

ÖNERİ SİSTEMİ İLE ÜRÜN TAVSİYESİ

Product Recommendation with the Recommender System

Oğulcan AKCA

İzmir Demokrasi Üniversitesi, Elektrik Elektronik Mühendisliği Bölümü, akca_ogulcan@hotmail.com

ÖZET

Son zamanlarda bir uygulama ya da sistem, onu kullanan kullanıcıların tercihleri ile zaman geçtikçe gelişmektedir. Buna araç olan öneri sistemlerine, kullanıcı ile ürün ilişkisi olan tüm alanlarda mutlaka rastlanmaktadır. Bu alanlara dijital içeriklerin yer aldığı platformları, perakende satış yapan kuruluşları ve finansal danışmanlık kuruluşları örnek gösterilebilir. Ayrıca öneri sistemleri, uygulanan alana göre kullanıcıların geçmiş deneyimleri, sisteme dahil edilerek gelecek için strateji geliştirme gibi birçok rol üstlenmektedir. Bu makalede perakende satış verileri kullanılarak müşteriler için ürün önerilmesi amaçlanmıştır. Üç yıllık bir müşteri-ürün ilişkisine dayalı veri kümesi incelenmiş ve KNN, NormalSVD, MidUserSVD, MidItemSVD, MidTotalSVD ve Apriori gibi çeşitli algoritmalar kullanılarak farklı türde öneri sistemleri oluşturulmuştur. Ayrıca, bu öneri sistemlerinin çoğunun birden fazla algoritma ve yöntemin birleşiminden oluştuğu göz önünde bulundurulmalıdır. Makalede, RFM analizi ve K-means gibi yöntemlerle gerçekleştirilen müşteri segmentasyonu sonuçları temel alınarak tasarlanan öneri sistemi ele alınmıştır. Bununla birlikte, bir diğer öneri sisteminde ise kural çıkarımı ve market sepeti analizi amacıyla Apriori algoritması kullanılmıştır. Daha sonra, RFM analizinden elde edilen sonuçlar BG/NBD modeline girdi olarak verilerek ürün önerisi yapılmıştır. Bu iki öneri sisteminin geliştirilme süreci CRISP-DM metodolojisi temel alınmıştır. Hazırlanan sistemlerden birkaçında ihtiyaç duyulan ve veri kümesinde yer almayan müşteri tarafından yapılan ürün derecelendirme bilgisi literatür araştırmasında bahsi geçen imputasyon tekniğiyle oluşturulmuştur. Burada Pearson Correlation Coefficient (PCC) yönteminden yardım alınmıştır. Çalışmadaki öneri sistemleri kendi içinde yer alan değerlendirme metrikleriyle veya analizlerle sistemlerin zayıf ve güçlü yönleri ele alınmış ve başarısı sınanmıştır. Makalede kitleye uygulanacak öneri sisteminin ihtiyacın ne olduğu belirlenip de tasarlanması ilk adım olmasının gerektiği görülmüştür.

Anahtar Kavramlar: Perakende Satış, CRISP-DM, RFM Analizi, Market Sepeti Analizi, Kural Çıkarımı, SVD

ABSTRACT

The preferences of the users who use an application or a system have been evolving with the passage of time. Recommendation systems, which are a tool for this, are encountered in all areas where there is a user-product relationship. These areas include digital content platforms, retail organizations and financial consultancy organizations. In addition, recommendation systems assume many roles such as developing strategies for the future by incorporating users' past experiences into the system according to the field of application. This paper aims to recommend products for customers using retail sales data. A three-year dataset based on customer-product relationship is analyzed and different types of recommendation systems are created using various algorithms such as KNN, NormalSVD, MidUserSVD, MidItemSVD and MidTotalSVD. It should also be noted that most of these recommender systems are a combination of multiple algorithms and methods. In this paper, the recommendation system designed based on the results of customer segmentation using RFM analysis and K-means is discussed. However, in another recommendation system, Apriori algorithm is used for rule extraction and grocery basket analysis. Then, the results obtained from RFM analysis are given as input to the BG/NBD model and product recommendation is made. The development of these two recommendation systems was based on the CRISP-DM methodology. The product rating information provided by the customer, which is needed in a few of the systems and is not included in the dataset, is created by the imputation technique mentioned in the literature research. Pearson Correlation Coefficient (PCC) method was used here. The weaknesses and strengths of the recommendation systems in the study were addressed and their success was tested with the evaluation

metrics or analyzes included in the study. In the article, it was seen that the first step should be to determine and design the suggestion system to be applied to the audience.

Keywords: Retail, CRISP-DM, RFM Analysis, Market Basket Analysis, Association Rule Mining, SVD

Makalede kullanılan kodlara erişim için: <https://www.kaggle.com/oulcanakca/recommender-systems>

1 GİRİŞ

İnternette yer alan çeşitli platformlar, perakende satış yapan kuruluşlar gibi alanlarda kullanılan öneri sistemlerinde, bu alanda işlem yapan kullanıcıların tercihleri ve davranışları analiz edilerek yeni tercihler, kullanıcının ilgisini çekebilecek öneriler sunulur. Böylece bu tercih veya öneriler kullanılarak sistem, kullanıcıların alışveriş veya içerik tüketme kararlarını kolaylaştırır ve onların taleplerine daha iyi öneri verme hedefine ulaşmayı amaçlamaktadır [1]. Bu çalışma, ürün öneri sistemlerinin geliştirilmesi konusunda çeşitli yaklaşımların nasıl kullanıldığını incelemektedir. Özellikle, incelenecek yaklaşımların performansını değerlendirmeye de odaklanmaktadır. Aynı zamanda bu çalışmada, literatür araştırmasında uygulanması kararlaştırılan yöntemlerin ve algoritmaların kullanıldığı, veri madenciliği ve yapay zekâ teknikleri yardımıyla ürün öneri sistemleri geliştirilecektir.

Öneri sistemleri, kullanıcıların bilgi arayış süreçlerinde istedikleri nesnelere temsil eden ve aynı zamanda bu öğelerle ilgili kullanıcı gereksinimlerini ve tercihlerini göz önünde bulundurarak öneriler sunan akıllı uygulamalar olarak tanımlanabilir [2]. Ayrıca öneri sistemlerinde bahsi geçen kullanıcılar mekanizmanın yapı taşı olan esas veride, çoğunlukla ürün ile ilişkilendirilse de çeşitli amaçlarca kullanıldığında farklı tanımlamada da bulunulabilmektedir. Sosyal medya veya çevrim içi platformlarda, kullanıcıların benzer ilgi alanlarına sahip kişileri bulmalarına yardımcı olmak için arkadaşlık önerileri sunmak, yemek tarifi sitelerinde, kullanıcıların geçmiş tarif tercihlerini göz önünde bulundurarak, onlara yeni ve ilgi çekici yemek tarifleri sunmak ve kullanıcılara film veya müzik önermek için, kullanıcıların izledikleri filmler, dinledikleri müzikler veya tercih ettikleri türler gibi verileri kullanarak kişisel öneriler sunmak gibi örnekler verilebilir [3]. Örnekler dikkate alındığında kullanıcılar, her gün karşlarına çıkan büyük miktardaki içeriklerle başa çıkmak zorunda kalıyorlar. Bu durum, kullanıcı deneyimini karmaşıklştırabilir ve kullanıcıların istedikleri içeriği bulmalarını zorlaştırabilir. İşte bu noktada, öneri sistemlerinin önemi ortaya çıkıyor. Öneri sistemleri, bu içerik yığını içerisinde kaybolan kullanıcılara yön göstererek, en ilgi çekici ve uygun seçenekleri sunma konusunda kritik bir rol oynuyor. Bu çalışmada, öneri sistemlerinin nasıl kullanıcı davranışlarını analiz ettiği, verileri nasıl işlediği ve bu verilere dayalı olarak nasıl öneriler sunduğu ayrıntılı bir şekilde incelenecektir.

1.1 Öneri Sistemlerinin Özellikleri

Birçok farklı sektörde kullanılmakta olan öneri sistemleri e-ticaret platformlarından medya yayıncılığına ve sosyal medya platformlarına kadar geniş bir yelpazede uygulama ihtiyacına göre çözüm olarak ele alınmaktadır [4]. Öneri sistemleri, üç temel özellik ile incelenir: kişiselleştirme, nesne sıralama ve tahmin [5]. Kişiselleştirme, her kullanıcının benzersiz tercihleri kullanılarak tasarlanan sistemin sunulan önerilerin ne kadar spesifik olabileceğini ifade eder. Nesne sıralama, kullanıcıya sunulan önerilerin her kullanıcı için önem sırasını belirleme sürecidir. Tahmin, kullanıcının ilgisini çekebilecek nesnelere seçme ve bu seçimleri önceliklendirme sürecini ifade eder [5].

Genellikle kullanıcı verileri ve nesne verileri gibi iki temel veri tipini kullanarak çalışan öneri sistemlerinde, [6] kullanıcı verileri, kullanıcının geçmiş tercihleri, izleme davranışları, derecelendirmeleri ve profil bilgilerini içerirken nesne verileri ise nesnelere tanımlayıcı özelliklerini içerir. Bu veri tipleri, öneri sistemlerinin çalışma prensiplerini belirlemek için kullanılır. Örneğin, kullanıcının benzer tercihlere sahip diğer kullanıcıların davranışları analiz edilerek, öneriler yapılabilir [7]. Ayrıca, 2001 yılında yayınlanan Sarwar, B., Karypis, G. ve diğer araştırmacıların yayınladığı makaleye

göre [8] nesnelerin özellikleri ve kullanıcı tercihleri arasındaki ilişkileri öğrenen modeller geliştirilerek daha etkili öneriler sunulabilir.

Öneri sistemlerinin çalışma prensiplerini kavramları ele alan bir çalışmaya göre [9] tahmin değerleri, öneri sistemlerinin kullanıcılara yapacakları önerileri belirlemek için kullandığı değerler olduğunu belirtmektedir. Bu değerler, kullanıcının ilgisini çekebilecek nesnelerin tahminini temsil eder. Böylece, kullanıcının tercihleri analiz edilerek ve benzer kullanıcıların tercihlerinden yola çıkılarak tahmin değerleri belirlenir. Tahmin değerleri, kullanıcılara en uygun önerilerin yapılmasında önemli bir rol oynar. Bunların yanı sıra başarı ölçütü ve tahmin doğruluğu ise öneri sistemlerinin performansını değerlendirmek için kullanılan metrikler olduğu makalede bahsedilmiştir. Başarı ölçütleri, öneri sistemlerinin ne kadar etkili ve kullanışlı olduğunu belirlemek için kullanılırken tahmin doğruluğu ise öneri sistemlerinin ne kadar doğru öneriler yaptığını gösterebilir. Buna örnek olarak yüksek tahmin doğruluğu, kullanıcılara daha uygun ve tatmin edici önerilerin sunulması verilebilir. Fakat çalışmada geliştirilecek sistemlerde de bahsedileceği üzere bu çıkarımı yapmak tek başına yeterli değildir. Çünkü sistemin etkileyeceği kitlenin ihtiyacına göre seçilen yöntemin ve yöntemi değerlendirilecek metriklerin belirlenmesi sonrasında analiz ile çıkarım yapılması gerekmektedir [9].

1.2 Çalışmada Kullanılan Modeller

Öneri sistemleri, kullanıcıların ihtiyaçlarına yönelik kişiselleştirilmiş öneriler sunarak kullanıcı deneyimini artırmayı amaçlayan önemli bir araştırma alanıdır [10]. Bu sistemler, kullanıcı-ürün ilişkilerinin analiz edilmesiyle temellendirilir ve satın alma davranışları, kullanıcı özellikleri, ürün sayısı gibi faktörler dikkate alınır. Öneri sistemleri, kullanıcılara çeşitli öğeleri keşfetme fırsatı sunarken, işletmelerin de müşteri memnuniyetini artırma ve satışlarını optimize etme potansiyelini barındırır [10]. Bu bölümde, çalışmada veri kümesi üzerinde çalışılması uygun görünen modeller için sırasıyla tanıtım yapılmıştır. Ayrıca burada bahsedilen öneri modelleri göz önünde bulundurulduğunda modellerin etkinliği, kullanılan algoritma ve yöntemlerin seçimine bağlıdır.

