OKAN ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI BAŞKANLIĞINA

ERYTHROMCIN İLACININ YAN ETKİLERİNİN ARAŞTIRILMASI ÜZERİNE VERİ MADENCİLİĞİ ÇALIŞMASI

YÜKSEK LİSANS TEZİ

ERHAN TAHMİNCİLER

tarafından

YÜKSEK LİSANS

derecesi şartını sağlamak için hazırlanmıştır.

OCAK 2014

ERYTHROMCIN İLACININ YAN ETKİLERİNİN ARAŞTIRILMASI

ÜZERİNE VERİ MADENCİLİĞİ ÇALIŞMASI

YÜKSEK LİSANS TEZİ

ERHAN TAHMİNCİLER

tarafından

OKAN ÜNİVERSİTESİ

Bilgisayar Mühendisliği Anabilim Dalı Başkanlığı'na

Yüksek Lisans

derecesi şartını sağlamak için sunulmuştur.

Onaylayan:

_____

Chair
Name
Danışman
Yrd. Doç. Dr. Pınar Yıldırım

_____

Member's
Name
Üye
Prof. Dr. B. Tevfik Akgün

_____

Member's
Name
Üye
Prof. Dr. A. Coşkun Sönmez

DATA MINING ON THE RESEARCH OF THE SIDE EFFECTS OF ERYTHROMYCIN

ABSTRACT

In this study, a research study has been performed by using the patient comments about Erythromycin on the website www.askapatient.com by utilizing data and text mining methods. The related comments have been analysed by some methods including text parsing and patients' gender and age, daily dosage information and duration of use have been systematized and the side effects of this drug on patients have been endeavoured to be determined.

As a result of the study, the side effects of the patient comments have been separated by text parsing methods and the frequencies of them have been calculated and the frequently seen side effects such as stomach ache, vertigo, vomiting and diearrhea have been determined. After analysing the fields such as gender, age, daily dosage, duration of use, the association rules (Apriori) algorithm has been used and hidden relationships among these attributes have been investigated

The results of this study will provide significant contributions to medical experts, researchers and pharmaceutical companies.

Keywords : Biomedical text mining; Erythromycin; Side effects; Apriori algorithm; Weka; Rapid Miner; N-Gram algorithm

TEŞEKKÜR

Tez çalışmalarım süresince, deneyimi ve bilgi birikimi ile yol haritamın oluşmasını sağlayan, fikirleri ve yol göstericiliği ile çalışmamın gelişmesine imkan tanıyan tez danışmanım Sayın Yrd. Doç. Dr. Pınar Yıldırım'a, yüksek lisansın başlangıcından itibaren, iş yerinde, gerek anlayışlı olmaları gerekse duyarlılıklarından ötürü çalışmamın büyük bir kısmındaki müdürüm Sayın Faik Sarı ve ikinci iş yerimdeki müdürüm Sayın Mehmet Kemal Erdog'a ve tüm çalışma arkadaşlarıma, hayatımın her alanında beni doğru yönde etkileyen ve maddi, manevi desteklerini hiçbir zaman esirgemeyen sevgili anneme, babama ve ağabeyim Erçin Tahminciler'e özel teşekkürlerimi sunuyorum.

İÇİNDEKİLER

TABLO LİSTESİ

ŞEKİL LİSTESİ

I. GİRİŞ

İnternet'in 1990'lı yıllardan sonra popüler hale gelmesi ile bilgiye ulaşma ve bilgiyi depolama gücü çok büyük oranda artmıştır.Hemen her sektörde artan bilgisayar kullanımı ve gelişen teknoloji ile birlikte elektronik ortamlara hızlı bir şekilde yüksek miktarlarda veriler depolanmaya başlanmıştır.

Diğer yandan bilgiye sahip olmanın ve onu verimli kullanmanın giderek önem kazandığı günümüzde güçler dengesi bilgi üzerinde yoğunlaşmaktadır.Bilginin bu derece önem kazanması gün geçtikçe çoğalan veri yığınları içerisinden hızlı bir biçimde anlamlı ve faydalı bilgilere ulaşılması çeşitli sektörlerde faliyet gösteren kurum ve kuruluşlar için yeni fırsatların ortaya çıkmasına olanak sağlamaktadır.

Bilim adamları ve araştırmacılar tarafından yapısal olmayan yüksek veri yığınlarından yararlanmak için herhangibi bir araç kullanmadan verileri analize tabi tutarak karar verme aşamasında kullanılmasının imkansız olduğu görülmüş ve bu aşamada karar verme bileşeni olarak ilişkisel veri tabanları ve veri ambarları yapıları ortaya çıkmıştır.

Veri Madenciliği, yüksek veri yığınlarına çeşitli analiz yöntemleri ve kuralların uygulanması ile saklı kalmış bilgilerin keşfedilmesi ve keşfedilen anlamlı bilgilerin karar verme süreci içerisinde kullanılmasına kadar olan bütün süreci kapsamaktadır.

Günümüzde veri madenciliği, pazarlama, bankacılık, sigortacılık, elektronik ticaret, sosyal medya, tıp ve sağlık hizmetleri gibi giderek artan bir çok sektörde aktif olarak kullanılmaktadır.

Tıp ve sağlık hizmetlerinde, her geçen yıl içinde içinde fizyolojik ve biyometrik hasta belirtilerinin tedavisi için ilaç kullanımında sürekli bir artış yaşanmaktadır.2009 yılında IMS Türkiye ve Türk Eczacıları Birliği'nde yayınlanan makalede Türkiye'de 2005 yılında 1 milyar 212 milyon, 2006 yılında 1 milyar 292 milyon, 2007 yılında 1 milyar 398 milyon, 2008 yılında, 1 milyar 476 milyon, 2009 yılı Ocak-Ekim ayları arasında 1 milyar 239 milyon kutu ilaç satışı olmuştur [1]. Bu rakamlar Türkiye ilaç pazarının 2005-2009 yılları arası değişimini göstermektedir.

Tedavilerde kullanılacak tıbbi ilaçların satış sürecinden önce klinik çalışmalarda güvenilirlik ve etkinlik değerlendirmeleri yapılmaktadır.Bu çalışmalar genellikle ilaç şirketleri tarafından yürütülür ve kısa sürede kesin sonuçlar almak için çok az insan içerir. İnternet kullanımının yaygınlaşması ile ilaç kullanan kişiler çesitli web ve forum siteleri aracılığı ile yaşadığı süreci ve görülen yan etkileri paylaşmaya başladılar.Bu bilgiler ilaç üretim şirketleri, doktorlar ve diğer hastalar için önemli bilgi kaynağı olmaktadır.

Bu tür siteler ve forumlarda yazılan yorumların bazılarında klinik dil ve terminoloji kullanılması bazılarında ise günlük konuşma dilini içermesi ve paylaşılan verilerde

---

1. Türk Eczacıları Birliği "İlaç Kullanımları Hakında Bilgiler," *http://www.e-kutuphane.teb.org.tr/pdf/raporlar/sag_ilac09/8.pdf,* 27 Aralık 2013.

gürültünün olması kullanılan dil farklılığını gidermek ve daha anlamlı bilgilere ulaşmak için veri madenciliği çalışmalarının yapılmasını zorunlu kılmaktadır.

Yapılan çalışmada erythromycin ilacının seçilme nedeni ise Helikobakter Pilori ve Mycobacterium Tuberculosis gibi dünyada yaygın olarak görülen bakterilerin tedavisinde etkin olarak kullanılmasından kaynaklanmaktadır. Helikobakter Pilori, mide ve oniki parmak barsağı ülserleri ile kronik gastritin en önemli etkeni olarak kabul edilmekte öte yandan Mycobacterium Tuberculosis ise günümüzde halen morbidite ve mortalitesi yüksek bir hastalık olarak seyreden tüberkiloz (verem) hastalığına yol açmaktadır.

Bu çalışmada, Erythromycin ilacının yorumları İnternet sitesinden çeşitli araçlar ile okunarak bir veritabanı oluşturulmuştur.Oluşturulan yapıda kural tabanlı metin ayrıştırma, eşleştirme ve çeşitli veri madenciliği yöntemleri ve araçları ile analiz edilmiş ve sonuçlar bir tıp uzmanına da danışılarak yorumlanmış ve paylaşılmıştır.

II. VERİ MADENCİLİĞİ VE ALGORİTMALAR

2.1. Veri Madenciliği

Bilişim ve bilgisayar teknolojilerinin çok hızlı bir şekilde gelişim gösterdiği günümüzde her türlü sayısal bilgi saklanabilir ve gerektiğinde yeniden kullanabilmektedir. Teknolojideki bu gelişme beraberinde bir takım sorunları da getirmiştir.Gelişen teknoloji giderek artış gösteren devasa büyüklükteki verileri saklamaya yeterli olabilir ancak bu veriler ne zaman, ne işe yarayacaktır? Biriken veriler gerçek anlamda "*bilgiye*" dönüştürülebilecek midir? Bu sorulara, veriler üzerinde çözümlemeler yapmak amacı ile çeşitli istatistiksel ve matematiksel yöntemleri kullanılarak yanıt vermek mümkün fakat giderek artan veri sayısı ile sorunlar ortaya çıkacak ve veri tabanları üzerinde bu çözümlemeri yapmak zorlaşacaktır.

Bu tür veriler üzerinde çözümleme yapabilmek için hem yeni veri tabanı kavramlarına hemde yeni çözümleme yöntemlerine gereksinim duyulmaktadır. Veriyi yönetmek için "*veri ambarı* " ve verileri çözümleyerek yararlı bilgiye erişilmesini sağlayan "*veri madenciliği*" kavramları ortaya çıkmıştır [2].

---

2. Ö.Yalçın, *Veri Madenciliği Yöntemleri : Veri Madenciliğine Giriş* Papatya, syf.37-38, 2012.

2.1.1. Veri madenciliği nedir?

Veri Madenciliği kavramı için herkesin üzerinde hem fikir olduğu bir ifade bulunmamakla birlikte tercih ediliş biçimlerine göre değişik tanımlar yapılmıştır.Başlıca tanımlar aşağıdaki gibi sıralanmıştır.

- Veri madenciliği, büyük veri tabanlarından gizli bilgilerin, önceden tahmin edilemeyen örüntülerin ve yeni kuralların keşfedilmesi sürecidir [3].

- Veri madenciliği, veri içerisinde gizli kalmış, önceden bilinemeyen ve potansiyel olarak kullanışlı olan anlamlı bilginin çıkarımıdır [4].

- Veri madenciliği, içerisinde varolan anlamlı örüntü ve kuralları ortaya çıkarmak amacıyla, büyük miktarlardaki verinin otomatik ve yarı otomatik araçlar yardımıyla incelenmesi ve analiz edilmesi sürecidir [5].

- Veri madenciliği, çeşitli mimarilerde depolanmış olan büyük miktarlardaki verilerden ilgi çekici bilginin keşfedilmesi sürecidir [6].

- Veri madenciliği, veri ambarında tutulan çok çeşitli verilere dayanarak daha önce keşfedilmemiş bilgileri ortaya çıkarma ve bu bilgileri, karar vermek ve eylem planını gerçekleştirmek için kullanma sürecidir [7].

- Veri madenciliği, veriye sahip olan kişi ya da kurum için, kuralları ve ilişkileri keşfederek, önceden bilinemeyen açık ve yararlı sonuçlar elde etmek amacıyla, çok miktardaki verinin seçilmesi, incelenmesi ve modellenmesi sürecidir [8].

---

3. P. Adrians, D. Zantinge, *Data Mining*, Addison, 1997.
4. C. Clifton, B.Thuraisingham, "*Emerging Standards for Data Mining*," Computer Standards & Interfaces, syf.187-193, 2001.
5. W.J.Flawley, G.Piatetsky-Shapiro, C.J.Matheus *Knowledge Discovery in Databases : An Overview*, AI Magazine, syf.57-70, 1992.
6. M.J.A.Berry, G.S.Linoff, *Mastering Data Mining : The Art and Science of CRM,* The Art and Science of CRM, syf.7, 2000.
7. J.Han, M.Kamber, a.g.e., syf.7.
8. R. Swift *Accelerating Customer Relationship*, Prentice Hall PTR, syf.93, 2001.

Yukarıda bahsedilen tanımlardan yola çıkarak basit bir çerçevede birleştirmek gerekirse, veri Madenciliği, büyük ölçekli veriler arasından "*değeri olan"* bir bilgiyi gelecekte kullanmak üzere elde etme sürecidir.

2.1.2. Veri madenciliği tarihçesi

Veri Madenciliği teknikleri üzerinde matematikçiler 1950'li yıllarda çalışmaya başlamışlar, mantık ve bilgisayar bilimleri alanlarında yapay zeka (artificial intelligence) ve makine öğrenme (machine learning) konularını geliştirmişlerdir.1960'lı yıllarda ise istatistikçiler yeni bazı algoritmalar üzerinde çalışmalar yapmışlardır.Örneğin regresyon analizi (regression analysis), maksimum olabilirlik tahmini (maximum likelihood estimates), sinir ağları (neural networks) gibi çeşitli yöntemler veri madenciliğinin ilk adımlarını oluşturmuştur.Bunlara ek olarak veri tabanları sistemleri giderek gelişmiş ve büyük sayıda metin dökümanlarının saklanması ve bilginin geri kazanılması sağlanmıştır[9].

1970, 1980, 1990'lı yıllarda yeni programlama dilleri ve yeni bilgisayar tekniklerinin geliştirilmesi, genetik algoritmalar (genetic algorithms), kümeleme yöntemleri (clustering methods) ve karar ağaçları (decision tree algorithms) gibi algoritmaları da içermiştir [10] [11].

---

9. M.Albayrak, *EEG Sinyallerindeki Epileptiform Aktivitenin Veri Madenciliği Süreci ile Tespiti*, Doktora Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, syf.74, 2008.
10. P.F.Brown, D.V.J.Pietra, P.V.DeSouza *Class-based n-gram models of Natural Language*, Computational Linguistics, vol.18, syf.467–479, 1992.
11. O.Engin, A.Fığlalı, *Akıs Tipi Çizelgeleme Problemlerinin Genetik Algoritma Yardımı ile Çözümünde Uygun Çaprazlama Operatörünün Belirlenmesi*, Doğus Üniversitesi Dergisi, s.6, syf.27-35, 2002.

1990 yılıyla beraber veri tabanında bilgi keşfinin ilk adımları oluşturulmuş ve veri ambarı veritabanları (database warehouse) geliştirilmiştir. Aynı zaman dilimi içerisinde yeni teknolojilerle beraber veri madenciliği geliştirilerek yaygın olarak kullanılan standart bir işin parçası haline gelmiştir.

1990'lı yıllardan itibaren veri madenciliği konusu veri yoğun araştırma alanlarında bilgi keşfi ismi ile anılmaya başlanmıştır. İlk yıllar çoğunlukla veri tabanındaki veriler üzerinden yürütülen çalışmalar zaman içerisinde veritabanında saklanmayan verileri de kapsayacak biçimde genişletilmiştir. Geçmiş, tüm bu çalışmaların değerlendirilmesi veri madenciliğinin geleceği konusunda fikir vermesi açısından önem taşımaktadır.

Günümüzde veri madenciliği bankacılıktan sigortacılığa, tıp ve sağlık hizmetlerinden pazarlamaya kadar bir çok alanda başarıyla kullanılmaktadır. Hızlı bir şekilde gelişen teknoloji yeni yöntemlerin oluşmasına zemin hazırlamış, karar ağaçları, yapay sinir ağları, istatistiksel yöntemler, algoritmalar ve varyans analizleri gibi çeşitli yapıların oluşması ile birlikte bilgiye ulaşmak daha da kolaylaşmıştır.

2.2. Veri Madenciliği Süreci

Sürekli olarak veri madenciliğinin bir süreç olduğundan bahsediyoruz, bu süreç, ele alacağımız işin tanımlanması ile başlamakta ve sırasıyla işin anlaşılması, o iş ile ilgili verilerin toplanması, verilerin hazırlanması, verilere ve işe uygun modelin oluşturulması, tasarımı yapılan modelin uygunluğunun ve yeterliliğinin değerlendirilmesi ile devam etmekte ve son olarak modelin uygulanması ile sonuca ulaşmaktadır.Süreçten de

anlaşılacağı üzere veri madenciliği sadece bir parçadan değil bir parçanın belirli bir sırada kullanılması ile oluşmaktadır.

2.2.1. Standart veri madenciliği süreçleri

Veritabanında bilgi keşfi yapmak için standartlaşmış farklı yöntemler bulunmaktadır.Altı Sigma (Six Sigma), SEMMA (Sample, Explore, Modify, Model, Assess) ve CRISP-DM (Cross- Industry Standard Process for Data Mining) metodojileri en yaygın olarak kullanılanlardır.

2.2.1.1. Altı Sigma (Six Sigma)

Operasyonlarda en sağlıklı işleyişin yapılandırılması amacıyla işletmelerde süreçlerin tanımlanması, ölçülmesi, analiz edilmesi, iyileştirilmesi ve kontrolü için kolay ve etkili istatistik yöntemlerinin kullanıldığı bir iş yönetim stratejisi olarak tanımlanmaktadır [12]. Altı Sigma ile süreçlerin istenilen kalitede olup olmadığı ve kalitenin sayısal durumu görüntülenebilir [13]. Bu yakşalım süreç performansını iyileştirerek bir milyonda 3.4 birim hata oranına ulaşmayı amaçlar [14].

Bu yaklaşım ;

- Müşteri memnuniyetini arttırma
- Çevrim süresini düşürme
- Hataları azaltma

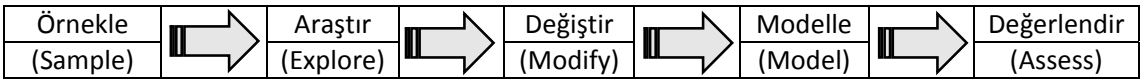
---

12. Y.S.Türkan, E.Manisalı, M.F.Çelikkol, *Evaluation of critical success factors effect on six sigma project success in Turkey's manufacturing sector*, Journal of Engineering and Natural Sciences, syf.105-117, 2009.
13. Y.Çabuk, S.Karayılmazlar, *Altı Sigma Yaklaşımı*, Bartın Orman Fakültesi Dergisi, s.94, 17 Aralık 2010.
14. A.Öztürk, *Kalite Yönetimi ve Planlaması*, Ekin Yayınevi, Bursa, ISBN 978-9944-141-79-6, 2009.

gibi üç konuya odaklanır [15]. Temel olarak Altı Sigma karar verme süreçlerinde deneyim yanında, doğru verinin doğru analizi ile oluşabilecek risklerin yönetimini, yönetsel ve istatistiksel araçlar ile yöneten kendini kanıtlamış bir metodolojidir.

2.2.1.2. SEMMA (sample, explore, modify, model, assess)

SEMMA, SAS şirketi tarafından yaratılan ve veri madenciliğinde yürütülen süreçlerin özünü gösteren, örnekle (**S**ample), araştır (**e**xplore), değiştir (**m**odify), modelle (**m**odel), değerlendir (**a**ssess) kelimelerinin baş harflerinden oluşmaktadır [16]. Bahsedilen SEMMA süreçlerinin aşamaları aşağıdaki gibidir.



Şekil 2.1. SEMMA İşlem Basamakları

Örnekle (Sample), modelleme için belirlenen yapıdan, veri seçme ve veri örnekleme ile başlar. Örneklenen veri seti verimli kullanılabilecek kadar küçük ama içerdiği bilgi açısından yeterince büyük olmalıdır. Bu aşama aynı zamanda veri bölümlemeyle de ilgilidir.

Araştır (Explore), bu aşama beklenmeyen eğilimleri arayarak veriler üzerindeki ilişkileri keşfeder, yeni bakış açısı ve fikirler kazandıran süreçi kapsar.

Değiştir (Modify), veri modelleme süreçi için hazırlık yaparak, veri oluşturma, veri değiştirme ve veri seçimi sürecini kapsar.

---

15. P.Pande, L.Holpp, *What is six sigma?*, McGraw-Hill, New York, ISBN 0-07-128185-6, 2002.
16. A.S.Koyuncugil, *Veri Madenciliği Ders Notları : Yönetim Bilişim Sistemleri*, http://www.koyuncugil.org/files/ders/bolum6.pdf , Ankara , 27 Aralık 2010.

Modelle (Model), aşamasında sonuçları alabilmek için hazırlanan veri seti üzerinde çeşitli veri madenciliği tekniklerinin uygulandığı süreçtir.

Değerlendir (Asses), son aşamada modelleme süreci sonuçlarının değerlendirilmesini, oluşturulan modellerin güvenilirlik ve kullanışlılık durumunu gösterir.

2.2.1.3. CRISP-DM (cross- industry standard process for data mining)

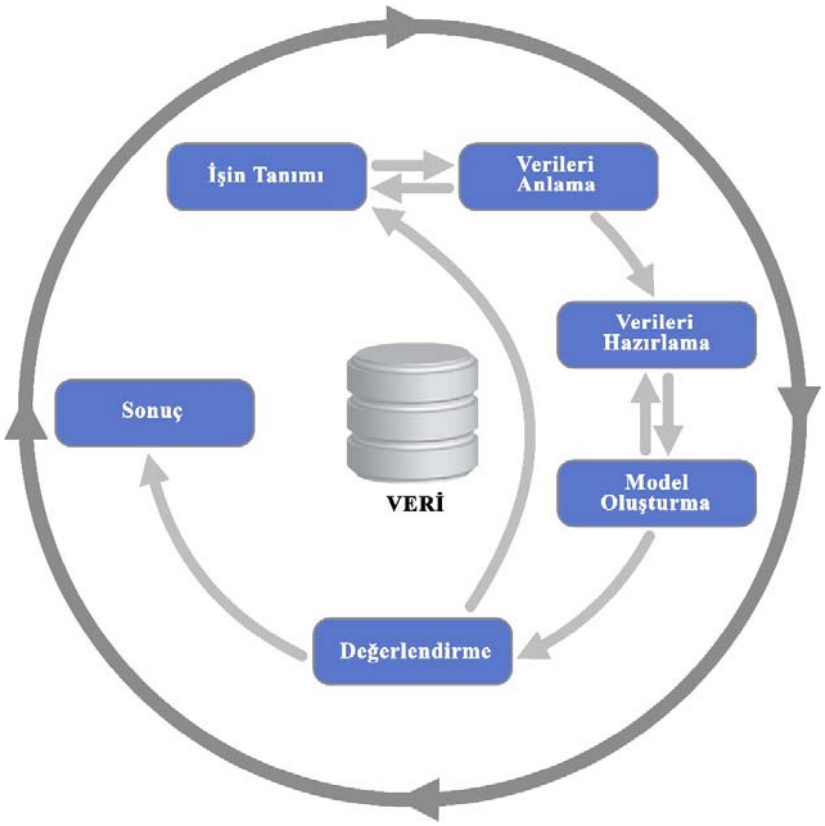
Daimler-Chrysler NCR Sistem Mühendisliği Kopenhag (Danimarka), SPSS (İngiltere) ve OHRA Verzegeringen en Bank Groep B.V (Hollanda) isimli firmalar tarafından tüm uygulayıcılara yönelik, veri madenciliği için iyi düşünülmüş, kimseye ait olmayan, ücretsiz bir standart süreç modeli hazırlamak üzere yola çıkılmış ve 1999 yılı sonunda veri madenciliği için Sanayilerarası Standart Süreç (VMSSS) (CRoss-Industry Standard Process for Data Mining - CRISP -DM) CRISP -DM 1.0 versiyonunu hazırlamış ve veri madenciliği için standart bir süreç modeli önermişlerdir. Bu sürecin adımları;

İşin tanımı, işletme açısından amaçları anlama ve bu bilgiyi bir veri madenciliği problemine dönüştürülmesidir. **Verinin anlaşılması**, veri kalitesini belirleme, verinin ilk kez anlaşılmasının keşfi için veri toplamaya başlanması sürecidir. **Verinin hazırlanması**, son veri setini oluşturmak için tüm faaliyetlerin kapsama alınmasıdır. **Modelleme**, değişik modelleme tekniklerinin seçilip, uygulanması ve ayarlanmasından oluşur. **Değerleme**, modelin kalitesinin değerlendirildiği aşamadır. **Sonuç**, karar verme sürecine yardım etmek için "güncel" bir modelin organizasyon içerisine uygulanması, adımlarından oluşur [17]. Şekil 2.2.'de CRISP-DM'nin süreci aşama aşama olarak gösterilmiştir.

