DENİZ KESER
comedy_series_imdb
keserdeniiz@gmail.com
07.10.2023

IMDB TV Dizileri Rating Tahmini Proje Raporu

1. Giriş

1.1 Proje Hedefi ve Amacı

Proje Hedefi: Bu proje, bir film veri seti üzerinde veri analizi ve rating tahmini yapmayı amaçlamaktadır.

Proje Amacı:

1. Veri seti üzerinde eksik verilerin tespit edilmesi ve doldurulması.

2. Filmlerin IMDb puanları ile çeşitli özellikler arasındaki ilişkinin incelenmesi.

3. Polinom Regresyon modeli kullanılarak rating tahmini yapılması.

4. Sentetik verilerin oluşturulması ve bu verilerin modellere verilerek tahminlerin alınması.

5. Elde edilen sonuçların analizi ve projenin sonuçlarının raporlanması.

1.2 Kullanılan Veri Seti

Proje için ilk adım olan veri seti seçimi Kaggle platformu üzerinden gerçekleştirilmiştir. Seçilen veri seti "IMDB TV Series Data" veri setidir. Bu veri seti başlık(title), IMDb ID, yayın yılı (release year), tür(genre), oyuncu kadrosu(cast), özet(synopsis), derecelendirme(rating), ekran süresi(runtime), sertifika(certificate), oy sayısı (number of votes) ve brüt gelir (gross revenue) bilgileri dahil olmak üzere IMDB'deki tv dizileri hakkında bilgi vermektedir. Veriler, IMDB'nin web sitesinden alınmış olunup her tür için ayrı csv dosyaları halinde düzenlenmiştir. Bu projede yalnızca "comedy_series.csv" dosyasında bulunan bilgiler kullanılmıştır.

Veri Seti Özellikleri (Feautures):

- Başlık(title): Dizinin adı.
- IMDb ID: Dizinin IMDb'deki benzersiz tanımlayıcısı.
- Yayın Yılı (Release Year): Dizinin yayınlandığı yıl.
- Tür (Genre): Serinin tür(ler)i.
- Oyuncular (Cast): Dizinin ana oyuncu kadrosu.
- Özet (Synopsis): Serinin kısa bir özeti veya açıklaması.
- Derecelendirme (Rating): Dizinin IMDb'deki ortalama puanı (1'den 10'a kadar ölçeklendirilmiş).
- Çalışma Süresi (Runtime): Her bölümün süresi veya serinin toplam çalışma süresi.
- Sertifika (Certificate): Diziye atanan içerik derecelendirmesi veya sertifika (örn. PG-13, TV-MA).
- Oy Sayısı (Number of Votes): Dizinin aldığı toplam oy veya derecelendirme sayısı.
- Brüt Gelir (Gross Revenue): Serinin (varsa) ürettiği toplam brüt gelir.

2. Veri Seti İncelemesi

Veri setinin yapısı şekil-1'de olduğu gibidir.

	Title	IMDb ID	Release Year	Genre	Cast	Synopsis	Rating	Runtime	Certificate	Number of Votes	Gross Revenue
0	Succession	tt7660850	2018–2023	Comedy, Drama	Stars:, Sarah Snook, , Nicholas Braun, , Brian...	The Roy family is known for controlling the bi...	8.9	60 min	TV-MA	203682	NaN
1	Ted Lasso	tt10986410	2020–2023	Comedy, Drama, Sport	Stars:, Jason Sudeikis, , Brett Goldstein, , B...	American college football coach Ted Lasso head...	8.8	30 min	TV-MA	274806	NaN
2	Barry	tt5348176	2018–2023	Action, Comedy, Crime	Stars:, Bill Hader, , Stephen Root, , Sarah Go...	A hit man from the Midwest moves to Los Angele...	8.4	30 min	TV-MA	101883	NaN
3	Guardians of the Galaxy Vol. 3	tt6791350	2023	Action, Adventure, Comedy	Director:, James Gunn, , Stars:, Chris P...	Still reeling from the loss of Gamora, Peter Q...	8.2	150 min	PG-13	160447	NaN
4	Dungeons & Dragons: Honor Among Thieves	tt2906216	2023	Action, Adventure, Comedy	Directors:, John Francis Daley, , Jonathan Gol...	A charming thief and a band of unlikely advent...	7.3	134 min	PG-13	123247	NaN

