İSTANBUL TİCARET
ÜNİVERSİTESİ

**FEN BİLİMLERİ ENSTİTÜSÜ**

# E-TİCARET İÇİN ÜRÜN TAVSİYE SİSTEM GELİŞTİRMESİ

**WALEED ABDULLAH**

**Danışman**

**Dr. Öğr. Üyesi Mustafa Cem KASAPBAŞI**

**YÜKSEK LİSANS TEZİ**

**BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**

**İSTANBUL - 2019**

## KABUL VE ONAY SAYFASI

**Waleed abdullah** tarafından hazırlanan **"E-TİCARET İÇİN ÜRÜN TAVSİYE SİSTEM GELİŞTİRMESİ "** adlı tez çalışması 08/02/2019 tarihinde aşağıdaki jüri üyeleri önünde başarı ile savunularak, İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü **BİLGİSAYAR MÜHENDİSLİĞİ Anabilim Dalı**'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

| | |
|---|---|
| **Danışman** | **Dr.Öğr. Üyesi Mustafa Cem KASAPBAŞI**<br>İstanbul Ticaret Üniversitesi |
| **Jüri Üyesi** | **Doç. Dr. Serhat Özekes**<br>İstanbul Üsküdar Üniversitesi |
| **Jüri Üyesi** | **Dr. Öğr. Üyesi Ali Boyacı**<br>İstanbul Ticaret Üniversitesi |

**Onay Tarihi :** 11.02.2019

**Prof. Dr. Necip ŞİMŞEK**
**Enstitü Müdürü**

## AKADEMİK VE ETİK KURALLARA UYGUNLUK BEYANI

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

- 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,
- ve bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

Tarih 11.02.2019

İmza

**Tez Yazarının Adı Soyadı**

WALEED ABDULLAH

# İÇİNDEKİLER

## ÖZET

**Yüksek Lisans Tezi**

**E-TİCARET İÇİN ÜRÜN TAVSİYE SİSTEM GELİŞTİRMESİ**

**Waleed Abdullah**

**İstanbul Ticaret Üniversitesi**
**Fen Bilimleri Enstitüsü**
**Bilgisayar Mühendisliği Anabilim Dalı**

**Danışman : Dr. Öğr. Üyesi Mustafa Cem KASAPBAŞI**

E-Ticaret çağında, öneri sistemleri, e-ticaret web sitelerinin temel gereksinimleri arsındadır. Bu sistemlerin doğruluğu ve verimliliği işin ana ilgi alanıdır. Bu faktörleri ölçmek için bazı popüler teknikler üzerinde analiz yapılmıştır. Bu çalışmada, 1023 adet ürün arasında yarım milyon adet Türk Özel İnşaat Perakende Satış İşletmesi satış hareketleri kullanılmıştır. Öğe-öğe işbirlikçi filtrelemenin (IF) ve sık örüntü madenciliğinin (Frequent Pattern Mining-FPM) detaylı bir değerlendirmesi, FPM için IF, Apriori ve FPGrowth algoritması için sırasıyla Kosinüs, Jaccard ve Pearson benzerlik işlevleri kullanılarak yapılmıştır. İlk olarak, benzerlik matrisleri daha sonra ham verilerle hesaplanır, veri modeline yeni artırılmış özellikler eklendikten sonra, benzerlik matrisleri tekrar hesaplanır. Hesaplanan benzerlik matrislerine ilişkin önerileri önermek için en yakın K komşu (KNN) algoritması uygulanır. Sonuçlar, önerilen veri modelimizi kullanarak Kosinüs ve Jaccard'taki hassasiyet skorunun sırasıyla 0.05 ve 0.2'lik önemli iyileşme kaydığını göstermiştir. WEKA Yazılımı ve GraphLab Kütüphanesi kullanılarak FPM'den yararlanmak için başka bir öneri karşılaştırması yapılmıştır. Sonuçlar, Jaccard benzerliği ve FP-Büyüme algoritmasının analizimizde en iyisi olduğunu göstermektedir.



***Anahtar Kelimeler:*** *E-Ticaret; En Yakın Komşu (EYK);* İşbirlikçi Filtrelemenin*; Sık Örüntü madenciliği Tavsiye Sistemler*

## ABSTRACT

**M.Sc. Thesis**

**PRODUCT RECOMENDATION SYSTEM DEVELOPMENT FOR E-COMMERCE**

**Waleed Abdullah**

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

**Supervisor: Assistant Prof. Dr. Mustafa Cem KASAPBAŞI**

In this new era of E-Commerce, recommendation systems are main requirement of e-commerce websites. Accuracy and efficiency of these systems are the core concern of business. To measure these factors, we have performed analysis on some of the popular techniques. In this study half a million transactions of Turkish Private Construction Retail company were used amongst 1023 products. A detail evaluation of item-item collaborative filtering (CF) and frequent pattern mining (FPM) have been carried out using Cosine, Jaccard and Pearson similarity functions for CF, Apriori and FPGrowth algorithm for FPM respectively. Initially, the similarity matrices are calculated with raw data later, after adding new augmented attributes to the data model similarity matrices are calculated again. K nearest neighbor (KNN) algorithm is applied to propose the recommendations regarding calculated similarity matrices. Results has shown the significant improvement shift of precision score in Cosine and Jaccard of 0.05 and 0.2 respectively by using our proposed data model. An other recommendation comparison is carried out to utilize FPM using WEKA Software and GraphLab Library. Results indicates that Jaccard similarity and FP-Growth algorithm were the best among our analysis.



***Key Words:*** *Collaborative Filtering; E-Commerce; Frequent pattern Mining; K nearest neighbor (KNN); Recommendation system;*

## TEŞEKKÜR

Bu yüksek lisans tezinin öğrenme süreci boyunca faydalı yorumlar, açıklamalar ve katılım için değerli danışman hocam Dr. Öğr. Üyesi Mustafa Cem KASAPBAŞI'na şükranlarımı sunarım.

Waleed Abdullah

ISTANBUL, 2019

## ŞEKİLLER

## ÇİZELGELER

**Sayfa**

## SİMGELER VE KISALTMALAR

| | |
|---|---|
| İF | İşbirlikçi Filtreme |
| PSA | Pazar Sepeti Analizi |
| SVD | Singular Vector Decomposition |
| RFM | Recency Frequency monetary |
| LSI | Latent Semantic Indexing |
| SAE | Saklı Anlamsal Endeksleme |
| PCA | Principal Component Analysis |
| KDD | Knowledge Discovery in Database |
| Arff | Attribute-Relation File Format |
| SQL | Structured Query Language |
| ETL | Extract Transform and Load |
| KNN | K-Yakın Komşu |
| OMH | Ortalama Mutlak Hatasıdır |
| MAE | Mean Absolute Error |
| FPM | Frequent Pattern Mining |
| İDF | İçeriğe Dayalı Filtreleme |
| CBF | Content Based Filtering |
| CF | Collaborative Filtering |
| ROC | Receiver Operator Characteristic |

# 1 GİRİŞ

Günümüzde, veriler elektronik ortamda kolayca bulunabilir ya sosyal medya veri setleri veya bir ticaret sitesi satış verisi olabilir. İnsanlar kararlar almak için sözlü kelimeler, tavsiye mektupları ve haber medyasından gelen haberler, anketler, seyahat rehberleri vb. gibi diğerlerinin önerilerine güvenirler. Tavsiye sistemleri, bu doğal sosyal sürece insanların kitap, makale, web sayfası, film, müzik, restoran, oyuncak, yiyecek ürünleri seçme konusunda yardımcı olmalarına yardımcı olmaktadır. Erken öneri sistemi duvar halısı, daha sonra İşbirlikçi Filtreleme (İF) olarak geliştiriciler arasında popülerlik kazanmış olan kural temelli danışmanlardan ve kullanıcı özelleştirmeden oluşur (Kumar, Rukmani, S,K 2010).
İF'nin temel kuralı, müşteri tarafından verilen benzer derecelendirme/öğeye göre hareket etmenin yanı sıra, ürünü tavsiye etmek (satın alma, görüntüleme, dinleme) gibi bazı temel özelliklere de tepki vermektir. Genellikle önerilen bir sistemin oluşturulmasında kullanılan, bir veya iki algoritma içeren öneri listesidir.

E-ticarette, Satış Veri Kümeleri, istatistiksel yöntemler ve yaklaşımlar kullanarak müşteriler, ürünler ve satışlar arasındaki örüntüleri bulmak için kolayca kullanılabilir. Veri madenciliği aynı zamanda geleneksel istatistiksel veri analizi uygulama alanıdır ve bir dizi analitik teknik içerir. Veri madenciliği, verilerin temizlenmesi, hazırlanması ve sonuçların görselleştirilmesi ve gerekli hedeflere ulaşılması için gelecekteki sonuçları öngörmek gibi çeşitli temel işlemlerden oluşur. (Mayuri Dalvi, Prof S.V Gumaste 2015)

Bu çalışmadaki problem beyanımız, inşaat sektöründe müşterilere ilgi duyabilecekleri ürünlerinlerin bulunmasını sağlayacak tavsiye sistemi geliştirilmesi olacaktır. Bu hedefle tüm veri setini farklı algoritmalar, teknikler kullanarak analiz etmek ve sistemin ürünü müşteriye başarıyla önerebileceği sonuçları almaktır.

Bu sorunun üstesinden gelmek için iki farklı yöntemler kullanarak sonuçlar çıkarılmıştır.

Birincisi, en iyi olanı önerme arayışındaki tavsiye algoritmalarını karşılaştırmak, ikincisi ise ürünleri müşterilere tanımlamak ve tavsiye etmektir. Bu analiz için öğe-öğe işbirliği filtrelemesini kullandık ve sonuçları karşılaştırmak için sıklık örüntü madenciliği kullandık.

Bu çalışmada, Türk Perakende Şirketinin elektrikli ürünler satan veri setini kullandık. Bu çalışmada, 1023 adet ürün arasında yarım milyon adet Türk Özel İnşaat Perakende Satış İşletmesine ait veriler kullanılmıştır. İlk başta, verileri SQL veritabanından bir seri ön işlemlere tabi tutulmuşlardır. Gereksiz veri tablolarını kaldırılmış ve bu veriler üzerinde algoritmaların yapılabileceği biçimde verileri temizlenmiş ve ayarlanmıştır.

İşbirlikçi filtreleme için, verileri Phyton içinde uygun veri yapılarına aktarıldıktan sonra ve filtreleme uygulanmıştır. Bu veri tabanı ile ilgili teknik bilgiler ilgili bölümde ayrıntılı şekilde verilmektedir.

# 2 LİTERATÜR ARAŞTIRMASI

Tavsiye sistemi bugünlerde o kadar yaygın bir şekilde kullanılmaktadır ve araştırmacılar için tercih edilen bir seçenek haline gelmiştir. Tavsiye sistemi üzerine ilk makale 1998 yılında yayınlanmıştır. O zamandan beri önemli sayıda makale yayınlandı. Bu çalışma temel olarak bir tavsiye sistemi geliştirmek üzerine olduğu için, literatür araştırmasında bu yöne ağırlık verilmiştir. Tavsiye sisteminin güvenilirliğini artırmak için farklı faktörler açıklanmıştır.

2005 yılında (John O'Donovan, Barry Smyth 2005), bir profilin genel olarak yaptığı doğru tahminlerin yüzdesi olarak (profil düzeyinde güven) veya belirli bir maddeye (madde düzeyinde güven) ilişkin olarak güven duymuştur. Yazarlar, bu farklı güven değerlerinin standart bir işbirlikçi filtreleme algoritmasına dahil edilebileceği ve her birinin denenmiş ve test edilmiş bir kıyaslama yaklaşımına ve standart bir veri setine göre değerlendirilebileceği çeşitli yollar tanımlamıştır. Bu, tahmin hatasını% 22 oranında azalttığını tespit edilmiştir.

2007 yılında, danışman sisteminde "sınırlayıcı algoritmasını etkileme" fikrini önerdi. Bu algoritma, araştırmamızın alakasız sonucunu gösteren saldırıları önler. Bu algoritma, bir saldırganın değiştirebileceği içerik sayısını sınırlar. (Paul Resnick, Rahul Sami 2007).

Aynı yıl University College Dublin'in çevrimiçi kayıt başvurusu için bir kurs danışmanlığı sisteminin geliştirilmesini önerdi. Bu uygulamayı destekleyen makalesinde, tarihsel öğrenci kayıt verilerini kullanarak devam eden yaklaşımın ampirik olarak değerlendirildiğini ve önerilen tasarımla umut verici performansın elde edildiğini göstermektedir. Yine 2007 yılında, anlamsal ağ için (Punam Bedi, Harmeet Kaur,Sudeep Marwaha 2007) tarafından güvene dayalı bir öneri sistemi önerildi. Ontolojiler biçiminde depolanan bilgileri kullanan bir öneri sisteminin tasarımının açıklaması verilmiştir ( Michael P. O'Mahony 2007 ).

Öneriler üretmek için akranlar arasındaki etkileşimler, aralarında var olan güven ağını temel alır. Akran ajanlar tarafından verilen bir ürünle ilgili tavsiyeler, üyelik derecesini, üyelik dışı ve belirsizliği kullanarak belirtilen Sezgisel Bulanık Kümeler şeklindedir. Literatürde, tavsiye veren sistemler öneriler üretmek için veri tabanlarını kullanır. Burada açıklanan tavsiye sistemi, ontolojilere Semantik Web için açıklamalı içerik oluşturmak için bir bilgi temsil tekniği kullanır.

Bhagya Ramesh ve Reeba R tarafindan 2017 yılında yazılan " Secure Recommendation System for E-Commerce Website " isimli makalede; tavsiye sistemlerinin doğruluğunu ve uygulanabilirliğini geliştirmek için işbirlikçi filtreleme kullanarak sosyal faktörleri, kişisel ilgiyi, benzerlikleri ve bireysel ürün puanlarını birleştirerek kişiselleştirilmiş bir öneri yaklaşımı önermiştir. (Bhagya Ramesh ve Reeba R, 2017).

Weikang Xue, Bopin Xiao ve Lin Mu 2015 yılında tarafından yazılan " Intelligent Mining on Purchase Information and Recommendation Systems for E-commerce " isimli makalenin hedefi ; üç popüler tavsiye modeli işbirlikçi filtreleme modelini, geliştirilmiş işbirlikçi filtreleme modelini ve hibrit öneri modelini karşılaştırmak. Böylece hibrit modelin daha uygulanabilir ve daha doğrulanmış olduğunu kanıtlanmıştır.(Weikang, Bopin ve Lin, 2015)

Hyunwoo Hwangbo, Yang Sok Kim, Kyung Jin Cha, tarafından yazılan" Recommendation system development for fashion retail e-commerce " adlı makalede Tipik ürün tabanlı işbirlikçi filtreleme algoritmasını genişleten, K-RecSys adlı yeni bir sistem önerdi. K-RecSys, müşterilerin çevrimiçi ve çevrimdışı tercihlerini yansıtmak için ağırlıklandırılmış çevrimiçi ürün tıklama verilerini ve çevrimdışı ürün satış verilerini birleştirir.