---

17. S.Walczak, *Review of The Handbook of Data Mining. Organizational Research Methods* ,s.7, syf.119-121, 2004.

Uygulamamın aşamaları da bu standart süreci temel alarak hazırlanacaktır.



Şekil 2.2. CRISP-DM Sürecinin İşlem Basamakları

2.2.2. İşin tanımı

İşin tanımı, veri madenciliği sürecinin ilk aşamasıdır. Bu aşamada işin çerçevelerinin belirlenmesi ve iş sorununun anlaşılması ile başlar. Veri madenciliği ve ilgili alanlardaki uzmanların belirli bir bakış açısı ile proje hedeflerini ve gereksinimlerini tanımlamak için

birlikte çalışırlar. Projenin amacı daha sonar veri madenciliği tanımına çevrilir.Bu aşamada veri madenciliği araçlarının kullanılması gerekli değildir.

2.2.3. Verileri anlama

Başlangıç verilerinin oluşturulması, verileri tanımaya ve anlamaya yönelik analizlerin yapıldığı veri madenciliği sürecidir.Ayrıca veri kalitesi sorunlarıda bu aşamada değerlendirilir.Veri madenciliği uzmanları ile işin tanımı aşamasındaki iş uzmanları ile çok sık bilgi alışverişinde bulunmak verilerin daha doğru anlaşılmasını sağlamaktadır.Veri araştırma aşamasında geleneksek veri analizi araçları, örneğin istatistiksel yöntemler verileri araştırmak için kullanılır.

2.2.4. Verileri hazırlama

Veri madenciliğinin bir diğer aşaması ise verileri hazırlama aşamasıdır.Verileri hazırlama yada aynı anlama gelen bir başka değişle verileri ön işlemeden geçirme sürecinde hangi değişkenleri ve hangi kayıtların analizde kullanılacağına karar verilir.Genellikle veriler birden çok farklı kaynaklardan toparlanarak oluşturulur.Gözlem sayısının çok fazla olması durumunda örnekleme yapılarak zaman ve çeşitli maliyetlerden kazanım sağlanabilir.

Veriler hazırlanırken, veri setinde eksik veya gürültülü verilerin olup olmadığı control edilir.Eksik veriler veri girişi sırasında elde edilmemiş olan değerlerdir.Veri setimizde bu şekilde veriler bulunuyorsa eksik verilerin zamanla tamamlama yoluna gidilebilir, kayıtlar tamamen silinebilir yada eksik değer tahmini bri değer ile doldurulabilir.Gürültülü veriler ise veri girişinin hatalı olması durumunda ortaya çıkmaktadır.Örneğin, kişinin yaşının 500

yada sahip olduğu çocuk sayısının 99 olması gibi veri setinde bulunan özel durumlar dikkate alınarak veri setinin düzenlenmesi gerekmektedir.Bunların dışında veri kodlama işleminin standart olup olmadığı da control edilmelidir.Veri kodlama aşamasında cinsiyetler için (0,1) veya (Erkek, Kadın) değerleri girilmiş olabilir.Veri hazırlama sürecinde veri setinde yapılan standardizasyon işlemleri ile analiz sırasında yaşanacak sorunlar engellenmektedir [18]. Veri hazırlama teknikleri yukarıda da bahsedildiği gibi 4 ana aşamada gerçekleştirilir.

- Veri Temizleme,
- Veri Birleştirme,
- Veri Dönüştürme,
- Veri İndirgeme

2.2.4.1. Veri temizleme

Veri seti üzerinde yapılan incelemelerde bazen bazen verilerin istenilen yapıya sahip olmadığı durumlar görülebilir.Örnek olarak eksik, tutarsız veya uygun olmayan veriler ile karşılaşılabilir.Bu duruma verilerde gürültünün bulunması durumu denilmektedir.Verilerde bulunan gürültünün temizlenmesi ve eksik verilerin yerine yenilerinin belirlenerek konulması gerekmektedir.Bunun için aşağıda belirtilen yöntemlerden bir yada bir kaçı beraber kullanılabilir.

- Eksik değer bulunduran kayıt yada kayıt kümeleri veri setinden atılabilir.
- Bilinmeyen değerlerin yerine her yerde kullanılabilen genel sabit kullanılabilir.

18. U.T.Ş.Gürsoy, *Uygulamalı Veri Madenciliği: Sektörel Analizler*, Ankara, s.4, 2012.

- İlgili nitelikte bulunan tüm verilerin ortalaması, eksik değerler yerine kullanılabilir.
- Verilere uygun bir tahmin yapılarak (regresyon yada karar ağacı modeli) eksik değer yerine kullanılabilir [19].

2.2.4.2. Veri birleştirme

Veri madenciliğinde, genellikle farklı veri tabanlarından yada veri kaynaklarından elde edilen verilerin birlikte değerlendirilmesi için farklı yapıdakı verilerin ortak bir yapıya dönüştürülmesi yani verilerin birleştirilmesi gerekmektedir.Örneğin bir very kaynağında girişler "tüketici-ID" şeklinde yapılmışken, bir diğerinde "tüketici-numarası" şeklinde olmuş olabilir.Bu tip şema birleştirme hatalarından kaçınmak için meta veriler kullanılır.Veri tabanlarında yada veri ambarlarında çoğunlukla meta verisi bulunmaktadır.Veri birleştirmede önemli bir başka konuda indirgemedir.Bir değişken bir başka tablodan türetilmişse fazlalık yaratabilmekte ve değişkendeki bu fazlalıklar sonuçta elde edilen verı kümesinde fazlalıklara neden olabilmektedir.Bu fazlalıklar korelasyon analizi ile araştırabilir.Örneğin "tüketici-ID" ile "tüketici-numarası" için korelasyon katsayısı bulunabilir.Eğer bulunan korelasyon katsayısı yüksek çıkıyorsa değişkenlerden biri veri setinden çıkarılarak indirgeme işlemi yapılabilir.

2.2.4.3. Veri dönüştürme

Veri dönüştürme ile veriler, veri madenciliği için uygun formlara dönüştürülürler. Veri dönüştürme; düzeltme, birleştirme, genelleştirme ve normalleştirme gibi değişik

19. Y.Özkan, *Veri Madenciliğine Giriş : Veri Madenciliği Yöntemleri*, İstanbul, syf.40-41, 2008.

işlemlerden biri veya bir kaçını içerebilir. Veri normalleştirme en sık kullanılan veri dönüştürme işlemlerinden birisidir. Veri normalleştirme tekniklerinden bazıları aşağıdaki biçimde sıralanabilir [20] :

- Min-Max
- Z Skor
- Ondalık Ölçekleme

Min-max normalleştirmesi ile orijinal veriler yeni veri aralığına doğrusal dönüşüm ile dönüştürülürler. Bu veri aralığı genellikle 0-1 aralığıdır.

Z Skor normalleştirmede (veya 0 ortalama normalleştirme) ise değişkenin her hangi bir y değeri, değişkenin ortalaması ve standart sapmasına bağlı olarak bilinen Z dönüşümü ile normalleştirilir.

Ondalık ölçekleme ile normalleştirmede ise, ele alınan değişkenin değerlerinin ondalık kısmı hareket ettirilerek normalleştirme gerçekleştirilir. Hareket edecek ondalık nokta sayısı, değişkenin maksimum mutlak değerine bağlıdır. Ondalık ölçeklemenin formülü aşağıdaki şekildedir:

Örneğin 900 maksimum değer ise, n=3 olacağından 900 sayısı 0,9 olarak normalleştirilir.

---

20. R.J.Roiger, M.W.Geatz, *Data Mining A Tutorial : Based Primer* ,Wesley, USA, syf.350, 2003.

2.2.4.4. Veri indirgeme

Veri indirgeme teknikleri, daha küçük hacimli olarak ve veri kümesinin indirgenmiş bir örneğinin elde edilmesi amacıyla uygulanır. Bu sayede elde edilen indirgenmiş veri kümesine veri madenciliği teknikleri uygulanarak daha etkin sonuçlar elde edilebilir. Veri indirgeme yöntemleri aşağıdaki biçimde özetlenebilir:

- Veri birleştirme veya veri küpü (Data Aggregation or Data Cube)
- Boyut indirgeme (Dimension Reduction)
- Veri Sıkıştırma (Data Compression)
- Kesikli hale getirme (Discretization)

Veriyi indirgeme aşamasında verilerin çok boyutlu veri küpleri biçimine dönüştürmek söz konusu olabilir.Böylece çözümlemeler sadece belirlenen boyutlara göre yapılır.Veriler arasında bir seçme işlemi yapılarak, gereksiz veriler veritabanından çıkarılır ve boyut azaltılması sağlanabilir.Veri sıkıştırma aşamasında büyük veri kümelerinin sıkıştırılarak daha az yer işgal etmeleri sağlanır.Örnekleme aşamasında ise büyük veri topluluğunun yerine onu temsil eden daha küçük veri kümelerinin oluşturulması amaçlanır.

2.2.5. Model oluşturma

Modelleme aşamasında veri setinin yapısına ve analizin amacına uygun olan modele karar verilmelidir.Seçilen modelin varsayımlarına dikkat edilerek uygun algoritmalarda seçilebilir.Veri madenciliğinde kullanılan modeller "Tanımlayıcı ve Tahmin Edici Modeller" olmak üzere ikiye ayrılmaktadır [21].

---

21. C. Bounsaythip, R.Rinta, *Overview of Data Mining for Customer Behavıour Modeling : VTT Information Technology Research Report* , TTEI, USA, 2001.

2.2.5.1. Tanımlayıcı modeller

Tanımlayıcı modellerde (Descriptice Models) amaç karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanmasını sağlamaktır. Örneğin aylık geliri 1000 ile 2500 TL arasında olan ve en az bir arabası olan çocuklu aileler ile çocuğu olmayan, aylık geliri 800 ile 2000 TL arasında olan ailelerin satın alma örüntülerinin benzerlik gösterip göstermediğinin belirlenmesi tanımlayıcı modele bir örnektir [22]. Tanımlayıcı modeller;

- İlişki analizi,
- Kümeleme analizi olmak üzere iki grupta incelenmektedir.

Birliktelik kuralları (association rules) ile ardışık zamanlı örüntüler (sequential patterns analysis) ilişki analizi kapsamında yer almaktadır.

2.2.5.2 Tahmin edici modeller

Tahmin edici modellerde (Predictive Models) sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak, sonuçları bilinmeyen veri kümeleri için sonuç değerlerinin tahmin edilmesi amaçlanmaktadır. Örneğin, bir banka önceki dönemlerde müşterilerine vermiş olduğu kredi bilgilerine sahip olabilir. Bu verilerde kredi alan müşterilerin özellikleri ve kredilerin zamanında ve eksiksiz olarak geri ödenip ödenmeği tutulmaktadır. Eldeki bu veriler üzerinde kuralacak bir model ile daha sonraki kredi taleplerinde müşteri özelliklerine göre verilecek olan kredinin geri ödenip ödenmeyeceğinin tahmininde kullanılabilmektedir.

---

22. H.Akpınar, *Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği* , İ.Ü.İşletme Fakültesi Dergisi, İstanbul, s.1, 2000.

Tahmin edici modeller,

- Sınıflandırma,
- İstatistiksel tahmin modelleri olmak üzere iki ana başlık altında incelenebilmektedir.

Karar ağaçları, yapay sinir ağları ve genetik algoritmalar en yaygın olarak kullanılan sınıflandırma teknikleri olarak bilinmektedir. İstatistiksel tahmin modelleri ise regresyon analizi, ayırma (diskriminant) analizi ve lojistik regresyon analizini kapsamaktadır [23].

2.2.6. Değerlendirme

Başlangıçta belirlenen problemin amacı ile uygulanan model sonucunda elde edilen çıktıların uygun olup olmadığı bu aşamada incelenmektedir. Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkün olabilmektedir. Bu nedenle veri hazırlama ve model kurma aşamaları en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir. Bir modelin doğruluğunun test edilmesinde kullanılan en basit yöntem basit geçerlilik testidir. Bu yöntemde tipik olarak verilerin % 5 ile % 30 arasındaki bölümü ile test verileri olarak ayrılır ve kalan kısım üzerinde modelin öğrenimi gerçekleştirildikten sonra bu veriler üzerinde test işlemi yapılmaktadır. Bir sınıflandırma modelinde yanlış olarak sınıflanan olay sayısının, tüm olay sayına oranı, hata oranını, doğru olarak sınıflanan olay sayısının tüm olay sayısına oranına ise doğruluk oranını vermektedir. Doğruluk oranı ile hata oranın toplamı 1'dir. Önemli bir diğer değerledirme ise ölçüt modelin anlaşılabilirliğidir.

---

23. F.Gürbüz, L.Özbakır, H.Yapıcı, *Türkiye'de Bir Havayolu İşletmesine Ait Parça Söküm Raporlarına İlişkin Veri Madencilği Uygulaması*, Gazi Üniv. Müh. Mim. Fak. Der. Cilt 24, No 1, 73-78, 2009.

Bazı uygulamalarda doğruluk oranlarındaki küçük artışlar çok önem arz etse de bir çok uygulamada ilgili kararların neden verildiğinin yorumlanabilmesi çok daha büyük önem taşıyabilmektedir. Çok az sayıda da olsa bazı uygulamalardaki kararların yorumlanamayacak kadar karmaşıklık içerseler bile genel olarak karar ağacı ve kural temelli sistemler model tahmininin altında yatan nedenleri daha başarılı bir şekilde ortaya koyabilmektedir [24].

2.2.7. Sonuç

Sürecin son aşamasında, yapılan çalışmanın başkaları tarafından da tekrarlanabilirliğini sağlamak ve sonuçlarını karar vericilere aktarabilmek üzere kapsamlı bir rapor oluşturulur. Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısı ile ürettikleri verilerde ortaya çıkabilecek değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve gerekirse yeniden düzenlenmesini gerekli kılacağından çalışmanın takibi önem arz etmektedir.

2.3. Veri Madenciliği Yöntemleri

Veri madenciliği yöntemleri temelde “sınıflandırma”, “kümeleme”, “birliktelik kuralları” olmak üzere üç ana başlık altında incelenmektedir.

2.3.1. Sınıflandırma

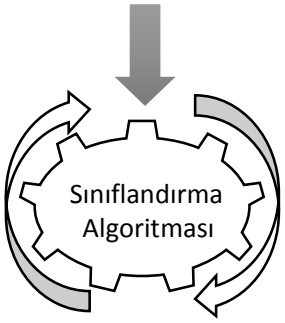
Veri madenciliğinin en çok kullandığı alandır. Var olan veri tabanının bır kısmı eğitim olarak kullanılarak sınıflandırma kuralları oluşturulur. Bu kurallar yardımı ile yeni bir durum ortaya çıktığında yeni bir durum ortaya çıktığında nasıl karar verileceği belirlenir.

---

24. W.DuMouchel, *Knowledge Discovery and Data Mining,* Proceeding of the Fifth ACM Conference, USA, syf 6-15, 1999.

Veri madenciliğinin sınıflandırma yöntemi içerisinde en sık kullandığı teknik karar ağaçlarıdır. Bunun yanında lojistik regresyon, diskriminant analizi, sinir ağları ve fuzzy setleri de kullanılmaktadır. İnsanlar verileri sürekli sınıflandırdıkları, kategorize ettikleri ve derecelendirdikleri için sınıflandırma, hem veri madenciliği hem de veri hazırlama aracı olarak da kullanılabilmektedir. Veri tabanındaki alan isimleri kullanılarak veri setine uygun bir model benimsenir.

Eğitim Verisi			
Müsteri	**Borç**	**Gelir**	**Risk**
C1	Yüksek	Yüksek	Kötü
C2	Yüksek	Yüksek	Kötü
C3	Yüksek	Düşük	Kötü
C4	Düşük	Yüksek	İyi
C5	Düşük	Düşük	Kötü
C6	Düşük	Yüksek	İyi



Sınıflayıcı Model
Eğer **Borç = Yüksek** ise **Risk = Kötü**
Eğer **Borç = Düşük** ve **Gelir = Düşük** ise **Risk = Kötü**
Eğer **Borç = Düşük** ve **Gelir = Yüksek** ise **Risk = İyi**

Şekil 2.3. Sınıflandırma Model Kurma Süreci

Sınıflandırma modelinin elde edilmesi için veri tabanından rasgele seçilen bir kısım veri, eğitim verisi olarak kullanılır. Sınıflandırma model kurulum süreci Şekil 2.3.' deki gibi yapılmaktadır.

Test verileri üzerinde sınıflandırma kuralları belirlendikten sonra söz konusu kurallar bu sefer test verileri baz alınarak sınanır. Örneğin C1 isimli yeni bir banka müşterinin bankadan kredi talebinde bulunduğunu varsayalım C1 müşterisinin risk durumunu belirlemek için örnek verilerden elde edilen karar kuralı doğrudan uygulanır. Bu müşteri için Borç=Düşük, Gelir=Yüksek olduğu biliniyorsa risk durumunun Risk=İyi olduğu kararı verilebilir.

Test Verisi			
Müsteri	**Borç**	**Gelir**	**Risk**
C7	Yüksek	Düşük	Kötü
C8	Düşük	Yüksek	İyi
C9	Düşük	Düşük	Kötü
C10	Yüksek	Yüksek	Kötü

Sınıflayıcı Model
Eğer **Borç = Yüksek** ise **Risk = Kötü**
Eğer **Borç = Düşük** ve **Gelir = Düşük** ise **Risk = Kötü**
Eğer **Borç = Düşük** ve **Gelir = Yüksek** ise **Risk = İyi**

Kredi Talep Eden Yeni Müşteri			
Müsteri	**Borç**	**Gelir**	**Risk**
CX	Düşük	Yüksek	**?**

Şekil 2.4. Modelin Uygulanması Süreci

Şekil 2.4.' deki gibi test sonucunda elde edilen modelin doğru olduğunu kabul edecek olursak, bu model diğer veriler üzerinde de uygulanabilir. Bu şekilde elde edilen model ile mevcut yada olası müşterilerin gelecekteki kredi talep risklerini belirlemek için kullanılabilir [25].

2.3.2. Kümeleme

Verilerin kendi aralarındaki benzerliklerin göz önüne alınarak gruplandırılması işlemidir ve kümeleme yöntemlerinin çoğu veriler arasındaki uzaklıkları kullanır. Hiyerarşik kümeleme yöntemleri en yakın ve en uzak komşu algoritmaları olarak tanımlabilir.Hiyerarşik olmayan kümeleme yöntemleri arasında ise k-ortalamalar yöntemi sayılabilir. Uygulamalarda çok sayıda kümeleme yöntemleri kullanılmaktadır. Bu yöntemler, değişkenler arasında benzerliklerden ya da farklılıklardan yararlanarak bir kümeyi alt kümelere ayırmada kullanılmaktadır. Hangi tekniğin kullanılacağı küme sayısına bağlı olmakla birlikte her iki tekniğin beraber kullanılması çok daha yararlıdır. Bu şekilde hem sonuçları hem de iki teknikten hangisinin daha uygun sonuçlar verdiğini karşılaştırmak mümkün olmacaktır.

Kümeleme analinizin amacı, gruplanmamış verileri benzerliklerine göre sınıflandırmak ve araştırmacıya özetleyici bilgiler elde etmede yardımcı olmaktır. Kümeleme analizinin uygulanabilmesi için verilerin normal dağılımlı olması varsayımı olmakla birlikte, bu varsayım teoride kalmakta ve uygulamalarda göz ardı edilmektedir. Bu varsayım sağlanması durumunda kümeleme analizinde Kovaryans matrisi için farklı bir varsayım

25. Ö.Yalçın, *Veri Madenciliği Yöntemleri : Veri Madenciliğine Giriş* Papatya, syf.40-45, 2012.

gerekmemektedir. “Küme, birbirine yakın (benzer) nesnelerin çok boyutlu uzayda oluşturdukları bulutlar benzetmesi” şeklinde tanımlanabilir [26]. Kümeleme analizi ise bu kümeleri oluşturma işidir.

Tablo 2.1. Örnek Kümeleme Analizi Verileri

Gözlemler	**Durum1**	**Durum2**
G1	1	1
G2	2	1
G3	4	5
G4	7	5
G5	5	5

Tablo 2.1.’ de belirtilen verilere “en yakın komşu algoritması” uygulandığında sonuçlar Tablo 2.2.’ de gösterildiği gibi oluşmaktadır.

Tablo 2.2. Kümeleme Algoritmasi Sonuçları

Kümeler	
Küme 1	1,2
Küme 2	4,5
Küme 3	3,4,5
Küme 4	1,2,3,4,5

2.3.3. Birliktelik Kuralları

Birliktelik kuralları, birbiriyle ilişkili olan değişkenlerin ortaya çıkarılması ve aralarındaki bağlantının büyüklüğünün tespit edilmesine yöneliktir. Birliktelik kuralları belirli türlerdeki veri yapıları arasındaki ilişkileri tanımlamaya çalışan bir yöntemdir.

26. H.Tatlıdil, *Uygulamalı Çok Değişkenli İstatistiksel Analiz*, Ziraat Matbaacılık, Ankara, syf.330, 2002.

Bağıntı analizi esasına dayanan ve veri madenciliği uygulamalarında çok kullanılan yöntemlerden birisi "sepet analizi" dir. Sepet analizi, özellikle işlemsel verileri ilişkilendirir. Birliktelik kurallarını bir örnek ile açıklamak gerekirse;

Bir mağazada alışveriş yapan müşterilerin alışveriş alışkanlıklarını belirlemek istediğimizi varsayalım.Örnek olarak beş adet müşterinin mağazadan hangi üzerinleri aldığını Tablo 2.3.' de gösterilmiştir.

Tablo 2.3. Alışveriş Sepet Bilgileri

Müşteri No	**Alınan Ürünler**
1	Un, Reçel, Diş Macunu, Parfüm
2	Un, Çikolata
3	Çikolata, Reçel, Diş Macunu, Yarabandı
4	Un, Çikolata, Reçel, Diş Macunu
5	Un, Çikolata, Reçel, Yarabandı

Tablo 2.3.' de gösterilen veriler üzerinde "Apriori" algoritması yardımı ile çalışma yapıldığında, sonuçları Tablo 2.4.' de gösterildiği gibi çıkmaktadır.

Tablo 2.4. Apriori Algoritması Sonuçları

{ Çikolata, Diş Macunu } --> { Reçel } (s=0.4, c=1.0)
{ Çikolata, Reçel } --> { Diş Macunu } (s=0.4, c=0.67)
{ Reçel, Diş Macunu } --> { Çikolata } (s=0.4, c=0.67)
{ Diş Macunu } --> { Çikolata, Reçel } (s=0.4, c=0.67)
{ Reçel } --> { Çikolata, Diş Macunu } (s=0.4, c=0.5)
{ Çikolata } --> { Reçel, Diş Macunu } (s=0.4, c=0.5)

Tablo 2.4.' de Apiori algoritmasının sonuçları, tablo satırlarının sırasıyla, çikolata ve diş macununu birlikte alanlar mutlaka reçelde alıyorlar, çikolata ve reçel satın alan müşteriler %67 olasılıkla diş macunuda alıyorlar, reçel ve diş macununu birlikte satın alanlar %67

olasılıkla çikolatada alıyorlar, diş macunu alanlar %67 olasılıkla çikolata ve reçel de satın alıyorlar, reçel satın alanlar %50 olasılıkla çikolata ve diş macunu da satın alıyorlar ve çikolata satın alanlar %50 olasılıkla reçel ve diş macunu da satın alıyorlar.

2.4. Veri Madenciliği Uygulama Alanları

Büyük hacimde veri bulunan her yerde veri madenciliği kullanmak mümkündür. Günümüzde karar verme sürecine ihtiyaç duyulan birçok alanda veri madenciliği uygulamaları yaygın olarak kullanılmaktadır. Örneğin pazarlama, bankacılık, sigortacılık, elektronik ticaret, tıp ve sağlık hizmetleri vb. birçok dalda başarılı uygulamalar görülmektedir [27] [28] [29].

