Film Öneri Sistemleri

Ege University 2021

Film Öneri Sistemleri 11.07.2021

Yaren Gündüz Birol Çiloğlugil Ünburak Öznur

[*Yarennn.gunduzzz@gmail.com*](mailto:Yarennn.gunduzzz@gmail.com)[*birol.ciloglugil@o365.ege.edu.tr*](mailto:birol.ciloglugil@o365.ege.edu.tr)[*burakoznur93@gmail.com*](mailto:burakoznur93@gmail.com)

Bilgisayar Mühendisliği Bölümü

Mühendislik Fakültesi, Ege Üniversitesi, İzmir

Özet

Öneri Sistemleri uzun yıllardır araştırmacıların odağında olan bir konudur. Kökleri bilgi elde etme sistemlerinde dayanan bu araştırma alanı, pratik uygulamalar konusunda da dikkat çekicidir.İlk bakışta oldukça basit bir konu gibi görünen öneri sistemleri, derinlemesine incelendiğinde her birisi ayrı bir araştırma konusu olabilecek pek çok zorluk içermektedir. Günümüzde öneri sistemleri pek çok alana uygulanmasına rağmen, tam verimli ve etkili çalışan bir öneri sistemi yapmak oldukça zordur ve makine öğrenmesinden içerik analizine kadar oldukça karmaşık yöntemler içermektedir. Bu çalışmada öneri sistemleri ve öneri sistemlerinde kullanılan yöntemler açıklanmış, pratikteki uygulamalarından örnekler verilmiş ve öneri sistemlerinin en ilginç uygulama alanlarından birisi olan haber öneri sistemleri açıklanmıştır.

**Anahtar Kelimeler:** Öneri Sistemleri , Film Öneri Sistmeleri, İşbirlikçi Filtreleme, İçerik Tabanlı Filtreleme, Bilgi Elde Etme

# Giriş

Öneri Sistemleri son yıllarda giderek daha popüler hale gelmiştir.E-ticaret sitelerinde bulunan ürünlerden müzik endüstrisine kadar birçok çeşitli alanda kullanılmaya başlanmıştır.eBay, Amazon, Alibab gibi e-ticaret sitelerinin çoğunluğu, müşterilere sevdikleri ürünlerle kişiselleştirilmiş bir hizmet verebilmek için öneri sistemlerinden faydalanmaktadır.Amazon’un gelirinin %35’inin öneri motorundan geldiği bilgisini verirsek, öneri sistemlerinin merak edilesi bir teknoloji olduğu konusunda hemfikir olabiliriz.

Kullanıcının seçeneklere vereceği yanıtları tahmin etmeyi içeren kapsamlı web uygulamaları sınıfına Öneri sistemleri denilmektedir.Öneri sistemleri genellikle tavsiye sistemleri olarak da adlandırılmaktadır.Büyük miktarda veri bulunan havuzdan faydalı içerikleri filtreleyerek kullanıcıya yönelik en anlamlı ve doğru ürünleri sağlamayı amaçlayan algoritmalardır.Öneri motorları da tüketicilerin seçimlerini öğrenerek veri setindeki veri desenlerini keşfeder ve kullanıcının ihtiyaçları ve ilgi alanları doğrultusunda sonuçlar üretir.

## Öneri Sistemleri Türleri

Birçok kaynağı inceledikten sonra temel olarak dört önemli öneri sistem türü mevcuttur.Bu filtreleme sistemleri:

* İşbirliğine Dayalı Filtreleme Sistemleri
* İçeriğe Dayalı Filtreleme Sistemleri
* Popülarite Tabanlı Öneri Sistemleri
* Hibrit Öneri Sistemleri

olarak sıralanır.