Aynı zamanda, zaman içindeki tercihlerdeki değişiklikleri yansıtmak için bir tercih edilen çürüme fonksiyonu benimsemiştir ve nihayet ürün kategorisi bilgisini kullanan ikame ve tamamlayıcı ürünler önermektedir. K-RecSys'i sadece

çevrimiçi verilerle uygulanan mevcut işbirlikçi filtreleme sistemi ile karşılaştırmak için gerçek işletim ortamında bir A / B testi yaptık. Deneysel sonuçlarımız, önerilen sistemin çevrimiçi alışveriş merkezindeki ürün tıklamaları ve satışları açısından üstün olduğunu ve bunun yerine verilen önerilerin tamamlayıcı önerilerden daha sık benimsendiğini göstermektedir. (Hyunwoo Hwangbo, Yang Sok Kim, Kyung Jin Cha, 2018)

Faryal Ali, Tauqir Ahmad, Aslam Muhammad ve Martinez-Enriquez A.M tarafından 2015 yılında yazılan " Data Mining Based Recommendation System using Social Websites " adlı makalede; öneri sağlamak için işbirlikçi ve içerik tabanlı filtreleme kullanarak bir kullanıcı tavsiye sistemini tanıttı. Sosyal ağ sitelerinin veri ayıklanması, soğuk başlangıç ve aşırı kişiselleştirme sorunları çözülmüştür. Sonuçlar, sırasıyla bilgi alma matrisleri Precision, Fallout ve F1 skoru kullanılarak değerlendirildi(Faryal, Tauqir, Aslam, 2015).

Tomasevic, Nikola & Paunović, Dejan & Vraneš, Sanja tarafindan 2019 yılında yazılan "User-based collaborative filtering approach for content recommendation in OpenCourseWare platforms " adlı makalede, İçerik öneri modülünün kavramsal fikrini sunmak, kullanıcı, ilgili etkinlikleri, tercihleri, türü ve içerik benzerliğini, vb. dikkate alarak ilgili güverteleri (sunumlar, eğitim materyalleri, vb.) önerme yeteneğine sahiptir. Özellikle kullanıcı odaklı İF yaklaşımının ve içerik değerlendirmeyle ilgili kullanıcı ile ilgili özelliklerin uygulanması için uygun teknikleri analiz eder. Önerilen yaklaşım, içerik önerisi için bütünsel ve etkili bir çözüm sağlamak amacıyla, kullanıcı tabanlı ve içerik tabanlı yaklaşımların bir kombinasyonu olarak bir karma öneri sistemi de öngörmektedir. Değerlendirme ve test amacıyla, SlideWiki geliştirme OCW platformunun bir parçası olarak belirlenmiş bir içerik öneri modülü uygulandı (Tomasevic, Nikola & Paunović, Dejan & Vraneš 2019).

Usmani, Shraddha, ve arkadaşları tarafından 2017 yılında yazılan “A Predictive Approach for Improved the sales of products in e-Commerce “ isimli makalede; tavsiyelerin tekniklerinin karşılaştığı çeşitli sorunları ve bunların üstesinden gelmek için önerilen çözümleri tartışır. Tavsiye tekniğini kullanma yaklaşımları sırasıyla “Ortak filtreleme için demografik sınıflar kullanma”, “Kullanıcıları farklı tiplere ayırma”, İlgili Sınıflandırma ve İlişkilendirme Kuralı Madenciliği (Apriori Algoritma) idi (Z. A. Usmani, Shraddha, Tahreem, Ayman 2017).

Xiaofeng Yuan, Lixin Han,Subin Qian,Guoxia Xu,Hong Yan 2019 tarafından yazılan “Singular value decomposition based recommendation using imputed data ” adlı makalede sezgisel verileri SVD çerçevesine dahil etmek için yeni bir yöntem (ISVD) önerdi. ISVD ayrıca, emsal veri oluşturmak için kullanıcıların veya komşuların etkili komşularını seçmek için yeni bir algoritma önermektedir. ISVD, tüm SVD tabanlı öneri yöntemleri için kullanışlıdır. Bu makalede dört gerçek veri kümesi üzerinde birkaç deney yapıyorlar: MovieLens 100k, MovieLens 1M, Netflix ve Filmtrust. Deney sonuçları, ISVD'nin modern CF'lerden daha iyi performans gösterdiğini ve ISVD'nin RMSE / MAE'lerinin diğer atfedici yöntemlere ve SVD bazlı yöntemlerden %10'dan daha iyi olduğunu göstermektedir (Xiaofeng ,Lixin,Subin,Guoxia ,Hong 2019).

Yuri Stekh, Mykhoylo Lobur ve Vitalij Artsibasov tarafindan 2015 yılında yazılan “Methods and Tools for Building Recommender Systems“ isimli makalenin hedefi tavsiye sistemlerinin mevcut durumlarını ve uygulama sistemlerini, model sistemlerini ve tavsiye sistemlerinin yapım yöntemlerini analiz etmek. 3 yöntem geliştirilmiştir. Birincisi, kompakt kümelenme kullanarak ürünlerin tahmin değerlerinin doğruluğunu artıran bulanık bir sistem geliştirmektir. İkincisi, tavsiyeler için arama ilişkilendirici öğeler sorununu aşmak için bir ilişkilendirme kural madenciliği yöntemi oluşturmaktır. Son kategorik kümeleme, kullanıcı ve yeni madde sorununu çözmek için geliştirilmiştir (Yuri Stekh, Mykhoylo Lobur, Vitalij Artsibasov 2015).

Kwei Tang, Yen-Liang Chen, ve Hsiao-Wei Hu tarafından 2008 yılında yazılan " Data Mining Based Recommendation System using Social Websites " adlı makalede; Çok mağaza ortamındaki işlem kayıtlarından ilişkilendirme kurallarını ayıklamak için yeni bir yaklaşım önermektedir. Uygulama için, algoritma, zaman ve mekânın konsept hiyerarşileri kullanılmıştır. Bu yaklaşımı, gerçek zamanlı Çin firma veri seti üzerinde gerçekleştirilmiştir. Konsept hiyerarşileri veya TP kafes inşa etmek için kümelenme ve segmentasyon araçlarını kullanma da dahil olmak üzere bu makalenin birkaç uzantısını önerdiler. Bu yaklaşımın faydası, kullanıcıyı yeterli zaman ve yer hiyerarşileri sağlaması için rahatlatmaktır (Kwei Tang, Yen-Liang Chen , Hsiao-Wei 2008).

Ammar Jabakji ve Hasan Dağ tarafından 2016 yılında yazılan " Improving item-based recommendations accuracy with user's preferences on Apache Mahout " isimli makalede, tüketici ürün incelemelerinin derecelendirme sistemini kullanan yeni bir veri modeli önermiştir. Performansı değerlendirmek için, gerçek bir kelime veri kümesi kullanarak Amazon MAHOUT öğe tabanlı benzerlik ölçülerini kullanılmıştır. Sonuç olarak, doğru sonuçlarda ortalama yüzde 4.6 iyileşme ve ortalama öneriler bulmuşlardır. Ek olarak, öklid uzaklığı benzerlik ölçüsünün, diğer madde madde temelli benzerlik ölçümleri arasında en iyi performans gösteren matris olduğunu da gözlemlemişlerdir (Ammar Jabakji, Hasan Dağ 2016)

Zhongyi Hu, Liangzhon ve Shengkai tarafından 2015 yılında yazılan makalede, Yaygın olarak kullanılan öneriler sistemi yöntemlerini analiz eder ve eksikliklerine dikkat çekerek e-ticaret için gelişmiş bir Apriori tabanlı personel tavsiye sistemi algoritması önerir. İlk önce, müşterek işbirlikli filtrelemeyi kullanarak benzer ilgi alanlarına sahip müşterileri bulur, ardından güncel sayfa ziyaretçileri ekleyerek sanal veri toplama havuzu olarak geçmişin sayfasını ziyaret edip ve ikinci olarak tüm veri kümesine ilişkilendirme kuralı madenciliği yapmaktadır. Bu şekilde, sistem kullanımının ilk aşamasında ziyaretçiyi sık kullanılan erişim sayfasını ilişkilendirme kuralları madenciliği ile alabilir. Kullanıcı

benzerlik matrisi için, kosinüs bazlı benzerlik, korelasyon benzerliği ve ayarlanmış kosinüs bazlı benzerlik kullanmışlardır (Zhong Hu, Liangzhon, Shengkai 2015).

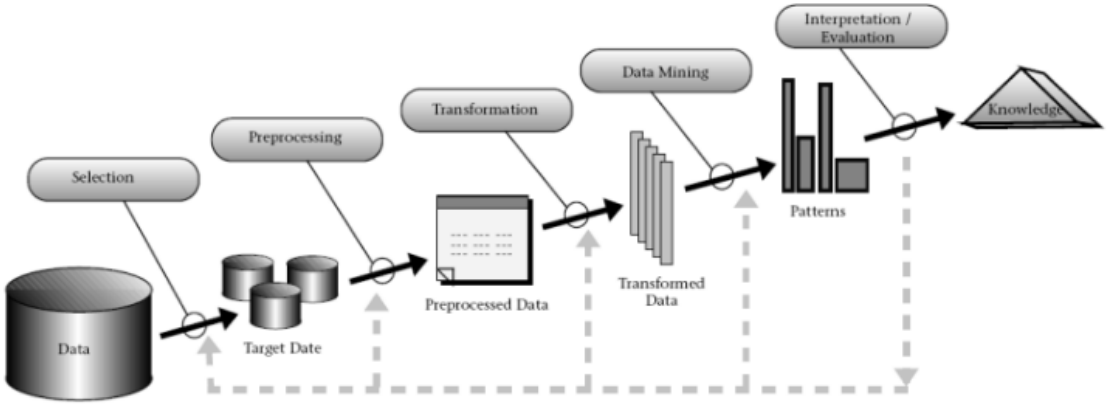
# 3 VERİ MADENCİLİĞİ

## 3.1 Veri Madenciliği Genel Bakış

Bu yeni teknoloji çağında, veri internet üzerinden kolayca kullanılabilir. Bu veriler istatistiksel yaklaşımları kullanarak geleceği öngörmek için de kullanılabilir. Veri madenciliği, geleneksel istatistiksel veri analizinin bir dalıdır ve farklı disiplinlerden alınan çok sayıda analitik teknik içermektedir. Veri madenciliği, tam bir veri analizi süreci, sonuçların temizlenmesi, hazırlanması ve görselleştirilmesini içerir ve gerçek zamanlı gelecek sonuçları tahmin etmek için gerekli hedeflerin yerine getirilmesini sağlar. Farklı veri türleri ve veri darlığı nedeniyle, Veri madenciliği, çok sayıda davranışta tarif edilebilir. Veri madenciliği, örüntü tanıma teknolojilerinin yanı sıra istatistiksel ve matematiksel teknikleri kullanarak, depolarda depolanan büyük miktarda veriyi eleyerek anlamlı yeni korelasyonları, örüntüleri ve eğilimleri keşfetme sürecidir. (Larose, pp.xi, 2005). Veri Madenciliği, diğerleri arasında Makine Öğrenimi, Veri Tabanı Teknolojisi, İstatistik, Matematik, Kümeleme ve Görselleştirme kavramlarını içeren disiplinler arası bir alan olarak düşünülebilir. (Sarabjot S. Anand and John G, 1998)

Bu tezin bölümlerini tanımlamak gerekirse, birinci bölüm, Veri madenciliği, veri madenciliğinin kullanımı, veri madenciliğinin kullanımı, veri madenciliğinin ne zaman kullanılacağı, tez konusu ile ilgili benzer çalışmalar ve veri madenciliğini kullanabileceğimiz bazı uygulamalardan oluşmaktadır. İkinci bölüm, bir inşaat web sitesinin e-ticaret verilerini analiz etmek için kullanılması gereken veri tekniklerini ve algoritmaları anlatılır. Üçüncü bölümde, eldeki veriler üzerinde yapılan veri işleme sürecinin, veri özelliklerini açıklanması ve sonuçlara ulaşmak için veri madenciliğinin temel aşamalarını uygulanması anlatılmıştır.

Verimli veri madenciliği elde etmek için, izlenecek bazı temel adımlar aşağıdaki şekil 3.1 gibi vardır.




Şekil 3.1: Veri madenciliğinin adımlar (Bharati and Ramageri, 2010)

### 3.1.1 Veri seçim

Veri Seçimi'nde, Verileri herhangi bir kaynaktan çıkarırız ve veriler ham formatta olabilir ve sadece veri madenciliğinde yararlı olabilecek verileri seçebiliriz. Ön işlem ve temizlik aşamasında, Veri özelliklerini iyice analiz ediyoruz, eksik veri, çift değerler, gürültü veya tutarsız veri problemlerini kontrol edilir. Bu sorunları gidermek için farklı teknikler uygulanır: eksik, tutarsız verilerin silinmesi, eksik değerlerin yerine ortalamayı temsil edecek değerler konuşması, eksik değerler için bir regresyon modeli oluşturulup tahminler yapılması vb. sayılabilir (Han et al. 2011).

### 3.1.2 Veri temizleme ve ön işleme

Veri temizleme, bir kayıt kümesinden, tablodan veya veri tabanından bozuk veya yanlış kayıtları algılama ve düzeltme (veya kaldırma) işlemidir ve verilerin eksik, yanlış, alakasız bölümlerini tanımlamak ve sonra değiştirerek, değiştirmek, veya kirli veya kaba verileri silebilmeyi içerir. Veri temizleme, veri yazma araçlarıyla etkileşimli olarak veya komut dosyası aracılığıyla toplu işlem olarak gerçekleştirilebilir.

Veri kalitesini iyileştirmek için bazı adımlar vardır.

**Ayrıştırma**: Sözdizimi hatalarının tespiti için kullanılır . Bir ayrıştırıcı ile, bir veri dizesinin izin verilen veri belirtiminde kabul edilebilir olup olmadığına karar verir. Bu, ayrıştırıcının dilbilgisi ve dilleriyle çalışmasına benzer.

**Yinelenen eleme**: Yinelenen algılama, verilerin aynı varlığın yinelenen temsillerini içerip içermediğini belirlemek için bir algoritma gerektirir. Genellikle, veriler daha hızlı tanımlama için çift girişleri birbirine yaklaştırır.

**İstatistiksel yöntemler**: Verileri ortalama, standart sapma, aralık veya kümeleme algoritmalarını kullanarak analiz ederek, bir uzmanın beklenmedik ve hatalı olan değerleri bulması mümkündür. Gerçek değer bilinmediği için bu tür verilerin düzeltilmesi zor olsa da, değerleri ortalama veya başka bir istatistiksel değere ayarlayarak çözülebilir. İstatistiksel yöntemler, genellikle kapsamlı veri büyütme algoritmalarıyla elde edilen bir veya daha fazla makul değerle değiştirilebilen eksik değerleri işlemek için de kullanılabilir.

### 3.1.3 Veri dönüşümü

Verileri bir formattan veya yapıdan başka bir format veya yapıya dönüştürme işlemidir. Veri dönüşümü, gerekli olan dönüşümün karmaşıklığına dayanarak ihtiyaç duyulduğunda uygulanabilir olan aşağıdaki adımlara ayrılabilir

- Veri bulma
- Veri haritalama
- Kod üretimi
- Kod yürütme
- Veri incelemesi

Adımlar şu şekilde tarif edilebilir:

**Veri bulma**, veri dönüştürme sürecindeki ilk adımdır. Tipik olarak veriler, profil oluşturma araçları kullanılarak veya bazen verilerin yapısını ve karakteristiklerini daha iyi anlamak ve nasıl dönüştürüleceklerine karar vermek için elle yazılmış profil oluşturma komut dosyaları kullanılarak profillenir.

