



**T.C. İSTANBUL T CARET
 NİVERSİTESİ**

Veri Madencilięi/E-Ticaret İ in  r n Tavsiye Sistemi Geliştirilmesi

Elif Şafak S VR 

Danışman

Yrd. Do . Mustafa Cem KASAPBAŐI

Eő Danışman

Do . Dr. Fethullah KARAB BER

**Y KSEK L SANS TEZİ
B LG SAYAR M HEND SL ę  ANAB L M DALI
 STANBUL - 2015**

KABUL VE ONAY SAYFASI

Elif Şafak SİVRİ tarafından hazırlanan " **Veri Madenciliği/E-Ticaret İçin Ürün Tavsiye Sistemi Geliştirilmesi**" adlı tez çalışması, aşağıdaki jüri üyeleri önünde İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü **Bilgisayar Mühendisliği Anabilim Dalı**'nda (YÜKSEK LİSANS veya ~~DOKTORA~~) **TEZİ** olarak başarı ile savunulmuştur.

Danışman

Yrd.Doç. Mustafa Cem KASAPBAŞI
İstanbul Ticaret Üniversitesi



Eş Danışman

Doç.Dr.Fethullah KARABİBER
Yıldız Teknik Üniversitesi



Jüri Üyesi

Yrd.Doç. Alper ÖZPINAR
İstanbul Ticaret Üniversitesi



Onay Tarihi : 30.09.2015

Prof. Dr. Doğan KAYA
Enstitü Müdürü

AKADEMİK VE ETİK KURALLARA UYGUNLUK BEYANI

İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmada,

- Tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- Görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- Başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- Atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- Kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- Bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

Tarih
30.09.2015

İmza


Tez Yazarının Adı Soyadı
Elif Şafak SİVRİ

İÇİNDEKİLER

Sayfa

İÇİNDEKİLER	I
ÖZET.....	II
ABSTRACT.....	İV
TEŞEKKÜR.....	V
ŞEKİLLER.....	VI
ÇİZELGELER.....	VII
SİMGELER VE KISALTMALAR.....	VIII
1. GİRİŞ	1
1. 1. VERİ MADENCİLİĞİNE GENEL BAKIŞ	1
1. 2. VERİ MADENCİLİĞİ KULLANIM ALANLARI	3
2. LİTERATÜR ÖZETİ	5
3. VERİ MADENCİLİĞİNDE KULLANILAN MODELLER VE TEKNİKLER ...	10
3. 1. TAHMİN EDİCİ MODELLER VE TEKNİKLER.....	10
3. 1. 1. Sınıflandırma	10
3. 1. 2. Regresyon analizi.....	11
3. 2. TANIMLAYICI MODELLER VE TEKNİKLER.....	14
3. 2. 1. Kümeleme analizi	14
3. 2. 1. 1. Kümeleme analizinde kullanılan uzaklık ölçütleri.....	15
3. 2. 2 Kümeleme yöntemleri.....	16
3. 2. 2. 1. Hiyerarşik kümeleme	16
3. 2. 2. 2. Hiyerarşik olmayan kümeleme	17
3. 2. 3. Birliktelik kurallarının tanımı	18
3. 2. 4. Birliktelik kurallarına örnek.....	19
3. 2. 5. Birliktelik kurallarında kullanılan terimler	19
3.2.6. Birliktelik kuralları performansı belirleyen durumlar	22
3.2.7. Birliktelik kurallarının belirlenmesinde kullanılan algoritmalar	23
3.2.7.1. Apriori algoritması	23
3.2.7.2. FP GROWTH algoritması.....	25
3. 2. 8. Benzerlik fonksiyonları (Jaccard, Cosine ve Pearson)	27
4. MARKET SEPET ANALİZİ VE BİRLİKTELİK KURALLARI İLE UYGULAMA.....	29
4. 1. MARKET SEPET ANALİZİ.....	29
4. 2. E-TİCARET PERAKENDE SİTESİNDE SEPET ANALİZİ ÖRNEĞİ	29
4. 3. VERİLERİN TOPLANMASI VE DÜZENLENMESİ.....	30
4.4. YÖNTEM VE ALGORİTMA.....	32
4. 5. AMAÇ	34
5. DEĞERLENDİRME.....	35
5. 1. BULGULAR.....	35
5. SONUÇ	48

5.1. SONUÇ VE ÖNERİLER	48
KAYNAKLAR	51
ÖZGEÇMİŞ	54

ÖZET

Yüksek Lisans Tezi

Veri Madenciliği/E-Ticaret İçin Ürün Tavsiye Sistemi Geliştirilmesi

Elif Şafak SİVRİ

İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Yrd. Doç. Dr. Mustafa Cem KASAPBAŞI

Eş Danışman: Doç. Dr. Fethullah KARABİBER
2015, 54 sayfa

Bu çalışmada,

Bir e-ticaret sitesi için, birliktelik analizi uygulaması ele alınmıştır. Hızlı değişen dünya koşullarından ötürü insanoğlunun ihtiyaçları gün geçtikçe artmakta ve değişmektedir. Bununla birlikte, insanların ihtiyaçları fazla zamanları kısıtlı olduğu için, e-ticaret alışverişinin kullanım ve tercih oranları da gün geçtikçe artmaktadır. Her 20 kişiden biri e-ticaret alışveriş sitelerini kullanarak, alışveriş yapmakta ve karşılıklarına alışveriş yaptıkları ürünler ile ilgili bir başka kişinin bu ürünle birlikte neler aldığı seçenekleri çıkmaktadır.

Özel bir şirkete ait e-ticaret perakende giyim verileri kullanılarak, yaş ve cinsiyet parametreleri de işin içine katılarak herhangi bir ürün satın alan müşterilerin başka hangi ürünleri satın aldıkları tespit edilmiştir. Birliktelik kuralları olarak Apriori ve FP-Growth algoritmaları kullanılmıştır. İki müşteri arasındaki benzerlik bulunurken Cosine, Pearson ve Jaccard benzerlikleri kullanılmıştır. Kümeleme analizi metotlarından k-ortalamlar kullanılarak; müşteriler yaş ve cinsiyet niteliklerine göre kümelendirilmiştir. Weka programı ile gerçek veriler üzerinde birliktelik kuralları elde edilmiştir ve bir kullanıcı ara yüzü ile benzerlik ve kümeleme analizleri yapılmıştır.

Anahtar Kelimeler: Apriori algoritması, FP-Growth algoritması, birliktelik kuralları, e-ticaret, veri madenciliği.

ABSTRACT

M.Sc. Thesis

Data Mining/Improving A Product Recommendation System For E-Commerce

Elif Şafak SİVRİ

**İstanbul Commerce University
Graduate School of Applied and Natural Sciences
Department of Computer Engineering**

Supervisor: Assist. Prof. Dr. Mustafa Cem KASAPBAŞI

**Co-Supervisor: Assoc. Prof. Dr. Fethullah KARABİBER
2015, 54 pages**

Since the world is changing the needs of the human beings are changing and increasing as day passes. E-commerce sites are becoming popular since they fit the contemporary daily life. One out of every twenty people uses e-commerce site for shopping and they are recommended with some products which other users also interested in or ordered.

In this study, Association analysis is carried out for an e-commerce site. With analyzing cloth sales data for the e-commerce site of a private Turkish retail company, concerning gender and age parameters, the association rules for which products are bought together is discovered. FP-Growth and Apriori are used for association rules. While discovering similarities between two customers Cosine, Pearson, and Jaccard are utilized. K-means is used as Clustering analysis, to cluster customers according age and gender properties of customers. WEKA software is used to discover the association rules in data and using a developed GUI clustering and association analysis are further performed.

Keywords: Apriori algorithms, FP-Growth algorithms, association rules, e-commerce, data mining.

TEŞEKKÜR

Bu araştırma için beni yönlendiren, karşılaştığım zorlukları bilgi ve tecrübesi ile aşmamda yardımcı olan değerli Danışman Hocam Yrd. Doç. Mustafa Cem KASAPBAŞI ve Doç. Dr. Fethullah KARABİBER'e teşekkürlerimi sunarım.

Tezimin uygulama aşamasındaki desteklerinden dolayı Hakan CENGİZ ve Dilek UÇAR'a teşekkür ederim.

Tezimin her aşamasında beni yalnız bırakmayan aileme sonsuz sevgi ve saygılarımı sunarım.

Elif Şafak SİVRİ
İSTANBUL, 2015

ŞEKİLLER

	Sayfa
Şekil 1. 1. Veri tabanı Bilgi Keşfi	3
Şekil 3. 1. FP-Growth Algoritmasının Yürütme Süreci.....	27
Şekil 4. 1. WEKA Datası Oluşumu.....	31
Şekil 4. 2. WEKA Arff Dosyası.....	32
Şekil 4. 3. WEKA Arff Datası	32
Şekil 4. 4. WEKA Giriş Ekranı.....	33
Şekil 4. 5. WEKA Data Yükleme Ekranı.....	33
Şekil 5. 1. Apriori Birliktelik Kuralları Sonuçları.....	35
Şekil 5. 2. FP Growth Birliktelik Kuralları Sonuçları.....	35
Şekil 5. 3. Jaccard Fonksiyonu ile Birliktelik Kuralları Sonuçları	37
Şekil 5. 4. Pearson Fonksiyonuna Birliktelik Kuralları Sonuçları	38
Şekil 5. 5. Pearson Fonksiyonu ile Birliktelik Kuralları Sonuçları.....	38

ÇİZELGELER

	Sayfa
Çizelge 3. 1. Küçük Veritabanı	26
Çizelge 5. 1. Apriori Algoritması İlk 20 Kural	36
Çizelge 5. 2. FP-Growth Algoritması İlk 20 Kural	36
Çizelge 5. 3. Apriori Algoritması ile Benzerlik Fonksiyonları Sonuçları.....	39
Çizelge 5. 4. FP-Growth Algoritması ile Benzerlik Fonksiyonları Sonuçları	41
Çizelge 5. 5. Apriori Algoritması ile Diğer Algoritma Karşılaştırmaları	43
Çizelge 5. 6. FP-Growth Algoritması ile Diğer Algoritma Karşılaştırmaları	44
Çizelge 5. 7. Yaş ve Cinsiyete Göre Kümeleme	44
Çizelge 5. 8. Yaş ve Cinsiyete Göre En Çok Alınan Ürün	45
Çizelge 5. 9. 25 Yaş Erkek Müşteri Benzerlik Fonksiyonu Sonuçları.....	45
Çizelge 5. 10. 25 Yaş Kadın Müşteri Benzerlik Fonksiyonu Sonuçları	46
Çizelge 5. 11. Analiz Çalışma Testi.....	46

SİMGELER VE KISALTMALAR

GA	Genetik Algoritmalar
VM	Veri Madenciliği
VT	Veri Tabanı

1. GİRİŞ

1. 1. Veri Madenciliğine Genel Bakış

Günümüzde ileri teknoloji sayesinde birçok alanda veriler toplanıp, yığınlar halinde depolanabilmektedir. Böylesine büyük ve gün geçtikçe büyüyen bu veri yığınının bir anlam çıkarılmaya çalışılmaktadır (Gürgen, 2008). Olası anlamların, örüntülerin saptanması büyük boyutlu veriler için insan eliyle işlenmesi mümkün olmamaktadır (Aras, 2010). Firmaların ya da sektörlerin bu veri yığınlarından insan eliyle işlemesiyle anlam ya da örüntü çıkarmasının mümkün olmama durumu, veri madenciliğini ve veri tabanlarında bilgi keşfini ortaya çıkarmaktadır.

Veri madenciliği için birçok farklı tanım yapılmaktadır. Bunlardan en çok kullanılanlar şu şekildedir;

- Veri madenciliği; büyük veri yığınlarından bilgisayar programlarını kullanarak gelecekle ilgili tahmin yapabilecek bağıntı ve kuralların çıkarılmasıdır.
- Veri madenciliği; veride aranacak ya da bulunacak sorun ya da çözümlerin yarı veya tam otomatik bir sistem ile keşfedilmesidir.
- Veri madenciliği; veri yığınlarına uygulanan çeşitli analizler ile bilgiye ulaşmaktır.

Yukarıdaki tanımlar incelendiğinde; bu tanımların iki ortak paydada buluştuğunu görüyoruz; 1) Bilgi elde etmek, 2) Çok fazla verinin, veri tabanlarında depolanması.

Veri Madenciliğinde bilgi keşfi; var olan verilerden doğru, işe yarayacak, yeni bilgiler, örüntüler ve modeller elde etmekte kullanılan özel bir süreçtir. Şekil 1.1’de veri madenciliği süreci gösterilmiştir.

Bu tez çalışmasının birinci bölümünde, veri madenciliği tanımı, ne amaçla kullanıldığı, veriye uygulanan ön işlemler, çalışmada geçen diğer bölümlerin açıklamaları ve daha önce vm ile yapılmış diğer çalışmalardan bahsedilecektir.

Çalışmanın ikinci bölümünde; bir alışveriş sitesine ait, müşteri alışveriş kayıtlarından aldıkları ürünler ve daha sonra alabilme potansiyeli olabilecek ürünleri tespit etmek

amaçlı, veriye uygulanacak algoritmaların açıklamaları, nasıl çalıştıkları yer almaktadır.

Çalışmanın üçüncü bölümü yani uygulama kısmında; veriden sağlıklı ve doğru bilgi çıkarabilmek için yapılan işlemlerin sırası ile anlatımı bulunmaktadır.

Bu adımlar ve adımlarda gerçekleşen işlemlerin özeti aşağıdaki gibidir;

- Veri Seçimi
- Veri Ön işleme
- Veri Dönüştürme
- Veri Madenciliği Tekniği
- Sonuç Değerlendirme (Gürgen, 2008)

Veri seçimi basamağında, öncelikle elde edilmek istenen bilgi tespit edilmiş daha sonra bu bilgiyi bize sunabilecek bir veri kümesi seçilmiştir.

Veri ön işleme basamağında; veride aynı ürünü ifade eden ama farklı isimlerle belirtilen ürün çeşitleri tek bir isim altında birleştirilmiştir. Bunun amacı aynı manaya gelen ama farklı adlandırılan ürün isimlerini aynı ana grup altında toplamak ve çıkacak birliktelik kurallarının daha sağlıklı olmasını sağlamaktır. Kümeleme analizinde müşterinin yaşını kullanabilmek için; doğum tarihinden, müşterinin yaşını çıkarmak gibi veri ön işleme basamakları uygulanmıştır.

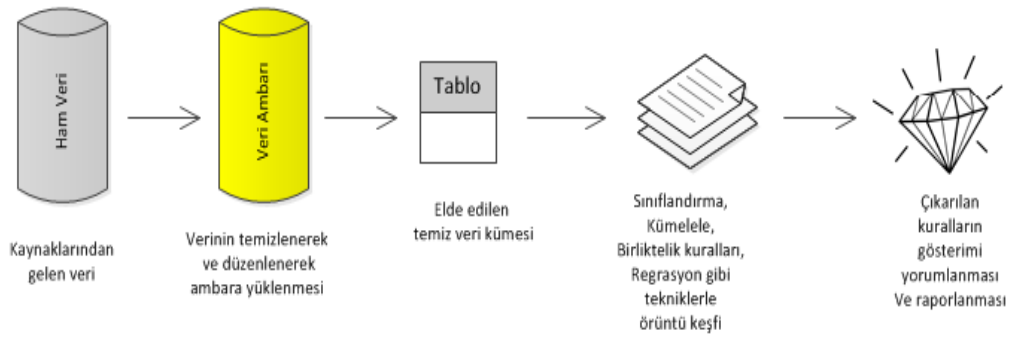
Veri dönüştürme basamağında; birliktelik kuralları elde edebilmek için, veri analizinin gerçekleştirileceği programda ve seçilen algoritmaya göre dönüştürme işlemi gerçekleştirilmiştir. Bu basamakta birliktelik kurallarını çıkarmada kullanılan WEKA programında, analizi gerçekleştirebilmek için verideki ürünlerin kolonlara göre sıralanması ve alınan ürünler ile alınmayan ürünlere değer atama işlemleri yapılarak, veri analize hazırlanmıştır.

Veri madenciliği tekniği basamağında, veriye Apriori ve FP-Growth algoritmaları uygulanarak birliktelik kuralları tespit edilmiştir. Amaç hangi ürünü alanlar hangi ürün ya da ürünleri de almıştır bilgisini çıkarabilmektir. İlk birliktelik kuralı çıkarılırken müşterinin yaş ya da cinsiyeti göz önünde bulundurulmadan analiz uygulanmıştır. Analizin ikinci aşamasında, C# programında yazılan program ile benzerlik fonksiyonları (Cosine, Jaccard, Pearson) kullanılarak birliktelik kuralları

tespit edilmiş; Apriori ve FP-Growth algoritmaları ile bulunan sonuçlarla karşılaştırılmıştır. Üçüncü aşamasında ise, yine C# programında yazılan programdan faydalanılarak, kümeleme yapılıp, hangi yaş grubu ve hangi cinsiyet kümesinde bulunanların; hangi ürünü, hangi ürün ya da ürünler ile birlikte aldıkları tespit edilmiştir.

Sonuç değerlendirme basamağı olan tezin dördüncü bölümünde, yapılan analiz sonuçlarının sağlanması gerçekleştirilmiştir. Analizin yüzde kaç doğrulukla çalıştığı belirlenmiştir.

Tez çalışmasının son kısmı olan sonuç bölümünde, elde edilen sonuçlar, bilgiler ve bu sonuçların nerede, hangi amaç için kullanılacağı belirtilmiştir.



Şekil 1. 1. Veri Tabanı Bilgi Keşfi (Uygun, 2014)

1. 2. Veri Madenciliği Kullanım Alanları

- Veri tabanı analizi ve karar verme desteği,
- Hedef pazar, müşteriler arası benzerliklerin saptanması, sepet analizi, çapraz pazar incelemesi,
- Kalite kontrol, rekabet analizi, öngörü, sahtekârlıkların saptanması,
- Belgeler arası benzerlik: haber kümeleri, e-posta,
- Müşteri kredi risk araştırmaları,
- Kurum kaynaklarının en optimal biçimde kullanımı,
- Geçmiş ve mevcut yapı analiz edilerek geleceğe yönelik tahminlerd bulunma,
- Müşterilerin satın alma örüntülerinin belirlenmesi,
- Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması,

- Posta kampanyalarında cevap verme oranının artırılması,
- Mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması,
- Pazar sepeti analizi (*Market Basket Analysis*),
- Müşteri ilişkileri yönetimi (*Customer Relationship Management*),
- Müşteri değerlendirme (*Customer Value Analysis*),
- Satış tahmini (*Sales Forecasting*),
- Farklı finansal göstergeler arasında gizli korelasyonların bulunması,
- Kredi kartı dolandırıcılıklarının tespiti,
- Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi,
- Kredi taleplerinin değerlendirilmesi,
- Yeni poliçe talep edecek müşterilerin tahmin edilmesi,
- Sigorta dolandırıcılıklarının tespiti,

2. LİTERATÜR ÖZETİ

Bu bölüm, tez çalışması ile ilgili konu olan tavsiye sistemleri ve birliktelik kuramlarının şu ana kadar hangi konularda ve nasıl uygulandığının, öne çıkan yayınlarla incelenmesini içermektedir. Alışveriş gerçekleştiren müşterilerin kayıtlarında SPSS Clementine kullanılarak, k-ortalamlar yöntemi ve birliktelik kuralları analizi yapılmıştır. K-ortalamlar yöntemi ile müşteri profili ortaya çıkarılmış ve Apriori algoritmasıyla müşterinin satın alma profilleri ortaya çıkartıldığı belirtilmiştir (Timör ve Ezerçe, 2011).

