ÇOKLU ÖLÇÜT OY DEĞERLERİ ÜZERİNDEN VERİ MADENCİLİĞİ

Yüksek Lisans Tezi

Tuğba TÜRKOĞLU

Eskişehir, 2016

ÇOKLU ÖLÇÜT OY DEĞERLERİ ÜZERİNDEN VERİ MADENCİLİĞİ

Tuğba TÜRKOĞLU

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Yard. Doç. Dr. İbrahim YAKUT

Eskişehir

Anadolu Üniversitesi

Fen Bilimleri Enstitüsü

Haziran, 2016

Bu Tez Çalışması BAP Komisyonunca kabul edilen 1502F062 no.lu proje kapsamında desteklenmiştir.

JÜRİ VE ENSTİTÜ ONAYI

Tuğba Türkoğlu'nun **"Çoklu Ölçüt Oy Değerleri Üzerinden Veri Madenciliği"** başlıklı tezi **20/06/2016** tarihinde aşağıdaki jüri tarafından değerlendirilerek "Anadolu Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği"nin ilgili maddeleri uyarınca, **Bilgisayar Mühendisliği** Anabilim dalında Yüksek Lisans tezi olarak kabul edilmiştir.

	Unvanı-Adı Soyadı	İmza
Üye (Tez Danışmanı)	: Yard. Doç. Dr. İbrahim YAKUT	…….
Üye	: Yard. Doç. Dr. Alper Kürşat UYSAL	…….
Üye	: Yard. Doç. Dr. Mehmet KOÇ	…….

Enstitü Müdürü

ÖZET

ÇOKLU ÖLÇÜT OY DEĞERLERİ ÜZERİNDEN VERİ MADENCİLİĞİ

Tuğba TÜRKOĞLU

Bilgisayar Mühendisliği Anabilim Dalı

Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, Haziran, 2016

Danışman: Yard. Doç. Dr. İbrahim YAKUT

Bilgi ve iletişim teknolojilerinin gelişmesi, müşterilerin ürün ve hizmetler hakkında görüş, yorum ve değerlendirmelerini internet üzerinden paylaşma imkânı sunmuştur. Müşterilerin bu ürünleri değerlendirirken birden fazla ölçütü dikkate alarak değerlendirmesi yaygın bir uygulamadır ve bu şekilde müşterilerden toplanmış veriler de mevcuttur. Çoklu ölçüt müşteri değerlendirmelerinde veri madenciliği yöntemleri kullanılarak, müşteri beklentilerinin ve profillerinin etkili bir şekilde analizi gerçekleştirilebilir.

Bu çalışmada çoklu ölçüt oy değerlerinin veri madenciliği yöntemleri kullanılarak nasıl inceleneceğine odaklanılmıştır. Havayolu yolcularının uçuş deneyim değerlendirmeleri ele alınarak, yolcu eğilimleri tespit edilmeye çalışılmış ve yolcu profillerinin nasıl oluşturulacağı tartışılmıştır. Özellik tabanlı ve benzerlik tabanlı kümeleme yaklaşımlarıyla veriler incelenmiştir. Özellik tabanlı yaklaşımda, müşteriler seçilen özelliklere göre gruplandırılırken ikinci yaklaşımda probleme özgü benzerlik tabanlı kümeleme algoritmaları önerilmiştir. Önerilen kümeleme yöntemlerinde her bir kullanıcı için hesaplanması gereken benzerlik skoru tanımlanmış ve bu skora göre karakteristik kullanıcılar belirlenmiştir. Yolcular, karakteristik kullanıcılara olan benzerliklerine dayanarak, gruplandırılmıştır. Bu yöntemle elde edilen her bir küme için ReliefF algoritması uygulanarak niteliklerin yolculara göre önem sırası belirlenmiştir.

Anahtar Sözcükler: Benzerlik tabanlı kümeleme, Pearson korelasyon katsayısı, Çoklu-ölçüt oylar, ReliefF algoritması, Skytrax,

ABSTRACT

DATA MINING ON MULTI-CRITERIA RATING VALUES

Tuğba TÜRKOĞLU

Department of Computer Engineering

Anadolu University, Graduate School of Sciences, June, 2016

Supervisor: Asst. Prof. Dr. İbrahim YAKUT

The development of information and communication technologies offers the possibility of sharing on customer views, comments and ratings about products and services over the Internet. Customers evaluate services or products by taking into account multiple criteria and in this context there are datasets collected from customers. Customer expectations and profiles can be effectively analyzed using data mining techniques over multi criteria customer reviews.

In this study, we focus on how multi criteria rating values will be investigated using data mining techniques. Using in-flight experience reviews of airline passenger, passenger trends are tried to be identified and how passenger profiles can be formed is discussed. Data are examined with feature-based and similarity-based clustering approaches. While feature-based approach grouped customers according to selected features, in the second approach novel similarity-based clustering algorithms are proposed in the view of research problem. In the proposed clustering methods, similarity score is defined to be computed for each users and according to this score characteristic users are determined. Passengers are clustered based on similarity between passenger and characteristic users. Then, ReliefF algorithm is applied for each obtained cluster, features are ranked according to importance in the view of passengers.

Keywords: Similarity-based clustering, Pearson correlation coefficient, Multi-criteria ratings, ReliefF algorithm, Skytrax

TEŞEKKÜR

Yüksek lisans tezim süresince benden yardımlarını esirgemeyen, karşılaştığım sorunlarda bilgi ve deneyimlerini benimle paylaşan değerli hocam ve tez danışmanım Yard. Doç. Dr. İbrahim YAKUT' a teşekkürü bir borç bilirim.

Yüksek lisans öğrenimim boyunca benden desteklerini esirgemeyen ve her daim yanımda olan aileme sonsuz teşekkürler.

Tuğba Türkoğlu

Haziran, 2016

20.06.2016

ETİK İLKE KURALLARINA UYGUNLUK BEYANNAMESİ

Bu tezin bana ait, özgün bir çalışma olduğunu, çalışmanın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamaların bilimsel etik ilke ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilemeyen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; bu çalışmanın Anadolu Üniversitesi tarafından kullanılan "bilimsel intihal tespit programı"'yla tarandığında ve hiçbir şekilde "intihal içermediğini" beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçlara razı olduğumu bildiririm.

Tuğba Türkoğlu

İÇİNDEKİLER

Sayfa

TABLOLAR DİZİNİ

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

$\boldsymbol{\sigma}$: Standart Sapma

$\boldsymbol{\varepsilon_q}$: Simülasyon Hata Değeri

B : İterasyon Sayısı

$\boldsymbol{C_i}$: i Kümesi

$\boldsymbol{cu_i}$: i Karakteristik Kullanıcısı

$\boldsymbol{E_n^*}$: *n* Boyutundaki Beklenen Değer

E(C_i) : Her Bir Küme İçin Entropi Değeri

$\boldsymbol{h}$: Etiket Sayısı

$\boldsymbol{j}$: Alt Nitelikler

$\boldsymbol{k}$: En Yakın Komşu

P(C_i) : Her Bir Küme İçin Saflık Değeri

$\boldsymbol{sim_{u,a}}$: *u* ve *a* Kullanıcıları Arasındaki Benzerlik

simratio : Benzerlik Skoru

$\boldsymbol{q}$: İdeal Küme Sayısı

$\boldsymbol{w}$: Ağırlık Değeri (ReliefF)

Wq : q Kümesi İçin Ağırlık Değeri (Gap İstatiksel Metot)

1. GİRİŞ

İnsanlar yaşam süreleri boyunca kişisel gereksinimlerine ek olarak toplumsal ihtiyaçlarını da karşılamak için karar verme yetilerini sürekli olarak kullanmaktadırlar. Karar aşamasında, ilk olarak problem belirlenir ve ne olduğu tanımlaması yapılır. Daha sonra alınacak kararın amacı, kriterleri varsa alt kriterleri, etkileri bilinmesi gerekmektedir. Tüm bunlar belirlendikten sonra karar verici için birden fazla alternatifin içinden en uygun karar belirlenmeye çalışılır (Karakaşoğlu, 2008). Her bir insan aynı durum, problem karşısında farklı kararlar alabilir. Çünkü her bir insan için bu karar alma aşamasındaki kriterlerin önem düzeyleri farklılık gösterebilir. Bu karar sürecinde de insanların kararını etkileyen doğa koşulları, amaçlar, seçenekler ve seçeneklerin sonuçları gibi birçok dış faktör olabilir (Kahraman, Cebeci ve Ruan, 2004, s. 171). Bilinçli ya da bilinçsiz olarak yapılan her şey alınan kararların sonuçlarıdır. Yaşam boyunca edinilen bilgiler, olaylar karşısındaki tutumu ve kararı etkiler. Fakat bu edinilen tüm bilgilerin kullanışlı olacağı anlamına gelmez. Bu bilgi bombardımanını önlemek için, bazen kullanılması gereken miktar kadar etkili bir analiz ve yöntem uygulaması yaparak, ilginç faydalı örüntüler ortaya çıkarmak şarttır.

İnsanlar günlük yaşamlarında aldıkları kararların veya işletmelerde karşılaştıkları problemlerin yanı sıra amaçlarını gerçekleştirmek için de karar mekanizmasını kullanmaktadır. İşletmeler bu kararları alırken doğru, güvenilir verilere ihtiyaç duymaktadırlar. Ancak bu karar probleminde farklı önceliklere sahip hatta birbiriyle çelişebilen birden fazla kriter de yer alabilmektedir. Örneğin, bir müşteri bir ürün almak istediğinde ürünün fiyatının yanında konforuna, dayanıklılığına da bakacaktır (Xu, Yang, 2001, s. 1). İşte bu aşamada çok kriterli karar verme yöntemleri devreye girmektedir. Çok kriterli karar verme yöntemleri 1960'lı yıllarda karar vermeye yardımcı olacak bir takım araçların gerekli görülmesiyle ortaya çıkmıştır. Çok kriterli karar verme yöntemleri, çoklu ve birbiriyle uyuşmayan kriterlerin olduğu bir probleme çözüm getirebilir ve doğru tercihin yapılmasını sağlar. Bu alanda yapılan çalışmalara bakıldığında Bülbül ve Köse, (2009) Türkiye'de gıda, içki ve tütün sanayinde faaliyet gösteren ve İstanbul Menkul Kıymetler Borsası'nda işlem gören 19 işletmenin finansal performansını, çok kriterli karar verme yöntemlerinden Topsis ve Electre yöntemleri kullanılarak değerlendirip, performans sıralaması yapmışlardır. Yapılan sıralamalar sonucunda her iki yöntemde de benzer sonuçlar elde edilmiştir. Sonuçlarda en iyi ve en

kötü performansı gerçekleştiren şirketlerin aynı olduğu görülmektedir. Gilliams vd., (2009, s.142) tarafından tarım arazilerinin ağaçlandırma probleminin çözümünde Promethee II, Electre III ve AHP metotlarının karşılaştırılması yapılmıştır. Sonuçlara bakıldığında, Promethee II diğerlerine göre daha iyi sonuç verdiği çıkarımı yapılmıştır.

Teknolojinin gelişmesi ve değişen çevre şartları nedeniyle veriden ziyade bilginin önemi zaman geçtikçe artmaktadır. İnternetin gelişmesiyle, her ne kadar artan veriler, veri yığınını ortaya çıkarsa da, bilgiye özellikle de faydalı bilgiye erişim problem olmaktadır. Kişi internet üzerinden herhangi bir konuda bilgi almak istediğinde, içinde kullanışsız bilgilerinde yer aldığı çok fazla sonuç sunulabilmektedir. Bu da kişinin karar verme mekanizmasını etkilemektedir. Burada önemli nokta ise, bu elde edilen sonuçlardan kişi için yararlı, kullanışlı bilgilerin önerilmesidir. Elde edilen veriler kullanışlı bilgiye dönüştürülmediği sürece fayda sağlamamaktadır. Bu veriler belirli bir amaç doğrultusunda işletildiği zaman değer kazanır. Bu değere ulaşma aşamasına birde rekabet eklenince, önem daha da artmaya başlamaktadır ve üstünlük yarışı ortaya çıkmaktadır. Fakat bu üstünlüklerin ise geçici değil kalıcı olması önem taşımaktadır. İşletmeler bu kalıcılığı sağlamak için çeşitli stratejiler geliştirmeye çalışmaktadır. Geliştirilen bu stratejilerin hepsi insan odaklıdır. Çünkü her insan farklı karakteristik özelliklere sahiptir. Bu karakteristik özelliklerin tespitinde ise veri madenciliği devreye girmektedir. Veri madenciliği ile kişilerin beklentileri ve ihtiyaçları analiz edilebilir. Şirketler bu analizleri kullanarak, uygun stratejiler geliştirip iş süreçlerini kolaylaştırabilir ve başarılarını uzun yıllar sürdürebilirler. Müşterilerin ne düşündüğü, beklentilerinin ihtiyaçlarının ne olduğu ve bu beklentilere nasıl karşılık verileceği belirlenmesi durumunda, hem müşteri hem de şirket açısından olumlu sonuçlar doğuracaktır (Karakaşoğlu, 2008).

Bu çalışmanın amacı, kullanıcıların yaşadıkları durumlar, edindikleri tecrübeler sonucunda paylaştığı değerlendirmeleri incelemektir. İncelenen bu değerlendirmeler sonucunda elde edilen bulgular hem şirketler hem de müşteriler için olumlu sonuçlar doğuracaktır. Müşteriler tecrübe edindikleri bir servis ya da ürün hakkındaki görüşlerini yazılı yani yorum şeklinde belirtebilir. Bu durum bazen sayısal oy şeklinde de olabilir. İnternet üzerinden Skytrax (airlineequality.com), TripAdvisor (tripadvisor.com), Booking (booking.com) gibi birçok farklı uygulamada bu tip değerlendirmeler görülebilir. Eğer bu genel yorumlar bilgi keşfi süreci açısından etkili bir şekilde analiz

edilebilirse iş ve market süreci için birçok ilginç örüntüler ortaya konabilir. Şekil 1.1'de TripAdvisor sitesinde yer alan bir kullanıcı yorumu görülmektedir. Burada kullanıcının memnuniyet durumunu belirten oy değeri, yolculuk esnasındaki memnuniyet durumunu anlatan bir yorum ve seyahati ile ilgili bilgiler yer almaktadır. Bu tez kapsamında belirtilen internet sitelerinden elde edilen çoklu ölçüt oyların incelenmesine odaklanılmıştır.

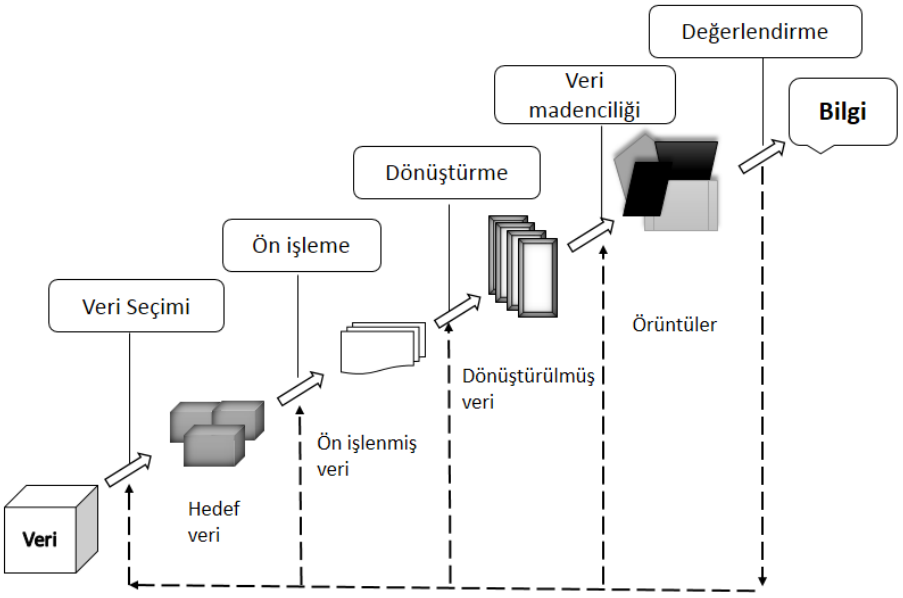
Şekil 1.1. *TripAdvisor Sitesinden Bir Kullanıcı Yorumu*

Bu çalışmada, kullanıcı yorumlarının yer aldığı internet sitelerinden biri olan Skytrax sitesindeki veriler ele alınmıştır. Burada kullanıcılar havayolu seyahatlerinde edindikleri deneyimleri, gelecekteki kullanıcılara ışık tutması için yorumlar ve oy olarak ifade etmişlerdir. Buradaki yorumlar sayısal ve sözel açıdan incelenebilecek, veri analizi yapılabilecek bir veri kümesi oluşturmaktadır. Bu analizlerin gerçekleştirilmesi ile kullanıcıların eğilimlerini belirlemek mümkün olabilir. Bu noktada akla ilk gelen soru *yolcuların seyahatlerinde dikkat ettikleri faktörler nelerdir*? Olacaktır. Bu çalışmada her bir yolcu için eğilimleri belirlemek yerine yolcular gruplar olarak ele alınacaktır. Çünkü grup içindeki bireyler aynı davranışları, tutumları sergiler. Bu noktada cevap aradığımız diğer sorular ise, bu gruplama işlemi nasıl gerçekleştirilecektir? Gruplama için kullanılacak bir metot var mıdır ve bu grupların belirlenmesindeki merkezi, kritik nokta nedir? Bu soruların cevaplarının arandığı ve tartışıldığı bu tez 4 bölümden oluşmaktadır. Veri madenciliği ile ilgili bilgiler Bölüm 2'de yer verilirken; çalışma boyunca kullanılan yöntemler, algoritmalar hakkında bilgiler Bölüm 3'te verilmiştir. Bölüm 4'te ise önerilen yöntemlerin ve yaklaşımların

uygulanması ve sonuçları sunulmuştur ve son bölümde ise çalışmadan genel olarak çıkartılan sonuç ve önerilere yer verilmiştir.