Son 20 yıldır Amerika Birleşik Devletleri'nde çeşitli veri madenciliği algoritmalarının gizli dinlemeden, vergi kaçakçılıklarının ortaya çıkartılmasına kadar çeşitli uygulamalarda kullanıldığı bilinmektedir. Kaynaklar incelendiğinde veri madenciliğinin en çok kullanıldığı alan olarak tıp, biyoloji ve genetik görülmektedir.

2.4.1. Pazarlama

Pazarlama alnında, veri madenciliğinin başlıca uygulama alanları aşağıda maddeler halinde belirtilmiştir.

---

27. O. İnan, *Veri Madenciliği : Yüksek Lisans Tezi*, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, 2003.
28. M.Albayrak, *EEG Sinyallerindeki Epileptiform Aktivitenin Veri Madenciliği Süreci ile Tespiti : Doktora Tezi*, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, 2008.
29. Ö.Akgöbek, F.Çakır, *Veri Madenciliğinde Bir Uzman Sistem Tasarımı*, Akademik Bilişim 09, 11-13 Şubat Harran Üniversitesi, Şanlıurfa, syf.801-806, 2009.

- 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)

2.4.2. Bankacılık

Bankacılık alnında, veri madenciliğinin başlıca uygulama alanları aşağıda maddeler halinde belirtilmiştir.

- Farklı finansal göstergeler arasında gizli korelasyonların bulunması,
- Kredi kartı dolandırıcılıklarının tespiti (Fraud Detection),
- Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi,
- Kredi taleplerinin değerlendirilmesi.

2.4.3. Sigortacılık

Sigortacılık alnında, veri madenciliğinin başlıca uygulama alanları aşağıda maddeler halinde belirtilmiştir.

- Yeni poliçe talep edecek müşterilerin tahmin edilmesi,
- Sigorta dolandırıcılıklarının tespiti,

- Riskli müşteri örüntülerinin belirlenmesi.

2.4.4. Tıp ve sağlık hizmetleri

Tıp ve sağlık alanındaki verilerinin elektronik ortamlarda saklanması sonucunda, veri madenciliği teknikleri ile bulunabilecek sorular ve başlıca uygulama alanları aşağıda maddeler halinde belirtilmiştir.

- Hastalıkları etkileyen faktörlerin ortaya çıkartılması.
- Hastalıklara erken teşhis koyularak sağlığın korunması ve doğru tedavi yöntemlerinin seçilmesi.
- Sağlık hizmetlerinin kalitesinin artırılması ve geleceğe dönük doğru sağlık politikalarının oluşturulması.
- Koruyucu hekimliğin yaygınlaştırılması ve sağlık harcamalarının düşürülmesi.
- Salgın hastalıkların tespit edilmesi gerekli önlemlerin alınması.
- Sağlık harcamalarındaki hileli işlemlerin ortaya çıkartılması, maliyetlerin düşürülmesi.
- İlaç geliştirici firmaların, sağlık veritabanlarından yararlanarak doğru ilaçları geliştirmesi.
- Sağlık hizmetlerinde kalitenin artırılması [30].

30. A.S.Koyuncugil, N.Özgülbaş, *Veri madenciliğinin Tıp ve Sağlık Alanında Kullanımı*, Bilişim Teknolojileri Dergisi, sayı.2, syf.2, 2009.

III. TIP VE SAĞLIK HİZMETLERİNDE VERİ MADENCİLİĞİ

Özellikle mühendislik alanında yoğun olarak kullanılan veri madenciliği, son yıllarda tıp alanında verilerin çok büyük boyutlara ulaşması ile bu alanda geliştirilmeye başlanmış ve sağlık çalışanlarına büyük destek olmaya başlamıştır. Büyük veri yığınları içinden, veri madenciliği ve istatistiksel analiz yöntemleri ile değerli bilgiler elde edilmektedir. Bu bilgiler bilgisayar destekli tanı çalışmalarında, doktorların doğru karar vermelerine yardımcı olmakta ve sağlık uygulamalarının gelişmesine çeşitli katkılar sağlamaktadır [31].

3.1. Tarihçe

Sağlık bilgi sistemlerindeki veri madenciliği tekniklerinin ilk kullanımı 1970'lerde ve daha sonraki yıllarda geliştirilen uzman sistemlerle olmuştur. Uzman sistemlerin tıp alanında güçlü araçlar sunmasına rağmen, bu alandaki verilerin hızlı değişmesi ve uzmanlar arasındaki görüş farklılıkları nedeniyle çok yaygınlaşmamışlardır [32].

Daha sonraki yıllarda özellikle 1990'lı yıllarda hastaların gelecekteki sağlık durumları ve maliyet tahminleri gibi konuları araştırmak için sinir ağları kullanılmaya başlanmıştır.

---

31. M. Makinacı, C. Güneşer, *Göğüs Kanseri Verilerinin Sınıflandırılması*, Elektrik-Elektronik-Bilgisayar Mühendisliği 12. Ulusal Kongresi ve Fuarı Bildirileri Kitabı, s.1, 2007.
32. P.Yıldırım, M.Uludağ, A.Görür, *Hastane Bilgi Sistemlerinde Veri Madenciliği*, Çanakkale Üniversitesi Akademik Bilişim Dergisi, s.1, 2008.

Tıp alanında, 1984 yılında Londra'da baş gösteren kolera salgınında çok sayıda ölümler gerçekleşmiş.(10675 kişi) John Snow bir harita üzerinde ölen kişilerin yerlerini işaretlediğinde kayıpların bazı bölgelerde yoğunlaştığını fark ediyor ve o bölgede su pompalarında yapılan incelemeler sonrasında ölümlerin su tesisindeki problemden dolayı kolera hastalığından kaynaklandığını tespit ediyor.Su tesisinin iyileştirilmesi ile kolera salgını son buluyor.Bu çalışma sağlık alanında kağıt kalem ile yapılan ilk veri madenciliği çalışmalarındandır.

3.2. Tıpta Veri Madenciliği

Günümüzde bilgi sistemleri ve iletişim teknolojilerindeki gelişmeler sayesinde tıp ve sağlık alanındaki birçok veri sayısal ortamda saklanabilmekte ve kolaylıkla erişilebilmektedirler. Hastane bilgi sistemleri hastalara ait demografik bilgiler, hastalık ve tedavi durumları, yapılan tetkikler, faturalama ve idari işlere ait bilgileri içerirler. Sağlık ve tıp, çağımızın en önemli bilimsel araştırma alanları olduğu için bu alandaki bilgi sistemleri de araştırmalar için en büyük veri kaynaklarıdır. Son otuz yılda dünyada sağlık bilgi sistemlerinde büyük gelişmeler yaşanmıştır. Sağlık Bilişiminin yeni bir alan olmasına rağmen özellikle bilgi modelleme ve tanı araçlarında hızlı yenilikler yapılmıştır. Sağlık bilgi sistemlerindeki veri madenciliği tekniklerinin ilk kullanımı 1970'lerde ve daha sonraki yıllarda geliştirilen uzman sistemlerle olmuştur. Uzman sistemlerin tıp alanında güçlü araçlar sunmasına rağmen, bu alandaki verilerin hızlı değişmesi ve uzmanlar arasındaki görüş farklılıkları nedeniyle çok yaygınlaşmamıştırlar. Daha sonraki yıllarda özellikle 1990'lı yıllarda hastaların gelecekteki sağlık durumları ve maliyet tahminleri gibi konuları araştırmak için sinir ağları kullanılmaya başlanmıştır.

Sağlık ve tıp, günümüzün en çok bilgi ihtiyacı olan araştırma alanlarıdır. Son yıllarda özellikle sağlık veri modelleri, standartlar ve kodlama sistemlerindeki yenilikler sayesinde hastanelerde ve sağlık merkezlerinde kullanılan bilgi sistemlerinde önemli gelişmeler yaşanmıştır. Bu gelişmeler daha çok ve çeşitli verinin saklanabilmesini sağlamış ve beraberinde bilgi keşfi ihtiyacını ortaya çıkarmıştır. Veri Madenciliği, sağlık ve tıp alanındaki büyük veritabanlarından değerli bilgileri ortaya çıkartarak, hem tıp açısından hem de hizmet kalitesinin artırılması açısından büyük katkılar sağlar. Günümüzde uluslararası ortak projeler kapsamında geliştirilen ve biyoloji verilerinin saklandığı veritabanları, bu veritabanlarına erişim ve veri madenciliği sistemleri de klinik araştırmaların önemli bir parçası haline gelmişlerdir [32].

IV. BİYOMEDİKAL METİN MADENCİLİĞİ

Dünya üzerindeki verilerin neredeyse % 90'ı yapılandırılmamış formatta bulunmaktadır [33]. Yapılandırılmamış veriler, bilgisayar sistemleri tarafından kolayca tanınmayan ve veri yapısına sahip olmayan müzik, resim, video ve serbest formatta bulunan metinlerdir (e-postalar, dökümanlar, web sayfaları vb) [34]. Metin madenciliği belirli bir yapıya sahip olmayan, yazı tipindeki veriler içerisinden gizli kalmış anlamlı bilgilerin çıkarılması, düzensiz halde bulunan bulunan verilerin belirli bir yapıya uyarlanması, formatlanması sürecidir [35]. Metin madenciliği, yeni bir terim olmasına rağmen bilgi erişim sistemleri ve doğal dil işleme ile ilgili yapılan araştırmalara bağlı olarak ortaya çıkmıştır.1960'lı yıllarda bilgiye erişim ile ilgili çalışmalar başlamış, doğal dili anlamaya ve sayısallaştırmaya yönelik uygulamalar yazılmaya çalışılmıştır. 1990'lı yıllar gelindiğinde ise metinlere erişim, metinler üzerinden bilgilerin çıkarılması, metinlerin kategorize edilmesi ve metinleri yapısal hale getirmeye yönelik çalışmalar hız kazanmıştır.

Tıptaki verilerin çok büyük bir kısmı karmaşık, yapılandırılmamış formatta, kağıt tabanlı veya elektronik ortamlarda serbest metin olarak saklanmaktadır. Hekimler karar verme yada araştırma süreçlerinde hasta raporları, klinik çalışmalar, araştırma raporları, web sayfaları ve hastane kayıtları gibi serbest metin halinde bulunun yada kağıt tabanlı saklanan metinleri kullanmaktadır.

---

33. C.Bhatt, *Mining the Medical Literature*, http://ai.stanford.edu/~serafim/CS374_2004/Lecture%20Notes/lecture6.pdf, 27 Aralık 2013.
34. *Unstructured Data*, http://en.wikipedia.org/wiki/Unstructured_data, 27 Aralık 2013.
35. R.Hwa, *An Overview of Text Mining*, http://www.umiacs.umd.edu/~hwa/textmining.ppt, 27 Aralık 2013.

Yüksek boyutlarda bulunan belirli bir yapıya sahip olmayan verileri insan gücüyle analiz etmek ve istenilen bilgiye ulaşmak oldukça zorlu bir süreç almakta ve zaman kaybına yol açmaktadır. Hastayla ilgili karar verme sürecinde doğru ve eksiksiz bilgiye ulaşmanın öneminin ve bu verileri kullanarak istenilen sonuçlara ulaşmanın zorluğu göz önünde bulundurulduğunda bu tür sistemlere olan ihtiyaç daha belirgin hale gelmiştir [37].

4.1. Metin Madenciliği Nedir?

İnternet ve çeşitli boyutlardaki kişisel bilgisayarın yaygınlaşması ile gün geçtikçe büyüyen hacme sahip doküman yığınları oluşmaktadır. Bu yığınlar içerisindeki önemli bilgiler kaybolup giderken değerli bilgilere ulaşmak için dökümanların içeriğinin belirlenmesi, verilerinlerin formatlanarak sorgulanabilir yapıda olması ihtiyacı kendini hissettirmektedir. Durum böyle olunca metin madenciliği kavramı doğmuştur fakat günümüzde tam olarak tanımlanmamış olsa da son on yılda büyük gelişimler sağlanmış bir alanlardır. Metin madenciliği için yaygın olarak kullanılan tanım, belirli bir formata sahip olmayan, yazı tipindeki veriler içerisinden gizli kalmış, anlamlı bilgilerin çıkarılması bir başka deyişle düzensiz haldeki verilerin formatlanması süreci olduğuna yöneliktir [36].

Losiewicz metin madenciliğini, metin koleksiyonlarından bilgiye erişmeyi, bireysel metinlerde bilgi çıkarmayı, veri tabanlarından bilgi keşfini, organizasyonlarda bilgi yönetimini ve verinin,

---

36. B.Oğuz, U.Bilge, O.Saka, *Tıpta Metin Madenciliği*, Biyoistatistik ve Tıp Bilişimi AD, Akdeniz Üniversitesi, Antalya, syf.1,2.
37. M.Sharp, *Text Mining*, http://www.scils.rutgers.edu/~msharp/text_mining.htm, 27 Aralık 2013.

bilginin görselleştirilmesi aşamalarını birleştiren bir yapı olarak tanımlamıştır [38].

Metin madenciliği, metin icindeki kalıpları tanımlayıp bilinmeyen bilgiyi ortaya çıkararak, var olan yapısı ile metinleri bilgiye dönüştüren anahtar bir süreçtir. Bu süreçte metin ilk başta bir ön işlemden (cümle ve kelime analizleri) geçer, anlamsız kelimeler çıkarılır, metinler kategorilendirilir ve sonuçta geleneksel veri madenciliği yöntemleri kullanılarak (kümeleme, yapay sinir ağları, karar ağaçları, regresyon analizleri vb.) geniş hacimli metinler analiz edilir ve daha sonra elde edilen sonuçlar değerlendirilir. Farklı dillerde binlerce döküman, web sayfa içerikleri, yayınlar ve özetler göz önüne alındığında erişilmek istenen bilgilere ulaşmanın güçlüğü bilinmektedir. Araştırmacılar düzenli haldeki verileri analiz ettikleri gibi (yaş, cinsiyet, kilo, kolesterol, nabız, tansiyon vb) tıbbi raporlardan, internet sayfalarından, makalelerden, fatura bilgilerinden buldukları metin verileri de analiz edebilmektedirler. Bu metinlerin kısa sürede analiz edilmesi ve nitelikli bilgilere çok kısa sürede erişilmesi için metin madenciliği yöntemi kullanılmaktadır [37].

4.2. Metin Madenciliği Süreci

Metin madenciliği süreci, genel olarak altı adımdan oluşmaktadır.

- **Metin toplama**: İstenilen konular için bilgi erişim sistemleri kullanılarak metin koleksiyonu oluşturma sürecidir. Bu süreç, son zamanlarda genel olarak internet üzerinden, çeşitli arama motorları kullanılarak gerçekleştirilmektedir.

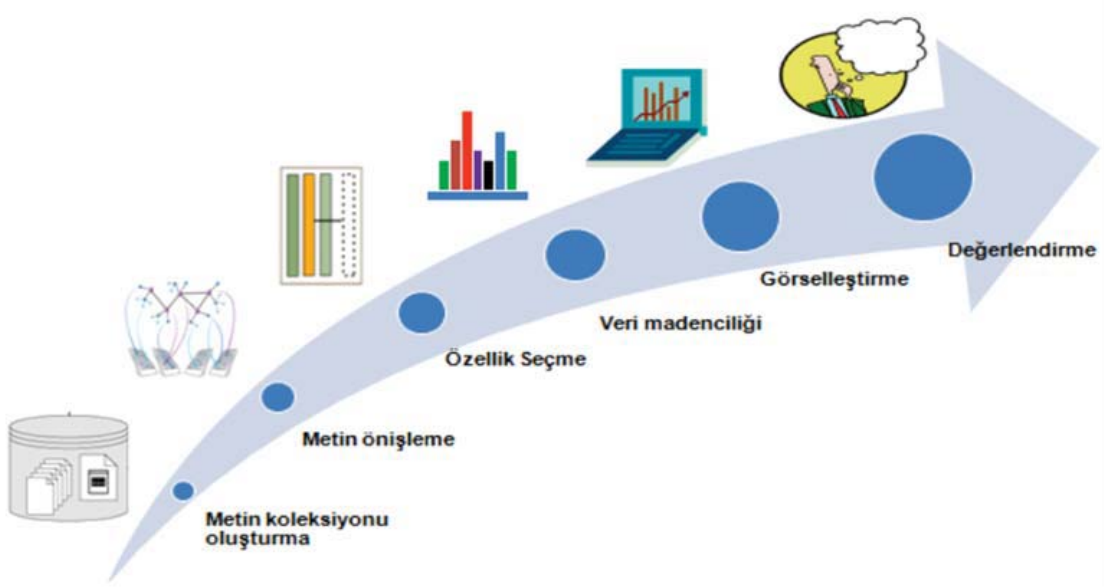
---

38. Cerrito, *Inside text mining: text mining provides a powerful diagnosis of hospital quality rankings - Data Warehousing/Mining*, http://findarticles.com/p/articles/mi_m0DUD/is_3_25/ai_114167705/pg_2 , 2006.

Tıp alanında ise metin toplama süreci yaygın olarak "PubMed" çevrim içi veritabanı üzerinden yapılmaktadır.

- **Metin önişleme**: Metni kelimelere ayırma, kelimelerin anlamsal değerlerini bulma (isim, sıfat, fiil, zarf, zamir vb.), kelimeleri köklerine ayırma ve gereksiz kelimeleri ayıklama, dökümanı fazlalıklardan arındırmak, gereksiz bilgileri cıkarmak, yazım kurallarına uygunluğunu tespit etmek, ekleri ve genel kelimeleri cıkarmak, önek ve sonekleri ayırmak, kelime anlamlarını belirlemek (kelimenin hangi anlamı kullanılıyor?), ilişki analizi (A'nın B ile bir ilişkisi varsa, B'nin de C ile bir ilişkisi varsa, A ile C arasında da potansiyel bir ilişki vardır) yapmak gibi metin belgelerin yapı taşı olan kelimelerle ilgili işlemleri içeren süreçtir.
- **Özellik seçme:** Bu aşamada ön işlemden geçen metinlerdeki önemli kelimeleri belirleme (isimler, tamlamalar, bileşik kelimeler, kısaltmalar, sayılar, tarihler, para birimleri vb.) ve ilişkili olmayan özelliklerin çıkarılması işlemleri yapılmaktadır.
- **Veri madenciliği**: Yapılandırılmış format haline getirilen metinlerin geleneksel veri madenciliği teknikleriyle (karar ağacları, yapay sinir ağları, kümeleme vb.) analizi sürecidir. Hem veri madenciliğinde hem de metin madenciliğinde gizli bilgilere bakılmakta ve genel yapay zeka, makine öğrenimi ve istatistik algoritmaları kullanılmaktadır. Veri madenciliğinde yapılandırılmış sayısal veri kullanılırken metin madenciliği yapılandırılmamış metinlerle ilgilidir. Veri madenciliğinde kullanılan veriler veri ambarlarından çıkartılmış, dönüştürülmüş ve yüklenmiş durumda bulunan verileri kullanırken metin madenciliği kesin olmayan verileri modellemeye çalışmaktadır [39].

39. M.Konchady, *Text Mining Application Programming. 1st ed.* ,Charles River Media, 2006.



Şekil 4.1. Metin Madenciliği Süreci

- **Görselleştirme:** Çalışmaların sonunda elde edilen sonuçların kullanıcılara gösteriminde en etkin ve anlaşılır biçimde (grafik, tablo vb. kullanılarak) sunumunun yapılması aşamasıdır.
- **Değerlendirme**: Genel olarak sistemlerin değerlendirilmesinde duyarlılık (precision), anma (recall) ya da ikisinin birleşiminden oluşan F-score ölçütü kullanılmaktadır. Duyarlılık, erişim çıktısındaki ilgili belge sayısının erişim çıktısındaki belge sayısına oranıdır. Anma ise, erişim çıktısındaki ilgili belge sayısının belgeler kümesinde ilgili belgeler sayısına oranı olarak tanımlanmaktadır.

4.3. Biyomedikal Metin Madenciliği

Biyomedikal metin madenciliğinin tıp alanındaki kullanımları son birkaç yılda büyük oranda artış göstermiştir. Tıptaki verilerin büyük bir kısmının serbest metin halinde bulunması önemli bilgilerin gözden kaçmasına, bilgiye hızlı ve doğru bir şekilde erişimin zorlaşmasına neden olmaktadır. Özellikle sağlık ile ilgili kayıtların elektronik ortamlarda tutulması Sağlık Bilgi Yönetiminin son yıllardaki en önemli hedeflerinden birisiyken böyle bir sistemin başarısının, klinik dökümantasyonun serbest metin halinde tutulmasından dolayı sınırlanmış durumda olması bu tür sistemlere olan ihtiyacı ortaya çıkarmıştır. Yapılan klinik çalışmalar, araştırma raporları, doktor notları, hastane kayıtları, prosedürler, labaratuar sonuçları ve faturalar gibi dökümanlar tıptaki en önemli veri kaynakları sayılmaktadır ve bu verilerin büyük bir kısmı serbest metin formatında bulunmaktadır.

Metin madenciği, tıp alanında özellikle tıbbi araştırmalarda, semptomlarla hastalıklar ve ilaçlarla kimyasal maddeler arasında nedensel bağları bulmada, hasta kayıtlarının analiz edilmesinde, gen-gen ve protein-protein ilişkilerinin tanımlanmasında, tanı ve tedavileri geliştirmek, servis kalitesini ve faydayı arttırmak, maliyetleri kontrol etmek için kullanılmaktadır [33].

4.4. Biyomedikal Metin Madenciliği Alanında Yapılmış Çalışmalar

Bu kısımda, biyomedikal metin madenciliği ile ilişkili yapılmış başlıca çalışmalara yer verilmiş ve çalışma bilgileri kısa özetler halinde anlatılmıştır. Shatgay'ın çalışmasında, DNA mikroarray deneylerinde genler arasındaki fonksiyonel ilişkilerin keşfedilmesi için biyomedikal yayınlar taranmış ve erişilen makalelerin özetlerinin içeriğine dayanarak ilişkiler bulunmaya çalışılmıştır [40].

Johnson ve arkadaşları, serbest formatta bulunan radyoloji raporlarında bulunan bilgileri yapılandıran ve çıkartan RADA adlı bir sistem tasarlamışlardır. Sistem, serbest formatta bulunan metinleri yapılandırmış hale dönüştürürken doğal dil işleme tekniklerini kullanmaktadır. Sistemin değerlendirilmesi için göğüs onkolojisi bölümünden 100 adet radyoloji raporu rastgele seçilmiştir. Raporlardaki bir cümlenin içerdiği ortalama kelime sayısı 14.97 ve sistemin bir raporu analiz etme süresi yaklaşık olarak 1-2 dakika olarak bulunmuştur. Ayrıca sistemden elde edilen sonuçların değerlendirilmesinde altın standart olarak göğüs radyolojisi uzmanından yardım alınmıştır. Sonuçta sistemin anma oranı % 85, duyarlılığı % 89 bulunmuştur [41].

---

40. H.Shatkay, S.Edwards, W.Wilbur, M.Boguski, *Genes, themes and microarrays, using information retrieval for large-scale gene analysis*, In Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, syf.317-328, 2000.
41. D.B.Johnson, R.K.Taira, A.F.Cardenas, D.R.Aberle, *Extracting Information from Free Text Radiology Reports*, Int J Digit Libr 997, sayı.1,syf.297-308.