1.2.1 Popülerlik Tabanlı Filtreleme Modelleri

Popüler olan ürüne göre çalışan bir tür tavsiye sistemi olan bu sistemler, trend olan veya kullanıcılar arasında en popüler olan ürünü kontrol eder ve doğrudan bunları önerir [11]. Kullanıcı bazlı öneri yapmayan popülerlik tabanlı öneri sisteminde en sık işlem gören (satın alma gibi) ürünlerden n tanesi seçilerek en popüler ürünler olarak adlandırılır. Bu ürünler kullanıcılara önerilir. Öneri sisteminin tasarlanma sürecine herhangi bir özel algoritma kullanılmasına gerek yoktur. Sistemin tasarlanması ve uygulanması zor olmazken popüler ürünler geniş bir kitle tarafından satın alındığı için genellikle öneriler müşteriler için uygun olabilir. Fakat müşteri özelinde kişiselleştirme bu sistemin zayıf noktalarından biridir. Bu nedenle ilk olarak bu modelin tasarlanması iyi bir başlangıç olabilir.

1.2.2 İçerik Tabanlı Öneri Modelleri

İçerik Tabanlı Öneri Modelleri, ürünlerin içeriklerini ve kullanıcı tercihlerini inceler, öneriler oluşturur [12]. Bu yaklaşım, ürünlerin açıklamaları, etiketleri ve kategorileri gibi içerik özelliklerini kullanır ve süreç boyunca kullanılan bu özellikler için benzerlik hesaplamaları gerçekleştirir. Ayrıca, kullanıcının geçmiş tercihlerini ve davranışlarını analiz ederek, kullanıcının ilgi alanlarına uygun şekilde özelleştirilmiş öneriler sunabilir. Bu yöntem, özellikle yeni ürünlerin tanıtılması veya özel niş tercihlerin göz önünde bulundurulması gereken durumlarda etkili bir şekilde kullanılabilir.

1.2.3 İşbirlikçi Filtreleme Modelleri

İşbirlikçi Filtreleme, kullanıcıların geçmiş değerlendirmeleri ve ilişkili verileri kullanır. Böylece bu yöntem kullanıcı benzerliklerini belirleyerek öneri yapar [13]. Tasarlanmış olan bu model, kullanıcılar benzer veya aynı ürünleri değerlendirdiğinde, bu kullanıcıların diğer ürünlerle ilgili görüşlerini tahmin etme yeteneğini öneri sunmak amacıyla kullanır [13]. “İşbirlikçi” terimi, müşteri ve ürün etkileşiminin gerçekleşmesiyle bunların beraber analizinin yapılması olarak ifade edilebilir [14].

1.2.4 Hibrit Öneri Modelleri

Hibrit Öneri Modelleri, çeşitli öneri tekniklerinin bir araya getirilerek daha kapsamlı ve doğru öneriler sunmayı amaçlar [15]. Diğer filtreleme yaklaşımlarının birleştirilmesiyle oluşturulan hibrit modeller, her bir tekniğin güçlü yönlerini bir araya getirerek kullanıcı deneyimini iyileştirir, kullanıcı tercihlerini belirler ve ürün önerileri oluşturur [16]. Bu yaklaşım, ürünlerin özelliklerini ve kullanıcı ihtiyaçlarını karşılaştırarak ürünlerin uygunluğunu değerlendirir [16] ve böylece kullanıcının geçmiş tercihleri ve davranışlarına dayandırılarak kullanıcıların ihtiyaçlarına en iyi şekilde cevap verebilecek ürünleri belirleyen sistem oluşturulur. Örneğin İşbirlikçi Filtreleme yöntemlerinde Apriori veya FP-Growth gibi ilişki kural çıkarma algoritmalarının kullanılması, kullanıcıların benzer tercihlerini veya davranışlarını tahmin etmekte yardımcı olabilir [17]. Bahsi geçen algoritmalar, veri kümesindeki sık nesnelere belirlemek için kullanılır ve öneri sistemlerinin temel taşlarından birini oluşturur [18].

2 LİTERATÜR ARAŞTIRMASI

2022 yılında yayınlanan makalede [19] ürünler, ürünleri satın alan müşterilerinin başka hangi ürünleri aldığını baz almadan çalışan popülerlik tabanlı öneri sistemini ele almışlardır. Başka bir deyişle herhangi bir benzerlik kurulmadan ve müşteri ayırt emeksizin ürün önerilmiştir. 2020 yılında yayınlanan bir makalede [20], K-Nearest Neighbors (KNN) algoritması ve Kosinüs benzerliği kullanılarak içerik tabanlı ve işbirlikçi filtreleme tabanlı öneri sistemi geliştirilmiştir. Sonuçlar ise Mean Absolute Error (MAE), F1 gibi metriklerle gösterilmiştir. Cut Fiarni, Herastia Maharani, Nathania Calista yayınladığı bir makalede [21], Kosinüs benzerliği ile içerik tabanlı işbirlikçi filtreleme ve RFM analizinin yer aldığı müşteri önerimi için işbirlikçi filtreleme kullanılarak iki farklı öneri sistemi geliştirilmiştir. Sistemin test aşamasında ürün benzerliği ve müşteri önerimi için Precision, Recall, Accuracy gibi metrikleri ile değerlendirilmiştir. Bu makalede kullanılan yöntemlere benzer şekilde Begüm Uyanık ve Günce Keziban Orman’ın 2023 yılında yayınladığı makalede [22] turizm müşterilerinin otel rezervasyonları için ihtiyaçlarına uygun özelliklere sahip otelleri çevrimiçi olarak seçmelerine yardımcı olmak için yeni bir öneri sistemi önermektedir. Müşterinin davranışı RFM ile analiz edilip Kosinüs benzerliği ile işbirlikçi filtrelemenin bulunduğu öneri sistemi tasarlanmıştır. Ayrıca müşteri bilgisinin analizi kullanılarak öneri sistemi tasarlanmasının başarılı olacağını belirtmişlerdir. Y. S. Cho, S. C. Moon, S. C. Noh ve K. H. Ryu, 2012’de yayınladıkları makalede [23] RFM analizi ve K-Means algoritması ile müşteri segmentasyonu sonucu işbirlikçi filtreleme ile öneri sistemi geliştirmişlerdir. Müşteri segmentlerine göre önerim ile daha kişiselleştirilebilir önerilerin sunulmasını belirtmişlerdir. Sonuçlar F1 skoru gibi değerlendirme metrikleriyle gösterilmiştir.

2018 yılında yayınlanan bir makalede [24] kullanıcı profillerini ve ürün özelliklerini modellemek amacıyla matris çözümlemesi temelli bir yöntem önerilmiştir. Singular Value Decomposition (SVD) tabanlı öneri sistemlerindeki veri seyrekliği sorununu çözmek için bu çalışmada önerilen yöntem olan imputasyon tabanlı SVD (ISVD) algoritmasında, Pearson Correlation Coefficient (PCC) ile benzer müşterileri kullanarak müşteri tarafından yapılan ürün derecelendirmeleri kolonuna veri imputasyonu yapılmıştır. Kısaca veri seyrekliği, ürün değerlendirmelerinin veya müşteri geri bildirimlerinin hiç bulunmaması durumudur ve öneri sistemlerinde büyük bir sorun olarak görülmektedir [11]. Makalede, eksik verilerin

derecelendirmeye göre tamamlanmasında SVD'nin kullanımına dayanan bu algoritma, imputasyon –diđer bir deęişle eksik derecelendirmeleri tamamlama- temelli SVD olarak adlandırılmıştır. Zeinab Sharifi, Mansoor Rezghi, Mahdi Nasiri, yayınladıkları makalede [25] SimpleSVD, NormalSVD gibi SVD algoritmaları ile eksik müşteri derecelendirmelere imputasyon uygulaması yapmışlardır ve imputasyonsuz SVD'ye göre daha başarılı bir sistem tasarlandığını savunmuşlardır. Ayrıca SVD algoritmasının k boyutuna göre RMSE, MAE değerlerini karşılaştırmışlardır.

Bir diđer çalışmada [26] Doae Mensouri ve Abdellah Azmani, şirketlerin hedeflenmiş ve öneri sistemi ile kişiselleştirilmiş pazarlama stratejileri oluşturmalarına yardımcı olmak amacıyla öneri sistemi için bir yöntem geliştirmiştir. Bu yöntem ile müşterilerin gelecekteki satın alacağı ürünlerin satın alma olasılığını belirlemek, bir sonraki satın alınan ürünü tahmin etmek ve müşteri memnuniyetini hesaplamak amacıyla RFM analizi veri kümesine uygulanmıştır. Ardından birlikte satın alınan ürünleri belirlemek ve müşteri satın alma davranışını tahmin etmek amacıyla kullanılan pazar sepeti analizi (Market Basket Analysis - MBA) için apriori algoritmasıyla birlikte sisteme uygulanmıştır. Hibrit bir yaklaşım ile daha iyi bir sistemin geliştirileceğini düşünmüşlerdir. Yöntemin değerlendirildiği aşamada ise her bir müşterinin gelecekte süreç içinde işlem gerçekleştirme olasılığını tahmin etmek için BG/NBD (Beta-Geometric/Negative Binomial Distribution) modeli kullanılmıştır. Değerlendirmede bu model bir deneyin sonuçları olan beklenen ürünün -ya da başka bir deęişle deęişkenin ortalamasını- satın alım sayısı hesaplanır. Buradaki amaç olan olasılık hesaplanmasıyla birlikte müşterilerin satın alım işlemleri incelenir, önerileri kişilerin özeline indirgemiş, başka bir deęişle kişiselleştirmiş olur [27].

Çalışmada kullanılacak veri kümesinde ürünleri değerlendiren puan bilgileri bulunmadığı durumda, ilk akla gelen imputasyon tekniđi, müşterinin ilgili ürüne verdiđi puanın o ürünün alınan sayı olarak eşleme yapmaktır. Bunun üstüne sanki veri kümesinde bilgisi zaten varmış gibi literatürde geçen PCC kullanılarak imputasyonun uygulanması hedeflenmektedir. Bunun sebebi ürün alınmasına rağmen derecelendirme bulundurmeyen ürün var olmasıdır. Öneri sistemleri hakkında yapılan literatür araştırmasına ve tanımlamalara dayanarak bu konuya temel seviyede giriş yapmak amaçlı herhangi bir özel algoritma kullanılmadan tasarlanabilen en sık satın alınan ürünleri müşterilere önermeye olanak tanıyan popülerlik tabanlı öneri sistemi geliştirilmesi kararlaştırılmıştır. Ardından müşterilerin aldıđı ürünlerle arasındaki benzerliđi başka bir deęişle vektörel ortamda o nokta ile arasındaki uzaklıđı en az olan isteđe bađlı olarak seçilen adet kadar ürünü müşteriye sunan sistem geliştirilecektir. Bu sistem içerik tabanlı öneri sistemi olarak adlandırılacak olmasının yanında benzerlik hesaplaması için Kosinüs benzerliđi, en yakın nokta ya da başka bir deęişle komşu seçmek içinse KNN algoritması kullanılacaktır. İşbirlikçi filtreleme tabanlı olarak adlandırılacak sistemde ise beş farklı SVD algoritması veri kümesi üzerinde uygulaması yapılması düşünülmektedir. Bunlar SimpleSVD, NormalSVD, MidUserSVD, MidItemSVD ve MidTotalSVD algoritmalarıdır. SVD tabanlı sistemden önce geliştirilmesi kararlaştırılan sistemlerde bahsi geçen iki imputasyon tekniđi kullanılması hedeflenirken bu sistemde, ürün sayısı ile puanlama oluşturulduktan sonra hâlâ puanlama bulunmayan ürünlere SVD algoritmalarıyla oluşturulması hedeflenmektedir.

Bahsi geçen SVD algoritmalarının ortak noktası algoritmaya verilen matrisin (örneğin ürün derecelendirmesi ve ürün matrisi) boyutu azaltılmasıdır. Tekil deđer matrisi olarak adlandırılan bu matris ile birlikte kullanıcı ve ürün özelliklerinin yer aldıđı matrisler de çıktı alınmış olunur. Sadece tekil deđer matrisi kullanılarak SimpleSVD algoritması ile ürün puanlama bilgisi tahmin edilir. NormalSVD ve MidUserSVD algoritmasında ise tekil deđer matrisi kullanıcı özelliđinin yer aldıđı matrisle nokta çarpımı yapıldıktan sonra puanlama bilgisi tahmin edilir. NormalSVD algoritmasını diđerinden ayırdığı nokta ise kullanıcı özelliđi matrisi normalize edilmesidir. MidItemSVD algoritmasında ise tekil deđer matrisi ürün özelliklerinin yer aldıđı matris ile nokta çarpımı yapılır. Son olarak MidTotalSVD algoritmasında ise tekil deđer matrisi her iki özellik matrisiyle nokta çarpımı yapılır ve puanlama tahmin edilir. Böylece bir nevi eksik puanlama bilgisine imputasyon işlemi uygulandıktan sonra ürün önerisi aşamasına geçilmesi hedeflenmektedir. Bir sonraki geliştirilmesi

hedeflenen işbirlikçi filtreleme tabanlı sistemde ise RFM analizi kullanılarak müşteri segmentasyonu sonucu öneri yapılması hedeflenmektedir. Kosinüs benzerliği bu çalışmada da yer alabilir. Çalışmadaki son olarak hibrit yaklaşım ele anılarak RFM ve apriori algoritmasıyla market sepeti analizi ile kural çıkarımı yapılarak BG/NBD modeli olasılık yardımıyla müşteriye öneri sunulacaktır. Sistemlerin başarısı literatür araştırmasında yer alan kendi metrikleriyle değerlendirilecektir.