2. VERİ MADENCİLİĞİ

Veri madenciliği 1980'lerin sonuna doğru ortaya çıkmış ve 1990'larda büyük bir gelişme göstermiş ve bu gelişme zaman içerisinde devam etmiştir. Veri madenciliği, büyük miktarlardaki verinin içinden kullanışlı bilgilerin çıkartılması işlemidir (Kalıkov, 2006). Veri madenciliği kullanılarak veriler üzerinde analizler yapılır ve bu analizlere göre sonuçlar üretilir. Bu sonuçlara göre de belli çıkarımlar yapılır. Veri madenciliği ile karıştırılan farklı ifadelerde mevcuttur. Bunlar arasında en popüler olanı veritabanları bilgi keşfidir (Knowledge discovery from data (KDD)). Aslında veri madenciliği veritabanları bilgi keşfinin adımlarından biridir. Veritabanları bilgi keşfinin adımları veri seçimi, ön işleme, dönüştürme, veri madenciliği, değerlendirme ve sunum olarak Şekil 2.1'de gösterilmektedir.



Şekil 2.1. *Veri Tabanları Bilgi Keşfi Adımları*

Şekil 2.1'de görüldüğü gibi ilk aşama veri seçme aşamasıdır. Bu aşamada yapılacak olan analizle ilgili ihtiyaç duyulan veri alınır. Veri tabanları bilgi keşfinin ikinci aşaması ön işleme aşamasıdır. Bu aşamada elde edilen veriler gürültülü olabilir. Gürültülü veri, veri tabanında yer alan hatalı ve tutarsız verilerdir. Bu gürültülü verilerden kurtulmak için, eksik değer içeren bilgi, kayıt atılabilir veya sabit bir değer ile doldurulabilir. Bu sayede gürültülü verilerin etkisi indirgenip, ortadan kaldırılabilir. Bu bölümde ayrıca gerekiyorsa farklı kaynaklardan gelen verilerin tek çatı altında

toplanması da gerçekleşir. Üçüncü aşama olan veri dönüşü aşamasında ise, verinin, veri madenciliği tekniklerinin kullanılabilecek hale dönüşümünün gerçekleştirildiği bölümdür. Veri madenciliği veri tabanları bilgi keşfinin beşinci aşamasını oluşturmaktadır. Burada veri hazır hale geldikten sonra veri madenciliği metotlarının uygulanmasıdır. Veri madenciliği sonucunda gereğinden fazla bilgi elde edilebilir. Bu bilgilerin kullanışlı olup olmadığını, ilginç olup olmadığını belirlemek için örüntü değerlendirme kullanılır. Bu değerlendirme ise altıncı aşamayı oluşturmaktadır. Veri tabanları bilgi keşfinin son aşamasını sunum oluşturmaktadır. Bu aşama da bulunan örüntülerin, bilgilerin sunulmasıdır.

Veri madenciliği ile ilgili kullanılan pek çok yöntemin yanı sıra gün geçtikçe yeni yöntemler ve algoritmalar eklenmektedir. Bu yöntemleri gördükleri işlevlere göre başlıca dört gruba ayırabiliriz:

1) Sınıflama
2) Kümeleme
3) Birliktelik kuralları
4) Aykırı değer analizi

Bu yöntemleri, denetim durumları ve özelliklerine göre Tablo 2.1'deki gibidir.

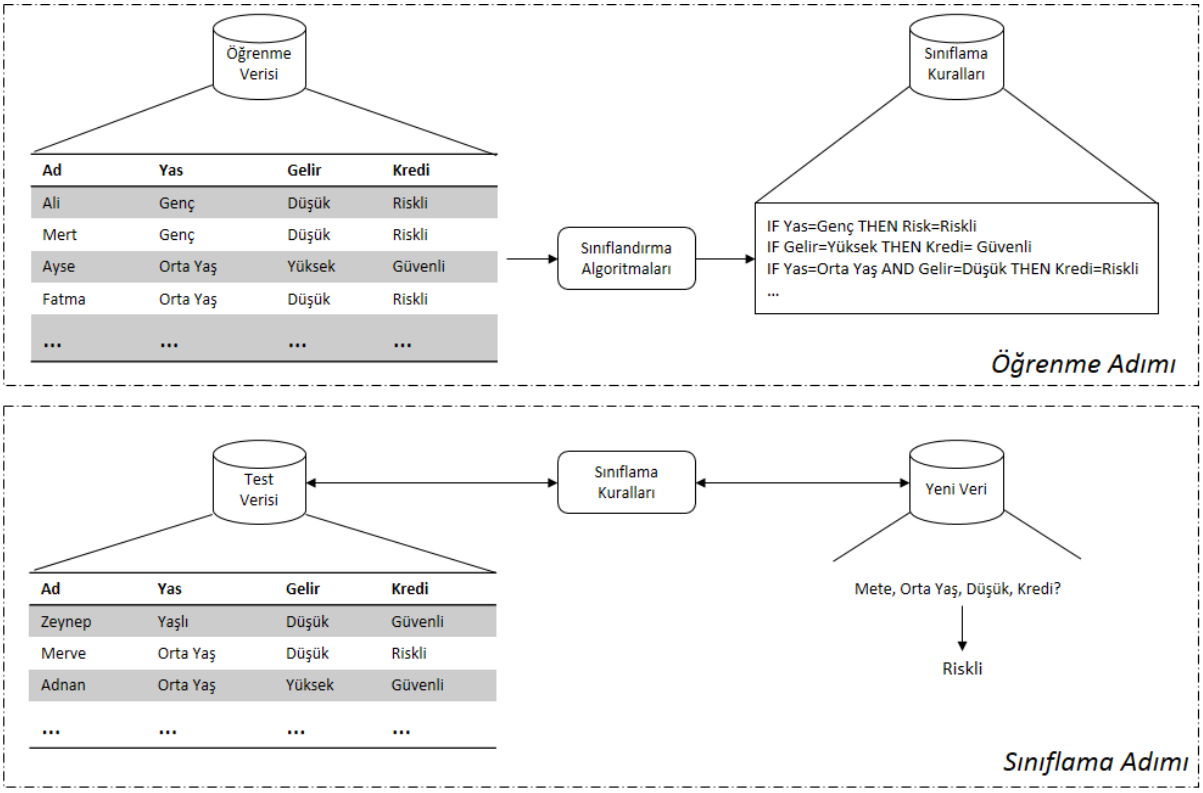
Tablo 2.1. *Başlıca Veri Madenciliği Yöntemleri*

Yöntem	Denetim Durumu	Özellik
Sınıflama	Denetimli	Tahmin edici
Kümeleme	Denetimsiz	Tanımlayıcı
Birlitelik Kuralları	Denetimsiz	Tanımlayıcı
Aykırı Değer Analizi	Denetimli/Denetimsiz	Tanımlayıcı

Tablo 2.1'de verilen özelliklerden tahmin edici, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesidir. Bir diğer özellik olan tanımlayıcı, karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır (Özekes). Denetimsiz öğrenmede örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklere göre sınıfların tanımlanması amaçlanır. Denetimli öğrenmede, ilk aşamada verinin bir kısmı modelin öğrenilmesi, diğer kısmı ise modelin geçerliliğinin test edilmesi için ayrılır.

2.1. Sınıflama

Sınıflama en yaygın kullanılan veri madenciliği tekniğidir. Veri sınıflama sürecinde, test verisi ve öğrenme verisi olarak iki veri, öğrenme ve sınıflama olarak da iki adım vardır. Öğrenme adımında, öğrenme verisi sınıflama algoritmaları kullanılarak analiz edilir. Daha sonra var olan verilerin bazıları kullanılarak kurallar oluşturulur. Yeni bir veri geldiği zaman bu kurallar kullanılarak o veri hakkında nasıl karar verileceği belirlenir. Sınıflama adımında ise, test verisi sınıflandırma kurallarının doğruluğunu tahmin etmek için kullanılır. Eğer doğruluk kabul edilebilir bir ölçütte ise o zaman kurallar yeni verinin sınıflandırılmasında kullanılabilir. Şekil 2.2 incelenecek olursa, öğrenme verisi üzerine sınıflama algoritmaları uygulanıp buradan sınıflama kuralları çıkarılır. Test verisi üzerinden bu kuralların doğruluğu ölçülüp yeni gelen veri hakkında karar oluşturulur. Sınıflamada amaç bir niteliğin değerini diğer nitelikleri kullanarak belirlemektir. Sınıf etiketleri olduğu için denetimli öğrenme metoduna girmektedir. Sınıflama yöntemlerine dolandırıcılık tespitinde ve kredi risk uygulamalarında sıkça başvurulur.



Şekil 2.2. *Veri Sınıflama Süreçleri*

Sınıflama modelinde kullanılan başlıca teknikler şunlardır; karar ağaçları, yapay sinir ağlar, genetik algoritmalar, *k*-en yakın komşu ve Naiive Bayestir. Kısaca bahsetmek gerekirse, karar ağaçları, isminden de anlaşılacağı üzere yapraklardan ve sınıf etiketlerinden oluşan bir ağaç yapısıdır. Yapay sinir ağları, insan beyninden esinlenip ortaya çıkartılan bir model iken, genetik algoritmalar optimizasyon problemi olarak tanımlanmaktadır. *k*-en yakın komşu metodunda ise, sınıflama işlemi sınıflandırılmak istenen verinin *k*- en yakın komşusu dikkate alınarak yapılır. Naiive Bayes ise olasılığı temel alan bir sınıflandırma yöntemidir.

2.2. Kümeleme

Kümeleme, veriyi gruplara veya kümelere ayırma işlemidir (Han, Karypis ve Kumar, 1999, s. 68-75). Aynı kümedeki elemanlar birbirleriyle benzerlik gösterirken, başka kümelerdeki elemanlar ile farklılık gösterirler. Bu kümeleme modelinde, sınıflama modelindeki gibi sınıf etiketleri yoktur. Bu bakımdan denetimsiz öğrenme metoduna girmektedir. Kümeleme yöntemlerinin çoğu veriler arasındaki uzaklıkları kullanır. Bu metot, gruplandırılmış veriler sayesinde özetleyici bilgiler elde edilmesine yardımcı olmaktadır. Kümeleme metodu ilk kez 1984 yılında Londra'daki meydana gelen kolera salgının çözümü için ortaya atılmıştır. Çok ciddi ölümlerin olması üzerine John Snow adlı bir kişi bir harita üzerinde ölen kişilerin yerlerini işaretlediğinde kayıpların bazı bölgelerde yoğunlaştığını fark ediyor. O bölgede su pompalarına bakılıp, atık su tesisindeki problem tespit edilerek koleradan meydana gelen ölümler engellenmiştir. Ana sokaklardan birindeki su pompasının sapını çıkarmak kolera salgınının sonlanması için yeterli olmuştur.

Kümeleme analizi, incelenen veriler arasındaki benzerlikleri, uzaklıkları dikkate alarak belirli gruplar içinde toplayarak sınıflandırma yapan, bu sayede birimlerin ortak özelliklerini ortaya koyan ve bu sınıflar ile ilgili genel tanımlar yapmayı sağlayan bir yöntemdir (Kaufman ve Rousseuw, 1990). Kümeleme analizinde amaç, gruplandırılmamış verileri benzerliklerine veya farklılıklarına göre gruplandırıp elde edilen bu gruplardan ilginç örüntüler ortaya çıkarmayı sağlamaktır. Her bir grup küme olarak adlandırılır ve bu küme içerisindeki elemanlar birbirleriyle benzerliklere sahip iken diğer küme elemanları ile farklılıklara sahiptir. Kümeleme analizi istatistik, örüntü tanıma ve makine öğrenme gibi birçok alanda başarılı araştırmaların yapılıp, ilginç sonuçlar elde edilmesini sağlar.

Literatürde pek çok kümeleme algoritması bulunmaktadır. Kullanılacak olan kümeleme algoritmasının seçimi, veri tipine ve amaca bağlıdır. Genel olarak başlıca kümeleme yöntemleri bölümleme, hiyerarşik, yoğunluk tabanlı, ızgara tabanlı ve model tabanlı olarak sınıflandınlabilir (Han, Kamber ve Pei, 2000). Bu modellerden hiyerarşik yöntem, elemanlar arasındaki uzaklıkları esas alıp kümeleme işlemini gerçekleştirir. Bölümleme yönteminde ise istenilen küme sayısına göre ayırma işlemi gerçekleştirilir. Örnek olarak k-means gösterilebilir (Han ve Kamber, 2016). Yoğunluk tabanlı kümeleme de, bir eşik değeri belirlenir ve bu değeri aşan bölgeler küme kabul edilir. Izgara tabanlı kümeleme, verileri incelemek için ızgara yapılarının kullanıldığı bir modellemedir.

2.3. Birliktelik Kuralları

Birliktelik kuralları veri madenciliğinde kullanılan ilk tekniklerden birisidir (Agrawal, Imielinski ve Swami, 1993, s. 207). Birliktelik kuralı ile geçmişteki alışkanlıklar, yapılanlar analiz edilerek geleceğe yönelik çalışmalar yapılmasını sağlayan bir yaklaşımdır (Özçakır ve Çamurcu, 2007, s. 21). Bu kuralda amaç, birbiriyle ilişkili olan değişkenlerin ortaya çıkarılması ve aralarındaki bağıntının büyüklüğünün tespit edilmesidir. Yoğun bir şekilde kullanılan internet ve mobil haberleşme teknolojileriyle birlikte gün geçtikçe veri yığınları oluşmaktadır. Birliktelik kuralları bu veri yığınlarını kullanarak veri kümeleri arasındaki birliktelik ilişkileri bulmaya çalışır. Biriken bu verilerden şirketler kaliteli sonuçlar elde etmek istemektedir. Şirketler bu veriler üzerinde birliktelik kuralları oluşturarak, karar alma aşamasını daha verimli hale getirmektedir.

Birliktelik kuralları pazarlama alanında uygulama alanı bulmuştur. Kullanıldığı en tipik örnek market sepeti uygulamalarıdır. Burada müşteri alışkanlıkları belirlenmeye çalışılır. Böylece firmalar pazarlama stratejisini bu bilgilere dayanarak düzenleyebilir.

Tablo 2.2'de bir örnek gösterilmiştir. Burada her bir satır, müşteri tarafından alınan elemanları göstermektedir. TID ise eşsiz tanımlayıcıları nitelendirmektedir. Bu tablodan örnek birliktelik kuralları çıkartılacak olursa;

{Bebek Bezi}→ {Su},

{Süt, Ekmek}→ {Yumurta, Kola},

{Su, Ekmek}→{Süt}

Bu kurallarda, bebek bezi alan müşteriler çoğunlukla yanında su aldığı, süt ve ekmek alanların yumurta, kola aldıkları ve son olarak su, ekmek alan müşterilerin ise yanında sütte aldığı tespit edilmiştir. Firmalar elde edilen bu kuralları kullanarak hangi ürün, hangi ürün ile birlikte alınıyorsa, bunları yan yana koyarak satış oranını arttırabilir. Basit bir veri üzerinde bu veri madenciliği teknikleri uygulandığında hız konusunda sıkıntı olmaz ama eğer milyonlarca veri üzerinde veri madenciliği teknikleri uygulanırsa, birliktelik sorgusu için kullanılan algoritmalar hızlı olması gerekmektedir (Agrawal ve Srikant, 1995, s. 3).