Bayesian Network ve birliktelik kuralları kullanılarak bir tavsiye sistemi geliştirilmiştir. Bu sistemin hedefi, müşterinin alışveriş anında aldığı ürünler ile alakalı ürünleri tavsiye etmek olarak belirtilmiştir (Weng vd. (2011)).

Öğrencilerin staj verileri üzerine uygulanan Apriori Algoritması ile birliktelik kurallarının çıkarılması ve staj eğiliminin belirlenmesi adlı makalede; Üniversite veri tabanına kayıtlı öğrenim görmüş ya da gören öğrencilerin, zorunlu donanım staj kayıt bilgileri kullanılarak, öğrencilerin staj eğilimleri belirlenmeye çalışılmıştır. Çalışmada SPSS Clementine programı kullanılarak CRISP-DM metodundan yararlanılmıştır. Elde edilen bilgilerden staj yapan öğrenciler hakkında bilgi toplanması hedeflenmiştir. Elde edilen sonuçlarla bölümdeki staj stratejileri hakkında alınan kararlara destek olduğu olarak belirtilmiştir (Taş vd. (2013)).

Apriori algoritması ile öğrenci başarıları analizi adlı makalede; üniversitenin öğrenci notları kullanılarak başarı analizi çıkarılmaya çalışılmıştır. Veriler, sql server veri tabanından elde edilmiş ve analize uygun şekilde, gerekli veri dönüşümleri yapılmıştır. Ardından MATLAB programı kullanılarak Apriori algoritması uygulanarak sonuçların elde edildiği belirtilmiştir (Karabatak ve İnce, 2015).

Birliktelik kuralları ile Van Gölü için mekansal-zamansal veri madenciliği isimli tezde; Van Gölü'ne ait geçmiş yıllardaki göl seviyesi, göle düşen yağış miktarı, göl yüzeyindeki buharlaşma miktarı gibi değerler kullanılarak Apriori algoritması ve mekansal-zamansal veri madenciliği uygulamaları gerçekleştirilmiş. Çalışmanın amacı gölün kıyı kısımlarında oluşan taşmalar ve yeni kıyı arazilerini tespit edilebilmeyi amaçlamış. Bu tespitin sonucunda da yapılabilecek önlemler bulunmaya çalışıldığı olarak belirtilmiştir (Alaeddinoğlu, 2012).

Metin madenciliği yöntemiyle Türkçede en sık kullanılan ve birbirini takip eden harflerin analizi ve birliktelik kuralları isimli tezde; C# program kullanılarak, Türkçe metinlerden sık kullanılan harfler, iki tane harfin birbirini takip etme sıklığı, üç harfin birbirini takip etme sıklıkları ortaya çıkarılmış. Apriori algoritması ile AIS ve SETM algoritmaları karşılaştırılmış. Apriori algoritmasının performans açısından da daha iyi olduğu belirtilmiştir (Çelikyay, 2010).

Birliktelik kuralları ile web siteleri için tavsiye motoru uygulaması isimli tezde; web sitesi ziyaretçilerin tıklama oranları yüksek olanların, tespit edilmesi ve tıklama oranlarının saptanması hedeflenen tez çalışmasında C# programı kullanılmış. Bu amacı sağlamak için birliktelik kuralları kullanıldığı belirtilmiştir (Aras, 2010).

Birliktelik kuralları ile sepet analizi ve uygulaması isimli tezde; Türkiye'deki market zincirlerinden birinin 7 günlük alışveriş fişleri kullanılmıştır. Fişlerdeki ürünler arası ilişki SPSS Clementine programı kullanılarak apriori algoritması uygulanmıştır. Ürünler arasında bulunan ilişkiler, market gelirinin artması amacı için kullanılacağı belirtilmiştir (Gürgen, 2008).

Müşteri ilişkileri yönetimi (Customer Relationship Management- CRM) ve veri madenciliği (data mining): tekstil sektöründe bir uygulama isimli tezde; SPSS Clementine programı kullanılarak, hazır giyim perakende sektörüne ait bir firmadan alışveriş yapan müşterilerine birliktelik kuralı analizi uygulanarak alışveriş alışkanlıkları, kümeleme analizi uygulanarak müşterilerin demografik özellikleri ve karar ağacı uygulanarak alışveriş alışkanlıklarını en çok etkileyen değişkenler tespit edilmesi hedeflenmiştir. Yapılan çalışma ile ilgili firmanın CRM departmanına anlatıldığı belirtilmiştir (Ezerçe, 2008).

Veri madenciliği ve Apriori algoritması ile süpermarket analizi isimli tezde, SPSS Clementine programı kullanılarak bir süpermarkete ait alışveriş kayıtlarındaki birlikte satın alma bilgilerini tespit etmektir. Bu bilgiler doğrultusunda süpermarketteki ilgili ürünler arasında yeni bir yerleşim düzeni önerisi planlanmıştır. Analiz için Apriori algoritması ve veri düzenlemeleri için SQL kullanıldığı belirtilmiştir (Ay, 2009).

Otomobil yetkili servislerinde birliktelik kuralları belirlenmesinde Apriori ve FP-Growth algoritmalarının karşılaştırılması isimli makalede, otomotiv sektöründeki bir yetkili servise ait müşterilerin alışveriş bilgileri ile Apriori ve FP-Growth Algoritmaları kullanılarak analizler yapılmıştır. Sonuç olarak firma karını arttırmayı amaçlayan kampanya ve promosyonlar yapılması hedeflenmiştir. Çalışma WEKA ve Excel programları yardım ile yürütüldüğü belirtilmiştir (Erpolot, 2015).

Sigortacılık sektöründe müşteri ilişkileri yönetimi için birliktelik kuralı kullanılması isimli çalışmada, bir sigorta firmasına ait müşteri verileri Apriori algoritması kullanılarak analiz edilmiştir. Verilerin hazırlanması için Oracle veri tabanı yönetim sistemi kullanılmıştır. Birliktelik kuralı analizi için SPSS Clementine programı kullanılmıştır. Analiz sonucunda müşterilerin daha çok hangi ürün gruplarını bir arada aldığı tespit edilmiş ve bu sonuçlar doğrultusunda daha etkin geri dönüş alınabilecek kampanyalar ve pazarlama stratejileri geliştirilmesinin mümkün olduğu belirtilmiştir (Doğan vd. (2014)).

Ege Bölgesi'ndeki bir araştırma ve uygulama hastanesinin acil hasta verilerinin veri madenciliği ile analiz edilmesi isimli makalede; belirli bir dönemde bir hastanenin acil servisine başvuran hastaların verileri kullanılarak birliktelik kuralları analizi ile veri setindeki anlamlı ilişkiler ortaya çıkarılmaya çalışılmıştır. Çalışma sonucunda hastaların başvuru nedenleri ve hasta profilleri ile ilgili bilgiler ortaya çıkmıştır. Bu bilgiler, acil servis bölümlerinin yeniden yapılandırma çalışmalarına da farklı bir açıdan yol göstermek amaçlı katkıda bulunduğu olarak belirtilmiştir (Erdem ve Özdağoglu, 2008).

Yazılan makalede, bir Web sunucusunda toplanan veriler ile birliktelik algoritmaları karşılaştırılmıştır. Bunlardan bazıları; IP adresleri, sayfa referansları ve kullanıcıların erişim süreleri bilgileri ile çalışma yürütülmüştür. Bu araştırma çalışması özellikle sunucu log dosyalarından web sitelerinin kullanım kalıplarını keşfetmek ve web kullanım madenciliği üzerinde yoğunlaşmaktadır. Apriori algoritması ve Sık Desen (FP) Büyüme algoritması kullanılarak bellek kullanımı ve zaman kullanımı karşılaştırıldığı belirtilmiştir (Kumar ve Rukmani, 2010).

Yazılan makalede; Apriori ve FP büyümenin temsili aritmetiği anatomiksel olarak sık desen öge kümesinin madencilik sürecini açıklar. FP ağacının inşaa yöntemi

ortaya konulmuş ve ilişkisel madencilik kurallarını nasıl etkilediği tartışılmıştır. Deneysel sonuçlar bu algoritmanın Apriori gibi mevcut birçok örneğine göre yürütme zamanı, bellek kullanımı ve CPU kullanımı gibi konularda daha yüksek madencilik verimliliğine sahip olduğu olarak belirtilmiştir (Zhang vd.(2008)).

Türkiye’de tarımsal kooperatifçiliğin iller açısından çok boyutlu analizi isimli makalede; kooperatif üye ve yöneticilerine yapılan anketlerden elde edilen verilerle yapılmıştır. Analizde tanımlayıcı istatistikler, çok boyutlu ölçekleme analizi ve kümeleme analizi uygulanmıştır. Çok boyutlu ölçekleme ve kümeleme analizine göre, iller, kooperatif ortakları, işletme özellikleri, yönetici özellikleri ve kooperatif özellikleri bakımından farklılaştığının tespit edildiği olarak belirtmiştir (Şahin vd. (2013)).

Plastik boru sektörünün verileri kullanılarak veri madenciliği teknikleriyle bir makale yazılmıştır. Bu sektördeki satış verilerinden yararlanılarak satılan ürünlerin arasında birliktelik kuralı olup olmadığını araştırmaya çalışmıştır. Bu kapsamda, birliktelik kuralları yöntemi kullanılmıştır ve analizler mevcut veriler kullanılarak Apriori algoritması ile WEKA’da yapılmıştır. Bu analizlerin sonucunda satılan ürünler arasında çeşitli birliktelik kuralları olduğu tespit edildiği belirtilmiştir (Demir, 2015).

Huang tarafından 2007 yılında yazılan makalede, internet üzerinde bireyselleştirilmiş bir alışveriş sitesi tasarımında FP-Growth ve Apriori algoritmalarının uygulaması hedeflenmiştir. FP-growth algoritması olarak adlandırılan gelişmiş bir Apriori algoritması tanıtılmaktadır. Bu algoritma Apriori algoritmasında darboğaza yol açan iki temel soruna çözüm getirmektedir ve orijinale göre daha verimli olduğunu öne sürmektedir. Son olarak, ünlü bir iş web sitesinden alınan büyük tüketim verisi üzerinden FP-growth algoritması uygulanarak İnternet üzerinde bireysel sanal mağaza dizayn edilmektedir. İnternet üzerinde bireyselleştirilmiş bir sanal mağaza FP-growth algoritması uygulanarak bireyselleştirilmiş sanal mağaza tasarımı oluşturulması amaçlanmıştır. Bu makaledeki, araştırma sonucu sadece referans için kullanıldığı belirtilmiştir (Huang, 2007).

Yazılan makalede araştırma ve lisansüstü öğrencileri için veri madenciliği çalışmalarını kolaylaştıran veri madenciliği Apriori algoritmasına dayalı eğitim odaklı bir uygulama geliştirmesi amaçlanmaktadır. Uygulama veri külliyyatının ilginç

desenlerini bulmak için ve sorun kısıtlamalarının bir fonksiyonu olarak (destek ve güven değişkenlerinin değeri) veya işlem veritabanının boyutuna yürütme hızına etkisini ölçmek için kullanılacağı belirtilmiştir. Çalışma, Apriori algoritmasının kısa bir özetini, adım-adım süreci kullanarak algoritma uygulanması konusunda tüm yönleri, eğitim odaklı kullanıcı ara yüzü hakkında bir tartışmayı ve bir test işlem veritabanına ait veri madenciliği sürecini kapsamaktadır. Algoritmanın hızı üzerinde bazı kısıtlamaların etkisi yürütme hızını artırmak için farklı yaklaşımların sistematığı gözden geçirilmeden deneysel bir yolla ölçülmüştür. Uygulamanın olası kullanım alanları, bunun yanı sıra sınırları kısaca gözden geçirilmiştir (Roşca ve Radoiu, 2015).

3. VERİ MADENCİLİĞİNDE KULLANILAN MODELLER VE TEKNİKLER

3. 1. Tahmin Edici Modeller ve Teknikler

Tahmin edici modeller; mevcut değerleri kullanarak bir model geliştirir. Bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuçların tahmin edilmesi işlemidir.

3. 1. 1. Sınıflandırma

Sınıflandırma, veri madenciliğinde çok sık kullanılan yöntemlerden biridir. Sınıflandırma denildiği zaman en basit örnekler; canlı türlerine göre, elementlere göre sınıflara ayırmak gibi düşünülebilir. Sınıflandırma işlemi; nereye ait olduğu bilinmeyen bir nesnenin özelliklerinin incelenip, özellikleri belli bir gruba atanması işlemi olarak tanımlanabilir.

Sınıflandırma işlemi veriyi, önceden belirlenmiş sınıf ya da sınıflara dahil eder. Bir niteliğin değerini diğer nitelikleri kullanarak belirler. Sınıflandırılacak veriler, veri tabanına eklenecek yeni kayıtlar şekline dönüştürülebilir. Sınıflandırma işi sonrasında, veri tabanındaki tabloya yeni bir sınıf koduyla yeni bir kolon eklenmiş olunur. Sınıflandırma yönteminde önemli olan verilerin eğitim veri kümesi olmasıdır. Sınıflandırma yöntemindeki amaç, sınıfı belirli olmayan bir veri geldiğinde, bu verinin hangi sınıfa ait olduğunu bulabilecek bir model oluşturmaktır.

Bu modelde olması lazım gelen değişkenler bağımlı değişken ve diğerleri tahmin edici (bağımsız) değişkenler olarak adlandırılır. Kurulan modeldeki amaç, girdi olarak tahmin edici değişkenlerin yer aldığı modelde, sonucun yani bağımlı değişkenin değerinin bulunduğu anlamlı bir model kurabilmektir. Eğer bağımlı değişken sayısı sayısal değilse bu durumda problem sınıflama problemidir ama bağımlı değişkenin sayısal olduğu yerlerde problem regresyon problemine girer (Ergün, 2008).

Aşağıda birkaç sınıflama örneği verilmiştir;

- Tümör hücreleri iyi huylu mu, kötü huylu mu tahmin edip sınıflandırmak,
- Kredi kartıyla alışveriş yapan kişi; kart sahibi mi, sahtekâr mı, ?

- Proteinin yapısını öğrenmeye çalışmak,
- İnternetteki karışık haberlerin, spor mu, politika mı, magazin mi sınıflandırması

Sınıflandırma yöntemi ayrıca bir ürünün özellikleri ile müşterinin özelliklerinin eşlenmesi içinde kullanılabilir. Böylece bir ürün için ideal müşteri profili ya da müşteri profili için ideal ürün belirlenebilir.

Örneğin, genç bayanlar küçük araba satın alır; yaşlı ve zengin erkekler büyük ve lüks araba satın alırlar. Böyle bir sınıflandırma modelinden firmalar pazarlama stratejisi olarak yararlanabilir, ilgili kitleye daha rahat ulaşılabilir ve kar oranlarını daha rahat yükseltebilirler (Gürgeç, 2008).

3. 1. 2. Regresyon analizi

İlişki analizi (relationship analysis), regresyon analizi (regression analysis) ya da tahmin teknikleri (forecasting analysis) ismiyle geçen bu tekniğin iki temel görevi vardır; bunlardan birincisi tahmin, diğeri ise karar vericiye yol seçmekte yardımcı olmaktır.

Regresyon analizi, herhangi bir değişkenin (bağımlı değişken) bir veya birden fazla değişkenle (bağımsız değişken) arasındaki ilişkinin matematiksel bir fonksiyon şeklinde yazılmasıdır. Bu fonksiyona regresyon denklemi denir (Orhunbilge, 2002).

Sınıflandırma ile regresyon arasındaki farka bakılacak olunursa; sınıflandırma kategorik değerleri tahmin eder, belli özelliklere göre gruplanmış değerleri tahmin eder. Regresyon ise süreklilik gösteren değerlerin tahmin edilmesinde kullanılır.

Regresyon analizine örnek vermek gerekirse; 10 yıl boyunca 6 Kasım günü hava durumuna bakılır. Önümüzdeki sene de, hava böyle olacak mı tahmin edilir. Sınıflandırma ve regresyon tekniklerinin karşılaştırmasını bir örnek ile açıklamak gerekirse; sınıflandırma modelinde sigorta uygulamalarının güvenli veya riskli olmalarını kategorize etmek amacıyla kurulurken, regresyon modeli yaşı ve cinsiyeti verilen potansiyel müşterilerin ev ürünleri alırken yapacakları harcamaları tahmin etmek için kullanılır (Özgeç, 2003).

Sınıflandırma ve regresyon modellerinde kullanılan tekniklerden bazıları aşağıdaki gibidir;

- Karar Ağaçları

Karar ağaçları tahmin edici bir tekniktir. Uygulanması fazla maliyet gerektirmediği ve güvenilirlikleri iyi seviyede olduğu için veri madenciliğinde geniş bir kullanım alanına sahiptir (Arabacı, 2007). Karar ağacı; kök, karar düğümleri, dallar ve yapraklardan oluşmaktadır. Karar ağacında kökten yapraklara doğru gidilerek kurallar yazılabilir. Bu kurallar if-then rules olarak da geçmektedir. Karar ağaçlarında her dal kendinden sonra evet ya da hayır şeklinde 2 dala ayrılmaktadır. Her dalın belirli bir olasılığı vardır.

Karar düğümleri yapılacak testleri temsil etmektedir. Verinin ağaç dallarında ayrılması bu testin sonucuna göre belirlenir. Bütün düğümlerde sırasıyla ayrılma işlemi gerçekleşir. Eğer dalın ucunda bir sınıflandırma işlemi yapılmıyorsa, karar düğümü oluşur. Gruba ayrılma sonucunda sınıf oluşuyorsa, dalın sonunda yaprak vardır. Yaprak, veri üzerinde belirlenecek olan sınıfı temsil etmektedir. Karar ağaçlarında işlem sırası kök düğümünden başlayıp yaprağa ulaşana kadar devam etmektedir (Arabacı, 2007).

Karar ağaçları sınıflandırmanın genel özelliklerini taşımaktadır. İlk önce öğrenme daha sonra model oluşturulur.

- Yapay Sinir Ağları (YSA)

Genelleme, tahmin etmek, sınıflandırma gibi kabiliyetlerinden dolayı veri madenciliği alanında yaygın olarak kullanılmaktadır. Yapay sinir ağları, oldukça yüksek sınıflandırma doğrulukları elde etmelerine rağmen elde ettikleri bilgileri kullanıcıların anlayabileceği şekilde sunamamaktadır ve tanımlama kabiliyetleri oldukça sınırlıdır. Bu nedenle günümüzdeki veri madenciliğinde yapay sinir ağlarının kullanımında eğilim, yapay sinir ağlarından kural çıkarımına yöneliktir. Tanımlama kabiliyetlerinin sınırlı olma nedenlerinin en temel nedeni; YSA’larda gizli olan bilginin aktivasyonlar ve nöron bağıntılarına dağılmış olmasıdır (Özbakır ve Baykasoğlu, 2011).