**Veri haritalama**, istenen nihai çıktıyı üretmek için bireysel alanların nasıl haritalandığını, değiştirildiğini, birleştirildiğini, filtreleneceğini, toplanmasını vb. tanımlama sürecidir. Geliştiriciler veya teknik veri analistleri, geleneksel olarak, dönüşüm kurallarını (ör., Görsel ETL (Extract transform load) araçları, dönüştürme dilleri) tanımlamak için belirli teknolojilerde çalıştıkları için veri haritalamayı gerçekleştirir.

**Kod üretimi**, verileri istenen ve tanımlanmış veri eşleme kurallarına dayalı olarak dönüştürecek, çalıştırılabilir kod (ör. SQL, Python, R veya diğer yürütülebilir komutlar) oluşturma işlemidir. Tipik olarak, veri dönüştürme teknolojileri, geliştiriciler tarafından tanımlanan tanımlara veya meta verilere dayanarak bu kodu üretir.

**Kod yürütme**, üretilen kodun istenen çıktıyı yaratmak için verilere karşı çalıştırıldığı adımdır. Yürütülen kod, dönüştürme aracına sıkıca entegre edilebilir veya geliştirici tarafından oluşturulan kodu manuel olarak yürütmek için ayrı adımlar gerektirebilir.

**Veri incelemesi**, çıktı verilerinin dönüşüm gereksinimlerini karşıladığından emin olmaya odaklanan sürecin son adımıdır. Bu adım, genellikle, bu adımı gerçekleştiren verilerin iş kullanıcısı veya son son kullanıcısıdır. Dönüştürme sürecinde uygulanacak yeni gereksinimler olarak geliştiriciye veya veri analistine geri gönderilen ve iletilen verilerdeki herhangi bir anormallik veya hata.

### 3.1.4 Veri madenciliği teknikleri

Veri madenciliği altı ortak sınıftan oluşur.

**Anomali tespiti** Olağandışı veri kayıtlarının tanımlanması, ilginç olabilecek veya daha fazla araştırma gerektiren veri hataları olabilir.

**Birliktelik kural öğrenimi** Değişkenler arasındaki ilişkileri arar**.** Örneğin, bir süpermarket müşteri satın alma alışkanlıkları hakkında veri toplayabilir. Birliktelik kuralı öğrenimini kullanarak, süpermarket hangi ürünlerin sıkça satın alındığını belirleyebilir ve bu bilgileri pazarlama amacıyla kullanabilir.
Bu bazen market sepeti analizi olarak adlandırılır.

**Kümelenme** verideki bilinen yapıları kullanmadan, bir şekilde veya başka bir "benzer" olan verilerdeki veri ve yapıları keşfetme görevidir.

**Sınıflandırma** yeni verilere uygulanacak bilinen yapıyı genelleştirme görevidir. Örneğin, bir e-posta programı bir e-postayı "meşru" veya "spam" olarak sınıflandırmaya çalışabilir.

**Regresyon** Veri veya veri kümeleri arasındaki ilişkileri tahmin etmek için, verileri en az hata ile modelleyen bir işlev bulmaya çalışır.

**Özetleme** görselleştirme ve rapor oluşturma dahil olmak üzere veri kümesinin daha kompakt bir temsilini sağlamak için kullanılır.

### 3.1.5 Değerlendirme

Beş aşamalı sürecimizin bu son aşaması, bilginin, temel sayısal sayıları, doğrudan değer karşılaştırması veya belirli öğeleri seçmek için grup karşılaştırması gibi daha eşit niteliklere dönüştürülmesini içerir. Örneğin, otel

odası puanlarıyla basit bir sıralama yaygındır, daha karmaşık karşılaştırmalı sıralama ise ürünlerle kullanılabilir. Bireysel ürünler benzer özelliklere sahip eşit gruplara göre veya üst satıcılar ile karşılaştırılabilir. Daha önceki aşamalarda çıkardığınız veriler nihai sonuca birleştirilebilir.

Veri madenciliği basit bir süreç değildir ve veriye sistematik ve matematiksel bir şekilde yaklaşmaya dayanır. Ancak, aynı zamanda esnek olmaya ve iyi organize edilmiş ve sıralı bir biçime uymayabilecek veriler almaya dayanır.

## 3.2 Veri Madenciliği Uygulama Alanları

Veri madenciliği sıklıkla aşağıda belirtilen sektör ve alanlarda kullanılır.

**Ticaret :**

- Yakınlık, Frekans, Parasal (Recency Frequency Monetary)
- Müşteri değeri analizi (Customer Value Analysis)
- Satış tahmini (Sales Forecasting)
- Market Sepeti analizi (Market Basket Analysis)
- Tavsiye sistemleri (Recommendation Systems)
- Çapraz satış
- Hedef Pazarlama
- Müşteri ilişkileri pazarlama

**Risk değerlendirmesi :**

- Kredi kartı güncellemeleri
- Ev kredileri
- Müşteri tutma
- Kredi derecelendirme

**Dolandırıcılık Tespiti :**

- Kredi kartı sahteciliği
- İç denetimler
- Depo tadilatı

# 4 TAVSİYE SİSTEMLERİNİN TEKNİKLERİ VE MODELLERİ

## 4.1 Tavsiye Sistemi

Günlük hayatta insanlar konuşulan kelimeler, referans mektupları, haber medyasından haberler, genel anketler, seyahat rehberleri vb. gibi diğer insanlardan gelen tavsiyelere güvenirler. Tavsiye sistemleri, insanların mevcut kitaplar, makaleler, web sayfaları, filmler, müzik, restoranlar, espriler, bakkal ürünleri ve benzerleri ile en ilginç ve değerli bilgileri bulmalarına yardımcı olmak için bu doğal sosyal sürece yardımcı olur ve destekler. İlk öneri sistemlerinden birinin geliştiricileri (Tapestry), (daha önceki öneri sistemleri, kural tabanlı danışmanlar ve kullanıcı özelleştirme içerir), öneri sahipleri tarafından yaygın olarak benimsenen "işbirlikçi filtreleme (İF)- collaborative filtering (CF)" ibaresini ortaya koydu. İF'nin temel varsayımı, kullanıcılar ve benzer öğeleri benzer şekilde veya benzer davranışları varsa (örneğin, satın alma, izleme, dinleme) ve dolayısıyla diğer öğeleri benzer şekilde derecelendirecek veya bunlara göre davranacaklardır (Kumar, Rukmani, S,K 2010).

Sonraki kısımlarda İşbirlikçi Filtreleme ve Sıklık Örüntü Model Madenciliği gibi veri madenciliğinde bazı popüler öneriler tekniklerini tartışılacaktır.

## 4.2 İşbirlikçi Filtreleme:

Tavsiye sistemi yapımında en başarılı yaklaşımlardan biri olarak, işbirlikçi filtreleme (İF), diğer kullanıcılar için bilinmeyen tercihlerin önerilerini veya tahminlerini yapmak için bir grup kullanıcının bilinen tercihlerini kullanır. İF teknikleri, kullanıcıların yeni bir kullanıcının beğenebileceği ek konuları veya ürünleri tahmin etmeleri için kullanıcılar tarafından tercihler veri tabanı kullanır. Tipik bir İF senaryosunda, $m$ kullanıcı için listesi $\{u_1 + u_2,,,,,+u_m\}$ ve $n$ öğelerinin listesi $\{i_1 + i_2,,,,,+i_n\}$ ve her kullanıcı, $u_i$ öğelerinin bir listesi vardır.

$Iu_i$, kullanıcının oy verdiği veya tercihlerinin davranışları aracılığıyla çıkarıldığı, derecelendirmeler, açık göstergeler olabilir, örneğin 1-5 ölçekte veya satın almalar veya tıklamalar gibi gizli göstergeler olabilir (B.N.Miller, J.A. Konstan, J.Riedl 2004). Örneğin Şekil 3.1 deki, insanların ve beğendikleri veya sevmedikleri filmlerin listesini kullanıcı-öğe derecelendirme matrisine dönüştürebiliriz. Burada Tony, tavsiyelerini yapmak istediğimiz aktif kullanıcıdır. 4.1'nin Matriste, kullanıcıların belirli öğeler için tercihlerini vermediği eksik değerler vardır.

**(a)**

**Alice**: (beğendi) Shrek, Snow White, (beğenmedi) Superman
**Bob**: (beğendi) Shrek, Snow White, (beğenmedi) Superman
**Chris**: (beğendi) Spiderman, (beğenmedi) Snow White
**Tony**: (beğendi) Shrek, (beğenmedi) Spiderman

**(b)**

| | **Shrek** | **Snow White** | **Spider-man** | **Super-man** |
|---|---|---|---|---|
| **Alice** | beğendi | beğendi | | beğenmedi |
| **Bob** | | beğendi | beğenmedi | beğendi |
| **Chris** | | beğenmedi | beğendi | |
| **Tony** | beğendi | | beğenmedi | ? |

Şekil 4. 1: Kullanıcı-öğe derecelendirme matrisine ( B.N.Miller, J.A. Konstan, J.Riedl 2004)

Erken nesil işbirlikçi filtreleme sistemleri, kullanıcılar veya öğeler arasındaki benzerlik veya ağırlığı hesaplamak için kullanıcı derecelendirme verilerini kullanır ve hesaplanan benzerlik değerlerine göre tahminler veya öneriler yapar. Bellek tabanlı İF yöntemleri, özellikle kullanımı kolay ve oldukça etkili oldukları için özellikle http://www.amazon.com/ ve Barnes and Noble gibi ticari sistemlere yerleştirilmiştir (G. Linden, B. Smith, J. York 2003). Her bir kullanıcı için İF sistemlerinin özelleştirilmesi, kullanıcılar için arama çabasını azaltır. Ayrıca, daha fazla müşteri sadakati, daha yüksek satışlar, daha fazla reklam geliri ve

hedeflenen promosyonların yararı için daha iyi olacağı anlaşılmıştır (A. Ansari, S. Essegaier, R. Kohli 2000).

Bununla birlikte, bellek tabanlı İF tekniklerinde, benzerlik değerlerinin ortak öğelere dayandığı ve bu nedenle verilerin seyrek olduğu ve ortak öğelerin bu az olduğu durumlarda güvenilmez olduğu gibi bazı sınırlamalar vardır. Daha iyi tahmin performansı elde etmek ve bellek tabanlı İF algoritma eksikliklerini gidermek için model tabanlı İF yaklaşımları araştırılmıştır. Model tabanlı İF teknikleri, tahmin yapmak, bir modeli belirlemek veya öğrenmek için saf derecelendirme verilerini kullanırlar (J. Breese, D. Heckerman, C. Kadie 1998). Model bir veri madenciliği veya makine öğrenimi algoritması olabilir. İyi bilinen model tabanlı İF teknikleri,( X. Su, T. M. Khoshgoftaar 2006) Bayes inanç filterleri (BNs) İF modelleri, İF modellerini (L. H. Ungar, D. P. Foster 1998) ve saklı semantik İF modellerini içerir. Bir MDP (Markov Decision Process Markov Karar Süreci) temelli İF sistemi , tavsiye uygulamamış bir sistemden çok daha yüksek bir kâr üretir (G. Shani, D. Heckerman, R. I. Brafman 2005).

İşbirlikçi filtrelemenin yanı sıra, içerik tabanlı filtreleme de önemli bir tavsiye sistemidir. İçeriğe dayalı öneri sistemleri, metinsel içeriğin içeriğini analiz ederek ve içerikte düzenlilikleri bularak önerilerde bulunur. İF ve içerik tabanlı öneri sistemleri arasındaki en büyük fark, İF'nin yalnızca kullanıcı-öğe derecelendirme verilerinde tahminler ve öneriler yapmak için kullanmasıdır; içerik tabanlı öneri sistemleri ise kullanıcıların özellikleri ve tahminler için kullanılan öğelere dayanır. Hem içerik tabanlı öneri sistemleri hem de İF sistemleri sınırlamalara sahiptir. İF sistemleri açık bir şekilde özellik bilgisini içermese de, içerik tabanlı sistemler bilgileri tercihen bireyler arasında tercih benzerliğine dahil etmemektedir. Bir kaç İF teknikleri aşağıdaki Şekil 4.2 verildiği gibidir.

| İF kategorileri | Temsil Edilen Teknikler | ana avantaj | ana eksiklikler |
|---|---|---|---|
| **bellek tabanlı İF** | * Komşu tabanlı İF<br>* Öğe tabanlı en yüksek tavsiyeler | * Kolay uygulama<br>* Yeni veriler kolayca eklenebilir<br>* Ortak puanlama öğeleriyle iyi ölçeklendirme | * İnsani notlara bağlı<br>* Yeni kullanıcılar için tavsiye edemez<br>* Veri seyrek olduğunda performans düşmesi sınırlı ölçeklenebilirlik |
| **model tabanlı İF** | * Bayesian inanç ağları İF<br>* Clustering İF<br>* MDP-tabanlı İF<br>* Gizli anlamsal İF | * Seyrekliği ele almak daha iyi<br>* Tahmin performansını iyileştirmek<br>* Öneriler için sezgisel bir rasyonel vermek | * Pahalı model oluşturma<br>* Tahmin performansı ve ölçeklenebilirlik arasında takas<br>* Boyutsallık teknikleri için yararlı bilgileri kaybeder |
| **Karma tavsiye modelleri** | * İçerik tabanlı İF<br>*İçerik artırıldı İF<br>* Hibrit İF birleştiren teknikler | * İF sınırlamalarının üstesinden gelmek<br>* Tahmin performansını iyileştirmek<br>* Gri koyun gibi sorunların üstesinden gelmek | * Uygulama için karmaşıklığı ve maliyeti arttırmış olması<br>* genellikle mevcut olmayan harici bilgilere ihtiyaç duymak |

Şekil 4. 2: İF teknikleri (G. Shani, D. Heckerman, R. I. Brafman 2005)

İF algoritmalarını değerlendirmek için, İF uygulama türlerine göre uygun ölçümleri kullanmamız gerekir. Sınıflandırma hatası yerine, İF'nin tahmin performansı için en çok kullanılan değerlendirme ölçeği, Ortalama Mutlak Hatasıdır (OMH) (MAE-Mean Absolute Error). Keskinlik (Precision) ve geri çağırma (Recall), bilgi çıkarım araştırmasında iade edilen öğelerin sıralı listeleri için yaygın olarak kullanılan metriklerdir. ROC (Receiver Operator Characteristic) duyarlılığı genellikle karar destek doğruluğu ölçümü olarak kullanılır, 1 e yakın değerler tercih edilir.

### 4.3 İşbirlikçi Filtrelemenin Özellikleri

E-ticaret tavsiyesi algoritmaları, özellikle eBay ve Amazon gibi büyük çevrimiçi alışveriş şirketleri için zordur. Genellikle, hızlı ve doğru tavsiyelerde bulunan bir öneri sistemi müşterilerin ilgisini çekecek ve şirketlere fayda sağlayacaktır. İF sistemleri için, yüksek kaliteli tahminler veya öneriler üretmek, İF görevlerinin özellikleri olan zorlukların ne kadar iyi ele alındığına bağlıdır. Bu zoruluklar veri

seyrekliği, ölçeklenebilirlik, eş anlamlılık ve gri koyun belirlenmiştir. Alt kısımlarda kısaca bu zorluklar hakkında bilgi verilecektir.