3 METODOLOJİNİN BELİRLENMESİ

Veri bilimi iş geliştirme sürecinin, planlama dahilinde daha kontrollü yürütülmesini sağlayan KDD, CRISP-DM ve SEMMA gibi yöntemler, veri bilimi süreçlerinin düzenli ve sistemli bir şekilde yönetilmesi için hayata geçirilmiş metodolojilerdir. Öneri sistemi geliştirilirken geliştirme sürecine aşamalar halinde odaklanabilmeyi sağlayan ve günlük hayata kullanılabilirliği daha kolay olan CRISP-DM (Cross-Industry Standard Process for Data Mining) ile evrensel bir yaklaşım sergileme amaçlanacaktır [28]. CRISP-DM, bazı avantajlara sahip olmasından dolayı veri madenciliği projeleri için yaygın olarak kullanılan bir metodolojidir. Proje tamamlandığında, elde edilen sonuçlar ve geri bildirimler kullanılarak süreç yeniden başlatılabilir. Bu, sürekli iyileştirmeyi ve tekrarlanabilirliği sağlar. Sahip olduğu her aşama, projeye özgü ihtiyaçlara ve zorluklara göre şekillendirilebilir. CRISP-DM metodolojisi tekrarlanabilir altı aşamadan oluşmaktadır. Bu yüzden öneri sistemiyle ürün tavsiyesi için CRISP-DM aşamaları sırasıyla altı başlıkta incelenecektir. Ek olarak burada, bir sonraki bölümde CRISP-DM metodolojisi ile geliştirilecek iki öneri sistemi için kısaca bağdaştırma ile giriş yapılmış planlanmaktadır.

3.1 İş Anlayışı (Business Understanding)

Öneri sistemi geliştirme sürecinde, ilk aşamada problem tanınır, amaçlar ve hedefler belirlenir. Literatür araştırması yapıldığında, kişiselleştirilmiş ürün tavsiyeleri sunmak için kullanılacak algoritma ve teknikler belirlenmiş ve uygulanmaya karar verilmiştir. Kullanıcının satın alma geçmişi ve işlem geçmişi analiz edilerek RFM analizi yapılması planlanmaktadır. Bu analiz son iki öneri sisteminde yer alacaktır. İlkinde müşteri segmentasyonu yapılarak müşteri segmentlerine göre ürün önerimi yapılacakken ikincisinde bu analiz sonucunda üretilen bazı değerler kullanılacak başka modele girdi bilgisinin olması beklenmektedir. Belirlenen algoritma ve teknikler uygulanarak "Bir müşteri için en iyi ürün hangisidir?" sorusuna cevap verilmesi hedeflenmektedir. Böylece CRISP-DM metodolojisinin bu aşamasında tasarlanacak uygulama veya sistem için amaç anlatılır, genel çerçevede hedefler bahsedilir ve oluşabilecek veya karşılaşılabilecek sorunlar öngörülürse belirtilir.

3.2 Veri Anlayışı (Data Understanding)

İkinci aşamada kullanılabilir veri kaynakları keşfedilir ve incelenir. Karar verilen veri kümesi proje için uygun ise proje geliştirilmesi devam eder. Araştırmada işlenmek üzere veri kümeleri incelenmiş ve Kaggle.com'daki perakende satış veri kümesinde [29] karar verilmiştir. Kaggle tarafından hesaplanan kullanılabilirlik skorunda 10/10 derecelendirilmesi bulunmakta ve veri kümesinin sağlam bir temele sahip olduğunu, eksiksiz ve güvenilir olduğunu kanıtlamaktadır. Ayrıca veri kümesinin uyumluluğu, lisansı, güncelleme sıklığı ve dosya formatı gibi değerlendirmeler göz önünde bulundurulmuş olup veri kümesinin nasıl kullanılacağı ve işlenebileceği hakkında ek bilgiler sağlamıştır. Değerlendirmeler sonucunda öneri sistemi ile ürün tavsiyesi araştırmasında bahsi geçen veri kümesinin kullanımı kararlaştırılmıştır. Böylece bu aşamada birçok veri kümesinin yer aldığı veri tabanları taranır ve uygulanması kararlaştırılan veri kaynakları üzerinde çalışmalar yapıldıktan sonra neden uygulanacağı gibi alınan kararlar ve düşünceler aktarılır.

3.3 Veri Hazırlığı (Data Preparation)

Bu aşamada veri kümesi, ön işleme süreçlerinden geçirilir. Verinin hazırlanması ve işin gereksinimlerini anlama sürecinden gelen verilerin uygun bir şekilde işlenmesine başlanılır. Araştırma sürecinde modelleme aşamasına ihtiyaç dahilinde işlenmiş veri kümesi iletilir. Veri inceleme sürecinde veri temizliği, veri entegrasyonu ve gerekirse özellik mühendisliği gibi işlemler yapılması öngörülmektedir. Bunun yanı sıra veri kümesinde müşteri geçmişi, ürün özellikleri vb. gibi veriler değerlendirilerek kişiselleştirilmiş öneri sunma amaçlanmaktadır. Böylece bu aşamada veri kümesi, literatür araştırması sonucu uygulanması kararlaştırılan algoritmalara ve tekniklere hazır hâle getirilir. Eğer ki CRISP-DM metodolojisinin geri kalan aşamalarında veri kümesinde herhangi bir işlemin yapılması gerek görülürse bu aşamaya geri dönülebilir. Bu tarz metodolojik yaklaşımların sağladığı faydalardan biri herhangi geriye dönük bir değişiklik gerekirse genel yapı korunarak işlemlerin yapılmasıdır.

3.4 Modelleme (Modelling)

Dördüncü aşamada ön işleme yapılmış verileri ve belirlenen makine öğrenimi algoritmalarını kullanarak modeller ya da istatistiksel modeller oluşturulur. Veriler, gerekli kütüphaneler veya platformlar üzerinde işlenir ve önceden belirlenmiş bir eğitim ve test seti olarak bölünür. Eğitim verileri kullanılarak sistem eğitilir ve test aşamasında ise algoritmanın daha önce hiç görmediği veriler üzerinde test yapılır. Öneri sistemi modeli inşa edilirken geliştirilmek istenen sistem modeli işlem tarihine, benzer kullanıcı tercihlerine veya ürün özelliklerine dayalı olacağı öngörülmektedir. Literatür araştırması yapıldıktan sonra kullanılması kararlaştırılan algoritmaların bu aşamada tasarlanması ve veri kümesine uygulanması bu aşamada gerçekleşir. Herhangi bir veri işleme adımı gerekirse bir önceki adıma geri dönülerek eksik tamamlanır ve tekrar modelleme aşamasına dönülebilir. Böylece modelin çıktı alınca geçiş süreci, bu aşamada idame ettirilir.

3.5 Değerlendirme (Evolution)

Modelin iş sürecine entegre edilmeden önce, modelleme aşamasında oluşturulan modelin yorumlanması bu aşamada gerçekleştirilir. Yapılan değerlendirmeler olası hatalara neden oluyorsa başa dönüp süreçler tekrar işletilebilir. Oluşturulacak öneri sisteminin performansının değerlendirilmesi amaçlanmaktadır. Aynı problem ve algoritmalar için bile farklı metrikler yorumlama sürecinde kullanılabilir. Bundan dolayı farklı algoritmaların kullanılmasıyla birlikte, değerlendirme süreci de değişiklik gösterebilir ve her öneri sistemi türü için farklı bir yorumlama yöntemi uygulanması öngörülmektedir. Müşterilere yönelik deneyler yaparak önerilerin doğruluğunu ve etkinliğini ölçmek hedeflenmektedir. Modelleme adımında kullanılan algoritmaların kendi metrikleriyle değerlendirilmesine olanak tanımak amaçlı analizler yapılması ve daha sonrasında çıkarımlar yapılarak bunların desteklenmesi iyi bir değerlendirme olabilir. Dolayısıyla çıkan sonuçların birbiriyle karşılaştırılabilecek ortamın oluşturulması hazırlanan çalışma için de iyi bir yöntem olarak karşılanabilir.

3.6 Canlıya alma (Deployment)

Son aşamada, değerlendirilmiş ve amaçlarına uygun olduğu doğrulanan sistem modeli iş süreçlerine entegre edilir. Öneri sisteminin geliştirilmesiyle, kullanıcıya en iyi ürünleri sunmayı hedeflenirken, ürün-müşteri ilişkisine sahip perakende satış yapan noktaların ihtiyaçlarının karşılanmasında kullanılması düşünülmektedir. Kısaca oluşturulan sistem, geliştirme ve test ekiplerince denenir ve sistemi kullanılacak ekibin hata ile karşılaşmaması için kullanılabilir hale getirilir. Ardından sistem, belirli ekibin sisteme her zaman ulaşabilmesi için altyapı ve ortam hakkındaki teknik detaylar dikkate alınarak bulut sisteme yüklenir. Böylece sistem hizmete sunulur. Anlatılanın yanında bu çalışmada geliştirilmesi hedeflenen sistemlerden sadece son iki sistemde CRISP-DM metodolojisinin uygulanması hedeflenmektedir.

4 ÖNERİ SİSTEMİ GELİŞTİRİLMESİ

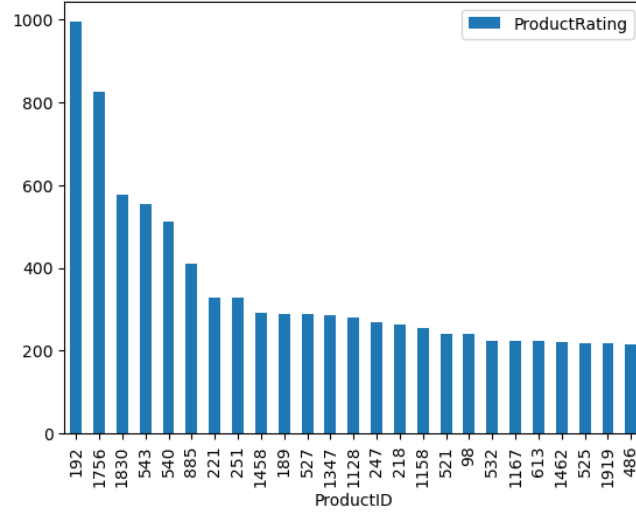
Literatür araştırmasında bahsedilen ve veri kümesinde başarılı olabileceđi düşünölen ve kendi içinde birçok algoritma ve teknik barındıran ayrı iki yöntemin CRISP-DM metodolojisi ışığında aşamalarının detayları verilere anlatılması hedeflenmektedir. Geri kalan yöntem veya algoritmalar tek bir başlık altında incelenecektir. Makalenin geliştirilen sistemi canlıya alma gibi bir hedefi olmadığı için CRISP-DM metodolojisindeki Deployment aşamasında üzerinde durulmayacaktır. Herhangi bir gereksinim olmasına karşın kısaca bir önceki bölümde bahsedilmiştir. Ürün tavsiyeleri yapmak ve öneri sistemini geliştirmek amacıyla Kaggle platformunda yer alan “Retail Data Set” [29] isimli perakende satış veri kümesi, önceki bölümde bahsedildiđi gibi geliştirme süreci için uygun görölmüştür. 8 kolon bulunan veri kümesinde kolon isimleri, kolonların veri türleri ve ne anlama geldiđinin kısaca açıklaması Tablo 1’de belirtilmiştir. Algoritma ve yöntemlerin kendi kısımlarında gerekli bilgi üretimi ihtiyacı geliřirse detaylıca açıklanacaktır.