A picture containing text, clock

Description automatically generated

*Şekil 1. Öneri Sistemleri*

### **İşbirliğine Dayalı Fltreleme Sistemleri**

Bu filtreleme yöntemi ile aslında uygulamayı veya hizmeti kullanan kullancılara kendi davranışları,tercihleri hakkında bilgi toplamaya, analiz etmeye ve diğer kullanıcılarla olan benzerliğini temel alıp nelerden hoşlanabileceklerini ,neleri yapmak hoşlarına gidebilir bunu tahmin etmeye çalışıyoruz. En önemli avantajlarından biri analiz için içeriğe ihtiyaç duymaması ve bu nedenle nesnenin kendisini anlamasını gerektirmeden karmaşık öğeleri doğru bir şekilde önerebilmesidir.İşbirliğine dayalı filtreleme, geçmişte benzer tercihlere sahip olan bireylerin gelecekte aynı tercihleri yapacakları varsayımına dayanmaktadır.Diagram

Description automatically generated

*Şekil 2.İşbirlikçi Filtreleme*

İşbirliğine dayalı filtreleme algoritmalarına örnekler;

* ***Kullanıcı – Kullanıcı İşbirliğine Dayalı Filtreleme***: Benzer müşteriler ilşkilendirilmeye çalışılır ve müşterilerin seçtği ürünlere dayanarak ürünler sunulur.Etkili bir yöntemdir fakat zaman ve kaynak analizi dikkatli yapılmalıdır.Bu nedenle büyük platformlar için bu algoritmayaı uygulamak zordur.
* ***Öğe*-Öğe İşbirliğine Dayalı Filtreleme:** Önceki algoritmaya çok benzerdir fakat müşteriler arasında benzerlik bulmak yerine ürün benzerliğine odaklanır.Bu algoritma ile birlikte herhangi bir öğeyi satın alan müşteriye, benzer öğeleri kolaylıkla önerebiliriz. Kullanıcı-Kullanıcı işbiriliğine dayalı filtrelemeye kıyasla daha az kaynak ve zaman gerektirmektedir.
* ***Diğer Basit Algoritmalar:*** Hem kullanıcı-kullanıcı hem de öğre-öğe tabalı yaklaşımlardan farklı olarak daha yüksek tahmin gücüne sahip olmayan yaklaşımlar mevcuttur.

### **İçeriğe Dayalı Filtreleme Sistemleri**

Bu filtreleme yöntemi, bir ürünün açıklamasına ve kullanıcının tercih ettiği seçeneklerin içeriğine dayanmaktadır.İçerik tabanlı bir öneri sisteminde, anahtar kelimeler ürünü tanımlamak için kullanılmaktadır.Algoritmalar, bir kullnıcının geçmişte beğendiklerine benzer ürünler önermeye çalışır.İçeriğe dayalı filtreleme fikri, bir ürünü beğendiyseniz bu ürüne benzer başka bir ürünü de beğeneceğinize dayanmaktadır.İçeriğe dayalı filtreleme ile ilgili en büyük sorun, ürünlerin içerik kaynağından kullanıcı tercihlerinin öğrenip öğrenilemeyeceğidir.

A picture containing chart

Description automatically generated

*Şekil 3.İçerik Tabanlı Filtreleme*

### **Popülarite Tabanlı Öneri Sistemleri**

Temelde trend olan mevcut ürünleri kullanır.Örneğin, sıklıkla her yeni kullanıcı tarafından satın alınan herhangi bir ürün, yeni kayıt olan bir kullanıcıya önerilebilinir.Popülerlik tabanlı öneri sistemler ile ilgili bazı sorunlar mevcuttur.Popülerliğe dayalı öneri sistemlinin en büyük sorunu verilerin kişiselleştirme için kullanılamamasıdır.Kullanıcının davranışını bilinse bile, ürünleri bu seçenklere göre önerme imkanımız yoktur.