Tablo 2.2. *Market Sepeti Kayıtları*

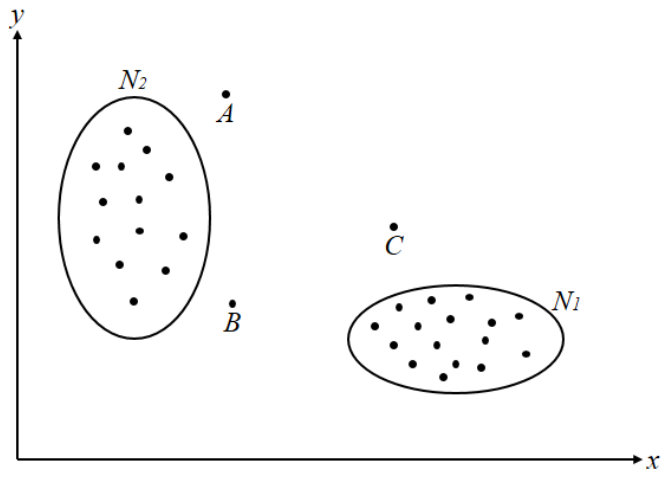
TID	**Elemanlar**
1	Ekmek, Süt
2	Ekmek, Bebek Bezi, Su, Yumurta
3	Süt, Bebek Bezi, Su, Kola
4	Ekmek, Süt, Bebek Bezi, Su
5	Ekmek, Süt, Bebek Bezi, Kola

Birliktelik kuralları için kullanılan algoritmalardan Apriori, CHARM, FP-Growth, vb. Bu algoritmalar içerisinde en popüler olanı Apriori'dir. Bu algoritma Agrawal ve Srikan (1994, s. 487) tarafından geliştirilmiştir. Apriori algoritması, eğer bir ürün kümesi sık değilse onun süper kümeleri de sık değildir prensibine dayanır ve birleştir, buda olarak iki adımı vardır. Algoritmanın çalışmasında veri tabanı baştan sona sürekli olarak tarandığı için dezavantaj olarak görülür. İkinci algoritma olarak CHARM, Apriori'den farklı olarak sık niteliklerin bulunmasında dikey olarak alınır. Avantaj olarak bakıldığında, destek sayısını bulmak için veri tabanı sadece bir defa taranır. Onun yerine kesişimler bulunur. Hipp, Güntzer ve Nakhaeizadeh (2000, s. 58) tarafından yapılan bir çalışmada birliktelik kuralları madenciliğinden olan Apriori, partition, fp-growth ve eclat algoritmalarının benzer ve farklı yönlerinin karşılaştırılması yapılmıştır. Bu karşılaştırmalar da zayıf güçlü yönlerinin belirlenmesine ek olarak, çalışma zamanı açısından da deneyler yapılmıştır. Her ne kadar algoritmalarda farklılık olsa da, çalışma zamanı, performans açısından değerlendirilme yapıldığında, şaşırtıcı bir şekilde aralarında büyük farklar olmadığı sonucuna varmışlardır.

2.4. Aykırı Değer Analizi

Aykırı değerler normal davranış sergilemeyen örüntülere sahip verilerdir. Bu aykırı veriler mevcut değerlerden oldukça farklı uç değerler olabileceği gibi, gürültülü

veriler olarak da düşünülebilir. Şekil 2.3'te aykırı verilere bir örnek gösterilmiştir. Burada N_1 ve N_2 birçok gözlemlerin yer aldığı küme olarak gösterilebilecek iki bölge iken, *A, B* ve *C* bu bölgelerden oldukça uzaktır. Bu yüzden *A, B* ve *C* aykırı veri olarak ifade edilebilir.



Şekil 2.3. *İki Boyutlu Düzlemde Aykırı Veri Örnekleri*

Aykırı değer analizde amaç, normal davranış sergilemeyen verilerin bulunmasıdır. Askeri denetim, kredi kartı sahtekârlığı gibi birçok alanda kullanılmaktadır. Aykırı değer analizinin üç geniş kategoride incelemek mümkündür: denetimsiz, denetimli ve yarı denetimli aykırı değer analizi. Denetimsiz aykırı veri analizinde etiketlenmemiş veri seti üzerindeki gürültülerin tespiti yapılır. Denetimli aykırı veri analizde, "normal" ve "anormal" olarak etiketlenen veri seti gereklidir. Yarı denetimli aykırı veri analizinde ise normal eğitici veri setinden normal davranışlar sergileyen model oluşturulur. Birçok aykırı veri tespiti teknikleri vardır. Bunlar, yoğunluk tabanlı teknikler, istatistiksel ve kümeleme tabanlı teknikler, destek vektör makineleri (SVM) ve yapay sinir ağları gibi tekniklerden de faydalanılmaktadır (Singh ve Upadhyaya, 2012, s. 307-323).

2.5. Literatür Taraması

Kullanıcı tutumlarının değerlendirilmesi ve kullanıcı profillerinin belirlenmesi ile ilgili birçok çalışma vardır (Vela ve Garcia, 2010, s. 235, Liou, Tzeng, 2010, s. 2230). Bu bağlamda Miranda ve Henriques (2013, s. 1) kümeleme metodunu kullanarak

havayolu yolcu verilerini analiz ettikten sonra farklı kampanya stratejileri geliştirmek isteyen şirketlere odaklanmışlardır. Bu çalışma için, onlar *k*-ortalamalar, öz düzenleyici haritalar (SOM) ve hiyerarşik öz düzenleyici algoritmaların performanslarını değerlendirmiştir. Analizler 20.000 yolcuya ait veri üzerinde gerçekleştirilmiştir. Sonuçlar incelendiğinde, *k*-ortalamalar en iyi olmasının yanı sıra öz düzenleyici haritalar ile benzer sonuçlar vermiştir. Vela ve Garcia (2010, s. 235), farklı uçuş nitelikleri ve gezi ile ilgili karakteristik özellikler hakkındaki yolcu değerlendirmeleri üzerine kümeleme işlemlerini gerçekleştirmişlerdir. Yaptıkları çalışmada, yolcuların dört segmentte var olduğu sonucuna varmışlardır. Bu segmentler, fiyat duyarlı, varış noktası ve uçuş bilinçli, duyarlı olmayan ve business, eğitici ve ikinci konut turistlerdir. Wang vd., (2002, s. 394) benzerliğin daha genel bir yapısını keşfetmeye çalışmışlardır. Onlar, pCluster olarak adlandırılan yeni bir model ortaya koymaya çalışmışlardır. Çalışma ekibi, kullanıcı tarafından belirtilen eşik değerinden daha büyük boyutta olan bütün pClusterları etkili bir şekilde ortaya çıkarabilen derinlik öncelikli algoritmayı tasarlamışlardır. Başka bir çalışmada Strehl vd., (2000, s. 58) dört benzerlik ölçütü olan öklid, kosinüs, pearson ve genişletilmiş jaccard kullanarak kümeleme kalitesini gözlemleyip performans değerlendirmesini yapmışlardır. Rastgele, öz düzenleyici haritalar, hiper-graf bölümleme, genel *k*-ortalamalar, ağırlıklı grafik bölümleme algoritmalarının karşılaştırmaları sonucunda, hiper-graf bölümlemenin performans açısından daha iyi olduğu sonucuna varmışlardır. Jarvis ve Patrick (1973, s. 1025) nonparametrik yolla verilerin nasıl kümeleneceği problemi üzerine odaklanmışlardır. En kısa arama ağacı kümeleme metodu ile görünüşte benzerliklere sahip bir metot sunmuşlardır. Balcan, Blum ve Vempala (2008) liste kümeleme ve hiyerarşik kümeleme uygulayarak, küme işlemini doğru bir şekilde yapmada yeterli olacak benzerlik fonksiyonunun özelliklerini incelemişlerdir. Liste kümeleme de amaç, kümelerin küçük bir listesini oluşturmaktır. Hiyerarşik kümeleme de ise kümelere budama işlemlerini uygulayarak hiyerarşik bir yapı oluşturulmak istenmiştir. Yang ve Wu (2004, s. 434) önerdikleri benzerlik kümeleme algoritmasının güçlülüğünü incelemeye çalışmışlardır. Onlar, beş algoritmanın birleşimi olan benzerlik kümeleme algoritması ile bulanık c-ortalamalar ve olası *c*-ortalamalarla karşılaştırmasını yapmışlardır. Sonuç olarak, seçilen toplayıcı hiyerarşik kümeleme metodu ile benzerlik tabanlı kümeleme, güçlü kümeleme sonuçları vermiştir. Arora vd., (2011, s. 761) benzerlik matrisinden küme olasılıklarını tahmin etmek için left-stochastic non-negative matris faktorizasyonu

(LSD) problemi ortaya konulmuş ve bu problem için de döndürme tabanlı algoritma öne sürülmüştür. Onlar, 13 veri seti üzerinden dört metrik açısından elde edilen sonuçlar incelenmiştir. Bu metrikler, küme benzerliği, hatalı sınıflandırma, karışıklık ve çalışma zamanıdır. Deneylerde LSD ile hiyerarşik LSD kümeleme algoritmalarının 9 diğer kümeleme algoritması ile karşılaştırmaları yapılmıştır. Sonuçlara bakıldığında LSD kümeleme ve hiyerarşik LSD birçok veri seti için en iyi sonuçları vermiştir. Liou ve Tzeng (2010, s. 2230) Tayvan havayolu marketi müşterilerinin basit ve çoktan seçmeli sorulardan oluşan ankete verdikleri cevaplar, kaba set tabanlı sınıflandırma ile incelemiştir. Sonuçlarda, müşterilerin karar aşamasını etkilen iki baskın kriter bulmuşlardır; güvenlik ve ücret. Ke-wu vd. (2007, s. 1284) havaalanı terminalinde bekleyen soruşturma verilerini incelemişler ve havayolu müşterilerini ID3 algoritmasını kullanarak sınıflandırmışlardır. Farklı müşteri grupları için farklı market stratejileri geliştirmenin şirketler için yararlı olacağı sonucuna varmışlardır. Velu ve Kashwan (2012, s. 151) bulanık mantık, genetik algoritma ve yapay sinir ağlarını uygulayarak belli özelliklere göre müşterileri incelemişlerdir. Onların çalışmaları esnasında, müşterilerin özelliklerini keşfetmeye çalışmışlardır.

Bu çalışmada odak nokta, kullanıcı yorumlarını incelemek ve Liou ve Tzeng (2010, s. 2230), Yakut ve Türkoğlu (2015, s.1) çalışmalarındaki gibi kullanıcıların belli servislere oy verdikleri zaman bu oyları vermelerindeki baskın kriterin ne olduğunun belirlenmeye çalışılmasıdır. Yakut ve Türkoğlu (2015, s.1) ve Liou ve Tzeng (2010, s. 2230) yaptıkları çalışmada müşteriler için baskın kriterlerin sırayla güvenlik ve ücret; parasal değer ve personel servisi hizmetlerinin olduğu sonuçlarına varmışlardır. Lacic vd. (2016) çalışmalarında yolcu tatmini için oy ve yorum özelliklerinden hangisinin daha belirleyici olduğunu tespit etmeye çalışmışlardır. Yazarlar çalışma boyunca Skytrax portalında havaalanları ve havayolu şirketleri hakkında paylaşılan yolcu değerlendirmelerini inceleyerek havaalanında kuyrukta bekleme zamanı, bekleme salonunda konfor, havayolu için kabin personeli ve koltukta diz mesafesinin genel müşteri memnuniyetine katkısının olduğunu saptamışlardır.

Bu çalışmada, var olan metotları uygulamanın yanı sıra ikinci bir çalışma olarak kullanıcı benzerliklerine dayanan yeni bir metot geliştirilmiştir. Bu esnada, bu kümeleme metodunu kullanarak temel kullanıcı profilleri elde edilmeye çalışılmıştır. Vela ve Garcia (2010, s. 235) ve Liou ve Tzeng (2010, s. 2230) çalışmalarındaki gibi,

müşteri verilerini incelemede kümeleme metotlarının kullanışlı bir araç olduğu kanaatine varılmıştır. Her grup içindeki kullanıcılar benzer davranışlar sergileyeceği için, kullanıcıları bireysel incelemekten ziyade, grup şeklinde incelemek daha yararlı olacaktır. Benzerliğe dayalı yapılan kümeleme çalışmaları: Jarvis ve Patrick (1973, s. 1025), Balcan vd. (2008), Yang ve Wu (2004, s. 434). Bu çalışmada da benzerlik ölçütleri esas alınarak yolcular gruplandırılmaya çalışılacaktır. Son olarak da havayolu servis kalitesi ile kullanıcı memnuniyetini etkileyen faktörler belirlenmeye çalışılmıştır.

Yolcu seçim analizini kolaylaştırmak için, segmentasyon pazarlama için önemli araçlardan biridir. Müşteri, yolcu ayrımını yapmak başarılı pazarlama için anahtardır. Çünkü bu şekilde müşteri memnuniyeti önemli seviyelere yükseltilmiş olur (Migueis, Camanho ve Cunda, 2012, s. 9359). Bu süre içerisinde market segmentasyonu aynı segment içerisinde benzer davranışlar sergileyen müşterileri inceler. Müşteri davranışlarını analiz etmek ve müşteri seçimlerine dayanarak uygun stratejiler geliştirmek, müşteri segmentleri üzerinde çok etkilidir (Dickson, 1982 s. 56). Bütün müşteriler uygun segmentlere atandıktan sonra, her bir segmentteki müşteriler davranış yönünden analiz edilebilir. Bu çalışmada kullanıcı modellerinin belirlenmesi için iki temel yaklaşım, segmentasyon uygulandı; özellik tabanlı veri analizi ve kümeleme tabanlı veri analizidir. İlk olarak müşteri, kullanıcı segmentleri belirlendi daha sonra bu segmentlerdeki kullanıcıların davranışları incelendi. İkinci yaklaşımda ise kullanıcılar arasındaki benzerliklere dayanan yeni kümeleme algoritmaları sunulmuştur.

3. YÖNTEM

3.1. Kullanılan Yöntemler

Bu bölümde literatürde önerilmiş ve çalışma kapsamında kullanılan yöntemler yer almaktadır.

3.1.1. Pearson korelasyon katsayısı

Pearson korelasyon katsayısı değişkenler arasındaki ilişkinin yönünü, derecesini ve önemini ortaya koyan, en yaygın olarak kullanılan istatistiksel yöntemdir. Bu analiz ile iki değişken arasındaki lineer ilişkinin güçlülüğü, derecesi ölçülür. Korelasyon katsayısında sonuç -1 ile +1 arasında değişmektedir. Eğer işaret pozitif ise değişkenler biri artarken diğerinin de arttığı, negatif ise değişkenlerden biri artarken diğerinin azaldığı anlamına gelmektedir. Bulunan sonucun +1 veya -1 olması arada mükemmel bir ilişki olduğunu gösterirken 0 olması arada doğrusal bir ilişki olmadığını gösterir (Sarvar vd., 2001, s. 285). Bu da, değerin mutlak değeri arttıkça iki değişken arasındaki ilişki de daha kuvvetlidir anlamına gelir. Korelasyon katsayısı kullanıcı u ve a arasındaki $sim_{u,a}$ olarak belirtilir. Formül aşağıda verilmiştir:

$$sim_{u,a} = \frac{\sum_{i=1}^{n}(r_{u,i}-\overline{r_u})(r_{a,i}-\overline{r_a})}{\sqrt{\sum_{i=1}^{n}(r_{u,i}-\overline{r_u})^2}\sqrt{\sum_{i=1}^{n}(r_{a,i}-\overline{r_a})^2}} \quad (3.1)$$

Burada $r_{u,i}$ kullanıcı u'nun i ögesi üzerinde verdiği oyu, $r_{a,i}$ a kullanıcısının i ögesine verdiği oyu, $\overline{r_u}$ ve $\overline{r_a}$ ise u. ve a. kullanıcıların ortalama oylarını temsil eder.

3.1.2. Gap istatistiksel metodu

Gap istatistiksel metodu Stanford araştırmacıları olan Tibshirani, Walther ve Hastie tarafından geliştirilmiştir. Bu metot veri seti içerisindeki kümelerin ideal sayısını belirlemek için kullanılan standart bir metottur. {X_{ij}}, i=1, 2, … , n, j=1, 2, …, p, n bağımsız örnekleri, p ise özellikleri ifade eden bir veri kümesi, C_1, C_2,...,Cq verilerin atandığı kümeler olsun.

$$D_r = \sum_{i,i' \in C_r} d_{ii'} \quad (3.2)$$

Burada $d_{ii'}$ i ve i' örnekler arasındaki uzaklıkları işaret eder. $d_{ii'}$ ifadesini belirlemek için en yaygın uzaklık ölçütü olan Öklid uzaklığı kullanılır.

$$W_q = \sum_{r=1}^{\gamma} \frac{1}{2n_r} D_r \quad (3.3)$$

n_r değeri her bir kümedeki yer alan eleman sayılarını ifade etmektedir. Eğer d uzaklık hesabı için Öklid kullanılacak olursa, W_q değeri bu karelerin toplamının küme boyutlarına bölümünden elde edilecektir. Öne sürülen bu yaklaşımın altındaki düşünce, verinin uygun boş referans dağılımının altında beklenen değer ile karşılaştırmalar yaparak $\log(W_q)$'nın grafiğini standardize etmektir. Optimal küme sayısı olan q değerinin belirlenmesi için $\log(W_q)$ değerinin bu referans eğimin en uzak altına düştüğü nokta, ideal q değeri olarak değerlendirilir (Tibshirani, Walther ve Hastie, 2001, s. 411). Bu yüzden;

$$Gap_n(q) = E_n * \{\log(W_q)\} - \log(W_q) \tag{3.4}$$

Burada E_n^* n boyutundaki beklenen değer ve bu $E_n^*\{\log(W_q)\}$ değeri, $\log(W_q)$ değerinin ortalamalarının alınmasıdır. Bu ortalamada, $\log(W_q)$ değeri, belirlenen bir iterasyon sayısına göre bulunan değerlerin ortalamasıdır. Bu iterasyon sayısı B ile ifade edilir. Bu değerlere ek olarak, birde elde edilecek olan hata değeri vardır. Bu simülasyon hata değeri,

$$\varepsilon_q = \sqrt{1 + {}^1/_B}\,\sigma_q \tag{3.5}$$

σ_q , q kümesi için standart sapmayı ifade etmektedir.