- Genetik Algoritmalar ve Evrimsel Algoritmalar

Evrimsel algoritmalar; Evrim Teorisine dayanan, popülasyon tabanlı bir algoritma olup, Charles Darwin tarafından ortaya konulmuştur. Evrimsel algoritmaların temel olarak çalışma sıralaması; birinci adımda, popülasyondaki birey kalitelerini belirlemek amacıyla uygunluk fonksiyonun belirlenmesi, ardından yeni popülasyonları belirleyecek dizinin seçilmesi ve operatörlerin tanımlanması şeklindedir. Evrimsel algoritmaların mantığına göre çözülecek problem sanki doğanın kendisiymiş gibi kabul edilir ve ayrıca problem içerisinde yer alan her bir birey doğadaki bir canlıymış gibi düşünülür. Bu şekilde kabul edilmesinin amacı, doğaya en iyi şekilde uyum sağlayabilen bireyleri bulabilmektir (Şahin, 2014).

Genetik algoritmalar Evrimsel Algoritmaların bir bölümüdür. Evrimsel algoritmalar yapay zekânın bir alt bölümüdür. 1960 yıllarda I. Rechenberg'in "Evrimsel Stratejileri" çalışmasıyla ortaya çıkmıştır. Evrimsel hesaplamalardaki amaç; en iyi olan durumun gerçekleşebilmesi için gerçek hayattaki problemlere uygulamaya çalışmaktır. Genetik algoritmalar daha çok optimizasyon ve makine öğrenmesi alanlarında sıkça kullanılmıştır.

Genetik algoritmalar matematiksel model bilgilerini kullanmak yerine sadece giriş-çıkış bilgilerini kullanır. Bu çalışma şekli Kompleks problemlerin optimizasyonunda bir avantaj sağlamaktadır ve bu tarz karışık problemlerin çözümünde tercih edilen bir yöntemdir. Genetik algoritmalar büyük parametreleri tolere edebilen bir teknik olmasıyla birlikte değerleri önceden bilinmeyen parametrelerin kontrol sistemlerinde de çok güzel sonuçlar verebilmektedir.

Genetik algoritmaları diğer yöntemlerden ayıran avantaj ve dezavantajlar aşağıdaki gibi sıralanmaktadır;

- GA, birçok arama noktası üzerinde ve arama uzayında global bir arama yapar,
- İşlemler olasılıksal kurallar üzerine yapılır, olabilme ihtimali olan durumlar sadece algoritmada yer alır.
- GA'lar matematiksel modeller kullanmak yerine, bireylerin uygunluk değerlerinin bulabilmek için arama uzayında uygunluk fonksiyonu kullanmaktadır. Matematiksel modelleri kullanmaya ihtiyaç duymaz.

Çünkü karmaşık işlemler yapmadığından, türevsel bilgilere de gerek duymamaktadır.

- Ayrık ve sürekli parametreleri optimize edebilme ve çok sayıdaki parametre ile çalışabilme yeteneği bulunmaktadır.
- Birden fazla parametrenin optimum çözümünü yapabilmektedir (Tuğ, 2005).

3. 2.Tanımlayıcı Modeller ve Teknikler

Tanımlayıcı modeller, karar vermeye yardımcı olacak verilerdeki, örüntüleri tanımlamada kullanılır.

3. 2. 1. Kümeleme analizi

Kümeleme analizi, gruplanmış verileri benzerliklerine göre sınıflandırmada sıklıkla kullanılan bir yöntemdir. Amacı; birey ya da nesnelerin temel özelliklerini dikkate alarak onları gruplandırmaktır. Diğer bir deyişle kümeleme analizi; gruplanmış verileri benzerliklerine göre gruplandırarak, araştırmacıya özetleyici bilgiler sunmaktır. Kümeleme analizi, araştırmada gözlenen bireylerin ya da nesnelerin ölçülen tüm değişkenler üzerindeki değerlerini hesaplayarak ortaya çıkacak kümelere veya gruplara odaklanmaktadır. Bireyler veya nesneler arasındaki benzerlikleri saptamak amacıyla uzaklık ölçüleri, korelasyon ölçüleri veya nitelik verilerinin benzerlik ölçüleri kullanılmaktadır (Kalaycı, 2009).

Kümeleme işleminde dikkat edilmesi gereken en önemli unsur; sınıf içi benzerliğin maksimum, kümeler arası benzerliğin ise minimum olmasıdır.

Kümeleme analizi temel olarak belli başlı amaçlara yönelik uygulanan bir yöntemdir. Bu temel amaçlar şu şekildedir;

- x sayıdaki nesnenin y sayıdaki değişkene göre, belirlenen özelliklerine göre kendi içinde homojen ve kendi aralarında heterojen kümelere ayırmak için,
- x sayıdaki nesneyi, y sayıdaki değişkene göre, ortak faktörlerini belirlemek için,

- Nesneler ve değişkenler birlikte değerlendirildiğinde x nesnesini y değişkenine göre ortak olacak alt kümeler ayırmak amacıyla, kullanılmaktadır.

Değişken sayısı ve gözlem sayısı çok fazla olduğu zaman, tanımlama yapmak zor olmaktadır. Bu tarz veri setlerinde bilgiye ulaşabilmek için eldeki gözlemleri gruplara ayırmamız gerekmektedir. Gruplama işleminde dikkat edilmesi gereken unsur; yapılan gruplamada gruplar kendi içinde homojen özellik gösterirken, gruplar arasında hetorejen bir özellik olması gerekir. Bu işlem belli miktarda bilgi kaybına sebebiyet verebilir ama kümeleme analizi sayesinde bu bilgi kaybı minimum seviyeye indirgenmiş olur.

3. 2. 1. 1. Kümeleme analizinde kullanılan uzaklık ölçütleri

Kümeleme analizindeki esas unsur, nesneler arasındaki benzerlikleri ya da uzaklıkları buma dayanmaktadır. Benzerlik ile uzaklık birbiri ile ters orantılı ölçütlerdir. Eğer iki nesne arasındaki benzerlik ölçüsü büyük çıkarsa bu, nesnelerin birbirine yakın olduğunun göstergesidir. Tam tersi olarak baktığımızda da eğer iki nesne arasındaki benzerlik ölçüsü küçük çıkarsa, bu da nesnelerin birbirinden uzak olduğunun göstergesidir. Kümeleme analizinde eldeki değişkene göre nesneler arasındaki uzaklıkları hesaplamak için çeşitli ölçü birimleri vardır.

Bu hesaplamalar değişkene göre ya da hem değişken hem de nesneler arasındaki uzaklıklara bakılmasına göre değişmektedir. Eğer değişkenler arasındaki uzaklıklar hesaplanacak ise “Minkowski” uzaklık ölçüsü kullanılarak hesaplanır. Bu uzaklık aşağıda yer alan 3. 1 formülü ile hesaplanmaktadır (Ergün, 2008);

$$d(x, y) = [\sum_{i=1}^p |x_i - y_i|^m]^{1/m} \quad (3.1)$$

Minkowski uzaklığının özel bir durumu olarak, m=2 durumunda Öklid uzaklığı ya da karesel Öklid uzaklığı olarak değişmektedir. a*p boyutlu bir veri setindeki Öklid uzaklığı aşağıda yer alan 3.2 formülü gibi hesaplanmaktadır (Ergün, 2008);

$$d(x, y) = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2} \quad (3.2)$$

Öklid; nesneler ve değişkenler arasındaki hem uzaklıkları hem de benzerlikleri hesaplamada kullanılabilen bir ölçüttür.

3. 2. 2 Kümeleme yöntemleri

Kümeleme analizinde kullanılan yöntemler, uzaklık ve benzerlik ölçütlerinden yararlanılarak nesneler veya değişkenlerin kendi içerisinde homojen ve kendi aralarında da heterojen olarak gruplara ayrılmasına yarayan yöntemlerdir. Kümeleme yöntemleri, grupların nasıl ayırdıklarına göre hiyerarşik ve hiyerarşik olmayan kümeleme yöntemleri olarak ikiye ayrılırlar (Ergün, 2008).

3. 2. 2. 1. Hiyerarşik kümeleme

Hiyerarşik kümelemede, nesneler arasındaki yakınlık-uzaklık ilişkisine göre oluşan kümelerden bir ağaç oluşturur. Hiyerarşik kümelemeye ait özellikler aşağıdaki gibidir;

- Veri seti birkaç kümeye ayrılır. Bu ayrıştırma dendogram denilen ağaç sayesinde oluşturulur.
- Dendogramın kurulum sıralaması, gövdeden yapraklara doğru olduğu gibi; yapraklardan gövdeye doğru da olabilir ve dendogramlar istenilen zamanda sonlandırılarak kümeler elde edilebilir.

Yapraktan gövdeye doğru olan hiyerarşik kümeleme şu şekildedir;

- Başlangıcı her nesne için farklı bir grup oluşturulmasıdır,
- Birleştirilme işlemi bazı kurallara dayanmaktadır. Bu kurallara örnek olarak; merkezler arası uzaklık, ortalama vb durumlar sayılabilir.
- Birleştirme işleminin bitmesi gruplayacak başka küme kalmaması halinde olur. Bir başka deyişle, eldeki bütün nesneler sadece bir küme içinde kalana dek veya istenen küme sayısı elde edilene kadar birleştirme işlemine devam edilir.

Gövdeden yaprağa inen hiyerarşik kümeleme ise şu şekildedir;

- Başlama noktasının, aynı küme içinde yer alan nesnelerle başlaması,
- Kümenin daha küçük kümelere bölünmesi,
- Sonlandırma durumu ortaya çıkana kadar işleme devam edilmesi. Bir başka deyişle, istenilen küme sayısı elde edilene kadar veya her nesne ayrı bir küme oluşturana kadar işleme devam edilmesidir (Darakçı, 2011).

3. 2. 2. 2. Hiyerarşik olmayan kümeleme

Hiyerarşik olmayan kümeleme yöntemleri, hiyerarşik yöntemlere göre daha az zaman almaktadır. Eğer kişi, veri setine göre anlamlı çıkması muhtemel küme sayısına karar verebiliyorsa ya da küme sayısı önceden belirli ise hiyerarşik olmayan kümeleme yöntemleri uygulanır. Hiyerarşik olmayan yöntemlerde hedef, nesnelerin uygun olduğu kümeler içerisinde toplanması ve n tane nesnenin m tane kümeye ayrışması beklenmektedir. Ayrılacak küme sayısı belirlendikten sonra, hangi nesnenin hangi kümede yer alacağı belirtilir. Eldeki veri seti daha önceden belirlenmiş sayıdaki kümeye ayrıştırma işlemi gerçekleştikten sonra, kümelerin merkez noktaları hesaplanır. Bu işlemler her bir gözlemin bir kümeye atılması gerçekleşene kadar devam eder. Herhangi bir nesnenin kümeye atama işlemi gerçekleştikten sonra yeri değişmez. Hiyerarşik olmayan kümeleme yapısında ise küme sayısı başlangıçta zaten belirlenir bu nedenle de kümeleme algoritması bitene kadar herhangi bir nesne ilk atandığı kümeden başka bir kümeye de atanabilir.

Hiyerarşik olmayan kümelemede veri seti x adet gruba ayrılmakta ve her grup bir kümeyi temsil etmektedir.

Hiyerarşik olmayan kümeleme analizindeki sıralama aşağıdaki gibidir;

1. Belirlenen küme sayısı kadar küme ortalaması, araştırmacı tarafından belirlenir.
2. Eldeki gözlemlerin ortalamaları hangi kümenin ortalamasına daha yakın ise o kümeye dahil edilir.
3. Kümelerin ortalama değerleri tekrardan hesaplanır.
4. Küme içinde yer alan elemanlarda bir değişiklik oluşmuyorsa işleyiş son bulur. Ama elemanlarda değişiklikler meydana geliyorsa ikinci adıma dönülerek işlem sürdürülür.

Hiyerarşik olmayan kümeleme içerisinde birçok yöntem bulunmaktadır. Bunlar k-means (k ortalamalar yöntemi), metot kümeleme, bulanık kümeleme gibidir. Fakat bu kümeleme çeşitlerinden en çok kullanılanı k-ortalamalar (k-means) yöntemidir (Ergün, 2008).

En eski kümeleme algoritmalarından olan K-ortalamlar, 1967 yılında MacQueen tarafından geliştirilmiştir (MacQueen, 1967). En yaygın kullanılan öğreticisiz öğrenme yöntemlerinden birisi olan K-ortalamların atama mekanizması, her verinin sadece bir kümeye ait olabilmesine izin verir. Bu nedenle, keskin bir kümeleme algoritmasıdır. Merkez noktanın kümeyi temsil etmesi ana fikrine dayalı bir yöntemdir (Han ve Kamber, 2011).

K-ortalamlar yöntemindeki işleyiş adımları aşağıdaki gibidir;

- Küme sayıları ve merkezlerin başlangıç değerleri rastgele olarak belirlenir,
- Her bir değişken için belirlenmiş olan başlangıç değerlerine olan uzaklıkları hesaplanır,
- Her bir değişken hesaplanan uzaklığa göre, kendisine en yakın olan kümeye atanır,
- Amaç fonksiyonun en küçüklenmesi işlemidir,
- Küme merkezleri yeniden hesaplanır ve yeni kümeler belirlenir,
- Kümeye ait elemanlarda bir değişiklik olmayacak duruma gelinceye kadar, işleyişe devam edilir (Fırat vd 2012).

3. 2. 3. Birliktelik kurallarının tanımı

Birliktelik kuralları; müşterilerin alışveriş anında satın aldıkları ürünlerin veritabanındaki kayıtlarından veriler alarak ve birliktelik analizi uygulanarak, satın aldıkları ürünler arasındaki birliktelik ya da diğer bir deyişle ürünler arasındaki ilişkileri bularak, müşterilerin satın alma alışkanlıklarını ortaya koymayı amaçlamaktadır. Öğrenilen bu birliktelik bağıntıları sayesinde, kuruluşlar daha kazançlı satışlar yapabilme olanağına sahip olmaktadır. Birliktelik kurallarını bir başka tanımla anlatacak olursak; bir ilişkide bir niteliğin alındığı değerler arasında ne kadar bağımlılık olduğu ve yer almayan diğer niteliklere göre gruplandırma yapılmış veriler üzerinde çalışılarak bulunulan kurallardır. Ortaya çıkan bilgiler, veri kümesinde sıklıkla birlikte geçen nitelikler arasındaki ilişkileri göstermektedir.

Birliktelik kurallarının kullanıldığı en klasik örnek “market sepet analizi”dir. Bu işlem, müşterilerin satın almış oldukları ürünler arasındaki bağımlılıkları keşfederek, müşterilerin satın alma alışkanlıklarını ortaya koyar (Gürgen, 2008).

3. 2. 4. Birliktelik kurallarına örnek

Büyük bir alışveriş merkezinin alışveriş işlemlerinin bulunduğu bir veritabanına sahip olduğumuzu düşünecek olursak; bu veri tabanında cips → kuru yemiş şeklinde bir birliktelik kuralının olduğunu varsayalım. Cips ve kuru yemiş müşterilerin aldıkları ürünlerdir. Birliktelik kuralı, “cips satın alan müşteriler %15 olasılıkla kuru yemişde satın alırlar ve tüm satış hareketlerinin %10’u cips ve kuru yemişi içermektedir” anlamını taşımaktadır. Buradaki %15 değeri desteği (support), %10 değeri güven (confidence) değerlerini göstermektedir (Sezer, 2008). Destek; ortaya çıkan kuralın kullanışlılığını, güven ise; doğruluğunu göstermektedir. Verilen örnekteki destek değeri, analiz edilen tüm alışverişlerden %15’inde kuruyemişin cips ile birlikte alındığını gösterir. %10 oranındaki güven değeri ise; aynı alışverişte cips alanların %10’unun kuru yemişde aldığını gösterir (Kalikov, 2006). Birliktelik kuralları “A ürünü ne ile birlikte en çok satılır?” sorularına cevap bulabilmemizi sağlamaktadır (Sezer, 2008). Büyük veri tabanlarında birliktelik kuralları bulunurken genel olarak aşağıdaki adımlar takip edilir;

1) Öncelikle sık tekrarlanan öge kümeleri bulunur. Yani bu öğelerden her biri en az önceden belirlenmiş olan minimum destek sayısı kadar tekrar etmektedir,

2) Sık tekrarlanan kümelerin öğeleri daha güçlü birliktelik kuralı oluşturmaktadır: yani bu kurallar minimum destek ve minimum güven değerlerini karşılayabilmektedir.

Sık tekrarlanan öğeleri bulmak için en yaygın ve klasik olarak uygulanan yöntem “Apriori Algoritması” dır (Kalikov, 2006).

3. 2. 5. Birliktelik kurallarında kullanılan terimler

Market sepet analizinde; nesneler, müşterilerin satın aldığı ürünler ve işlem (transaction), beraber alınan bütün nesnelerin kümesini ifade etmektedir. Birliktelik kurallarında sıklıkla geçen bir kaç önemli terim bulunmaktadır. Bunlar aşağıdaki gibi sıralanmıştır;

- Kuralın sol tarafını ifade eden önce (antecedent),
- Kuralın sağ tarafını ifade eden sonuç (consequent),

- Destek değeri (support),
- Güven (confidence),
- Lift,
- Minimum Destek,
- Minimum Güven,
- Nesne küme,
- Yaygın nesne kümesi,
- Aday nesne kümesi

Birliktelik Kuralları iki aşamadan oluşmaktadır; 'tüm yaygın nesne kümelerinin bulunması' Yani; nesne kümelerinin yaygın olarak yer alabilmesi için; önceden tanımlanmış minimum destek (support) değerinin destek değerinden küçük olması gerekir. İkinci aşamada ise; 'yaygın nesne kümelerinden güçlü birliktelik kurallarının elde edilmesi' Yani; üretilen kuralların, minimum destek (support) ve minimum güven (confidence) durumlarını sağlaması gerekir (Çelikyay, 2010).

Birliktelik kuralları probleminin biçimsel tanımlanması aşağıdaki gibi yapılabilir (Han ve Kamber 2001):

Tanım 1: $I = \{I_1, I_2, \dots, I_m\}$ veri tabanındaki farklı elemanların kümesi olarak gösterilir. D , her biri birincil bir anahtara sahip olan (TID) ve T olarak adlandırılan kayıtların oluşturduğu veri tabanı olup $T \subseteq I$ olacak şekilde eleman kümelerini içerir. X ve Y ' nin $X \subset I$, $Y \subset I$ ve $X \cap Y = \emptyset$ koşullarını sağlayacak şekilde eleman kümeleri olduğu varsayılırsa birliktelik kuralı $X \rightarrow Y$ şeklinde gösterilebilir.