### **Hibrit Öneri Sistemleri**

Son araştırmalar, işbirlikçi ve ieçriğe dayalı öerilerin birleştirilmesinin daha etkili olabileceğini göstermektedir.Hibrit yaklaşımlar, iki öneri sistemi ayrı ayrı yapıalrak ve ardından birleştirilerek yugulanabilir.Ayrıca, işbirliğine dayalı bir yakaşlıma içerik tabanlı özellikler ekleyerek veya içerik tabanlı bir öneri sistemeine işbirliğine dayalı özellikler ekleyerek bu yaklaşımlar tek modelde bileştirilebilmektedir.Hibrit öneri ssiteminin performansı saf işbirlikçi ve içerik temelli yöntemlerle karşılaştırmaya odaklanan saf işbirlikçi ve içerik temelli yöntemlerle karşılaştırmaa odaklanan birkaç çalışma yapılmış ve hibrit yöntemlerin saf yaklaşımlardan daha doğru tavsiyeler sunabileceği kanıtlanmıştır.Hibrit öneri sistemi, veri yetersizliği gibi yaygın sorunların üstesinden gelmek için kullanılabilir.

## Film Öneri Sistemi Yapılandırılması

Kullanıcılarımıza öneriler almak için henüz izlemedikleri filmler için derecelendirmelerini tahmin edeceğiz. Filmler daha sonra dizine eklenir ve bu tahmin edilen derecelendirmelere göre kullanıcılara önerilir.Bunu yapmak için, gelecekteki derecelendirmelerini tahmin etmek için geçmiş film kayıtlarını ve kullanıcı derecelendirmelerini kullanacağız. Bu noktada, gerçek dünyada muhtemelen yeni kullanıcılarla veya geçmişi olmayan filmlerle karşılaşacağımızı belirtmekte fayda var. Bu tür durumlara soğuk başlangıç ​​problemleri denir. Soğuk başlatma sorunları, meta bilgilere dayalı önerilerle ele alınabilir. Yeni kullanıcılar için önerileri tahmin etmek için konumlarını, yaşlarını, cinsiyetlerini, tarayıcılarını ve kullanıcı cihazlarını kullanabiliriz.Yeni filmler için, kullanıcıları hedeflemek üzere önermek için tür, oyuncu kadrosu ve ekibi kullanabiliriz.

Tavsiye sistemimiz için, yukarıda bahsedilen ve okuduğumuz makalelerden çıkarım yaparak her iki tekniği de kullanarak dikkate almaya çalıştık.İçerik tabanlı ve işbirliğine dayalı filtreleme. İçerik tabanlı yöntemimiz için filmler arasındaki benzerliği bulmak için, bir kosinüs benzerlik işlevi kullanacağız. İşbirliğine dayalı filtreleme yöntemimiz için bir matris çarpanlara ayırma tekniği kullanacağız.

Bunun için ilk adım, matris çarpanlara ayırmaya dayalı bir model oluşturmaktır. Son modele girdi sağlamak için bu modelin çıktısını ve el yapımı birkaç özelliği kullanacağız. Temel süreç şöyle görünecek:

* Adım 1: Matris çarpanlarına dayalı bir model oluşturun
* Adım 2: El yapımı özellikler oluşturun
* Adım 3: Nihai modeli uygulayın

### **Matris Ayrıştırma Tabanlı Algoritma:**

Matris çarapanlara ayırma, öneri sistemlerinde kullanılan bir işbirliğe dayalı filtreleme algoritmaları sınıfıdır.Bu yöntem ailesi, ne kadar etkili olduğu için Netflix ödül yarışması sırasında yaygın olarak tanındı. Matris çarpanlara ayırma algoritmaları, kullanıcı-film etkileşim matrisini iki düşük boyutlu dikdörtgen matrisin, örneğin U ve M'nin ürününe ayrıştırarak çalışır. Ayrıştırma, ürünün kullanıcı-film etkileşim matrisine neredeyse benzer değerlerle sonuçlanacağı şekilde yapılır . Burada, U kullanıcı matrisini, M film matrisini, n kullanıcı sayısını ve m film sayısını temsil etmektedir. Kullanıcı matrisinin her satırı bir kullanıcıyı temsil eder ve film matrisinin her sütunu bir filmi temsil eder.Diagram

Description automatically generated

*Şekil 4.Matris Çarpanlara Ayırma*

Kullanıcı-film etkileşim matrisindeki boş olmayan hücrelere dayalı olarak U ve M matrislerini elde ettikten sonra, U ve M'nin çarpımını gerçekleştiriyoruz ve kullanıcı-film etkileşim matrisindeki boş olmayan hücrelerin değerlerini tahmin ediyoruz. Matris çarpanlarına ayırmayı uygulamak için, öneri sistemleri oluşturmak ve test etmek için Sürpriz adlı basit bir Python kitaplığı kullanıyoruz. Veri çerçevesi, Sürpriz kitaplığı tarafından kabul edilecek bir veri seti formatı olan bir tren setine dönüştürülür.

