T.C.
AKDENİZ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
Biyoistatistik ve Tıp Bilişimi Anabilim Dalı

METİN MADENCİLİĞİ TEKNİKLERİ KULLANILARAK KULAK BURUN BOĞAZ HASTA BİLGİ FORMLARININ ANALİZİ

Başak OĞUZ

Yüksek Lisans Tezi

Antalya, 2009

T.C.
AKDENİZ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
Biyoistatistik ve Tıp Bilişimi Anabilim Dalı

METİN MADENCİLİĞİ TEKNİKLERİ KULLANILARAK KULAK BURUN BOĞAZ HASTA BİLGİ FORMLARININ ANALİZİ

Başak OĞUZ

Yüksek Lisans Tezi

Tez Danışmanı: Yrd. Doç. Dr. Uğur BİLGE

"Kaynakça Gösterilerek Tezimden Yararlanılabilir"

Antalya, 2009

Sağlık Bilimleri Enstitüsü Müdürlüğüne;

Bu çalışma jürimiz tarafından Biyoistatistik ve Tıp Bilişimi Anabilim Dalı'nda Tıp Bilişimi Yüksek Lisans tezi olarak kabul edilmiştir. …./…./….

Tez Danışmanı: Yrd. Doç. Dr. Uğur BİLGE
Akdeniz Üniversitesi
Tıp Fakültesi
Biyoistatistik ve Tıp Bilişimi Anabilim Dalı

Üye: Prof. Dr. Osman SAKA
Akdeniz Üniversitesi
Tıp Fakültesi
Biyoistatistik ve Tıp Bilişimi Anabilim Dalı

Üye: Yrd. Doç. Dr. Neşe ZAYİM
Akdeniz Üniversitesi
Tıp Fakültesi
Biyoistatistik ve Tıp Bilişimi Anabilim Dalı

Üye: Yrd. Doç. Dr. K. Hakan GÜLKESEN
Akdeniz Üniversitesi
Tıp Fakültesi
Biyoistatistik ve Tıp Bilişimi Anabilim Dalı

Üye: Doç. Dr. Levent DÖNMEZ
Akdeniz Üniversitesi
Tıp Fakültesi
Halk Sağlığı Anabilim Dalı

ONAY:

Bu tez, Enstitü Yönetim Kurulunca belirlenen yukarıdaki jüri üyeleri tarafından uygun görülmüş ve Enstitü Yönetim Kurulu'nun …./…./2009 tarih ve ……../..….. sayılı kararıyla kabul edilmiştir.

Prof. Dr. İsmail ÜSTÜNEL

Enstitü Müdürü

ÖZET

Dünyadaki verilerin yaklaşık olarak %90'ı yapılandırılmamış formatta bulunmaktadır. Bu tip veriler üzerinde işlem yapılması, verilerin yönetilmesi veya bu verilere erişim zordur. Bu yüzden 1960'lı yıllardan itibaren veriyi yapılandırıp makine tarafından işlenebilir hale dönüştürme amacıyla sistemler geliştirilmeye başlanmıştır. Metin madenciliği, özellikle 2000'li yıllarda daha fazla ilgi gören, serbest formatta bulunan metinler içerisindeki daha önceden bilmediğimiz bilgileri ortaya çıkarmamızı sağlayan işlemler bütünüdür.

Metin madenciliği tekniklerinin tıpta kullanımı son birkaç yılda büyük oranda artmıştır. Yapılan klinik çalışmalar, araştırma raporları, hastane kayıtları, doktor notları ve faturalar gibi serbest formatta bulunan metinler tıptaki en önemli veri kaynaklarıdır. Fakat yapılandırılmamış formatta bulunan bu geniş veri yığınlarını insan gücüyle analiz etmek ve istenilen bilgiye ulaşmak hem zordur hem de zaman kaybına yol açmaktadır. Hastayla ilgili karar verme süresinin, doğru verilere erişmenin ve bu verileri kullanarak istenilen bilgilere ulaşmanın zorluğu göz önünde bulundurulduğunda bu tür sistemlerin önemi ön plana çıkmaktadır.

Bu çalışmada, Akdeniz Üniversitesi Hastanesi Kulak Burun Boğaz Hastalıkları Anabilim Dalı'ndan alınan ameliyat geçiren hastalara ait 600 adet hasta bilgi formunu yapılandırılmış formata dönüştürmek, hekimlerin hasta ile ilgili ihtiyaç duydukları bilgilere erişimini kolaylaştırmak, hasta bilgi formlarından klinik verileri çıkartmak ve bu verileri analiz etmek amacıyla bir yazılım geliştirilmiştir. Önce Microsoft Office Word belge formatında bulunan hasta bilgi formlarındaki veri alanları ön işlemden geçirilerek veri tablosu haline dönüştürülmüştür. Veri tablosundaki veriler, Microsoft Office Excel'e gönderilebilmekte veya XML olarak veritabanına kaydedilebilmektedir. Hazırlanan metin sorgu formuyla birlikte hekimlerin hasta bilgi formlarında aradıkları özellikteki hastalara erişimlerinde kolaylık sağlanmaktadır. Ayrıca her alana özgü oluşturulan anahtar kelime listeleriyle metin içerikleri kodlanabilmekte ve bu veriler üzerinde veri madenciliği teknikleri uygulanabilmektedir. Bu çalışmada, varlıklar/ kavramlar arasındaki ilişkilerin tanımlanabilmesi için veri madenciliğinde kullanılan ilk tekniklerden biri olan Birliktelik Kuralı yöntemi uygulanmıştır. İlerleyen zamanlarda kazanılan deneyimlerle diğer anabilim dallarında da kullanılabilecek daha kapsamlı ve profesyonel bir yazılım geliştirilmesi planlanmaktadır.