Tablo 1: Retail Veri Kümesi Özellikleri

Kolon İsmi	Açıklam	Veri Türü
Unnamed: 0	Dizin numarası	int64
InvoiceID	Benzersiz işlem kimliđi	int64
Date	İşlemin gerçekleştiđi tarih	object
ProductID	Benzersiz ürün kimliđi	int64
TotalSales	İşlemden elde edilen gelir	float64
Discount	İşlemden uygulanan indirim	float64
CustomerID	Benzersiz müşteri kimliđi	int64
Quantity	İşlem gören ürün sayısı	int64

4.1 Popülerlik Bazlı Öneri Sistemi

Veri kümesindeki derecelendirme bilgisinin eksikliđi bu yöntem için türetilme gereksinimi duyulmuştur. Ürünü satın alan müşterinin derecelendirmesi bir müşterinin aldığı ürünün sayısı ile türetilir. Fakat ilk bilgi için bu yöntem kullanılıp sonra literatür araştırmasında bahsedilen imputasyon tekniđiyle derecelendirmeler düzenlenmiştir. Bu teknik, her müşterinin diđer müşteriler için Pearson benzerliđi hesaplanıp o ürün için ortalama benzerlikten fazla olanlar ile korelasyon matrisi oluşturulmuştur. Sonra ilk derecelendirme için kullanılmış müşterinin aldığı ürün sayısının sıralanmış matrisi ile derecelendirme oluşturulmuştur. Bu derecelendirmeler kusurlu durumdayken 0.5’lik dilimlerle sınıflandırılmış ve derece bilgisi düzenlenmiştir.



Şekil 1: Popülerlik tabanlı sistemin önerdiği ürünler

Tablo 1’de yer alan kolonlardan CustomerID, ProductID ve imputasyon tekniğiyle oluşturulan ProductRating kolonu kullanılıp Şekil 1’de gösterilen en popüler veya başka bir değişle en çok derecelendirmeye sahip olan ürünler müşterilere öneri olarak sunulabilir. Bu sistemin önerdiği ürünler, önerildiği kitle tarafından bir karşılığı bulunması muhtemel olduğu düşündürmesinin yanı sıra, müşteri özelinde öneri yapılmaması sistemin daha detaylı ürün önerisi için tercih edilmesini pek olası kılmamaktadır.

4.2 KNN ve İçerik Tabanlı Filtreleme ile Öneri Sistemi

Bir önceki yöntemde olduğu gibi derecelendirme bilgisi bu yöntemde de oluşturulması gerekmiş ve aynı yöntem ile üretilmiştir. K değeri, her bir müşteri için kaç komşunun dikkate alınacağını belirtir ve 50 değerinin kullanılması denemeler sonucu kararlaştırılmıştır. Tablo 1’deki kolonlardan CustomerID, ProductID ve imputasyon tekniğiyle oluşturulan ProductRatings kolonu KNN algoritmasına verilmiştir. Tablo 2’de bazı müşterilere önerilen ürünler önerilmiştir.

Tablo 2: Bazı müşterilere önerilen 5 ürün

Müşteri Kimliği	Önerilen Ürünler
576	[448, 420, 270, 448, 230]
664	[227, 448, 301, 59, 413]

İçerik tabanlı filtreleme, müşterilerin geçmiş tercihleri ve ürün özellikleri gibi içerik bilgilerini kullanarak önerilerde bulunur. Genel anlamda bu aşamadan sonra KNN algoritması kullanılarak içerik tabanlı bir öneri sistemi oluşturulup kullanıcının önceden beğendiği öğeleri temel alarak, benzer özelliklere sahip diğer ürünleri önerildi. Kosinüs benzerliği hesaplama metodu kullanılarak bu benzerlik değerleri hesaplanmıştır. Üstelik, müşterilerin kendisine önerilen ürünleri nasıl değerlendirecekleri ile alakalı denemeler yapmak için, KNN algoritmasının hiperparametrelerini iyi seçmek ve model performansı iyi olması amacıyla GridSearch uygulanmıştır. Buradaki hiperparametreler, benzerlik için Kosinüs hesaplamasının mı yoksa ortalama kare farkının mı kullanılması, kullanıcı temelli mi yoksa ürün temelli analiz mi gibi

seçenekleri içerir. Üç katlamalı çapraz doğrulama yapıldıktan sonra ortalama kare farkı ile benzerlik hesaplanması, algoritmanın kullanıcı temelli olması en iyi kombinasyon olarak ortaya çıkmıştır. Algoritma başarısı test edilebilmesi için veri kümesi denemeler sonucu %30 test %70 eğitim olarak algoritmaya veri kümesi verilmeden önce rastgele bir şekilde bölünmesi uygun görülmüştür. Böylece sistem ile müşteriye ürün önerildi ve müşterinin önerilen ürünü beğenip beğenmediğinin bilgisini tutan ProductRatings değerinin tahmini yapılmıştır. Bu tahminlerin başarısı Tablo 3'teki değerlendirme metrikleriyle hesaplanmıştır.

Tablo 3: Değerlendirme metrikleri

RMSE	MAE	Accuracy	Precision	Recall	Fallout	Miss Rate	F1 Score
0.238	0.448	0.965	0.965	0.965	0.034	0.034	0.965

Bahsi geçen değerlendirme metrikleri kullanılan algoritma ve yöntemlere uygun açıklanması gerekirse Precision, pozitif olarak tahmin edilen örneklerin gerçekten pozitif olan örneklerin oranını ifade eder. Bu metrik kullanıcının alacağı ürünlerin doğru tahmin sayısının, müşterinin aldığı ürünlere bölünmesi şeklinde formülleştirilir. Recall, gerçekten pozitif olan örneklerin tam olarak kaçını yakalandığını gösterir. Önerilen ürünlerin alındığı sayısına alınan ürün sayısını bölünmesiyle formülleştirilir. Fallout (False Alarm Rate, False Positive Rate) ise gerçekten negatif olan örnekler arasında yanlış pozitiflerin oranını ifade eder. Önerilmeyip gerçekte de alınmayan ürünlerin sayısına alınmayan ürünlerin sayısı bölünür. Miss Rate (False Negative Rate), gerçekten pozitif olan örnekler arasında yanlış negatiflerin oranını ifade eder. Önerilmeyip de alınan ürünlerin sayısına alınan ürünlerin sayısı bölünerek formülleştirilir. F1 Skoru, hassasiyet ve duyarlılık arasında bir denge sağlayan bir ölçüttür. Bu metrik, doğru pozitifleri (TP) ve yanlış negatifleri (FN) dikkate alarak Precision ve Recall kullanılır. İki kere Precision ve Recall değerlerinin çarpımına Precision ve Recall toplamı bölünerek formülize edilir. Root Mean Square Error (RMSE), müşteri derecelendirmesinin tahmini gerçek değerlerden ne kadar uzak olduğunu ölçen bir metriktir. Aynı şekilde Mean Absolute Error (MAE) ise tahmin edilen müşteri derecelendirmesinin tahmini ile gerçek değerler arasındaki mutlak farkların ortalamasını ifade eder. Son olarak Accuracy (doğruluk), öneri sisteminin doğru tahminlerinin toplam tahminlere oranını ifade eder.

Değerler incelendiğinde, imputasyon tekniği sayesinde ürün önerimi sisteminin başarısı göz ardı edilemez. Ancak, bu yöntemin temeli belirli istatistiksel hesaplamalara dayandığı için, gerçek verilerle sistem çalıştırıldığı zaman benzer sonuçlar elde etmeyi garanti etmediğini unutmamak önemlidir.

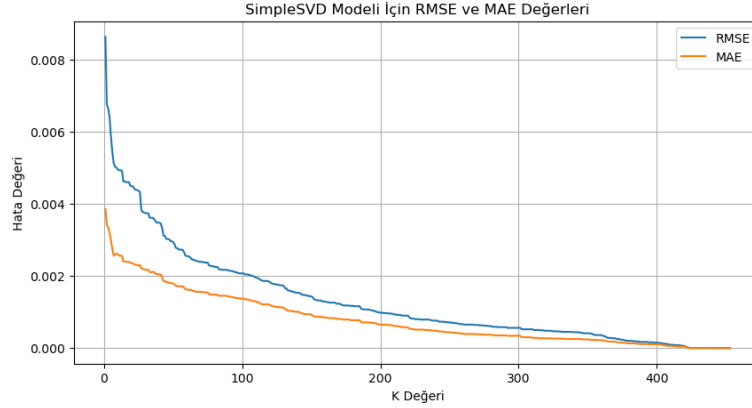
4.3 SimpleSVD, NormalSVD, MidItemSVD, MidUserSVD, MidTotalSVD Tabanlı İşbirlikçi Filtreleme ile Öneri Sistemi Geliştirilmesi

Ürün bazlı müşteri değerlendirme bilgisi veri kümesinde bulunmadığı için ilk iki yöntemde kullanılan imputasyon tekniği bu yöntemde de kullanılmıştır. Fakat iki yöntemden farklı olarak, literatür araştırmasında bahsedilen birtakım SVD algoritmaları ile hâlâ değerlendirme bulunmayan hücreler düzenlenmiştir. Bu imputasyon tekniğinden sonra ürün önerimi gerçekleşmiştir. Bahsi geçen eksik derecelendirmelerin bulunmasından dolayı SVD algoritmasının kullanılması ile veri seyrekliği sorununu azaltma amaçlı bir adım atılmış olunur. Algoritmaların başarısı test edilebilmesi için veri kümesi denemeler sonucu %30 test %70 eğitim olarak algoritmalara veri kümesi verilmeden önce rastgele bir şekilde bölünmesi uygun görülmüştür. Tablo 4'te literatür araştırmasında bahsi geçen çeşitli SVD algoritmalarının başarı değerlendirmeleri için RMSE ve MAE değerleri gösterilmektedir. Ayrıca SVD çıktıları algoritmaya verilen k değerine göre küçük değişimler gözlemlenmektedir. NormalSVD, MidItemSVD ve MidUserSVD algoritmalarındaki metriklerin neredeyse aynı çıkması algoritmalara parametre geçilen değerlerin aynı olmasından dolayı olduğu düşünülmektedir.

Tablo 4: Değerlendirme metrikleri

k değeri	SimpleSVD	NormalSVD	MidItemSVD	MidUserSVD	MidTotalSVD
	RMSE - MAE	RMSE-MAE	RMSE - MAE	RMSE - MAE	RMSE - MAE
10	0.049 – 0.002	11.66 – 3.69	11.66 – 3.69	11.66 – 3.69	1587.2 – 508.1
49	0.002 – 0.001	11.66 – 3.69	11.66 – 3.69	11.66 – 3.69	1587.2 – 508.1
189	0.001 – 0.0007	11.66 – 3.69	11.66 – 3.69	11.66 – 3.69	1587.2 – 508.2
1267	0.0006 – 0.0003	11.66 – 3.69	11.66 – 3.69	11.66 – 3.69	1587.2 – 508.2

SimpleSVD algoritması için değerler, modelin verileri çok iyi tahmin ettiğini ve hatta tam olarak eşleştirdiğini gösteriyor gibi görünüyor. Ancak, bu sonuçlar aşırı uyma (overfitting) ihtimalini düşündürülebilir. Veriye çok iyi uyan bir modelin, genelde yeni veriler üzerinde daha kötü performans gösterme eğilimi de olabilir. Bu yüzden diğer algoritmalarından NormalSVD, MidUserSVD ve MidItemSVD'nin değerleri, tahminlerin genel olarak kabul edilebilir derecede iyi olduğunu gösteriyor. Bu sonuçlar, modelin veriyi uygun şekilde genelleştirdiği ve aşırı uymadan kaçınmış olabilir. MidTotalSVD algoritmasının değerleri, bu modelin diğer modellere göre çok daha kötü performansı olduğunu gösteriyor. Bu durumda, modelin veriyi doğru bir şekilde yakalayamadığı veya aşırı uyma sorunu olduğu görülebilir. SimpleSVD için k boyutuna göre değişen RMSE ve MAE değerleri ise Şekil 2'de gözlemlenmektedir.



Şekil 2: SimpleSVD algoritmasının RMSE ve MAE değişim grafiği

Birtakım SVD algoritmaları ile yapılan denemeler sonucu ürün önerisi için SimpleSVD ve MidUserSVD karşılaştırılmıştır. CustomerID'si 20 olan müşteriye önerilen 5 ürün ile müşteri derecelendirme tahmini Tablo 5'te gösterilmektedir. En iyi k değeri SimpleSVD için 10, MidUserSVD için 49 belirlenmiştir. Değerlendirme bilgisi 0 ile 5 arasında tutulması kararlaştırıldığı için değerlendirme tahmini bu aralıkta olması beklenmektedir.

Tablo 5: SimpleSVD ile MidUserSVD için Önerilen Ürünler ve Müşteri Değerlendirme Tahmini

SimpleSVD - Önerilen Ürün	SimpleSVD - Rating	MidUserSVD - Önerilen Ürün	MidUserSVD - Rating
1894	1.27	357	4.72
218	1.2	94	4.69
221	1.15	1549	4.35

İlk olarak bir müşterinin aldığı ürün sayısı kullanılarak müşteri derecelendirme bilgisinin oluşturulmasından sonra hâlâ bu bilginin eksik olan müşterilere hangi ürünü alabileceğini ve nasıl değerlendireceği üzerinden öneri sistemi sonuçları yer

almaktadır. KNN ile tasarlanan içerik tabanlı öneri sistemi geliştirilmesinde de bahsedildiđi gibi imputasyon tekniklerinin temeli istatistiksel hesaplamalara dayandıđı için, gerçek verilerle sistem çalıştırıldıđı zaman benzer sonuçlar elde etmeyi garanti etmediđini unutmamak önemlidir.