### **El Yapımı Özellikler Oluşturma:**

Veri çerçevesi formatındaki verileri bir kullanıcı-film etkileşim matrisine çevirdik. Bu tür problemlerde kullanılan matrisler genellikle seyrektir çünkü kullanıcıların yalnızca birkaç filmi derecelendirme şansı yüksektir. CSR formatı olarak da adlandırılan seyrek matris veri formatının avantajları aşağıdaki gibidir:

* Verimli aritmetik işlemler: CSR + CSR, CSR \* CSR, vb.
* Verimli sıra dilimleme
* Hızlı matris vektör ürünleri

Bu seyrek matrisi kullanarak 3 set özellik oluşturacağız:

1. Global ortalamaları temsil eden özellikler
2. Benzer ilk beş kullanıcıyı temsil eden özellikler
3. En iyi beş benzer filmi temsil eden özellikler

Kullanacağımız üç küresel ortalama.

* Tüm kullanıcılar tarafından verilen tüm filmlerin ortalama derecelendirmeleri. Tüm kullanıcılar tarafından verilen belirli bir filmin ortalama derecelendirmeleri.
* Belirli bir kullanıcı tarafından verilen tüm filmlerin ortalama derecelendirmeleri

Bu özellik setinde, belirli bir kullanıcı tarafından derecelendirilen ilk 5 benzer filmi elde ederiz. Bu benzerlik, filmler arasındaki kosinüs benzerliği kullanılarak hesaplanır. İşte içeriğinin daha ayrıntılı bir dökümü:

* GAvg: Tüm derecelendirmelerin ortalama puanı
* Bu filmin benzer kullanıcı derecelendirmesi: sur1, sur2, sur3, sur4, sur5 (o filmi derecelendiren ilk 5 benzer kullanıcı)
* Bu kullanıcı tarafından derecelendirilen benzer filmler: smr1, smr2, smr3, smr4, smr5 (kullanıcı tarafından derecelendirilen en iyi 5 benzer film)
* UAvg: Kullanıcı Ortalaması derecelendirmesi
* MAvg: Bu filmin ortalama değerlendirmesi
* derecelendirme: Bu filmin bu kullanıcı tarafından derecelendirilmesi.

### **Nihai Model Uygulaması:**

Bir öneri sisteminin performansını değerlendirmenin iki ana yolu vardır: Kök Ortalama Karesel Hata (RMSE) ve Ortalama Mutlak Yüzde Hatası (MAPE). RMSE kayıp karesini ölçer, MAPE ise mutlak kaybı ölçer. Daha düşük değerler, daha düşük hata oranları ve dolayısıyla daha iyi performans anlamına gelir.Bu ölçümleri hesaplayarak model performanısı hakkında bilgi edinebiliriz.

## Film Öneri Sistemleri Uygulama Alanları

Film öneri sistemleri günümüzde eğlence sektöründe oldukça karşımıza çıkan tavsiye sistemleridir.Bu sistemlerin baz aldığı filtreleme yaklaşımları ile hem kullanıcı hem de firma ihtiyaçlarını karşılayabilmeyi hedef almışızdır.