### 4.3.1 Veri Seyrekliği

Uygulamalarda, çok büyük ürün gruplarını değerlendirmek için birçok ticari tavsiye sistemi kullanılmaktadır. İşbirlikçi filtrelemede kullanılan kullanıcı-öğe matrisi bu nedenle son derece seyrek olacaktır ve İF sistemlerinin tahminlerinin veya tavsiyelerinin performanslar etkileyecek. Veri kısıtlılığı zorluğu birkaç durumda ortaya çıkar, özellikle yeni bir kullanıcı veya öğe sisteme girdiğinde gözükür, bu da soğuk başlatma problemi olarak bilinen problemi ortaya çıkarır. Çünkü yeni eklenen öğe yada kullanıcı içi benzer bilgileri bulmak zordur. Bazı literatürlerde, soğuk başlangıç problemi yeni kullanıcı problemi veya yeni öğe sorunu olarak da adlandırılır ( G. Adomavicius and A. Tuzhilin 2005).

Veri seyrekliği sorununu hafifletmek için birçok yaklaşım önerilmiştir. Bazı kullanıcılar değerlendirene kadar, yeni öğeler önerilemez ve yeni kullanıcılara iyi öneriler verilmez (K. Yu, A. Schwaighofer, V. Tresp, X. Xu, and H.-P. Kriegel 2004). Tekil Değer Ayrıştırma (SVD-Singular Vector Decomposition) gibi boyut azaltma teknikleri, kullanıcı öğesi matrisinin boyutlarını doğrudan azaltmak için temsilci, önemsiz kullanıcıları veya öğeleri kaldırır. Bilgi almada kullanılan patentli Saklı Anlamsal Endeksleme (LSI-Latent Semantic Indexing), SVD'ye dayanmaktadır ( S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman 1990, D. Billsus and M. Pazzani 2004).

Burada, kullanıcılar arasındaki benzerlik, azaltılmış alandaki kullanıcıların temsili ile belirlenir. Goldberg ve diğleri ilk olarak 1901'de Pearson tarafından tanımlanan ve boyut azaltmak için tanımlanan yakın ilişkili bir faktör analiz tekniğini (PCA-Principal Component Analysis) uygulayan özdeğeri geliştirmiştir. Bununla birlikte, belirli kullanıcılar veya öğeler atıldığında, bunlarla ilgili

tavsiyeler için yararlı bilgiler kaybolabilir ve öneri kalitesi düşebilir (K. Goldberg, T. Roeder, D. Gupta, and C. Perkins 2001).

### 4.3.2 Ölçeklenebilirlik

Mevcut kullanıcıların ve öğelerin sayısı muazzam bir şekilde büyüdüğünde, geleneksel İF algoritmaları, hesaplama kaynaklarının pratik veya kabul edilebilir düzeylerin ötesine geçmesiyle birlikte, ciddi ölçeklenebilirlik sorunlarına maruz kalacaktır. Örneğin, on milyonlarca müşteri (M) ve milyonlarca farklı katalog öğesi (N) ile yaklaşık O(M+N) karmaşıklığı olan bir İF algoritması zaten çok büyüktür.
Bunun yanı sıra, birçok sistemin, bir İF sisteminin yüksek ölçeklendirilebilir olmasını gerektiren satın alma ve derecelendirme geçmişine bakılmaksızın, çevrimiçi gerekliliklere anında tepki vermesi ve tüm kullanıcılar için önerilerde bulunması gerekir (G. Linden, B. Smith, and J. York 2003).

Öğeye dayalı Pearson korelasyon İF algoritması gibi bellek tabanlı İF algoritmaları, tatmin edici ölçeklenebilirlik sağlayabilir. Öğelerin tüm çiftleri arasındaki benzerlikleri hesaplamak yerine, öğe tabanlı Pearson İF, bir kullanıcı tarafından yalnızca eşleştirilen öğeler çifti arasındaki benzerliği hesaplar (B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl 2001). Naive Bayes İF algoritması, gözlenen değerlere dayanarak tahminler yaparak ölçeklenebilirlik problemini ele alır ( K. Miyahara and M. J. Pazzani 2002). İF algoritmalarının kümelenmesi gibi model tabanlı İF algoritmaları, tüm veri tabanı yerine küçük ve oldukça benzer kümeler içinde kullanıcılara öneri sunarak ölçeklenebilirlik sorununu ele alınabilir. Ölçeklenebilirlik ve tahmin performansı arasında bir denge durumu olduğu gözükmektedir (S. H. S. Chee, J. Han, and K. Wang 2001).

### 4.3.3 Eş anlamlılık

Eş anlamlılık, aynı veya çok benzer öğelerin sayısının farklı adlara veya girdilere sahip olma eğilimine işaret eder. Çoğu öneri sistemleri bu gizli ilişkiyi keşfedemez ve dolayısıyla bu ürünleri farklı şekilde ele alabilir. Örneğin, görünüşte farklı olan “çocuklar filmi” ve “çocuk filmi” öğeleri aynı öğedir, ancak bellek tabanlı İF sistemleri, benzerlikleri hesaplamak için aralarında hiçbir eşleşme bulamazlar.
Gerçekten de, tanımlayıcı terim kullanımındaki değişkenlik derecesi, yaygın olarak şüphelenilenden daha büyüktür. Eş anlamlılıkların yaygınlığı, İF sistemlerinin öneri performansını azaltır.

Önceki eş anlamlılık problemini çözme girişimleri, entelektüel veya otomatik terim genişlemesine ya da bir eş anlamlıların inşasına bağlıydı. Bazı ilave terimlerin amaçlanandan farklı anlamlara sahip olabileceği ve böylece tavsiye performansının hızlı bir şekilde bozulmasına yol açabileceği tam otomatik yöntemler için dezavantaj oluşturmaktadır (S. K. Jones 1972).

SVD teknikleri, özellikle de LSI yöntemi, eş anlamlı problemlerle başa çıkabilme yeteneğine sahiptir. SVD, bir terim-belge ilişkilendirme matrisi alır ve birbiriyle yakından ilişkili olan terimlerin ve belgelerin birbiriyle yakından bağlantılı olduğu bir semantik alanı oluşturur.

SVD, semantik alandaki düzenlemenin verilerdeki ana ilişkisel modelleri yansıtmasına ve daha küçük, daha az önemli olanları görmezden gelmesine izin verir. LSI'nin eşanlamlılık problemini ele almadaki performansı, keskinliğin genellikle oldukça düşük olduğu daha yüksek geri çağırma seviyelerinde etkilidir, dolayısıyla büyük orantılı gelişmeleri temsil eder. Bununla birlikte, LSI yönteminin en düşük geri çağırma düzeyindeki performansı zayıftır (S.

Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman 1990)

### 4.3.4 Gri koyun

Gri koyun, görüşleri herhangi bir grup insana tutarlı olarak katılmayan ve dolayısıyla işbirlikçi filtrelemeden faydalanamayan kullanıcıları ifade eder. Gri koyun, kendine özgü tatları neredeyse imkansız kılan karşıt gruptur. Bu, öneri sistemindeki bir başarısızlık olmakla birlikte, elektronik olmayan öneri sahiplerinin de bu durumlarda büyük problemleri vardır, bu yüzden gri koyun kabul edilebilir bir başarısızlıktır (M. Claypool, A. Gokhale, T. Miranda et al 1999)

Claypool ve diğ. içerik temelli tahminlerin ve İF tahmininin ağırlıklı ortalaması üzerine bir tahminde bulunmak suretiyle, içerik tabanlı ve İF önerilerini birleştiren bir karma yaklaşım sağlamıştır. Bu yaklaşımda, içeriğe dayalı ve İF öngörülerinin ağırlıkları, kullanıcı başına belirlenir ve sistemin, gri koyun sorununu çözmeye yardımcı olmak için her kullanıcı için içeriğe dayalı ve İF'nin en uygun karışımını belirlemesine olanak tanır (M. Claypool, A. Gokhale, T. Miranda et al 1999)

## 4.4 Bellek Tabanlı İşbirlikçi Filtreleme Teknikleri

Bellek tabanlı İF algoritmaları, bir tahmin oluşturmak için, kullanıcı öğesi veritabanının tamamını veya bir örneğini kullanır. Her kullanıcı benzer ilgi alanlarına sahip bir grup insanın parçasıdır. Yeni bir kullanıcının (veya aktif kullanıcının) sözde komşularını tanımlayarak, kendisi için yeni öğeler üzerinde tercihlerin bir tahmini üretilebilir.

Komşu tabanlı İF algoritması, yaygın bellek tabanlı İF algoritması, aşağıdaki adımları kullanır:

İki kullanıcı veya iki öğe arasındaki mesafeyi, korelasyonu veya ağırlığı yansıtan $w_{i,j}$ benzerliği veya ağırlığı hesaplanır, $i$ ve $j$; Kullanıcı veya öğenin belirli bir öğe veya kullanıcı üzerindeki tüm ağırlıklarının ağırlıklı ortalamasını alarak veya basit ağırlıklı bir ortalama kullanarak etkin kullanıcı için bir tahmin üretebilir (B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl 2001 ).

Bir N adet önerisi oluşturulması istediğinde benzerliklerin hesaplanmasından sonra, en benzer kullanıcı veya öğeleri (en yakın komşuları) bulunlamı ve daha sonra N adet sık kullanılan öğeleri tavsiye olarak almak için komşuları bir araya getirilebilir.

### 4.4.1 Benzerlik hesaplaması

Öğeler veya kullanıcılar arasındaki benzerlik hesabı, bellek tabanlı işbirlikçi filtreleme algoritmalarında önemli bir adımdır. Öğe tabanlı İF algoritması için öğe $i$ ve $j$ arasındaki benzerlik hesabının temel fikri, bu iki öğenin de önermiş kullancılar üzerinde çalışmak ve sonra $w_{i,j}$ benzerliği hesaplamak için bir benzerlik hesabını iki ortak önerilmiş öğe kullancıları arasında uygulanır. Kullanıcı tabanlı bir İF algoritması için, önce $u$ ve $v$ aynı kullanıcılara sahip olan öğelere arasındaki $w_{u,v}$ benzerliğini hesaplanır (B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl 2001). Kullanıcılar veya öğeler arasında benzerlik veya ağırlık hesaplamak için birçok farklı yöntem bulunmaktadır. Bunların başlıcaları, Kosinüs Benzerliği, Jaccard Benzerliği ve Pearson Benzerliğidir. Sonraki alt kısımlarda bu benzerlikler anlatılacaktır.

#### 4.4.1.1 Kosinüs benzerliği

İki belge arasındaki benzerlik, her bir belgenin, sözcük frekanslarının bir vektör olarak işlenmesi ve frekans vektörlerinin oluşturduğu açının kosinüsünün hesaplanmasıyla ölçülebilir. Bu yöntem, kelime frekansları yerine belge, ayrıca

derecelendirme yerine kullanıcı veya öğeleri kullanan işbirlikçi filtrelemede benimsenebilir.

Biçimsel olarak, eğer $R(mXn)$, $m > n$ kullanıcı-öğe matrisi ise, o zaman, $A$ve $B$, iki öğe arasındaki benzerlik, matrisin $A$. ve $B$. sütununa karşılık gelen $n$ boyutlu vektörlerin kosinüsü olarak denklem 4.1 gibi hesaplanır (Zhu, X. et al 2018) (Pang et al. 2013) .

$$w_{A,B} = cos(A,B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (4.1)$$

“.”, iki vektörün nokta çarpınını gösterir. İstenen benzerlik hesaplamasını elde etmek için, $n$ öğeler için bir $n * n$ benzerlik matrisi hesaplanmıştır. Örneğin, vektör $A = \{x_1 + y_1\}$ise, vektör $B = \{x_2 + y_2\}$ ise , A ve B arasındaki vektör kosinüs benzerliği aşağıda gösterildiği denkleme 4.2 gibidir

$$w_{A,B} = cos(A,B) = \frac{A \cdot B}{\| A \| \| B \|} = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2}\sqrt{x_2^2 + y_2^2}} \quad (4.2)$$

#### 4.4.1.2 Jaccard benzerlik

Jaccard benzerliği (Jaccard 1902, Jaccard 1912) ikili değişkenler için ortak bir endekstir. İki nesne arasındaki kesişen nokta ile ikili olarak karşılaştırılan değişkenlerin birliği arasındaki bölüm olarak tanımlanır. Sonlu örnek kümeleri arasındaki benzerliği ölçer ve kesişimin boyutu, örnek kümelerinin birliği boyutuna bölünmüş olarak tanımlanır: Jaccard benzerliği formülü denkleme 4.3 gibidir (Sven Kosub, 2016) (Pang et al. 2013) .

$$J(A,B) = \frac{AveBninkesişenöğesayısı}{Kesişmeyenöğelerçıkarıldığındakalanlarıntümü} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} \quad (4.3)$$

$f_{11}$ A ve B'nin kesişen öğe sayısını

$f_{10}$ A'nın ilişkili olduğu ve B'nin olmadığı öğe sayısını

$f_{01}$ A'nın ilişkili olmadığı ve B'nin ilişkili olduğu öğe sayısını temsil etmektedir.

#### 4.4.1.3 Pearson benzerlik

X ve Y gibi iki değişken arasındaki doğrusal korelasyonun bir ölçüsüdür. +1 ve −1 arasında bir değere sahiptir, burada 1 toplam pozitif doğrusal korelasyon, 0 doğrusal korelasyon olmadığını ve −1 toplam negatif lineer korelasyon olduğu anlamındadır. Pearson benzerliği formülü denkleme 4.4 gibidir (Pang et al. 2013).

$$corr(X,Y) = \frac{cov(X,Y)}{\sigma x \sigma y} \tag{4.4}$$

cov X ve Y nin kovaryansıdır ve σx, X'in standart sapmasıd ve σy, Y'nin standart sapmasıdır.

### 4.4.2 En yakın komşu algoritması

K-En Yakın Komşu algoritması (KNN), sınıflandırma veya regresyon için çok basit bir algoritmadır. KNN, denetlenen öğrenme algoritması ailesine düşer. Bu, eğitim gözlemlerinden $(x, y)$ oluşan etiketli bir veri kümesi verdiğimiz anlamına gelir ve $x$ ile $y$ arasındaki ilişkiyi belirlenmesi amaçlanmaktadır. Amacımız bir işlevi öğrenmek, $h: X \rightarrow Y$, görünmeyen bir gözlem x verildiğinde $x$ karşılık gelen çıkış $y$ 'yi güvenle tahmin edebilmektir.

KNN sınıflandırıcı ayrıca parametrik olmayan ve örnek tabanlı bir öğrenme algoritmasıdır.

- Parametrik olmayan araçlar, h'nin fonksiyonel formu hakkında açık bir varsayım yapmaz, verilerin altta yatan dağıtımını yanlış modellemenin tehlikelerinden kaçınır. Örneğin, verilerimizin Gauss olmayan bir derece olduğunu varsayalım, ancak seçtiğimiz öğrenme modeli bir Gauss formu olduğunu varsayar. Bu durumda algoritmamız çok zayıf tahminler yapar.