Biyomedikal alanda metin madenciliği ile ilgili özellikle isimlendirilmiş varlıkları tanıma (named-entity recognition, NER) alanında çalışmalar yapılmıştır. İsimlendirilmiş varlıkları tanıma, metin koleksiyonlarında bulunan tüm isim varlıkların (ilaç isimleri, hastalık isimleri, simgeler) tanımlanması işlemidir. NER, metin içindeki varlıkların tanımlanması varlıklar arasındaki ilişkilerin bulunmasına, anahtar kavramların tanımlanmasına ve bu kavramların uygun bir şekilde sunulmasına olanak sağlamaktadır. Biyomedikal alanda NER ile ilgili yapılan çalışmalarda serbest metinlerdeki gen ve protein isimlerini tanımaya odaklanılmıştır. Biyolojik isim varlıkların tüm tiplerinin tam olarak belirtildiği bir sözlük bulunmaması, biyolojik varlıların çok kelimeli olabilmesi, aynı varlığın birden fazla isim alabilmesi vb. yaşanan problemler arasındadır. Bu alanda geliştirilen sistemlerin değerlendirilmesi doğruluk ve duyarlılık ölçümleri ile yapılmaktadır. Bu iki ölçümün birleştirilmesi ile oluşturulan ölçüm ise F-score (harmonik ortalama) ölçümü ise yaygın olarak tercih edilmektedir [42].

Schadow ve arkadaşları serbest formatta bulunan cerrahi patoloji raporlarından numuneler ve numunelerle ilgili bulgular hakkındaki bilgilere erişmeyi sağlayan bir metod geliştirmişlerdir. Patoloji laboratuarından elde edilen 622 adet rapor otomatik olarak eleme yapan bir sistem tarafından taranmış ve içinde "tanı" kelimesi geçmeyen raporlar ayıklanmıştır. Sonuç olarak geriye kalan 275 rapor XML formatına dönüştürülmüş ve doku tipi, yeri, toplama methodu ile ilgili bilgiler elde edilmiştir [44].

---

42. A.M.Cohen, W.R.Hersh, *A Survey of Current Work in Biomedical Text Mining, Briefings in Bioinformatics*, sayı.6, syf.57–71, 2005.
43. G.Schadow, C.J.Mcdonald, *Extracting Structured Information from Free Text Pathology Reports*, AMIA Annu Symp Proc., syf.584, 2003.

Swanson'un çalışmasında ise, metin madenciliği teknikleri hastalıklar ve semptomları arasındaki ilişkilerin ve bağıntıların bulunması için kullanılmıştır. Tıbbi araştırma sayfaları, makaleler, haberler kullanılarak semptomlar, ilaçlar, hastalıklar, kimyasallar arasında ilişki örüntülerine bakılmıştır [33]. Medikal başlık ve özetler incelenerek belirlenen bir problemin (az görülen bir hastalık) nedenleri karşılaştırılmıştır. Başlıklar arasında nedensel ilişkinin bulunması için ARROWSMITH adlı bir yazılım geliştirilmiştir [44]. Swanson'ın sisteminde MEDLINE'da iki başlık kullanılarak arama yaptırılır (magnezyum ve migren) ve sonuçlar (başlıklar ve özetler), yaygın olarak bulunan önemli kelime ve tamlamaları liste şekline getiren ARROWSMITH programına atılır. Migrene bağlı magnezyum eksikliği problem olarak belirlendikten sonra beslenmeyle ve migrenle ilgili literatür taraması yapılmıştır. Sonuç olarak tüm elde edilen veriler incelendikten ve modellendikten sonra magnezyum eksikliğinin migren ağrılarına neden olabileceği bulunmuştur [45].

Lindsay ve arkadaşları Swanson'ın yaklaşımını, metin madenciliği olarak adlandırmadan genişletmişlerdir. Lindsey ve Gordon bu yaklaşıma genel kelimeler ve tamlamaları bulmak için kelime sıklığı istatistiklerini de eklemişlerdir. Fakat Swanson'ın yaklaşımındaki gibi hala birkaç noktada "insan filtrelere" ihtiyaç duyulmaktadır [37].

---

44. T.Bekhuis, *Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy*, Biomedical Digital Libraries, 3:2, 2006.
45. N. Gürsakal, *Sözcük ve Sayı*, www.20.uludag.edu.tr/~gursakal/down/say.ppt, 05 Kasım 2006.

Tanabe ve arkadaşları tarafından geliştirilen AbGene sistemi, biyomedikal metinlerdeki gen ve protein isim varlıkların tanımlanması için oluşturulan en başarılı kural tabanlı yaklaşımlardan birine sahiptir. Sistemde, gen isimleri ile gen olmayan isimlerin ayrımının yapılması için gereken kuralları oluşturmada tümevarımsal mantıksal programlama kullanılmıştır. Sistemin duyarlılık % 85.7, anma oranı % 66.7'dir. Hanisch çalışmasında gen ve protein isimlerinin yer aldığı ve kelimelerin anlamsal olarak sınıflandırıldığı geniş bir sözlük kullanmıştır. Sistemin doğruluğu % 0.95, duyarlılığı % 0.90'dır [46].

46. A.M.Cohen, W.R.Hersh, *A Survey of Current Work in Biomedical Text Mining, Briefings in Bioinformatics*, syf.38.

V. ERYTHROMYCIN İLACININ YAN ETKİLERİNİN ARAŞTIRILMASI İÇİN WEB'DEKİ HASTA YORUMLARININ ANALİZİ

Her geçen yıl içinde içinde fizyolojik ve biyometrik hasta belirtilerinin tedavisi için ilaç kullanımında sürekli bir artış yaşanmaktadır.2009 yılında IMS Türkiye ve Türk Eczacıları Birliği'nde yayınlanan makalede Türkiye'de 2005 yılında 1 milyar 212 milyon, 2006 yılında 1 milyar 292 milyon, 2007 yılında 1 milyar 398 milyon, 2008 yılında, 1 milyar 476 milyon, 2009 yılı Ocak-Ekim ayları arasında 1 milyar 239 milyon kutu ilaç satışı olmuştur [1]. Bu rakamlar Türkiye ilaç pazarının 2005-2009 yılları arası değişimini göstermektedir.

Tedavilerde kullanılacak tıbbi ilaçların satış sürecinden önce klinik çalışmalarda güvenilirlik ve etkinlik değerlendirmeleri yapılmaktadır. Bu çalışmalar genellikle ilaç şirketleri tarafından yürütülür ve kısa sürede kesin sonuçlar almak için çok az insan içerir. İnternet kullanımının yaygınlaşması ile ilaç kullanan kişiler çesitli web ve forum siteleri aracılığı ile yaşadığı süreci ve görülen yan etkileri paylaşmaya başladılar. Bu bilgiler ilaç üretim şirketleri, doktorlar ve diğer hastalar için önemli bilgi kaynağı olmaktadır. Bu tür siteler ve forumlarda yazılan yorumların bazılarında klinik dil ve terminoloji kullanılması bazılarında ise günlük konuşma dilini içermesi ve paylaşılan verilerde gürültünün olması kullanılan dil farklılığını gidermek ve daha anlamlı bilgilere ulaşmak için veri madenciliği çalışmalarının yapılmasını zorunlu kılmaktadır.

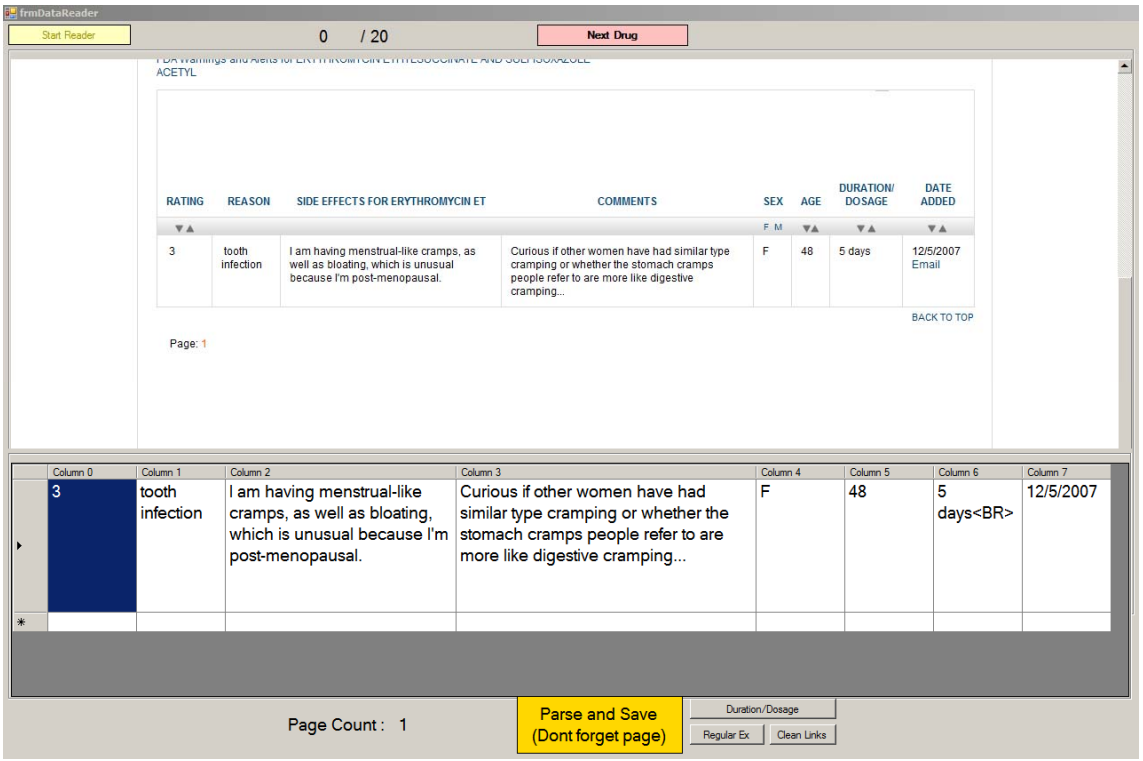
5.1. Neden Erythromycin ?

Erythromycin, bir makrolit antibiyotiğidir. İlacının seçilme nedeni ise Helikobakter Pilori ve Mycobacterium Tuberculosis gibi dünyada yaygın olarak görülen bakterilerin tedavisinde etkin olarak kullanılmasından kaynaklanmaktadır. Helikobakter Pilori, mide ve oniki parmak barsağı ülserleri ile kronik gastritin en önemli etkeni olarak kabul edilmekte öte yandan Mycobacterium Tuberculosis ise günümüzde halen morbidite ve mortalitesi yüksek bir hastalık olarak seyreden tüberkiloz (verem) hastalığına yol açmaktadır.

Bu çalışmada, erythromycin ilacına ait web üzerinde bulunan yorumların çeşitli veri madenciliği ve biyomedikal metin madenciliği yöntemleri ile analiz edilerek elde edilen sonuçlar, hasta, doktor ve ilaç üreten şirketlere yol gösterir olacaktır.

5.2. Hasta Yorumlarından Veritabanının Oluşturulması

Hastaların kullandıkları ilaçlar hakkkında, kullanım nedeni, görülen yan etkiler, yaş ve cinsiyet gibi bilgilerin paylaşıldığı, 10 yılı aşkın süredir faliyet gösteren ve bir çok ilaç hakkında bilgi içeren www.askapatient.com web sitesindeki erythromycin ilacı hakkında yapılan yorumlar, C#.Net dilinde geliştirilen ve Şekil 6.1’ de arayüzü bulunan WindowsForm uygulaması ile verilerin okunması sağlanmış ayrıca Ek-J’ de programda kullanılan html veri okuma ve parçalanmasını içeren methotlar, Ek-K’ de Kullanım Süresi ve Dozaj niteliklerinin ayrıştırılmasına ait kodlar, Ek-L’ de ise bu işlemlerin kurgulandığı veri tabanı modeli gösterilmiştir.



Şekil 6.1. Html Sayfası Okuyucu Programın Ekran Görüntüsü

Bu uygulamanın içerisinde bulunan webbrowser nesnesi ile ilgili İnternet sayfası html formatta okunarak Tablo 6.1.' de belirtilen yapıda toplamda yedi adet nitelik olacak şekilde kural tabanlı metin parçalama yöntemleri uygulanarak veri tabanına kayıt işlemi yapılmış ve toplam 711 adet kayıt bulunan veri seti oluşturulmuştur.

Tablo 6.1. Veriseti

Nitelik Adı	**Nitelik Açıklaması**	**Nitelik Tipi**	**Örnek Veri**
CommentID	Kayıt No	integer	1
Rating	Değerlendirme	integer	3
Reason	Kullanım nedeni	text	Tonsillitis (bademcik iltihabı)
Sideeffects	Yan Etkiler	text	Severe chest pain (şiddetli göğüs ağrısı)
Sex	Cinsiyet	bool	Female (Kadın)
Age	Yaş	integer	19
DurationDosage	Kullanım süresi ve Dozu	text	3 Days / 1000 mg 1xD

5.3. Veri Hazırlama Süreci

Veri hazırlama yöntemlerinden veri temizleme, birleştirme, dönüştürme, indirgeme ve bu teknikler ile ilişkili çeşitli algoritmalar kullanılarak aşağıdaki işlemler yapılmıştır.

5.3.1. Kullanım süresi ve dozunun iki ayrı nitelik olarak ayrılması

Kullanım süresi ve dozu alanında "1 days
 500 mg 4XD" şeklindeki gibi birlikte yer alan bilgiler parçalama methodu kullanılarak Kullanım Süresi (1 days) ve Dozaj (500 mg 4XD) olacak şekilde ayrıştırma işlemi yapılmıştır. İşlem sonucunda eksik ve hatalı veri girişinden kaynaklanan Tablo 6.2. 'deki yapı kural tabanlı veri taşıma ile Tablo 6.3.'deki gibi olacak şekilde, 426 kayıt için gerçekleştirilmiştir.

Tablo 6.2. Hatalı Veri Yapısı

Kullanım Süresi	**Kullanım Dozu**
500 mg 4XD	null
1 days 	null

Tablo 6.3. Doğru Veri Yapısı

Kullanım Süresi	**Kullanım Dozu**
null	500 mg
1 days	null

5.3.2. Kullanım süresi niteliğinin veri dönüştürme işlemleri

Yazılan yorumları daha anlaşılır ve standart bir yapıya getirmek amacı ile Tablo 6.5.' deki gibi bulunan Kullanım Süresi niteliğindeki veriler için parçalama işlemi ve Tablo 6.4. 'daki parametre tablosu kullanılarak dönüştürme işlemi yapılmıştır. Sonuç olarak Tablo 6.6. 'de gösterilen şekilde Kullanım Süresi Sayısal verisi, Kullanım Süresi Zamanı ve dönüşüm sonrası Yeni Kullanım Süresi Sayısal verisi olacak şekilde ayrılmıştır ve standartizasyon sağlanmıştır.

Tablo 6.4. Standardizasyon Parametreleri

Giriş Parametresi	Dönüştürülmüş Parametreler
1 month	30 days
1 week	7 days
1 year	365 days

Tablo 6.5. Standardizasyon Öncesi

Kullanım Süresi
3 days
1 months
6 weeks
1 years

Tablo 6.6. Standardizasyon Sonrası

Kullanım Süresi Sayısal	Kullanım Süresi Zaman	Kullanım Süresi Sayısal (Yeni)
3	D	3
1	M	30
6	W	45
1	Y	365

5.3.3. Kullanım dozu niteliğinin veri dönüştürme işlemleri

Dozaj niteliğinde “250 mg 2XD” şeklinde bulunan verinin dönüşümü ve standardizasyon çalışmasında Tablo 6.7. ’deki gibi üç ayrı niteliğe ayrılmıştır.

Tablo 6.7. Dozaj Standardizasyon Ayrıştırılması

Dozaj	Dozaj (Kullanım miktarı)	Mg (Dozaj Kullanım Birimi)	Kullanım Sıklığı	Dozaj Kullanım Periyodu
250 mg 2XD	250	mg	2	D (Gün)
500 mg 4XD	500	mg	4	D (Gün)
500 2XD	500	mg	2	D (Gün)
250 mg 4	250	mg	4	D (Gün)
4 perday	-	mg	4	D (Gün)
2 tabs 2X	-	mg	2	D (Gün)
8XD	-	mg	8	D (Gün)

Veri dönüşümü sırasında Dozaj alanında bulunan verilerde gürültü ile karşılaşılmıştır. Tablo 6.7. ‘deki gibi gürültülü örnek kayıtlarda veri dönüştürme işlemi yapılmıştır.

Dönüşüm işleminden sonra Dozaj (Kullanım miktarı) verisi olmayan kayıtlar için Dozaj (Kullanım miktarı) verisi bulunan kayıtların günlük kullanımına bakılarak frekans hesaplaması yapılmış ve frekansı en yüksek olan kayıt baz alınarak Dozaj (Kullanım miktarı) verisi bulunmayan sekiz kayıt güncellenmiştir. Tablo 6.8. ’de günlük Dozaj (Kullanım miktarı) frekans hesaplamaları gösterilmiştir.

Tablo 6.8. Günlük Kullanım Frekans Tablosu

Frekans (Adet)	Dozaj (mg)
4	500
3	250
1	1000
1	30

5.3.4. Kullanım nedeni niteliğinin veri dönüştürme işlemleri

Tablo 6.9. ve Tablo 6.10.‘ da gösterilen ilaç kullanım nedeni niteliğinin yazım hatalarının ve aynı anlama gelebilecek kelimelerin normalizasyon çalışması yapılmıştır. Bu çalışmadan sonra ilacın kullanım nedenlerinin frekansı hesaplanmış ve ilk 15 tanesi Tablo 6.11.’ de grafiksel gösterimide Şekil 6.2.’ deki gibi gösterilmiştir.

Tablo 6.9. Kullanım Nedeni Niteliğinin Normalizasyon Tablosu

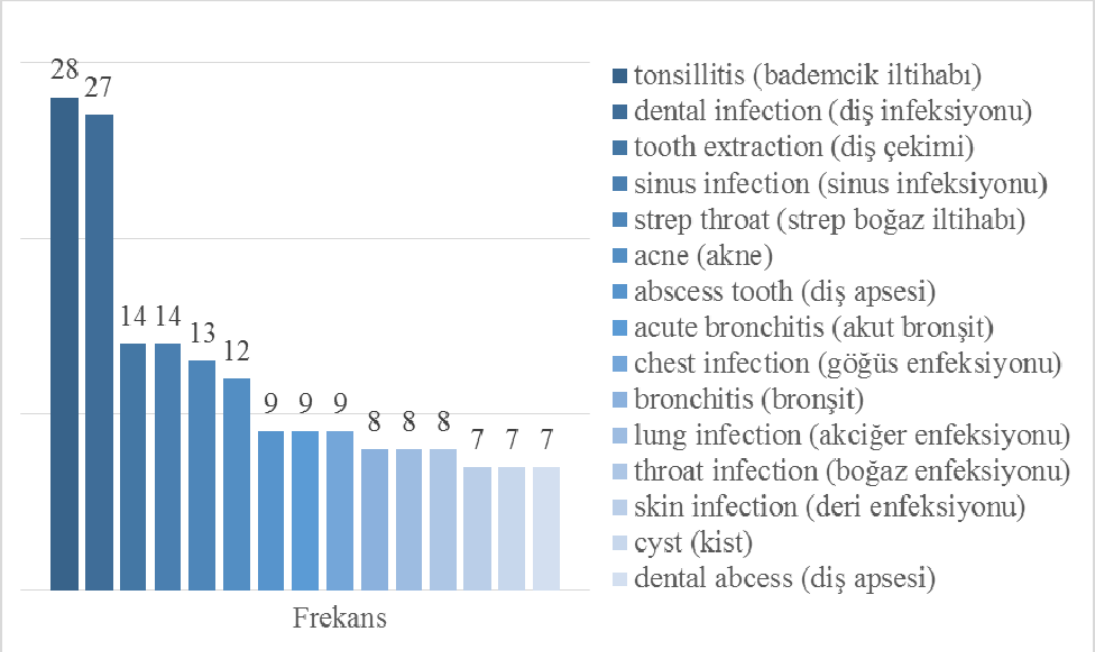
Normalizasyon Giriş	**Normalizasyon Çıkış**
abcesss	abscess
ance	acne
bacteria	bacterial
cellulitus	cellulitis
sunus	sinüs
tonsillitus	tonsillitis
roseaca	rosacea

Tablo 6.10. Kullanım Nedeni Niteliği İçin Aynı Anlama Gelebilecek Kelimeler

Kabul Edilen	**Aynı Anlama Gelebilecek Kelimeler**
dental infection (diş enfeksiyonu)	tooth ache / infection (diş ağrısı / enfeksiyonu)
	tooth infection (diş enfeksiyonu)
	dental premedication (diş premedikosyonu)
acne (sivilce)	acne on back (sırtta sivilce)
	scalp acne (kafada sivilce)
	body acne (vücutta sivilce)
	severe acne (şiddetli sivilce)
	slight acne (hafif sivilce)
dental extraction (diş çekimi)	dental surgery (diş cerrahisi),
	before dental work (çekim öncesi)
	following tooth extraction (çekim öncesi)
	post oral surgery (ağız cerrahisi sonrası)
	tooth extaction (diş alımı),
	wisdom teeth extraction (yirmi yaş dişleri çıkarma)

Tablo 6.11. En Çok Tekrar Eden İlaç Kullanım Nedenleri ve Frekansları

İlaç Kullanım Nedenleri	Frekans
tonsillitis (bademcik iltihabı)	28
dental infection (diş infeksiyonu)	27
tooth extraction (diş çekimi)	14
sinus infection (sinus infeksiyonu)	14
strep throat (strep boğaz iltihabı)	13
acne (akne)	12
abscess tooth (diş apsesi)	9
acute bronchitis (akut bronşit)	9
chest infection (göğüs enfeksiyonu)	9
bronchitis (bronşit)	8
lung infection (akciğer enfeksiyonu)	8
throat infection (boğaz enfeksiyonu)	8
skin infection (deri enfeksiyonu)	7
cyst (kist)	7
dental abcess (diş apsesi)	7



Şekil 6.2. En Çok Tekrar Eden İlaç Kullanım Nedenleri ve Frekansları

5.3.5. Yan etkiler niteliğinin veri dönüştürme işlemleri

Tablo 6.12. ve Tablo 6.13.‘ de gösterilen ilaç yan etkiler niteliğinin yazım hatalarının ve aynı anlama gelebilecek kelimelerin normalizasyonu yapılmıştır.

Tablo 6.12. Yan Etkiler Niteliğinin Normalizasyon Tablosu

Normalizasyon Giriş	**Normalizasyon Çıkış**
stomache	stomach
stomaceh	
stomacah	
stomacha	
stomcah	

Tablo 6.13. Yan Etkiler Niteliği İçin Aynı Anlama Gelebilecek Kelimeler

Kabul Edilen	**Aynı Anlama Gelebilcek Kelimeler**
numbness (uyuşukluk)	numbness (uyuşukluk)
	lethargy (uyuşukluk)
	sluggishness (uyuşukluk)
weakness (yorgunluk zayıflık)	weakness (halsizlik)
	fatigue (yorgunluk)
	tiredness (yorgunluk)
	exhaustion (yorgunluk)
	extreme fatigue (aşırı yorgunluk)

5.3.6. Yan etkiler isim listesinin oluşturulması

Yan etkiler isimlerinin listesi, Avrupa Moleküler Biyoloji Laboratuvarının bünyesinde çalışmalarını yürüten Sider2 (Side Effect Resource – Yan Etki Kaynakları) araştırma grubunun veritabanından elde edilmiştir [47]. Sider2 ilaç yan etkileri araştırma grubu,

47. Yan Etkiler Araştırma Grubu, *EMBL : Avrupa Moleküler Biyoloji Laboratuvarı Sider2*, http://sideeffects.embl.de/, 27 Aralık 2013.