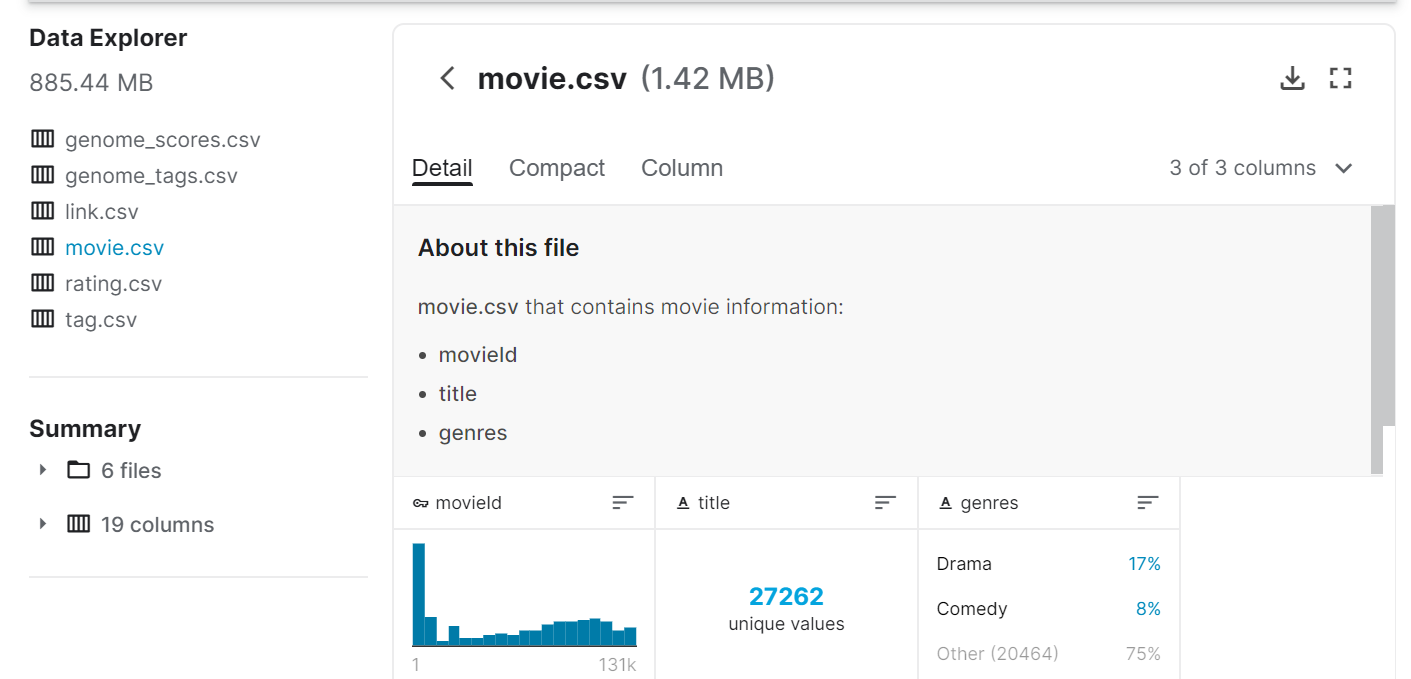
* Netflix
* AmazonPrime
* Taste
* Charmy
* Raf

Gibi uygulamalarda bu sistemlerden beslenerek karşıdaki bireyin sevebileceği veya izlemekten keyif alacağı sistemler oluşturabilir.

## Makale Gözlemleri

# İncelediğimiz makalelerden ve araştırdığımız yapılardan incelediğimize göre kendi yöntemlerimizi uygulayarak hem işbirlikçi hem de içerik tabanlı filtrelem yöntemlerini surprise kütüphanesini kullanarak yapmayı planladık.

# Veri Kümesi

Veri kümesi güncel olarak movielens veri kümesinin 20M veri gözlemi içeren tabular formatını kullandık.Bunun içerisinde birbiri ile ilgili 6 tane ayrı ama ilşkili csv dosyaları vardır.Bu dosyalarda birbiri ile ilgili bağıntılar bulduk.Her bir bağıntıyı birleştirerek son çalışacağımız formu oluştruduk.

*Şekil 5. Veri Kümesi Yapısı*

Verisetleri üzerinde yaptığımız modifikasyondan sonra Film ID’si, adı ve tiplerini ve Puanlarını bir arada topladık.

