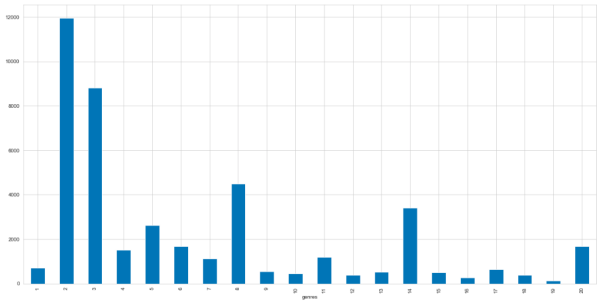
ŞEKİL 3. Yaş Sınırı Niteliğinin Durumu

3) Film Türü: Filmin türüne ait bilgileri içermektedir. Veri setinde Json formatında tutulmaktadır. Bu nitelik değeri için tüm farklı film türleri bulunarak kategorik değer haline getirilerek bir önişlemeden geçmiştir. Filmler için 20 farklı kategori bulunmaktadır. Tablo 2'de filmlerin türlerine karşılık gelen kategorik değerler gösterilmektedir.

TABLO II
FILM TÜRLERINE KARŞILIK GELEN KATEGORIK DEĞERLER

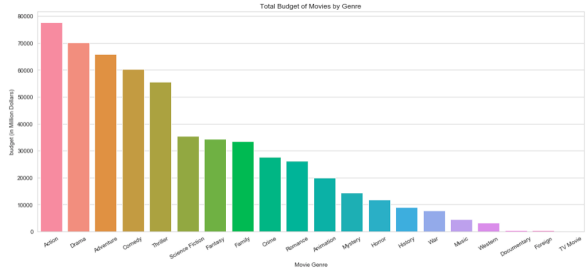
Kategorik Değer	***Film Türü***
1	Fantasy
2	Drama
3	Comedy
4	Adventure
5	Horror
6	Crime
7	Animation
8	Action
9	Mystery
10	Western
11	Romance
12	War
13	Family
14	Documentary
15	Music
16	History
17	Science Fiction
18	TV Movie
19	Foreign
20	Thriller

Veri kümesinde film türlerinin sayılarına ait bilgiler Şekil 4'de gösterilmektedir.



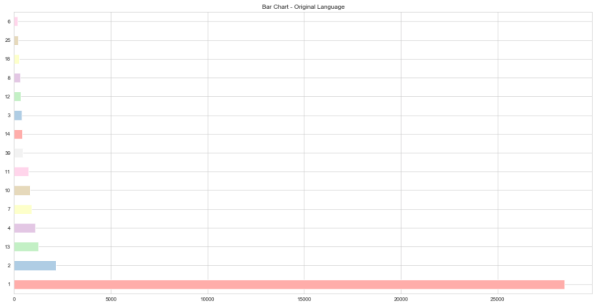
ŞEKİL 4. Film türünün veri kümesi içindeki dağılımı

4) Film Bütçesi: Filmin bütçesinin ne olduğu bilgisini göstermektedir. Bütçede 0 olan 34163 kayıt bulunmaktadır. Bu kayıp değerler veri kümesinde ortalama değer ile doldurularak bir önişlemeden geçirilmiştir. Filmlerin türlerine göre toplam bütçe fiyatları Şekil 5'de gösterilmektedir.



ŞEKİL 5. Film türlerine göre toplam bütçe tutarları

5) Filmin Dili: Filmin hangi dilde olduğu bilgisini içermektedir. Veri kümesinde json formatında tutulmaktadır. 85 farklı dilde film veri kümesinde bulunmaktadır. Veri önişleme adımı ile bu değerler kategorik hale getirilmiştir. Filmdeki dillere ait film sayıları Şekil 6'da gösterilmektedir.



ŞEKİL 6. Filmlerde Kullanılan Dillere Ait Film Sayıları

Filmlerde en çok kullanılan 5 dil türünün ve bu dillere ait veri kümesindeki film sayıları aşağıdaki Tablo 3'de gösterilmektedir.

TABLO III
VERI KÜMESINDE EN ÇOK BULUNAN DILLER

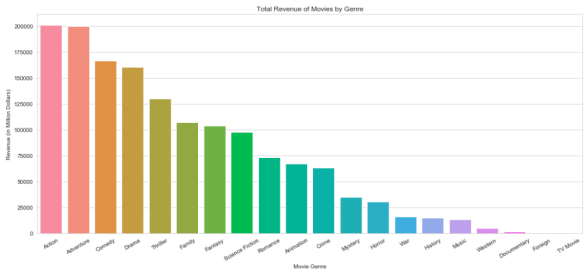
Filmin Dili	***Toplam Film Sayısı***
İngilizce	28852
Fransızca	2287
Japonca	1245
İtalyanca	1240
Almanca	971

6) Filmin Durumu: Filmin durumu hakkında bilgi verir. Json formatında tutulmaktadır. Veri önişleme adımı ile kategorik verilere dönüştürülmüştür. Tablo 4'te tüm durum kategorilerine ait film sayıları gösterilmiştir.