$$\bar{W} = \frac{1}{B}\sum_{i=1}^{B} \log(W_{qi}) \tag{3.6}$$

$$\sigma_q = \sqrt{{}^1/_B \sum_i (\log(W_{qi}) - \bar{W})^2} \tag{3.7}$$

Simülasyon hata değeri de hesaplandıktan sonra, en uygun q değerini belirlemek için (3.4) kullanılarak gap istatistiksel metodu hesaplanır. Bu hesaplamalardan sonra, aşağıda belirtilen (3.8) şartı sağlandığı takdirde, eldeki tüm kümeler için ideal küme sayısı q olacaktır.

$$Gap_n(q) \geq Gap_n(q+1) - \varepsilon_{q+1} \tag{3.8}$$

3.1.3. Özellik seçme yöntemleri

Relief algoritması, en çok bilinen özellik seçme metotlarından biridir. Oldukça başarılı ve etkili nitelik tahminleri yapabilen bir algoritmadır. Bu nitelik tahminler, özelliklerin ya da niteliklerin ağırlıklandırılması ile gerçekleştirilir. Nitelik ağırlıkları,

konvex optimizasyon problemini çözerek belirlenir. Ama Relief algoritmasının dezavantajı, tamamlanmamış veri ile baş edememesi ve iki sınıflı problemlerle sınırlı olmasıdır. Bu ve diğer problemleri çözmek için Relief algoritmasının genişletilmiş hali olan ReliefF algoritması öne sürülmüştür. Bu genişletilmiş algoritma çok güçlü ve gürültülü, tamamlanmamış verilerin üstesinden gelebilir. ReliefF algoritmasının çalışma mantığına bakılacak olursa, ilk olarak rastgele bir şekilde bir örnek R_i seçilir, ama daha sonra H_j olarak adlandırılan, aynı sınıftan ona en yakın *k* adet komşu ve $M_j(C)$ olarak adlandırılan farklı sınıfların her birinden de *k* adet yakın komşu seçilir. R_i, H_j ve $M_j(C)$'nin değerlerine bağlı olarak tüm *A* nitelikleri için *w*[*A*] değeri güncellenir. Nitelik ağırlıkları -1 ile +1 arasında değişmektedir. En büyük pozitif değerler niteliğin önemli olduğu anlamına gelmektedir. Bu süreç kullanıcı tarafından belirlenen sayı kadar devam eder. Bu algoritmada güncelleştirme formülü, bütün kaçırma (miss) ve vurmaların (hits) ortalama katkılarıdır (Robnik Sikonja ve Kononenko, 2003, s. 23). Algoritma aşağıda verilmiştir:

Algoritma 1: ReliefF algoritması

Girdi: Sınıf degeri ve nitelik değerlerinin her bir eğitim örneği

Çıktı: Niteliklerin kalitesini belirten *w* vektörü

1: Bütün ağırlık değerlerini oluştur *w*[A]:=0.0

2: **for** *i*=1 **to** *m* **do**

3: Rastgele bir R_i örneği seç.

4: *k* en yakın H_j bul.

5: **for** her bir sınıf C≠class(R_i) **do**

6: *C* sınıfından *k* en yakın $M_j(C)$ bul.

7: **for** *A*=1 **to** *a*

7.1: $\mathrm{w[A]} = \mathrm{w[A]} - \sum_{j=1}^{k} \mathrm{diff}(A, R_i, H_j)/(m * k) +$

$\sum_{C \neq \mathrm{class}(R_i)} \left[\frac{P(C)}{1 - P(\mathrm{class}(R_i))} \sum_{j=1}^{k} \mathrm{diff}\left(A, R_i, M_j(C)\right)/(m * k)\right]$

8: **end for**

Diff fonksiyonu ile örnekler ve nitelikler arasındaki farklılıklar yani uzaklıklar hesaplanır. Bu fonksiyonun hesaplanması, niteliklerin yazılı veya sayısal olmasına göre

değişir. I_1 ve I_2 örnek, A ise nitelik olsun. Eğer nitelikler yazılı ise bu durumda hesaplama;

$$diff(A, I_1, I_2) = \begin{cases} 0; & value(A, I_1) = value(A, I_2) \\ 1; & diğer\ durumlarda \end{cases} \quad (3.9)$$

Sayısal nitelikler var ise hesaplama;

$$diff(A, I_1, I_2) = \frac{|value(A,I_1) - value(A,I_2)|}{\max(A) - \min(A)} \quad (3.10)$$

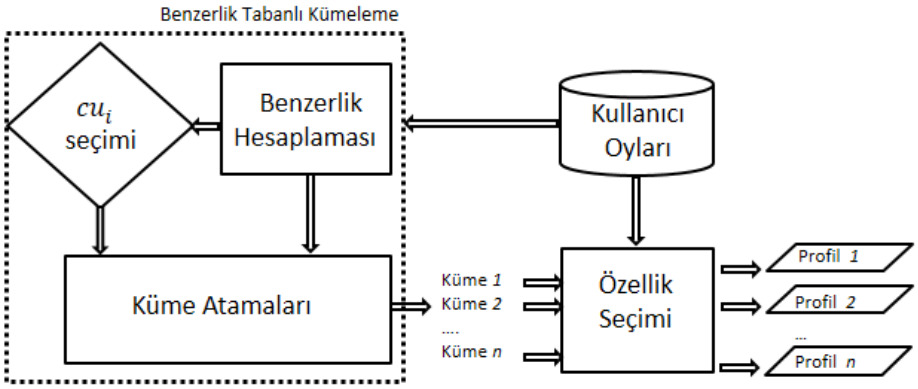
şeklinde olur. *k*'nın seçimi algoritmanın gürültülü verilere karşı güçlülüğünü arttırır. Bu değer kullanıcı tarafından belirlenebilir ama eğer *k*, 1 seçilirse algoritma gürültülü verilere karşı hassas olacaktır. Birçok çalışmada, *k* değeri 10 olarak seçilmiş ama *k* değerinin farklı şekilde seçilmesi, niteliklerin önem düzeylerinin incelenmesinde daha faydalı olacaktır. Son olarak *k* değerinin çok küçük seçilmesi de benzer kötü sonuçlara sebep olacaktır.

3.2. Önerilen Metot: Benzerlik Tabanlı Kümeleme

Bu çalışmada odak nokta yolcu yorumlarının, oylarının incelenerek yolcu eğilimlerinin belirlenmesidir. Burada öne çıkan soru; "yolcular seyahatlerini değerlendirirken hangi faktör ağır basmaktadır?". Çünkü bu faktörler yolcuların eğilimlerini temsil etmektedir. Bu işleme geçmeden önce, benzer tutumlara sahip yolcuları gruplandırmak daha anlamlı sonuçlara ulaştıracaktır. Yapılan bu çalışmada, baş edilmesi gereken ilk zorluk, bu yolcuların nasıl gruplandırılacağı sorusuna cevap bulmaktır. Bu gruplar nasıl oluşturulacaktır ve bu grupların belirlenmesindeki önemli noktalar neler olacaktır? Bu yüzden, yolcuları gruplara ayırmak için kümeleme işlemi önerilmiştir. Bu gruplandırma düşüncesine dayanarak birçok çalışma yapılmıştır (Lu ve Lin, 2002, s. 1, Kwan, Fong ve Wong, 2005, s. 189). İkinci baş edilmesi gereken zorluk ise, gruplar belirlendikten sonra her bir yolcunun tutumunun nasıl belirlenmesi gerektiğidir.

Bu bölümde bu çalışma için geliştirilen metotlar yer almaktadır. Öne sürülen bu metot, yolcular arasındaki benzerlik ilişkisini temel almaktadır. Bu benzerlikten kasıt, yolcuların kriterlere verdikleri önem düzeylerinin benzer olmasıdır. Şekil 3.1'de önerilen metodun işlem akışı gösterilmiştir. Bu önerme de ilk olarak, benzerlik tabanlı kümeleme metoduna dayanarak yolcular arasındaki benzerlikler hesaplanır. Bu yolcular arasında kümeleme işlemi yapmak için karakteristik kullanıcılar belirlenmeye çalışılır.

Bu karakteristik kullanıcılar temel alınarak, geri kalan kullanıcıların ataması yapılır. Bu atama işleminde, geride kalan kullanıcıların bu karakteristik kullanıcılarla olan benzerliklerine bakılıp, en son işlem olarak, karakteristik kullanıcılara göre oluşturulan bu kümeler üzerinde, profil ortaya çıkarmak için nitelik seçme işlemi uygulanır. Bu nitelik seçme işlemi her bir küme için gerçekleştirilir.



Şekil 3.1. *Önerilen Metodun Blok Diyagramı*

Algoritma 2'de benzerlik tabanlı kümeleme metodunun adımları verilmiştir. İlk olarak, korelasyon katsayısı formülü yani (3.1) kullanılarak tüm yolcular arasındaki benzerlikler hesaplanır. Her bir kullanıcı için benzerlik skoru hesaplaması yapılarak karakteristik kullanıcılar bulunmaya çalışılır. Benzerlik skorunu, *simratio* olarak göstereceğiz. Herhangi bir kullanıcı için benzerlik skoru, bu kullanıcıyla bütün pozitif benzerliklere sahip kullanıcıların benzerlikleri toplamının, negatif benzer kullanıcıların benzerlik toplamının mutlak değerine bölünmesi ile aşağıdaki denklemde verildiği gibi hesaplanır:

$$simratio_u = \frac{\sum^{n}_{(sim_{u,i}>0)} sim_{u,i}}{\sum^{n}_{(sim_{u,i}<0)} |sim_{u,i}|} \tag{3.11}$$

Algoritma 2'de yolcu benzerlikleri ve benzerlik skorları hesaplandıktan sonra, bu benzerlik skorlarına dayanarak, sınır çizgisi olacak iki yolcu seçilir. En büyük benzerlik skoruna sahip olan, ilk sınır kullanıcı olarak seçilirken, en az benzerlik skoruna sahip olan ise ikinci sınır kullanıcı olarak seçilir. İlk sınır kullanıcı cu_1 ve ikinci sınır kullanıcı ise cu_2 olarak ifade edilir. Bu algoritmanın 6. adımına bakıldığında her iki sınır kullanıcı karakteristik kullanıcı olarak seçilir ve bu kullanıcılar daha sonra bütün karakteristik kullanıcıların yer alacağı *S* kümesine eklenir. Bu *S* kümesinde yer alan

kullanıcılara, zıt kullanıcılar bulunmaya çalışılır. Diğer karakteristik kullanıcıların bulunması için *S* kümesindeki tüm kullanıcılara zıt kullanıcılar seçilmiştir. Bu şekilde bulunan her bir kullanıcı *S* kümesine eklenir. Bu işlemler 7. adımdaki `while` döngüsü tarafından gerçekleştirilir ve *S* kümesindeki kullanıcıların hepsi ile negatif benzerliğe sahip kullanıcı, olmayana kadar devam edecektir. Kümeleme işleminin son adımında ise, Adım 8' de yer alan `for` döngüsü ile her bir kullanıcı, en çok benzediği karakteristik kullanıcının kümesine atanır.

Algoritma 2: Benzerlik tabanlı kümeleme (Yaklaşım I)

Girdi: *n* kullanıcının *m* özellik ile ilgili (*nxm)* oy değerleri

Çıktı: Kullanıcıların kümelere atanması

1: *n* x *m* boyutlu oy matrisini yükle

2: $S = Ø$;

3: Denklem (3.1) kullanarak kullanıcı benzerliklerini hesapla.

4: Denklem (3.11) kullanarak tüm *u* kullanıcıları için $simratio_u$ hesapla

5: $cu_1 = \max(simratio_u)$ ve $cu_2 = \min(simratio_u)$ sahip cu_1 ve cu_2 kullanıcılarını seç.

6: $S = S \cup \{cu_1, cu_2\}$

7: **while** ($\exists_i\ sim_{u,i} < 0, \forall u \in S$) **do**

7.1: $\min(\sum_{u \in S} sim_{u,i})$ sahip kullanıcı *i* seç

7.2: $S \cup cu_i$

7.3: **end while**

8: **for** each *u* **do**

8.1: $\max(sim_{u,cu_j})$ sahip *u* kullanıcısını *j* kümesine ata.

8.2: **end for**

Algoritma 2'de karakteristik kullanıcılar zıtlıklara dayanarak belirlenmiştir. İlk iki karakteristik kullanıcının belirlenmesi dışındaki diğer karakteristik kullanıcılar zıtlık esas alınarak yapılmıştır. Bu karakteristik kullanıcılar seçilirken *S* kümesinde yer alan karakteristik kullanıcılara zıt ortak kullanıcı, diğer karakteristik kullanıcı olarak seçilmiştir. Algoritma 2'ye alternatif olarak farklı karakteristik kullanıcıların Algoritma 3 ile de seçimiyle kümeleme işlemi gerçekleştirilebilir. Algoritma 3'de, ilk iki karakteristik kullanıcı, Algoritma 2'de olduğu gibi maksimum ve minimum benzerlik skoruna sahip kullanıcılar seçilir. Bundan sonraki karakteristik kullanıcıları bulmak için,

kendisinden önceki iki karakteristik kullanıcılara zıt olan kullanıcılar bulunmaya çalışılır. Bulunan bu kullanıcılardan maksimum ve minimum benzerlik skoruna sahip kullanıcılar, diğer karakteristik kullanıcılar olarak seçilip *S* kümesine eklenir. Son aşamada Algoritma 2'de olduğu gibi Adım 8'de yer alan `for` döngüsü ile her bir kullanıcı en çok benzediği karakteristik kullanıcının kümesine atanır.

Algoritma 3: Benzerlik tabanlı kümeleme (Yaklaşım II)

Girdi: *n* kullanıcının *m* özellik ile ilgili (*nxm)* oy değerleri, küme sayısı *q*

Çıktı: Kullanıcıların kümelere atanması

1: *n* x *m* boyutlu oy matrisini yükle

2: $S = Ø$;

3: Denklem (3.1) kullanarak kullanıcı benzerliklerini hesapla.

4: Denklem (3.11) kullanarak tüm *u* kullanıcıları için $simratio_u$ hesapla

5: $cu_1 = \max(simratio_u)$ ve $cu_2 = \min(simratio_u)$ sahip cu_1 ve cu_2 kullanıcılarını seç.

6: $S = S \cup \{cu_1, cu_2\}$

7: $T = S$;

8: *cnt*=2;

9: **while** (cnt<*q and* $\exists_u\ sim_{u,u+1} < 0, \forall u \in T$) **do**

9.1: $\min(\sum_{u \in T} sim_{u,i})$ sahip kullanıcı *i* seç

9.2: $T \cup cu_i$

9.3: *cnt++*

9.4*:* **if** *cnt*<*q*

9.5: $\max(\sum_{u \in T} sim_{u,j})$ sahip kullanıcı *j* seç

9.6: $T \cup cu_j$

9.7: *cnt++*

9.8: **end if**

9.9: **else** *break;*

9.10: **end while**

10: **for** each *u* **do**

10.1: $\max(sim_{u,cu_j})$ sahip *u* kullanıcısını *j* kümesine ata.

10.2: **end for**

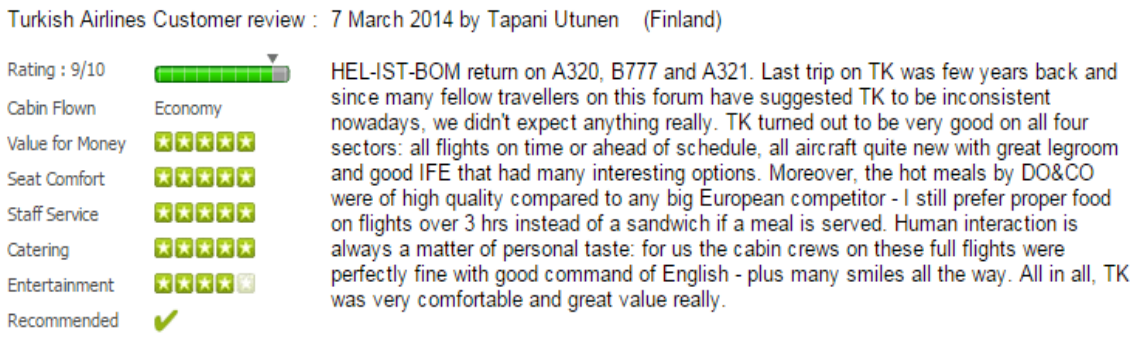
Burada amaç, uçuş servislerinin baskın karakteristik özelliklerini belirlemek olduğu için, kullanıcılara göre faktörlerin, niteliklerin öneminin belirlenmesi gerekir.

Diğer bir deyişle, kullanıcılardan elde edilen veriler incelenerek, her bir kullanıcının eğilimlerini belirlemektir. Kullanıcı eğiliminin hangi yönde olduğu, kullanıcı için hangi niteliğin daha önemli olduğu gibi sorular nasıl cevaplanabilir? Bu soruların cevabı ReliefF algoritması ile verilmiştir. Son işlem olarak ReliefF algoritması ile niteliklerin önem sırası belirlenmiştir. Bu şekilde, her bir küme için baskın nitelikler ortaya çıkartılmıştır.