Şekil 1: Veri Seti Yapısı

Veri seti 11777 satır ve 11 sütundan oluşmaktadır. Veri setine ait sütun adları 'Title', 'IMDb ID', 'Release Year', 'Genre', 'Cast', 'Synopsis', 'Rating', 'Runtime', 'Certificate', 'Number of Votes', 'Gross Revenue' şeklindedir. Veri setine ait bu bilgiler elde edildikten sonra isna().sum() ile eksik veri kontrolü yapılmıştır. Bu kontrol sonucunda Cast, Synopsis, Runtime, Certificate ve Gross Revenue sütunlarında eksik veriler olduğu tespit edilmiştir. Cast sütunundaki eksik veri sayısı (2) az olduğundan silinebileceğine, silinen bu verilerin veri setine zarar vermeyeceğine karar verilmiştir. Runtime sütunun yalızca sayısal veriler değil string veriler içerdiği (örnek: 30 min) tespit edilmiştir. Bu sebeple öncelikle Runtime sütunundaki virgüller kaldırılmış ardından değerler float değerlere dönüştürülmüştür ve Runtime_minutes isimli yeni bir sütuna aktarılmıştır. Eksik değerler de ortalama değerlerle doldurulmuştur. Certificate ve Gross Revenue sütunundaki eksik veri sayısı çok fazla olduğundan sonuçları etkilememesi için 'unspecified' olarak doldurulmuştur.

3. Veri Keşfi ve Görselleştirmesi

Bu adımda rating dağılımı, yıl bazında çıkış grafiği, tür dağılım grafiği, rating ve oy sayısı ilişki grafiği, sertifika dağılım grafiği, süre dağılımı ilişkileri görselleştirilmiş ve incelenmiştir. Rating Dağılımı grafiğinde 6-8 rating aralığında daha fazla dizi olduğu görülmektedir. Yıl bazında çıkış grafiği incelendiğinde son yıllara doğru bir artış olduğu gözlenmektedir. En çok görülen 10 tür grafiği ise; komedi türünde bir inceleme yapılmasına rağmen veri seti içindeki dizilerin yalnızca komedi türünde değil farklı türler içerdiği de tespit edilmiş olduğundan dağılım grafiğine dökülme ihtiyacı doğmuştur. Rating ve oy sayısı arasında bir ilişki olup olmadığı da Rating ve Oy Sayısı grafiğinde görselleştirilerek incelenmiştir.

4. Veri Analizi Soruları

Proje amacı kapsamında istenilen 10 adet başlangıç ve orta düzeydeki sorular aşağıdaki gibidir.

Başlangıç seviye:

1.Dizilerin türleri nelerdir ve her türde kaç dizi bulunmaktadır?

2.Dizilerin IMDb puanlarının dağılımı nasıldır? Hangi puan aralıkları daha yaygındır?

3. En düşük ve en yüksek IMDb puanına sahip diziler

4.Hangi yıllarda daha fazla dizi çıkmıştır?

5.Dizilerin sürelerinin (runtime) dağılımı nasıldır? Hangi süre aralığı daha yaygındır?

6.Dizilerin sertifika dağılımı nasıldır? Hangi sertifikasyonlar daha yaygındır?

7. Dizilerin oy dağılımı nasıldır?

8.Hangi diziler en fazla oy almıştır?

9.Dizilerin özetlerinde en çok geçen kelimeler nelerdir?

10.Dizilerin oyuncu kadroları içerisinde en çok hangi oyuncular yer almıştır?

11.Dizilerin yönetmenleri arasında hangi yönetmen en fazla dizi yönetmiştir?

Orta Seviye:

1. Rating değeri 8.5'un üzerinde olanlar için Runtime dağılımı nasıl değişiyor?

2.Rating değeri 2'nin altında olanlar için Runtime dağılımı nasıl değişiyor?

3.En yüksek rating oranına sahip dizinin oyuncu kadrosu nasıldır?

4.En düşük rating oranına sahip diziyi dizinin oyuncu kadrosu nasıldır?

5.Number Of Votes sayısı 60k nın üzerinde ve altında olanlar için Rating dağılımı nasıl değişiyor?

6.En yüksek ve en düşük rating değerine sahip olan dizileri kim yönetmiştir?

7.En yüksek rating değerine sahip dizinin türü nedir?

8.En yüksek ortalamaya sahip 10 türü ve ortalama puanlarını yazdırın?

9. Son 5 yılda çıkan dizilerin rating dağılımları nasıl?

10. 100.000 oydan fazla oylanan dizilerin Certificate dağılımı nasıldır?