TABLO IV
FILMIN DURUM TÜRLERINE KARŞILIK GELEN FILM SAYILARI

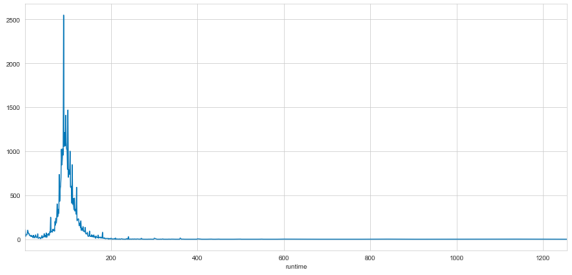
Filmin Durumu	***Toplam Film Sayısı***
Released	45228
Rumored	210
Post Production	85
In Production	11
Planned	11
Canceled	1

7) Film Gelir Bilgisi: Filmden elde edilen gelir bilgisidir. Gelir için min değer 1 dolar, max değer 2.787.965.087,0 dolar olarak görülmektedir. Kutu grafiği yöntemi ile veri aykırı değerlerden temizlenmiştir. Şekil 7'de film türlerine ait toplam gelir bilgisi gösterilmektedir.



ŞEKİL 7. Film türlerine göre toplam gelir tutarları

8) Film Süresi: Filmin süresinin gösterildiği nümerik bir değerdir. 260 film için süre bilgisi yer almamaktadır. 1558 film için de süreler 0 olarak görünmektedir. Süre bilgileri olmayanlar veri setinden temizlenmiştir. Filmlere ait minimum süre 1 dakika ile 'La pyramide de Triboulet' filmi, maximum süre 1256 dakika ile 'Centennial' filmidir. Filmin süresine ait dağılım Şekil 8'de gösterilmektedir.



ŞEKİL 8. Film Süresinin Veri İçindeki Dağılımı

9) Filmleri Oylayan Kişi Sayıları: Filmleri oylayan kişi sayılarının tutulduğu nümerik bir değerdir. Veri kümesinde 2438 film kimse tarafından oylanmamıştır. Bu veriler veri kümesinden silinmiştir. Filmler en az 1 kişi, en fazla 14075 kişi tarafından oylanmıştır.

10) Filmlerin Popülerlik Bilgisi: Filmlerin tıklanma, yorum alma durumuna göre IMDB tarafından hesaplanmış nümerik bir değerdir. Veri kümesinde minimum 0, maximum 547 değerini almaktadır. Veri içerisinde eksik olan popülerlik

bilgileri veri kümesinin ortalama değeri ile doldurulmuştur. Popülerlik niteliği için min-max normalizasyon yöntemi ile normalize edilmiştir.

11) Filmlere Ait Oy Ortalama Değerleri: Filmin imdb ortalama puanı bilgisidir. Bu bilgi nümerik bir değerdir.Veri kümesinde 0 ile 10 arasında değerler almaktadır. Bu çalışmada filmlere ait ortalama değeri sınıf etiketi olarak kullanılmaktadır. Veri kümesi farklı sınıf sayılarına bölünerek sınıflandırma başarıları kontrol edilmiştir. Verinin etiketlenmesi Tablo 4'de yer alan kurallara göre oluşturulmuştur.

TABLO V
FARKLI SINIF ETIKETLERINE BÖLÜNME KURALLARI

Sınıf Bilgisi	1	2	3	4
4 sınıfa ayırma kriteri	0-3	3-5	5-7	7-10
3 sınıfa ayırma kriteri	0-5	5-8	8-10	-
2 sınıfa ayırma kriteri	0-5	5-10	-	-

B. Veri Önişleme

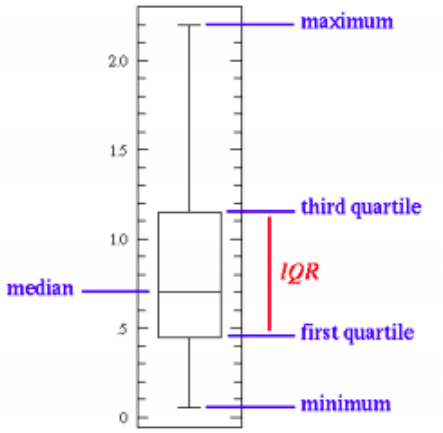
Veri önişleme adımında kategorik nitelikler için One Hot Encoding yöntemi, nümerik değerler için min-max normalleştirme yapılmıştır. Tüm nitelikler için aykırı değerler tespit edilmiş ve veri kümesinden silinmiştir.