- Örnek tabanlı öğrenme, algoritmamızın açık bir şekilde bir model öğrenmediği anlamına gelir. Bunun yerine, tahmin aşaması için daha sonra “bilgi” olarak kullanılan eğitim örneklerini ezberlemeyi seçer. Somut olarak, bu, yalnızca veri tabanımıza bir sorgu yapıldığında (yani, bir girdi verilen bir etiketi tahmin etmesini istediğimizde), algoritmanın, eğitim örneklerini bir cevabı tükürmek için kullanacağı anlamına gelir.

Sınıflandırma ayarında, K-en yakın komşu algoritması esasen, benzeri olmayan bir gözleme en benzer (yakın) K örnekler arasında, çoğunluk oyu oluşturmaya dayanır. Benzerlik, iki veri noktası arasındaki bir uzaklık metriğine göre tanımlanır. Bu metrik için popüler bir seçim, Öklid uzaklığının kullanılmasıdır.

## 4.5 İçeriğe dayalı filtreleme

İDF (CBF-Content Based Filtering), belirli bir kullanıcının geçmişte beğendiklerine benzer öğeler önermeye çalışır. Temel işlem, kullanıcı tercihlerini öğe özellikleriyle eşleştirerek gerçekleştirilir. Bu nedenle, bu sistemler öğeleri temsil etmek ve kullanıcı tercihlerini belirlemek için uygun tekniklerin yanı sıra, kullanıcı tercihlerini öğe temsilleriyle karşılaştırmak için stratejiler gerektirir.

İDF algoritmaları, kullanıcılara öğelerin tanımlarına ve kullanıcı tercihlerine göre uygun öğeleri önerir. Ayrıca, İDF algoritmaları yalnızca aktif kullanıcı için profil

bilgilerini veya derecelendirmeleri kullanır ve bu nedenle diğer kullanıcılardan gelen derecelendirmelerin sayısı yeterince büyük olmasa bile doğru öneriler üretebilirler. Bununla birlikte, İDF 'nin bazı sakıncaları vardır. Bir öğe için analiz edilen içerik kategorizasyon için uygun bilgi içermiyorsa, İDF uygun öneriler üretemez. Bu sınırlamayı ele almak için, İDF için farklı özelliklere farklı önem dereceleri atanan özellik ağırlıklandırma önerilmiştir (Debnath, Ganguly, & Mitra, 2008).

Örneğin, bir cep telefonu seçerken, fiyat renginden daha önemli olabilir. Bununla birlikte, bu yaklaşım, genellikle uzmanlık sorunu olarak adlandırılan sadece birkaç özelliğe dayanarak, benzerliği tekrar tekrar hesaplayarak öğeler önerildiğinde kullanıcılara taraflı sonuçlar sağlama eğilimindedir.
Bu sorunları gidermek için bazı çalışmalar anlam bilimsel analizlere ontolojik bilgileri dahil etmeye çalışmıştır. Ancak, ölçeklenebilirlik ve seyreklik sorunu, büyük miktarda veri matris tabanlı yaklaşımlar kullanılarak hesaplandığında ortaya çıkabilir (Di Noia et al. 2012)

## 4.6 İlişkilendirme Kuralları Veri Madenciliği

Son yıllarda iletişim teknolojisi ve internet teknolojisi hızla gelişmektedir. Bu bağlamda, insanlar özellikle verilerin zamana ve değerle ilgili bilgilerinin önemine ve bu bilgilerin bilgiye ulaşmasına ve veri madenciliğine daha fazla önem veriyorlar. Veri madenciliği teknolojisinin birçok önemli dalı vardır. İlişkilendirme kuralları veri madenciliği teknolojisinin en büyük avantajı, büyük miktarda bilgiden değerli, ancak daha az anlamlı olan kalıpları türetme kabiliyetidir. İlişkilendirme kural madenciliğinin temel problemi, sıklıkla beraber olan öğe kümelerinin madenciliğinin yapılmasıdır. İlişkilendirme kuralının temel amacı, bir dizi veri arasındaki önemsiz bağıntıları bulmak ve daha sonra anlamlı ve değerli eğilimleri, örüntüleri ve benzerlerini analiz etmektir. Veri tabanındaki Bilgi Keşfinin (KDD) hayati bir parçasıdır.

Market sepeti analizi problemi, ilişkilendirme kulları veri madenciliğinin uygulama alanının tipik bir örneğidir. Aynı müşteri tarafından satın alınan belirgin ilişkileri içermeyen farklı ürünleri inceleyerek, aralarındaki ilişkiyi analiz ettikten sonra analiz sonuçları emtia tavsiyesi, envanter düzenlemesi, kargo yeri vb. düzenlemeler, kampayalar veya istekler gerçekleştirilebilir.

1993 SIGMOD uluslararası konferansında, birleşme kuralı madenciliği kavramını, Agrawal ve ark. örüntü keşfi ve iki veya daha fazla değişken arasındaki özellik kurallarının keşfini ve açıklamasını kolaylaştırmak için önerdi.

İlişkilendirme kurallarının temelini oluştururken, sıklıkla beraber kullanılan ürün kümeleri, alışveriş arabaları konusundan bu yana yoğun ilgi gördü ve klasik algoritmalar arasındadır (Aggarwal, C.C. and Yu, P.S., 1999).

### 4.6.1 Sık öğeler madenciliği algoritmaları

#### 4.6.1.1 Apriori algoritması

Agrawal tarafından önerilen, bir dizi veriden sıklıkla beraber olan öğe kümlerini bulmak için kullanılır. Algoritma, her verinin destek değerini belirlemek için veri tabanındaki tüm verilerin taranmasını gerektirir; Denklem 3.5 de hesaplanan destek değeri belirlenen eşikten büyük olan verilere sıklık 1- öğekümesi denir. Böylece işlem veri setlerinden sıklık 1-öğekümesi elde edilebilir. Algoritma daha sonra 1-öğe kümesinden gelen 2-öğe kümesi adayını bulur. Benzer şekilde, algoritma, önceki adımda üretilen sık k-1 bileşeninden yeni bir aday k-öğe kümesi elde edebilir. Son olarak, aday öğe kümesinin destek değeri sayılmalıdır.

Bu adımı başarmak için, algoritmanın veri kümesini tekrar tekrar taraması ve destek değeri eşik değerinden daha az olan tüm öğeleri silinmesi ve yeni sık kullanılan öğeler bulunana kadar, sıklık k-öğe kümesi elde edilir. Algoritmanın

uygulanmasından, algoritmanın tekrar tekrar veri kümesini birkaç kez taraması gerektiği görülebilir (Aggarwal, C.C. and Yu, P.S., 1999).

Apriori algoritmasını uygularken dikkate alınması gereken bazı yararlı kavramlar vardır:

- Destek: Öğe kümelerinin önemli bir özelliği de destek sayısıdır (Denklem 4.5), buda belirli öğe kümesi içeren bir hareket sayısını ifade eder. Burada T tüm hareketleri , $t_i$ ilgili öğe kümesini ifade etmektedir.

$$\sigma(X) = |\{t_i \vee X \subseteq t_i, t_i \in T\}| \tag{4.5}$$

  Bir kuralın desteği, X→Y, olarak gösterilir ve denklem 4.6 daki gibi hesaplanır (Samaraweera, Wishma, Chekaprabha, Uma, 2016) (Pang et al. 2013).

$$DestekX \rightarrow Y: \frac{\left(\sigma(X \cup Y)\right)}{N} \tag{4.6}$$

- Güven: Bir kuralın güveni X→Y, T içerisindeki işlemlerin yüzdesi X, ayrıca Y içerir. Bu koşullu olasılıktır P(Y|X). X→Y kuralının güvenirliği formülü 4.7 gibi hesaplanır (Samaraweera, Wishma, Chekaprabha, Uma, 2016) (Pang et al. 2013).

$$G\ddot{u}venX \rightarrow Y: \frac{\left(\sigma(X \cup Y)\right)}{\left(\sigma(X)\right)} = P(Y \vee X) \tag{4.7}$$

- Kaldırma Oranı : Kaldırma, bir modelin performansını ölçen parametredir. X→Y kuralının güvenirliği formülü 4.8 gibi hesaplanır (Pang et al. 2013).

$$Kaldırma(X \rightarrow Y) = \frac{Güven(X \rightarrow Y)}{Destek(Y)} \quad (4.8)$$

#### 4.6.1.2 Fp - growth algoritması

Bu bölüm, sık öğe setlerini keşfetmeye radikal biçimde farklı yaklaşım getiren FP-büyüme adı verilen alternatif bir algoritma sunmaktadır. Algoritma Apriori'nin olştur ve test et paradigmasına bağlı çalışmaz. Bunun yerine bıir FP ağacı adı verilen kompakt bir veri yapısı kullanarak veri kümesini kodlar ve sık sık beraber olan öğe kümelerini doğrudan bu yapıdan ayıklar. Bu yaklaşım detayları (Pang et al. 2013 ) da yer verilmiştir.

# 5 MARKET SEPETİ ANALİZİ VE İŞBİRLİKÇİ FİLTRELEMENİN GELİŞTİRİLMESİ

## 5.1 Market Sepeti Analizi

Pazar sepeti analizi (PSA), müşterilerin gelecekteki satın alma kararlarını öngörmek için bir iş zekası tekniğidir. Alışveriş sepetindeki mevcut ürünlerle birlikte satın almayı tercih edeceğini tahmin etmek için müşterilerin satın alma kalıplarını ve tercihlerini inceler. Örneğin, bir müşterinin 5 katından 3'ünün un ve şekerle birlikte yumurta satın alması halinde (muhtemelen kek pişirmek için), o zaman market sepeti analizi, bu iki ürünle birlikte sunulduğunda yumurta satın alma olasılığını tahmin edebilir.

Market sepeti analizi, çoğunlukla, örneğin ilişkiler şeklinde tarif edilir:

- Un satın alınırsa şeker de satın alınır.
- Şeker satın alınırsa un da satın alınır.
- Hem un hem de şeker satın alınırsa, yumurta% 60 oranında satın alınır.

**Destek:** Destek, analiz altındaki etkinlik lehine olasılıkları göstermektedir. Eğer% 50'den az ise, ilişki daha az verimli kabul edilir.

**Güven:** Kuralın operasyonel verimliliğini ifade eder.
**Kaldırma oranı:** Kaldırma oranı, rastgele seçim işlemlerine kıyasla sonuç bulmada kuralın verimliliğini hesaplar. Genel olarak, birden fazla olan bir kaldırma oranı, kuralın bazı uygulanabilirliğini göstermektedir.

Un, şeker ve yumurta gibi 3 ürünle ilgili bir örnek aldığımızda bu kolaydı ancak bakkal, kişisel hijyen, kıyafet, yiyecek ve içecek, banyo aksesuarları, kırtasiye, elektronik, çanta gibi farklı ürünlerden veri setlerini birleştirirken ne kadar

karmaşık olacağını anlaşılabilir. Walmart'ın resmi web sitesine göre, mağazasında 142.000 farklı ürün var. Bu öğeler çok sayıda olası alt kümeye neden olabilir. Eğer 100 kümelik küçük bir setten 3 set oluşturmaya başlarsak, 161.700 kombinasyon kurmak mümkündür. 142.000 maddeden en iyi kombinasyonları bulmak için analiz edilmesi gereken veri miktarının ne kadar büyük olduğunu düşünün. Ayrıca, bu hesaplamada 2 maddeden 2000 maddeye kadar veri seti olabilir. E-ticarette, bu sorun daha geniş bir ürün yelpazesi nedeniyle daha da büyüyebilir. Export-x'e göre, 2015 itibarıyla Amazon'da 488 milyon ürün bulunuyordu (Grey P. 2015) .

## 5.2 Market Sepeti Analizi Ve İşbirlikçi Filtrelemenin Uygulanması Ve Karşılaştırılması

Bu tez çalışmasında, market sepeti analizi ve işbirlikçi filtrelemeyi hesaplamak için Türk perakende İnşaat Firması veri setini kullandık. FP-Growth ve Apriori algoritması, sık örüntü madenciliği (FPM) için kullanılmış ve işbirliği filtreleme analizi için Kosinüs, Jaccard ve Pearson benzerlik fonksiyonları kullanılmıştır. İlk önce, ürünler arasında FPM kullanarak ilişkilendirme kurallarını çıkarılmıştır ve web sitesinin performans parametrelerine (doğruluk, verimlilik, güvenilirlik) bağlı olarak tavsiye sistemi için en iyi algoritmayı önerdik.

İşbirlikçi Filtreleme için, ilk veri kümesinden üç benzerlik matrisini hesaplanmıştır, daha sonra veriler için gelişmiş parametreler (tahmini puanlar) eklenmiş ve benzerlik matrislerini tekrar hesaplanmıştır. Bu uygulandıktan sonra önerileri almak için benzer öğeleri almak için en yakın komşu algoritmasını kullanılmıştır.

### 5.2.1 Veri toplanması ve düzenlenmesi:

Kullandığımız veri seti, bir Türk inşaat perakende şirketinin 3 yıllık satış verisini (2008'den 2010'a kadar) içermektedir. Veriler ham halde 300 den fazla bir biri ile

ilişkisi olamayan ilişkisel veri tabanı tabloları halinde 9GB boyutunda elde edilmiştir. Veri ön işleme yöntemlerine başlamadan önce gerekli SQL (Structured Query Language) sorguları ile ilgili veri tablolarından excel dosyalarına aktarılmıştır. Aşağıdaki bu tablolardan birisine ait ekran görüntüsü ham görünümünü Şekil 5.1 gösterilmektedir.


Şekil 5.1: Verinin SQL Tablosu

Aşağıda, çıkardığımız verilerin bazı özellikleri bulunmaktadır. Şekil 5.2'sinde verinin ilk biçim vardır. Bu veri tabanı incelendiğinde aşağıdaki özet tanımlayıcı bilgilere ulaşılmıştır:

- Toplam Müşteri Sayısı: 1.826
- Toplam Ürün Sayısı: 1.315
- Toplam Sipariş Sayısı: 42.123
- Toplam İşlem Sayısı: 194.902

| Firma | Donem | Tarih | CH Kodu | CH Unvani | Malzeme Kodu | Malzeme Adi | Miktar | Birim Fiyat | KDV | Net_Satir_Tutari | FISNO |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 208 | 01 | 2008-08-19 | TR.34.13.1003 | ABC | 45.60.00.00001 | XYZ | 25 | 15.0 | 18 | 755.124 | 00000015 |

Şekil 5.2: Verinin SQL biçim

Analiz edilirken, eksik değerler, çift değerler vb. verilerle ilgili bazı sorunlarla karşılaşılmıştır. Bu problemlerin üstesinden gelmek için bazı temel çözümleri uygulanmıştır. İncelendiğinde kayıp değerlerin düşük oranı nedeniyle, eksik değerleri ve yinelenen değerleri araştırma öğe kümesinden çıkarılmasına karar verilmiştir. Bu işlemler gerçekleştirmek için Phyton programlama dilinden içinde bulunan Pandas kütüphanesinden yararlanılmıştır.