4.4 RFM Analizi ve Kosinüs Benzerliđi ile Kullanıcı Tabanlı İşbirlikçi Filtreleme

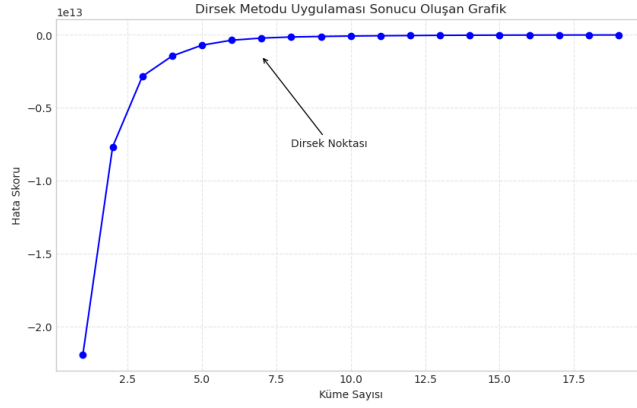
Bir önceki bölümde bahsi geçen iş süreci yürütme yöntemi olan CRISP-DM'in temel alınması ile bu bölümde, yapılan literatür araştırması ışığında tanımlanan öneri sisteminin geliştirilmesi amaçlanmaktadır. Dolayısıyla her bölüme göre geliştirme süreci anlatılarak herhangi bir aşamaya geri dönülebilir yol izlenecektir. Süreç boyunca kullanılacak algoritmalar Python dilinde uygulanması planlanmaktadır.

4.4.1 İş Anlayışı

Öneri sistemleri alanında yapılan literatür arařtırmaları, etkili bir ürün öneri sisteminin geliştirilmesinin, kullanıcı-ürün ilişkisini detaylı bir şekilde analiz edilmesine dayandıđını ortaya koymaktadır. Öneri sistemi kullanıcı tabanlı işbirlikçi filtreleme yapılarak geliştirilmesi kararlařtırılmıřtır. Böylece RFM analizi ile müşteri segmentasyonu yapılarak kullanıcılar arası benzerlik hesaplanmasında yardımcı olacađı düşünölmektedir. Kullanıcı ile ürün ilişkisinin analizi yapılarak kullanıcıların geçmiş satın alma davranışları ve işlem geçmişinin kullanılmasıyla müşteri davranışlarını anlamayı hedeflemektedir. Bu amaçla kullanılan ve çalışmada da yer alacak yöntemlerden biri olan RFM analizidir. RFM analizi, Recency (yenilik), Frequency (sıklık) ve Monetary (parasal deđer) kelimelerinin baş harfleri ile temsil edilen yöntem, müşterileri son satın alım tarihine, satın alım sıklığına ve satın alım sayısına göre gruplara ayırmayı amaçlamaktadır. Bu gruplandırma sayesinde müşteri davranışları alınan ürünler baz alınarak analiz edilebilir. Bunun üzerine öneri sistemi kurulduktan sonra müşteri bazlı öneri doğruluđu incelendikten sonra genel anlamda önerilen ürünlerin test kümesinde varlığı üzerinden deđerlendirmeler literatür arařtırmasında bahsi geçen metriklerle ele alınacaktır.

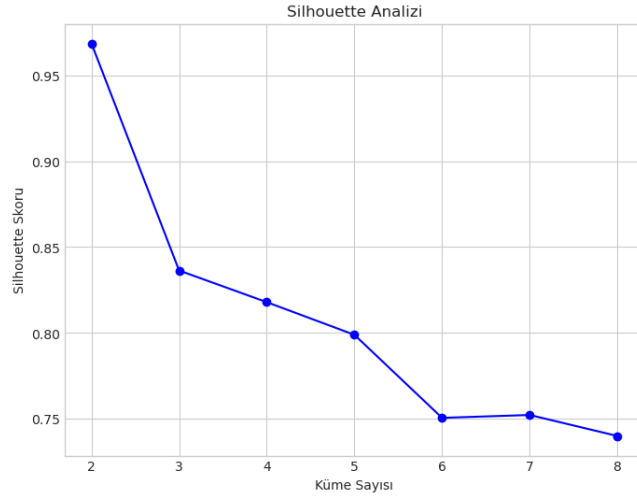
4.4.2 Veri Anlayışı

Ürün tavsiyesi yapmak için müşteri segmentasyonu yapılmıřtır. Bunun nedeni öneri sisteminde ürünleri, müşteriler ve onların aldıkları ürünlerin benzerliđi ile tavsiye etmek hedef olarak belirlenmiřtir. Kolonlardan ProductID, CustomerID ve InvoiceID kullanılarak veri kümesinde yer alan müşterilerin geçmişe dayalı verilerinden müşteriler kümelere ayrılmıřtır. Recency, Frequency ve Monetary skorları hesaplandıktan sonra kümeler isimlendirilmiřtir. Kümeleme yöntemi olarak K-Means kullanılmıř olup kümeleme hatalarını en aza indirmek ve optimal küme sayısını belirlemek için dirsek yöntemi ve silhouette skoru kullanılmıřtır.



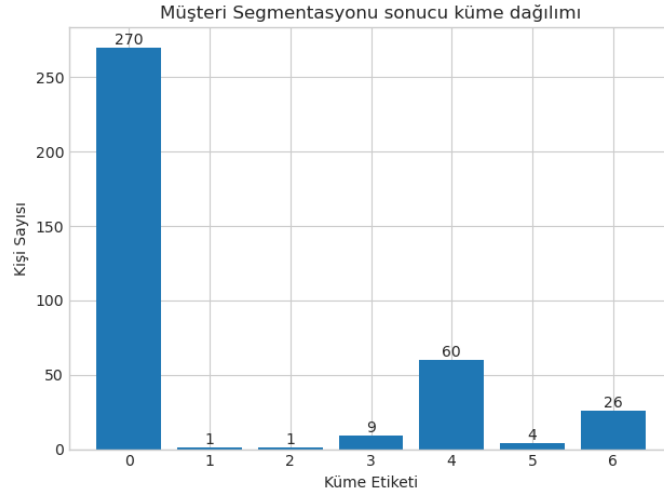
Şekil 3: Dirsek Yöntemi ile optimal küme sayısı belirlenmesi

Şekil 3'te bulunan değer gösterildiği gibi makul küme sayısı belirlenmek amacıyla bir metrik ile sağlama ihtiyacının olduğu görülmüştür. İkinci bir değerlendirme metriği olarak da Silhouette skoru hesaplanmış ve küme sayısı artarken Silhouette skoru da azalmıştır. Bir önceki yöntem ile örtüşmesi sonucu küme sayısı 7 olarak karar verilmiştir. Oluşan grafik Şekil 4'te sunulmuştur.



Şekil 4: Silhouette skoru ile optimal küme sayısı belirlenmesi

Silhouette skorunu kullanarak optimal küme sayısını belirlerken, yüksek skora sahip olan kümeleme sayısı genellikle tercih edilir. Ancak, segmentasyon analizinde kullanılan veri setine, işletme hedeflerine ve diğer değerlendirme yöntemlerine de bağlı olarak optimal küme sayısı değişebilir. Dirsek yönteminden çıkan sonuç ve Şekil 4'te sunulan grafikten yola çıkılarak optimal küme sayısının yedi olmasına karar verilmiştir. Ayrıca müşteri segmentasyonu yedi küme için müşteri sayısına göre grafik Şekil 5'te gösterildiği gibidir.



Şekil 5: RFM analizinde müşteri segmentasyonu sonucu oluşan küme dağılımı

Örneğin kümesi 0,4 ve 6 olan müşterilerin Recency, Frequency, Monetary değerleri gösterilen Tablo 6'da cluster kolonu veya başka bir deyişle kümesi sıfır olan müşteriler uzun bir süredir gelmemiş görünüyor. Bu nedenle, onları tekrar kazanmak ve dikkatlerini çekmek için indirim gibi teşvikler sağlanabilir.

Tablo 6: Kümesi 0, 4, 6 Olanların Ortalama Recency, Frequency, Monetary Değerleri

Recency	Frequency	Monetary	Cluster
518,56	6,67	9844,41	0
115,38	65,03	100675,75	4
84,38	130,65	271096,83	6

Küme 4'teki kısım orta seviye müşterilere işaret ediyor gibi görünüyor çünkü daha düşük sıklıkta alışveriş yapıyorlar ve ortalama harcama miktarları da kümesi altı olan müşterilere göre bu nedenle daha düşük kalıyor. Onları daha sık alışveriş yapmaları için teşvik etmek önemli olabilir. Küme altı en gözde müşteriler olarak tanımlanabilir. Yüksek miktarda para harcıyorlar ve Recency değerleri de oldukça iyi ve bu nedenle bu müşteriler kaybedilmemeli, onları elde tutmak için özel teklifler veya hizmetler sunulabilir.

4.4.3 Veri Hazırlığı

Veri kümesinde olağandışılık söz konusu olup olmadığı incelenmiş, süreçte kullanılması düşünülen kolonlar belirlenmiş ve birkaç değişiklik gerektiği görülmüştür. Pandas kütüphanesi tarafından sağlanan DataFrame sınıfı, verileri tutmak için bir index (dizin) sütunu içerir. Varsayılan olarak, DataFrame zaten bir index sütunuyla birlikte gelir. Bu index sütunu, her satıra benzersiz bir kimlik sağlamak için kullanılır ve DataFrame'in satırlarını etiketlemek ve sorgulamak için kullanılabilir. Bu yüzden dizin bilgisi halihazırda tutulduğu için ayrıca veri kümesinde bulunması gereksiz olduğundan Unnamed:0 kolonunun silinmesine karar verilmiştir. Ayrıca DataFrame'lerde tarih ve saat bilgisini temsil etmek için genellikle datetime türünü kullanır. Bunun nedeni, datetime türünün, tarih ve saat aritmetiği yapmak, tarihler üzerinde filtreleme yapmak ve zaman serileri analizi gibi işlemleri gerçekleştirmek için daha fazla işlevselliğe sahip olmasıdır. Belirlenen ölçütler ve proje amacı zaman serisi analizine pek değinmese de object türünde bulunan Date kolonunun veri türü datetime olarak değiştirilmiştir.

Müşteri segmentasyonu yapmak için RFM analizi uygulanmıştır. CustomerID'lere göre Date kolonları gruplandırılarak müşterinin son satın alım işlem tarihi için LastPurchaseDate kolonu oluşturulmuş ve CustomerID ile Recency skoru hesaplanmıştır. Bir sonraki değer kolonu olan Frequency için CustomerID'lere göre InvoiceID kolonu gruplandırılarak Frequency ile hesaplama yapıp veri kümesi işlenmiştir. Son olarak Monetary değer kolonunu hesaplamak için TotalSales kolonu CustomerID'lere göre gruplandırılıp toplanmıştır. Böylece modelleme aşamasına aktarılmak üzere hazırlanan veri kümesi son olarak K-Means algoritmasıyla kümeleme yapılarak Cluster kolonu oluşturulmuştur. Yeni oluşturulan Cluster kolonu ile InvoiceID, ProductID ve CustomerID dışındaki diğer kolonlar araştırmanın devamında işe yaramayacağı için silinmiştir.

4.4.4 Modelleme

Literatür araştırmasında bahsedilen uygulanması öngörülen ve iş anlayışı aşamasında uygulanması kararlaştırılan algoritmalar ve yöntemler modelleme aşamasında işleme tabii tutulacaktır. Veri kümesinde bu aşamaya kadar sadece RFM analizi ile müşteri segmentasyonu yapıp kümeye uygulanmıştır. Değerlendirme aşamasında oluşturulan modeli değerlendirmek amaçlı veri kümesinin 2/3'lük kısmı eğitim, 1/3'lük kısmı test olarak ayrıldı. Bununla birlikte veri kümesi geçmişten 2023-03-25 tarihine kadar sıralı bir şekilde bulunmaktadır. Böyle bir bölümlendirmede eğitim kümesi 2019/01/02-2021/01/25 tarihler arası denk gelmiş, test kümesi 2021/01/25-2023-03-25 tarihleri arası denk gelmiştir. Tahmin edileceği gibi müşteri kaybı bu yöntemde müşteri kaybı olabileceğidir. Başka bir deęişle train kümesinde yer alan müşterilerden bazıları test kümesinde yer almayabilir. Bunun tespiti için bir örnek vermek gerekirse Şekil 5'te kendisi ile Kosinüs benzerliği 1 olması beklenen CustomerID'si 10 olan müşterinin 9. indekste bulunması gerekirken 4. indekste bulunuyor. Değerlendirme yapmak için veri kümesini bölmemizden dolayı CustomerID'si 0, 3, 5, 7, 8 olan müşteriler eğitim kümemize denk gelmemiştir. Bu yüzden CustomerID'si 10 olan müşteri 4. indekste bulunuyor. Bunun sonucu olarak birçok müşteri bu şekilde test kümesine denk gelmediğinden yer almayacaklardır.