Veri içerisinde keşfsel veri analiini yapıp, istatistiksel olarak dağılımlarını minimum,maximum değerleri ile filmlerin veri kümesi içerisindeki dağılımlarını kontrol ettik.

Film isimlerine göre puanlama miktarlarını ve veri içerisinde çıkarabileceğimiz görselleştirme adımlarını tamaladık.Kullanıcıların filmlere verdiği puanları bir pivot table’da toplayarak her film için verilen puanlamaları toplayıp, hiç puan verilmeyen veya puan olarak en düşük filmleri veri kümesini kalabalıklaştırmamak adına çıkarabileceğimizi düşündük.

En sonunda da K-Nearest Neighbors algoritması kullandık. bu algoritmalar denetimli bir makine öğrenmesi algoritmasıdır. Bu parametrik değildir. Veri setinden herhangi bir varsayım öngörmeyen, Lazy algoritması KNN'nin herhangi bir eğitim adımı yoktur. Bu algoritma hem sınıflandırma hem de regresyon problemlerinde kullanılır.

KNN, gerçek ideal değer için K olarak bir değer seçme yolu gibi çalışır. Daha sonra K noktaları ve tüm gözlem noktaları arasındaki tüm mesafeleri hesaplarız. Bundan sonra, yeni veri noktalarına KNearestNeighbors'u bulabiliriz. Sınıflandırma problemleri için, her kategorideki veri noktalarının sayısını K değerleri arasında sayarız. Regresyon için, yeni veri noktasının değeri K Komşularının ortalaması olacaktır.mMesafe kullanılarak hesaplanabilir.

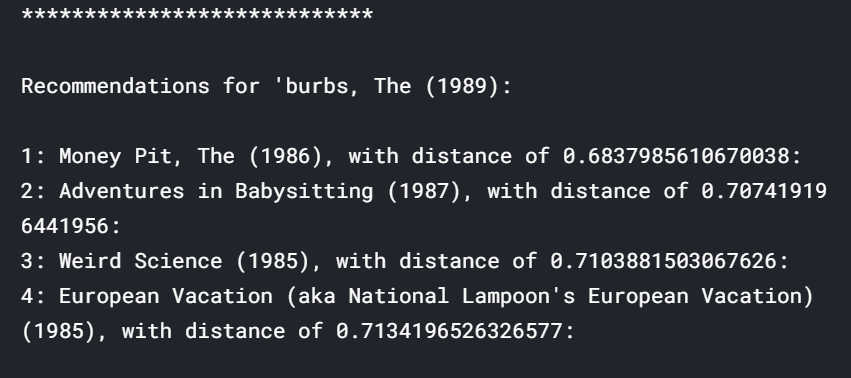
* Öklid mesafesi,
* Manhattan mesafesi,
* Hamm Mesafesi,
* Minkowski Mesafesi.

Öklid uzaklığı, bağımsız ve bağımlı değerlerin farklarının karesi ve tahminleri alınarak hesaplanır. Manhattan mesafesi, gerçek ve tahmin edilen değerler arasındaki farkların mutlak değeri alınarak hesaplanır. Hamming mesafesi kategorik değerler gruplanarak hesaplanır. Kategorik değerlerim büyük ve orta olarak gruplanırsa 1, aksi takdirde büyük-büyük kombinasyon sadece 0 döndürür. Minkowski mesafesi, gerçek ve tahmin edilen değerler arasındaki farkların mutlak değerlerinin gruplanmasıyla hesaplanır. Toplam tekrar sayısı şu şekilde verir: bir üstel.Sonuçlar üstel de 1/p olacak.KNN artıları ve eksileri de sağlar.

***Avantajları***: Basit ve uygulaması kolay. Parametrik olmayan - altta yatan verilerle ilgili varsayım yok. Hem sınıflandırma hem de regresyon. Eğitim adımı daha hızlıdır.

***Dezavantajları***: KNN, yeni nokta tahminleri için pahalıdır. Aykırı değerlere duyarlı. Doğruluk, gürültü ve alakasız veriler nedeniyle hedeften sapabilir.

