**Veri Madenciliğinde Türkiye Genelinde Sinemaları Verilerinin Değerlendirilmesi**

**Dursun DEMİR**

Mehmet Akif Ersoy Üniversitesi- Bilgisayar Mühendisliği

dursunzaman2261@gmail.com

**Özet**

Veri madenciliği yöntemleri ile eldeki veriler sınıflandırılarak, gruplandırılarak ya da veriler arasında ilişkiler, bağıntılar, istatistiksel sonuçlar oluşturularak modeller oluşturulur. Oluşturulan model, oluşturulduğu veri kümesinde olmayan yeni bir kayıt geldiğinde, yeni gelen kayıt hakkında tahminleme yapma imkanı verir. Yapılan tahminlerin doğruluk derecesi oluşturulmuş olan modelin veri üzerindeki başarımını ortaya koyar. Dolayısı ile bir veri madenciliği uygulamasında hangi algoritma ile daha iyi sonuçlar üretildiği uygulamanın başarımı açısından önemlidir. Ayrıca sürekli geliştirilmekte olan yeni algoritmaların başarım derecesinin var olan algoritma sonuçları ile karşılaştırılması yeni geliştirilen algoritmanın kabul edilebilirliğini ortaya koyması açısından önemlidir. Bu makale WEKA (Waikato Environment for Knowledge Analysis) programı ile tuik 2019 yılı içerisindeki verileri türkiye genelinde sinema salonlarının değerlendirmesi yapılmıştır.

**Anahtar Kelimler:** Veri Madenciliği, weka, sinema

**Abstract**

Relationships are attractive, with models that they can achieve, whether by teaching giving, by pointing out, grouping, or looking from the road. The created model may not provide information about the incoming record in a new session that is not in the new dataset it was created. It reveals the efficiency of the model that applies the charges of the estimates made. A performance is important about how good results are produced. At the same time, it is important to demonstrate the acceptability of renewing existing achievements in breeding achievements. This article is an evaluation of movie theaters across Turkey regarding the WEKA (Waikato Environmental Information Analysis) program and the monogram 2019 data.

**Keywords:** Data mining , weka ,cinema

**1.Giriş**

Gelişen bilgisayar teknolojileri ile birlikte her geçen gün kullanılmakta olan veri miktarı da hızla büyümekte; hızla artan veri miktarı da bu verilerin analiz edilmesini zorlaştırmaktadır. Bir tahmine göre her 20 ayda bir dünyadaki veri miktarının ikiye katlandığı ifade edilmektedir[1]. Son on yıllarda verileri analiz ederek yararlı ve anlamlı bilgilere ulaşmak için kaynağını istatistik, yapay zeka ve makine öğrenmeden alan veri madenciliği disiplini oluşturulmuş, endüstri, ekonomi, ve iş çevrelerinden gelen yoğun ilgi ile beraber akademik alanda da ilgi odağı olmuştur. Sağlık, temel bilimler, bankacılık, finans, pazar araştırmaları gibi pekçok alanda kullanılan veri madenciliği yöntemleri çok sayıda algoritma barındırır. Veri madenciliğinde amaç, analiz edilerek bilgi çıkarılması zor olan büyük veri yığınlarını analiz ederek anlamlı, gizli ve faydalı olabilecek bilgi çıkarmak; bu bilgileri içerisinde barındıran bir model oluşturarak yeni gelecek bir veri nesnesi hakkında yorum yapmayı ve bu veri hakkında tahminde bulunmayı sağlamaktır. Veri madenciliği algoritmalarının empirik yollarla karşılaştırılarak değerlendirilmesi tüm bilimsel çalışmalarda yapılagelmiş olsa da akademik literatürde böyle çalışmaların objektif ve kesin sonuçlar üretmeyeceğine dair eleştiriler mevcuttur. Bu eleştiriler model oluşturuma basamaklarında uygulanan veri önişleme, parametre seçimi, test ve öğrenme kümelerinin oluşturulması gibi adımların uygulamayı yapan kullanıcının insiyatifinde olmasından dolayı oluşturulan model başarımlarının uygulamayı yapan kullanıcıya bağlı olmasından kaynaklanmaktadır. Bir başka eleştiri de yeni bir algoritmanın var olan eski algortmalarla kıyaslandığı akademik çalışmalarda yapılan uygulamalarda karşılaştırmaların geliştiricinin yanlı duruşu nedeni ile objektif bir sonuç vermeyeceği yönündedir[2] .Tüm bu eleştirilere rağmen algoritmaların karşılaştırılması gerekliliği ortak bir görüş olarak kabul edilmiş, gerek uygulama gerekse geliştirme anlamında yapılan akademik çalışmalarda ve güncel uygulamalarda yer edinmiştir.

**2.Veri Madenciliği Süreci**

Veri madenciliği süreci dört aşama ile tanımlanabilir. İlk aşamada problem tanımlanarak veri kaynakları değerlendirilir. İkinci aşamada veriler kullanıma uygun hale getirilmek için hazırlanır. Arkasından model kurulur ve nihai aşamada model değerlendirilerek kullanıma hazır hale getirilir.