4. VERİ ANALİZİ

4.1. Veri Seti

Skytrax, Londra'da bulunan çoğunlukla havacılık sektöründe araştırma ve danışmanlık yapan bir firmadır (Perezgonzalez ve Gilbey, 2011, s. 336). Skytrax, hava yolu seyahatlerinde en iyi personeli, havayolu, havaalanı ve birçok diğer faktörleri bulmak için havayolu için araştırmalar yapar. Bu çalışmada, Skytrax interaktif web sitesi olan www.airlinequality.com 'daki verilerden yararlanılmıştır. Burada en son olarak 681 havayolu firması ve 728 havaalanı için, oyları ve seyahat değerlendirmeleri yer almaktadır. Çok fazla yolcu verisi ve havayolu firması olması nedeniyle belirli havayolu firmaları seçilip ve bu havayolu firmalarına ait yolcu verileri kullanılmıştır. Şekil 4.1'e bakıldığında örnek bir havayolu için müşteri yorumları ve oyları yer almaktadır. Çalışmada kullanılan veriler, sayısal verilerdir. Şekil 4.1'de verilen örneğin, sağ tarafındaki yorumlardan ziyade ilgilenilen kısım sol taraftır. Yolcunun hangi kabinde seyahat ettiği, havayolu firmasını tavsiye edip etmediği gibi kullanışlı veriler yer almaktadır.



Şekil 4.1. *Uçuş Servisleri Hakkında Yolcu Yorumları*

Yolcular seyahat ettiği havayolu firmasının 5 özelliğine memnuniyet durumuna göre oylar vermiştir. 5 alt özellik, parasal değer (value of money), koltuk konforu (seat comfort), personel servisi (staff service), ikram (catering) ve eğlencedir (entertainment). Bu oylar 1 ile 5 arasında değişen yıldız şeklinde gösterilmektedir. Bu alt özelliklere oyu verdikten sonra, 1 ile 10 arasında değişen genel oy (rating) veren yolcu, ayrıca tavsiye edip etmediğini de (recommended) belirtmiştir. Skytrax içerisinde çok fazla havayolu ve kullanıcı yorumları yer aldığı için bunun bir alt kümesinin oluşturulması gerekiyordu. Bu yüzden Star Alliance'ın en geniş üyelerinden oluşan bir alt küme

oluşturulmaya karar verilmiştir. Star Alliance önemli, güçlü seviyede iş birliği yapmak için 2 ya da daha fazla havayolu firması arasındaki anlaşmadır. Aynı zamanda çok sayıda havayolu firması, günlük uçuşlar, hedefler ve ülkeler ile dünya çapında bir havayolu ağıdır (Holtbrügge, Wilson ve Berg, 2006, s. 306). Aynı ittifaktan ilişkili havayolları tercih edilmiştir. Alt kümeyi seçmek için aynı ittifak içerisindeki havayolu firmaları, benzer karakteristik özelliklere sahiptirler ve aynı standartlarla hizmet verirler. Bundan dolayı aynı ittifak içerisindeki havayolu firmalarının verilerini seçmek tutarlı olacaktır. Bu çalışmada Star Alliance üyesi olan 5 havayolu firması seçilmiştir. Bu havayolu firmaları şunlardır, United, Lufthansa, Air China, Turkish, All Nippon Airways(ANA). Veri seti olarak, 1 Ocak 2014 ile 31 Aralık 2014 arasındaki kullanıcı yorumları ele alınmıştır. Bu seçilen havayolları yıllık taşıdıkları yolcu sayısı Star Alliance bakımından en geniş üyeleridir (www.staralliance.com). Seçilen havayollarının her biri farklı bölgelerdendir; Kuzey Amerika (ABD'den United), Avrupa (Almanya'dan Lufthansa), Orta Doğu (Turkiye'den Turkish), Uzak Doğu (Çin'den Air China ve Japonya'dan ANA). Toplam da 1494 yolcuya ait veri yer almaktadır. Bu verilerin dağılımına bakılırsa; 545 (United), 434 (Lufthansa), 98 (Air China), 341 (Turkish) ve 76 (ANA) şeklindedir.

4.2. Değerlendirme Ölçütleri

Küme değerlendirme ölçütlerinden biri olan saflık (purity), denetimsiz öğrenme algoritmalarının performansını değerlendirmede en yaygın kullanılan metotlardan biridir (Zhao ve Karypis, 2004, Zhao ve Karypis, 2002). Başlangıçta her bir küme, kümede en çok bulunan kategori ile etiketlendirilir. Bu etiket diğer kümeler içinde baskın bir nitelik ise diğer kümeler de bu kategori ile etiketlendirilir. Bu etiketlere dayanarak saflık hesaplaması yapılır. Saflık ölçütü, kümenin uyumunu değerlendirir. Saflık şu şekilde hesaplanır. Öncelikle her bir küme C_i için $P(C_i)$ hesaplanır.

$$P(C_i) = max_h({n_i}^h) \tag{4.1}$$

Burada $max_h({n_i}^h)$ değeri C_i kümesindeki baskın olan kategorinin sayısını belirtirken, ${n_i}^h$ise h kategorisinin, niteliğinin C_i kümesindeki sayısını belirtir. Genel saflık hesaplamasına bakılırsa,

$$Saflık = \frac{1}{n}\sum_{i=1} P(C_i) \tag{4.2}$$

olarak belirlenir. Burada n değeri verinin toplam boyutunu belirtmektedir.

Saflık bir sınıflandırma oranı olarak yorumlanabilir. Her bir kümede baskın bir kategorinin olduğu ve bu kümedeki tüm örneklerin bu baskın kategorinin bir üyesi olduğu tahmin edilebilir. İdeal bir küme için, sadece basit bir kategoriye sahiptirler ve bu küme için saflık değeri 1'dir. Saflık değeri ile küme kalitesi arasında bir doğru orantı vardır. Saflık değerindeki artış, küme kalitesinin de yüksek olduğu anlamına gelmektedir. Genel bir ifade olarak, saflık değeri ne kadar yüksek ise kümeleme kalitesi de o kadar iyidir (Huang, 2008).

Kümeleme kalitesinin ölçümü entropi (entropy) kullanılarak belirlenebilir. Bu ölçüt, her bir kümedeki sınıfların nasıl dağıldığına dayanır (Zhao ve Karypis, 2002). n_i boyutlu C_i kümesinin entropi değeri,

$$E(C_i) = -\frac{1}{logc}\sum_{h=1}^{k}\frac{{n_i}^h}{n_i}\log\left(\frac{{n_i}^h}{n_i}\right) \quad (4.3)$$

olarak tanımlanır. Burada *c* veri setindeki kategorilerin toplam sayısını belirtir.

Genel çözümün ortalama entropi değeri, her bir kümenin entropi değerlerinin toplamı olarak tanımlanır. Bu ifade şu şekildedir,

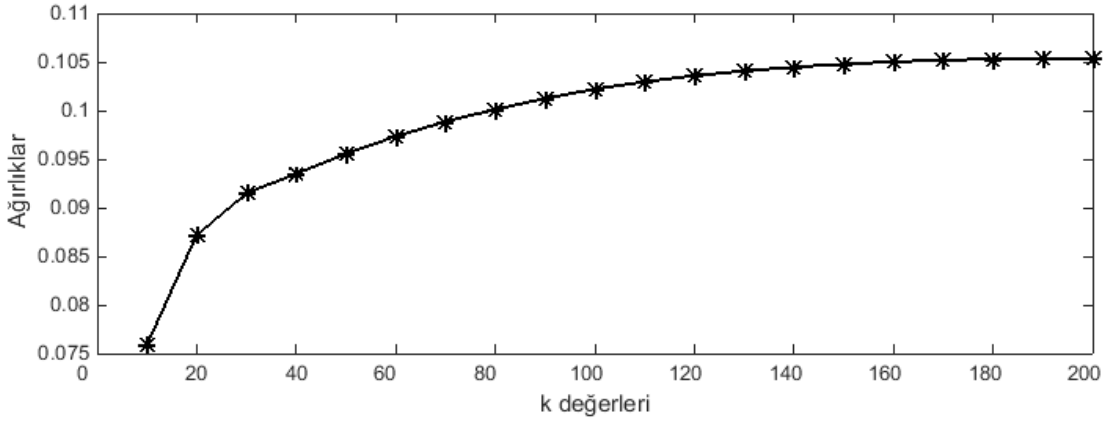
$$Entropi = \sum_{i=1}^{k}\frac{n_i}{n}E(C_i) \quad (4.4)$$

Entropi, saflık ölçütü arasında bir karşılaştırma yapılacak olursa, entropi ölçütünün saflık ölçütünden daha kapsamlı olduğu söylenebilir. Saflık ölçütünün aksine, eğer küme basit bir kategoriye sahip ise entropi değeri 0 olacaktır. Entropi değeri ile küme kalitesi arasında ters bir orantı vardır. Entropi değerindeki artış küme kalitesi değerini ters etkilemektedir. Başka bir deyişle entropi değerindeki artış, küme kalitesinin düşük olduğu ifadesini taşımaktadır. Genel bir ifade olarak, entropi değeri ne kadar yüksek ise kümeleme kalitesi de o derece düşüktür (Huang, 2008).

4.3. Özellik Tabanlı Veri Analizi

Özellik tabanlı modellemede, yolcular seyahat ettikleri kabin sınıfına, havayolu firmasına ve yolcuların havayolu firmasını tavsiye edip etmediğine göre gruplandırma işlemi yapılmıştır. İlk yaklaşım da, yolcuların seyahat ettikleri kabin sınıfı dikkate alınarak yapılmıştır. Bu veri için yolcuların seyahat ettikleri kabin sınıfı sayısı premium ekonomi, business, ekonomi ve first olarak 4 adettir. Burada önemli bir nokta vardır, first sınıfında seyahat eden yolcuların sayısı çok az olduğu için bu sınıftaki veriler, business sınıfına dâhil edilmiştir. Business sınıfında yer alan yolcu sayısı 381, ekonomi

sınıfındaki yolcu sayısı 1002 iken premium ekonomi sınıfındaki yolcu sayısı 111'dir. Buradaki amaç, bu sınıflarda yer alan yolcuların her biri için seyahatleri sırasında baskın etmenlerin ne olduğunu belirlemektir. Bunun için etmenlerin önem sırasını belirlemeyi sağlayan ReliefF algoritması kullanılmıştır. Bu algoritma her bir sınıf için tek tek uygulanmıştır. Bu algoritma için en yakın komşu sayısı olan *k* değerinin belirlenmesi önemli bir sorundur. Bunu belirlemek için, *k* değerini 10'dan 200'e kadar arttırarak, her bir nitelik için ağırlık değerleri bulunmuştur. Bulunan bu ağırlık değerleri her bir *k* değeri için ortalaması alınıp, bu ortalamalara göre *k*'nın en büyük değere sahip olduğu nokta ideal *k* noktası olacaktır. Artan *k* değerlerine göre, elde edilen nitelik ağırlık değerleri Şekil 4.2'de gösterilmiştir. Bu şekle göre, bu çalışma için *k* değeri 190 iken maksimum seviyeye ulaştığı görülmüştür. Bu nokta bu çalışma için ideal *k* değeri olduğunu göstermektedir.



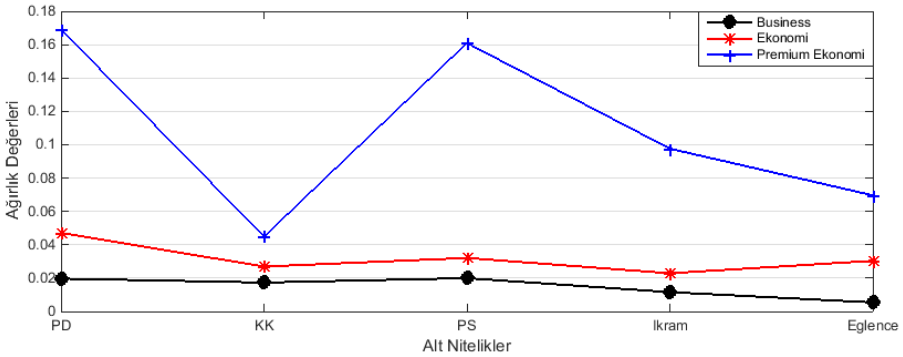
Şekil 4.2. *Değişen k Değerlerine Göre Ağırlıklar (Uçuş Sınıfı)*

Seçilen *k* değerine göre, belirlenen gruplara ReliefF algoritması uygulanmıştır. Sonuçlara bakıldığında, Tablo 4.1'de görüldüğü gibi business sınıfı için öncelik değerin parasal değer olduğu görülmektedir. Diğer sınıflara bakıldığında bunlar için de, parasal değerin öncelik taşıdığı görülmektedir. Business ve ekonomi sınıflarında yer alan yolcuların, benzer eğilimlere sahip olduğu sonucu çıkartılabilir. Parasal değerden sonra önemli olan etmenler, personel servisi ve koltuk konforu olduğu çıkarımı yapılabilir. Bu sonuç her iki sınıf için de aynıdır. Şekil 4.3'e göre, premium ekonomi sınıfına bakıldığında, diğer iki sınıftan farklı olduğu Şekil 4.3'te görülmektedir. Bu sınıf için paradan sonra baskın nitelikler, personel servisi ve ikramdır. Diğer sınıflar için önemli

etmenler arasına giren koltuk konforu, bu sınıf için önem sırasını yitirmekte ve en son sırada yer almaktadır.

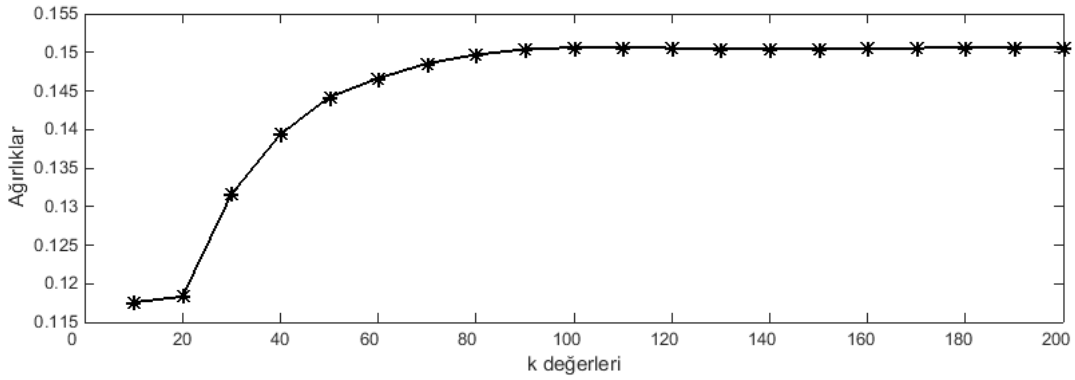
Tablo 4.1. *Her Bir Kümeye Göre Ağırlık Değerlerinin Dağılımı*

	w_1	w_2	w_3	w_4	w_5
Alt-oylar	*P.Değer*	*K.Konforu*	*P.Servisi*	*İkram*	*Eğlence*
Business	0.0527	0.0256	0.0367	0.0235	0.0128
Ekonomi	0.0384	0.0137	0.0230	0.0105	0.0109
Pre. Eko.	0.1692	0.0449	0.1610	0.0976	0.0696



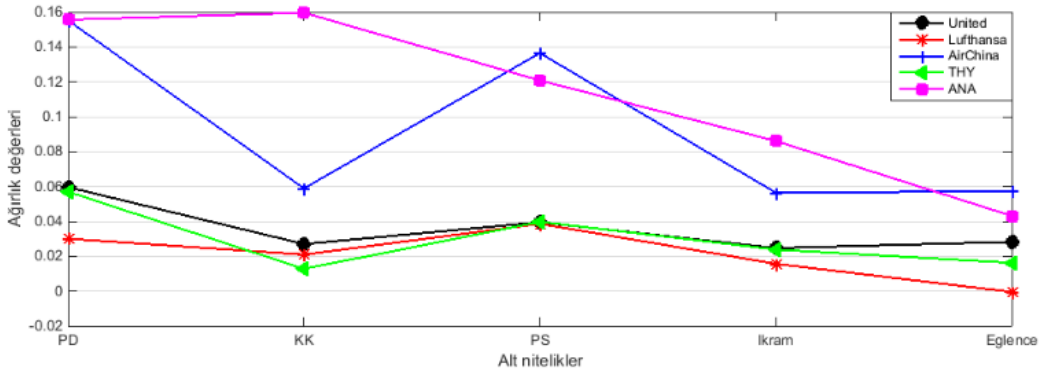
Şekil 4.3. *Her Bir Sınıf Uçuş Sınıfı İçin Niteliklerin Ağırlıkları*

İkinci çalışma olarak seyahat edilen havayolu firmalarına göre gruplandırma işlemi yapılmıştır. İlk çalışmadaki gibi bu verilere de, yolcular için önemli kriterleri belirlemek için ReliefF algoritması uygulanmıştır. Bir önceki çalışmadaki gibi en yakın komşu değeri olan *k*'yı belirlemek için benzer işlemler yapılmıştır. Bu işlemler sonucunda elde edilen *k* değerleri Şekil 4.4'te gösterilmiştir.