1) One Hot Encoding Yöntemi: One Hot Encoding yöntemi, kategorik değişkenlerin ikili olarak temsil edilmesi anlamına gelmektedir. Bu yöntemde ilk önce kategorik değerlerin tamsayı değerleriyle eşlenir. Daha sonra, her bir tamsayı değeri, 1 ile işaretlenmiş tamsayı indeksi dışındaki tüm değerleri sıfır olan bir ikili vektör olarak gösterilmektedir. Makine öğrenmesi algoritmalarında verimli bir kullanım sağladığından bu yöntem tercih edilmiştir.

2) Min-Max Normalleştirme: Min-Max normalleştirme orijinal veri üzerinde doğrusal bir dönüşüm yapılmasını sağlamaktadır. Bu yöntem aracılığıyla veriler genellikle [0-1] aralığına dönüştürülür. Min.-Max normalleştirme işlemi alan değerinin minimum değerden ne kadar büyük olduğuna bakar ve bu farkları sıralar. Denklem 1'de min-max normalleştirme formülü gösterilmektedir. [4]

$$X' = \frac{X - min(X)}{max(X) - min(X)} \quad (1)$$

3) Aykırı Değer Analizi: Aykırı değer analizi için veri kümesindeki nümerik değerlerin minimum, 1.çeyrek, medyan, 3.çeyrek ve maksimum değerlerinin bulunarak veri kümesi 4 parça haline bölünür. Veri kümesinin 4 parça haline bölünmüş hali Şekil 9'da gösterilmektedir.



ŞEKİL 9. Veri Kümesi Boxplot Gösterimi

N elemanlı bir veri kümesinde 1.çeyrek değeri n/4 ya da (n+1)/4 sıradaki değerdir. 3.çeyrek değeri n/4 ya da (n+1)/4 sıradaki değerdir. 3.çeyrek değerinden 1.çeyrek değeri çıkarılarak çeyrekler aralığı hesaplanır. Çeyrek aralığı, verilerin medyan hakkında nasıl yayıldığını gösterir.

$$altsınır = Q1–(1.5 * IQR) \quad (2)$$

$$üstsınır = Q3 + (1.5 * IQR) \quad (3)$$

Alt sınır ve üst sınır bilgileri denklem 2 ve 3'de gösterilmiştir. Bu sınırların dışında kalan alanlar aykırı değer olarak belirlenir. Bu değerler veri kümesinden silinmiştir.

C. Öznitelik Seçimi

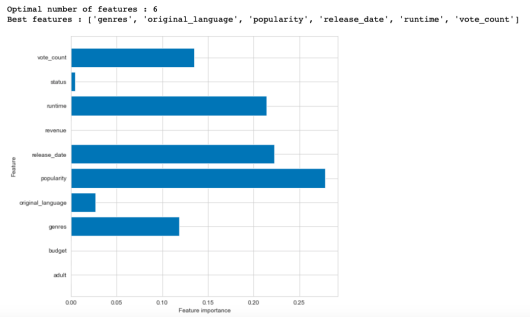
Öznitelik seçimi, veri kümesi içerisinde en yararlı öznitelikleri seçme ve bulma sürecidir. Bu işlem makine öğrenmesi modelinin performansını çok fazla etkilemektedir. 3 farklı sınıflandırma metoduna göre öznitelik önemlerinin sıralaması Tablo 6'da gösterilmektedir.

TABLO VI
ÖZNITELIK ÖNEM DEĞER SIRALAMASI

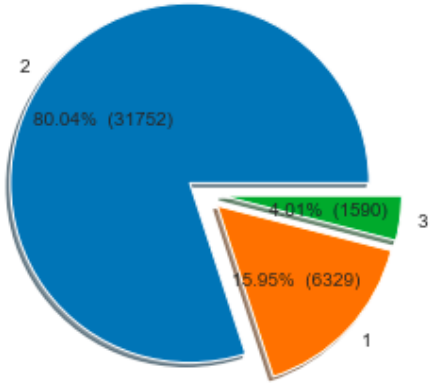
Nitelik Adı	Logistic Regression	SVM	*Naive Bayes*
Popularity	0.288643	0.317745	0.110405
Release Date	0.222923	0.218092	0.194681
Runtime	0.211124	0.216027	0.119087
Vote Count	0.130345	0.115996	0.209809
Genre	0.113410	0.099331	0.301519
Language	0.028928	0.028691	0.060070
Status	0.004302	0.004119	0.003753
Adult	0.000203	0.000000	0.000676
Budget	0.000000	0.000000	0.000000
Revenue	0.000000	0.000000	0.000000

En optimal özniteliklerin hangileri olduğunu bulmak için sınıflandırma yöntemleri kullanılarak sonuçları karşılaştırılmıştır. Karar ağacı sınıflandırma yönteminin

sonucuna göre en uygun nitelikler Şekil 10'da gösterilmektedir.