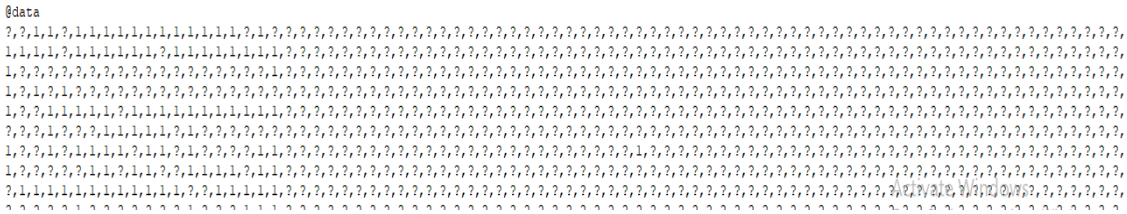
Temizlik ve düzenleme işleminden sonra, öğe-öğe işbirlikçi filtrelemeyi gerçekleştirmek için, veri tabanımızı Pandas veri yapısından, GraphLab adı verilen veri yapılarına dönüştürülmüştür.

Sık kullanılan örüntü madenciliği algoritmaları için, veri kümesi Weka veri madenciliği aracında işlenmek üzere excel dosyası .arff biçimine dönüştürülmüştür.

Şekil 5.3 ve 5.4 de Weka dosyasının biçimleri aşağıda görününmektedir.

```
1   @relation APandFP
2
3   @attribute 10.10.00.13105 {1}
4   @attribute 10.10.00.16105 {1}
5   @attribute 10.11.00.26900 {1}
6   @attribute 10.12.00.00121 {1}
7   @attribute 10.12.00.01442 {1}
8   @attribute 10.13.00.01100 {1}
9   @attribute 10.13.00.18023 {1}
10  @attribute 10.13.00.23023 {1}
11  @attribute 10.15.00.02050 {1}
12  @attribute 10.16.00.01100 {1}
13  @attribute 10.17.00.80580 {1}
14  @attribute 10.18.00.01500 {1}
15  @attribute 10.19.00.01850 {1}
16  @attribute 10.20.00.00350 {1}
17  @attribute 10.21.00.01000 {1}
18  @attribute 10.21.00.01200 {1}
19  @attribute 15.10.00.04440 {1}
```

Şekil 5.3: Verinin Arff dosyası

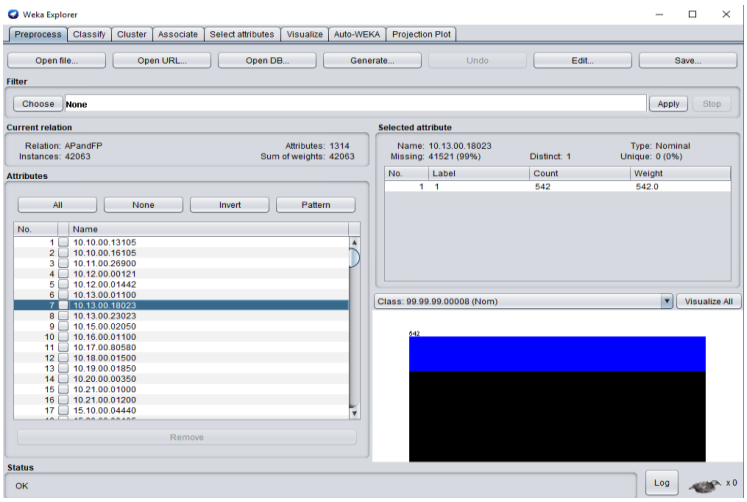



Şekil 5.4: Veri arff dosyası görünümü

## 5.3 Yöntem ve algoritması

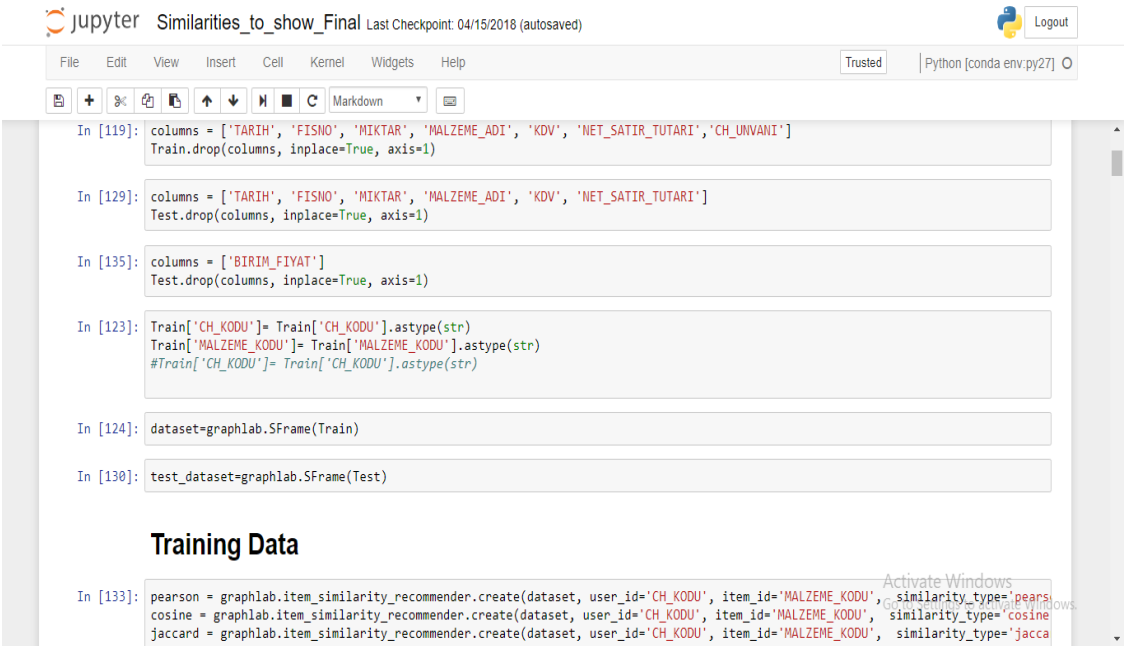
Ürünler arasındaki ilişkiyi bulmak için sırasıyla Apriori Algoritması ve Fp-Growth Algoritması uygulanmıştır. Öğe öğesi işbirlikçi filtreleme için, benzerlik matrisleri oluşturulup Kosinüs, Jaccard ve Pearson benzerlik ölçütleri hesaplanmıştır. Daha sonra en yakın komşu algoritmasını uygulayarak tavsiye edilen ürünleri alınmıştır. Veri setimizi eğitim ve test veri kümelerine ayırmak için Sk-Learn adlı kütüphanesi Çapraz Doğrulama tekniği kullanılmıştır. Eğitim için %70 veri kullanılmış ve test için %30 veri ayrılmıştır. Bu algoritma, önceki bölümlerde kısaca açıklanmıştır.

Analiz için Weka ve Python GraphLab kütüphanesi kullanılmıştır. Weka, “Bilgi Analizi için Waikato ortamı” nı temsil ediyor. Weka veri madenciliği görevleri için bir makine öğrenme algoritmaları topluluğudur. Algoritmalar doğrudan bir veri kümesine uygulanabilir veya kendi Java kodunuzdan çağrılabilir. Weka, veri ön işleme, sınıflandırma, regresyon, kümeleme, ilişkilendirme kuralları ve görselleştirme için araçlar içerir. Ayrıca yeni makine öğrenimi programları geliştirmek için çok uygundur ve veri madenciliği alanında önemli bir rol oynar. Aşağıdaki Şekil 5.5’inde Weka’da veri açtığın zaman nasıl gösterdiğini gösterilmiştir.




Şekil 5.5 : Weka veritabani gösterimi

İşbirlikçi filtreleme'de, Python kullanarak GraphLab kütüphanesi kullanılmıştır. Python uygulaması Jupyter-notebook içinde geliştirilmiştir. Şekil 5.6'de gösterdiği gibi uygulaması için kullanılmıştır.




Şekil 5.6: Jupyter-Notebook uygulaması

## 5.4 Amaç

Bu analizin ana fikri, çeşitli veri madenciliği teknikleri ve algoritmalarını karşılaştırmaktır. Ancak amaç, ürünler arasındaki ilişkiyi bulmak ve belirli bir kullanıcıya göreceli bir ürün önermektir. Ayrıca yeni kullanıcı için ürünü tavsiye etmektedir.

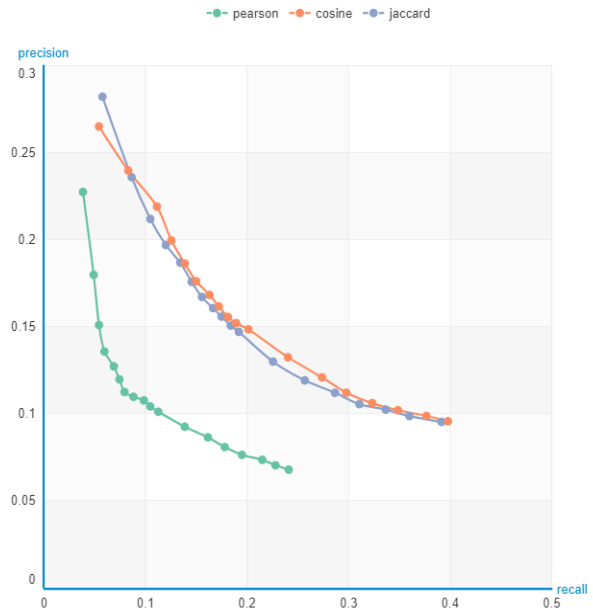
Bu amaca göre tezimizin hipotezi şu şekilde olmaktadır.

“En iyi tavsiye algoritmasını önermek ve ürünler arasında bir ilişki bulunarak bazı popüler veri madenciliği algoritmalarını (FPM ve İF Tekniği) kullanıcılara önermektedir”.

# 6 DEĞERLENDİRME

## 6.1 Bulgular

Verilen veri setine öğe-öğe işbirlikçi filtreleme uygulayarak Şekil 6.1 de gösterildiği gibi aşağıdaki sonuçlar elde edilmiştir.



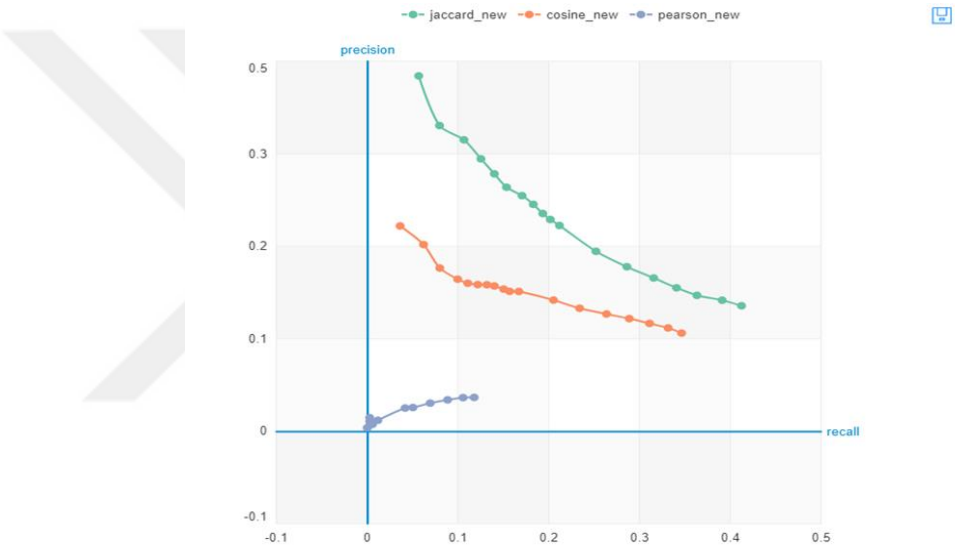

Şekil 6. 1: Öğe-Öğe İF Sonuçları

Şekil 6.1 deki eksenleri temsil eden Keskinlik (Precision) seçilen elemanın alalı olduğunun olasılığını verir ve Geri Çağırma (Recall) seçili alakalı ürünlerin, var olan tüm alakalı ürünlerin toplamına oranını ifade eder.

Kırmızı çizgi Kosinüs, mavi renk ise Jaccard benzerliği endeksini, yeşil çizgiler ise Pearson'u benzerlik ölçütlerini göstermektedir. Bunu veri setindeki ilk teste uygulayarak, 0.28 en yüksek hassasiyete ve 0.39 geri çağırma puanına sahip olmuştur. Bu veri setinden elde edilen sonuçları daha da iyileştirmek için Prediktif derecelendirme eklenmiştir.

Başlangıç değerleri ile test ettikten sonra, Veri tabanımızı, gelişmiş derecelendirme adlı veritabanına başka bir sütun ekleyerek geliştirilmiştir. Bunu başarmak için, her bir ürün için satış sayısını ekleyip ve ardından uygun şekilde puanlanmıştır. Örneğin, en yüksek satın alma ürünü aynı şekilde en yüksek dereceye sahip olmuştur, düşük popüler ürün daha düşük dereceye sahip olmuştur.

Prediktif derecelendirme ekledikten sonra, Şekil 6.2'te göstererek keskinlik skorunu iyileştirdik.




Şekil 6.2: Öğe-öğe İF yeni sonuçlar

Veri kümesindeki yeni geliştirilmiş eklenen parametrelerde Performans değişimini Jaccard benzerliğindeki gibi açıkça görebiliyoruz. Bu puanları aldıktan sonra, önerilen maddeleri almak için KNN uyguladık. Benzerlik skorları en yüksek tavsiye edilen maddeleri almak için rastgele bir madde (45.70.20.00450-Teleskopik Ray en 4,5cm) aşağıda şekil 5.3te gösterildiği gibi kullandık.

| MALZEME_KODU | similar | score | rank |
|---|---|---|---|
| 45.70.20.00450 | 45.70.20.00400 | 0.55153876543 | 1 |
| 45.70.20.00450 | 45.70.20.00350 | 0.530429899693 | 2 |
| 45.70.20.00450 | 45.70.20.00500 | 0.513858377934 | 3 |
| 45.70.20.00450 | 45.70.20.00300 | 0.494557321072 | 4 |
| 45.70.20.00450 | 45.70.20.00550 | 0.39774531126 | 5 |

Şekil 6.3 (a): KNN kullanarak Jaccardin benzerlik sonuçlar

| MALZEME_KODU | similar | score | rank |
|---|---|---|---|
| 45.70.20.00450 | 45.70.20.00400 | 0.37209302187 | 1 |
| 45.70.20.00450 | 45.70.20.00350 | 0.322834670544 | 2 |
| 45.70.20.00450 | 45.70.20.00500 | 0.305555582047 | 3 |
| 45.70.20.00450 | 45.70.20.00300 | 0.2909091115 | 4 |
| 45.70.20.00450 | 45.70.20.00550 | 0.196581184864 | 5 |

Şekil 6.3 (b): KNN kullanarak Kosinüsün benzerlik sonuçlar

| MALZEME_KODU | similar | score | rank |
|---|---|---|---|
| 45.70.20.00450 | 45.70.20.00500 | 0.383015215397 | 1 |
| 45.70.20.00450 | 45.70.20.00400 | 0.343911826611 | 2 |
| 45.70.20.00450 | 45.70.20.00350 | 0.319993913174 | 3 |
| 45.70.20.00450 | 45.70.20.00300 | 0.165979027748 | 4 |
| 45.70.20.00450 | 45.70.20.00550 | 0.156271457672 | 5 |

Şekil 6.3 (c): KNN kullanarak Pearsonin benzerlik sonuçlar

Uygulama ekran çıkıtılarından da anlaşılacağı gibi 4.5cm en e sahip Teloskopik ray için, farklı enlerdeki (4.0cm, 3.5cm, 3.0cm ve 5.5cm) eninde Teloskopik raylar önerildiği anlaşılmaktadır. Bu sonuçta ilginç olan farklı metotların farklı sıralarda ürünleri önermiş olmaları, teleskopik rayların öğe-öğe birbirlerine çok benzedikleri bilinse de en sıklıkla beraber kullanılan enler hakkında bilgi ortaya çıkarılmıştır.