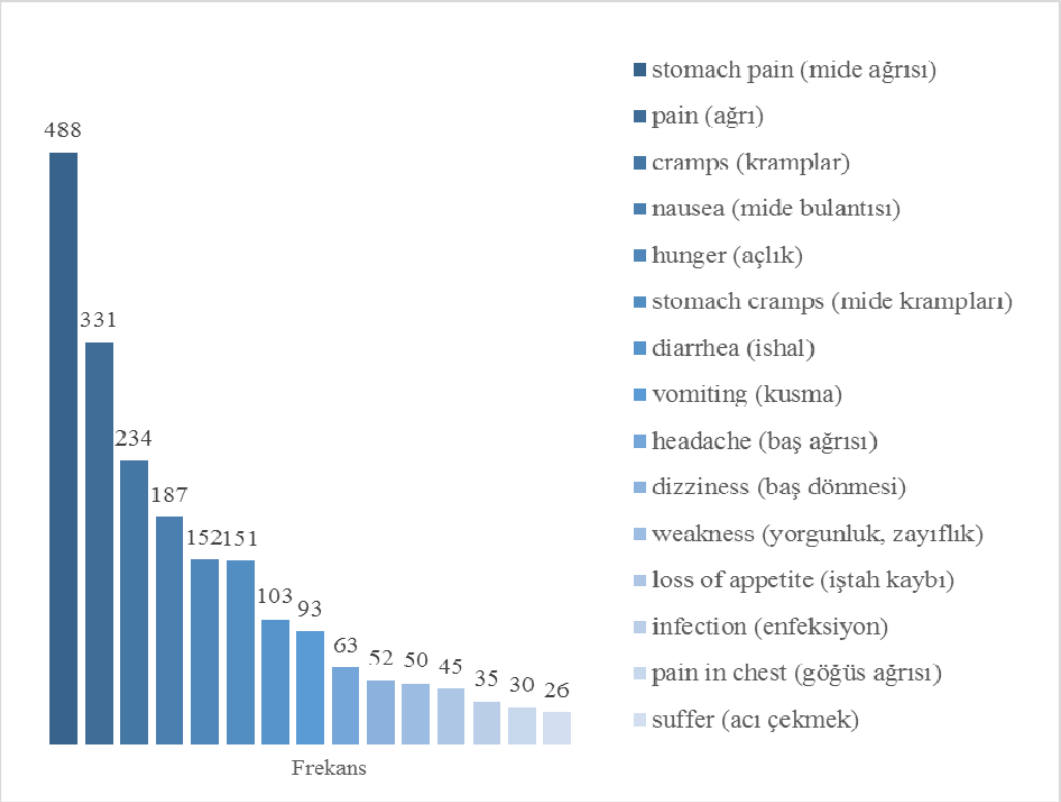
kamudan bu bilgilerin toplanması ve ilaç yan etkileri veritabanının oluşturulmasına yönelik çalışmalar yapan bir araştırma grubudur.

5.3.7. Yan etkiler niteliğinin frekans hesaplanması

Yan Etkiler niteliğinde bulunan verilerde her bir kayıt için Avrupa Moleküler Biyoloji Lavoratuvarının bünyesinde yer alan Sider2 ilaç yan etkileri araştırma grubundan alınan yan etki isimleri ile karşılaştırılmış ve rastlanan yan etkilerin frekans hesaplaması yapılmıştır. Hesaplama sonucunda frekansı en yüksek olandan en düşüğe doğru sıralama yapılarak ilk 15 tanesi Tablo 6.14.‘ de ve grafiksel olarakta Şekil 6.3.‘ de gösterilmiştir.

Tablo 6.14. En Çok Tekrar Eden Yan Etkiler ve Frekansları

Yan etkiler	**Frekans**
stomach pain (mide ağrısı)	488
pain (ağrı)	331
cramps (kramplar)	234
nausea (mide bulantısı)	187
hunger (açlık)	152
stomach cramps (mide krampları)	151
diarrhea (ishal)	103
vomiting (kusma)	93
headache (baş ağrısı)	63
dizziness (baş dönmesi)	52
weakness (yorgunluk, zayıflık)	50
loss of appetite (iştah kaybı)	45
infection (enfeksiyon)	35
pain in chest (göğüs ağrısı)	30
suffer (acı çekmek)	26



Şekil 6.3. En Çok Tekrar Eden Yan Etkiler ve Frekansları

5.3.8. Yaş niteliğinin kategorilere ayrılması

Veri kümesinde bulunan yaş niteliği NCBI Pubmed (National Center for Biolotechnology Information Advances Science and Health) den alınan bilgiler doğrultusunda Tablo 6.15.' deki gibi kategorilere ayrılmıştır [48].

48. NCBI Pubmed, *National Center for Biolotechnology Information Advances Science and Health*, http://www.ncbi.nlm.nih.gov/, 27 Aralık 2013.

Tablo 6.15. Yaş Niteliğinin Kategorileri

Yaş Aralığı	Kategori
0-6	Okul öncesi çocukluk
07-12	Çocukluk
13-24	Genç
25-43	Yetişkin
44-64	Orta Yaşlı
65 ≤	Yaşlı

5.4. Kullanılan Algoritmalar

Bu çalışmada, Apriori ve N-Gram algoritmaları ile web sitesine yapılan yorumların içeri sinden, klinik dil ve terminolojiye ait kelimelerin kelimelerin ve birliktelik kuralları çerçevesinde anlamlı ilişkileri ve örüntüleri saptamaya çalışılmıştır.

5.4.1. Apriori algoritması

Veri madenciliğinde kullanılan ve veri kümeleri veya veriler arasındaki ilişkiyi çıkarmak için geliştirilmiş algoritmalardan bir tanesidir. Algoritmanın ismi, kendinden önceki çıkarımlara bağlı olduğu için, latince, önce anlamına gelen “prior” kelimesinden gelmektedir.Apriori algoritması, özellikle çok büyük ölçekli veri tabanları (VLDB, very large databases) üzerindeki veri madenciliği (datamining) çalışmalarında geliştirilmiştir. Genel anlamda münasebet kuralı (association rule, birliktelik kuralı) çıkarımında kullanılan bir algoritmadır. Algoritmanın amacı, veri tabanında bulunan satırlar arasındaki bağlantıyı ortaya çıkarmaktır.

Algoritma yapı olarak, aşağıdan yukarıya (bottom-up) yaklaşımı kullanmakta olup, her seferinde tek bir elemanı incelemekte ve bu elemanla diğer adaylarla münasebetini ortaya çıkarmaya çalışmaktadır.Ayrıca algoritmanın her eleman için çalışmasını, bir arama

algoritmasına benzetmek mümkündür [49]. Algoritma, bu anlamda sığ öncelikli arama (breadth first search) yapısında olup, sanki adayları birer ağaç (tree) gibi düşünerek bu ağaç üzerinde arıyor kabul edilebilir.Ağaç yapısında, k elemanlı bir aday listesinden k-1 elemana baktıktan sonra, alt frekans örüntüsü yetersiz olan elemanları budamakta ve kalan elemanların üzerinden arama yapmaya devam etmekte ve bu şekilde birbiriyle ilişkili olan elemanlar saptanmaktadır. Şekil 6.4.‘ de Apriori pseudon (sahte kod) kodu gösterilmiştir.

$$
\begin{array}{l}
\text{Apriori}(T, \epsilon) \\
\quad L_1 \leftarrow \{\text{large } 1-\text{itemsets}\} \\
\quad k \leftarrow 2 \\
\quad\quad \textbf{while } L_{k-1} \neq \ emptyset \\
\quad\quad\quad C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \in \bigcup L_{k-1} \wedge b \notin a\} \\
\quad\quad\quad \textbf{for } \text{transactions } t \in T \\
\quad\quad\quad\quad C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\} \\
\quad\quad\quad\quad \textbf{for } \text{candidates } c \in C_t \\
\quad\quad\quad\quad\quad count[c] \leftarrow count[c] + 1 \\
\quad\quad\quad L_k \leftarrow \{c \mid c \in C_k \wedge \ count[c] \geq \epsilon\} \\
\quad\quad\quad k \leftarrow k + 1 \\
\quad\quad \textbf{return } \bigcup_k L_k
\end{array}
$$

Şekil 6.4. Apriori Pseudon Kodu

Çalışmada ilk olarak veriseti, java programlama dilinde geliştirmiş olduğum, kodları Ek-M’ de bulunan apriori algoritması ile çalıştırılmıştır fakat elde edilen sonuçlarda destek (support) ve güven (confidence) skorları tam olarak oluşturulamadığı için açık kaynak kodlu, genel kabul görmüş ve akademik çalışmalarda kullanılan WEKA programında bulunan apriori algoritmasının kullanılmasına karar verilmiştir.

---

49. E.Şeker, *Bilgisayar Kavramları*, http://www.bilgisayarkavramlari.com/2011/09/07/apriori-algoritmasi/, 27 Aralık 2013.

5.4.2. N-Gram algoritması

Elde bulunan bir metnin hangi dile ait olduğunu, ancak o metin içerisindeki harflerin ve kelimelerin birbirleri ile ilişkilerinin bilinmesi ile mümkün olabilmektedir.Bu ilişkilerin ortaya konması için, istatistiksel yöntemler (olasılık) kullanılmaktadır.

İstatistiksel dil modellemede amaç, sıradaki kelimeyi, daha önce karşılaşılan kelimeler aracılığıyla tahmin etmektir. İlk çalışmalardan biri Shannon'a (1951) aittir ve "Shannon Game" ile bir metindeki sonraki harfi tahmin etmeye çalışmıştır. Bu çalışmayı takip eden

birçok farklı çalışma literatürde yer alsa da N-gram modeli, dil modellemede en basit ve en başarılı temeli oluşturmuştur [50]. N-gram dil modelleri sıradaki kelimenin görülme olasılığının ondan önceki n-1 kelimeye dayandığını varsayar. Karakter n-gram yöntemi aynı yaklaşımı karakterlerin görülme sıraları için yapar.

N-gram yöntemi, dokümanların benzerliklerinin incelenmesinde ve kümeleme çalışmalarında kullanıldığı gibi genelde büyük boyutlu metinlere uygulanır ve metin içinde kullanılan her kelimenin olasılıkları hesaplanarak elde edilen sonuçlar, takip eden kelimelerin görülme olasılıklarına yansıtılır.

Bir metin içerisinde bulunan bir harfin o metinde bulunma olasılığı Denklem 5.1.' de ve aynı şekilde bir harf dizisinin (kelimenin) bir metin içinde bulunma olasılığı da Denklem 5.2.' deki gibi hesaplanmaktadır.

---

50. X.Huang, F.Peng, A.An, D.Shuurmans, N. Cercone, *Applying Machine Learning to Text Segmentation for Information Retrieval*, Information Retrieval sayı.6, syf.333–362, 2003.

$$\text{Olasılık} = \frac{\text{Metin içindeki “seçilen harf” sayısı}}{\text{Metin içindeki toplam harf sayısı}} \tag{5.1}$$

$$\text{Olasılık} = \frac{\text{Metin içindeki “kelime” sayısı}}{\text{Metin içindeki toplam kelime sayısı}} \tag{5.2}$$

Yukarıda verildiği gibi bir cümleyi oluşturan kelimelerin o cümle içinde bulunma olasılıkları $P(w_1, w_2, w_3 .... w_{n-1}, w_n)$ olduğunda, zincir kuralının kullanılması ile bu olasılıkların biraraya getirilmeleri münkün olmaktadır ve Denklem 5.3.’ de gösterilmiştir [51].

$$P(w_1^n) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1^2)...P(w_n \mid w_1^{n-1}) \tag{5.3}$$

$$P(w_1^n) = \prod_{K=1}^{N} P(w_n \mid w_1^{n-1}) \tag{5.4}$$

Denklem 5.4. ’deki hesaplama sonucunda da, kelimelerin ve kelime içindeki harflerin ardalanmalarının olasılıkları hesaplanmaktadır.

Bir cümle içerisindeki bir harf dizisinin bulunabilme olasılığını belirleyen algoritmalar, aynı zamanda tamamlanmış bir cümle içinde bulunan ve gelmesi muhtemel bir sonraki kelimenin belirlenmesinde de kullanılabilir. Kelime tahmininde kullanılan modeller “N-Gram” olarak adlandırılmaktadır. N-gram modeli yukarıda da belirtildiği gibi bir sonraki

51. P.Garret, *Making-Breaking Codes,* ISBN 0-13-030369-0, Prentice Hall, 2001.

kelimenin o metin içerisinde bulunabilme olasılığını belirleyebilmek için önceki n-1 adet kelimeyi kullanmaktadır.Konuşma tanıma işleminde, bu türdeki kelime ardalanmalarının istatistiksel modellerini belirtmek için Dil Modeli – DM (Language Model - LM) terimi kullanılmaktadır [52].

5.4.3. K-Means kümeleme algoritması

En eski kümeleme yöntemlerinden biri olan k-means 1957 yılında ilk kez Hugo Steinhaus'un öne sürdüğü bir fikir olmasına rağmen 1967 yılında J.B. MacQueen tarafından geliştirilmiştir [53]. K-means algoritmasının genel mantığı n adet veri nesnesinden oluşan bir veri setini, giriş parametresi olarak verilen k adet kümeye bölümlemektir. Amaç, gerçekleştirilen bölümleme işlemi sonunda elde edilen kümelerin, küme içi benzerliklerinin maksimum ve kümeler arası benzerliklerinin minimum olmasını sağlamaktır. Küme benzerliği, kümenin ağırlık merkezi olarak kabul edilen bir nesne ile kümedeki diğer nesneler arasındaki uzaklıkların ortalama değeri ile ölçülmektedir [54][55].

En yaygın kullanılan gözetimsiz öğrenme yöntemlerinden birisi olan k-means' in 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. Eşit büyüklükte küresel kümeleri bulmaya eğilimlidir [56].

---

52. D.Jurafsky, J.H.Martin, *Speech and Language Processing*, Prentice Hall, 2000.
53. http://en.wikipedia.org/wiki/K-means_clustering, 27 Aralık 2013.
54. J.Han, M.Kamber, *Data Mining Concepts and Techniques*, Morgan Kauffmann Publishers Inc., 2001.
55. P.Berkhin, *Survey of Clustering Data Mining Techniques*, San Jose,California, USA, Accrue Software Inc., 2002.
56. M.Işık, A.Y.Çamurcu, *K-means, K-medoids ve Bulanık C-means Algoritmalarının Uygulamalı Olarak Performanslarının Tespiti*, 2007.

Algoritmaya k-means adı verilmesinin nedeni, algoritmanın çalışmasından önce sabit bir küme sayısına ihtiyaç duyulmasıdır. Küme sayısı k ile gösterilir ve elemanlarının birbirlerine olan yakınlıklarına göre oluşacak grup sayısını ifade eder.
Buna göre k önceden bilinen ve kümeleme işlemi bitene kadar değeri değişmeyen sabit bir pozitif tam sayıdır [57].

Bazı kümeleme algoritmaları bazı verilerde daha iyi sonuçlar vermesine rağmen k-means kümeleme algoritması her çeşit veride kabul edilebilir sonuçlar verir. Algoritmanın en büyük dezavantajı yerel optimumlarda kalarak genel optimumlara ulaşamamasıdır [64].

Çok yaygın kullanımı olan bu algoritmanın aşağıda belirtildiği gibi birtakım zayıf yanları da bulunmaktadır [58]:

- Algoritmanın başlangıcında giriş parametresi olarak bir k sayısına ihtiyaç vardır. Elde edilecek olan sonuçlar k sayısına göre değişkenlik gösterebilir. Eğer küme sayısı belirli değil ise deneme yoluyla en uygun sayı bulunur.

- Aşırı gürültü ve istisna veriler algoritmayla hesaplanan ortalamayı değiştirdiği için k-means algoritması gürültü ve istisnaya karşı çok duyarlıdır.
Algoritma uygulanmadan önce veriler gürültü veya istisnadan temizlenebilir.

- Çakışan kümelerde iyi sonuç vermez.

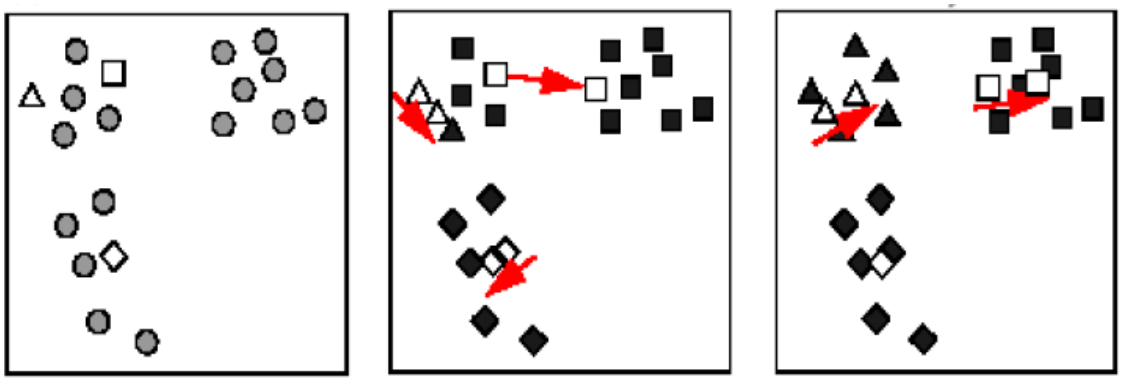
- Her eleman aynı anda verilen bir kümenin içindedir veya dışındadır.

- K-means algoritması sadece sayısal veriler ile kullanılabilmektedir. Kategorik verilerin kümelenmesi için k-means algoritması bir çözüm sunmaz [56].

---

57. E.Dinçer, *Veri Madenciliğinde K-means Algoritması ve Tıp Alanında Uygulanması*, s.24-64, 2006.
58. Y.Yünel, *K-means Kümeleme Algoritmasının Genetik Algoritma Kullanılarak Geliştirilmesi*, s.1,2, 2010.

K-means algoritmasının çalışma şekline bir örnek Şekil 6.5' de görülmektedir. Bu örnekte k=3 olarak seçilmiş ve beyaz simgelerle Şekil 6.5(a)' da rastgele seçilen küme merkezlerini temsil etmektedir. Şekil 6.5(b)' de geri kalan noktalar (bu kez aynı simgelerin siyah olanları) aynı şekilli ve beyaz renk olan küme merkezlerine dahil edilerek ilk kümeler oluşturulur. Bu işlem sonunda küme merkezleri her kümedeki elemanların ortalaması dikkate alınarak tekrar hesaplanır. Değişen küme merkezleri Şekil 6.5(b)' de oklar ile gösterilmiştir. Şekil 6.5(c)' de aynı işlem tekrar edildiğinde küme merkezlerinin değişimi görülmektedir. Bu şekilde başlangıç durumunda rastgele seçilen küme merkezleri, sürekli yinelemeler ile gerçek kümelenme alanlarının ortasına doğru yaklaşır. Bu işleme merkeze yakınsama denir. Merkeze yakınsama minimum seviyeye geldiğinde veya durduğunda kümeleme işlemi sona erer [57].



(a) İlk Küme Merkezleri (b) İlk Ortalama Hesabı (c) Merkeze Yakınsama

Şekil 6.5. K-Means Kümeleme Algoritması

Bu işlemleri matematiksel olarak ele aldığımızda k-means aralık ölçekli veriler için uzaklık yada komşuluk mesafesi hesaplamada üç çeşit uzaklık formülü kullanır.

1. Öklid Uzaklığı : En sık kullanılan yöntemdir. İki ya da daha çok boyutlu düzlemde kolaylıkla kullanılabilir ve Denklem 5.5.' de gösterilen

$$d(i,j)=\sqrt{(x_{i1}+x_{j1})^2+(x_{i2}+x_{j2})^2+...+(x_{ip}+x_{jp})^2} \tag{5.5}$$

ifadesi ile verilmektedir. Burada *i* ve *j* ifadeleri *p* boyutlu veri nesnelerini temsil etmektedir.

2. Manhattan Uzaklığı : *p* boyutlu uzayda herhangi iki noktanın karşılıklı her bir koordinat değerinin farkı alınarak bulunur ve bu ifade Denklem 5.6.' da gösterilmiştir.

$$d(i,j)=\left|x_{i1}-x_{j1}\right|+\left|x_{i2}-x_{j2}\right|+...+\left|x_{ip}-x_{jp}\right| \tag{5.6}$$

3. Minkowski Uzaklığı : Öklid ve Manhattan uzaklığının genelleştirilmiş hali olarak

$$d(i,j)=(\left|x_{i1}-x_{j1}\right|^q+\left|x_{i2}-x_{j2}\right|^q+...+\left|x_{ip}-x_{jp}\right|^q)^{1/q} \tag{5.7}$$

Denklem 5.7.' deki gibi ifade edilir. *q* bir pozitif tam sayı olmak üzere bu ifade q = 1 için Manhattan uzaklığını, q = 2 için Öklid uzaklığını belirtir. *q* değişkeninin değeri artırıldıkça daha hassas uzaklık ölçüm ifadeleri elde edilir.

5.5. Modelin Yapısı

Bu kısımda, önceki kısımlarda proje hedefleri ve veri madenciliği amaçları doğrultusunda şekillendirilen Erythromcin ilacına ait hasta yorumları verileri üzerinde veri hazırlama süreçleri ile oluşturulan veriseti, birliktelik kuralları ve metin madenciliği yöntemleri ve algoritmalarıyla (Apriori, N-Gram, K-Means vb.) uygulanarak analiz edilecektir.

5.5.1. Verilerin weka üzerinde apiori algoritması ile analizi ve sonuçları

Veri kümesinde WEKA yazılımı kullanılmıştır. WEKA, makine öğrenimi için yazılımlar, veri madenciliği çalışmaları için çeşitli algoritmalar ve veri ön işleme gibi (sınıflandırma, regresyon, kümeleme, birliktelik kuralları, sonuç görselleştirme) bir çok araçları barındıran açık kaynak kodlu bir yazılımdır [59].

---

59. M.Hall, E.Holmes, G.Pfahringe, P.Reutemann & I.E.Witten, *The WEKA data mining software : an update*, , ACM SIGKDD Explorations Newsletter, sayı.11, no.1, 2009.

Oluşturulan veri kümesine WEKA 3.6.6 yazılımı kullanılarak Apiori algoritmasının uygulanması ile bir çok kural oluşturulmuştur [60]. Bu kuralların bazıları şöyledir ;

1. Rating=1 hunger=Yes 80 ==> Sex=Kadin 75 conf:(0.94)
2. AgeScaleStatu=Adult hunger=Yes 79 ==> Sex=Kadin 72 conf:(0.91)
3. AgeScaleStatu=Adult DosageMG=500 80 ==> Sex=Kadin 72 conf:(0.9)
4. Rating=1 nausea=Yes 107 ==> Sex=Kadin 96 conf:(0.9)
5. Rating=1 DosageMG=500 86 ==> Sex=Kadin 76 conf:(0.88)
6. AgeScaleStatu=Adult nausea=Yes 103 ==> Sex=Kadin 91 conf:(0.88)
7. hunger=Yes 138 ==> Sex=Kadin 121 conf:(0.88)
8. Rating=1 stomach_cramps=Yes 108 ==> Sex=Kadin 94 conf:(0.87)
9. nausea=Yes 187 ==> Sex=Kadin 162 conf:(0.87)
10. AgeScaleStatu=Adult DosageXTimes=4 95 ==> Sex=Kadin 82 conf:(0.86)
11. AgeScaleStatu=Adult stomach_cramps=Yes 92 ==> Sex=Kadin 79 conf:(0.86)
12. Rating=1 DurationValueDays=2 110 ==> Sex=Kadin 93 conf:(0.85)
13. Rating=1 DosageXTimes=4 95 ==> Sex=Kadin 80 conf:(0.84)
14. vomiting=Yes 93 ==> Sex=Kadin 78 conf:(0.84)
15. DosageMG=500 154 ==> Sex=Kadin 129 conf:(0.84)
16. stomach_cramps=Yes 166 ==> Sex=Kadin 138 conf:(0.83)
17. Rating=1 AgeScaleStatu=Adult 196 ==> Sex=Kadin 162 conf:(0.83)
18. Rating=1 AgeScaleStatu=Young 103 ==> Sex=Kadin 85 conf:(0.83)
19. DosageMG=500 DosageXTimes=4 87 ==> Sex=Kadin 71 conf:(0.82)
20. stomach_pain=Yes 103 ==> Sex=Kadin 84 conf:(0.82)
21. Rating=1 377 ==> Sex=Kadin 306 conf:(0.81)
22. Sex=Kadin DurationValueDays=2 116 ==> Rating=1 93 conf:(0.8)
23. DosageXTimes=4 181 ==> Sex=Kadin 145 conf:(0.8)
24. DurationValueDays=2 145 ==> Sex=Kadin 116 conf:(0.8)
25. AgeScaleStatu=Adult 371 ==> Sex=Kadin 294 conf:(0.79)
26. DosageMG=250 113 ==> Sex=Kadin 89 conf:(0.79)
27. vomiting=Yes 93 ==> Rating=1 73 conf:(0.78)
28. DurationValueDays=2 145 ==> Rating=1 110 conf:(0.76)
29. Rating=2 100 ==> Sex=Kadin 75 conf:(0.75)
30. diarrhea=Yes 100 ==> Sex=Kadin 75 conf:(0.75)
31. DurationValueDays=7 135 ==> Sex=Kadin 100 conf:(0.74)
32. AgeScaleStatu=Young 189 ==> Sex=Kadin 139 conf:(0.74)
33. Rating=3 117 ==> Sex=Kadin 85 conf:(0.73)
34. AgeScaleStatu=Middle_Aged 122 ==> Sex=Kadin 88 conf:(0.72)
35. DosageMG=250 113 ==> DosageXTimes=4 81 conf:(0.72)
36. Sex=Kadin stomach_cramps=Yes 138 ==> Rating=1 94 conf:(0.68)
37. stomach_cramps=Yes 166 ==> Rating=1 108 conf:(0.65)
38. DurationValueDays=2 145 ==> Rating=1 Sex=Kadin 93 conf:(0.64)
39. Sex=Kadin hunger=Yes 121 ==> Rating=1 75 conf:(0.62)
40. Sex=Kadin AgeScaleStatu=Young 139 ==> Rating=1 85 conf:(0.61)

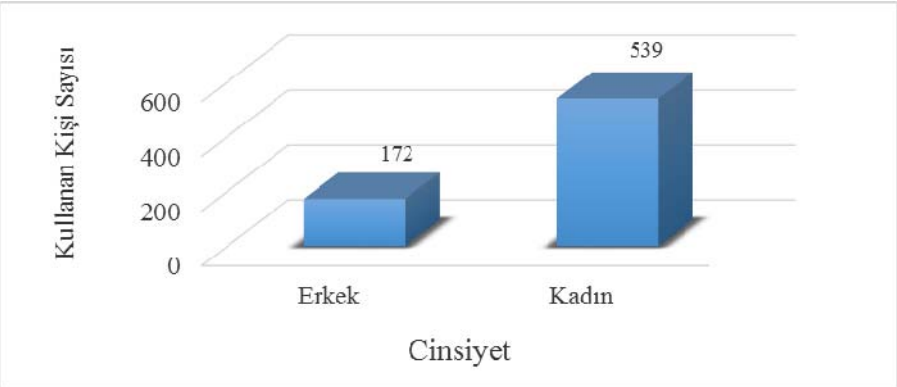
60. WEKA, *Weka 3: Data Mining Software in Java* » http://www.cs.waikato.ac.nz/ml/weka, 27 Aralık 2013.