ŞEKİL 10. Karar ağacı yöntemine göre en uygun nitelikler



ŞEKİL 11. Verinin Dağılımı

D. *Sınıflandırma*

Niteliklerden yetişkinlere uygun olup olmama durumunu ifade eden nitelik, filmlerin gelir ve bütçe bilgilerinin önişleme adımında sınıflandırma başarısında bir etkisi olmadığından bu nitelikler kullanılmamıştır. Sınıflandırmada veri kümesi %70 eğitim %30 test veri kümesi olarak ayrılarak 3 farklı sınıf etiketi kuralı için 6 farklı sınıflandırma yöntemine göre başarıları ölçülmüştür.

Farklı sınıflandırma yöntemlerine göre elde edilen kesinlik ve anma metriklerinin sonuçları Tablo 7'de gösterilmektedir.

TABLO VII
SINIFLANDIRMA SONUÇLARI

Sınıflandırma Yöntemi	4 sınıf etiketi		3 sınıf etiketi		2 sınıf etiketi	
Logistic Reg.	Kesinlik	0.51	Kesinlik	0.64	Kesinlik	0.71
	Anma	0.62	Anma	0.8	Anma	0.84
SVM	Kesinlik	0.53	Kesinlik	0.67	Kesinlik	0.71
	Anma	0.62	Anma	0.8	Anma	0.84
Naive Bayes	Kesinlik	0.38	Kesinlik	0.64	Kesinlik	0.71
	Anma	0.62	Anma	0.8	Anma	0.84
Random Forest	Kesinlik	0.6	Kesinlik	0.76	Kesinlik	0.82
	Anma	0.64	Anma	0.8	Anma	0.84
KNN	Kesinlik	0.53	Kesinlik	0.72	Kesinlik	0.78
	Anma	0.53	Anma	0.75	Anma	0.81
Gradient B.	Kesinlik	0.53	Kesinlik	0.77	Kesinlik	0.82
	Anma	0.53	Anma	0.81	Anma	0.85

Sınıflandırma sonuçlarına göre 4 sınıf etiketi kullanıldığında en iyi başarıyı Random Forest sınıflandırma yöntemi ile %64 , 3 sınıf etiketi kullanıldığında en iyi başarıyı Gradient Boosting sınıflandırma yöntemi ile %81, 2 sınıf etiketi kullanıldığında en iyi başarıyı Gradient Boosting sınıflandırma yöntemi ile %85 olarak görülmektedir.

3 sınıf etiketi için verinin dağılımı Şekil 11'de gösterilmektedir.

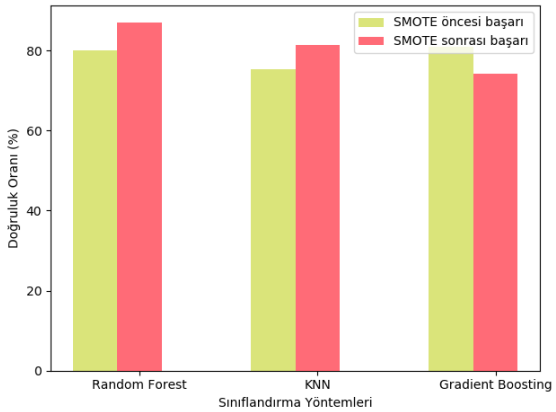
Sınıflandırma başarısını arttırmak için veri kümesinin dengeli hale getirilerek ve hiper-parametre kestirimi ile farklı parametrelerle sınıflandırma başarıları ölçülmüştür.