Tanım 2: Bir birliktelik kuralının destek değeri (support), $X \cup Y$ ' yi birlikte içeren kayıt sayısının veri tabanındaki toplam kayıt sayısına oranıdır. Örneğin bir kuralın desteğinin %5 olması, veri tabanındaki tüm kayıtların %5'inin $X \cup Y$ 'yi içermesi anlamına gelir. Destek değeri bir kuralın istatistiksel anlamını belirtir. Genellikle birliktelik kuralları için yüksek destek değerleri cazip olsa da bu her zaman mümkün olmamaktadır. Örneğin, iletişim ağlarındaki bir hatayı tahmin etmeye yönelik bir uygulamada, hata öncesi gerçekleşen olaylar ve bunların birliktelikleri kısıtlı sayıda olsa bile, destek değeri düşük olan bu birliktelikler hatayı bulabilmek için göz ardı edilmemelidir. $X \cup Y$ öğeler kümesinin bulunduğu hareket sayısının toplam hareket sayısına oranıdır. Destek aşağıda yer alan 3. 3 formülü ile gösterilebilir;

$$\text{Destek } (X \rightarrow Y) = n(X \cup Y) / n(D) \quad (3.3)$$

Tanım 3: Bir birliktelik kuralının güven değeri (confidence), $X \cup Y$ 'yi birlikte içeren kayıt sayısının X 'i içeren kayıt sayısına oranıdır. Örneğin bir birliktelik kuralının güven değerinin %85 olması X içeren kayıtlarının %85'inin aynı zamanda Y 'yi de içermesi anlamına gelir. Güven değeri veri kümesi içerisinde X ve Y arasındaki bağıntının derecesine karşılık gelir. Güven değeri kuralın güçlülüğünün ölçüsüdür. Birliktelik kuralları için genellikle yüksek güven değeri gereklidir. Eğer bir ürün nadiren başka bir ürünle beraber satılıyorsa, yani güven değeri düşükse, bu kural yararlı bir kural değildir. Güven değeri şu şekilde ifade edilebilir; $X \cup Y$ öğeler kümesinin bulunduğu hareket sayısının X öğeler kümesi bulunan hareket sayısına oranıdır (Sezer, 2008).

Güven değeri aşağıda yer alan 3. 4 formülü ile gösterilir;

$$\text{Güven } (X \rightarrow Y) = n(X \cup Y) / n(X) \quad (3.4)$$

Lift değeri, x ve y ürünlerinin birlikte ne sıklıkla geçtiğini ifade etmektedir. Lift değeri aşağıda yer alan 3. 5 formülü ile gösterilir;

$$\text{Lift } (X \rightarrow Y) = \text{Supp}(X \cup Y) / [\text{Supp}(X) * \text{Supp}(Y)] \quad (3.5)$$

Veri tabanlarından birliktelik kuralları elde etme işlemi, kullanıcı tarafından tanımlanmış minimum destek ve minimum güven olarak adlandırılan eşik destek ve güven değerini sağlayan tüm kuralların bulunmasını içerir. Eşik değerlerini sağlayan kurallara güçlü kurallar denir. Genel olarak bu değerler sayısal 0-1 aralığından ziyade %0 - %100 yüzde aralığı olarak ifade edilmektedir (Sezer, 2008).

Birliktelik Kuralları Madenciliği süreci iki alt sınıfa ayrılır:

1. Tüm sık geçen nesne kümelerinin bulunması: Bu aşamada tanımlı olan destek değerini kullanarak bu değeri sağlayan nesne kümelerinin bulunması gerçekleştirilir. Eşik değerini aşan nesne kümeleri “sık geçen” ya da “büyük” nesne kümeleri, eşik değerinin altında kalan nesne kümeleri ise “küçük” nesne kümeleri olarak adlandırılır.

Eğer x , veri tabanından elde edilmiş küçük bir nesne kümesi ise x 'i içeren üst kümeler de küçük nesne kümeleridir denilebilir. Bunun tam tersi de doğrudur. Yani x büyük bir nesne kümesi ise X 'in alt kümeleri de göz önüne alınması gereken büyük nesne kümeleri olabilir.

2. Sık geçen nesne kümelerinden güçlü birliktelik kurallarının bulunması: Bu aşamada, birinci aşamada bulunan büyük nesne kümelerini ve önceden tanımlı güven değerini kullanarak güçlü birliktelik kuralları bulunur. Büyük ölçekli veri tabanlarında sık geçen nesne kümeleri bulunması işlemi, birliktelik kuralları bulma sürecinin en zahmetli ve maliyetli bölümü olduğundan çoğu araştırma, sürecin birinci aşamasını çözmeye yönelik verimli algoritmalar geliştirmeye odaklanmıştır (Han ve Kamber, 2011).

3.2.6. Birliktelik kuralları performansı belirleyen durumlar

Veri madenciliği teknikleri çok büyük boyutlu veriler üzerinde uygulandığından, birliktelik analizi için kullanılan algoritmalar da hızlı çalışmalıdır. Birliktelik kuralları algoritmalarının performansını etkileyen durum, tüm yaygın nesne kümelerinin bulunmasıdır.

Minimum güven ve destek değerlerini sağlayan birliktelik kuralları problemi; yaygın geçen nesne kümelerinin bulunması ve yaygın nesne kümelerinden güçlü birliktelik kuralları oluşturulur şeklinde iki adıma bölünmüştür.

Bunlardan birincisi olan, "yaygın geçen nesne kümelerinin bulunması": Nesne küme, kullanıcılar tarafından belirlenmiş minimum destek değerini sağlayan nesne kümeler denilmektedir.

Bunlardan ikincisi olan, "Sık tekrarlanan kümelerin öğeleri daha güçlü birliktelik kuralı oluşturmaktadır": kullanıcı tarafından belirlenen minimum güven değerini sağlayan birliktelik kurallarının yaygın nesne kümeler kullanılarak bulunmasıdır. Minimum güven değerine göre bulunan birliktelik kuralları, kullanıcının ilgilendiği ve önemli bilgi içeren kurallar olacaktır. Birliktelik kurallarının performansını etkileyen esas adım birinci adımdır.

Sepet analizinin başarılı olduđu kısımlara değinecek olursak; kolay sonuç üretir ve üretilen sonuçlar anlaşılır düzeydedir. Farklı boyuttaki veriler üzerinde çalışabilir.

Sepet analizinde seçilen katsayılara göre işlem adedi ve üretilen kuralların sayısı artsa da algoritmanın her adımında yapılan işlemler diğer yöntemlere (genetik algoritmalar, yapay sinir ağları vb.) göre çok daha basit olmaktadır (Çelikyay, 2011).

3.2.7. Birliktelik kurallarının belirlenmesinde kullanılan algoritmalar

Birliktelik kurallarının birçok çeşidi bulunmaktadır. Bunlar; kuralda kullanılan değerlerin tiplerine göre, kuralın içerdği verinin boyutlarına göre olmak üzere iki başlık altında incelenir.

- Kuralda Kullanılan Değerlerin tiplerine Göre: Oluşan kural nesnelerin yokluğu ya da var olması durumuna göre bir birliktelik kuralı olarak türemiş ise, bu tür kurallara “mantıksal birliktelik” kuralları denir. Bu tarz kurallar genel olarak market sepet analizinden elde edilir. Eğer oluşan kurallar nesnelerin nicelikleri ya da özellikleri arasındaki birliktelik ilişkilerini açıklıyorsa, bunlar nicel birliktelik kuralı olarak tanımlanır.
- Kuralın İçerdği Verinin Boyutlarına Göre: Oluşan birliktelik kurallarındaki özellikler ya da nesneler kaç boyutu temsil ediyorsa, kaç boyutlu birliktelik kuralı oldukları söylenir. Örneğin bir nesne ve özellikleri iki boyutu temsil ediyorsa, o zaman iki boyutlu birliktelik kuralıdır denir (Çelikyay, 2010).

3.2.7.1. Apriori algoritması

Apriori algoritması bilgisayar bilimleri ve veri madenciliğinde ilişkisel kuralları öğrenmek için kullanılan klasik bir yöntemdir (Kumar ve Rukmani,2010). Apriori algoritması ismi çalışma prensibine göre adlandırılmıştır. Bu algoritmada, bilgiler bir önceki adımdan geldiği için “pior” anlamına gelen Apriori ismi kullanılmıştır. Bu algoritmanın çalışma prensibi tekrarlayan bir çalışma stiline sahiptir. Veri tabanlarında en yaygın geçen öge kümelerinin keşfedilmesinde kullanılmaktadır. Apriori algoritmasının temelinde; z adet elemana sahip bir öge kümesi, minimum destek değerini sağlıyorsa; bu öge kümesinin alt kümeleri de minimum destek değerini karşılayacaktır.

Birliktelik kuralı iki aşamadan oluşmaktadır. İlk adımda, en yaygın geçen öğeler bulunur. İkinci adımda ise, bu öğelerden en geçerli birliktelik kuralları üretilir. Apriori algoritması birliktelik kurallarının ilk aşamasında en klasik ve popüler olarak kullanılan algoritmadır (Çelikyay, 2010).

Bu algoritma veritabanı üzerinde bir dizi geçiş içerir. k geçişi sırasında, algoritma minimum destek gereksinimi karşılayan k uzunluğundaki L_k sık öğe seti kümesini bulmaktadır. L_k boş iken algoritma sonlandırılır. Bir budama adımı daha küçük bir alt kümesi olan herhangi bir aday ortadan kaldırır (Zhang vd, 2008).

Apriori algoritması; veritabanındaki hareketlerle ilgilenmeden, bir önceki geçişte oluşan yaygın nesne kümelerini birleştirir ve aday nesne kümelerini oluşturmuş olur. Önceki geçişte oluşan alt kümelerden küçük olanlarını da siler. Bir önceki geçişte oluşmuş yaygın nesne kümelerinin de dikkate alınmasıyla, en sık geçen aday nesne kümelerin sayısında işimizi kolaylaştıracak ve anlamlı bir yönde azalma olur (Çelikyay, 2010).

Algoritmanın aşamaları ayrıntılı olarak aşağıdaki gibidir;

- İlk aşamada, destek ve güven değerlerinin karşılaştırılabilmesi için eşik değerleri belirlenir. Analiz sonucu elde edilen sonuçların eşik değerlerinden büyük ya da eşit olması beklenir.
- İkinci adımda, analize dâhil edilecek ürünler için ne kadar tekrar ettikleri yani destek sayıları hesaplanır ve bunun için de veri tabanı taranır. Destek sayıları elde edildikten sonra eşik destek sayısı ile karşılaştırılır. Ürünün elde edilen destek değeri, eşik destek sayısından küçük ise, ilgili satırlar analizden çıkarılır. Geriye sadece uygun olan kayıtlar göz önünde tutulur.
- Üçüncü adıma geldiğimizde, ikinci basamakta seçilen ürünler ikişerli gruplandırılarak, destek sayıları yani tekrar etme sayıları elde edilir. Bir önceki basamaktaki olduğu gibi bu sayılar da destek sayıları ile karşılaştırılır. Eşik değerinden küçük olan satırlar analizden çıkartılır.
- Dördüncü basamakta gruplama sayıları git gide bir sayı arttırılarak, üçerli, dörderli gruplarında eşik değerleri karşılaştırılır. Tekrar etme sayıları, eşik değerinin üstünde veya eşit olduğu sürece işleme devam edilir.

Son aşamada ürün grubu belirlendikten sonra kural destek ölçütüne bakılarak birliktelik kuralları üretilir. Kuralların her biri için güven ölçütleri hesaplanır (Şekeroğlu, 2010).

3.2.7.2. FP GROWTH algoritması

Birliktelik kuralları analizini gerçekleştirmek üzere geliştirilmiş yöntemlerden bir tanesi de FP-Growth (Frequent Pattern Growth) Algoritması'dır . Diğer algoritmalara göre daha yüksek performans gösteren algoritma tüm veri tabanını FP-Tree (Frequent Pattern Tree) adı verilen sıkıştırılmış bir ağaç veri yapısında tutar ve veri tabanını sadece iki kez tarar. Birinci taramada, tüm nesnelerin destek değerlerini hesaplar, ikinci taramada ise ağaç veri yapısını oluşturur. Önemli özellikleri, yaygın nesne kümelerini aday nesne kümeleri üretmeden test edebilmesi, büyük veri kümelerinde hızlı çalışabilmesi ve sistem kaynaklarını verimli kullanabilmesidir. FP-Algoritması'nda öncelikle, veri tabanındaki her bir nesnenin destek değerleri hesaplanır. Destek değerleri, algoritmaya girdi olarak verilen destek eşik değerinden büyük ve eşit olan nesneler, büyüktür küçüğe sıralanarak bir liste içerisine konulur. Böylelikle yaygın olmayan nesnelerin ağaca eklenmesi önlenmiş olur. Daha sonra, veri tabanındaki her bir hareket kaydı nesnelerin destek değerlerine göre sıralanarak ağaca sıkıştırılmış biçimde eklenir. Sıralama işlemi sayesinde destek değeri daha büyük olan nesneler köke daha yakın olur. Sıkıştırma işlemi çok tekrarlı nesnelerin ilk ekler olarak birleştirilmesi ile gerçekleştirilir. Bu metot arama maliyetini önemli ölçüde azaltır. Hareket kaydı içerisinde yer alan bir nesne ağaçta yoksa o nesne için yeni bir düğüm oluşturulur ve destek değeri 1 olarak atanır. Eğer o nesne daha önce ağaçta oluşturulmuş ise sadece o düğümün destek değeri 1 arttırılır. Nesnelerin ağaçtaki başlangıç noktaları, başlık tablosu içerisinde tutulur. Ağaç oluşturulduktan sonra, üzerinde Growth Algoritması çalıştırılır. Growth Algoritması'nda öncelikle ele alınan nesnenin içerisinden geçtiği dallar belirlenir. Eğer tek bir dal varsa yaygın nesneler kümesi, dalı oluşturan nesnelerin kombinasyonudur. Eğer birden fazla dal varsa, destek değeri o daldaki minimum destek değeri olarak belirlenir. Daha sonra bu dallar, o nesne için koşullu örüntü temelini oluşturur. Her bir koşullu örüntü temelinden koşullu örüntü ağacı elde edilir. Daha sonra bu şartlı örüntü ağacı üzerinden algoritma özyinelemeli olarak yeniden çalıştırılır (Han vd, 2000). Birlikte

sıklıkla görünen nesneler kümesini belirleyen FP-Growth Algoritması, ‘böl ve yönet’ yaklaşımına uygun olarak büyük yaygın nesnenin kendi içinde daha küçük nesne kümelerine ayrılmasını sağlar. Bu nedenle oluşturulan FP-Tree veri yapısı asıl veri kümesinden daha büyük olamaz.

Örnek: Çizelge 3. 1 'de gösterildiği gibi altı ögeli küçük bir veritabanı düşünün. Veritabanını taranarak ve frekansı eşik değere eşit ya da daha büyük olan öğeleri tespit edilir.

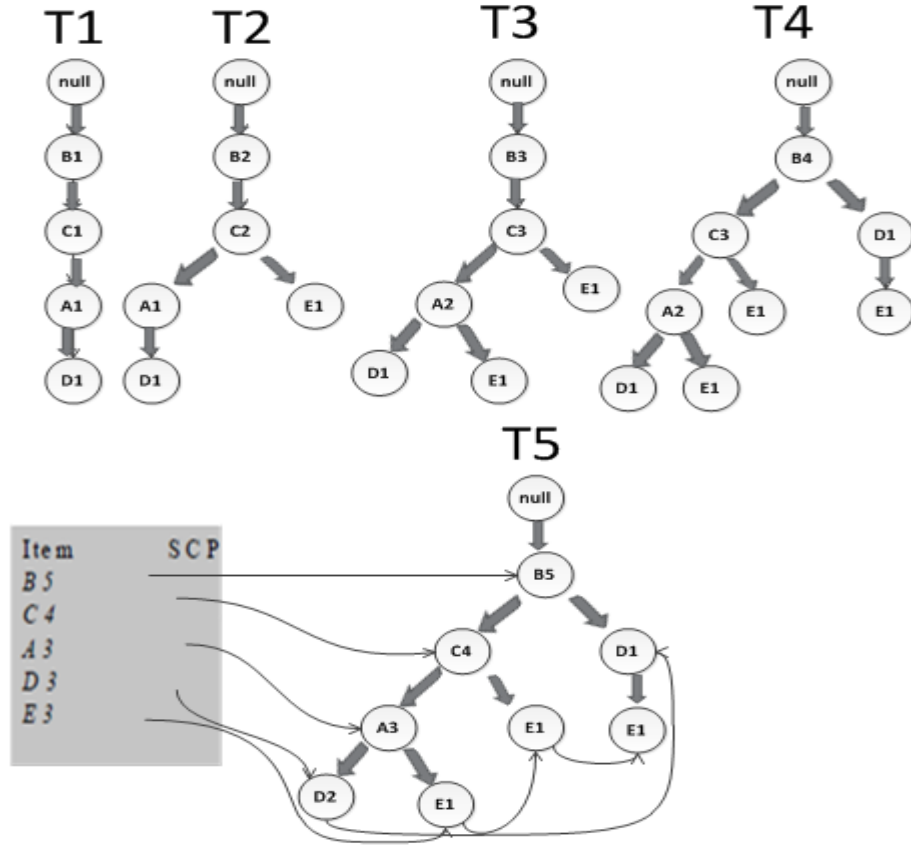
- 1) Sık öğeleri azalan sıraya göre dizilir, {B5, C4, A3, D3, E3}.
- 2) Sadece kökü olan bir ağaç inşa edilir.

Çizelge 3. 1.Küçük Veritabanı

TID	Itemset
1	A, B, C, D, F
2	B, C, E
3	A, B, D
4	A, B, F
5	C, D, F
6	E, F

- 3) Veritabanını yeniden taranarak, her numune için;
 - a) Sadece sık öğeleri (adım 1’de bulunan öğeleri) kullanarak numuneden alınan öğeler ağaca eklenir.
 - b) Tüm numuneler işlenene kadar a adımı tekrarlanır (Zhang vd,2008).

Şekil 3. 1’de FP-Growth algoritmasının çalışma mantığı bulunmaktadır.



Şekil 3. 1. FP-Growth Algoritmasının Yürütme Süreci (Zhang vd, 2008)

3. 2. 8. Benzerlik fonksiyonları (Jaccard, Cosine ve Pearson)

Kosinüs Benzerliği: Doküman karşılaştırmasında kullanılmaktadır. Bir dokümanda bulunan diğerinde bulunmayan özellik için bulunmayan dokümanda o özellik 0 olarak alınır (Başak, 2015). U ve U_i olmak üzere verilen iki kullanıcının benzerliğini karşılaştırmak istediğimizde, $\text{Sim}(u; u_i)$ fonksiyonu bu benzerliğin sonucunu 0-1 aralığında ifade edecek şekilde kullanılır. Bu vektörler her bir kullanıcı için tüm d_i ($i=1, 2, \dots, n$) öğeleri için, önceden elde edilmiş derecelendirme puanlarından oluşur. İki kullanıcının benzerliğini ölçmek için, bu iki vektör arasındaki açının büyüklüğü kullanılabilir. Eğer açı değeri 0 olursa o zaman vektörlerin eş olduğu söylenir. Burada açı büyüdükçe benzerlik oranı da azalır. Bu ters ilişkidten dolayı benzerlik ölçüsü olarak açı yerine açının kosinüs değerini kullanmak daha uygun olabilir. Basitlik için tüm benzerlik puanlarının pozitif olduğu ve açının maksimum değerinin $\frac{\pi}{2}$ olduğu varsayılacaktır. Böylece 0 değerinin en az benzerliği ve 1 değerinin en

yüksek benzerliği ifade ettiği bir fonksiyon elde etmiş olunur. İki vektör arasındaki kosinüs değeri aşağıda yer alan 3. 6 formülü ile hesaplanır (Sorensen, 2015).

$$\cos(\emptyset) = \frac{A.B}{\|A\|\|B\|} \quad (3.6)$$

Burada A ve B birer vektördür. A. B ise bu iki vektörün nokta çarpımına denk gelir. $\|A\|$ ise Euclidean büyüklük olup, aşağıda yer alan 3.7 formülü ile hesaplanır (Sorensen, S).

$$\|A\| = \sqrt{\sum_{i=1}^n a_i^2} \quad (3.7)$$

Burada a_i A vektörünün i'ninci bileşenini ifade eder.