Benzerlik hesaplanması için müşteri bazlı filtreleme yapılmıştır. Başka bir deęişle, müşterinin aldığı ürünleri CustomerID ile gruplanarak yeni bir Dataframe'e eşlendi. Böylece CustomerID değeri ile o müşterinin müşteri matrisiyle diğer müşteri matrislerinin benzerlikleri döngüye alınarak tüm müşteriler için benzerlikler hesaplandı. Şekil 7'de CustomerID'si 10 olan müşterinin InvoiceID, CustomerID, ProductID ve cluster kolonlarıyla diğer kullanıcılar ile arasındaki benzerliği göstermektedir.

0.37989451	0.34049415	0.27580451	0.32970506	0.32970506	1.
0.64203932	0.16648815	0.13584646	0.60246378	0.15830603	0.29051442
0.17058609	0.59438211	0.45106497	0.59938732	0.21845412	0.35465831
0.52640233	0.2159947	0.74144242	0.62224665	0.65254985	0.32970506
0.30274867	0.59536583	0.6832603	0.32970506	0.32970506	0.32970506
0.32970506	0.32970506	0.32970506	0.06460176	0.16495207	0.46397446
0.32970506	0.40771326	0.35774217	0.56552711	0.93767304	0.03145106
0.32970506	0.15244143	0.68400311	0.61278208	0.22240525	0.12966975
⋮	⋮	⋮	⋮	⋮	⋮
0.19486555	0.53028717	0.32970506	0.00854643	0.25302326	0.32970506
0.16775086	0.68323489	0.32970506	0.32970506	0.44932967	0.53175956
0.32970506	0.83259069	0.32970506	0.67791332	0.32970506	0.08794601
0.32970506	0.53408345	0.14841566	0.21183881	0.32970506	0.35468924
0.90626507	0.53028717	0.21393862	0.25942842	0.32970506	0.17153925
0.53032476	0.24661407	0.3680497	0.33339727	0.08077612	0.10862351
0.32970506	0.32970506	0.53028717	0.32445369	0.32970506	0.53028717

Şekil 6: CustomerID değeri 10 olan müşterinin diğer müşterilerle arasındaki konsinüs benzerlik matrisi

Şekil 6'daki CustomerID'si 10 olan müşterinin 4. indeks'teki müşterini ile benzerliği 1'dir, başka bir deyişle tavan değeridir. Sadece bu durum için konuşulduğunda CustomerID değeri 10 olan matriste 4. indekste bulunmaktadır. Sonuç olarak benzerlik matrisi, tüm müşteriler için bir müşterinin diğer tüm müşteriler arasındaki benzerliği içeriyor. Önerilecek ürün sayısı isteğe bağlı olarak beş seçilmiş olup aldığı ürünler dışındaki en yüksek benzerliğe sahip 5 ürün önerilecek şekilde tavsiye sistemi geliştirilmiştir. Tablo 7'de altı müşteriye önerilen beş ürün gösterilmektedir.

Tablo 7: Bazı müşterilere önerilen 5 ürün

Müşteri Kimliği	Cluster	Önerilen Ürünler
120	0	[0, 1026, 4, 518, 521]
230	2	[7, 521, 524, 525, 526]
458	3	[0, 1026, 521, 1547, 524]
10	0	[1026, 7, 521, 525, 527]
501	4	[515, 521, 525, 526, 527]
181	6	[1158, 521, 268, 1678, 527]

Geliştirilen öneri sistemi aracılığıyla ürün önerisi gerçekleştirilmiştir. Burada önerilen ürün sayısı spesifik bir anlam taşımadığı göz önünde bulundurulması gerekmektedir. Bu yöntemde ve bu çalışmada isteğe bağlı olarak önerilecek ürün sayısı 5 seçilmiştir. Bu durumun en makul açıklaması, araştırmada kullanılan algoritmalar ve yöntemler ile geliştirilen sistemlerin sonuçlarının kolayca karşılaştırılabilir olmasını sağlamaktır. Bu yaklaşım ile çalışmanın sonuçlar bölümünde daha etkili bir karşılaştırma yapılabilmesi amaçlanacaktır.

4.4.5 Değerlendirme

Geliştirilen öneri sistemi ile bazı müşterilere önerilen ürünlerin analizinin sonucunda segmentlerin en çok aldığı ürünler Tablo 8'de gösterilmiştir. Dolayısıyla Tablo 8'de yer alan müşteri segmentlerinin en çok aldığı ürünlerin Tablo 7'de yer alan bazı müşterilerin yanı sıra çalışmada analizi yapılan diğer müşterilere önerilen ürünlerde baskın gelmediği görülmüştür. Fakat yapılan analizler sonucu müşterilere önerilen ürünlerle segmentlere önerilen ürünler birbirine yakın olduğu da saptanmıştır. Bu önerilen ürünlerdeki yakınlık kişiselleştirilebilir bir sistemin başarısı olduğunu gösterilebilmesinin yanı sıra, alınan ürünün sayısı ile bir ilişkinin olmadığı durum ortaya çıkmış olabilir.

Tablo 8: Bazı müşterilere önerilen 5 ürün

Cluster	Segmentlerin En Çok Aldığı Ürünler
0	[1347, 613, 1756, 1128, 885, 1167, 1724, 617, 1915, 1830]
1	[545, 1341, 559, 1830, 22, 1578, 9, 49, 192, 541]
2	[1291, 628, 900, 624, 647, 1368, 1398, 1290, 1400, 1548]
3	[1347, 1756, 192, 540, 617, 300, 1830, 885, 1724, 527]
4	[885, 613, 1347, 1167, 1128, 617, 1756, 1724, 1830, 540]
5	[613, 1830, 885, 189, 1756, 192, 540, 1128, 1347, 300]
6	[1347, 1756, 1724, 1167, 1128, 613, 885, 617, 192, 543]

Karşılaşılan durumun nedeni sistemdeki modele verilen kolonlarda Quantity bilgisinin verilmemesi olduğu düşünüldükten sonra InvoiceID, ProductID, CustomerID ve cluster kolonlarına ek olarak Quantity kolonu eklenmiştir. Sonuçta ise müşterilerin aldığı ürünlerin sayısı Quantity kolonunun eklenmemiş hâline göre modele farklılık yaratacak kadar katkı yapmadığı saptanmıştır. Bu yüzden müşteri segmentlerinin alabileceği ürünlerin bulunduğu segmentin en çok aldığı ürünler ile bir yakınlık olmadığı görülmüştür.

Makalede literatür araştırmasında bahsedilen bu yöntemde RFM analizini, analiz sonucu müşteri segmentasyonunu ve özellik çıkarımını ve kullanıcı tabanlı işbirlikçi filtreleme ile Kosinüs benzerliğini içermektedir. Tablo 9’da literatür araştırmasında bahsi geçen, öneri sistemi geliştirilmesinde kullanılan ilk yöntemin değerlendirme metrikleri gösterilmiştir.

Tablo 9: İlk yöntemin değerlendirme metrikleri

Precision	Recall	Fallout	Miss Rate	F1 Score
0.5	0.5	0.597345	0.5	0.5

Tasarlanan sistemde tavsiye edilen ürünler benzerliğe göre sıralanmış bir şekilde çıktı sunuluyor. Sistem, Tablo 7’de CustomerID’si 501 olan müşteriye 521, 525 ve 527 ürünlerinde içinde olduğu 5 ürün önermiştir ve 2021/01/25 tarihinden sonra sadece bu 3 ürünü müşteri satın almıştır. Aynı şekilde 181 CustomerID’li müşteri 1678 ve 527 ürünlerinin içinde olduğu ürünleri önermiş ve sadece bu 2 ürünü önceki örnekte bahsi geçen tarihten sonra satın almıştır. Ürün önerme sayısı 5 seçilmiş olup müşteriler 899 tanesini almamış, 46 tanesini almıştır. Sonuçta çıkan verilere dayanarak sistem alınan ürünlerin yaklaşık %4,86 kadarını doğru tahmin edebilmiştir. Tablo 9’da literatür araştırmasında bahsi geçen, öneri sistemi geliştirilmesinde kullanılan ilk yöntemin değerlendirme metrikleri gösterilmiştir.

4.5 RFM ve Market Sepeti Analizi ile Kural Çıkarımı İçin Apriori Yardımıyla Hibrit Öneri Sistemi Geliştirilmesi

Hibrit bir yaklaşım sergilenen bu yöntemde, CRISP-DM metodolojisi baz alınarak RFM analizi uygulanıp üretilen bazı çıktılar BG/NBD modeline verilmesiyle müşterilerin alabileceği ürünlerin önerimi olasılıkla desteklenecektir. Bunun yanı sıra sepet analizi uygulanarak kural çıkarımı için Apriori algoritması kullanılacaktır. Bu yöntemde uygulama dili Python uygun görülmüştür. Önceki yöntemden farklı olarak RFM analizi çıktıları ile müşteri segmentasyonu yapılmayacak, BG/NBD modeli için girdi olacaktır. Ancak süreç boyunca ihtiyaç öngörülürse müşteri segmentasyonu yapılacaktır.

4.5.1 İş Anlayışı

Bir önceki yöntemde kullanılan RFM analizi bu yöntemde de kullanılacaktır. Bunun yanı sıra perakende satış verilerini inceleyerek müşteri alışverişini anlamak ve ürün ilişkilerini keşfetmek için kullanılan Apriori algoritması veri kümesine uygulanacaktır. Bu algoritma bir veri madenciliği tekniği olan sepet analizinde kullanılır ve algoritmanın çalışmasında dikkate alınacak uygulanma şekli ise eğer bir müşteri belirli bir ürünü satın alıyorsa, o müşteri aynı zamanda belirli başka ürünleri de satın alma eğiliminde olduğudur. Bu prensip, kural çıkarımından sonra tekrar eden satın alma davranışı

serilerini modellemek ve müşteri davranışını tahmin etmek için kullanılan BG/NBD modeli kullanılacaktır. Ardından BG/NBD sonuçları Apriori algoritmasının çıktıları ile karşılaştırılacaktır.

4.5.2 Veri Anlayışı

RFM analizi uygulanmış olan veri kümesinde Tablo 1'deki kolonlara ek olarak Recency, Frequency ve Monetary kolonları üretilmiştir. Hesaplanan "Recency" değeri, müşterinin son alışverişinin ne kadar süre önce olduğunu ifade eder. BG/NBD modeli, müşterinin ne sıklıkla alışveriş yapacağını tahmin ederken bu değeri kullanacaktır. Frequency değeri müşterinin belirli bir zaman aralığında kaç defa alışveriş yaptığını belirtir. Bu da BG/NBD modelinin müşterinin gelecekteki alışveriş sıklığını tahmin etmesine yardımcı olacaktır. Aynı şekilde Monetary değeri, müşterinin toplam harcamasını ifade eder. BG/NBD modeli, müşterinin gelecekteki harcamalarını tahmin ederken bu değeri de kullanacaktır.

4.5.3 Veri Hazırlığı

Bu yöntemde bir önceki yöntemin veri hazırlığı aşamasında bahsedilen işlemlerden geçen veri kümesinin varolan hâline ek işlemler yapılacaktır. Çünkü aynı veri kümesinin kullanılmasının yanında iki yöntem sadece RFM analizi işleminden sonra birbirinden farklı yöntemler oluşturuyor. Bir önceki sistemde RFM analizinde hesaplanan Frequency, Recency ve Monetary bilgilerine ek olarak T değeri de hesaplanmış ve Tablo 10'daki durum ortaya çıkmıştır.