5. Makine Öğrenmesi ve Test Edilmesi

Bu adım, projenin ana hedefi olan rating tahminini gerçekleştirmek için veri setinin hazırlanması ve modelin oluşturulması aşamalarını içerir.

Öncelikle veri seti, proje gereksinimlerine uygun hale getirilmiştir. Bu aşamada şu sütunlar veri setinden çıkarılmıştır: 'Genre', 'Cast', 'Synopsis', 'Runtime', 'Certificate', 'Gross Revenue', 'Stars', 'Directors'. Bu sütunların projenin amacına katkı sağlayabileceği düşünülmüş ancak verilen bu sütunların list yapısı olması sebebiyle One-Hot Encoding uygunalamdığından projenin devamlılığı açısından veri setinden çıkarılmıştır.

Veri seti, polinom regresyon algoritması kullanılarak bir model oluşturmak için hazırlanmıştır. Rating tahmini yapabilmek için bağımlı değişken olarak 'Rating' kullanılmıştır. Bağımsız değişkenler ise 'Release Year' (Yayın Yılı), 'Number of Votes' (Oy Sayısı) ve 'Runtime_minutes' (Çalışma Süresi Dakika cinsinden) olarak seçilmiştir.

Polinom regresyon modeli, verilerin ikinci dereceden polinomlarını dikkate alarak eğitilmiştir. Bu, değişkenler arasındaki karmaşık ilişkileri yakalamaya yardımcı olur.

Modelin başarısını değerlendirebilmek için 10 adet sentetik veri oluşturulmuştur. Bu veriler, gerçek verilere benzer özelliklere sahiptir. Oluşturulan sentetik veriler, polinom regresyon modeli kullanılarak tahmin edilmiştir.

Gerçek hedef değerleri ile modelin tahminleri karşılaştırılarak performans değerlendirmesi yapılmıştır. Modelin başarı ölçütleri olan MAE (Mean Absolute Error), MSE (Mean Squared Error) ve RMSE (Root Mean Squared Error) hesaplanmıştır. Bu değerler, modelin tahminlerinin gerçek değerlerden ne kadar sapma gösterdiğini gösterir.

Sonuçlar aşağıdaki gibidir:

Polinom Regresyon MAE: 1.3790854168627402
Polinom Regresyon MSE: 2.0942964304754765
Polinom Regresyon RMSE: 1.4471684181447149

6. Sonuç

Bu proje, IMDb TV dizilerinin rating tahmini yapma amacıyla başladı ve başlangıçta veri seti seçimi ve eksik verilerle ilgili bazı zorluklarla karşılaşıldı. Ancak, projenin ilerleyen aşamalarında bu zorluklar aşıldı ve başarıyla tamamlandı.

Veri seti, Kaggle platformunda bulunan "IMDB TV Series Data" veri seti olarak seçildi. Veri seti, dizilerin başlık, IMDb ID, yayın yılı, tür, oyuncu kadrosu, özet, derecelendirme, çalışma süresi, sertifika, oy sayısı ve brüt gelir gibi önemli özellikleri içeriyordu. Ancak veri setinin bazı eksik verileri vardı ve bazı sütunlar proje amacına uygun olmadığı için çıkarıldı.

Veri seti üzerinde yapılan veri analizi ve görselleştirmeler, projenin amaçlarına ulaşmada önemli bir rol oynadı. Rating dağılımı, çıkış yıllarının grafiği, türlerin dağılımı, rating ve oy sayısı ilişkisi, sertifika dağılımı ve süre dağılımı gibi faktörler incelenerek projenin daha iyi anlaşılması sağlandı.

Makine öğrenimi aşamasında, polinom regresyon modeli kullanılarak rating tahmini yapıldı. Modelin başarısını değerlendirmek için sentetik veriler oluşturuldu ve gerçek hedef değerleri ile karşılaştırıldı. Elde edilen sonuçlar, modelin kabul edilebilir bir başarı elde ettiğini gösterdi.

Sonuç olarak, bu proje, IMDb TV dizilerinin rating tahminini başarıyla gerçekleştirdi ve veri analizi, makine öğrenimi ve görselleştirmeler gibi çok sayıda veri bilimi becerisi kullanılarak tamamlandı. Başlangıçta yaşanan zorluklar, projenin sonunda başarılı bir şekilde aşıldı ve projenin amacına ulaşılmasını sağladı.

7. Referanslar

https://www.kaggle.com/datasets/suraj520/imdb-tv-series-data?select=comedy_series.csv