Pearson; Pearson çarpım-moment korelasyon katsayısı (Pearson korelasyonu) iki rassal değişken arasındaki bağımlılığının derecesini ifade eden bir ölçümdür. X ve Y rassal değişkenleri arasındaki korelasyon aşağıda yer alan 3.8 formülü ile hesaplanır.

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\alpha_X \alpha_Y} \quad (3.8)$$

Burada $\text{cov}(x, Y)$ X ve Y'nin kovaryansını, α_X ise X'in standart sapmasını ifade eder. Burada X ve Y değişkenleri yerine U ve U_i değerlerini kullanıldığını varsayarsak Pearson korelasyonunu hesaplanır. U ve U_i değerlerine rassal değişkenler gibi davranılır ve her biri eşit olasılığa sahip çıktı değerleri üreten $R(u, d_j)$ ve $R(u_i, d_j)$ fonksiyonları tarafından üretilirler (Sorensen, 2012).

Jaccard Benzerlik katsayısı (Jaccard indeksi) verilen iki küme arasındaki benzerlik katsayısını istatistiksel olarak değerlendirir. Burada benzerliği hesaplamak için iki kümedeki kesişen eleman sayısı birleşimin eleman sayısına bölünür (Başak,2015). Aşağıda yer alan 3. 9 formülünde, nasıl hesaplandığı bulunmaktadır.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (3.9)$$

Jaccard indeksinin değerinin yüksek çıkması iki kümenin yani stringin birbirine daha çok benzediği anlamına gelir. İki küme birbirine tamamen eşitse Jaccard indeksi 1 olarak bulunur. İki kümenin hiç ortak özelliği yoksa da değer 0 olarak bulunacaktır. Jaccard indeksi 0 ile 1 arasında değişen bir değer alabilir (Şeker,2013).

4. MARKET SEPET ANALİZİ VE BİRLİKTELİK KURALLARI İLE UYGULAMA

4. 1. Market Sepet Analizi

Market sepet analizi, geçmişteki yapılan hareketleri anlamak, karar destek sistemlerinde verilen kararların kalitesini arttırmak için kullanılan bir yaklaşımdır. 1990'lı yıllara kadar, teknik yetersizliklerden ötürü, satışın yapıldığı anda değil, ancak seçilmiş belirli bir zaman aralığında (günlük, haftalık, aylık ya da yıllık) yapılmış satış hareketlerinin hepsinden yola çıkarak veriler elektronik ortamlarda tutulamaktaydı. Fakat bu durum barkot teknolojisinin gelişmesi ile değişime uğradı. Barkot sistemindeki gelişmelerle birlikte, yapılan bir alışveriş hareketi anında toplanmaya ve elektronik ortamda tutulabilecek hale gelmeye başladı. Genel olarak bu hareketlere ait satış verileri süpermarketlerin satış noktalarında toplanmaya başlandığı için, bu tür veriler “market sepet verisi” olarak adlandırılmıştır. Market sepet verisinde yer alan bir kayıta; tek olan hareket numarası (transaction number), hareketin yapıldığı tarih, satın alınan ürünlerin kodu ve isim bilgileri, satın alınan ürün miktarı, satın alınan ürünün fiyatı (orijinal, indirimli, kampanya fiyatları) ve müşterinin bilgileri tutuluyorsa müşteriye ait kişisel bilgiler (ad, soyadı, yaş, medeni durum, adres, meslek vb.) tutulur.

Market sepet analizinde amaç; satışlar arasındaki ilişkileri bulmak ve buna bağlı olan kuralları çıkarmaktır. Eğer A ürününü alan müşteri kitlesi, B ürününü de yüksek olasılıkla alıyorsa gibi kuralları çıkarabilmek, şirketin satışlarını yükseltmek amacı ile kullanılabilir değerli bir bilgi haline gelebilir.

Bu tür bilgilerin elde edilmesiyle birlikte sadece şirket karlarını arttırmak değil, birlikte satın alınan ürünler için kampanya uygulamaları ya da müşterileri satın aldıkları ürünlere göre gruplandırma işlemleri yapılabilir ve bu sayede ürünleri alacak potansiyel müşteriler belirlenebilir (Çelikyay, 2010).

4. 2. E-Ticaret Perakende Sitesinde Sepet Analizi Örneği

Bir e-ticaret perakende sitesi için birliktelik kuralları analizi konulu yüksek lisans çalışmasında, Türkiye'deki bir perakende sektörüne ait bir e-ticaret sitesinde müşteriler tarafından satın alınmış ürünler göz önünde tutularak birliktelik kuralları

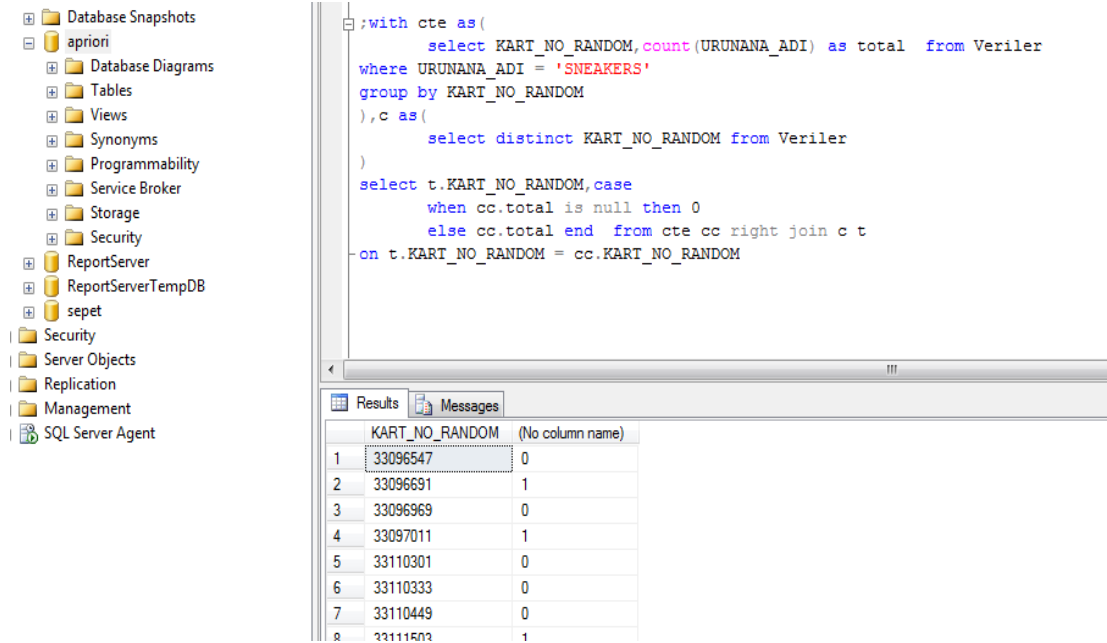
uygulanması yapılarak birbiri ile ilişkisi olan ürünlerin tespit edilmesi hedeflenmiştir ve ardından kümeleme analizi uygulanmıştır. Apriori algoritması, FP-growth algoritmaları k-means, Jaccard, Cosine ve Pearson benzerlik fonksiyonları ile veri analizi yapılmış ve ürünler arasındaki birliktelik kuralları tespit edilmiştir. Daha sonra k-means kümeleme yöntemi ile müşterilerin yaş ve cinsiyetlerine göre aldıkları ürünler ve aynı yaş ve cinsiyet grubunda olan müşterilerin alabileceği ürünler tespit edilmiştir. Ardından yapılan çalışmanın yüzde kaç doğrulukla çalıştığı tespit edilmiştir.

4. 3. Verilerin Toplanması ve Düzenlenmesi

Bu uygulamada kullanılan veri seti Türkiye’deki perakende sektöründe faaliyet gösteren alışveriş sitelerinden bir tanesine ait, 3 aylık zaman diliminde toplanan alışveriş hareketlerinden oluşmaktadır. Gizlilik prensipleri nedeni ile alışveriş sitesi belirtilmeyecek, ürün isimleri ise genel geçen isimler şeklinde ifade edilecektir.

Veriler pl/sql üzerinden yazılan sorgularla çekilerek, Excel’e aktarıldı. Verilere cinsiyet olarak filtre uygulandı ve sadece erkek ürünleri analize alındı. Ardından en çok satın alınmış ürünler tespit edildi ve 43 adet ürün analize sokulmaya karar verildi. Analiz WEKA programında gerçekleştirileceği için, veriler WEKA’ya uygun bir formata getirilmeye çalışıldı. Sonuçların sağlıklı çıkması ve yanıltıcı olmaması adına 10 taneden az ürün satın alanlar veri setinden çıkarıldı. Veride 10 taneden az alınmış ürünlerin bulunması, sonuçlar konusunda yanlış ve yanıltıcı bir sonuca sürükleyebilir. Daha sonra aynı manaya gelen ama farklı isimlerde geçen ürün isimleri sql sorgularıyla güncellenerek ürünler tek bir başlık altında birleştirildi. Sadece ayakkabı çeşitleri ayakkabı altında birleştirilmedi. Çünkü spor ayakkabısı alan bir kişi ile klasik ayakkabı alan kişilerin alacağı ürünler farklı yönde olabilmektedir. Müşteri id’leri bir kolonda aşağı kadar ve aynı şekilde satın aldıkları ürünler yanlarında yazmaktadır. Bu veri setini WEKA’nın algılayabileceği formata getirmek gerekmektedir. Bu nedenle veriler SQL Server Management Studio 2008 programına atıldı. Şekil 4.1’de yazılan sorgu ile veri 1 ve 0 şekline getirildi. 1’ler ürünün o müşteri tarafından satın alındığını, 0’ler ise o ürünün müşteri tarafından satın alınmadığını göstermektedir. Fakat veride 0 olduğu zaman program bu ürünü almayanlar bu ürünü de almazlar gibi bir sonuç ürettiği için, 0’lar ‘?’ ile değiştirildi. Daha sonra Şekil 4.1’deki sorgu ile teker teker her bir ürün için çalıştırılarak WEKA

için bir excel dosyası elde edildi. Ardından oluşan excel dosyası arff formatına dönüştürülerek, Şekil 4.2 ve Şekil 4.3’de WEKA’ya yüklenecek biçime getirildi. Apriori ve FP Growth algoritmalarına ait sonuçlar WEKA program ile elde edildi. Daha sonra sonuçları test etmek için Jaccard, Cosine ve Pearson algoritmaları C# program kullanılarak yazıldı ve bir ara yüz geliştirildi. Veride bulunan yaş ve cinsiyet değişkenleri C# program kullanılarak k-means algoritmasıyla kümelendi. Daha sonra yaş, cinsiyet ve satın alınan ürünler birlikte kümelendi. Hangi yaş grubunda ve hangi cinsiyetler, hangi ürünlerle, hangi kümede bulunmaktadır şeklinde kümelere ayrıştırıldı. C# ile geliştirilen ara yüzde yeni gelen müşterinin yaş, cinsiyet, aldığı ürünler ve bu ürünlerden kaç adet satın aldıkları ile seçimler bulunmaktadır. Ayrıca Jaccard, Cosine ve Pearson algoritmaları ile birliktelik kuralları tercihinde bulunup sistemin kişiye tavsiyede bulunabileceği bir sistem ve k-means’e göre kümelenen kişiler içinde bu algoritmaları seçip uygulatarak ürün tavsiye eden bir ara yüz geliştirilmiştir.



```

;with cte as(
    select KART_NO_RANDOM, count(URUNANA_ADI) as total from Veriler
    where URUNANA_ADI = 'SNEAKERS'
    group by KART_NO_RANDOM
), c as(
    select distinct KART_NO_RANDOM from Veriler
)
select t.KART_NO_RANDOM, case
    when cc.total is null then 0
    else cc.total end from cte cc right join c t
on t.KART_NO_RANDOM = cc.KART_NO_RANDOM

```

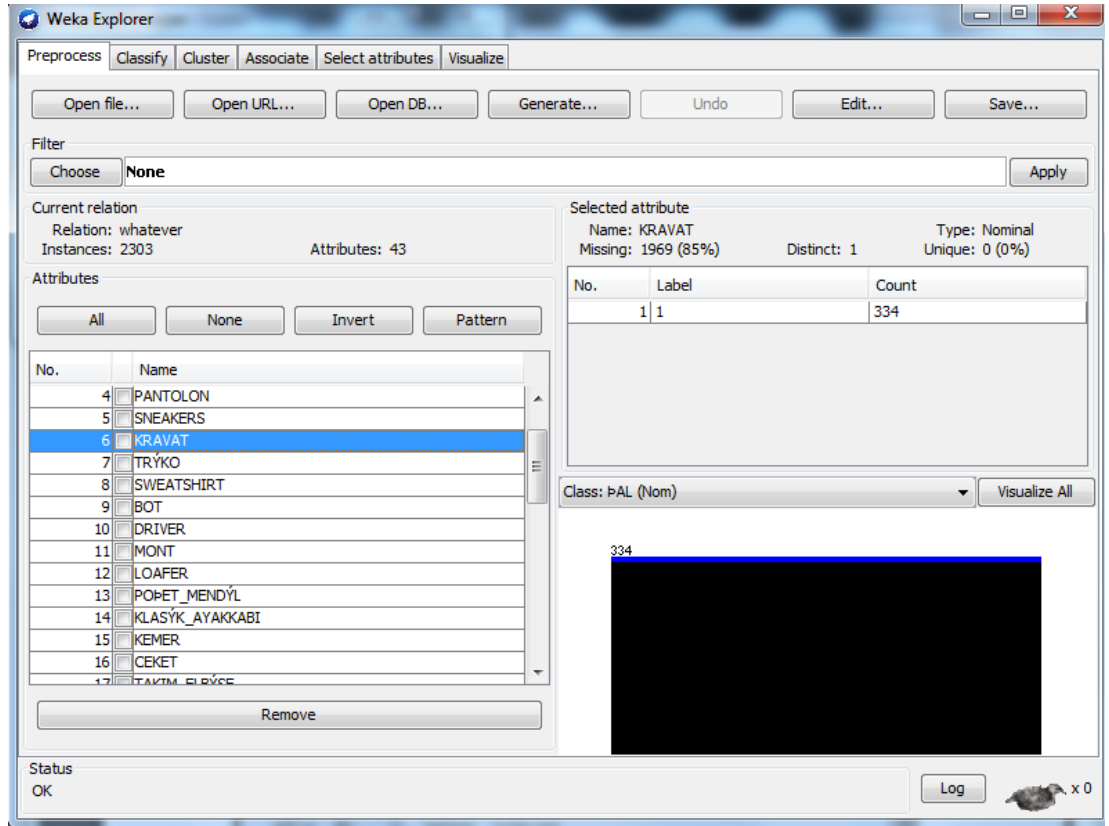
	KART_NO_RANDOM	(No column name)
1	33096547	0
2	33096691	1
3	33096969	0
4	33097011	1
5	33110301	0
6	33110333	0
7	33110449	0
8	33111502	1

Şekil 4. 1. WEKA Datası Oluşumu

Temel olarak sınıflandırma (classification), kümeleme (clustering) ve ilişkilendirme (association) işlemlerini yapmaktadır. WEKA'ya ait giriş ekranı Şekil 4.4'de ve veri yükleme ve analiz yapma ara yüzü Şekil 4.5'deki gibidir.



Şekil 4. 4. WEKA Giriş Ekranı



Şekil 4. 5. WEKA Data Yükleme Ekranı

4. 5. Amaç

Veri setinin içerisinde bir biri ile ilgili olan ürünleri tespit etmek, alışveriş yapan müşterileri yaş ve cinsiyetiyle birlikte aldıkları ürünlere göre kümeleyip, yeni gelen bir müşteriye yaşı ve cinsiyetine göre ürün tavsiye etmektir.