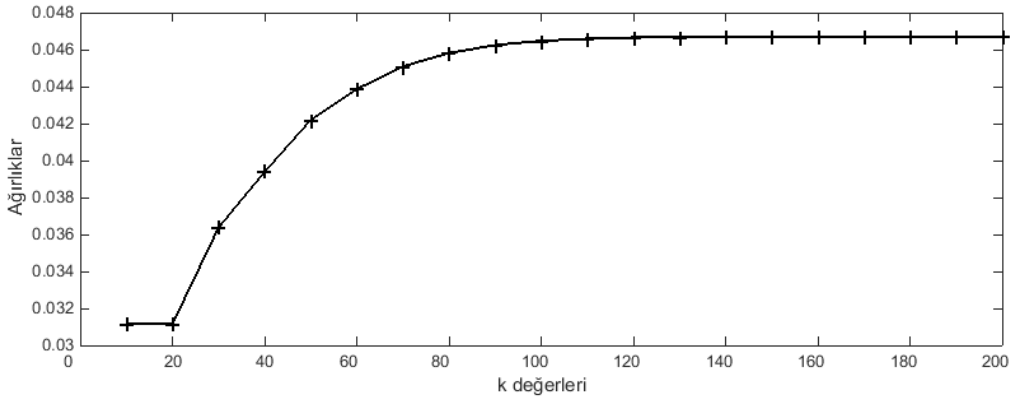
Şekil 4.4. *Değişen k Değerlerine Göre Ağırlıklar (Havayolu Firması)*

Şekil 4.4'te görüldüğü gibi *k*=100 iken maksimum ağırlık değeri elde edilmiştir. Bundan sonraki işlem, her bir havayolu yolcu grubu için ReliefF algoritmasının uygulanmasıdır. *k*=100 seçilerek uygulanan özellik seçimi sonucu, Şekil 4.5'te görülmektedir. Bu şekle göre, ANA haricindeki tüm havayolu firmaları için parasal değeri ilk sırada yer almaktadır. ANA havayolu firması incelendiğinde koltuk konforu ilk sırada iken, parasal değer ikinci sırada ve diğer niteliklerin önem düzeyi ardışık olarak azalan bir sıraya sahiptir. United, Air China ve THY firmalarına bakıldığında ikisi içinde ilk sıralarda, parasal değer ve personel servisi yer almaktadır. Lufthansa için de benzer bir durum söz konusudur. Parasal değer ve personel servisi ilk sıralarda yer alır fakat United ve THY'den farkı, personel servisinin ilk sırada, parasal değerin ikinci sırada yer almasıdır. Genel olarak bakıldığında ikram ve eğlence niteliklerinin son sıralarda görüldüğü, önem yönünden düşük olduğu çıkarımı yapılabilir. THY firması için koltuk konforu niteliğinin genel oy üzerinde çok az bir etkiye sahip olduğu sonuçlarına varılabilir.



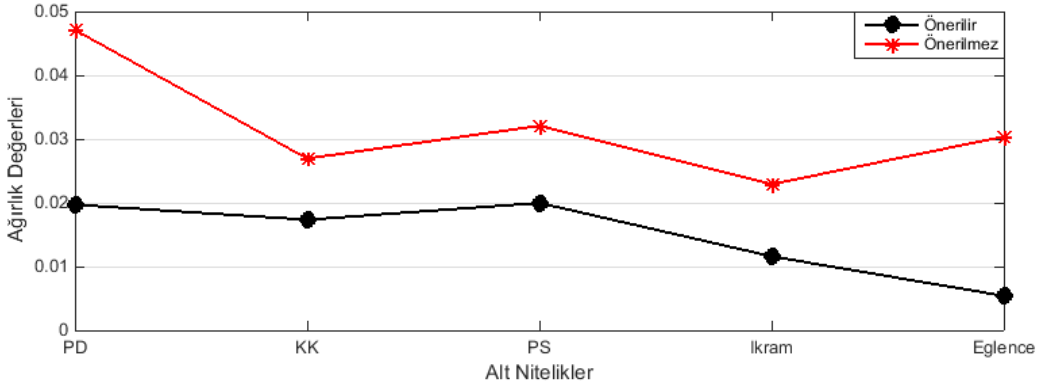
Şekil 4.5. *Her Bir Havayolu Firması İçin Niteliklerin Ağırlık Değerleri*

Özellik tabanlı modellemenin son çalışmasında ise, yolcuların havayolu firmalarını önerip önermediği incelenmiştir. Öneri, yolcu memnuniyeti ile doğru orantılıdır. Çünkü yolcu eğer memnunsa, etrafındaki arkadaşlarına, ailelerine ve iş arkadaşlarına şirket hakkında tavsiyelerde, önerilerde bulunabilir (Finn, Wang ve Frank, 2009, s. 209). Eldeki verilere bakıldığında, havayolu firmalarında 797 öneren ve 697 adet önermeyen yolcu yer almaktadır. Yüzde olarak değerlendirildiğinde yolcuların %56'sı memnun iken %44'ü memnun değildir. Yolcular öneren ve önermeyen şeklinde gruplandırıldıktan sonra her bir yolcu grubuna ReliefF algoritması uygulanmıştır. Bu algoritma uygulaması içinde, önceki çalışmalarda yapıldığı gibi ideal *k* değeri seçilmiştir. Şekil 4.6'da görüldüğü gibi, bu çalışma için ideal *k* değeri 120 olarak belirlenmiştir.



Şekil 4.6. *Değişen k Değerlerine Göre Ağırlıklar (Öneri Durumu)*

Elde edilen sonuçlar Şekil 4.7'de gösterilmiştir. Bu şekle göre, öneren yolcuların birçok nitelik için ağırlık değerleri, önermeyen yolculardan düşüktür. Şekil 4.7'de görüldüğü gibi, öneren yolcular için memnuniyet durumlarını etkileyen en temel faktörler paranın değeri ve personel servisidir. Aslında her iki yolcu grubu için bu sonuç geçerlidir. Öneren yolcular için eğlence kriterinin çok fazla önemli olmadığı görülürken, önermeyen yolcular için tam tersi durum söz konusudur. Bu yolcu grubu için bu kriter önemli faktörlerden biridir. Bu grup için en az öneme sahip olan kriterin ne olduğuna bakılacak olursa, bu kriterin ikram özelliği olduğu görülür.



Şekil 4.7. *Öneri Durumlarına Göre Niteliklerin Ağırlık Değerleri*

4.4. Kümeleme Tabanlı Veri Analizi

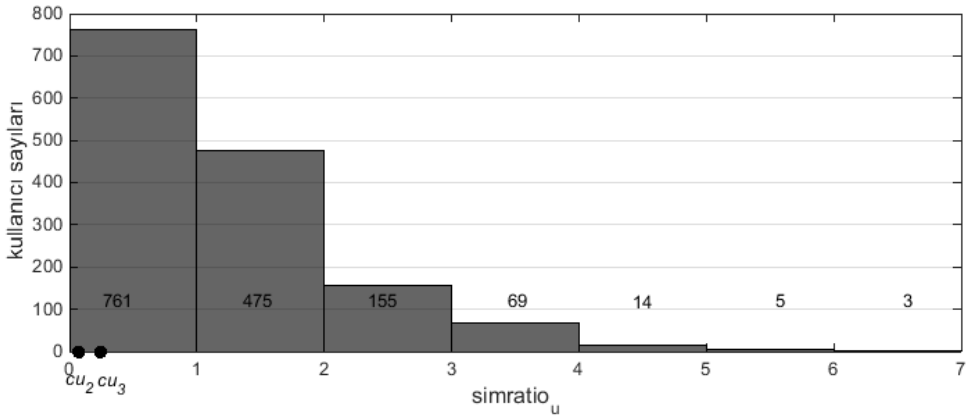
Bu çalışmada kullanıcı eğilimlerinin belirlenmesinde bölüm 3'te önerilen iki yaklaşım uygulanmıştır. Seçilen özelliklere göre yolcu gruplandırılmasına dayanan yaklaşım I iken yolcu benzerliklere göre gruplandırılma işlemi ise yaklaşım II olarak verilmiştir.

Yaklaşım I: Eldeki 1494 yolcuya ait veriye ilk olarak Algoritma 2 uygulanmıştır. Pearson korelasyon katsayısı kullanılarak veri seti içerisindeki yolcular arasındaki benzerlikler hesaplanmıştır. Daha sonra (3.1) kullanılarak, her bir yolcu için benzerlik skoru hesaplaması yapılmıştır. Elde edilen benzerlik skorlarının dağılımı Şekil 4.8'de gösterilmiştir. Şekil incelenecek olursa, 0 ve 1 arasında benzerlik skoruna sahip kullanıcı sayısı 761 ve yaklaşık olarak %51'lik bir çoğunluğa sahiptir. Şekil 4.8'e göre benzerlik skorundaki artma ile kullanıcı sayıları arasında ters orantı olduğu görülmektedir. Benzerlik skorunda artma meydana gelirken, kullanıcı sayısında azalma meydana gelmektedir. Benzerlik skoru hesaplandıktan sonra sıra karakteristik kullanıcıların belirlenmesine gelmiştir. Bu karakteristik kullanıcıların belirlenmesinde benzerlik skoru kullanılacaktır. Maksimum ve minimum benzerlik skoruna sahip kullanıcılar cu_1 ve cu_2 olarak seçilmiştir. cu_1 ve cu_2 diğer kullanıcılarla en çok ve en az benzerliğe sahip kullanıcı özelliklerine sahiptir. Bu iki karakteristik kullanıcı belirlendikten sonra diğer karakteristik kullanıcıların belirlenmesi için Algoritma 2'de belirtildiği gibi, cu_1 ve cu_2 'ye benzemeyen kullanıcılar listelenmiştir. Bu kesişimde toplamda 42 tane benzemeyen kullanıcılar ortaya çıkmıştır. Bu kullanıcılar içerisinde minimum benzerliğe sahip olan, üçüncü karakteristik kullanıcı olarak seçilmiştir. Üçüncü karakteristik kullanıcıda belirlendikten sonra, Algoritma 2' nin 7. adımında

herhangi bir sonuç alınmadığı için `while` döngüsü durup, karakteristik kullanıcı sayısı üç ile sınırlı kalmıştır. Bu bulunan karakteristik kullanıcılardan ikisi, Şekil 4.8'de gösterilebilir. Şekil 4.8'de önemli bir nokta vardır, x-ekseni [0-7] arasında sınırlandırılmıştır. Çünkü 7' den sonra 11 adet kullanıcı bulunmaktadır. cu_1 maksimum benzerlik skoruna sahip olduğu için ve bu skor da 40.32 olduğundan, Şekil 4.8'de gösterilememektedir. Elde edilen tüm karakteristik kullanıcılar hakkında bazı bilgiler Tablo 4.2'de verilmiştir. Bu bilgiler kullanıcı oyları ve benzerlik skorlarıdır.

Tablo 4.2. *Karakteristik Kullanıcılar Hakkında Bilgiler*

	G.Oy	**P.Değ.**	**K.Kon.**	**P.Ser.**	**İkram**	**Eğlence**	***simratio***
$\boldsymbol{cu_1}$	8	10	6	8	10	6	40.32
$\boldsymbol{cu_2}$	3	2	8	6	2	8	0.02
$\boldsymbol{cu_3}$	10	8	10	8	8	6	0.16



Şekil 4.8. *Kullanıcıların Benzerlik Skoru Dağılımları*

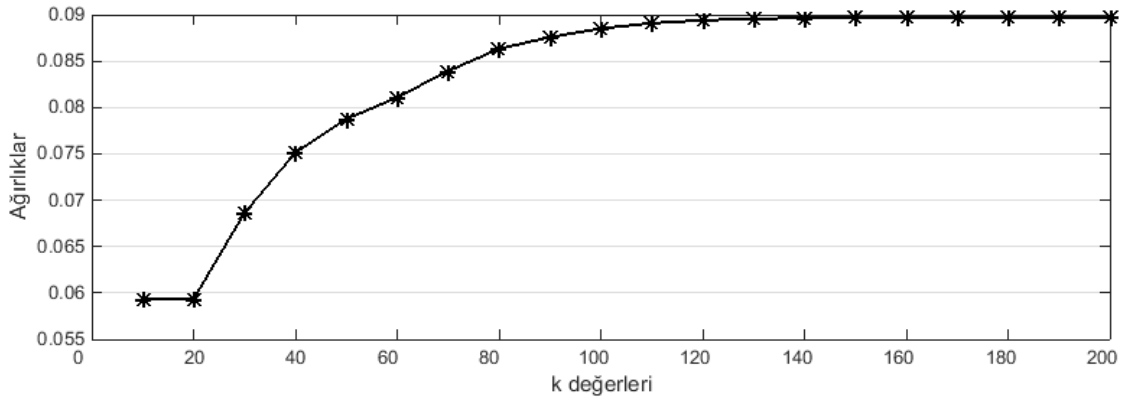
Belirlenen bu üç karakteristik kullanıcı aslında kullanıcıların kaç kümeye ayrılacağı ipucunu da vermektedir. Bu üç kullanıcı haricindeki diğer kullanıcıların her biri hangi karakteristik kullanıcıya daha çok benziyorsa o kümeye dâhil edilmiştir. Bu işlemler veri seti içerisindeki tüm kullanıcılar atanasıya kadar devam etmiştir. Atama işlemleri yapıldıktan sonra, sonuçlara bakıldığında, kümelerde makul dağılımlar elde edildiği görülmüştür. Her bir küme için ortalama oy değerleri Tablo 4.3'te gösterilmiştir. Bu bilgilere ek olarak, her bir küme içerisinde kullanıcıların memnuniyet durumu, hangi sınıfta seyahat ettikleri ve küme içerisinde kaç adet yolcunun yer aldığı gibi bilgilerde eklenmiştir. Pemium ekonomi ve first sınıflarındaki yolcu sayılarına

bakıldığında, çok az bir sayı olduğu görülmektedir. Bunun için premium ekonomi, ekonomi (E) sınıfına; first ise business (B) sınıfına dâhil edilerek yüzde hesaplaması yapılmış ve tabloda gösterilmiştir. Tablo 4.3 incelendiğinde, en kalabalık küme, toplam kullanıcı sayısının %41.3'lük bir oran ile 2 numaralı küme olarak görülmektedir. Yalnız çoğunluk olarak kalabalık olan bu kümedeki yolcular, memnuniyet açısından en düşük yüzdeye sahiplerdir. Tablo 4.3'teki değerlere bakılacak olursa, C_3 kapasite olarak en düşük iken %79'lük bir oranla yolcu memnuniyeti açısından ilk sıradadır. Küme 1 ise hem oy değerleri hem de yolcu memnuniyeti açısından ortada yer almaktadır.

Tablo 4.3. *Küme Kapasiteleri ve Ortalama Değerleri*

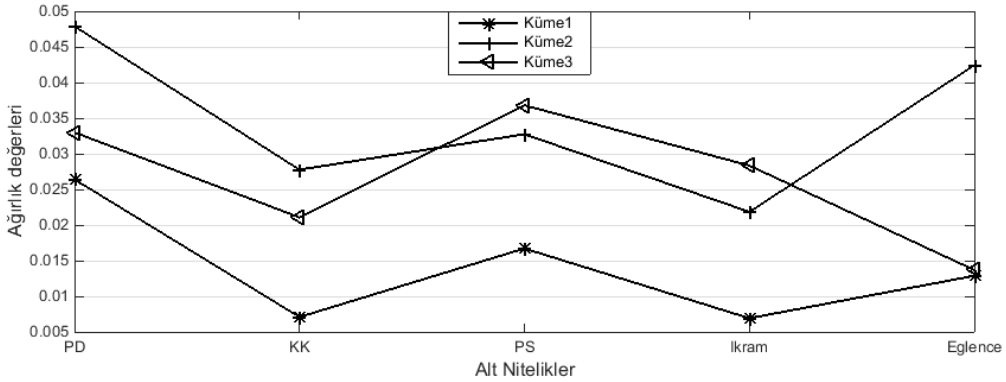
Küme no.		C_1	C_2	C_3
Ortalama Oylar	G.Oy	5.52	3.60	7.54
	P.Değ.	7.15	4.66	7.68
	K.Kon.	4.88	5.54	7.99
	P.Ser.	6.98	5.25	7.60
	İkram	6.98	4.55	6.77
	Eğlence	5.23	5.85	5.28
Tavsiye Durumu (%)		58	32	79
Kabin Sınıfı	Business	93	150	126
	Ekonomi	388	467	270
	E/Toplam (%)	0.81	0.76	0.68
Küme Boyutları		481	617	396

Kullanıcı grupları belirlendikten sonra, son işlem olarak her bir grup için önemli niteliklerin belirlenmesidir. Bunun için, ReliefF algoritması kullanılarak kullanıcılara göre her bir özelliğin önemi, önceliği belirlenmiştir. Fakat özellik seçimini gerçekleştirmek için, en yakın komşu sayısı (k) ne olacaktır? Bunu belirlemek için, k değerlerini 10'dan 200'e kadar arttırarak ReliefF algoritması çalıştırılmıştır. Şekil 4.9'a göre k değeri 140 iken, her bir niteliğin ağırlıklarının maksimum değere ulaştığı görülmüştür.