Tablo 10: Modele verilmek üzere işlemde geçirilecek veri kümesi

CustomerID	Recency	Frequency	Monetary	T
317	58	79	123897.3	504
185	74	169	275071.31	547

Veri kümesi RFM analizi uygulanmadan model sonuçlarını test edebilmek için 1/3 ve modeli eğitmek için 2/3 oranında ayrılmıştır. BG/NBD modeli, birkaç kolona ek olarak bir müşterinin son alışveriş tarihini (Recency) ve tahmin yapılacak olan zaman aralığını (T) kullanarak gelecekteki satın alma davranışını tahmin eder. T değeri mantıklı seçebilmek önemlidir ve bu değere test veri kümesinde yer alan işlemin en son gerçekleştiği tarih ile eğitim kümesinde yer alan en son işlem tarihi arasındaki gün farkı eşlenmiştir. Tabii ki bu değer her işlem tarihi için farklıdır. Ayrıca önemli bir noktadan bahsedilmesi gerekirse Recency değeri, T değerinden büyükse, müşteri son alışverişini tahmin yapılacak olan zaman aralığı sonrasında yapmış. Bu durum, modelin tahminlerinin gerçekçi olmasını zorlaştırabilir çünkü müşterinin satın alma davranışını tahmin etmek için son alışverişini kullanmak önemlidir. Bu yüzden Recency değeri T değerinden büyük olan satırlar veri kümesinden düşürülmüştür.

4.5.4 Modelleme

Veri hazırlığı yapıldıktan sonra BG/NBD modeline girdi olarak verilmeye hazır veri kümesi bulunuyor. Bu model lifetimes kütüphanesinden BetaGeoFitter sınıfı yardımıyla projeye dahil edilir. BetaGeoFitter, bir müşteriye yönelik gelecekteki satın alma davranışını tahmin etmek için kullanılan istatistiksel model olan BG/NBD modelini uygulamak için kullanılan bir sınıftır. BG/NBD, Beta-Geometric dağılımının Negative Binomial dağılımına bölümü olarak ifade edilir.

$$f(t|r, \alpha, \beta) = \frac{\Gamma(r+t)}{\Gamma(r)\Gamma(t+1)} \frac{\alpha^r \beta^t}{(\alpha+\beta)^{r+t}} \quad (1) \quad P(X = x|\lambda, p) = \binom{x+p-1}{x} \left(\frac{\lambda}{\lambda+p}\right)^x \left(\frac{p}{\lambda+p}\right)^p \quad (2)$$

BG dağılımı, müşterilerin satın alma aralarındaki süreleri modellemek için kullanılır. Bu dağılım, bir müşterinin ne kadar süre boyunca sadık kalabileceğini ve bir süre sonra ne zaman kaybedebileceğini gösterir. Denklem 1'de gösterilen BG formülünde: t ilk ve son satın alma arasındaki dönemi, r bir müşterinin en az bir satın alma yapmış olma sayısını, α ve

β Beta dağılım parametrelerini ve Γ , gamma fonksiyonunu temsil eder. Veri kümemizde bulunan T değeri t parametresine, Frequency değeri r parametresine eşlenecek. α ve β parametreleri ise istatistiksel optimizasyon teknikleri kullanılarak veriye uygun şekilde tahmin edilir. Böylece modelin gerçek verilere ne kadar iyi uyduğunu değerlendiren bir uyum metriđi (genellikle log-likelihood) üzerinden parametreleri optimize eder. Gamma fonksiyonu ise müşterinin BG fonksiyonundaki t dönemi (iki ardışık satın alma arasındaki süre) dağılımını modellemek için kullanılır. NBD dağılımında ise müşterinin satın alma frekansını modellemek için kullanılır. Denklem 2’de gösterilen NBD formülünde: x, belirli bir zaman aralığında yapılan satın alma sayısı iken veri kümemizdeki Frequency değerine eşlenecektir. λ parametresi ortalama satın alma sayısını belirtirken, p parametresi satın alma davranışının varyansını ifade eder. x’in ortalaması lambda parametresidir ve p ile birlikte model hesaplamasını yapar. Sonuç olarak NBD, BG dağılımına bölünmesiyle model temel anlamda oluşturulmuş olur. Overfitting eğilimini sınırlamak ve daha genelleştirilebilir sonuçlar üretmek için kullanılan penalizer_coef parametresi girdi olarak verilmiştir. Bu parametre, veri kümesi üzerinde deneme yoluyla 0,8 değerine eşlenmiştir. Model eğitilmiş ve çıktı olarak $\lambda = 0.12$, $\alpha = 0.25$, $\beta = 0.41$ ve $r = 0.58$ parametre değerleri hesaplanmıştır.

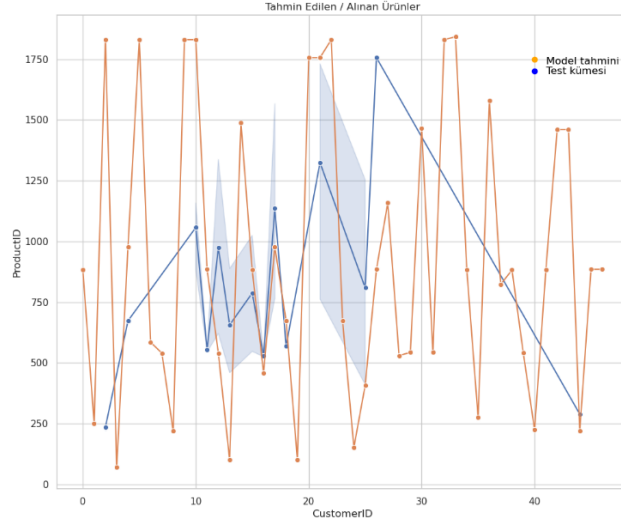
4.5.5 Deđerlendirme

İkinci yöntem veri kümesine uygulanmış ve model sonucu hesaplanan olasılıklar test veri kümesi ile karşılaştırılmıştır. Tablo 11’de model sonucu en fazla satın alma olasılığına sahip 5 CustomerID için ProductID kolonundaki ürünü alma olasılığı yüzdesel olarak PredictedPurchases kolonnda gösterilmektedir.

Tablo 11: En fazla satın alım olasılığına sahip 5 işlem

PredictedPurchases	CustomerID	ProductID
2.14	398	543
2.38	430	1830
2.44	88	544
2.55	141	243
3.02	387	1919

Tabloda gösterilmeyen diđer işlem olasılık değerleri veri kümesinde sona gittikçe sıfıra yakınıyor. Ayrıca bu değerler modelin veri kümesindeki etkisi ile Apriori algoritması ile yapılan market sepeti analizini anlamlandırmayı zorlaştırmaktadır. Ortaya çıkan duruma rağmen Şekil 7’de veri kümesindeki müşterilerin bir kısmının aldığı ürünleri ve modelin hesapladığı ürün satın alabilme olasılığına sahip ürünleri göstermektedir.



Şekil 7: Bir kısım müşterinin satın alabilme olasılığı olan ürünler üzerinde modelin tahmini

Literatür araştırmasında bahsi geçen yöntem veri kümesine uygulanırken veri hazırlığı aşamasında, veri kümesinde müşterilerin ürün için puanlamanın bulunmaması ve eğitim ile test kümelerinin arasında müşteri kaybının var olabileceği bir etken olduğu düşünülmektedir. Fakat önceki algoritmalarda da kullanılan imputasyon tekniği bu yöntemde de kullanılmıştır ve puanlama bilgisi elde edilmiştir. Ayrıca bu puanlamanın sadece modeli değerlendirme kısmında kullanılma ihtiyacı gelişmiştir. BG/NBD modelini değerlendirmek için, CLV değeri ve S değeri hesaplanmıştır. S değeri bir müşterinin üründen ne kadar memnun olduğunu ölçmek için kullanılır ve bu hesaplama için ürün puanlaması baz alınmıştır. CLV değeri ise müşterinin satın alma sıklığı, ortalama sipariş değeri ve müşteri sadakati gibi faktörleri dikkate alınarak hesaplanan müşterinin beklenen geliri tahmin edilmek için kullanılan bir metriktir. Yüksek CLV değerleri, müşterilerinizin uzun vadeli değerini yansıtabilirken aynı şekilde yüksek S değerleri, müşterilerin sık sık alışveriş yapma eğiliminde olduğunu gösterebilir. Bu durum, müşteri sadakati ve tekrarlayan işlemler için olumlu bir işaret olabilirken ortalama CLV ve S değerleri, müşteri davranışları uygun şekilde hareket edildiğini düşündürülebilir. Son olarak düşük CLV ve S değerleri, müşteri sadakatinin düşük olduğunu veya müşterilerin nadiren tekrar alışveriş yaptığını gösterebilir. Bu durumda müşteri ilişkilerini güçlendirmek ve pazarlama stratejilerini gözden geçirmek gerekebilir [30]. Bu iki değer, Doae Mensouri ve arkadaşının yayınladığı makalede [26] 0 ile 1 arasında normalize edilmesi uygun görülmüştür, şirketlerin müşteri hizmetleri ve ürün tekliflerinde iyileştirme alanlarını belirlemelerine yardımcı olabilir ve bu da müşteri memnuniyetinin ve sadakatinin artmasına yol açabileceğinden bahsetmişlerdir. Bu çalışmadaki BG/NBD modelinin değerlendirmeleri Tablo 12’de gösterilmiştir.

Tablo 12: BG/NBD modelinin değerlendirme metrikleri

CustomerID	CLV	S
120	0.3	0.08
12	0.3	0.02
18	0.27	0.03
13	0.27	0.07
139	0.26	0.019

Bu çalışmadaki ortalama CLV değeri 0.254 civarıyken, Tablo 8'deki müşteriler ortalama değere yakın olduğu görülmektedir. Bu müşterilere uygulanan strateji, önerilen ürünlere ve bunlara indirimlere uygulamaya devam devam edilmeli yorumu yapılabilir. Bunların yanı sıra Apriori algoritması için her InvoiceID için ProductID gruplanmıştır. En sık rastlanan öge kümelerini bulmak ve en güçlü birliktelik kuralını belirlemek için support, confidence ve lift değerleri kullanılacaktır [23]. Bu çalışmada analizi gerçekleştirecek kolonlardan "Antecedents" ürünleri, hangi ürünlerin birlikte satın alındığını gösterirken, "consequents" ürünleri, hangi ürünlerin sıkça birlikte satın alındığını gösterir. Bir kuralın "support" değeri, "antecedents" ve "consequents" kümesinin ürünleri ne sıklıkta birlikte satın alındığını ifade ederken, "confidence" değeri, "antecedents" öğelerinin "consequents" ürünlerinin içinde ne sıklıkta yer aldığını gösterir. "Lift" değeri ise, "antecedents" ürün kümesinin alınmasıyla "consequents" ürün kümesinin alınma olasılığını ne kadar artırdığını belirtir. Örnek olarak Apriori uygulanması sonucu bahsi geçen kolonlar için 192 ProductID'li ürünün kural çıkarımı Tablo 13'te gösterilmiştir.

Tablo 13: Apriori algoritmasının bazı çıktıları

antecedents	consequents	supprt	confidence	lift
(1756,189,1654)	(192)	0.000185	0.5	7.57
(1641,218)	(192)	0.000278	0.6	9.08
(873,126,221,1894)	(192,885)	0.000185	1	771.14
(192, 248, 218, 995)	(243,252, 894, 251)	0.000185	1	5398

Apriori algoritması 361.036 tane kural çıkarımı yapmıştır ve en çok göze çarpan ProductID'si 192 olan ürünün çoğu sepette yer aldığıdır. Tablo 13'te gösterilmemesine karşın çoğunlukla antecedents kümesinde yer alan bu ürün, müşterilerin bir ürün alırken sıklıkla aldığına işaret etmektedir. Bu kurallar daha sonra müşterilerin satın alma geçmişlerine dayalı olarak kişiselleştirilmiş teklifler oluşturmak için kullanılabilir, Apriori algoritmasının çıktısını analiz ederek hangi ürünlerin sıklıkla birlikte satın alındığını belirleyebilir ve bu bilgileri hedefli promosyonlar, satış stratejileri geliştirmek için kullanılabilir [23]. Örneğin sadece Tablo 13'te görülene göre 192 ve 218 ProductID'li ürünler için birlikte satın alınabilecek kampanyalar tasarlanabilir, indirim uygulanabilir.

Tablo 14: İkinci yöntemin değerlendirme metrikleri

Accuracy	Precision	Recall	Fallout	Miss Rate	F1 Score
0.447	0.456	0.438	0.543	0.56	0.447

Sonuç olarak model, 19402 tane satın alım işlemi yapılacağını ve 1224 farklı ürünün müşteriler tarafından alınabileceğini tahmin etmiştir. Ayrıca test kümesinde 6468 tane satın alım işlemi, 716 farklı ürün bulundururken bunlardan 263 farklı ürün satın alınmış ve 1303 satın alım işlemi yapılmıştır. Hem test hem de eğitim kümesinde yapılan işlem sayısı ile satın alınan ürün sayısı karşılaştırıldığında sadece 1303 işlemde 263 ürün satın alımı gerçekleşmiş ve model yaklaşık olarak %20 doğru müşteri için doğru ürün tahmini yapmıştır. Bunun yanı sıra müşteri bazlı tahmin yapılmadan doğru ürün tahmini için değerlendirme metrikleri Tablo 14'te belirtilmiştir.