Apriori algoritmasının sonuçlarına göre en çok birlikte görülülen yan etkiler, stomach cramps (mide krampları), stomach pain (karın ağrısı), hunger (açlık), nausea (bulantı), vomiting (kusma), diarrhea (ishal) dır. İlaç dozaj kullanım miktarları ise ağırlıklı 500 mg ve 250 mg olarak görülmektedir. İlaç kullanım sonrası verilen puanlamada (Rating=1) kullananlar %47,4 oran ile olumsuz puanlama yapmışlardır ayrıca Young (Genç), Adult (Yetişkin) ve kadın kullanıcıların toplam ilaç kullananlara oranı %71,7 olduğu, stomach cramps (mide krampları) ve rating puanlaması "1" olan kayıtlar birlikte görülmüştür. Ayrıca genç kadınlar ilaç puanlamasına hep "1" puan vermişlerdir.

Veri kümesinden elde edilen bilgiler içerisinde Erythromycin ilacının yaş ve cinsiyete göre kullanım sayıları Tablo 6.16. ve Tablo 6.17.' de, grafiksel dağılımları ise Şekil 6.6. ve Şekil 6.7.' deki gibi gösterilmiştir, ayrıca puanlama bilgisine göre kullanıcı sayıları Tablo 6.18.'de, grafiksel görünümü ise Şekil 6.8. üzerinde gösterilmiştir.

Tablo 6.16. Cinsiyete Göre Kullanım Sayısı

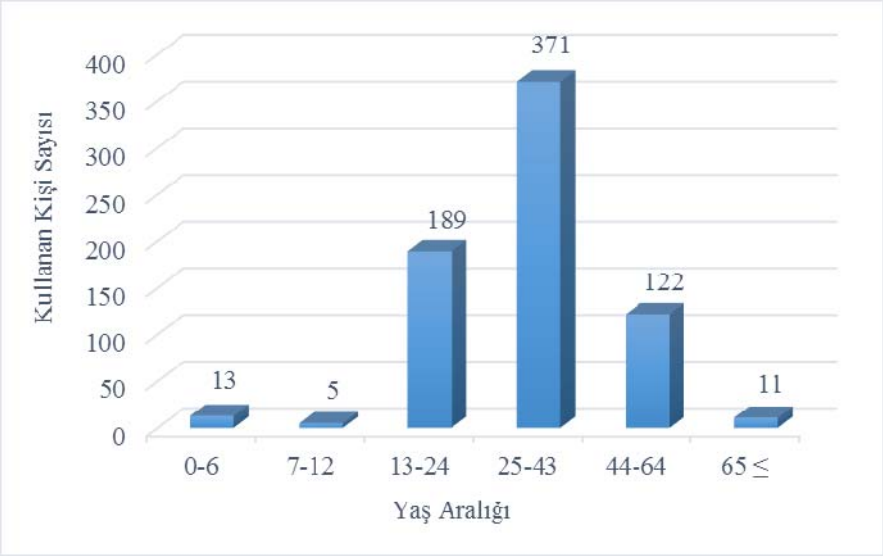
Cinsiyet	**Sayı**
Erkek	172
Kadın	539



Şekil 6.6. Cinsiyete Göre Kullanım Sayısı

Tablo 6.17. Yaş Aralığına Göre Kullanım Sayısı

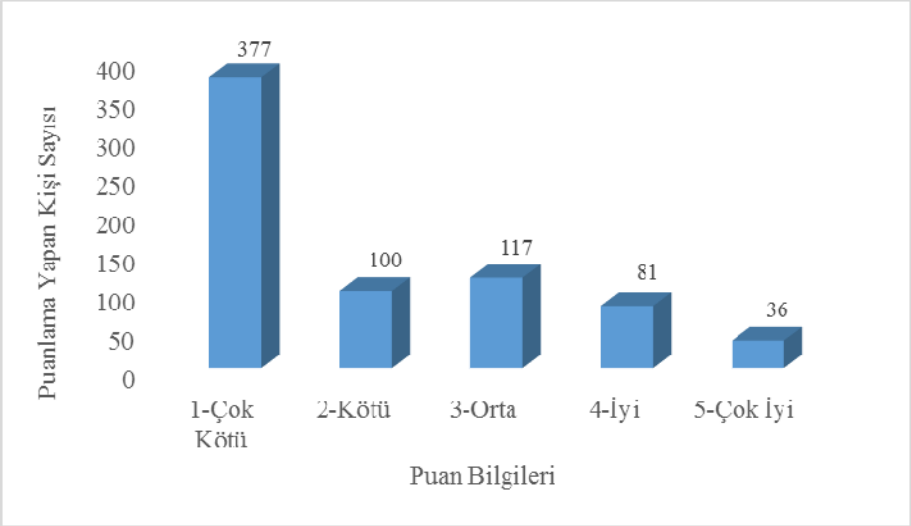
Yaş Aralığı	Sayı
0-6	13
7-12	5
13-24	189
25-43	371
44-64	122
65 ≤	11



Şekil 6.7. Yaş Aralığına Göre Kullanım Sayısı

Tablo 6.18. Puanlama(Rating) Bilgisine Göre Kullanıcı Sayısı

Puan Bilgisi	Puan Bilgisi
1-Kötü	377
2-Zayıf	100
3-Orta	117
4-İyi	81
5-Pekiyi	36



Şekil 6.8. Puanlama Bilgisine Göre Kullanıcı Sayısı

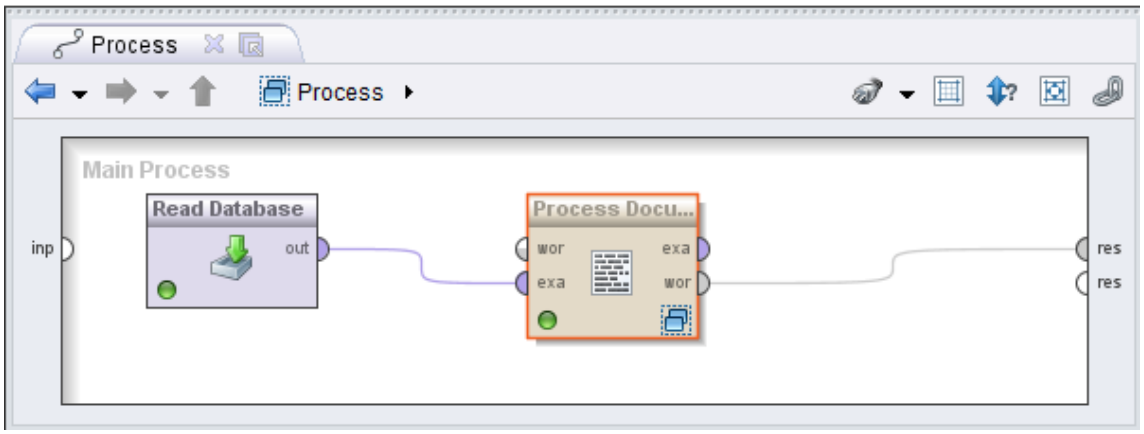
5.5.2. Verilerin Rapidminer üzerinde n-gram algoritması ile analizi ve sonuçları

İlk zamanlarda YALE (Yet Another Learning Environment – Yeni bir öğrenme ortamı) adını taşıyan, 2001 yılında Dortmund Teknik Üniversitesi, Yapay Zeka Bölümünde görev yapan Ralf Klinkenberg, Ingo Mierswa ve Simon Fischer tarafından geliştirilmeye başlanmıştır. 2006 yılında şirket kurulum çalışmalarına başlayan ve aynı yıl Ingo Mierswa ve Ralf Klinkenberg tarafından kurulan şirket, 2007 yılında yazılımın adını Rapidminer olarak değiştirerek, Rapid-I GmbH şirketini kurmuşlardır.

Rapidminer programı, makine öğrenmesi, veri madenciliği, metin madenciliği, tahmine dayalı analitik ve iş analitiği için entegre bir ortam sağlamaktadır, aynı adı taşıyan Rapid-I GmbH tarafından geliştirilen bir yazılım platformudur. Program, endüstriyel uygulamaların yanı sıra, araştırma, eğitim, öğretim, hızlı prototipleme ve uygulama geliştirme için kullanılan ve sonuçları görselleştirme, doğrulama ve optimizasyonu da dahil olmak üzere veri madenciliği sürecinin tüm adımları desteklemektedir.

RapidMiner bir iş üzerinde geliştirilen çekirdek ve yazılımın önceki sürümleri için kaynak modeli oluşturabilen OSI sertifikalı açık kaynak lisansına sahiptir [61].

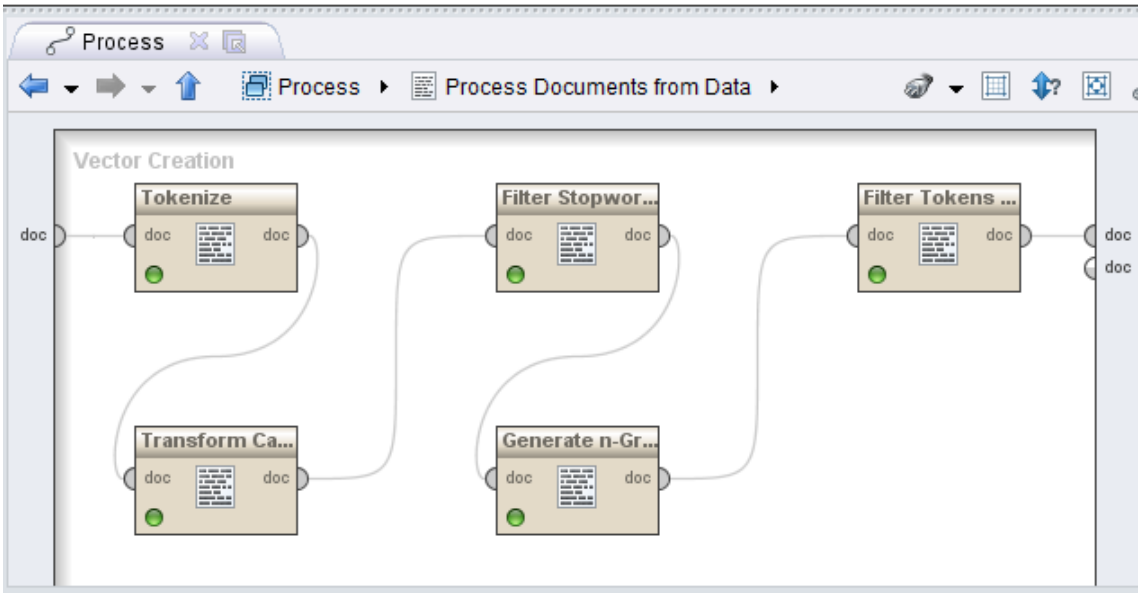
Oluşturulan veriseti, Rapidminer programı üzerinde n-gram algoritması ve çeşitli metin ayrıştırma (tokenization - dizgeciklere ayırma), dönüşüm (transform case), temizleme araçları kullanılarak analiz edilmiştir. Şekil 6.9.' da kurulan modelin yapısı bulunmaktadır.



Şekil 6.9. Rapidminer Modelinin Yapısı

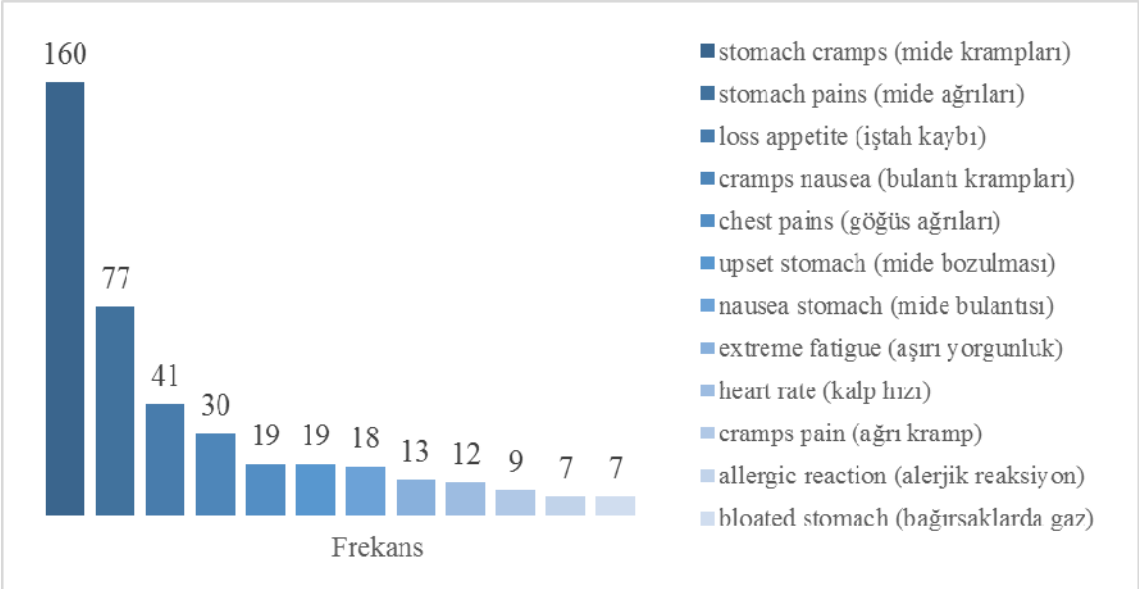
"Read Database" bileşeni ile MS SQL Server 2012 veritabanında bulunan ve Tablo 6.1.' de ki gibi oluşturulan "SideEffects-Yan Etkiler" niteliğinin okuma işlemi gerçekleştirilmektedir. Şekil 6.10.' da detayları bulunan bir sonraki bileşen olan "Process Documents From Data" ile sırası ile tokenize, transform case, filter stopwords, generate n-grams ve filter tokens by length (2-999) algoritma ve metin işleme araçları kullanılarak analiz edilmiştir.

61. M.Hofmann, F.Klinkenberg, *RapidMiner: Data Mining Use Cases and Business Analytics Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)*, CRC Press, 2013.



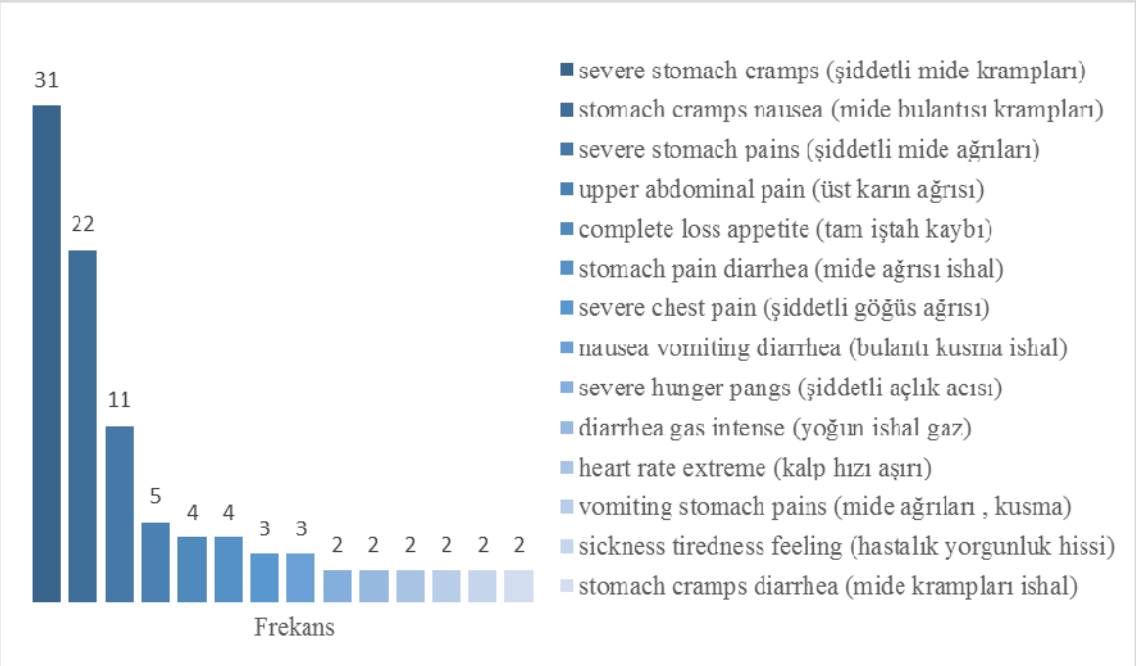
Şekil 6.10. Rapidminer Modelinin Detay Yapısı

2–Gram algoritması sonuçları, Ek A. – Ek C.' de listelenmiştir. Listelenen sonuçlarda aynı anlama gelebilecek frekansı en yüksek kayıtlar ele alınarak grafiksel gösterimi Şekil 6.11.' de sunulmuştur.



Şekil 6.11. 2-Gram Algoritması Sonuçları

3–Gram algoritması sonuçları ise Ek D. – Ek I.' da listelenmiştir. Listelenen sonuçlarda aynı anlama gelebilecek frekansı en yüksek kayıtlar ele alınarak grafiksel gösterimi Şekil 6.12.' de sunulmuştur.



Şekil 6.12. 3-Gram Algoritması Sonuçları

5.5.3. Verilerin weka üzerinde simple k-means algoritması ile analizi ve sonuçları

Olusturulan veri seti ile en çok tekrar eden 15 adet yan etki isimleri, Weka programı üzerinde "simple k-means" algoritması ile analiz edilmiş ve sonuçları Tablo 6.16.' da gösterilmiştir.

Çıkan sonuçları yorumlamak gerekirse, k-means algoritması oluşturulan veri setindeki 711 adet kaydı kullanarak verileri 6 gruba (cluster - küme) ayırdı. Birbirinden farklı 6 gruptan;

Tablo 6.19. K-Means Kümeleme Algoritması Sonuçları

Cluster#							
Attribute	Full Data	0	1	2	3	4	5
	(711)	(178)	(178)	(136)	(130)	(33)	(56)
Rating	1	1	1	1	1	1	3
Sex	Kadın	Kadın	Erkek	Kadın	Kadın	Kadın	Kadın
AgeScaleStatu	Adult	Middle_Aged	Adult	Adult	Young	Adult	Adult
DurationValueDays	2	3	2	7	7	1	2
DosageMG	500	500	500	500	500	500	500
DosageXTimes	4	4	4	4	4	4	4
stomach_pain	Yes	Yes	Yes	Yes	Yes	Yes	Yes
nausea	Yes	Yes	Yes	Yes	Yes	Yes	Yes
hunger	Yes	Yes	Yes	Yes	Yes	Yes	Yes
stomach_cramps	Yes	Yes	Yes	Yes	Yes	Yes	Yes
diarrhea	Yes	Yes	Yes	Yes	Yes	Yes	Yes
vomiting	Yes	Yes	Yes	Yes	Yes	Yes	Yes
headache	Yes	Yes	Yes	Yes	Yes	Yes	Yes
dizziness	Yes	Yes	Yes	Yes	Yes	Yes	Yes
loss_of_appetite	Yes	Yes	Yes	Yes	Yes	Yes	Yes
infection	Yes	Yes	Yes	Yes	Yes	Yes	Yes
pain_in_chest	Yes	Yes	Yes	Yes	Yes	Yes	Yes
suffer	Yes	Yes	Yes	Yes	Yes	Yes	Yes

- Birincisine ait kayıtların genel özelliklerine baktığımızda, 178 adet "Kadın" cinsiyetinde olan ve "Middele_Aged (44-64)" yaş aralığındaki kişilerin ilacı 3 gün boyunca günde 4 kez ve 500 mg olarak kullandığı ve "1" puan verdiği,
- İkincisine ait kayıtların genel özelliklerine baktığımızda, 178 adet "Erkek" cinsiyetinde olan ve "Adult (25-43)" yaş aralığındaki kişilerin ilacı 2 gün boyunca günde 4 kez ve 500 mg olarak kullandığı ve "1" puan verdiği,
- Üçüncüsüne ait kayıtların genel özelliklerine baktığımızda, 136 adet "Kadın" cinsiyetinde olan ve "Adult (25-43)" yaş aralığındaki kişilerin ilacı 7 gün boyunca günde 4 kez ve 500 mg olarak kullandığı ve "1" puan verdiği,

- Dördüncüsüne ait kayıtların genel özelliklerine baktığımızda, 130 adet "Kadın" cinsiyetinde olan ve "Young (13-24)" yaş aralığındaki kişilerin ilacı 7 gün boyunca günde 4 kez ve 500 mg olarak kullandığı ve "1" puan verdiği,
- Beşincisine ait kayıtların genel özelliklerine baktığımızda, 33 adet "Kadın" cinsiyetinde olan ve "Adult (25-43)" yaş aralığındaki kişilerin ilacı 1 gün boyunca günde 4 kez ve 500 mg olarak kullandığı ve "1" puan verdiği,
- Altıncısına ait kayıtların genel özelliklerine baktığımızda, 56 adet "Kadın" cinsiyetinde olan ve "Adult (25-43)" yaş aralığındaki kişilerin ilacı 2 gün boyunca günde 4 kez ve 500 mg olarak kullandığı ve "3" puan verdiği, görülmektedir.

Ayrıca tüm gruplardaki bu kullanımlarla beraber stomach_pain (mide ağrısı), nausea (bulantı), hunger (açlık), stomach_cramps (mide krampları), diarrhea (ishal), vomiting (kusma), headache (baş ağrısı), dizziness (baş dönmesi), loss_of_appetite (iştah kaybı), infection (enfeksiyon), pain_in_chest (göğüste ağrı), suffer (acı çekmek) yan etkileri birlikte görülmektedir.