Şekil 4.9. *Değişen k Değerlerine Göre Ağırlıklar (Yaklaşım I)*

Burada alt oy değerlerinin genel oy değerlerine bağlı olduğu hipotezine dayanarak, bu özellik seçimi gerçekleştirilmiştir. Her bir küme için ReliefF algoritması çalıştırıldıktan sonra, Şekil 4.10'da çalıştırılan algoritma sonucunda, her bir bağımsız niteliğin elde edilen ağırlık değerleri verilmektedir. Her bir küme için niteliklerin önem derecesi incelendiğinde, parasal değer ve personel servisi ilk sıralarda yer almaktadır. Diğer niteliklere bakıldığında, Şekil 4.10'da görüldüğü gibi koltuk konforu ve ikram tüm kümeler içinde en az öneme sahiptir. Her bir küme için bu nitelikler incelenecek olursa, C_3 için personel servisi ilk sırada yer almaktadır. C_1 ve C_2 için paranın değeri en önemli faktör iken, bu özellik C_3 için ikici sırada yer almaktadır. Eğlence faktörü C_2 haricindeki diğer kümeler için en az önem derecesine sahip iken, C_2 kümesinde baskın nitelik olarak kendini göstermektedir. Eldeki sonuçlar genel olarak incelendiğinde, parasal değer ve personel servisi niteliklerinin en önemli özellikler olduğu çıkarımı yapılabilir. Bu özelliklerin müşteri memnuniyeti açısından önemli bir etkiye sahip olduğu söylenebilir. Çünkü sadece C_2 kümesi, temel olarak memnun olmayan müşterilerden oluşmaktadır.



Şekil 4.10. *Her Bir Kümedeki Niteliklerin Ağırlıkları*

Oluşturulan üç küme için küme değerlendirme ölçütleri olan saflık ve entropi hesaplaması yapılmıştır. Bu ölçütleri kullanmak için sınıf etiketleri gereklidir. Eldeki veri incelendiğinde sınıf etiketi olarak, yolcunun seyahat ettiği havayolu firmasını tavsiye edip etmediğini belirttikleri özellikler baz alınmıştır. Bu kümelerde sınıf etiketleri; yolcunun havayolu firmasını tavsiye ettiği "evet" ve tavsiye etmediği "hayır" etiketidir. Sınıf etiketleri de belirlendikten sonra, (4.2) ve (4.4) kullanılarak saflık ve entropi hesaplaması yapılmıştır. Bu çalışma için 3 küme mevcut ve bu kümelerdeki eleman sayıları 481, 617 ve 396'dır. Saflık hesaplaması için değişkenlerin ne olduğunu açıklamak gerekirse, *n* toplam veri sayısını, *h* ise sınıf etiketlerini temsil etmektedir. Toplam 1494 veri vardır. *n* değeri 1494 iken *h* değeri 2'dir. Her bir küme için elde edilen sınıf etiketleri sayısı, sahip oldukları maksimum etiket sayısı ve toplam sayıları Tablo 4.4'te gösterilmiştir.

Tablo 4.4. *Saflık ve Entropi Ölçütleri İçin Kümelerdeki Sınıfların Dağılımı*

Küme no.		C_1	C_2	C_3	***Toplam***
Etiketler	Evet	281	203	313	797
	Hayır	200	414	83	697
$\|C_i\|$		481	617	396	1494
$P(C_i)$		281	414	313	1008
$E(C_i)$		1.41	1.31	1.06	3.78

Yukarıdaki tablo yorumlanacak olursa, C_1 kümesindeki yolcuların 281 tanesi memnun iken, 200 tanesi bu yolculuktan memnun ayrılmamıştır. C_2 için ise yolcuların çoğunluğu yolculuklarını tavsiye etmedikleri yönünde bir öneri sunmuşlardır. Bu küme için, önermeyen yolcuların oluşturduğu bir grup denilebilir. Son olarak C_3 kümesi

incelenecek olursa, öneren yolcu sayısı 313'tür ve bu küme öneren yolculardan oluşmaktadır. Saflık hesaplaması için bu sınıf etiketlerinden, küme içinde maksimum değere sahip olan sınıf alınacaktır. İlk küme için maksimum değere sahip olan, "evet" etiketi ile 281, ikinci küme için ise 414 yolcuya sahip olan, "hayır" etiketi ve son küme için 313'lük bir sayıyla "evet" etiketi alınacaktır. Bu maksimum değerlerin toplanıp, toplam veri sayısına bölünmesiyle genel saflık değeri elde edilecektir. Genel saflık değeri 0.6747 olarak bulunmuştur.

İkinci küme değerlendirme ölçütü olarak entropi ölçütü kullanılmıştır. Bu ölçütte 2 adet sınıf kategorisi vardır. Yolcunun önerip önermediğini belirten etiketlerdir. Her bir küme için etiket sayıları elde edilmiş ve bu etiket sayıları kullanılarak ayrı ayrı entropi değerleri hesaplanmıştır. Bu hesaplamalar sonunda elde edilen sonuçlar Tablo 4.4'te gösterilmiştir.

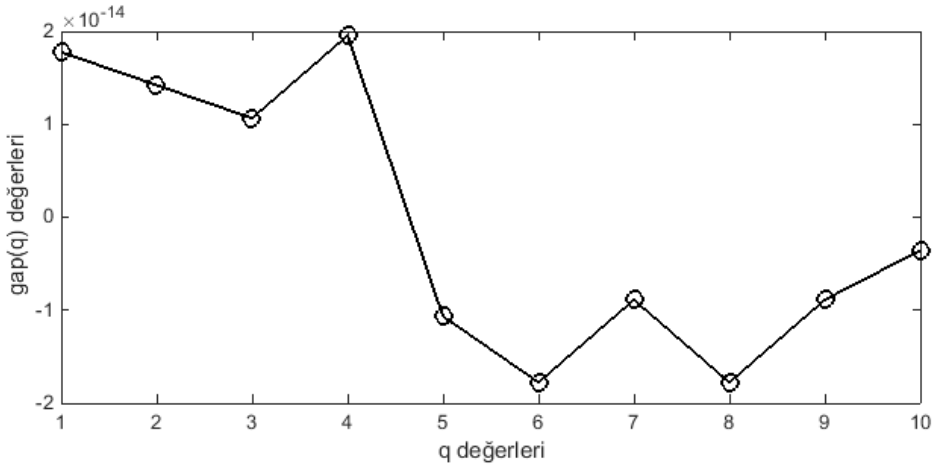
Denklem (4.4) kullanılarak genel entropi hesaplanmıştır. Bu hesaplama için, her bir kümenin entropi değerleri kullanılmıştır. Her bir kümenin entropi değeri ile küme boyutları çarpılmış ve tüm sonuçlar toplanmıştır. Elde edilen bu son sonuç ise toplam veri sayısına bölünmüştür. Genel entropi değeri 1.2827 olarak bulunmuştur.

Yaklaşım II: Bu yaklaşım var olan veri üzerinde Algoritma 3 uygulanarak gerçekleştirilmiştir. Denklem (3.1) kullanılarak ilk olarak kullanıcılar arasındaki benzerlik hesaplaması yapılmıştır. Daha sonra bu kullanıcılar arasındaki benzerlik skoru ilişkisi belirlenmeye çalışılmıştır. Bunun için (3.11) kullanılarak, her bir kullanıcı için benzerlik skoru ortaya çıkarılmıştır. Bu işlemler ilk yaklaşımın, ilk aşamaları ile benzerliğe sahiptir. Benzerlik skor hesaplaması bulunduktan sonra ilk iki kullanıcının belirlenmesi işlemi gerçekleştirilmiştir. İlk iki karakteristik kullanıcı maksimum ve minimum benzerlik skoruna sahip kullanıcılar seçilmiştir. Burada maksimum ve minimum benzerlik skorundan kasıt, diğer kullanıcıların hepsine en çok ve en az benzerliğe sahip olduğu anlamı taşımaktadır. cu_1 ve cu_2 belirlendikten sonra, sıra başka karakteristik kullanıcılar var mı bunun belirlenmesine gelmiştir. cu_1 ve cu_2'ye benzemeyen kullanıcılar listelenmiştir. Bu liste sonucunda toplam da 42 zıt kullanıcılar elde edilmiştir. Diğer iki karakteristik kullanıcının belirlenmesi için, bu liste içerisinden seçilecek kullanıcı için, benzerlik skorları dikkate alınmıştır. Diğer kullanıcılarla en çok ve en az benzeyen yani maksimum ve minimum benzerlik skoruna sahip kullanıcılar, diğer iki karakteristik kullanıcı olarak seçilmiştir. cu_3 ve cu_4'te belirlendikten sonra

cu_5 ve cu_6'nın belirlenmesine sıra gelmiştir. cu_5 ve cu_6'nın belirlenmesinde, önceki iki karakteristik kullanıcının dikkate alınması gerekir. cu_3ve cu_4'e zıt olan kullanıcılar dikkate alınıp bu kullanıcılar içerisinde benzerlik skoru en yüksek ve en düşük olan sonraki iki karakteristik kullanıcı olan cu_5 ve cu_6 olacaktır. Bu işlemler kullanıcı bulunmayana kadar devam ettirilmiştir. Fakat kullanıcı sayısı çok fazla olduğundan 7 karakteristik kullanıcı ile sınırlandırılmıştır. Bu işlemlerde, bulunan karakteristik kullanıcılar ile ilgili bilgiler Tablo 4.5'te gösterilmiştir. Tablo 4.5'te her bir karakteristik kullanıcının vermiş olduğu oy değerleri ve sahip olduğu benzerlik skoru yer almaktadır.

Tüm bu işlemlerden sonra, ikinci adım olarak kümeleme işlemi gerçekleştirilecektir. Fakat burada önemli bir nokta vardır; kümeleme işleminde gruplar nasıl belirlenecektir ve elde edilen tüm karakteristik kullanıcı sayısı dikkate alınarak küme sayısı nasıl belirlenebilir? İşte bu esnada istatistiksel bir metot olan gap devreye girmektedir. Bu istatistiksel metotta *k*-means, linkage gibi yöntemlerde kullanıcılar arasındaki uzaklıklar dikkate alınmaktadır. Bu çalışmada bu yöntemlerin aksine benzerlik dikkate alınarak yapılmıştır. Tablo 4.5'te verilen karakteristik kullanıcılara göre atama işlemi gerçekleştirilecektir. Bu atama işleminin nasıl yapıldığına bakılacak olursa, bu merkezi karakteristik kullanıcılar ile her bir kullanıcı arasındaki benzerliklere bakılır. Kullanıcı hangi karakteristik kullanıcıya daha çok benzer ise o kümeye dâhil edilir. Fakat öne sürülen bu metotta merkezi noktalar, Algoritma 3 kullanılarak elde edilen karakteristik kullanıcılar seçilmiştir. Değiştirilen bu istatistiksel metodun uygulaması incelendiğinde, ilk olarak birinci karakteristik kullanıcı alınır ve bu karakteristik kullanıcıya, tüm kullanıcılar atanır. Sadece bir küme oluşturulmuş olur. Bu atama işleminden sonra küme içerisindeki her bir kullanıcı ile küme merkezi noktası olan, karakteristik kullanıcı arasındaki uzaklıklar hesaplanır. Bu hesaplamadan sonra W_q değeri hesaplanır. Hesaplanan W_q değerinden sonra logaritmik sonuç bulunur. Bu sonuç tek bir karakteristik kullanıcı yani tek küme için geçerlidir. Elde edilen sonuç kümeye atanır ve sonraki işlem, ilk iki karakteristik kullanıcı için yapılacaktır. İki karakteristik kullanıcı olan cu_1 ve cu_2, merkezi kullanıcılar olarak seçilir ve diğer kullanıcılarla uzaklık durumlarına göre atama işlemi yapılır. Kullanıcı hangi karakteristik kullanıcıya daha çok yakın ise o karakteristik kullanıcının kümesine dâhil edilir. Yapılan bu işlemlerin bazı basamakları ilk yaklaşımı anımsatmaktadır. Tüm bu kullanıcı atama işlemleri herhangi bir kullanıcı kalmayana kadar devam edecektir. Sonuç olarak, iki küme oluşturulmuş ve bu kümenin ideal küme olup olmadığını belirlemek için gerekli

olan logaritmik sonuç da elde edilir. Bu sonuç, sonradan hesaplamalarda kullanılmak üzere, bir önceki sonucun atıldığı kümeye atanır. Daha sonra diğer karakteristik kullanıcılar içinde aynı işlemler uygulanır. Küme sayıları seçilen karakteristik kullanıcılara göre ardışık olarak artmaya devam eder. 7. karakteristik kullanıcıya gelindiğinde 7 küme oluşturulacak demektir. Bu yedi kümenin içerisinde kendisinden önceki karakteristik kullanıcılarda yer almaktadır. Bir sonraki karakteristik kullanıcının seçimi demek, kendisinden önceki karakteristik kullanıcılarında dikkate alınacağı anlamına gelmektedir ve kullanıcı atama işlemleri bu karakteristik kullanıcılar da dikkate alınarak yapılacak demektir. Her bir karakteristik kullanıcı ile 1'den 7'ye kadar oluşturulan kümelerde elde edilen logaritmik sonuçların her biri, eldeki kümeye atanır. Bu atama işlemleri bittikten sonra logaritmik sonuçtan sonra ihtiyaç duyulan bir diğer sonuç, beklenen logaritmik sonuçtur. Matematiksel işlemler sonucunda, (3.4) kullanılarak $Gap_n(q)$ değeri elde edilir. Daha sonra standart sapma değerleri bulunur. Bulunan bu gap değerlerinden standart sapma değeri çıkartılarak bir önceki küme sayısında bulunan gap değerleri ile arasındaki büyüklük küçüklük ilişkisine bakılır. Eğer bulunan gap değerinden çıkarılan sapma değerli sonuç önceki gap değerinden küçük ise o zaman ideal küme sayısı, bu büyüklük küçüklük ilişkisindeki en küçük "*q"* değeridir. Bulunan gap değerlerinin küme sayısına göre grafiği Şekil 4.11'de gösterilmiştir.



Şekil 4.11. *Küme Sayısına Göre Elde Edilen Gap Değerleri*

Tablo 4.5. *Karakteristik Kullanıcıların Oy Değerleri ve Benzerlik Skorları*

Karakteristik Kullanıcılar	G.Oy	P.Deg.	K.Kon.	P.Ser.	İkram	Eğlence	*simratio*
cu_1	8	10	6	8	10	6	40.32
cu_2	3	2	8	6	2	8	0.02
cu_3	7	6	6	8	6	6	0.33
cu_4	10	10	10	8	10	10	2.84
cu_5	3	4	8	6	2	6	0.13
cu_6	1	2	4	2	2	2	37.45
cu_7	8	8	8	10	8	6	0.32

Şekil 4.11'de belirtilen *k* değerleri, kümelerin nasıl elde edildiği incelenecek olursa, *k*=1 iken kullanılan merkezi nokta Tablo 4.5'de verilen cu_1'dir. cu_1 diğer kullanıcılar ile en yüksek benzerlik skoruna sahiptir. Bu küme için elde edilen gap değeri Şekil 4.11'de gösterilmiştir. *k*=2 iken iki küme oluşturulduğu anlamını taşımaktadır. Bu küme oluşturulurken iki karakteristik kullanıcı olan cu_1ve cu_2 kullanılmıştır. cu_2 diğer kullanıcılar ile en az benzerlik skoruna sahip olan kullanıcıdır. Bu iki karakteristik kullanıcıya olan uzaklıklarına göre diğer tüm kullanıcıların atama işlemi gerçekleştirilmiştir. Kullanıcı hangi karakteristik kullanıcıya daha çok yakın ise o zaman o karakteristik kullanıcının kümesine dâhil edilecektir. İki küme oluşturulduğu zaman elde edilen gap değerinin, cu_1 kullanılarak oluşturulan tek kümeden düşük olduğu görülmektedir. cu_1, cu_2, cu_3 ve cu_4 kullanılarak oluşturulan dört adet küme için, bu gap değerindeki azalma tersine dönmüştür. Bu değer, şu ana kadar maksimum gap değerine sahip olan *k*=1'i geçmiştir. *k*=5 iken oluşturulan 5 küme için, gap değerinde büyük bir düşme yaşanmaktadır. Bundan sonraki oluşturulan kümelerde dalgalanmalar yaşanmaktadır. Denklem (3.8)'de verilen şartın sağlandığı *k* değeri, 7 olarak belirlenmiştir. Bu *k* değer sayısı, ideal küme sayısı olarak verilmektedir.

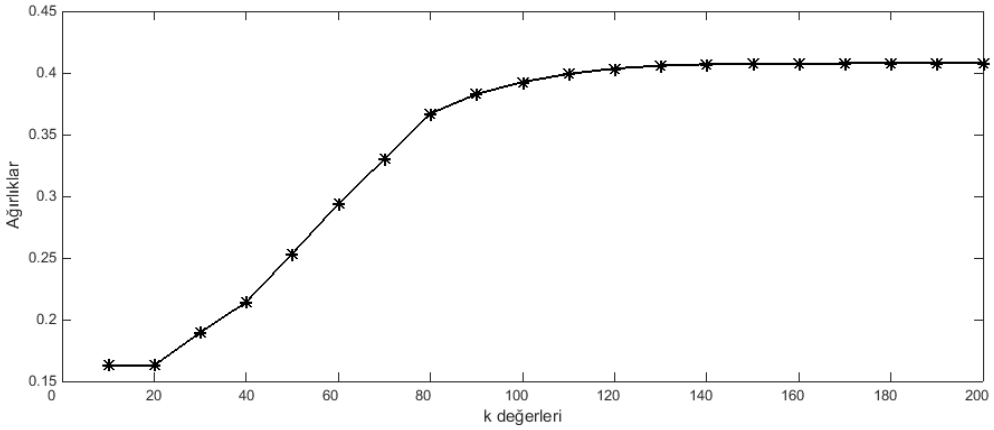
1'den 7'ye kadar küme oluşturulurken, (4.2) kullanılarak ilk olarak saflık değeri hesaplaması yapılmıştır. Bu hesaplama da ilk olarak iki küme için saflık değeri hesaplanmış ve sonra bu hesaplama ardışık şekilde artarak elde edilmiştir. Daha sonra (4.4) kullanılarak diğer değerlendirme metriği olan entropi hesaplaması yapılmıştır. Her iki değerlendirme metriği içinde sonuçlar bulunup Tablo 4.6'da verilmiştir.