1) Veri Kümesinin Dengeli Hale Getirilmesi: Şekil 11 'e bakıldığında sınıfların eşit dağılmadığı verinin %80 gibi büyük bir oranı 2 numaralı sınıf etiketinin oluşturduğu görülmektedir. Dengesiz sınıf dağılımların bulunduğu veri kümeleri üzerinde sınıflandırıcılar istenilen düzeyde performans gösteremeyebilir. Veriyi dengeli hale getirmek için sentetik gözlemler üreten sentetik azınlık aşırı örnekleme (Synthetic Minority Oversampling Technique-SMOTE) algoritması kullanılmıştır. [5] Veri dengeli hale getirildikten sonra Random Forest, KNN ve Gradient Boosting sınıflandırma yöntemlerinde her sınıf etiketi için F-ölçütü sonuçları Tablo 8 'de gösterilmektedir.

TABLO VIII
SMOTE SONRASI SINIFLANDIRMA SONUÇLARI

Sınıflandırma Yöntemi	0. Sınıf		1.Sınıf		2.Sınıf	
Random Forest	SMOTE öncesi	0.33	SMOTE öncesi	0.89	SMOTE öncesi	0.11
	SMOTE sonrası	0.84	SMOTE sonrası	0.84	SMOTE sonrası	0.92
KNN	SMOTE öncesi	0.28	SMOTE öncesi	0.86	SMOTE öncesi	0.08
	SMOTE sonrası	0.82	SMOTE sonrası	0.71	SMOTE sonrası	0.89
Gradient Boosting	SMOTE öncesi	0.21	SMOTE öncesi	0.89	SMOTE öncesi	0.09
	SMOTE sonrası	0.72	SMOTE sonrası	0.73	SMOTE sonrası	0.78

SMOTE öncesi ve sonrası Random Forest, KNN ve Gradient Boosting sınıflandırma yöntemlerine ait başarı yüzdeleri Şekil 12'de gösterilmektedir.



ŞEKİL 12. SMOTE öncesi ve sonrası sınıflandırma yöntemlerinin başarı yüzdeleri

2) Hiper-parametre Kestirimi: Sınıflandırma yöntemlerinin parametrelerinin optimal değerlerini bulmak için kullanılan bir yöntemdir. Yöntemin uygulanması için sklearn kütüphanesinden GridSearch kullanılmıştır. Random Forest ve KNN için farklı parametrelere ait başarı değerleri Tablo 9'daki gibidir.

TABLO IX
FARKLI PARAMETRE SEÇIMLERINE GÖRE SINIFLANDIRMA SONUÇLARI

Sınıflandırma Yöntemi	Parametre Değeri	Başarı Yüzdesi	Parametre Değeri	Başarı Yüzdesi
Random Forest	max_features 3	0.8746	n_estimators 100	0.8767
	max_features 4	0.8724	n_estimators 200	0.8744
	max_features 5	0.8711	n_estimators 300	0.8762
	max_features 6	0.8644	n_estimators 400	0.8760
KNN	n_neighbors 3	0.8148	p=1 (manhattan dist.)	0.8513
	n_neighbors 5	0.7942	p=2 (euclidean dist.)	0.8336

IV. SONUÇ

Filmlerin derecelerinin tahmin edilmesi üzerine bir çalışma yapılmıştır. Çalışma IMDB veri kümesi kullanılarak gerçeklenmiştir. Logistic Regression, SVM, Naive Bayes, Random Forest, KNN ve Gradient Boosting sınıflandırma yöntemleri kullanılarak farklı sınıf sayıları için başarıları kesinlik ve anma metrikleri ile ölçülmüştür. Veri kümesi üzerinde verinin dengelenmesi için SMOTE yöntemi ve hiper-parametre yöntemi ile de farklı parametrelere göre sınıflandırma başarıları gösterilmiştir. Çalışmaya göre en başarılı sonucu sınıflandırmayı %87.67 ile random forest yöntemi ile ölçülmüştür. Çalışma python dilinde yazılmışır. Sklearn ve imblearn kütüphaneleri kullanılmıştır.

REFERENCES

[1] S. Gaenssle, O. Budzinski, and D. Astakhova, "Conquering the box office: Factors influencing success of international movies in russia," 2018.

[2] W. R. Bristi, Z. Zaman and N. Sultana, "Predicting IMDb Rating of Movies by Machine Learning Techniques," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019, pp. 1-5, doi: 10.1109/ICCCNT45670.2019.8944604.

[3] Banik,R.The Movies Dataset,https://www.kaggle.com/rounakbanik/the-movies-dataset (2017).

[4] Larose, D.T. Discovering Knowledge In Data: An Introduction to Data Mining. New Jersey: John Wiley and Sons Inc. (2005).

[5] E. Kartal And Z. Özen, "Dengesiz Veri Setlerinde Sınıflandırma," In Mühendislikte Yapay Zeka ve Uygulamaları , Sakarya: Sakarya Üniversitesi Kütüphanesi Yayınevi, 2017, pp.109-131.