Bu şekilde eriştiğimiz sonuçlaradan bir kesit sunacak olursak Şekil 6’da göreceğimiz gibi bir filmin benzerlerine olan yakınlığı hesaplanığ gerekli hesaplamalar kurabiliyoruz ve bunu bir web hizmeti olarak da sunabiliyoruz.



*Şekil 6.Öneri Sistemi Sonuç Kesiti*

# Sonuç

Öneri sistemleri günlük yaşantımızda biz farkında olmasak da internette çokça karşımıza çıkmaktadır.Öneri sistemleri daha çok film, müzik, alışveriş, sosyal medya ve haber alanlarında yaygın olarak kullanılmaktadır. Şimdiden öneri sistemleri hayatımızda çokça rol almaya başlamıştır ve gelecekte daha etkin sistemler yapılacak, daha da yaygınlaşacaktır. Bu bildiride öneri sistemleri ve öneri sistemlerinde kullanılan yöntemler açıklanmış.Film öneri sistemleri özellikle de eğlence sektörünün vazgeçilmezleridir ve her an geliştirilmeye daha açıktır.Bu yüzden bu hizmetlerin herhangi bir noktada sonlaancağı fikrine kapılmak doğru bir yere ulaştırma bizi. Bu yüzden üzerinde çalışılarak daha iyi geliştirmeler yapılabilinir.

# Teşekkür

Çalışmalarım sırasında bu projeyi azimle gerçekleyen Yaren Gündüz & Ünburak Öznur’a ve bize danışmanlık yapan değerli Dr.Öğretim Üyesi Birol Çiloğlugil’e teşekkürlerimizi borç biliriz.

# Kaynaklar

[1]<https://www.kaggle.com/grouplens/movielens-20m-dataset?select=movie.csv>

[2]<https://dl.acm.org/doi/abs/10.1145/3025171.3025209?casa_token=GnypkiVBlOQAAAAA:DZU7GdUEYQdTc4U7YWXjrtyuTSTZBW0TsU4yzsDorDy6Cgj695JHpYHPB7a9mToy9jGhsqmk7NqTqw>

[3]<https://d1wqtxts1xzle7.cloudfront.net/33554940/V3I4201494-with-cover-page-v2.pdf?Expires=1625924969&Signature=CsL66fFXctAAZjsekjN~twh4ie0XAmbdsT9QuN-cavlVuXV2mTz-IjJy29shFAO0M~DroVav-POmH1Ky2bd9pkhYJyZjvumHaR~EMcDjI433~T4sMG5jh4OiHQ~uqnZsMSz1exSonOEtCNV0wC2EXRH-IrWH3v71L9ee-jYEfc5E0Z~KiyqK13LwFQb9JuiQbZ31FKXtXhq~uV7p4g-YLGCSLUVLV4BjbhIUcj~S54z55GvDIi5SAQR2zw5ACWow3-WeaZIchgtbTq64YC0zx686wzW720~LrBPme4SgTAp-ZNESw8vPY8DezvQ0CKlvRez7NgsyPlKB~gD7SNOsVA__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA>

[4]<https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.20.9955&rep=rep1&type=pdf>

[5] <https://ieeexplore.ieee.org/abstract/document/6382910>

[6] <https://calteches.library.caltech.edu/4333/>

[7]<https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.149.2642&rep=rep1&type=pdf>

[8]<https://dl.acm.org/doi/abs/10.1145/2623330.2623758?casa_token=aqEvKWqaI90AAAAA:pIzmDcR-vDQFpnad8WmaRoWGfFDOnQ_SqHl9FP6D0Zi5OV-Z75BF68yYFR_S8IkMAsTFPFNAQIBxsQ>

[9]<https://www.sciencedirect.com/science/article/pii/S1045926X14000901>

[10] <https://link.springer.com/article/10.1007/s11042-006-0082-7>

[11]<https://www.sciencedirect.com/science/article/abs/pii/S0957417412001509>