Şekil 6.3 (a) 'da görebileceğimiz gibi, Jaccard bize ürün benzerliği ile en yüksek benzerlik puanını vermektedir.

### 6.1.1 Apriori kuralları

Öğe-öğe işbirlikçi Filtreleme ile bitirdikten sonra, Weka'da sık sık kalıp madenciliği uyguladık. Şekil 6.4 te gösterdiği gibi Apriori algoritması tarafından bulunan en iyi kuralları göstermektedir. Kaldırma(Lift) özelliği, öğenin birlikte satın alınma olasılığını gösterir. Öğe 1., 2. sırasıyla ilk öğenin ve ikincisinin öğe adıdır. Eğer bu iki kalem de alınmışsa, 3. öğenin birlikte alınacağı yönünde % 84,9 oranında bir artış söz konusudur. Weka'da FP-Growth ve Apriori algoritmasını uyguladıktan sonra, aşağıdaki sonuçları bulunmuştur (Çizilge 6.1, Şekil 6.4).

```
Best rules found:

 1. 50.90.61.00200=1 50.90.63.10045=1 422 ==> 50.90.40.35140=1 421    <conf:(1)> lift:(84.95) lev:(0.01) [416] conv:(208.52)
 2. 60.10.00.10098=1 60.20.00.00037=1 447 ==> 60.12.00.09008=1 431    <conf:(0.96)> lift:(45.11) lev:(0.01) [421] conv:(25.73)
 3. 50.90.40.35140=1 50.90.63.10045=1 438 ==> 50.90.61.00200=1 421    <conf:(0.96)> lift:(44.92) lev:(0.01) [411] conv:(23.81)
 4. 60.12.00.09008=1 60.20.00.00125=1 452 ==> 60.20.00.00037=1 434    <conf:(0.96)> lift:(79.5) lev:(0.01) [428] conv:(23.5)
 5. 60.20.00.00037=1 60.20.00.00125=1 456 ==> 60.12.00.09008=1 434    <conf:(0.95)> lift:(44.53) lev:(0.01) [424] conv:(19.4)
 6. 60.10.00.10098=1 60.20.00.00125=1 450 ==> 60.12.00.09008=1 428    <conf:(0.95)> lift:(44.5) lev:(0.01) [418] conv:(19.15)
 7. 60.12.00.09008=1 60.20.00.00125=1 452 ==> 60.10.00.10098=1 428    <conf:(0.95)> lift:(74.03) lev:(0.01) [422] conv:(17.85)
 8. 50.90.63.10045=1 463 ==> 50.90.40.35140=1 438    <conf:(0.95)> lift:(80.55) lev:(0.01) [432] conv:(17.6)
 9. 60.12.00.09008=1 60.20.00.00037=1 464 ==> 60.20.00.00125=1 434    <conf:(0.94)> lift:(73.54) lev:(0.01) [428] conv:(14.78)
10. 90.40.10.00120=1 90.40.10.00180=1 460 ==> 90.40.10.00150=1 429    <conf:(0.93)> lift:(49.34) lev:(0.01) [420] conv:(14.1)
```

Şekil 6.4: Apriori Birliktelik ilk on kuralı

Aşağıda gösterildiği gibi, bu kodlar her benzersiz ürünü temsil eder.

Çizilge 6.1 Apriori Algorithması ilk 5 kural {Öğe1, Öğe2}→Öğe3

| Öğe 1 | Öğe 2 | Öğe 3 | Kaldırma % | Güven % |
|---|---|---|---|---|
| Çekiç ( Ağaç Sapli ) 200 g | Çekiç Yedeği (Plastik) 45 mm | Yan Keski 140-5,5"(Polish) | 84.9 | 100 |
| Panç Bi Metal 98 mm | Matkap Ucu ( Hss ) 3,7 mm | Panç Adaptörü 32 - 152 mm (Sds Plus) 9008 | 79.5 | 0.96 |
| Yan Keski 140-5,5"(Polish) | Çekiç Yedeği (Plastik) 45 mm | Çekiç ( Ağaç Saplı ) 200 g | 74.03 | 0.95 |
| Panç Bİ Metal 98 mm | Matkap Ucu ( Hss ) 3,9 mm | Panç Adaptörü 32 - 155 mm (Sds Plus) 9005 | 73.54 | 0.94 |
| Yan Keski 140-4,5"(Polish) | Çekiç Yedeği (Plastik) 45 mm | Çekiç ( Ağaç Saplı) 200 g | 49.34 | 0.93 |

Çizelge 6.1 de verilen ilişkilendirme kuralları şu şekilde yorumlanabilir.

Çekiç (Ağaç Saplı) 200g ve Çekiç yedeği 45 mm beraber alanların, %100 güven ile Yan Keski 140 5,5” de aldıkları anlaşılmıştır. Sepetinden Çekiç (Ağaç Saplı) 200g ve Çekiç yedeği 45 mm beraber olması, Yan Keski 140 5,5” de alınma olasılığını %84,9 arttırmaktadır.

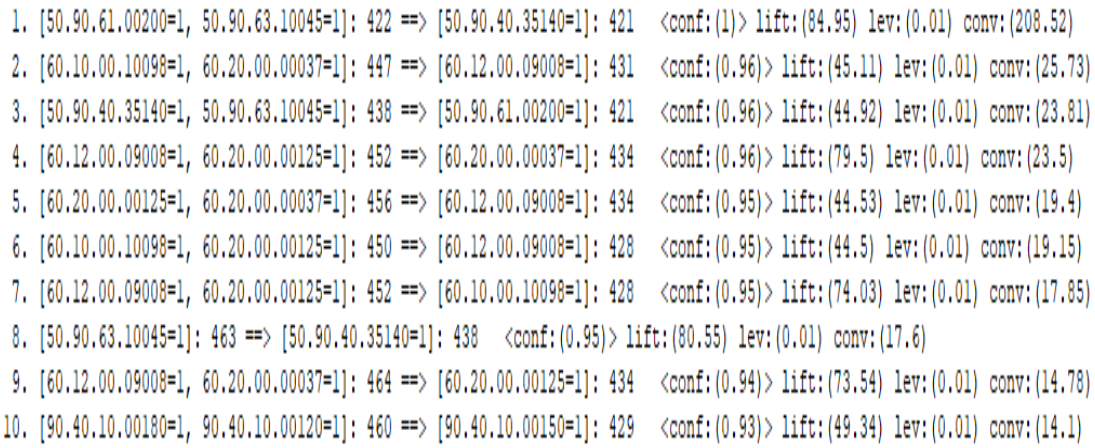
Panç Bi Metal 98 mm ve Matkap Ucu ( Hss ) 3,7mm beraber alanların, %96 güven ile Panç Adaptörü 32-152mm (Sds Plus) 9008 de aldıkları anlaşılmıştır. Sepetinden Panç Bi Metal 98mm ve Matkap Ucu(Hss) 3,7mm beraber olması, Panç Adaptörü 32-152mm (Sds Plus) 9008 de alınma olasılığını %79,5 arttırmaktadır.

Yan Keski 140-5,5"(Polish) Ugr ve Çekiç Yedeği (Plastik) 45mm beraber alanların, %95 güven ile Çekiç ( Ağaç Saplı ) 200g de aldıkları anlaşılmıştır. Sepetinden Yan Keski 140-5,5"(Polish) Ugr ve Çekiç Yedeği (Plastik) 45mm beraber olması, Çekiç ( Ağaç Saplı )200g de alınma olasılığını %74,03 arttırmaktadır.

Panç Bİ Metal 98 mm ve Matkap Ucu ( Hss ) 3,9mm beraber alanların, %94 güven ile Panç Adaptörü 32-155 Mm (Sds Plus) 9005 de aldıkları anlaşılmıştır. Sepetinden Panç Bİ Metal 98 mm ve Matkap Ucu ( Hss ) 3,9mm beraber olması, Panç Adaptörü 32 - 155 Mm (Sds Plus) 9005 de alınma olasılığını %73,54 arttırmaktadır.

Yan Keski 140-4,5"(Polish) Ugr Çekiç Yedeği (Plastik) 45 mm beraber alanların, %93 güven ile Çekiç ( Ağaç Sapli ) 200 Gr de aldıkları anlaşılmıştır. Sepetinden Yan Keskü 140-4,5"(Polish) Ugr Çekiç Yedeği (Plastik) 45 mm beraber olması, Çekiç ( Ağaç Saplı ) 200g de alınma olasılığını %49,34 arttırmaktadır.

### 6.1.2 Fp-growth kuralları




Şekil 6.5 : FP- Growth Birliktelik kuralları Sonuçları

Kurallar, Çizilge 6.1 ve Çizilge 6.2 de gösterildiği gibi en büyükten en küçük değere Lift değerlerine bağlı olarak sınıflandırılır.

Çizilge 6.2 FP-Growth Algoritması ilk 5 kural

| Öğe 1 | Öğe 2 | Öğe 3 | Kaldırma % | Güven % |
|---|---|---|---|---|
| Çekiç ( Ağaç Saplı ) 200 Gr | Çekiç Yedeği (Plastik) 45 Mm | Yan Keski 140-5,5"(Polısh) g | 84.95 | 100 |
| Panç Bi Metal 98mm | Matkap Ucu ( Hss ) 3,7mm | Panç Adaptörü 32 - 152mm (Sds Plus) 9008 | 79.5 | 0.96 |
| Yan Keski 140-5,5"(Polısh) Ugr | Çekiç Yedeği (Plastik) 45 mm | Çekiç ( Ağaç Saplı ) 200 g | 74.5 | 0.95 |
| Panç Bi Metal 98 Mm | Matkap Ucu ( Hss ) 3,9mm | Panç Adaptörü 32 - 155mm (Sds Plus) 9005 | 73.54 | 0.94 |
| Yan Keski 140-4,5"(Polısh) g | Çekiç Yedeği (Plastik) 45mm | Çekiç ( Ağaç Saplı ) 200 g | 49.34 | 0.93 |

Çizelge 6.2 de verilen ilişkilendirme kuralları şu şekilde yorumlanabilir.

Çekiç (Ağaç Saplı) 200g ve Çekiç yedeği 45 mm beraber alanların, %100 güven ile Yan Keski 140 5,5” de aldıkları anlaşılmıştır. Sepetinden Çekiç (Ağaç Saplı) 200g ve Çekiç yedeği 45 mm beraber olması, Yan Keski 140 5,5” de alınma olasılığını %84,9 arttırmaktadır.

Panç Bi Metal 98 mm ve Matkap Ucu ( Hss ) 3,7 Mm beraber alanların, %96 güven ile Panç Adaptörü 32 - 152 mm (Sds Plus) 9008 de aldıkları anlaşılmıştır. Sepetinden Panç Bi Metal 98 mm ve Matkap Ucu ( Hss ) 3,7 Mm beraber olması, Panç Adaptörü 32 - 152 mm (Sds Plus) 9008 de alınma olasılığını %79,5 arttırmaktadır.

Yan Keski 140-5,5"(Polish)g ve Çekiç Yedeği (Plastik) 45 mm beraber alanların, %95 güven ile Çekiç ( Ağaç Saplı ) 200g de aldıkları anlaşılmıştır. Sepetinden Yan Keski 140-5,5"(Polish) Ugr ve Çekiç Yedeği (Plastik) 45 mm beraber olması, Çekiç ( Ağaç Saplı )200g de alınma olasılığını %74,03 arttırmaktadır.

Panç Bİ Metal 98 mm ve Matkap Ucu ( Hss ) 3,9 mm beraber alanların, %94 güven ile Panç Adaptörü 32 - 155mm (Sds Plus) 9005 de aldıkları anlaşılmıştır. Sepetinden Panç Bİ Metal 98 mm ve Matkap Ucu ( Hss ) 3,9 mm beraber olması, Panç Adaptörü 32 - 155mm (Sds Plus) 9005 de alınma olasılığını %73,54 arttırmaktadır.

Yan Keski 140-4,5"(Polish) Ugr Çekiç Yedeği (Plastik) 45 mm beraber alanların, %93 güven ile Çekiç ( Ağaç Saplı ) 200 Gr de aldıkları anlaşılmıştır. Sepetinden Yan Keski 140-4,5"(Polish)g Çekiç Yedeği (Plastik) 45 mm beraber olması, Çekiç ( Ağaç Sapli ) 200 Gr de alınma olasılığını %49,34 arttırmaktadır.

FP-Growth'un kurduğu kurallar Apriori'ye benzerlik göstermektedir. Lift değeri ve güven, aşağıdaki bölümdeki sonuçlarla hemen hemen aynıdır.
Her iki algoritmayı da analiz ettikten sonra, istenen veri setimize bağlı olarak sonuçlar neredeyse aynı olduğu gözlemlenmiştir. Ancak parametre ve veri setinin boyutunu değiştirdiğimizde önemli bir performans değişikliği farkedilmiştir.
Bu çalışmalar aşağıdaki makine konfigürasyonunda çalıştırıldığında çalışma süreleri ile ilgili bilgi aşağıdaki çizelgede verilmiştir.

Çizilge 6.3 Bilgisayar özellikleri

| İşlemci | İntel (R) core i3-2310M CPU 2.10GHz |
|---|---|
| Ram | 12 GB |
| Sabit Disk Kapasitesi | 500 GB |
| Ekran Kart | 128 MB |
| işletim sistemi | Windows 10 – 64 bit |

Bu algoritmaların her ikisinden de sonuç elde etmede büyük bir fark vardı. Performansın karmaşıklığı ve veri kümesinin boyutuna ve Algoritmaların çalışma farklarına bağlı olduğu söylenebilir.

# 7 SONUÇ

Bu çalışmada, Türk Perakende Şirketinin elektrikli ürünler satan veri setini kullanılmıştır. Bu veri kümesinde, 1023 adet ürün arasında yarım milyon adet Türk Özel İnşaat Perakende Satış bilgileri kullanılmıştır. İlk başta, verileri SQL veritabanından excel sayfalarına yönetilebilir boyutlara getirilmiştir. Gereksiz veri tablolarını kaldırmış ve bu veriler üzerinde algoritmaların yapılabileceği biçimde verileri temizlenip ayarlanmıştır.
İşbirlikçi filtreleme için, verileri Phyton Panda veri yapılarına aktarılarak bunlara ortak filtreleme uyguladık.