Tablo 4.6. *Farklı Küme Sayıları için Saflık ve Entropi Değerleri*

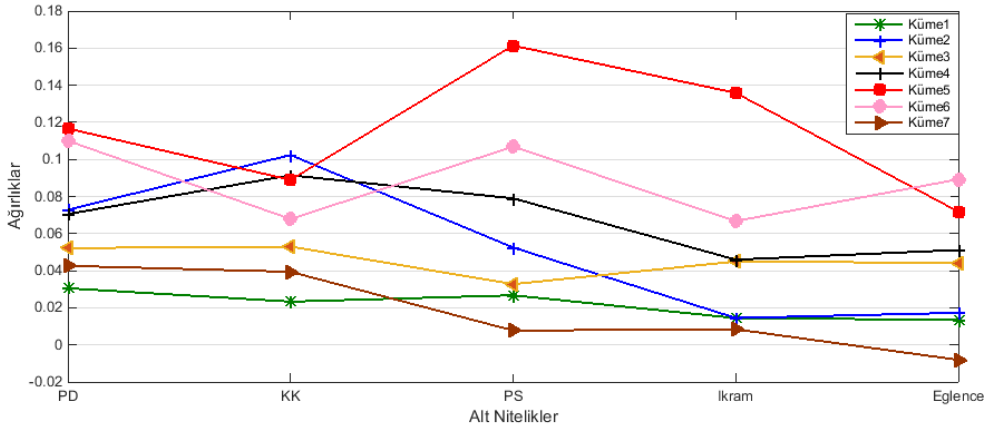
Kümeler	Saflık	Entropi
C_1, C_2	0.6017	1.39
C_1, C_2, C_3	0.6479	1.3402
$C_1, \dots, C_4$	0.6714	1.3035
$C_1, \dots, C_5$	0.6827	1.2832
$C_1, \dots, C_6$	0.7162	1.17
$C_1, \dots, C_7$	0.7149	1.1644

Saflık değeri ne kadar yüksek ise küme kalitesi de o kadar yüksek olduğu, entropi değeri ne kadar yüksek ise küme kalitesinin de o derece düşük olduğu önceki bölümlerde ifade edilmiştir. Tablo 4.6 incelendiğinde, C_1, C_2 ifadesinde iki küme oluşturulduğu gösterilmek istenmiştir. Bu iki küme için merkezi noktalar Tablo 4.5'de verilen karakteristik kullanıcılardan cu_1 ve cu_2 kullanılarak oluşturulmuştur. 7 adet küme üzerinde, başlangıç küme oluşumundan, son küme oluşumuna kadar tek tek saflık ve entropi hesaplaması yapılmıştır. Tabloya bakıldığında, küme sayısı arttıkça saflık değerinde de çoğunluk olarak artış olduğu gözlenmektedir. Entropi değerlerine bakıldığında, küme sayısındaki artış entropi değerinde bir düşme meydana getirmektedir. Bu sonuçlardan küme sayısında ne kadar büyük bir artış olursa, bu metrik değerlerinde de o derece inişler ve çıkışlar olacağı kanaatine varılabilir. Fakat tek ölçüt olarak bunun yanı sıra gap istatistiksel metodu da seçildiği için ideal küme sayısı 7 olarak kesinleşmiştir.

Bundan sonra yapılacak işlem, ideal küme sayısı belirlendikten sonra, nitelikler arasındaki önem sırasının belirlenmesidir. Bu önem sırasının belirlenmesi için ReliefF algoritması kullanılacaktır. Başarılı bir özellik seçme işlemi gerçekleştirmek için, en yakın komşu sayısının nasıl belirleneceği önemli bir problemdir. Bu çalışmada ideal en yakın komşu (k) değerinin belirlenmesi için, k değerini 10'dan 200'e kadar arttırarak ağırlık değerleri elde edilmeye çalışılmıştır. Şekil 4.12'ye göre k değeri 160 iken maksimum noktaya ulaşmaktadır. Bu noktadaki değer ideal k değeri olarak seçilmektedir. Bundan sonraki işlemde bulunan bu değer kullanılacaktır. k=160 iken nitelikler arasındaki önem sırası Şekil 4.13'te gösterilmektedir.



Şekil 4.12. *Değişen k Değerlerine Göre Ağırlıklar (Yaklaşım II)*



Şekil 4.13. *Her Bir Kümedeki Niteliklerin Ağırlıkları*

Şekil 4.13 incelendiğinde, parasal değer ve koltuk konforu niteliklerinin genel olarak önemli oldukları görülmektedir. Küme 1, 6 ve 7 için parasal değer en temel faktör iken küme 2, 3 ve 4 için koltuk konforu birinci sırada görülmektedir. Küme 2 ve 4 nitelik sıralaması olarak benzerlik göstermektedir. İkram ve eğlence niteliklerinin en son sıralarda yer aldığı görülmektedir. Buradan genel olarak yolcuların bu iki niteliğe önem vermedikleri ifadesi çıkarılabilir. Küme 5 için personel servisi ve ikram nitelikleri önem derecesi artarken parasal değerin ve koltuk konforu niteliklerinin önem sıralarında düşüş meydana gelmektedir.

5. SONUÇ VE ÖNERİLER

Geçmişte pazarlama, şirketlerin üretim portföyüne dayanmaktaydı. Şirketin ürettiği bir ürün var ise, bu ürünü satabileceği bir müşteri olmalıydı. Bunun için şirketler her ürettiği ürün için mutlaka bir müşteriye ihtiyaç duymaktadır. Bu nedenden dolayı şirketler ilk olarak yetenekleri doğrultusunda ürün üretip daha sonradan bu ürünü satın alacak müşterilere ulaşmaya çalışırdı. Zaman içerisinde bu kural, ürünün ne olduğu değil de bu ürünün kimin için olduğu şeklinde değişmiştir. Şimdilerde, şirketler ihtiyaçları karşılamayan ürünler üretmek yerine, müşteri odaklı olmuşlardır. Başarılı pazarlama gerçekleştirmek için, şirket yöneticileri müşteri tercihleri hakkında kullanışlı, yararlı bilgilere erişme ihtiyacı hissetmişlerdir. Eğer yöneticiler müşterileri tercih bilgilerini dikkate alıp, bunun farkındalığına varırlarsa, diğer şirketler üzerinde avantaj sağlamış olacaktır. Bu nedenlerden dolayı müşteri verilerinin elde edilip uygun bir şekilde analiz edilmesiyle farklı stratejiler geliştirilebilir. Geliştirilen bu stratejilerle firmalar verimliliklerini arttırabilirler.

Bu çalışmada yolcu yorumları veri analiz teknikleri kullanılarak irdelenmeye çalışıldı. Yolculara ait çoklu ölçüt oy verileri üzerinden kümeleme analizi ve özellik seçimi ile elde edilen bulgulardan çıkarımlar yapılmıştır. Bu bağlamda, yolcular arasında benzerlik ilişkisine ve benzerlik değerlendirme ölçütüne dayanan yeni kümeleme metotları öne sürülmüştür. Benzerlik ölçütlerine dayanarak elde edilen karakteristik kullanıcılarla oluşturulan kümelerde, özellik seçme yöntemi uygulanarak kullanışlı, yararlı sonuçlar elde edilmiştir. Öne sürülen ilk kümeleme metodu biraz daha geliştirerek yeni bir metot daha öne sürülmüştür. Geliştirilen bu metotta küme sayısının tespitinde istatistiksel metoda başvurulmuştur. Bu verilere kümeleme ve özellik seçilimi uygulanıp sonuçlar incelendiğinde, parasal değer ve personel servisi niteliklerin genel itibari ile ilk üç sırada yer aldıkları, koltuk konforu ve ikram özelliklerinin ise seyahat eden yolcular için pek önem arz etmediği görülmektedir. Elde edilen tüm bu sonuçlarla, havayolu müşteri ilişkileri kalitesini arttırmak için yolcu servis önceliklerini dikkate alıp tavsiyelerde bulunmak ve bu sayede havayolu firmalarının farklı taktikler geliştirebileceği çıkarımları yapılabilir.

Çalışmamızda yeni kümeleme yöntemleriyle birlikte var olan özellik seçimi ve diğer istatistiksel yöntem ve ölçütlerden yararlanılmıştır. Çalışma kapsamında önerilen yöntemler ve elde edilen bulgular yayın olarak araştırmacıların ilgisine sunulmuştur

(Yakut ve Türkoğlu, 2015, s. 1, Yakut ve Türkoğlu, 2016, s. 1). Gelecekte incelenebilecek konular hala mevcuttur. Örneğin, öne sürülen benzerlik tabanlı yöntemi geliştirmek için farklı yöntemler var olabilir. Kullanılan ölçütler ve değerlendirme kriterleri geliştirilip, kullanıcı oylarını kümelemede değerlendirilebilir. Ayrıca elde edilen sonuçlar, pazarlama ve yönetim açısından sosyal bilimcilerin için tartışmaya açıktır. Havayolu yolcu yorumlarına ek olarak, birçok alan için kullanıcı yorumları da ele alınabilir. Yorum verilerinin bu türleri için analiz metotları geliştirilerek ve geliştirilen bu metotlar kullanılarak, ilginç sonuçlar ortaya çıkarılabilir.

KAYNAKÇA

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. *Proceedings of the 20th International Conference on Very Large Databases*, USA, 487-499.

Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. *11th International Conference on Data Engineering*, Taipei, Taiwan, 3-14.

Agrawal, R., Imielinski, T. and Swami, A. (1993). Mining association rules between sets of items in large databases. *In Proceedings of the ACM SIGMOD International Conference on Management of Data*, Washington, USA, 207-216.

Arora, R., Gupta, M., Kapila, A., Fazel, M.(2011). Similarity-based clustering by left-stochastic matrix factorization. *In: ICML*, 761-768.

Balcan, M.F., Blum, A. and Vempala, S. (2008). Clustering via similarity functions: Theoretical foundations and algorithms. *Proceedings of the 40th ACM Symposium on Theory of Computing (STOC).*

Bülbül, S. ve Köse, A. (2009). Türk gıda şirketlerinin finansal performansının çok amaçlı karar verme yöntemleriyle değerlendirilmesi. *10. Ekonometri ve İstatistik Sempozyumu, Atatürk Üniversitesi.* Erzurum.

Dickson, P. R., (1982). Person–situation: Segmentation's missing link. *Journal of Marketing*, 46(4) , 56−64.

Finn, A., Wang, L. and Frank, T. (2009). Attribute perceptions, customer satisfaction and intention to recommend e-services. *Journal of Interactive Marketing.* 23, 209-220.

Gilliams, S., Raymaekers, D., Muys, B., Orsahovan, J.V. (2005). Comparing multiple criteria decision methods to extend a geographical information system on afforestation. *Computers and Electronics in Agriculture*, 49, 142-158.

Han, J., Kamber, M. and Pei, J. (2000). Data mining concepts and techniques. *Morgan Kaufmann Publishers*, 1st Ed., San Francisco, USA.

Han, J. and Kamber, M. (2006). Data mining concepts and techniques. *Morgan Kaufmann Publishers*, Second Ed., San Francisco, USA.

Hipp, J., Güntzer, U. and Nakhaezadeh, G. (2000), Algorithms for association rule mining- general survey and comparison. *ACM SIGKDD Explorations Newsletter*, 2(1), USA, 58-64.

Holtbrügge, D., Wilson, S. and Berg, N. (2006). Human resource management at Star Alliance: Pressures for standardization and differentiation. *Journal of Air Transport Management.* 12, 306-312.

Huang, A. (2008). Similarity measures for text document clustering. *Proceedings of the 6th New Zealand Computer Science Research Student Conference*.

Jarvis, R.A. and Patrick, E.A. (1973). Clustering using a similarity measure based on shared near neighbors. *IEEE Transaction on Computers*, C22, 1025-1034.

Kahraman, C., Cebeci, U. and Ruan, D. (2004), Multi-attribute comparison of catering service compines using fuzzy AHP: The Case of Turkey, *International Journal of Production Economics*, 87, 171-184.

Kalıkov, A. (2006). Veri madenciliği ve bir e-ticaret uygulaması. *Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü*.

Karakaşoğlu, N. (2008). Bulanık çok kriterli karar verme yöntemleri ve bir uygulama, *Yüksek Lisans Tezi, Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü, Denizli*.

Karypis, G., Han, E. and Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *IEEE Computer*, 32, 68-75.

Kaufman, L. and Rousseeuw, P.J. (1990). Finding groups in data: An introduction to cluster analysis. New York, John Wiley.

Ke-wu, Y., Jin-fu, Z. and Qiang, S.(2007). The application of ID3 algorithm in aviation marketing. *Proceedings of the IEEE International Conference on Grey Systems and Intelligent Services*, 1284-1288, Nanjing, China.

Kwan, I.S.Y., Fong, J. and Wong, H.K. (2005). An e-customer behavior model with online analytical mining for internet marketing planning. *Decision Support Systems*, 41, 189-204.

Lacic, E., Kowald, D., Lex, E. (2016). High Enough? Explaining and Predicting Traveler Satisfaction Using Airline Review. In Proceedings of the 27th ACM Conference on Hypertext and Social Media, Halifax, Canada.

Liou , J.J.H. and Tzeng, G.H. (2010). A dominance-based rough set approach to customer behaviour in the airline market. *Information Sciences*, 180, 2230-2238.

Lu, H. and Lin, J.C. (2002). Predicting customer behavior in the market-space: a study of Rayport and Sviokla's framework, *Information and Management*, 40, 1-10.

Migueis, V. L., Camanho, A.S. and e Cunha, J. F. (2012). Customer data mining for lifestyle segmentation. *Expert Systems with Applications.* 39, 9359-9366.

Miranda, H. S. and Henriques, R. (2013). Building clusters for CRM strategies by mining airlines customer data. *2013 8th Iberian Conference on Information Systems and Technologies (CISTI),* 1-5.

Özçakır, F.C. and Çamurcu, A.Y. (2007). Birliktelik kuralı yöntemi için bir veri madenciliği yazılımı tasarımı ve uygulaması. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 21-37.

Özekes, S. (2003). Veri Madenciliği Modelleri ve Uygulama Alanları. *İstanbul Ticaret Üniversitesi Dergisi, 2(3),* 65-82.

Perezgonzalez, J. D. and Gilbey, A. (2011). Predicting skytrax airport rankings from customer reviews. *Journal of Airport Management*. 5, 336.

Robnik Sikonja, M. and Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and Relief. *Machine Learning Journal*, 53, 23-69.

Sarvar, B., Karypis, G., Konstan, J., Rield, J. (2001). Item-based collaborative filtering recommendation algorithms. *Proceedings of the 10th international conference on the World Wide Web (ACM)*, 285-295.

Strehl, A., Ghosh, J. and Mooney, R. (2000). Impact of similarity measures on web-page clustering. *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, 58-64.

Singh, K., and Upadhyaya, S. (2012). Outlier Detection: Applications and Techniques. *International of Computer Science Issues*, 9, 307-323.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in data set via the gap statistic. *J.R. Statist. Soc. B*, 63, 411-423.

Vela, M. R., and Garcia, E. M. (2010). A Segmentation analysis and segments profile of budget air travelers. *Cuadernos de Turismo*, 26, 235-253.

Velu, C.M. and Kashwan, K. R.(2012). Pareto classification of data mining for customer relationship. *International Proceedings of Computer Science and Information Technology*, 37, 151-156.

Wang, H., Wang, W., Yang, J., Yu, P.S. (2002). Clustering by pattern similarity in large data sets. *SIGMOD 2002*, Madison, Wisconsin, USA, 394-405.

Xu, D.L., Yang, J.B. (2001), Introduction to multi-criteria decision making and the evidential reasoning approach, *Working Paper No.0206*, 1-21.

Yakut, I., Turkoglu, T. and Yakut, F. (2015). Understanding customers' evaluations through mining airline reviews. *Int. J. Data Mining and Knowledge Management Process*, 5, 1-11.

Yakut, I. and Turkoglu, T. (2016). Evaluating Multi-criteria Flight Reviews with Similarity-based Clustering. *6th World Conference on Soft Computing*, Berkeley, California, USA.

Yang, M.S. and Wu, K.L. (2004). A similarity-based robust clustering method. *IEEE Trans. Pattern Anal. Mach. Intell*., 434-448.

Zhao, Y. and Karypis, G. (2004). Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3).

Zhao, Y. and Karypis, G. (2002). Evaluation of hierarchical clustering algorithms for document datasets. *In Proceedings of the International Conference on Information and Knowledge Management*.