5 TARTIŞMA

Literatür araştırmasında bahsi geçen öneri sistemleri, bu çalışmada geliştirilmeye çalışılmıştır. Sistemlerin bulunduğu bölümlerde kendi içerisinde literatüre uygun şekilde değerlendirilmiştir. Popülerlik tabanlı öneri sisteminde, veri kümesinde ürün miktarı bilgisi kullanılarak ürün derecelendirmesi oluşturulduktan sonra en sık satın alınan dolayısıyla en çok ürün değerlendirmesi bulunan ürünler önerilmiştir. Bu sistemin geniş kitlelere sahip ihtiyaçlar için kullanılmasındaki gereklilik saptanmıştır. Çünkü müşteri veya ürün bazlı herhangi bir kişiselleştirme yapılmadan ürünler, müşteri ayır

etmeksizin önerilmiştir. Bir diđer tasarlanan sistem olan KNN algoritmasının kullanıldığı içerik tabanlı öneri sisteminde ise yine gerekli imputasyon tekniđi uygulandıktan sonra Kosinüs benzerliđi ile müşteri-ürün matrisleri arasında benzerlik hesaplaması sonucu en yakın komşu seçimi ile ürün önerimi gerçekleştirmiştir. Literatürde yer alan araştırmadaki KNN algoritmasının kullanıldığı içerik tabanlı öneri sisteminin sonuçlarına göre daha iyi bir sonuç elde edildiđi görülmüştür. Precision ve F1 deđeri o makalede 0,5 deđerlerindeyken 0,9 sonucu bu makalede bulunmuştur. Dolayısıyla aynı algoritmayla işbirlikçi filtreleme denemesi diđer üç sistemden ikisinde daha gelişmiş işbirlikçi filtreleme zaten yapılacak olması amacıyla yapılmamıştır. İşbirlikçi filtrelemenin yer aldığı sistemde SVD algoritmalarıyla imputasyon tekniđi ayrı ayrı uygulandıktan sonra öneri yapılmıştır. SimpleSVD algoritmasının deđerleri çok iyi gözükmesine karşın ürün veya müşteri matrisinin tekil deđer matrisiyle işleme tabii tutulmamasından kaynaklı sistem güven vermemektedir. Tekil deđer matrisi, NormalSVD ve MidUserSVD algoritmalarında müşteri matrisi, MidItemSVD algoritmasında ise ürün matrisiyle işleme tabii tutulmuştur. MidTotalSVD algoritmasında ise hem müşteri hemde ürün matrisi tekil deđer matrisiyle işleme tabii tutulmuştur. Çıkan sonuçlar literatür araştırmasında yer alan makaleyle karşılaştırıldığında deđerlerde bir yakınlık gözlenmemiştir. Bunun üstüne geliştirilen bir diđer işbirlikçi filtreleme tabanlı sistemde RFM analizi yapılarak müşteri segmentasyonu uygulanmış ve K-Means algoritması yer almıştır. Sistemdeki deđerlerin literatür araştırmasında yer alan makaleye göre çok yakın olduđu görülmüştür. Literatür araştırmasında yer alan makalede 0,5 civarında olan metrikler bu çalışmada da gözlemlenmiştir. Dolayısıyla işbirlikçi filtreleme modellerinden SVD tabanlı sisteme göre daha iyi sonuç veren bu model ve literatür araştırmasında da yer alan makaledeki deđerlere yaklaştığı için daha başarılı olduđu görülmektedir. Bunun nedeni RFM analizi ve müşteri segmentasyonu olduđu düşünülmektedir. Son sistemde ise hibrit yaklaşım ele alındığı bahsedilmiştir. Bu yaklaşımda Apriori algoritması ile kural çıkarımı sonucu ürün analizi ile RFM analizi sayesinde müşterinin geçmiş davranışının analizi birleştirilmiştir. RFM analizinde çıkan sonuçlar BG/NBD modeline girdi olarak verilmiş, ardından CLV ve S deđeri hesaplanmıştır. Bu deđerler literatür araştırmasında yer alan makaledeki deđerler ile karşılaştırıldığında oldukça yakın olduđu gözlemlenmiştir.

6 SONUÇ

Çalışma neticesinde ürün tavsiyesi yapmak için bazı algoritma ve yöntemlerle öneri sistemleri perakende satış verileri kullanılarak öneri sistemleri tasarlanmıştır. Geliştirme yapmadan önce literatür araştırması yapılmış ve makalede yer alan algoritma ve yöntemlerin kullanılması kararlaştırılmıştır. Tasarlanan sistemler, popülerlik bazlı, içerik tabanlı, işbirlikçi filtreleme ve işbirlikçi filtreleme ile içerik tabanlı öneri sistemlerinin birleştirilerek hibrit tabanlı öneri sistemleri geliştirilmiştir. Bu sistemlerde K-Means, KNN, SimpleSVD, NormalSVD, MidUserSVD, MidItemSVD, MidTotalSVD, Apriori algoritmaları, Kosinüs benzerliđi, PCC teknikleri ve RFM, market sepeti analizleri kullanılmıştır. CLV, F1 gibi deđerlendirme metrikleriyle sistemlerin başarıları, uygulanan algoritmaya göre sınıanmıştır. Benzer verilerle öneri sistemi tasarlanırken uygulanacak algoritma, teknikler ve analizler fikir verme amacı gütmektedir.

7 REFERANSLAR

- [1] Huang, Zan & Zeng, Daniel Dajun & Chen, Hsiu-chin. (2007). Analyzing Consumer-Product Graphs: Empirical Findings and Applications in Recommender Systems. Management Science. 53. 1146-1164. 10.2139/ssrn.906512.
- [2] Tariq Mahmood and Francesco Ricci. 2009. Improving recommender systems with adaptive conversational strategies. In Proceedings of the 20th ACM conference on Hypertext and hypermedia (HT '09). Association for Computing Machinery, New York, NY, USA, 73-82. <https://doi.org/10.1145/1557914.1557930>
- [3] Hasan Mahmud, A.K.M. Najmul Islam, Syed Ishtiaque Ahmed, Kari Smolander, What influences algorithmic decision-making? A systematic literature review on algorithm aversion, Technological Forecasting and Social Change, Volume 175, 2022, 121390, ISSN 0040-1625, <https://doi.org/10.1016/j.techfore.2021.121390>.
- [4] L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. 22, 1 (January 2004), 5-53. <https://doi.org/10.1145/963770.963772>

- [5] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. GroupLens: an open architecture for collaborative filtering of netnews. In Proceedings of the 1994 ACM conference on Computer supported cooperative work (CSCW '94). Association for Computing Machinery, New York, NY, USA, 175–186. <https://doi.org/10.1145/192844.192905>
- [6] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl. 1999. An algorithmic framework for performing collaborative filtering. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99). Association for Computing Machinery, New York, NY, USA, 230–237. <https://doi.org/10.1145/312624.312682>
- [7] F. Ricci, L. Rokach, B. Shapira and P. B. Kantor, "Introduction to Recommender Systems Handbook," In: F. Ricci, L. Rokach, B. Shapira and P. B. Kantor, Recommender Systems Handbook, Springer, Berlin, 2011, pp. 1-35. http://dx.doi.org/10.1007/978-0-387-85820-3_1
- [8] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th international conference on World Wide Web (WWW '01). Association for Computing Machinery, New York, NY, USA, 285–295. <https://doi.org/10.1145/371920.372071>
- [9] Aamir, M., & Bhusry, M. (2015). Recommendation System: State of the Art Approach. International Journal of Computer Applications, 120, 25-32.
- [10] Aggarwal, C. C. (2016). Recommender systems. Springer
- [11] A. Fanca, A. Puscasiu, D. -I. Gota and H. Valean, "Recommendation Systems with Machine Learning," 2020 21th International Carpathian Control Conference (ICCC), High Tatras, Slovakia, 2020, pp. 1-6, doi: 10.1109/ICCC49264.2020.9257290.
- [12] Pazzani, M.J., Billsus, D. (2007). Content-Based Recommendation Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds) The Adaptive Web. Lecture Notes in Computer Science, vol 4321. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-72079-9_10
- [13] Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative Filtering for Implicit Feedback Datasets. 2008 Eighth IEEE International Conference on Data Mining, 263-272.
- [14] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. GroupLens: an open architecture for collaborative filtering of netnews. In Proceedings of the 1994 ACM conference on Computer supported cooperative work (CSCW '94). Association for Computing Machinery, New York, NY, USA, 175–186. <https://doi.org/10.1145/192844.192905>
- [15] Burke, R., Alexander Felfernig and Mehmet H. Göker. "Recommender Systems: An Overview." AI Mag. 32 (2011): 13-18.
- [16] Smyth, B. (2007). Case-Based Recommendation. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds) The Adaptive Web. Lecture Notes in Computer Science, vol 4321. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-72079-9_11
- [17] S. Vijayarani and S. Sharmila, "Comparative analysis of association rule mining algorithms," 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2016, pp. 1-6, doi: 10.1109/INVENTIVE.2016.7830203.
- [18] Georg Groh and Christian Ehlig. 2007. Recommendations in taste related domains: collaborative filtering vs. social filtering. In Proceedings of the 2007 ACM International Conference on Supporting Group Work (GROUP '07). Association for Computing Machinery, New York, NY, USA, 127–136. <https://doi.org/10.1145/1316624.1316643>
- [19] F. Islam, M. S. Arman, N. Jahan, M. H. Sammak, N. Tasnim and I. Mahmud, "Model and Popularity Based Recommendation System- A Collaborative Filtering Approach," 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2022, pp. 1-5, doi: 10.1109/ICCCNT54827.2022.9984348.
- [20] M. Gupta, A. Thakkar, Aashish, V. Gupta and D. P. S. Rathore, "Movie Recommender System Using Collaborative Filtering," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 415-420, doi: 10.1109/ICESC48915.2020.9155879.
- [21] FIARNI, Cut; MAHARANI, Herastia. Product Recommendation System Design Using Cosine Similarity and Content-based Filtering Methods. IJITEE (International Journal of Information Technology and Electrical Engineering), [S.l.], v. 3, n. 2, p. 42-48, sep. 2019. ISSN 2550-0554.
- [22] B. Uyanık ve G. K. Orman , "A Manhattan distance based hybrid recommendation system", International Journal of Applied Mathematics Electronics and Computers, c. 11, sayı. 1, ss. 20-29, Mar. 2023, doi:10.18100/ijamec.1232090W.
- [23] Y. S. Cho, S. C. Moon, S. C. Noh and K. H. Ryu, "Implementation of personalized recommendation system using k-means clustering of item category based on RFM," 2012 IEEE International Conference on Management of Innovation & Technology (ICMIT), Bali, Indonesia, 2012, pp. 378-383, doi: 10.1109/ICMIT.2012.6225835.
- [24] Xiaofeng Yuan, Lixin Han, Subin Qian, Guoxia Xu, Hong Yan, Singular value decomposition based recommendation using imputed data, Knowledge-Based Systems, Volume 163, 2019, Pages 485-494, ISSN 0950-7051, <https://doi.org/10.1016/j.knsys.2018.09.011>.
- [25] Z. Sharifi, M. Rezghi and M. Nasiri, "New algorithm for recommender systems based on singular value decomposition method," ICCKE 2013, Mashhad, Iran, 2013, pp. 86-91, doi: 10.1109/ICCKE.2013.6682799.
- [26] Mensouri, Doae and Azmani, Abdellah. "A New Marketing Recommendation System Using a Hybrid Approach to Generate Smart Offers" Applied Computer Systems, vol.27, no.2, 3922, pp.149-158. <https://doi.org/10.2478/acss-2022-0016>
- [27] Rabia Yörük, Eda Eşmekaya, Customer Lifetime Value, YBS Ansiklopedi, Cilt 5 Ekim, 2018
- [28] Palacios, Herman Jair Gómez, Robinson Andres Jimenez Toledo, Giovanni Albeiro Hernandez Pantoja and Alvaro Alexander Martinez Navarro. "A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change." Advances in Science, Technology and Engineering Systems Journal 2 (2017): 598-604.
- [29] Sadi Evren SEKER. "Retail Data Set". Kaggle. 2023. <https://doi.org/10.34740/KAGGLE/DS/3067824>
- [30] Zehra Rıdvançoğulları, Şadi Evren Şeker, Müşteri Yaşam Boyu Değeri, YBS Ansiklopedi, Cilt 10, Sayı 1, Haziran 2022