Bu amaca göre tezimizin hipotezi şu şekilde olmaktadır;

“Veri madenciliği yöntemlerinden olan birliktelik kurallarına ait algoritmalar uygulanarak belirli ürünleri alanların, farklı ürünler arasındaki ilişkisi bulunabilir ve ürün tavsiyesinde bulunulabilir. ”

5. DEĞERLENDİRME

5. 1. Bulgular

Veri setinin WEKA'ya sokulmasından sonra Şekil 5. 1'de Apriori ve Şekil 5. 2'de FP Growth algoritmalarına ait sonuçlar sırasıyla aşağıdaki gibidir;

Best rules found:

1. PANTOLON=1 SNEAKERS=1 TRIKO=1 MONT=1 22 ==> GÖMLEK=1 22 conf:(1)
2. TSHIRT=1 PANTOLON=1 SNEAKERS=1 TRIKO=1 MONT=1 20 ==> GÖMLEK=1 20 conf:(1)
3. GÖMLEK=1 TSHIRT=1 TRIKO=1 KEMER=1 17 ==> PANTOLON=1 17 conf:(1)
4. AYAKKABI=1 SNEAKERS=1 TRIKO=1 MONT=1 16 ==> GÖMLEK=1 16 conf:(1)
5. MONT=1 JEAN_PANTOLON=1 15 ==> PANTOLON=1 15 conf:(1)
6. GÖMLEK=1 AYAKKABI=1 TRIKO=1 KEMER=1 15 ==> PANTOLON=1 15 conf:(1)
7. TSHIRT=1 AYAKKABI=1 SNEAKERS=1 TRIKO=1 MONT=1 15 ==> GÖMLEK=1 15 conf:(1)
8. AYAKKABI=1 PANTOLON=1 SNEAKERS=1 TRIKO=1 MONT=1 15 ==> GÖMLEK=1 15 conf:(1)
9. GÖMLEK=1 SNEAKERS=1 SWEATSHIRT=1 MONT=1 14 ==> PANTOLON=1 14 conf:(1)
10. GÖMLEK=1 PANTOLON=1 SWEATSHIRT=1 MONT=1 14 ==> SNEAKERS=1 14 conf:(1)
11. TSHIRT=1 AYAKKABI=1 PANTOLON=1 SNEAKERS=1 TRIKO=1 MONT=1 14 ==> GÖMLEK=1 14 conf:(1)
12. GÖMLEK=1 TSHIRT=1 AYAKKABI=1 PANTOLON=1 SNEAKERS=1 MONT=1 14 ==> TRIKO=1 14 conf:(1)
13. AYAKKABI=1 TRIKO=1 TAKIM_ELBİSE=1 13 ==> GÖMLEK=1 13 conf:(1)
14. SNEAKERS=1 MONT=1 JEAN_PANTOLON=1 13 ==> PANTOLON=1 13 conf:(1)
15. TSHIRT=1 SNEAKERS=1 TRIKO=1 JEAN_PANTOLON=1 13 ==> GÖMLEK=1 13 conf:(1)
16. GÖMLEK=1 SNEAKERS=1 TRIKO=1 KEMER=1 13 ==> PANTOLON=1 13 conf:(1)
17. JEAN_PANTOLON=1 YELEK=1 12 ==> PANTOLON=1 12 conf:(1)
18. GÖMLEK=1 SWEATSHIRT=1 LOAFER=1 12 ==> TSHIRT=1 12 conf:(1)
19. TSHIRT=1 MONT=1 JEAN_PANTOLON=1 12 ==> PANTOLON=1 12 conf:(1)
20. TRIKO=1 SWEATSHIRT=1 LOAFER=1 12 ==> TSHIRT=1 12 conf:(1)
21. TSHIRT=1 PANTOLON=1 TRIKO=1 LOAFER=1 12 ==> GÖMLEK=1 12 conf:(1)
22. AYAKKABI=1 PANTOLON=1 TRIKO=1 BOT=1 12 ==> GÖMLEK=1 12 conf:(1)
23. PANTOLON=1 SNEAKERS=1 KRAVAT=1 KEMER=1 12 ==> GÖMLEK=1 12 conf:(1)
24. GÖMLEK=1 SNEAKERS=1 KRAVAT=1 KEMER=1 12 ==> PANTOLON=1 12 conf:(1)
25. PANTOLON=1 SNEAKERS=1 TRIKO=1 CEKET=1 12 ==> GÖMLEK=1 12 conf:(1)

Şekil 5. 1. Apriori Birliktelik Kuralları Sonuçları

FPGrowth found 1590 rules (displaying top 50)

1. [PANTOLON=1, TRIKO=1, MONT=1]: 28 ==> [TSHIRT=1]: 25 conf:(0.89) lift:(3.14) lev:(0.01) <conv:(5.01)>
2. [PANTOLON=1, TRIKO=1, MONT=1]: 28 ==> [GÖMLEK=1]: 25 conf:(0.89) lift:(2.53) lev:(0.01) <conv:(4.53)>
3. [SNEAKERS=1, TRIKO=1, MONT=1]: 27 ==> [GÖMLEK=1]: 24 conf:(0.89) lift:(2.52) lev:(0.01) <conv:(4.37)>
4. [GÖMLEK=1, TRIKO=1, SWEATSHIRT=1]: 42 ==> [TSHIRT=1]: 36 conf:(0.86) lift:(3.01) lev:(0.01) <conv:(4.29)>
5. [TSHIRT=1, TRIKO=1, MONT=1]: 32 ==> [GÖMLEK=1]: 28 conf:(0.88) lift:(2.48) lev:(0.01) <conv:(4.14)>
6. [AYAKKABI=1, PANTOLON=1, KRAVAT=1]: 38 ==> [GÖMLEK=1]: 33 conf:(0.87) lift:(2.46) lev:(0.01) <conv:(4.1)>
7. [GÖMLEK=1, AYAKKABI=1, SNEAKERS=1, TRIKO=1]: 32 ==> [PANTOLON=1]: 27 conf:(0.84) lift:(3.56) lev:(0.01) <conv:(4.07)>
8. [TRIKO=1, CEKET=1]: 31 ==> [GÖMLEK=1]: 27 conf:(0.87) lift:(2.47) lev:(0.01) <conv:(4.01)>
9. [TSHIRT=1, PANTOLON=1, KRAVAT=1]: 31 ==> [GÖMLEK=1]: 27 conf:(0.87) lift:(2.47) lev:(0.01) <conv:(4.01)>
10. [TSHIRT=1, SNEAKERS=1, PANTOLON=1, TRIKO=1]: 43 ==> [GÖMLEK=1]: 37 conf:(0.86) lift:(2.44) lev:(0.01) <conv:(3.98)>
11. [AYAKKABI=1, TSHIRT=1, SNEAKERS=1, TRIKO=1]: 31 ==> [PANTOLON=1]: 26 conf:(0.84) lift:(3.54) lev:(0.01) <conv:(3.94)>
12. [GÖMLEK=1, TRIKO=1, MONT=1]: 33 ==> [TSHIRT=1]: 28 conf:(0.85) lift:(2.98) lev:(0.01) <conv:(3.94)>
13. [GÖMLEK=1, JEAN_PANTOLON=1]: 44 ==> [PANTOLON=1]: 36 conf:(0.82) lift:(3.45) lev:(0.01) <conv:(3.73)>
14. [TRIKO=1, LOAFER=1]: 28 ==> [GÖMLEK=1]: 24 conf:(0.86) lift:(2.43) lev:(0.01) <conv:(3.63)>
15. [GÖMLEK=1, SNEAKERS=1, SWEATSHIRT=1]: 33 ==> [PANTOLON=1]: 27 conf:(0.82) lift:(3.45) lev:(0.01) <conv:(3.6)>
16. [GÖMLEK=1, PANTOLON=1, TRIKO=1, SWEATSHIRT=1]: 30 ==> [TSHIRT=1]: 25 conf:(0.83) lift:(2.93) lev:(0.01) <conv:(3.58)>
17. [GÖMLEK=1, TSHIRT=1, PANTOLON=1, SWEATSHIRT=1]: 32 ==> [TRIKO=1]: 25 conf:(0.78) lift:(6.2) lev:(0.01) <conv:(3.5)>
18. [GÖMLEK=1, SNEAKERS=1, SWEATSHIRT=1]: 33 ==> [TSHIRT=1]: 27 conf:(0.82) lift:(2.88) lev:(0.01) <conv:(3.37)>
19. [AYAKKABI=1, SNEAKERS=1, TRIKO=1]: 44 ==> [PANTOLON=1]: 35 conf:(0.8) lift:(3.36) lev:(0.01) <conv:(3.36)>
20. [AYAKKABI=1, TSHIRT=1, SNEAKERS=1, TRIKO=1]: 31 ==> [GÖMLEK=1]: 26 conf:(0.84) lift:(2.38) lev:(0.01) <conv:(3.34)>
21. [GÖMLEK=1, SNEAKERS=1, MONT=1]: 35 ==> [PANTOLON=1]: 28 conf:(0.8) lift:(3.37) lev:(0.01) <conv:(3.34)>
22. [GÖMLEK=1, SNEAKERS=1, PANTOLON=1, TRIKO=1]: 46 ==> [TSHIRT=1]: 37 conf:(0.8) lift:(2.83) lev:(0.01) <conv:(3.29)>
23. [GÖMLEK=1, AYAKKABI=1, SNEAKERS=1, TRIKO=1]: 32 ==> [TSHIRT=1]: 26 conf:(0.81) lift:(2.86) lev:(0.01) <conv:(3.27)>
24. [GÖMLEK=1, TSHIRT=1, SNEAKERS=1, TRIKO=1]: 47 ==> [PANTOLON=1]: 37 conf:(0.79) lift:(3.32) lev:(0.01) <conv:(3.26)>
25. [GÖMLEK=1, SNEAKERS=1, SWEATSHIRT=1]: 33 ==> [TRIKO=1]: 25 conf:(0.76) lift:(6.02) lev:(0.01) <conv:(3.2)>

Şekil 5. 2. FP Growth Birliktelik Kuralları Sonuçları

Veri setinden Apriori algoritmasına göre lift değerleri en büyükten en küçüğe doğru ilk 20 kural Çizelge 5. 1'deki gibi sıralanmıştır;

Çizelge 5. 1. Apriori Algoritması İlk 20 Kural

	İlk Alışveriş	İkinci Alışveriş	Lift
1	gömlek, tshirt, pantolon(108)	triko, mont(23)	11,41
2	triko, mont(43)	gömlek, tshirt, pantolon(23)	11,41
3	gömlek, pantolon, triko(76)	tshirt, mont(23)	10,25
4	tshirt, mont(68)	gömlek, pantolon, triko(23)	10,25
5	tshirt, pantolon, triko(73)	gömlek, mont(23)	10,08
6	gömlek, mont(72)	tshirt, pantolon, triko(23)	10,08
7	pantolon, triko(121)	gömlek, tshirt, mont(23)	9,95
8	gömlek, tshirt, mont(44)	pantolon, triko(23)	9,95
9	gömlek, tshirt, triko(91)	pantolon, mont(23)	9,87
10	pantolon, mont(59)	gömlek, tshirt, triko(23)	9,87
11	tshirt, triko(143)	gömlek, pantolon, mont(23)	9,26
12	gömlek, pantolon, mont(40)	tshirt, triko(23)	9,26
13	tshirt, pantolon, triko(73)	gömlek, sweatshirt(25)	8,76
14	gömlek, sweatshirt(90)	tshirt, pantolon, triko(25)	8,76
15	gömlek, triko(155)	tshirt, pantolon, mont(23)	8,76
16	tshirt, pantolon, mont(39)	gömlek, triko(23)	8,76
17	gömlek, tshirt, sneakers(81)	pantolon, triko(37)	8,69
18	pantolon, triko(121)	gömlek, tshirt, sneakers(37)	8,69
19	tshirt, triko(143)	gömlek, pantolon, sweatshirt(25)	8,57
20	gömlek, pantolon, sweatshirt(47)	tshirt, triko(25)	8,57

Veri setinden FP-Growth algoritmasına göre lift değerleri en büyükten en küçüğe doğru ilk 20 kural Çizelge 5.2'deki gibi sıralanmıştır;

Çizelge 5. 2. FP-Growth Algoritması İlk 20 Kural

	İlk Alışveriş	İkinci Alışveriş	Lift
1	gömlek, tshirt, pantolon, sweatshirt(32)	triko(25)	6,2
2	gömlek, ayakkabı, sneakers, triko(32)	pantolon(27)	3,56
3	ayakkabı, shirt, sneakers, triko(31)	pantolon(26)	3,54
4	gömlek, jean_pantolon(44)	pantolon(36)	3,45
5	gömlek, sneakers, sweatshirt(33)	pantolon(27)	3,45
6	ayakkabı, sneakers, triko(44)	pantolon(35)	3,36
7	gömlek, triko, sweatshirt(42)	tshirt(36)	3,01
8	gömlek, triko, mont(33)	tshirt(28)	2,98
9	gömlek, pantolon, triko, sweatshirt(30)	tshirt(25)	2,93
10	gömlek, sneakers, sweatshirt(33)	tshirt(27)	2,88
11	pantolon, triko, mont(28)	gömlek(25)	2,53
12	sneakers, triko, mont(27)	gömlek(24)	2,52
13	tshirt, triko, mont(32)	gömlek(28)	2,48
14	triko, ceket(31)	gömlek(27)	2,47
15	tshirt, pantolon, kravat(31)	gömlek(27)	2,47
16	ayakkabı, pantolon, kravat(38)	gömlek(33)	2,46
17	tshirt, sneakers, pantolon, triko(43)	gömlek(37)	2,44
18	triko, loafer(28)	gömlek(24)	2,43

	İlk Alışveriş	İkinci Alışveriş	Lift
19	ayakkabı, tshirt, sneakers, triko(31)	gömlek(26)	2,38
20	tshirt, sneakers, mont(36)	gömlek(27)	2,13

Aşağıda Apriori algoritmasının lift sonuçlarına göre top 20 önerilerinin oluşturulan tavsiye sistemiyle karşılaştırmalarının ekranları bulunmaktadır. Bu karşılaştırmayı yapabilmemiz için birliktelik kuralının seçilmesi gerekmektedir. Ayrıca sistem FP Growth algoritması içinde aynı şekilde çalışmaktadır. Birliktelik kuralı seçilmiş halde 3 farklı benzerlik algoritmasının sonuçları aşağıda Çizelge 5. 3 ve 5. 4’de belirtilmiştir. Şekil 5. 3, 5. 4 ve 5. 5’de de üç farklı benzerlik fonksiyonu için ara yüz ekran görüntüleri mevcuttur.

Gömlek, tshirt, pantolon alanlar kişilerden triko ve mont da alanlar için aşağıda Şekil 5. 3’de Jaccard fonksiyonuna göre dönen sonuçlar bulunmaktadır.

Hoş Geldiniz ElifSafak

Ürün Adı: PANTOLON Adet: 1 Sepete Ekle

Sizin: ☒ Birliktelik ☐ Kümeleme Benzerlik Alg.: Jaccard

Sepeteki Ürünler

Ürün ID	Ürün Adı	Ürün Sayısı	Sil
0	GÖMLEK	1	İptal
1	TSHIRT	1	İptal
3	PANTOLON	1	İptal

Önerilen Ürünler

Ürün Kodu	Ürün Adı	Önizleme
2	AYAKKABI	Önizle
4	SNEAKERS	Önizle
6	TRİKO	Önizle

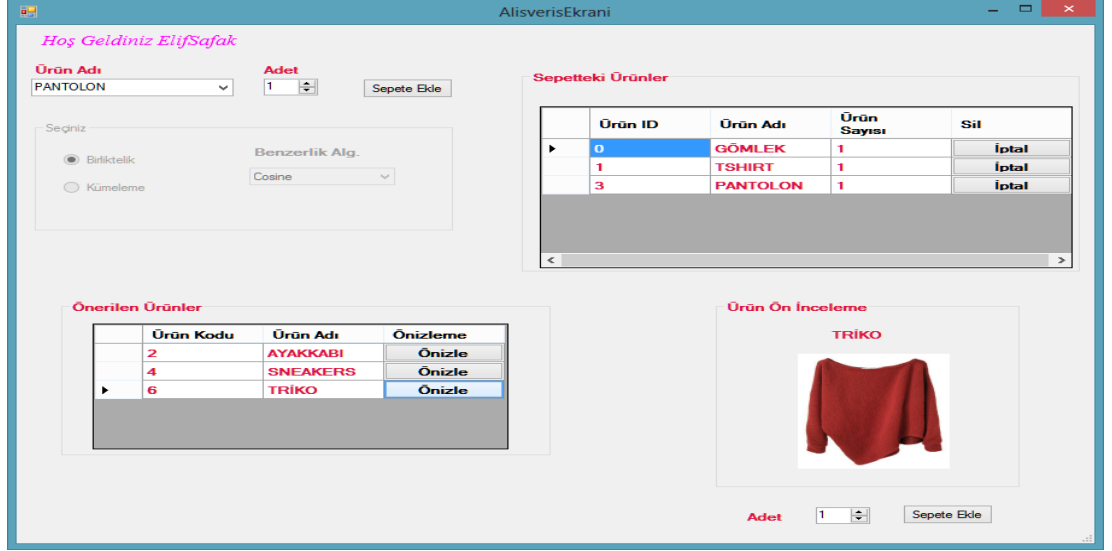
Ürün Ön İnceleme

TRİKO

Adet: 1 Sepete Ekle

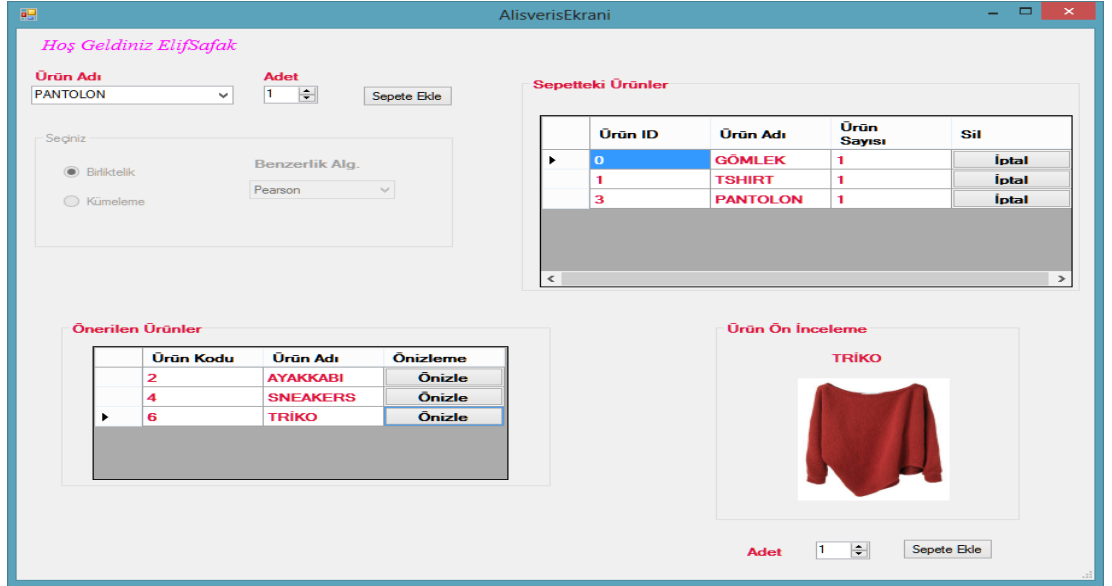
Şekil 5. 3. Jaccard Fonksiyonu ile Birliktelik Kuralları Sonuçları

Gömlek, tshirt, pantolon alanlar kişilerden triko ve mont da alanlar için aşağıda Şekil 5. 4’de Cosine fonksiyonuna göre dönen sonuçlar bulunmaktadır.



Şekil 5. 4. Cosine Fonksiyon ile Birliktelik Kuralları Sonuçları

Gömlek, tshirt, pantolon alan kişilerden triko ve mont da alanlar için aşağıda Şekil 5.5’de Pearson fonksiyonuna göre dönen sonuçlar bulunmaktadır.