Anahtar Kelimeler: kulak burun boğaz; metin madenciliği; veri madenciliği; birliktelik kuralları

ABSTRACT

Approximately 90% of the world's data is held in unstructured or semi-structured formats. Since 1960s, several methods have been developed to transform the data into a structured and machine processable format. Text mining has seen an increasing interest in the 2000s, for discovering unknown information and facts from the free-text data.

Recently, the number of text mining applications in medical sciences has grown with an increasing rate. Unstructured free-text data, such as patient discharge notes and reports, doctor's notes, clinical trials and studies, research reports, web pages and hospital records are some of the important data sources for physicians. To analyze and access this kind of data by human efforts is difficult and time consuming. Considering the time it takes for decision making, and accesing accurate and required information about patients, this kind of systems have become necessary.

In this study, we developed a software system to transform 600 discharge notes, from the Department of Otolaryngology of Akdeniz University, to a structured form, enabling physicians to access patient information, extracting clinical data from the discharge notes, and codifying them for analysis. First of all, discharge notes which are kept as Microsoft Office Word documents have been transformed into a data table after preprocessing. Data in the data table can be stored in XML format or as Microsoft Office Excel spreadsheets. A query form has also been designed for enabling physicians to access the patient data. To identify the significant content words within each section keyword lists have been used and content words have been converted into a predefined coded structure. Association Rules, that is one of the methods of the traditional data mining, has been applied to the coded data in order to discover the relations between entities/concepts. In the future, we plan to develop a more comprehensive and professional software that could be used in the other departments of the hospital.

Key Words: otolaryngology; text mining; data mining; association rules

TEŞEKKÜR

Bu tezin hazırlanmasında her türlü imkanı sağlayan, katkıları ve eleştirileriyle bana yol gösteren değerli hocalarım Prof. Dr. Osman SAKA, Yrd. Doç. Dr. Kemal Hakan GÜLKESEN, Yrd. Doç. Dr. Neşe ZAYİM'e, bana rehberlik eden danışmanım Yrd. Doç. Dr. Uğur BİLGE'ye, Kulak Burun Boğaz alanında bilgileriyle bana destek veren Yrd. Doç. Dr. Murat TURHAN'a, tez çalışması sürecinde teknik açıdan bana her zaman yardımcı olan arkadaşım Mehmet Kemal SAMUR'a, moral ve yardımlarını esirgemeyen mesai arkadaşlarım Deniz ÖZEL, Özgür TOSUN, Filiz İŞLEYEN, Anıl AKTAŞ SAMUR, Selen BOZKURT ve Yılmaz Kemal YÜCE'ye teşekkürlerimi sunarım.

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

VTYS	**:**	Veri Tabanı Yönetim Sistemi
KBB	**:**	Kulak Burun Boğaz
DDİ	**:**	Doğal Dil İşleme
NER	**:**	Named Entity Recognition
NLM	**:**	National Library of Medicine
UMLS	**:**	Unified Medical Language System
XML	**:**	Extensible Markup Language
W3C	**:**	The World Wide Web Consortium
GO	**:**	Gene Ontology

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

GİRİŞ

Bilgisayar teknolojilerindeki hızlı gelişmeler hayatın her alanında kolaylıklar sağlamakta ve kişilerin beklentilerini her geçen gün arttırmaktadır. Bu beklentilerin karşılanması amacıyla ihtiyaca göre birçok sistem geliştirilmekte ve bu sistemler günümüzde giderek daha fazla özelliğe sahip ve fonksiyonel hale gelmektedir. Bilgi sistemleri, kişilerin bu ihtiyaçları karşısında ortaya çıkan, son yıllarda çeşitli alanlarda yaygın olarak kullanılan ve verilerin veritabanlarında depolanmasını sağlayarak veri organizasyonunu kolaylaştıran sistemlerdir. Bu sistemlerle birlikte veriler çok hızlı bir şekilde veritabanlarına depolanabilmekte ve istenildiği zaman istenilen veriye erişilebilmektedir.