KNN kullanarak en yüksek benzerlik puanlarına sahip önerilen öğeleri almak için rastgele bir öğe (45.70.20.00450 Teleskopik Ray en 4.5cm) kullandık. Değerlendirmeden sonra, %55 benzerlik ile en yüksek benzerlik skorunu Jaccard benzerlik endeksi ile aldık.

| MALZEME_KODU | similar | score | rank |
|---|---|---|---|
| 45.70.20.00450 | 45.70.20.00400 | 0.55153876543 | 1 |
| 45.70.20.00450 | 45.70.20.00350 | 0.530429899693 | 2 |
| 45.70.20.00450 | 45.70.20.00500 | 0.513858377934 | 3 |
| 45.70.20.00450 | 45.70.20.00300 | 0.494557321072 | 4 |
| 45.70.20.00450 | 45.70.20.00550 | 0.39774531126 | 5 |

Şekil 7.1: KNN kullanarak Jaccardin benzerlik sonuçlar

Şekil 7.1 bize, sistemin rastgele ürünümüzü satın alan herkese 45.70.20.00400 (Teleskopik Ray en 4.0cm) ürünü önereceğini göstermektedir (45.70.20.00450 Teleskopik Ray en 4.5cm). Bu ürünü de alabilecekleri % 55 keskinlik vardır.

Daha sonra, üzerlerinde Apriori ve FP-Growth algoritmalarını uygulamak için veriler WEKA formatına dönüştürülür. Bundan sonra, lift parametrelerini

kullanarak verileri analiz ettik. Apriori algoritması kullanılarak, kaldırma parametreleri kullanılarak elde edilen sonuçlar aşağıdadır.

- ÇEKİÇ (AĞAÇ SAPLI) 200 GR ve ÇEKİÇ YEDEĞİ (PLASTİK) 45 MM alan 422 kişiden 421 kişi YAN KESKİ 140-5, 5"(POLISH) UGR ta almışlar.

- PANÇ Bİ METAL 98 MM ve MATKAP UCU ( HSS ) 3,7 MM alan 452 kişiden 434 kişi PANÇ ADAPTÖRÜ 32 - 152 MM (SDS PLUS) 9008 ta almışlar.

- YAN KESKİ 140-5,5"(POLISH) UGR ve ÇEKİÇ YEDEĞİ (PLASTİK) 45 MM alan 464 kişiden 434 kişi ÇEKİÇ ( AĞAÇ SAPLI ) 200 GR da almışlar.

- PANÇ Bİ METAL 98 MM ve MATKAP UCU ( HSS ) 3,9 MM alan 452 kişiden 428 kişi PANÇ ADAPTÖRÜ 32 - 155 MM (SDS PLUS) 9005 da almışlar.

- YAN KESKİ 140-4,5"(POLISH) UGR ve ÇEKİÇ YEDEĞİ (PLASTİK) 45 MM alan 460 kişiden 429 kişi ÇEKİÇ ( AĞAÇ SAPLI ) 200 GR da almişlar.

  FPGrowth algorithması ile lift değerine göre bulunan ilk 5 kural şu şekildedir.

- ÇEKİÇ (AĞAÇ SAPLI) 200 GR ve ÇEKİÇ YEDEĞİ (PLASTİK) 45 MM alan 422 kişiden 421 kişi YAN KESKİ 140-5, 5"(POLISH) UGR ta almışlar.

- PANÇ Bİ METAL 98 MM ve MATKAP UCU ( HSS ) 3,7 MM alan 447 kişiden 431 kişi PANÇ ADAPTÖRÜ 32 - 152 MM (SDS PLUS) 9008 ta almışlar.

- YAN KESKİ 140-5,5"(POLISH) UGR ve ÇEKİÇ YEDEĞİ (PLASTİK) 45 MM alan 438 kişiden 421 kişi ÇEKİÇ ( AĞAÇ SAPLI ) 200 GR da almışlar.

- PANÇ Bİ METAL 98 MM ve MATKAP UCU ( HSS ) 3,9 MM alan 452 kişiden 428 kişi PANÇ ADAPTÖRÜ 32 - 155 MM (SDS PLUS) 9005 da almışlar.

- YAN KESKİ 140-4,5"(POLISH) UGR ve ÇEKİÇ YEDEĞİ (PLASTİK) 45 MM alan 460 kişiden 429 kişi ÇEKİÇ ( AĞAÇ SAPLI ) 200 GR da almışlar.

Eğitim için %70 veri ve test için %30 veri kullandık ve bu kuralları aldıktan sonra, kaldırma oranı ve güven parametrelerine bağlı olarak müşteriye en yakın ürünü önerecektir.

Bütün bu öneri algoritmalarını karşılaştırdıktan sonra, Sonuçlar aşağıdaki gibiydi özetlenebilir. Öğe-öğe işbirlikçi filtrelemede, yeni eklenen parametrelerle Jaccard benzerliği, bu veri seti için %0,2 performans arttırarak ile bu analiz arasında en iyisi olduğu belirlendi.

Sık Kullanılan Örüntü Madenciliğinde, her iki algoritma arasında FPGrowth en iyi sonuçları verdiği tespit edildi. Her ne kadar sonuçlar benzer olmasına rağmen her iki algoritmada da çok fazla performans farkı vardı. FpGrowth algoritması işlemede çok hızlıydı, daha az bellek, verimli performans sağladı ve veritabanını sadece iki kez taraması yapmıştır. Hızlı öneri sistemleri için en iyi seçimdir. Bu çalışmadan elde edilen sonuçların gölgesinde gelecekteki çalışmalara gelince,

bu sonuçları öneri sistemlerinin karşılaştırılmasında melez işbirlikçi Filtreleme tekniği ile karşılaştırmak daha iyi olacaktır.

## KAYNAKLAR


A. Ansari, S. Essegaier, and R. Kohli, “Internet recommendation systems,” Journal of Marketing Research, vol. 37, no. 3, pp. 363–375, 2000.

Aggarwal, C.C. and Yu, P.S., 1999, Methodologies for Knowledge Discovery and Data Mining. Data Mining Techniques for Associations, Clustering and Classification. PAKDD '99 Proceedings of the Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining

Ammar Jabakji, Hasan Dag, 2016. Improving item-based recommendation accuracy with user’s preferences on Apache Mahout. Big Data (Big Data), 2016 IEEE International Conrerence

B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl, “Item-based collaborative filtering recommendation algorithms,” in Proceedings of the 10th International Conference on World Wide Web (WWW '01), pp. 285–295, May 2001.

B. N. Miller, J. A. Konstan, and J. Riedl, “PocketLens: toward a personal recommender system,” ACM Transactions on Information Systems, vol. 22, no. 3, pp. 437–476, 2004.

Bharati, M & , Ramageri. (2010). DATA MINING TECHNIQUES AND APPLICATIONS. Indian Journal of Computer Science and Engineering. 1.

Bhagya,Ramesh, & Reeba, R. (2017). Secure recommendation system for E-commerce website. 2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT), 1-5.

Di Noia,R. Mirizzi,V.C.Ostuni,D.Romito,M.Zanker “Linked open data to support content-based recommender systems” Proceedings of the 8th international conference on semantic systems(2012), pp.1-8

Debnath,N.Ganguly,P.MitraFeature weighting in content based recommendation system using social network analysis Proceedings of the 17th international conference on World Wide Web(2008), pp.1041-1042

D. Billsus and M. Pazzani, “Learning collaborative information filters,” in Proceedings of the 15th International Conference on Machine Learning (ICML '98), 1998.

DATA MINING TECHNIQUES AND APPLICATIONS-Scientific Figure on ResearchGate.Availablefrom:https://www.researchgate.net/figure/Knowledge-discovery Process_fig1_49616224 [accessed 7 Jan, 2019]

Faryal, A. Tauqir, A. M. Martinez-Enriquez and M. Aslam, "Data Mining Based Recommendation System Using Social Websites," 2015 IEEE / WIC / ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Singapore, Singapore, 2015, pp. 365-368. doi:10.1109/WI-IAT.2015.78

G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions,” IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734–749, 2005

G. Linden, B. Smith, and J. York, “Amazon.com recommendations: item-to-item collaborative filtering,” IEEE Internet Computing, vol. 7, no. 1, pp. 76–80, 2003.

G. Shani, D. Heckerman, and R. I. Brafman, “An MDP-based recommender system,” Journal of Machine Learning Research, vol. 6, pp. 1265–1295, 2005.

Han J., Kamber M., Pei J., 2011,Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems) 3 edition

Hyunwoo Hwangbo, Yang Sok Kim, Kyung Jin Cha, “Recommendation system development for fashion retail e-commerce” in Electronic Commerce Research and Applications,Volume 28, 2018,Pages 94-101,ISSN 1567 4223, https://doi.org/10.1016/j.elerap.2018.01.012.

J. Breese, D. Heckerman, and C. Kadie, “Empirical analysis of predictive algorithms for collaborative filtering,” in Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI '98), 1998.

John O'Donovan, Barry Smyth 2005. Trust in recommender systems, IUI '05 Proceedings of the 10th international conference on intelligent user interfaces

K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, “Eigentaste: a constant time collaborative filtering algorithm,” Information Retrieval, vol. 4, no. 2, pp. 133–151, 2001.

K. Miyahara and M. J. Pazzani, “Improvement of collaborative filtering with the simple Bayesian classifier,” Information Processing Society of Japan, vol. 43, no. 11, 2002

K. Yu, A. Schwaighofer, V. Tresp, X. Xu, and H.-P. Kriegel, “Probabilistic memory-based collaborative filtering,” IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 1, pp. 56–69, 2004.

Kumar, Rukmani, S, K 2010. Implementation of web usage mining using Apriori and FP Growth Algorithms. Int. J. of Advanced Networking and Applications, 6,400-404 (Kumar, Rukmani, S, K 2010)

Kwei Tang, Yen-Liang Chen, ve Hsiao-Wei Hu. 2008 Data Mining Based Recommendation System using Social Websites, IEEE Transactions on Knowledge and Data Engineering

Larose, D. T., 2005, Discovering knowledge in data: An introduction to data mining

L. H. Ungar and D. P. Foster, "Clustering methods for collaborative filtering," in Proceedings of the Workshopon Recommendation Systems, AAAI Press, 1998.

M. Claypool, A. Gokhale, T. Miranda et al., "Combining content-based and collaborative filters in an online newspaper," in Proceedings of the SIGIR Workshop on Recommender Systems: Algorithms and Evaluation, Berkeley, Calif, USA, 1999.

Mayuri Dalvi, Prof S.V Gumaste, 2015 Review Paper on Colloborative Filtering, International Research Journal of Engineering and Technology (IRJET) (Mayuri Dalvi, Prof S.V Gumaste 2015)

Michael P. O'Mahony, Barry Smyth 2007 A recommender system for on-line course enrolment: An initial study

Pang-Ning Tan, Michael Steinbach, Vipin Kumar, 2013, Introduction to Data Mining-Pearson

Paul Gray 2015, How Many Products Does Amazon Sell?, https://export-x.com/2015/12/11/how-many-products-does-amazon-sell-2015/ en son erişim 13 Mart 2019

Paul Resnick , Rahul Sami 2007. The Influence Limiter: Provably Manipulation-Resistant Recommender System

Punam Bedi Harmeet Kaur Sudeep Marwaha 2007, Trust Based Recommender System for Semantic Web.

Roşca, Radoıu, D, 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

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," Journal of the American Society for Information Science, vol. 41, no. 6, pp. 391–407, 1990.

S. H. S. Chee, J. Han, and K. Wang, “RecTree: an efficient collaborative filtering method,” in Proceedings of the 3rd International Conference on Data Warehousing and Knowledge Discovery, pp. 141–151, 2001.

Seven kosub, A note on the triangle inequality for the Jaccard distance, Department of Computer & Information Science, arXiv: 1612.02696v1 [cs: DM] 8 Dec 2016, University of Konstanz Box 67, D-78457 Konstanz, Germany

S. K. Jones, “A statistical interpretation of term specificity and its applications in retrieval,” Journal of Documentation, vol. 28, no. 1, pp. 11–21, 1972.

Samaraweera, Wishma & Waduge, Chekaprabha & Meththananda, Uma. (2016). Market Basket Analysis: A Profit Based Approach to Apriori Algorithm.

Sarabjot S. Anand, John G.Hughes, 1998, Hybrid Data Mining systems: The Next Generations

Tomasevic, Nikola & Paunović, Dejan & Vraneš, Sanja. (2019). User-based collaborative filtering approach for content recommendation in OpenCourseWare platforms.

T. Hofmann, “Latent semantic models for collaborative filtering,” ACM Transactions on Information Systems, vol. 22, no. 1, pp. 89–115, 2004.

T. Landauer, M. Littman, and Bell Communications Research (Bellcore), “Computerized cross-language document retrieval using latent semantic indexing,” US patent no. 5301109, April 1994.

Weikang Xue, Bopin Xiao, Lin Mu. 2015 Intelligent Mining on Purchase Information and Recommendattion System for E-Commerce, Industrial Engineering and Engineering Management (IEEM), 2015 IEEE International Conference

X. Su and T. M. Khoshgoftaar, “Collaborative filtering for multi-class data using belief nets algorithms,” in Proceedings of the International Conference on Tools with Artificial Intelligence (ICTAI '06), pp. 497–504, 2006.

Xiaofeng Yuan,Lixin Han,Subin Qian,Guoxia Xu,Hong Yan, “Singular value decomposition based recommendation using imputed data” in Knowledge-Based Systems, Elsevier jan,2019.

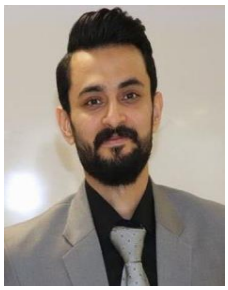
Yuri & Lobur, Mykhoylo & Artsibasov, Vitalij & Chystyak, Vitalij. (2015). Methods and tools for building recommender systems. 300-305. 10.1109/CADSM.2015.7230862.

Z. A. Usmani, S. Manchekar, T. Malim and A. Mir, "A predictive approach for improving the sales of products in e-commerce," 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), Chennai, 2017, pp. 188-192.
doi: 10.1109/AEEICB.2017.7972409

Zhongyi Hu, Liangzhong Shen, Shengkai Chen, An Improved Apriori-Based Personal Recommendation Algorithm for E-commerce, Pervasive Computing and Applications, 2008. ICPCA 2008. Third International Conference

Zhu, X., Su, S., Fu, M., Liu, J., Zhu, L., Yang, W., Jing, G., … Guo, Y. (2018). A Cosine Similarity Algorithm Method for Fast and Accurate Monitoring of Dynamic Droplet Generation Processes. Scientific reports, 8(1), 9967. doi:10.1038/s41598-018-28270-8

## ÖZGEÇMİŞ



Adı Soyadı : WALEED ABDULLAH

Doğum Yeri ve Yılı : PAKISTAN, 25/05/1992

Medeni Hali : (Bekar)

Yabancı Dili : İngilizce, urduca

E-posta : waleedabdullah928@gmail.com

**Eğitim Durumu**

| | | |
|---|---|---|
| Lise | : | Islamabad College For Boys G-6/3, ISB, Pakistan |
| Lisans | : | NUML Islamabad,Mühendislik Fakültesi, Bilgisayar Bilimi |
| Yüksek Lisans | : | İstanbul Ticaret Üniversitesi,<br>Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı |

**Mesleki Deneyim**

| | | |
|---|---|---|
| Software Developer Intern | : | NAVTTC, Pakistan (jan,2015 - jun,2015) |
| Android Developer | : | CODE Islamabad (2013-2014) |