Şekil 5. 5. Pearson Fonksiyonu ile Birliktelik Kuralları Sonuçları

Çizelge 5. 3. Apriori Algoritması ile Benzerlik Fonksiyonlarının Sonuçları

İlk alışveriş	İkinci alışveriş
GÖMLEK, TSHIRT, PANTOLON →	TRİKO, MONT
Birliktelik - Jaccard Fonksiyonu	ayakkabı,sneakers,triço
Birliktelik - Cosine Fonksiyonu	ayakkabı,sneakers,triço
Birliktelik - Pearson Fonksiyonu	ayakkabı,sneakers,triço
TRİKO, MONT →	GÖMLEK, TSHIRT, PANTOLON
Birliktelik - Jaccard Fonksiyonu	gömlék,tshirt,sneakers,pantolon,sweatshirt
Birliktelik - Cosine Fonksiyonu	gömlék,tshirt,sneakers,pantolon,sweatshirt
Birliktelik - Pearson Fonksiyonu	gömlék,tshirt,pantolon,sneakers,ayakkabı
GÖMLEK, PANTOLON, TRİKO →	TSHIRT, MONT
Birliktelik - Jaccard Fonksiyonu	tshirt
Birliktelik - Cosine Fonksiyonu	tshirt
Birliktelik - Pearson Fonksiyonu	tshirt
TSHIRT, MONT→	GÖMLEK, PANTOLON, TRİKO
Birliktelik - Jaccard Fonksiyonu	sneakers,gömlék,pantolon,triço,ayakkabı
Birliktelik - Cosine Fonksiyonu	sneakers,gömlék,pantolon,ayakkabı,triço
Birliktelik - Pearson Fonksiyonu	gömlék,sneakers,pantolon,ayakkabı,triço
TSHIRT, PANTOLON, TRİKO →	GÖMLEK, MONT
Birliktelik - Jaccard Fonksiyonu	gömlék
Birliktelik - Cosine Fonksiyonu	gömlék
Birliktelik - Pearson Fonksiyonu	gömlék
GÖMLEK, MONT→	TSHIRT, PANTOLON, TRİKO
Birliktelik - Jaccard Fonksiyonu	ayakkabı,pantolon,tshirt,sneakers,triço
Birliktelik - Cosine Fonksiyonu	ayakkabı,pantolon,tshirt,sneakers,triço
Birliktelik - Pearson Fonksiyonu	ayakkabı,pantolon,tshirt,sneakers,triço
PANTOLON, TRİKO→	GÖMLEK, TSHIRT, MONT
Birliktelik - Jaccard Fonksiyonu	tshirt,gömlék,ayakkabı,sneakers,kravat,bot,mont
Birliktelik - Cosine Fonksiyonu	tshirt
Birliktelik - Pearson Fonksiyonu	tshirt,gömlék,ayakkabı,sneakers,kravat,bot,mont
GÖMLEK, TSHIRT, MONT→	PANTOLON, TRİKO
Birliktelik - Jaccard Fonksiyonu	-
Birliktelik - Cosine Fonksiyonu	-
Birliktelik - Pearson Fonksiyonu	-
GÖMLEK, TSHIRT, TRİKO→	PANTOLON, MONT
Birliktelik - Jaccard Fonksiyonu	ayakkabı,pantolon
Birliktelik - Cosine Fonksiyonu	ayakkabı,pantolon
Birliktelik - Pearson Fonksiyonu	ayakkabı,pantolon
PANTOLON, MONT →	GÖMLEK, TSHIRT, TRİKO
Birliktelik - Jaccard Fonksiyonu	gömlék,tshirt,sneakers,ayakkabı,kravat,triço
Birliktelik - Cosine Fonksiyonu	tshirt,gömlék,sneakers,ayakkabı,kravat,triço
Birliktelik - Pearson Fonksiyonu	tshirt,gömlék,sneakers,ayakkabı,kravat,triço

İlk alışveriş	İkinci alışveriş
TSHIRT, TRİKO→	GÖMLEK, PANTOLON, MONT
Birliktelik - Jaccard Fonksiyonu	gömlek,sweatshirt,pnatolon,sneakers,bot,ayakkabı,mont
Birliktelik - Cosine Fonksiyonu	-
Birliktelik - Pearson Fonksiyonu	-
GÖMLEK, PANTOLON, MONT →	TSHIRT, TRİKO
Birliktelik - Jaccard Fonksiyonu	-
Birliktelik - Cosine Fonksiyonu	-
Birliktelik - Pearson Fonksiyonu	-
TSHIRT, PANTOLON, TRİKO→	GÖMLEK, SWEATSHIRT
Birliktelik - Jaccard Fonksiyonu	gömlek
Birliktelik - Cosine Fonksiyonu	gömlek
Birliktelik - Pearson Fonksiyonu	gömlek
GÖMLEK, SWEATSHIRT →	TSHIRT, PANTOLON, TRİKO
Birliktelik - Jaccard Fonksiyonu	tshirt,ayakkabı,pantolon,kravat
Birliktelik - Cosine Fonksiyonu	tshirt,kravat,pantolon,ayakkabı
Birliktelik - Pearson Fonksiyonu	tshirt,ayakkabı,pantolon,kravat
GÖMLEK, TRİKO →	TSHIRT, PANTOLON, MONT
Birliktelik - Jaccard Fonksiyonu	tshirt,ayakkabı,kravat,sneakers,mont
Birliktelik - Cosine Fonksiyonu	-
Birliktelik - Pearson Fonksiyonu	tshirt,ayakkabı,kravat,sneakers,mont
TSHIRT, PANTOLON, MONT→	GÖMLEK, TRİKO
Birliktelik - Jaccard Fonksiyonu	-
Birliktelik - Cosine Fonksiyonu	-
Birliktelik - Pearson Fonksiyonu	-
GÖMLEK, TSHIRT, SNEAKERS→	PANTOLON, TRİKO
Birliktelik - Jaccard Fonksiyonu	pantolon
Birliktelik - Cosine Fonksiyonu	pantolon
Birliktelik - Pearson Fonksiyonu	pantolon
PANTOLON, TRİKO→	GÖMLEK, TSHIRT, SNEAKERS
Birliktelik - Jaccard Fonksiyonu	tshirt,gömlek,ayakkabı,sneakers,kravat
Birliktelik - Cosine Fonksiyonu	tshirt
Birliktelik - Pearson Fonksiyonu	tshirt,gömlek,ayakkabı,sneakers,kravat
TSHIRT, TRİKO→	GÖMLEK, PANTOLON, SWEATSHIRT
Birliktelik - Jaccard Fonksiyonu	gömlek,sweatshirt,pantolon,sneakers,bot,ayakkabı,mont
Birliktelik - Cosine Fonksiyonu	-
Birliktelik - Pearson Fonksiyonu	gömlek,sweatshirt,pantolon,sneakers,bot
GÖMLEK, PANTOLON, SWEATSHIRT →	TSHIRT, TRİKO
Birliktelik - Jaccard Fonksiyonu	-
Birliktelik - Cosine Fonksiyonu	-
Birliktelik - Pearson Fonksiyonu	-

Çizelge 5. 4. FP-Growth Algoritması ile Benzerlik Fonksiyonlarının Sonuçları

İlk alışveriş	İkinci alışveriş
PANTOLON, TRİKO, MONT →	TSHIRT
FP-Growth - Jaccard Fonksiyonu	tshirt,gömlek
FP-Growth - Cosine Fonksiyonu	gömlek,tshirt
FP-Growth - Pearson Fonksiyonu	gömlek,tshirt
PANTOLON, TRİKO, MONT →	GÖMLEK
FP-Growth - Jaccard Fonksiyonu	tshirt,gömlek
FP-Growth - Cosine Fonksiyonu	gömlek,tshirt
FP-Growth - Pearson Fonksiyonu	gömlek,tshirt
SNEAKERS, TRİKO, MONT →	GÖMLEK
FP-Growth - Jaccard Fonksiyonu	gömlek,sweatshirt,şapka
FP-Growth - Cosine Fonksiyonu	gömlek,sweatshirt,şapka
FP-Growth - Pearson Fonksiyonu	gömlek,tshirt,pantolon,sweatshirt,şapka
GÖMLEK, TRİKO, SWEATSHIRT →	TSHIRT
FP-Growth - Jaccard Fonksiyonu	-
FP-Growth - Cosine Fonksiyonu	-
FP-Growth - Pearson Fonksiyonu	-
TSHIRT, TRİKO, MONT →	GÖMLEK
FP-Growth - Jaccard Fonksiyonu	-
FP-Growth - Cosine Fonksiyonu	-
FP-Growth - Pearson Fonksiyonu	-
AYAKKABI, PANTOLON, KRAVAT →	GÖMLEK
FP-Growth - Jaccard Fonksiyonu	gömlek
FP-Growth - Cosine Fonksiyonu	gömlek
FP-Growth - Pearson Fonksiyonu	gömlek
GÖMLEK,AYAKKABI, SNEAKERS, TRİKO →	PANTOLON
FP-Growth - Jaccard Fonksiyonu	tshirt,pantolon
FP-Growth - Cosine Fonksiyonu	pantolon
FP-Growth - Pearson Fonksiyonu	pantolon,tshirt
TRİKO, CEKET →	GÖMLEK
FP-Growth - Jaccard Fonksiyonu	gömlek,tshirt,pantolon
FP-Growth - Cosine Fonksiyonu	gömlek,tshirt,pantolon
FP-Growth - Pearson Fonksiyonu	gömlek,tshirt,pantolon,sneakers,ayakkabı
TSHIRT, PANTOLON, KRAVAT →	GÖMLEK
FP-Growth - Jaccard Fonksiyonu	-
FP-Growth - Cosine Fonksiyonu	-
FP-Growth - Pearson Fonksiyonu	-
TSHIRT, SNEAKERS, PANTOLON, TRİKO →	GÖMLEK
FP-Growth - Jaccard Fonksiyonu	sweatshirt,gömlek,ayakkabı,mont,ceket
FP-Growth - Cosine Fonksiyonu	sweatshirt,gömlek,ayakkabı,mont,ceket
FP-Growth - Pearson Fonksiyonu	gömlek,ayakkabı,sweatshirt,mont
İlk alışveriş	İkinci alışveriş

İlk alışveriş	İkinci alışveriş
AYAKKABI, TSHIRT, SNEAKERS, TRİKO →	PANTOLON
FP-Growth - Jaccard Fonksiyonu	gömlek,pantolon,sweatshirt
FP-Growth - Cosine Fonksiyonu	gömlek,pantolon,sweatshirt
FP-Growth - Pearson Fonksiyonu	gömlek,pantolon,sweatshirt,mont
GÖMLEK, TRİKO, MONT →	TSHIRT
FP-Growth - Jaccard Fonksiyonu	-
FP-Growth - Cosine Fonksiyonu	-
FP-Growth - Pearson Fonksiyonu	-
GÖMLEK, JEAN, PANTOLON →	PANTOLON
FP-Growth - Jaccard Fonksiyonu	pantolon,ayakkabı,sneakers,tshirt,atkı
FP-Growth - Cosine Fonksiyonu	pantolon,ayakkabı,sneakers,tshirt,atkı
FP-Growth - Pearson Fonksiyonu	pantolon,ayakkabı,tshirt,sneakers,kravat
TRİKO, LOAFER →	GÖMLEK
FP-Growth - Jaccard Fonksiyonu	gömlek,tshirt,pantolon,ayakkabı,sneakers
FP-Growth - Cosine Fonksiyonu	gömlek,tshirt,pantolon,ayakkabı,sneakers
FP-Growth - Pearson Fonksiyonu	gömlek,tshirt,pantolon,ayakkabı,sneakers
GÖMLEK, SNEAKERS, SWEATSHIRT →	PANTOLON
FP-Growth - Jaccard Fonksiyonu	-
FP-Growth - Cosine Fonksiyonu	-
FP-Growth - Pearson Fonksiyonu	-
GÖMLEK, PANTOLON, TRİKO, SWEATSHIRT →	TSHIRT
FP-Growth - Jaccard Fonksiyonu	tshirt,sneakers,ceket,ayakkabı,bot
FP-Growth - Cosine Fonksiyonu	tshirt,sneakers,ceket,ayakkabı,bot
FP-Growth - Pearson Fonksiyonu	tshirt,sneakers,ayakkabı,bot,driver
GÖMLEK, TSHIRT, PANTOLON, SWEATSHIRT →	TRİKO
FP-Growth - Jaccard Fonksiyonu	sneakers,triço
FP-Growth - Cosine Fonksiyonu	sneakers,triço
FP-Growth - Pearson Fonksiyonu	sneakers,triço
GÖMLEK, SNEAKERS, SWEATSHIRT →	TSHIRT
FP-Growth - Jaccard Fonksiyonu	-
FP-Growth - Cosine Fonksiyonu	-
FP-Growth - Pearson Fonksiyonu	-
AYAKKABI, SNEAKERS, TRİKO →	PANTOLON
FP-Growth - Jaccard Fonksiyonu -	-
FP-Growth - Cosine Fonksiyonu -	-
FP-Growth - Pearson Fonksiyonu -	-
AYAKKABI, TSHIRT, SNEAKERS, TRİKO →	GÖMLEK
FP-Growth - Jaccard Fonksiyonu	gömlek,pantolon,sweatshirt
FP-Growth - Cosine Fonksiyonu	gömlek,pantolon,sweatshirt
FP-Growth - Pearson Fonksiyonu	gömlek,pantolon,sweatshirt,mont

Weka ile elde edilen Apriori sonuçları ile C# ile geliştirilen tavsiye sistemindeki Jaccard, Pearson ve Cosine benzerlik fonksiyonlarına göre çıkan sonuçların karşılaştırılması Çizelge 5. 5’de gösterilmiştir.

Çizelge 5. 5. Apriori Algoritması ile Diğer Algoritma Karşılaştırmaları

Alınan Ürün	Apriori	Jaccard	Cosine	Pearson
Gömlek, Tshirt, Pantolon	Triko, Mont	Ayakkabı, Sneakers, Triko	Ayakkabı, Sneakers, Triko	Ayakkabı, Sneakers, Triko
Triko, Mont	Gömlek, Tshirt, Pantolon	Pantolon Sweatshirt	Sweatshirt	Sneakers, Ayakkabı
Gömlek, Pantolon, Triko	Tshirt, Mont	Tshirt	Tshirt	Tshirt
Tshirt, Mont	Gömlek, Pantolon, Triko	Sneakers, Gömlek, Pantolon, Triko, Ayakkabı	Sneakers, Gömlek, Pantolon, Triko, Ayakkabı	-
Tshirt, Pantolo, Triko	Gömlek, Mont	Gömlek	-	-
Tshirt, Mont	Gömlek, Pantolon, Triko	-	Gömlek, Sneakers, Pantolon, Ayakkabı	-
Tshirt, Pantolon, Triko	Gömlek, Mont	Gömlek	Gömlek	-
Gömlek, Mont	Tshirt, Pantolon, Triko	Ayakkabı, Pantolon, Tshirt, Sneakers, Triko	Ayakkabı, Pantolon, Tshirt, Sneakers, Triko	Ayakkabı, Pantolon, Tshirt, Sneakers, Triko
Pantolon, Triko	Gömlek, Tshirt, Mont	Tshirt, Gömlek, Ayakkabı, Sneakers, Kravat, Bot, Mont	Tshirt	Tshirt, Gömlek, Ayakkabı, Sneakers, Kravat, Bot, Mont
Gömlek, Tshirt, Mont	Pantolon, Triko	-	-	-
Gömlek, Tshirt, Triko	Pantolon, Mont	Ayakkabı, Pantolon	Ayakkabı, Pantolon	Ayakkabı, Pantolon
Pantolon, Mont	Gömlek, Tshirt, Triko	Ayakkabı, Kravat, Triko	Kravat, Triko	-
Tshirt, Triko	Gömlek, Pantolon, Mont	Gömlek, Sweatshirt, Pantolon, Sneakers, Bot, Ayakkabı, Mont	-	Gömlek, Sweatshirt, Pantolon, Sneakers, Bot, Ayakkabı, Mont
Gömlek, Pantolon, Mont	Tshirt, Triko	-	-	-
Tshirt, Pantolon, Triko	Gömlek, Sweatshirt	Gömlek	Gömlek	Gömlek
Gömlek, Sweatshirt	Tshirt, Pantolon, Triko	Tshirt, Ayakkabı, Pantolon, Kravat	Tshirt, Kravat, Pantolon, Ayakkabı	Tshirt, Ayakkabı, Pantolon, Kravat
Gömlek, Triko	Tshirt, Pantolon, Mont	Tshirt, Ayakkabı, Kravat, Sneakers, Mont	-	Tshirt, Ayakkabı, Kravat, Sneakers, Mont
Tshirt, Pantolon, Mont	Gömlek, Triko	-	-	-
Gömlek, Tshirt, Sneakers	Pantolon, Triko	Pantolon	Pantolon	Pantolon
Pantolon, Triko	Gömlek, Tshirt, Sneakers	Tshirt, Gömlek, Ayakkabı, Sneakers, Kravat	Tshirt	Tshirt, Gömlek, Ayakkabı, Sneakers, Kravat
Tshirt, Triko	Gömlek, Pantolon, Sweatshirt	Gömlek, Sweatshirt, Pantolon, Sneakers, Bot, Ayakkabı, Mont	-	Gömlek, Sweatshirt, Pantolon, Sneakers, Bot
Gömlek, Pantolon, Sweatshirt	Tshirt, Triko	-	-	-

Weka ile elde edilen FP sonuçları ile oluşturulan tavsiye sistemindeki Jaccard, Pearson ve Cosine algoritmalarına göre çıkan sonuçların karşılaştırılması Çizelge 5.6’daki gibidir.

Çizelge 5. 6. FP-Growth Algoritması ile Diğer Algoritma Karşılaştırmaları

Alınan Ürün	FP- Growth	Jaccard	Cosine	Pearson
Pantolon, Triko, Mont	Tshirt	Tshirt, Gömlek	Gömlek, Tshirt	Gömlek, Tshirt
Sneakers, Triko, Mont	Gömlek	Gömlek, Sweatshirt, Şapka	Gömlek, Sweatshirt, Şapka	Gömlek, Tshirt, Pantolon , Sweatshirt, Şapka
Gömlek, Triko, Sweatshirt	Tshirt	-	-	-
Tshirt, Triko, Mont	Gömlek	-	-	-
Ayakkabı, Pantolon, Kravat	Gömlek	Gömlek	Gömlek	Gömlek
Gömlek, Ayakkabı, Sneakers,	Pantolon	Tshirt, Pantolon	Pantolon	Pantolon, Tshirt
Triko, Ceket	Gömlek	Gömlek, Tshirt, Pantolon	Gömlek, Tshirt, Pantolon	Gömlek, Tshirt, Pantolon, Sneakers, Ayakkabı
Tshirt, Pantolon, Kravat	Gömlek	-	-	-
Tshirt, Sneakers, Pantolon, Triko	Gömlek	Sweatshirt,Gömlek,Ayakkabı, Mont,Ceket	Sweatshirt, Gömlek, Ayakkabı, Mont, Ceket	Gömlek, Ayakkabı, Sweatshirt, Mont
Ayakkabı, Tshirt, Sneakers, Triko	Pantolon	Gömlek, Pantolon, Sweatshirt	Gömlek, Pantolon, Sweatshirt	Gömlek, Pantolon, Sweatshirt, Mont
Gömlek, Triko, Mont	Tshirt	-	-	-
Gömlek, Jean, Pantolon	Pantolon	Pantolon, Ayakkabı, Sneakers, Tshirt, Atkı	Pantolon , Ayakkabı , Sneakers, Tshirt, Atkı	Pantolon,Ayakkabı,Tshirt,Sneakers,Kravat
Triko, Loafer	Gömlek	Gömlek, Tshirt, Pantolon, Ayakabı, Sneakers	Gömlek, Tshirt, Pantolon, Ayakabı, Sneakers	Gömlek,Tshirt,Pantolon,Ayakkabı,Sneakers
Gömlek, Sneakers, Sweatshirt	Pantolon	-	-	-
Gömlek, Pantolon, Triko,	Tshirt	Tshirt, Sneakers, Ceket,	Tshirt, Sneakers, Ceket,	Tshirt, Sneakers, Ayakkabı, Bot, Driver
Gömlek, Tshirt, Pantolon,	Triko	Sneakers, Triko	Sneakers, Triko	Sneakers, Triko
Gömlek, Sneakers, Sweatshirt	Tshirt	-	-	-
Ayakkabı, Sneakers, Triko	Pantolon	-	-	-
Ayakkabı, Tshirt, Sneakers, Triko	Gömlek	Gömlek, Pantolon, Sweatshirt	Gömlek , Pantolon, Sweatshirt	Gömlek,Pantolon,Sweatshirt,Mont

K-means yöntemi kullanılarak yapılan kümeleme sonuçları ise Çizelge 5. 7’ ve Çizelge 5. 8’deki gibidir;

Müşterilerin yaş ve cinsiyetlerine göre küme dağılımları Çizelge 5.7’de belirtilmiştir;

Çizelge 5. 7. Yaş ve Cinsiyete Göre Kümeleme

	Erkek	Kadın
10-20	11	3
20-30	207	83
30-40	461	168
40-50	335	73
50-60	124	39
60-70	37	6
70-80	1	0
80-90	0	0
90-90+	0	1

Tabloyu yorumlamak gerekirse, 10-20 yaş grubundan 11 erkek, 3 kadın; 20-30 arasındaki yaş grubundan 207 erkek, 83 kadın; 30-40 arasındaki yaş grubundan 461 erkek, 168 kadın vardır şeklinde yorumlanabilir.