VI. SONUÇ

Bu çalışma kapsamında, “www.askapatient.com” adresinde bulunan 06.07.2001 ile 01.05.2013 tarihleri arasında yazılmış Erythromycin ilacına ait yapılan yorumlar, (toplam 711 adet kayıt) C#.Net dilinde yazılan web sayfası okuyucu (html okuyucu) program kullanılarak elde edilen veriler Microsoft Sql Server 2012 veritabanı sisteminde bir araya getirilerek veri kümesi oluşturulmuş ve çok geniş bir çalışma alanı olan veri madenciliğinin birliktelik kuralları ve kümeleme algoritmaları üzerinde durulmuştur.

Çalışmada, birliktelik kurallarından Apriori ve N-Gram algoritmaları, kümeleme algoritmalarından ise Simple K-Means algoritması kullanılmıştır.

Apriori algoritması ile ilacın hangi cinsiyet ve yaş aralıklarında, hangi kullanım nedenleri ile hangi yan etkilerin, hangi kullanım sürelerinde, hangi dozaj kullanımlarında birlikte görüldüğü tespit edilmeye çalışılmış, “Kadın” cinsiyetinde, “Adult (25-43)” yaş aralığında, ”500 mg” dozaj kullanımlarında, hunger (açlık) ve stomach_cramps (mide krampları) yan etkileri gibi birlikte görülen sonuçlar elde edilmiştir.

N-Gram algoritması ile yan etkiler niteliği analiz edilmeye çalışılmış ve “n” değişkenin değeri “2” iken aynı anlama gelen bir çok kelimenin bulunma olasılıkları yüksek olarak görünmektedir. Analiz sonucunda, ”yorgunluk”, “kramp” ve “ağrı” gibi kelimeler çıkarken “n” değeri arttıkça “mide bulantısı”, “ağız kuruluğu”, “nefes darlığı” ve hatta “ishal ve gaz

yoğun mide krampları" gibi aynı anda görülen daha kesin ve anlamlı sonuçlar elde edilmiştir.

Bu çalışmanın zayıf yönleri, tıp terimlerinin fazlalığı ve aynı anlama gelen yada yazılımı aynı olupta farklı anlamlar içeren kelimelerin bulunması, çalışma sonucunda çıkan sonuçların sağlıklı bir şekilde yorumlanmasını zorlaştırmaktadır. İlacın üretiminde "30 mg" gibi bir ürün olmamasına rağmen veri setinin içerisinde bulunmasından dolayı bir takım gürültülü veriler bulunmaktadır. İlaca en çok kadınların yorum yapmış olması nedeniyle veri setinde düzensizlik yaratmış olduğundan dengeli bir dağılım bulunmamaktadır. Ayrıca analiz edilen verilerin, internet üzerinde bulunan web sitesinden elde edilmiş olması ve günümüz teknolojisi ile bu verilerin gerçek kişiler ve gerçek etkileri tamamen yansıttığı onaylanamadığı için verilerin doğruluğundan şüphe edilmesi çalışmanın eksik yönlerinden sayılabilir.

Bu çalışma kapsamında gelecekte, kurulan modelin en başından en sonuna kadar (veri toplama, veri işleme, dönüştürme ve indirgeme, algoritmaların işletilmesi ve sonuçlarının görselleştirilmesi) tek bir yapı üzerinde toparlanıp tamamen otomatik bir hale getirilerek diğer tüm ilaçlara uygulanması ile sürekli olarak yenilenen ilaç yan etki sonuçlarını görebileceğimiz bir çalışma başlatılabilir. Bunlara ek olarak günümüz internet teknolojilerinin gelişmesi ile orantılı olarak İnternette bulunan verilerin doğruluğu ve onaylanmış olması sağlanabilir. Örneğin bir web sitesinde yorum yapacak kişilerin Devlet tarafından sağlanan kişiye özel çipli kimlik kartları ile yorum yapabilmeleri, yapılabilecek sahte yorumların önüne geçecektir.

Sonuç olarak çalışmadan çıkan sonuçlar, ilaç üretimi yapan şirketlere büyük bir katkı sağlayacağı, doktorlar ve ilacı kullanmaya başlayacak hastalar için önemli bir bilgi kaynağı olması faydaları olarak görülebilir.

KAYNAKÇA

1. Türk Eczacıları Birliği "İlaç Kullanımları Hakında Bilgiler," *http://www.e-kutuphane.teb.org.tr/pdf/raporlar/sag_ilac09/8.pdf,* 27 Aralık 2013.

2. Ö.Yalçın, *Veri Madenciliği Yöntemleri : Veri Madenciliğine Giriş* Papatya, syf.37-38, 2012.

3. P. Adrians, D. Zantinge, *Data Mining*, Addison, 1997.

4. C. Clifton, B.Thuraisingham, "*Emerging Standards for Data Mining*," Computer Standards & Interfaces, syf.187-193, 2001.

5. W.J.Flawley, G.Piatetsky-Shapiro, C.J.Matheus *Knowledge Discovery in Databases : An Overview*, AI Magazine, syf.57-70, 1992.

6. M.J.A.Berry, G.S.Linoff, *Mastering Data Mining : The Art and Science of CRM,* The Art and Science of CRM, syf.7, 2000.

7. J.Han, M.Kamber, a.g.e., syf.7.

8. R. Swift *Accelerating Customer Relationship*, Prentice Hall PTR, syf.93, 2001.

9. M.Albayrak, *EEG Sinyallerindeki Epileptiform Aktivitenin Veri Madenciliği Süreci ile Tespiti*, Doktora Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, syf.74, 2008.

10. P.F.Brown, D.V.J.Pietra, P.V.DeSouza *Class-based n-gram models of Natural Language*, Computational Linguistics, vol.18, syf.467–479, 1992.

11. O.Engin, A.Fığlalı, *Akıs Tipi Çizelgeleme Problemlerinin Genetik Algoritma Yardımı ile Çözümünde Uygun Çaprazlama Operatörünün Belirlenmesi*, Doğus Üniversitesi Dergisi, s.6, syf.27-35, 2002.

12. Y.S.Türkan, E.Manisalı, M.F.Çelikkol, *Evaluation of critical success factors effect on six sigma project success in Turkey's manufacturing sector*, Journal of Engineering and Natural Sciences, syf.105-117, 2009.

13. Y.Çabuk, S.Karayılmazlar, *Altı Sigma Yaklaşımı*, Bartın Orman Fakültesi Dergisi, s.94, 17 Aralık 2010.

14. A.Öztürk, *Kalite Yönetimi ve Planlaması*, Ekin Yayınevi, Bursa, ISBN 978-9944-141-79-6, 2009.

15. P.Pande, L.Holpp, *What is six sigma?*, McGraw-Hill, New York, ISBN 0-07-128185-6, 2002.

16. A.S.Koyuncugil, *Veri Madenciliği Ders Notları : Yönetim Bilişim Sistemleri*, http://www.koyuncugil.org/files/ders/bolum6.pdf , Ankara , 27 Aralık 2010.

17. S.Walczak, *Review of The Handbook of Data Mining. Organizational Research Methods* ,s.7, syf.119-121, 2004.

18. U.T.Ş.Gürsoy, *Uygulamalı Veri Madenciliği: Sektörel Analizler*, Ankara, s.4, 2012.

19. Y.Özkan, *Veri Madenciliğine Giriş : Veri Madenciliği Yöntemleri*, İstanbul, syf.40-41, 2008.

20. R.J.Roiger, M.W.Geatz, *Data Mining A Tutorial : Based Primer* ,Wesley, USA, syf.350, 2003.

21. C. Bounsaythip, R.Rinta, *Overview of Data Mining for Customer Behavıour Modeling : VTT Information Technology Research Report* , TTEI, USA, 2001.

22. H.Akpınar, *Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği* , İ.Ü.İşletme Fakültesi Dergisi, İstanbul, s.1, 2000.

23. F.Gürbüz, L.Özbakır, H.Yapıcı, *Türkiye'de Bir Havayolu İşletmesine Ait Parça Söküm Raporlarına İlişkin Veri Madencilği Uygulaması*, Gazi Üniv. Müh. Mim. Fak. Der. Cilt 24, No 1, 73-78, 2009.

24. W.DuMouchel, *Knowledge Discovery and Data Mining,* Proceeding of the Fifth ACM Conference, USA, syf 6-15, 1999.

25. Ö.Yalçın, *Veri Madenciliği Yöntemleri : Veri Madenciliğine Giriş* Papatya, syf.40-45, 2012.

26. H.Tatlıdil, *Uygulamalı Çok Değişkenli İstatistiksel Analiz*, Ziraat Matbaacılık, Ankara, syf.330, 2002.

27. O. İnan, *Veri Madenciliği : Yüksek Lisans Tezi*, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, 2003.

28. M.Albayrak, *EEG Sinyallerindeki Epileptiform Aktivitenin Veri Madenciliği Süreci ile Tespiti : Doktora Tezi*, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, 2008.

29. Ö.Akgöbek, F.Çakır, *Veri Madenciliğinde Bir Uzman Sistem Tasarımı*, Akademik Bilişim 09, 11-13 Şubat Harran Üniversitesi, Şanlıurfa, syf.801-806, 2009.

30. A.S.Koyuncugil, N.Özgülbaş, *Veri madenciliğinin Tıp ve Sağlık Alanında Kullanımı*, Bilişim Teknolojileri Dergisi, sayı.2, syf.2, 2009.

31. M. Makinacı, C. Güneşer, *Göğüs Kanseri Verilerinin Sınıflandırılması*, Elektrik-Elektronik-Bilgisayar Mühendisliği 12. Ulusal Kongresi ve Fuarı Bildirileri Kitabı, s.1, 2007.

32. P.Yıldırım, M.Uludağ, A.Görür, *Hastane Bilgi Sistemlerinde Veri Madenciliği*, Çanakkale Üniversitesi Akademik Bilişim Dergisi, s.1, 2008.

33. C.Bhatt, *Mining the Medical Literature*, http://ai.stanford.edu/~serafim/CS374_2004/Lecture%20Notes/lecture6.pdf, 27 Aralık 2013.

34. *Unstructured Data*, http://en.wikipedia.org/wiki/Unstructured_data, 27 Aralık 2013.

35. R.Hwa, *An Overview of Text Mining*, http://www.umiacs.umd.edu/~hwa/textmining.ppt, 27 Aralık 2013.

36. B.Oğuz, U.Bilge, O.Saka, *Tıpta Metin Madenciliği*, Biyoistatistik ve Tıp Bilişimi AD, Akdeniz Üniversitesi, Antalya, syf.1,2.

37. M.Sharp, *Text Mining*, http://www.scils.rutgers.edu/~msharp/text_mining.htm, 27 Aralık 2013.

38. Cerrito, *Inside text mining: text mining provides a powerful diagnosis of hospital quality rankings - Data Warehousing/Mining*, http://findarticles.com/p/articles/mi_m0DUD/is_3_25/ai_114167705/pg_2 , 2006.

39. M.Konchady, *Text Mining Application Programming. 1st ed*. ,Charles River Media, 2006.

40. H.Shatkay, S.Edwards, W.Wilbur, M.Boguski, *Genes, themes and microarrays, using information retrieval for large-scale gene analysis*, In Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, syf.317-328, 2000.

41. D.B.Johnson, R.K.Taira, A.F.Cardenas, D.R.Aberle, *Extracting Information from Free Text Radiology Reports*, Int J Digit Libr 997, sayı.1,syf.297-308.

42. A.M.Cohen, W.R.Hersh, *A Survey of Current Work in Biomedical Text Mining, Briefings in Bioinformatics*, sayı.6, syf.57–71, 2005.

43. G.Schadow, C.J.Mcdonald, *Extracting Structured Information from Free Text Pathology Reports*, AMIA Annu Symp Proc., syf.584, 2003.

44. T.Bekhuis, *Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy*, Biomedical Digital Libraries, 3:2, 2006.

45. N. Gürsakal, *Sözcük ve Sayı*, www.20.uludag.edu.tr/~gursakal/down/say.ppt, 05 Kasım 2006.

46. A.M.Cohen, W.R.Hersh, *A Survey of Current Work in Biomedical Text Mining, Briefings in Bioinformatics*, syf.38.

47. Yan Etkiler Araştırma Grubu, *EMBL : Avrupa Moleküler Biyoloji Laboratuvarı Sider2*, http://sideeffects.embl.de/, 27 Aralık 2013.

48. NCBI Pubmed, *National Center for Biolotechnology Information Advances Science and Health*, http://www.ncbi.nlm.nih.gov/, 27 Aralık 2013.

49. E.Şeker, *Bilgisayar Kavramları*, http://www.bilgisayarkavramlari.com/2011/09/07/apriori-algoritmasi/, 27 Aralık 2013

50. X.Huang, F.Peng, A.An, D.Shuurmans, N. Cercone, *Applying Machine Learning to Text Segmentation for Information Retrieval*, Information Retrieval sayı.6, syf.333–362, 2003.

51. P.Garret, *Making-Breaking Codes,* ISBN 0-13-030369-0, Prentice Hall, 2001.

52. D.Jurafsky, J.H.Martin, *Speech and Language Processing*, Prentice Hall, 2000.

53. http://en.wikipedia.org/wiki/K-means_clustering, 27 Aralık 2013.

54. J.Han, M.Kamber, *Data Mining Concepts and Techniques*, Morgan Kauffmann Publishers Inc., 2001.

55. P.Berkhin, *Survey of Clustering Data Mining Techniques*, San Jose,California, USA, Accrue Software Inc., 2002.

56. M.Işık, A.Y.Çamurcu, *K-means, K-medoids ve Bulanık C-means Algoritmalarının Uygulamalı Olarak Performanslarının Tespiti*, 2007.

57. E.Dinçer, *Veri Madenciliğinde K-means Algoritması ve Tıp Alanında Uygulanması*, s.24-64, 2006.

58. Y.Yünel, *K-means Kümeleme Algoritmasının Genetik Algoritma Kullanılarak Geliştirilmesi*, s.1,2, 2010.

59. M.Hall, E.Holmes, G.Pfahringe, P.Reutemann & I.E.Witten, *The WEKA data mining software : an update*, , ACM SIGKDD Explorations Newsletter, sayı.11, no.1, 2009.

60. WEKA, *Weka 3: Data Mining Software in Java* » http://www.cs.waikato.ac.nz/ml/weka, 27 Aralık 2013.

61. M.Hofmann, F.Klinkenberg, *RapidMiner: Data Mining Use Cases and Business Analytics Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)*, CRC Press, 2013.

EKLER

EK A

YAN ETKİLER NİTELİĞİ 2-GRAM ALGORİTMASI SONUÇLARI (1-27) ARASINDAKİ KAYITLAR

Sıra No	2-Gram	2-Gram (Türkçe)	Frekans
1	stomach pains	mide ağrıları	77
2	stomach pain	mide ağrısı	58
3	abdominal pain	karın ağrısı	34
4	pains stomach	ağrıları mide	15
5	severe abdominal	şiddetli karın	15
6	pain stomach	ağrı mide	11
7	pain nausea	ağrı bulantı	10
8	stomach ache	mide ağrısı	10
9	painful stomach	ağrılı mide	9
10	sick stomach	hasta mide	5
11	stomach cramps	mide krampları	160
12	stomach cramping	mide krampları	18
13	abdominal cramps	karın krampları	10
14	stomach upset	mide rahatsızlığı	7
15	abdominal pains	karın bölgesinde ağrı	6
16	horrible stomach	korkunç mide	6
17	stomach discomfort	rahatsızlık mide	5
18	loss appetite	iştah kaybı	41
19	hunger pains	açlık ağrıları	31
20	hunger pangs	açlık acısı	21
21	feeling hungry	aç duygu	15
22	extreme hunger	aşırı açlık	13
23	hungry time	aç zamanı	11
24	severe hunger	şiddetli açlık	10
25	empty stomach	boş mide	9
26	lack appetite	eksikliği iştah	9
27	constant hunger	sürekli açlık	8

EK B

YAN ETKİLER NİTELİĞİ 2-GRAM ALGORİTMASI SONUÇLARI (28-58) ARASINDAKİ KAYITLAR

Sıra No	2-Gram	2-Gram (Türkçe)	Frekans
28	nausea stomach	mide bulantısı	18
29	extreme nausea	aşırı bulantı	14
30	severe nausea	şiddetli bulantı	11
31	nausea extreme	bulantı aşırı	9
32	pains nausea	ağrıları bulantı	7
33	diarrhea nausea	ishal bulantı	6
34	nausea feeling	bulantı hissi	6
35	cramping nausea	kramp bulantı	5
36	feeling nausea	duygu bulantı	5
37	cramps nausea	bulantı krampları	30
38	nausea vomiting	bulantı kusma	26
39	vomiting diarrhea	ishal kusma	6
40	nausea vomitting	bulantı -kusma	5
41	vomiting severe	şiddetli kusma	5
42	vomiting stomach	mide kusma	5
43	chest pains	göğüs ağrıları	19
44	chest pain	göğüs ağrısı	18
45	shortness breath	kısalık nefes	8
46	chest infection	göğüs enfeksiyonu	6
47	heart rate	kalp hızı	12
48	heart beat	kalp atışı	11
49	heart palpitations	kalp çarpıntısı	11
50	extreme fatigue	aşırı yorgunluk	13
51	light headedness	hafif sersemlik	9
52	extremely tired	son derece yorgun	8
53	cramps pain	ağrı kramp	9
54	cramps feeling	duygu krampları	7
55	pain feeling	ağrı hissi	6
56	pains cramps	ağrıları kramplar	7
57	upset stomach	mide bozulması	19
58	loose stools	gevşek dışkı	10

EK C

YAN ETKİLER NİTELİĞİ 2-GRAM ALGORİTMASI SONUÇLARI (59-62) ARASINDAKİ KAYITLAR

Sıra No	2-Gram	2-Gram (Türkçe)	Frekans
59	allergic penicillin	alerjik penisilin	8
60	allergic reaction	alerjik reaksiyon	7
61	bloated stomach	bağırsaklarda gaz	7
62	severe gas	ciddi gaz	6

EK D

YAN ETKİLER NİTELİĞİ 3-GRAM ALGORİTMASI SONUÇLARI (1-29) ARASINDAKİ KAYITLAR

Sıra No	3-Gram	3-Gram (Türkçe)	Frekans
1	abdominal pains cramps	karın ağrıları krampları	2
2	awful stomach cramps	korkunç mide krampları	2
3	bad stomach cramps	kötü mide krampları	4
4	cramping severe eat	yemek şiddetli kramp	2
5	diarrhea stomach pain	ishal mide ağrısı	2
6	diarrhoea severe stomach	ishal şiddetli mide	2
7	diarrhoea stomach ache	ishal mide ağrısı	2
8	experience stomach cramps	deneyim mide krampları	2
9	extreme stomach cramping	aşırı mide kramp	2
10	extreme stomach cramps	aşırı mide krampları	6
11	farting stomach cramps	mide krampları farting	2
12	mild stomach cramps	hafif mide krampları	3
13	pain stomach cramps	ağrı mide krampları	2
14	pains stomach cramps	ağrıları mide krampları	2
15	sever stomach cramps	mide krampları sever	2
16	severe abdominal cramps	şiddetli karın krampları	3
17	severe abdominal pain	şiddetli karın ağrısı	10
18	severe stomach cramping	şiddetli mide krampları	6
19	severe stomach cramps	şiddetli mide krampları	31
20	sick stomach cramps	hasta mide krampları	2
21	stomach cramps feeling	mide krampları duygu	6
22	stomach cramps loss	mide krampları kaybı	3
23	stomach cramps mild	mide hafif kramp	3
24	stomach cramps severe	mide şiddetli krampları	5
25	stomach cramps sick	mide krampları hasta	3
26	stomach pain cramps	ağrı kramp mide	3
27	terrible stomach cramps	korkunç mide krampları	4
28	wind stomach cramps	rüzgar mide krampları	2
29	abdominal pain nausea	karın ağrısı bulantı	4

EK E

YAN ETKİLER NİTELİĞİ 3-GRAM ALGORİTMASI SONUÇLARI (30-58) ARASINDAKİ KAYITLAR

Sıra No	3-Gram	3-Gram (Türkçe)	Frekans
30	cramping nausea vomiting	bulantı kusma , kramp	3
31	cramps nausea loose	bulantı gevşek krampları	2
32	cramps nausea loss	bulantı kaybı kramp	2
33	cramps nausea particularly	Özellikle bulantı krampları	2
34	cramps nausea vomiting	bulantı kusma krampları	2
35	extreme nausea extreme	aşırı bulantı aşırı	2
36	extreme nausea stomach	Aşırı mide bulantısı mide	3
37	nausea abdominal cramps	bulantı karın krampları	2
38	nausea abdominal pain	bulantı karın ağrısı	2
39	nausea loss appetite	bulantı iştahsızlık	7
40	nausea stomach cramps	bulantı , mide krampları	9
41	nausea stomach pain	bulantı , mide ağrısı	4
42	nausea stomach pains	bulantı , mide ağrıları	3
43	pain cramping nausea	ağrı , kramp , bulantı	2
44	pain cramps nausea	ağrı , bulantı krampları	2
45	severe nausea stomach	şiddetli bulantı mide	4
46	stomach cramping nausea	mide bulantısı , kramp	2
47	stomach cramps nausea	mide bulantısı krampları	22
48	stomach cramps sickness	mide bulantısı krampları	2
49	stomach pain nausea	mide ağrısı bulantı	3
50	stomach pains nausea	mide ağrıları bulantı	5
51	abdominal cramps nausea	karın bulantı krampları	2
52	abdominal pain cramping	karın ağrısı kramp	3
53	bloating stomach cramps	şişkinlik , mide krampları	2
54	eat stomach pains	mide ağrıları yemek	2
55	extreme stomach pain	aşırı mide ağrısı	3
56	extreme stomach upset	aşırı mide rahatsızlığı	2
57	pain upper stomach	ağrı üst mide	3
58	severe stomach ache	Şiddetli mide ağrısı	2

EK F

YAN ETKİLER NİTELİĞİ 3-GRAM ALGORİTMASI SONUÇLARI (59-87) ARASINDAKİ KAYITLAR

Sıra No	3-Gram	3-Gram (Türkçe)	Frekans
59	severe upper abdominal	Şiddetli üst abdominal	3
60	severe upper stomach	şiddetli üst mide	3
61	sharp pains stomach	keskin ağrılar mide	2
62	sickness stomach pains	hastalık mide ağrıları	2
63	stomach ache cramps	ağrı kramp mide	2
64	stomach cramps headache	mide ağrısı krampları	2
65	stomach pain amp	mide ağrısı amp	2
66	stomach pain cramping	mide ağrısı kramp	2
67	stomach pain kept	tuttu mide ağrısı	2
68	stomach pain unbearable	dayanılmaz mide ağrısı	2
69	stomach pains days	mide ağrıları gün	2
70	stomach pains feel	mide ağrıları hissediyorum	2
71	stomach pains feeling	mide ağrıları duygu	2
72	stomach pains stomach	mide ağrıları mide	3
73	terrible stomach pain	korkunç mide ağrısı	3
74	terrible stomach pains	korkunç mide ağrıları	2
75	upper abdominal pain	Üst karın ağrısı	5
76	upper abdominal tightness	Üst karında gerginlik	2
77	upper stomach pain	Üst mide ağrısı	2
78	upper stomach pains	Üst karın ağrıları	2
79	upset stomach backache	mide ağrısı	2
80	pain stomach pain	ağrı mide ağrısı	2
81	sleep stomach pain	uyku mide ağrısı	2
82	bad stomach pains	kötü mide ağrıları	6
83	extremely painful stomach	son derece ağrılı mide	2
84	horrible stomach pains	korkunç mide ağrıları	5
85	painful stomach cramps	ağrılı mide krampları	5
86	pains stomach pains	ağrıları , mide ağrıları	5
87	severe hunger pains	şiddetli açlık ağrıları	5