**2.1.Problemin Tanımlanması**

Amaç, işletme problemine verileri kullanarak çözüm getirmek olduğundan, ilk olarak ihtiyaç duyulan şey tam olarak tanımlanmalıdır. Bu problem, işletmenin ayrılmakta olan müşterisinin belirli özelliklerini tanımlayarak ona uygun davranmak olabildiği gibi, kendi kaynaklarını optimum kullanabilmek için yapacağı bir planlamada gelecek dönemdeki harcamalarını tahmin etmek şeklinde de olabilir. “Bu adımda ihtiyaç duyulan şeyin tanımlanması için cevaplanması gereken sorular neyin otomatize edilmeye değer olduğu ve neyin insan içeren süreçlere bırakılması gerektiği, amacın ne olduğu ve hangi performans kriterlerinin daha önemli olduğu, sürecin sonucunda elde edilecek çıktının keşif, sınıflandırma, özetleme gibi şeyler için kullanılıp kullanılmayacağı olabilir.” [3] Problemin tanımlanması durumunda ihtiyaç duyulan iş modelinin kalıbı da belirlenmiş olur.

**2.2.Modelin Kurulması**

Modelin kurulması aşamasında birçok model denenerek veriyi en iyi temsil eden model seçilir. Verileri temsil eden en iyi modeli bulabilmek için çok sayıda model kurulmalı, en iyi sonucu alana kadar denemeye devam edilmelidir. Modelin kuruluşu, amacımızın ne olduğuna, problemimizi ne şekilde çözmek istediğimize ve sonucun ne kadar işimize yarar olacağına göre değişebilir. Örneğin görmek istediğimiz gelecek dönemdeki tahmini ciromuz ise, sürekli bir değişkeni tahmin edeceğimiz doğrusal regresyon modelini; müşterilerimizin pasifleşme eğiliminde olup olmadıkları ise kategorik bir değişkeni tahmin edeceğimiz sınıflandırma modelleri olan karar ağaçlarını, yapay sinir ağını veya kategorik değişkenin olasılığını tahmin edeceğimiz lojistik regresyon modelini, hangi ürünlerimizin diğerlerine oranla daha çok beraber alındığı ise birliktelik analizi, beraber alınan bu ürünlerin hangi sırayla alındığı, nedensellikleri ise sıralı örüntü algoritmaları kullanılabilir. Ayrıca müşterilerimizin sahip oldukları alışveriş özelliklerine göre (gelme sıklıkları, uğradıkları mağazalar, satın aldıkları ürünler vb.) belirli gruplara ayırmak için kümeleme algoritmaları kullanılabilir. 11 Model kurulurken denetimli veya denetimsiz öğrenmeye göre farklı aşamalar uygulanmaktadır. Örneğin sınıflandırma algoritmaları kullanılırken tüm veri kümesi öğrenme ve test kümesi olarak ayrılmalı; modelin verilerden öğrenerek oluşturulması öğrenme kümesi, doğruluğunun kontrolü ise test kümesi ile gerçekleştirilmelidir. Kurulan modellerde birbiri ile ilişkili olan veya anlamsız olan değişkenlerin elenmesine dikkat edilmelidir. Amaç bilgi çıkarımı olduğundan ve birbiri ile ilişkili olan değişkenler bize ekstra bilgi vermediğinden, diğerine göre daha anlamlı olan değişkeni modele katmak faydamıza olacaktır.

**2.3Modelin Değerlendirilmesi**

Kurulan modellerin karşılaştırılarak veri kümesini en iyi temsil eden modelin seçildiği aşamadır. Karşılaştırma için, sınıflayıcının tahmin ettiği sınıfların oranını belirten doğruluk oranı kullanılır. Sınıflayıcının doğruluk oranının görece yüksek olması, diğer modellere göre veri kümesini daha iyi ifade ettiğini gösterebilir. Doğruluğun testi için kullanılan geçerlilik yöntemleri basit geçerlilik yöntemi, çapraz geçerlilik yöntemi, n-katlı geçerlilik yöntemi olarak sıralanabilir. Basit geçerlilik yönteminde verilerin bir kısmı test verisi olarak ayrılır, kalan kısım üzerinde modelin öğrenimi gerçekleştirildikten sonra ayrılan kısım üzerinde test işlemi yapılır. “Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile doğruluk oranı hesaplanır.” [4] Çapraz geçerlilik yöntemi daha az sayıda veri kümesine sahip olunduğu durumlarda kullanılabilir. Bu yöntemde veri kümesi rastgele seçilerek iki eşit gruba ayrılır, gruplar sırayla öğrenme ve test kümesi yapılarak elde edilen doğruluk oranlarının ortalaması kullanılır. N-katlı geçerlilik yöntemi de çapraz geçerlilik yöntemi gibi küçük veri kümeleri için

**3.Uygulama**

Uygulama Türkiye genelindeki tüm sinemaların salon verileri, salondaki mevcut koltuk sayıları, yıl içerisinde gösterime girmiş yerli ve yabancı filmleri , bu filmleri gitmiş seyirci sayıları değerlendirimiştir. Uygulamada bir karar ağacı algoritması olan ve temeli ID3 ve C4.5 algoritmalarına dayanan J48 ve simplecart uygulanmıştır. Uygulama neticesinde görülmüştür ki türkiye genelinde sinema noktasında iyileştirmeler yapılaması gerektedir.

**4.Kaynakça**

1. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Limited-Use Data (1973-2006), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2009, based on the November 2008 submission.

2. Witten, I. H. ; Frank, E. ; Data Mining, Practical Machine Learning Tools and Techniques ; Morgan Kaufmann , USA; 2005.

3.S. Sumathi, S.N. Sivanandam, Introduction to Data Mining and its Applications, New york, Springer 2006, s. 189

4. Yaralıoğlu, a.g.e., s.175