Yaş ve cinsiyete göre en çok alınan ürünler ise Çizelge 5.8’de belirtilmiştir;

Çizelge 5. 8. Yaş ve Cinsiyete Göre En Çok Alınan Ürün

	Erkek	Kadın
10-20	Tshirt	Tshirt, Ayakkabı, Triko
20-30	Tshirt	Sneakers
30-40	Gömlek	Ayakkabı
40-50	Gömlek	Ayakkabı, Gömlek
50-60	Ayakkabı, Gömlek	Ayakkabı, Pantolon
60-60+	Gömlek	Gömlek, Ayakkabı

Yukarıdaki tabloyu yorumlamak gerekirse; 10-20 yaş arasındaki erkekler tshirt almayı tercih ederken; 10-20 yaş arındaki kadınlar tshirt, ayakkabı ve trikoyu tercih etmektedir. 20-30 yaş grubundaki erkekler tshirt almayı tercih ederken, aynı yaş aralığındaki kadınlar sneakers almayı tercih etmektedirler. 30-40 yaş aralığındaki erkekler gömlek almayı tercih ederken, aynı yaş aralığındaki kadınlar ayakkabı almayı tercih etmektedirler. 40-50 yaş aralığındaki erkekler gömlek alırken kadınlar ayakkabı ve gömlek almaktadırlar. 50-60 yaş aralığındaki erkekler ayakkabı ve gömlek almayı tercih ederken, kadınlar ayakkabı ve pantolon almayı tercih etmektedirler. 60-60+ yaş grubundaki erkekler gömlek almayı tercih ederken, kadınlar gömlek ve ayakkabı almayı tercih etmektedirler.

Daha sonra Jaccard, Cosine ve Pearson benzerlik fonksiyonlarına göre yeni gelen bir müşteriye, yaş, cinsiyet ve geçmişteki alınmış ürünlere yakınlık açısından hangi kümeye girebileceği hesaplanarak ürün tavsiye edilmeye çalışılmıştır. Yeni gelen erkek müşteri için Çizelge 5. 9'da, yeni gelen kadın müşteri için Çizelge 5.10'da sonuçlar bulunmaktadır.

Çizelge 5. 9. 25 Yaş Erkek Müşteri Benzerlik Fonksiyonu Sonuçları

25 Yaş - Erkek Müşteri		
Benzerlik Fonksiyonu	İlk Ürün	İkinci Ürün
Jaccard Fonksiyonu	Triko	Gömlek, Tshirt, Pantolon, Sneakers, Sweatshirt
Cosine Fonksiyonu	Triko	Gömlek, Tshirt, Pantolon, Sneakers, Sweatshirt
Pearson Fonksiyonu	Triko	Gömlek, Tshirt, Pantolon, Sneakers, Sweatshirt

Çizelge 5. 10. 25 Yaş Kadın Müşteri Benzerlik Fonksiyonu Sonuçları

25 Yaş - Kadın Müşteri		
Benzerlik Fonksiyonu	İlk Ürün	İkinci Ürün
Jaccard Fonksiyonu	Triko	Gömlek, Tshirt, Pantolon
Cosine Fonksiyonu	Triko	Gömlek, Tshirt, Pantolon
Pearson Fonksiyonu	Triko	Gömlek, Tshirt, Pantolon

Yapılan analizlerin sonuçlarını test etmek amacıyla Çizelge 5.11'deki test sonuçları tüm veri kümesi ve kümelenmiş veri seti için; Jaccard, Cosine, Pearson, Apriori ve FP-Growth algoritmaları için test edilmiştir.

Çizelge 5. 11. Analiz Çalışma Testi

	Kümelenmiş Data	Tüm Data
Jaccard	Toplam: 274 ürün Önerilen: 178 ürün Önerilmeyen: 96 ürün %64.96 doğru önerme istatistiği.	Toplam: 274 ürün Önerilen: 159 ürün Önerilmeyen: 115 ürün %58.03 doğru önerme istatistiği.
Cosine	Toplam: 274 ürün Önerilen: 178 ürün Önerilmeyen: 96 ürün %64.96 doğru önerme istatistiği.	Toplam: 274 ürün Önerilen: 150 ürün Önerilmeyen: 124 ürün %54.74 doğru önerme istatistiği.
Pearson(Phi Katsayısı)	Toplam: 274 ürün Önerilen: 178 ürün Önerilmeyen: 96 ürün %64.96 doğru önerme istatistiği.	Toplam: 274 ürün Önerilen: 157 ürün Önerilmeyen: 117 ürün %57.30 doğru önerme istatistiği.
Apriori	Parametrelere bağlı olarak istatistikler değişmektedir.	Parametrelere bağlı olarak istatistikler değişmektedir.
FP-Growth	Parametrelere bağlı olarak istatistikler değişmektedir.	Parametrelere bağlı olarak istatistikler değişmektedir.

Yapılan test çalışmasına göre, kümelenmiş veride

- Jaccard: %65,
- Cosine: %65,
- Pearson: %65

Oranında doğru sonuç vermiştir.

Tüm veri kümesinde,

- Jaccard: %58,
- Cosine: %54,
- Pearson: %57

Oranında doğru sonuç vermiştir.

Apriori ve FP-Growth algoritmaları parametrelere göre değişiklik gösterdiği için istatistikleri değişmektedir.

6. SONUÇ

6.1. Sonuç ve Öneriler

Bu çalışmada, bir e-ticaret sitesine ait erkek alışveriş verileri kullanılmıştır. Uygulamaya, veri ön işleme ile başlanmıştır. Verideki alt grup ürün isimleri, ürün ile ilişkili olacak şekilde bir üst gruba taşınmıştır. Daha sonra veride 10 adetten az alışveriş yapmış olan kayıtlar tespit edilerek, veriden silinmiştir. Veri kümesi WEKA programında Apriori algoritması ve FP-Growth algoritması uygulanabilecek bir formata çevrilmiş, daha sonra lift değerine göre analize sokulmuştur. Apriori ve FP-Growth için ayrı ayrı birliktelik kuralları bulunmuştur. Apriori algoritması ile lift değerine göre bulunan kurallardan beş tanesi şu şekildedir;

- Gömlek, tshirt, pantolon alan 108 kişiden, 23'ü triko ve mont da almıştır,
- Gömlek, pantolon, triko alan 76 kişiden, 23'ü tshirt ve mont da almıştır,
- Tshirt, pantolon ve triko alan 73 kişiden, 23'ü gömlek ve mont da almıştır,
- Pantolon ve triko alan 121 kişiden, 23'ü gömlek, tshirt ve mont da almıştır,
- Gömlek, tshirt ve triko alan 91 kişiden, 23'ü pantolon ve mont da almıştır,

Fp-Growth algoritması ile lift değerine göre bulunan ilk on kural şu şekildedir;

- Gömlek, tshirt, pantolon, sweatshirt alan 32 kişiden, 25 'i triko da almıştır,
- Gömlek, ayakkabı, sneakers, triko alan 32 kişiden, 27'si pantolon da almıştır,
- Ayakkabı, tshirt, sneakers, triko alan 32 kişiden 26'sı pantolon da almıştır,
- Gömlek, jean, pantolon alan 44 kişiden, 36'sı pantolon da almıştır,
- Gömlek, sneakers, sweatshirt alan 33 kişiden 27'si pantolonda almıştır.

Daha sonra veriye, kümeleme analizi olan k-means uygulanmıştır. Kümeleme için verideki müşterilerin cinsiyet (kadın, erkek) ve yaş aralığı (20-30, 30-40..) değişkenleri kullanılmıştır. Aradaki benzerlikleri bulabilmek için, Jaccard, Cosine ve Pearson algoritmalarından yararlanılmıştır. Yapılan kümeleme işlemine göre sonuçlar şu şekildedir;

- 10–20 yaş aralığında 11, 20-30 yaş aralığında 207, 30-40 yaş aralığında 461, 40-50 yaş aralığında 335, 50-60 yaş aralığında 124, 60-70 yaş aralığında 37, 70-80 yaş aralığında 1 erkek müşteri bulunmaktadır.

- 10–20 yaş aralığında 3, 20-30 yaş aralığında 83, 30-40 yaş aralığında 168, 40-50 yaş aralığında 73, 50-60 yaş aralığında 39, 60-70 yaş aralığında 6, 90-90+ yaş aralığında 1 kadın müşteri bulunmaktadır.
- Erkek müşteriler için; 10-20 yaş aralığında tshirt, 20-30 yaş aralığında tshirt, 30-40 yaş aralığında gömlek, 40-50 yaş aralığında gömlek, 50-60 yaş aralığında ayakkabı ve gömlek, 60-60+ yaş aralığında gömlek alınmaktadır.
- Kadın müşteriler için; 10-20 yaş aralığında tshirt, ayakkabı, triko, 20-30 yaş aralığında sneakers, 30-40 yaş aralığında ayakkabı, 40-50 yaş aralığında ayakkabı, gömlek, 50-60 yaş aralığında ayakkabı ve pantolon, 60-60+ yaş aralığında gömlek ve ayakkabı alınmaktadır.

Analiz işlemleri gerçekleştikten sonra, yapılan analizin sonuçlarının doğruluğu test edilmiştir. Kasım, Aralık ve Ocak verileri test verisi olarak alınmıştır. Şubat ayındaki 100 müşteriye uygulanmıştır. Testte kümeleme verisi ve tüm veri için; pearson, jaccard ve cosine benzerlik fonksiyonları için uygulanmıştır ve şu sonuçlar elde edilmiştir;

- Kümelenmiş veride, Jaccard benzerlik fonksiyonuna göre;

Toplam 274 üründen 178 tanesi doğru önerilmiştir. Doğru önerme istatistiği %64.96'dır.

- Kümelenmiş veride, Cosine benzerlik fonksiyonuna göre;

Toplam 274 üründen 178 tanesi doğru önerilmiştir. Doğru önerme istatistiği %64.96'dır.

- Kümelenmiş veride, Pearson benzerlik fonksiyonuna göre;

Toplam 274 üründen 178 tanesi doğru önerilmiştir. Doğru önerme istatistiği %64.96'dır.

- Tüm veride, Jaccard benzerlik fonksiyonuna göre;

Toplam 274 üründen 159 tanesi doğru önerilmiştir. Doğru önerme istatistiği %58.03'tür.

- Tüm veride, Cosine benzerlik fonksiyonuna göre;

Toplam 274 üründen 150 tanesi doğru önerilmiştir. Doğru önerme istatistiği %54.74'tür.

➤ Tüm veride, Pearson benzerlik fonksiyonuna göre;

Toplam 274 üründen 157 tanesi doğru önerilmiştir. Doğru önerme istatistiği %57.30'dür.

Geçerli parametreler ile %71 doğru çalışma göstermektedir.

Çalışmadan elde edilen sonuçların firmanın pazarlama ekibi ile paylaşarak, ilgili ürünlerin ilgili kişilere tavsiye etmek üzerine bir kampanya çalışması yapılması ve satışları artırılması hedeflenmektedir.

KAYNAKLAR

- Alaeddinoğlu, M, F., 2012. Birliktelik Kuralları ile Van Gölü için Mekansal-Zamansal Veri Madenciliği, Atatürk Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi,74, Erzurum
- Aras, M, E., 2010. Birliktelik Kuralları ile Web Siteleri İçin Tavsiye Motoru Uygulaması, Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, 121, İstanbul
- Ay, D., 2009. Veri Madenciliği ve Apriori Algoritması ile Süpermarket Analizi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 137, Sakarya
- Başak, S., 2015. Türkçe Dokümanların Benzerliği. 27.04.2015
<http://www.selcukbasak.com/download/TurkceDokumanBenzerligi.pdf>
- Çelikyay, E, K., 2010. Metin Madenciliği Yöntemiyle Türkçede En Sık Kullanılan ve Birbirini Takip Eden Harflerin Analizi ve Birliktelik Kuralları, Beykent Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 95, İstanbul
- Darakçı, H, Ç., 2011.Kümeleme Analizi Kullanırken Benzin İstasyonlarının Operasyonel Değerlendirilmesi, Maltepe Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 96, İstanbul
- Demir, N., 2015. Association Rules Mining Plastic Pipe Sector. International Journal of Scientific Engineering and Applied Science, 4, 2395-3470
- Doğan, B., Erol, B., Buldu, A., 2014. Sigortacılık Sektöründe Müşteri İlişkileri Yönetimi İçin Birliktelik Kuralı Kullanılması, Marmara Fen Bilimler Dergisi,3, 105-114
- Erdem, S., Özdağoğlu, G., 2008. Ege Bölgesindeki Bir Araştırma ve Uygulama Hastanesinin Acil Hasta Verilerinin Veri Madenciliği ile Analiz Edilmesi, Anadolu Üniversitesi Bilim ve Teknoloji Dergisi, 9, 261-270
- Ergün, E., 2008. Ürün Kategorileri Arasındaki Satış İlişkisinin Birliktelik Kuralları ve Kümeleme Analizi ile Belirlenmesi ve Perakende Sektöründe Bir Uygulama, Afyon Kocatepe Üniversitesi, Sosyal Bilimler Enstitüsü, Doktora Tezi, 202, Afyonkarahisar
- Erpolot, S., 2015. Otomobil Yetkili Servislerinde Birliktelik Kurallarının Belirlenmesinde Apriori ve FP-Growth Algoritmalarının Karşılaştırılması. Anadolu Üniversitesi Sosyal Bilimler Dergisi,12, 151-166
- Fırat, M., Dikbaş, F., Koç, A., Güngör, M., 2012. K-Ortalamalar Yöntemi ile Yıllık Yağışların Sınıflandırılması ve Homojen Bölgelerin Belirlenmesi, İMO Teknik Dergi

- Gürgen, G., 2008. BirlikteKlik Kuralları ile Sepet Analizi ve Uygulaması, Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, 104, İstanbul
- Han, J., Kamber, M., 2011. Data Mining: Concepts and Techniques, Second Edition, Morgan Kaufmann Publishers
- Huang, L., 2007. FP-Growth Apriori Algorithm's Application In The Design Individualized Virtual Shop On The Internet. Proceedings of the Sixth International Conference on Machine Learning and Cybernetic. Hong Kong, 19-22 Ağustos.
- Kalikov, A., 2006. Veri Madenciliği ve Bir E-Ticaret Uygulaması, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 108, Ankara
- Karabatak, M., İnce, C., 2011. Apriori Algoritması ile Öğrenci Başarısı Analizi. Eleco International Conference on Electrical and Electronics Eng., 2004, Bursa.
- Kumar, S., Rukmani, K., 2010. Implementation of web usage mining using Apriori and FP Growth Algorithms. Int. J. of Advanced Networking and Applications, 6, 400-404
- Orhunbilge, N., 2002, Uygulamalı Regresyon ve Korelasyon Analizi, İstanbul Üniversitesi, 340, İstanbul
- Roşca, D., Radoiu, D., 2015. Step by Step Model For The Study Of The Apriori Algorithm For Predictive Analysis. Scientific Bulletin of the Petru Maior University of Tirgu Mureş, 12, 2286-3184
- Sezer, Ü., 2008. Karar Ağaçlarının BirlikteKlik Kuralları ile İyileştirilmesi, Kocaeli Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 77, Kocaeli
- Sorensen, S. Accuracy of Similarity Measures in Recommender Systems, 2012
- Şahin, A., Cankurt, M., Günden, C., Miran, B., 2013. Türkiye'de Tarımsal Kooperatifçiliğin İller Açısından Çok Boyutlu Analizi. Atatürk Üniversitesi Ziraat Fakültesi Dergisi, 44, 51-62
- Şahin, U., 2014. Trendden Arındırılmış Finans Verileri Üzerinde Evrimsel Algoritmalar ile Salınım-Tabanlı Teknik Analiz Göstergelerinin Parametre En İyilenmesi, TOBB Ekonomi ve Teknoloji Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 87, Ankara
- Şeker, S., 2013. Jaccard İndeksi (Jaccard Index). Erişim Tarihi: 20.04.2015. <http://bilgisayarkavramlari.sadievrenseker.com/2013/07/01/jaccard-indeksi-jaccard-index>
- Şekeroğlu, S., 2010. Hizmet Sektöründe Bir Veri Madenciliği Uygulaması, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 135, İstanbul

Taş, M., Adak, F., Yurtay, N., 2013. Öğrencilerin Staj Verileri Üzerine Uygulanan Apriori Algoritması ile Birliktelik Kurallarının Çıkarılması ve Staj Eğiliminin Belirlenmesi. 1st International Symposium On Innovative Technologies In Engineering And Science, 7-9 Haziran,2013, Sakarya

Timör, M., Ezerçe, A., Gürsoy, T., 2011. Müşteri Profili ve Alışveriş Davranışlarını Belirlemede Kümeleme ve Birliktelik Kuralları Analizi: Perakende Sektöründe Bir Uygulama. Yönetim-İstanbul Üniversitesi İşletme Fakültesi İşletme İktisadi Enstitüsü Dergisi, 68, 128-147

Uygun, O., 2014. Veri Madenciliği Nedir. Erişim Tarihi: 02.05.2015.
<http://www.muhendisbeyinler.net/veri-madenciligi-nedir>

Zhang, W., Liao, H., Zhao, N., 2008. Research on the FP Growth Algorithm About Association Rule Mining. International Seminar on Business and Information Management

Weng, S., Liu, S., Wu, T., 2011. Applying Bayesian Network and Association Rule Analysis For Product Recommendation

ÖZGEÇMİŞ

Adı Soyadı : Elif Şafak SİVRİ
Doğum Yeri ve Yılı : İstanbul, 11.06.1986
Medeni Hali :Bekar
Yabancı Dili : İngilizce
E-posta : elifsafaks@gmail.com

Eğitim Durumu

Lise : Marmara Koleji
Lisans : İstanbul Ticaret Üniversitesi, Fen Fakültesi, İstatistik
Yüksek Lisans : İstanbul Ticaret Üniversitesi,
Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim
Dalı

Mesleki Deneyim

Ipsos : Operasyon Raporlama Uzman Yardımcısı (2013-2014)
Boyner Holding : İş Zekası ve Raporlama Uzman Yardımcısı (2014-)