EK G

YAN ETKİLER NİTELİĞİ 3-GRAM ALGORİTMASI SONUÇLARI (88-116) ARASINDAKİ KAYITLAR

Sıra No	3-Gram	3-Gram (Türkçe)	Frekans
88	severe stomach pain	şiddetli mide ağrısı	8
89	severe stomach pains	şiddetli mide ağrıları	11
90	stomach cramps pain	mide ağrısı krampları	6
91	stomach pains cramps	ağrıları kramp mide	5
92	constant feeling hungry	aç sürekli duygu	2
93	empty feeling stomach	boş duygu mide	3
94	empty stomach feeling	Boş mide duygu	2
95	extremely hungry time	Son derece aç zamanı	2
96	feeling extreme hunger	Aşırı açlık hissi	2
97	feeling hungry time	aç zaman duygu	3
98	hollow hungry feeling	oyuk aç duygu	2
99	hunger pains stomach	açlık ağrıları mide	2
100	hunger pangs nausea	açlık bulantı acısı	2
101	hunger stomach cramps	açlık mide krampları	2
102	hungry feeling cramping	aç duygu kramp	2
103	pains hunger pangs	ağrıları açlık acısı	2
104	severe hunger pangs	şiddetli açlık acısı	2
105	stomach pains hunger	ağrıları açlık mide	2
106	take empty stomach	boş mide	2
107	diarrhea gas intense	yoğun ishal gaz	2
108	gas intense stomach	gaz yoğun mide	2
109	gassy upset stomach	gazlı mide	2
110	painful farting stomach	acı gaz çıkarma mide	2
111	pains gassy upset	ağrıları gazlı üzgün	2
112	stomach pains gas	mide ağrıları gaz	2
113	stomach pains gassy	gazlı mide ağrıları	2
114	tons painful farting	ton acı gaz çıkarma	2
115	chest abdominal pains	göğüs karın ağrıları	2
116	chest difficulty breathing	göğüs solunum güçlüğü	2

EK H

YAN ETKİLER NİTELİĞİ 3-GRAM ALGORİTMASI SONUÇLARI (117-145) ARASINDAKİ KAYITLAR

Sıra No	3-Gram	3-Gram (Türkçe)	Frekans
117	chest pains stomach	göğüs ağrıları mide	2
118	chest pains tiredness	göğüs ağrıları yorgunluk	2
119	feeling sick chest	Hasta göğüs duygu	2
120	severe chest pain	şiddetli göğüs ağrısı	3
121	stomach cramps chest	mide göğüs krampları	2
122	extreme nausea vomiting	aşırı bulantı kusma	2
123	nausea vomiting diarrhea	bulantı kusma ishal	3
124	nausea vomiting hunger	bulantı kusma açlık	2
125	nausea vomiting severe	mide bulantısı şiddetli kusma	3
126	nausea vomiting stomach	bulantı kusma mide	3
127	pain diarrhea nausea	ağrı ishal bulantı	2
128	complete loss appetite	tam iştah kaybı	4
129	cramps loss appetite	kaybı iştah krampları	2
130	fatigue loss appetite	yorgunluk iştah kaybı	2
131	feeling hungry unable	yapamaz aç hissi	2
132	loss appetite nausea	iştah kaybı , bulantı	2
133	loss appetite tiredness	iştah kaybı yorgunluk	2
134	cramping diarrhea nausea	kramp , ishal bulantı	2
135	stomach cramping diarrhea	kramp ishal mide	3
136	stomach pain diarrhea	mide ağrısı ishal	4
137	diarreah tons painful	acı ishal	2
138	stomach cramps vomiting	mide krampları , kusma	3
139	vomiting severe stomach	Şiddetli mide kusma	3
140	heart palpitations shortness	kalp çarpıntısı darlığı	2
141	heart rate extreme	kalp hızı aşırı	2
142	rapid heart beat	hızlı kalp atışı	2
143	heart rate went	kalp hızı gitti	2
144	stomach cramps vomitting	mide krampları kusma	2
145	varying amounts diarrhea	değişen miktarlarda diyare	2

EK I

YAN ETKİLER NİTELİĞİ 3-GRAM ALGORİTMASI SONUÇLARI (146-153) ARASINDAKİ KAYITLAR

Sıra No	3-Gram	3-Gram (Türkçe)	Frekans
146	vomiting stomach pains	mide ağrıları , kusma	2
147	rate extreme fatigue	oranı aşırı yorgunluk	2
148	stomach pain weakness	mide ağrısı halsizlik	2
149	sickness tiredness feeling	hastalık yorgunluk hissi	2
150	stomach cramps day	mide krampları gün	2
151	stomach cramps days	mide krampları gün	2
152	stomach cramps diarrhea	mide krampları ishal	2
153	stomach cramps diarrhoea	mide krampları ishal	2

EK J

HTML VERİ OKUYUCU PROGRAMIN KODLARI

```
private void pageOpener()
      {
         #region ilaç sayfası açılıyor
         bool document_complete = false;
         wbrowser.DocumentCompleted += delegate { document_complete = true; };
         while (!document_complete)
            Application.DoEvents();
         #endregion
      }

private void commentParser(int drugID)
      {
         #region yorumlar Ayrıştırılıyor
         string textResult = "";
foreach (HtmlElement pageElement in
wbrowser.Document.GetElementsByTagName("TABLE"))
            textResult = pageElement.Children[0].InnerHtml;

         DataTable dt = HtmlTableParser.ParseTable(textResult);

         for (int i = 0; i < dt.Rows.Count; i++)
            for (int j = 0; j < dt.Columns.Count; j++)
            {
               dt.Rows[i][j] = dt.Rows[i][j].ToString().Replace(" ",
"").Replace("<B>", "").Replace("</B>", "").
                  Replace("</FONT>", "");
               if (dt.Rows[i][7].ToString().Length >= 10)
               {
                  string date = dt.Rows[i][7].ToString().Substring(0, 10);
                  dt.Rows[i][7] = date.Replace("<", "").Replace("B", "");
               }
            }
         dt.Rows.RemoveAt(1);
         dt.Rows.RemoveAt(0);
         #endregion
         InserTableComments(dt, drugID);
      }
```

```
private void InserTableComments(DataTable dt, int drugID)
      {
         dgw.DataSource = null;
         dgw.DataSource = dt;
         for (int i = 0; i < dt.Rows.Count; i++)
         {
            mDrugComment gc = new mDrugComment();
            gc.DrugID = drugID;
            gc.Rating = fn.Int(dt.Rows[i][0]);
            gc.Reason = fn.Str(dt.Rows[i][1]);
            gc.SideEffectsForX = fn.Str(dt.Rows[i][2]);
            gc.Comment = fn.Str(dt.Rows[i][3]);
            gc.Sex = (fn.Str(dt.Rows[i][4]) == "F");
            gc.Age = fn.Int(dt.Rows[i][5]);
            gc.Duration = fn.Str(dt.Rows[i][6]);
            gc.AddDate = fn.DMY(fn.Str(dt.Rows[i][7]));
            hsDB.mDrugComments.InsertOnSubmit(gc);
            hsDB.SubmitChanges();
         }
      }

private List<HtmlElement> pageNumberSetGenericList()
      {
         var linkList = wbrowser.Document.Links.Cast<HtmlElement>().
                     Where(x => x.TagName == "A" &&
                          x.OuterHtml.Contains("viewrating.asp?drug=") &&
                          fn.Int(x.InnerText) != 0
                         ).ToList();
         return linkList;
      }
```

EK K

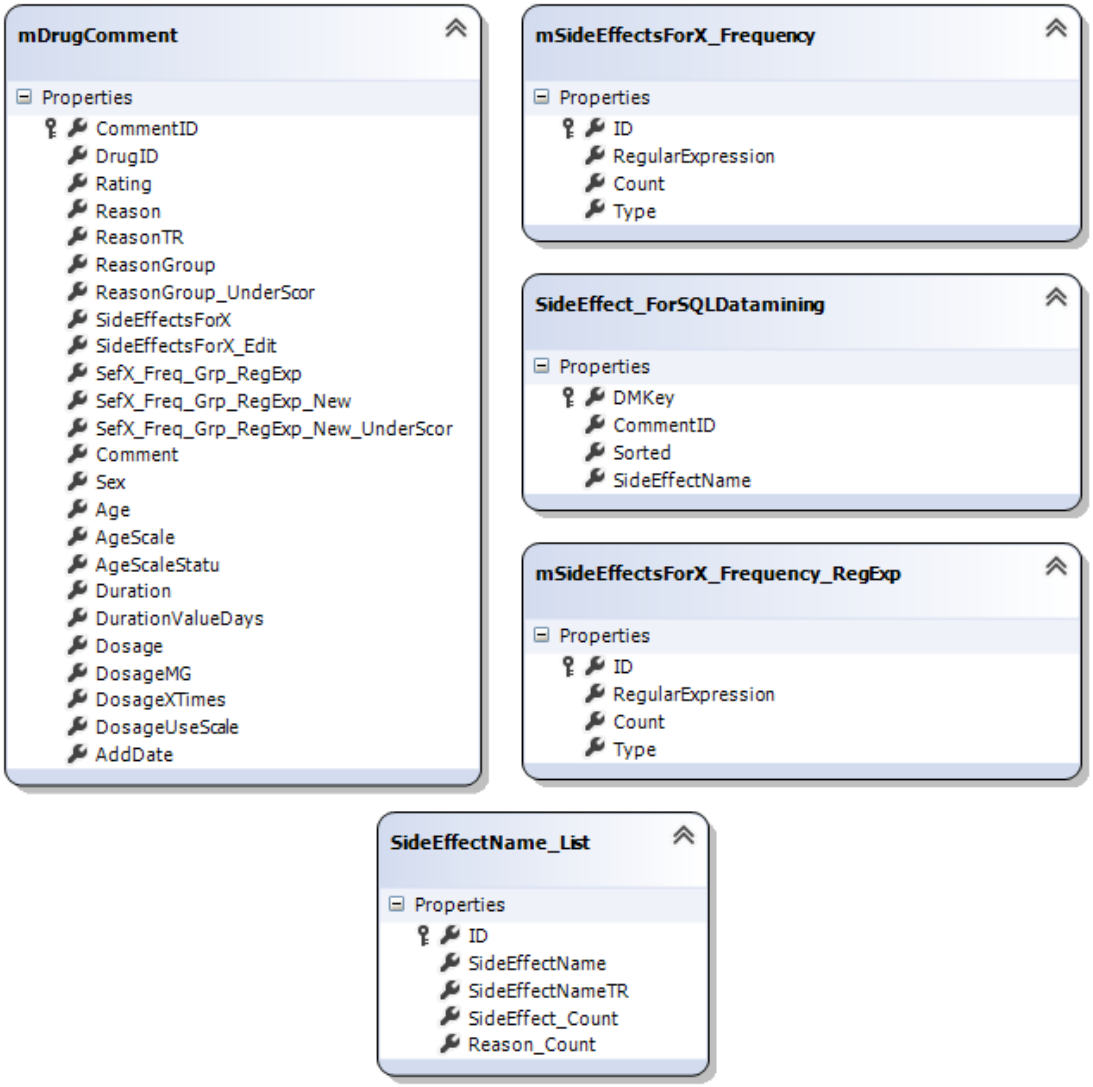
İLAÇ KULLANIM SÜRESİ VE DOZUNUN İKİ AYRI NİTELİK OLARAK AYRILMASI İŞLEMLERİNİN KODLARI

```
private void DurationNormalization_Click(object sender, EventArgs e)
{
   AskaPatientDataContext db = new AskaPatientDataContext();
   var list = db.mDrugComments.ToList();
   foreach (var item in list)
   {
      try
      {
         if (!(item.Duration.Contains("days") || item.Duration.Contains("weeks") ||
item.Duration.Contains("months") || item.Duration.Contains("years")))
            item.Dosage = item.Duration;
         else
         {

            var count = Convert.ToDouble(item.Duration.Split(' ')[0]);
            var scale = item.Duration.Split(' ')[1];
            if (scale == "months")
               count = count * 30;
            else if (scale == "weeks")
               count = count * 7;
            else if (scale == "years")
               count = count * 365;
            item.DurationValueDays = Convert.ToInt32(Math.Floor(count));
         }
         db.SubmitChanges();
      }
      catch (Exception)
      {
   }
   }
}


private void btnDosageParsing_Click(object sender, EventArgs e)
{
   AskaPatientDataContext db = new AskaPatientDataContext();
   var list = db.mDrugComments.Where(p => p.Dosage != null).ToList();
   foreach (var item in list)
   {
      item.Dosage = item.Dosage.ToUpper();
```

```
        if (item.Dosage.Contains("MG"))
        {
            var res = item.Dosage.Split(new string[] { "MG" }, StringSplitOptions.None);
            if (res.Length > 1)
            {
                try
                {
                    var mg = Convert.ToInt32(res[0].Trim());
                    var ress = res[1].Split('X');
                    var xTimes = 0;
                    if (ress.Length > 1)
                        xTimes = Convert.ToInt32(ress[0].Trim());
                    var useScale = "";
                    if (ress.Length > 1)
                        useScale = ress[1].Trim();

                    item.DosageMG = mg;
                    item.DosageXTimes = xTimes;
                    item.DosageUseScale = useScale;
                }
                catch (Exception)
                {
                }
            }
        }
        else
        {
            var xxx = item.Dosage.Split(' ');
      if (xxx.Length == 3 && item.Dosage.Contains("X") && item.Dosage.Length > 5)
            {
                try
                {
                    item.DosageMG = Convert.ToInt32(xxx[0].Trim('G').Trim());
                }
                catch (Exception)
                {
                    item.DosageMG = 0;
                }
                item.DosageXTimes = Convert.ToInt32(xxx[1].Trim('X'));
                item.DosageUseScale = xxx[2].Trim();
            }
        }
        db.SubmitChanges();
    }
}
```

EK L

OLUŞTURULAN VERİ SETİ YAPISI



EK M

JAVA DİLİNDE YAZILAN APRIORI ALGORİTMASI KODLARI

```
package AprioriAlgorithm;
import java.io.BufferedReader;
import java.io.FileInputStream;
import java.io.FileNotFoundException;
import java.io.IOException;
import java.io.InputStreamReader;
import java.util.Collections;
import java.util.Iterator;
import java.util.LinkedHashMap;
import java.util.Map.Entry;
import java.util.Set;
import java.util.Vector;
public class AprioriAlgorithm {
	static int min_sup;
	static Vector<Vector<String>> dataset = new Vector<Vector<String>>();
	static Vector<String> fieldListOneItem = new Vector<String>();
	public static int buyukAsalBul(int sayi){
		while(!asalmi(sayi))
		{
			sayi++;
		}
		return sayi;
	}
	public static boolean asalmi(int sayi){
		boolean asalmi = true;
		int k = 2;
		while(asalmi==true && k<sayi){
			if(sayi%k==0)
				asalmi=false;
			k++;
		}
		return asalmi;
	}
	public static class Itemsets{
		LinkedHashMap<Vector<String>, Integer> Candidates = new
LinkedHashMap<Vector<String>, Integer>();
		LinkedHashMap<Vector<String>, Integer> listOfItemsets = new
LinkedHashMap<Vector<String>, Integer>();
		Vector<Vector<String>> allListOfItemsets = new
Vector<Vector<String>>();
		public Itemsets(int hashTableSize){
```

```
                this.Candidates = new LinkedHashMap<Vector<String>,
Integer>(hashTableSize);
                this.listOfItemsets = new LinkedHashMap<Vector<String>,
Integer>(hashTableSize);
                }
                void generateCandidateOneItemsets()
                {
                        Collections.sort(fieldListOneItem);
                        for (String s : fieldListOneItem) {
                Vector<String> k = new Vector<String>();
                k.add(s);
                                Candidates.put(k, 0);
                        }
                }
                void generateCandidateFromOneItemsets()
                {
                        Candidates.clear();
                        Set<Vector<String>> listOfItemsetsKeys = listOfItemsets.keySet();
                        int x=0;
                        int sz = listOfItemsetsKeys.size();
                        String[] isets = new String[sz];
                        for (Vector<String> k : listOfItemsetsKeys) {
                                Collections.sort(k);
                                for (String s : k) {
                                        isets[x] = s;
                                }
                                x++;
                        }
                        for (int t = 0; t < sz; t++) {
                           for (int i = t+1; i < sz; i++) {
                                Vector<String> cs = new Vector<String>();
                                cs.add(isets[t]);
                                cs.add(isets[i]);
                                Candidates.put(cs, 0);
                                }
                        }
                }
                void generateCandidates(){
                        Candidates.clear();
                        Iterator<Entry<Vector<String>, Integer>> d =
listOfItemsets.entrySet().iterator();
                        boolean dond = true;
                        int rowNumber=0;
                        while(dond && d.hasNext())
                        {
                                int y=0;
```

```
                                Iterator<Entry<Vector<String>, Integer>> i =
listOfItemsets.entrySet().iterator();
                                while(i.hasNext() && y<=rowNumber)
                                {
                                        i.next();
                                        y++;
                                }
                                Entry<Vector<String>, Integer> entryd =
(Entry<Vector<String>, Integer>) d.next();
                                boolean isequal = true;
                                while(isequal && i.hasNext())
                                {
                                        int indexNumber =  0;
                                        Vector<String> newCandidate = new
Vector<String>();
                                        Entry<Vector<String>, Integer> entryi =
(Entry<Vector<String>, Integer>) i.next();
                                        Vector<String> ed = (Vector<String>)
entryd.getKey();
                                        Vector<String> ei = (Vector<String>)
entryi.getKey();
                                        while(dond && isequal && indexNumber<ed.size()-
1)
                                        {

        if(ed.get(indexNumber).hashCode()!=ei.get(indexNumber).hashCode())
                                                {
                                                        isequal = false;
                                                }
                                                newCandidate.add(ed.get(indexNumber));
                                                indexNumber++;
                                        }
                                        if(isequal)
                                        {
                                                newCandidate.add(ed.get(indexNumber));
                                                newCandidate.add(ei.get(indexNumber));
                                                if(controlSubsets(newCandidate))
                                                        Candidates.put(newCandidate,0);
                                        }
                                }
                                rowNumber++;
                        }
                }
                void generateListOfItemsets()
                {
                        Set<Vector<String>> listOfItemsetsKeys = listOfItemsets.keySet();
                        for (Vector<String> key: listOfItemsetsKeys) {
```

```
                allListOfItemsets.add(key);
        }
        listOfItemsets.clear();
        Set<Vector<String>> CandidateKeys = Candidates.keySet();
        for (Vector<String> k : CandidateKeys) {
                if(Candidates.get(k)>=min_sup)
                        listOfItemsets.put(k, Candidates.get(k));
        }
}
boolean controlSubsets(Vector<String> data)
{
        int sz = data.size();
        boolean issubset = true;
        String[][] subsets = new String[sz][sz-1];
        Vector<String> subset = new Vector<String>();
        int x = 0;
        int k = 1;
        int i = 1;
        while(x<(sz-1))
        {
                for (int j = 0; j < sz; j++) {
                        subsets[j][x] = data.get(k-1);
                        if(i%(sz-1)==0)
                                k++;
                        i++;
                }
                x++;
        }
        for (String[] s : subsets) {
                subset.clear();
                for (String c : s) {
                        subset.add(c);
                }
                if(!isSubsetInlistOfItemsets(subset))
                        issubset = false;
        }
        System.out.println(" ");
        return issubset;
}
void increaseCandidateValue(Vector<String> k)
{
        int a = Candidates.get(k)+1;
        Candidates.put(k, a);
}
boolean isSubset(Vector<String> main, Vector<String> sub)
{
        int x=0;
```

```
                boolean subset = true;
                while(subset && x<sub.size()) {
                        if(!main.contains(sub.get(x)))
                                subset=false;
                        x++;
                }
                return subset;
        }
        boolean isSubsetInlistOfItemsets(Vector<String> sub )
        {
                boolean subset = false;
                Iterator<Entry<Vector<String>, Integer>> i =
listOfItemsets.entrySet().iterator();
                while(subset==false && i.hasNext()) {
                        Entry<Vector<String>, Integer> entrya =
(Entry<Vector<String>, Integer>) i.next();
                        if(isSubset((Vector<String>) entrya.getKey(),sub))
                                subset = true;
                }
                return subset;
        }
        void countCandidatesOneItemsets()
        {
                for (Vector<String> vector : dataset) {
                        for (String s : vector) {
                                Vector<String> ts = new Vector<String>();
                                ts.add(s);
                                increaseCandidateValue(ts);
                        }
                }
                printCandidatesItemsets(1);
        }
        void countCandidatesItemsets()
        {
                Set<Vector<String>> CandidateKeys = Candidates.keySet();
                for (Vector<String> data : dataset) {
                        for (Vector<String> cd : CandidateKeys) {
                                if(isSubset(data,cd))
                                        increaseCandidateValue(cd);
                        }
                }
        }
        void printCandidatesItemsets(int sz)
        {
                System.out.println("List of Candidates frequent "+sz+"-Itemsets");
                Set<Vector<String>> CandidateKeys = Candidates.keySet();
                for (Vector<String> items : CandidateKeys) {
```

```
                    for (String s : items) {
                        System.out.print(s+",");
                    }
                    System.out.println(Candidates.get(items) );
                }
                System.out.println(" ");
            }
            void printListOfItemsets(int sz)
            {
                Set<Vector<String>> listOfItemsetsKeys = listOfItemsets.keySet();
                System.out.println("List of frequent "+sz+"-Itemsets");
                for (Vector<String> vector : listOfItemsetsKeys) {
                    Collections.sort(vector);
                    for (String c: vector) {
                        System.out.print(c+",");
                    }
                        System.out.println(listOfItemsets.get(vector));
                }
                System.out.println(" ");
            }
            void getItemsets()
            {
                generateCandidateOneItemsets();
        countCandidatesOneItemsets();
        generateListOfItemsets();
        printListOfItemsets(1);
        int cc=2;
        while(listOfItemsets.size()>0)
        {
            if(cc>2)
                generateCandidates();
            else
                generateCandidateFromOneItemsets();
            countCandidatesItemsets();
            printCandidatesItemsets(cc);
            generateListOfItemsets();
                printListOfItemsets(cc);
                cc++;
        }
        System.out.println("Found " + allListOfItemsets.size() + " of Itemsets");
            }

}
public static void main(String[] args) throws FileNotFoundException {
            InputStreamReader istream = new InputStreamReader(System.in) ;
            BufferedReader bufRead = new BufferedReader(istream) ;
            try
```

```
                {
                        System.out.println("Please Enter minimum support value : ");
                   min_sup = Integer.parseInt( bufRead.readLine()) ;
                   System.out.println("If minimum support is : " + min_sup + ";");
                   System.out.println(" ");

                }
                catch (IOException err)
            {
               System.out.println("Error reading line");
            }
                FileInputStream file_in = new FileInputStream("full_apiori.txt");
                BufferedReader data_in = new BufferedReader(new
InputStreamReader(file_in));
                try
        {
            while(data_in.ready())
            {
               String[] f;
               Vector<String> s = new Vector<String>();
               f = (data_in.readLine()).split(",");
               for (String g : f) {
                                    if(!s.contains(g))
                                    {
                                            s.add(g);
                                            if(!fieldListOneItem.contains(g))
                                    fieldListOneItem.add(g);
                                    }
                                }
               Collections.sort(s);
                        dataset.add(s);
            }
            Collections.sort(fieldListOneItem);
            int hashTableSize = buyukAsalBul(fieldListOneItem.size()*2+1);
            Itemsets itemset1 = new Itemsets(hashTableSize);
            itemset1.getItemsets();
        }
                catch (NumberFormatException e)
                {
                        e.printStackTrace();
                }
                catch (IOException e)
                {
                        e.printStackTrace();
                }
        }
}
```

ÖZGEÇMİŞ

Ad Soyad : Erhan Tahminciler,

Erhan Tahminciler 1984'te İzmir'de doğdu. İlk ve orta okulu Reşat Nuri Güntekin'de, liseyi Çimentaş Lisesi'nde okudu. 2001'de Süleyman Demirel Üniversitesi Bilgisayar Programcılığı bölümünde önlisans öğrenimini 2003 yılında tamamladı. 2005'de Anadolu Üniversitesi İşletme bölümünde başladığı lisans öğrenimini 2008 yılında tamamladı. Okan Üniversitesi Bilgisayar Mühendisliği programında tezli yüksek lisan öğrenimine 2011 yılında başladı. Halen Allianz Türkiye firmasında iş zekası uzmanı olarak veri madenciliği konularında çalışmalarını sürdürmektedir.