İlk olarak VTYS (Veritabanı Yönetim Sistemleri) ile birlikte birçok alanda veriler depolanmış, güncellenmiş ve yönetilmiştir. Fakat son yıllarda veritabanlarında depolanan verinin hacmi oldukça büyümüş [1] ve buna bağlı olarak verileri çeşitli yöntemlerle organize etme, bu kadar çok veri arasından gereken bilgiyi çıkartma, verileri analiz ederek bilgiye dönüştürme, veriler arasındaki ilişki örüntülerini tespit etme vb. ihtiyaçlar ortaya çıkmıştır. Artık sadece bilgiye erişmek değil, gerekli koşullarda bilgi üretmek de önemli hale gelmiştir. Bu noktada veri madenciliği kavramı karşımıza çıkmaktadır. Veri madenciliğinin ilk adımları 1960'lı yıllarda atılmış olmasına rağmen, kavramsal olarak 1990'lı yıllardan sonra tam olarak ortaya çıkmıştır. Veri madenciliği, veri ambarlarında yararlı olma potansiyeline sahip, aralarında beklenmedik/bilinmedik ilişkilerin olduğu verilerin keşfedilerek, hem anlaşılır hem de kullanılabilir bir biçime getirilmesine yönelik geliştirilmiş yöntemler topluluğudur [2]. Tanımda da belirtildiği üzere veri madenciliği, birçok alanda büyük yığınlar halinde bulunan verilerin analiz edilerek insanlara daha önceden bilmedikleri, beklenmeyen örüntülerin sunulmasını sağlamaktadır. Ne yazık ki veri madenciliği teknikleri sadece yapılandırılmış formatta bulunan sayısal veriler üzerinde uygulanabilmektedir. Fakat dünyadaki verilerin yaklaşık olarak % 90'ının yapılandırılmamış formatta bulunduğu [3] düşünüldüğünde geleneksel veri madenciliğinin yetersiz kaldığı gözlemlenmiştir. Metin madenciliği bu problemlere çözüm olarak sunulan, metin formatındaki verileri kullanarak içerisindeki bilgileri gün ışığına çıkaran ve özellikle 2000'li yıllardan sonra ilginin giderek arttığı önemli bir alandır [4].

Tıp alanında bulunan mevcut veri oldukça fazla ve hayati öneme sahiptir. Bilgi sistemleri sayesinde bu verilerin bazıları yönetsel ve klinik amaçlarla düzenli olarak kaydedilmekte ve bu veriler üzerinde yapılan veri madenciliği çalışmaları hem uzmanlar ve hastane yönetimine yardımcı olmakta hem de hastaların daha kaliteli bir hizmet almalarında etkin rol oynamaktadır [1]. Ne yazık ki hasta verilerinin birçoğu hala veri madenciliği yöntemlerinin uygulanamayacağı karmaşık, yapılandırılmamış formatta, kağıt tabanlı veya elektronik ortamda serbest metinlerde bulunmaktadır. Hekimler karar verme sürecinde veya araştırma yaparken hasta raporları, klinik

çalışmalar, araştırma raporları, web sayfaları ve hastane kayıtları gibi serbest metin formatında veya kağıt tabanlı olarak bulunan bu metinleri kullanmaktadır. Yapılandırılmamış formatta bulunan bu geniş veri yığınlarını insan gücüyle analiz etmek ve istenilen bilgiye ulaşmak hem zordur hem de zaman kaybına yol açmaktadır. Hastayla ilgili karar verme süresinin, doğru verilere erişmenin ve bu verileri kullanarak istenilen bilgilere ulaşmanın önemi göz önünde bulundurulduğunda bu tür metinler içerisindeki verilere erişmeyi ve verilerin analiz edilmesini sağlayan sistemlere olan ihtiyaç ön plana çıkmıştır.

Bu çalışmada, KBB (Kulak Burun Boğaz) Hastalıkları Anabilim Dalından alınan ve ameliyat geçiren hastalara ait hasta bilgi formları kullanılmış ve bu metinler üzerinde işlemler yapılmıştır. KBB burun, boğaz, kulak, baş ve boyun rahatsızlıkları ile ilgilenen [5], hastanelerde en fazla yoğunluğun yaşandığı bölümlerden biridir. Bu yüzden inanılmaz hacimde veri depolanmakta ve bu verilerin çoğu kağıt tabanlı olarak ya da elektronik ortamda yapılandırılmamış formatta tutulmaktadır. Hastalara ait verilerin yapılandırılmamış formatta bulunması hem verilere erişimi hem de verilerin analiz edilmesini zorlaştırmaktadır.

Bu çalışma üç temel amaç üzerine kurulmuştur. Bunlar;

1. Yapılandırılmamış formatta bulunan KBB hasta bilgi formlarının yapılandırılmış hale dönüştürülmesi,

2. Hekimlerin karar verirken ya da araştırma yaparken hasta ile ilgili ihtiyaç duydukları bilgilere erişimlerinin kolaylaştırılması ve hasta bilgi formlarını incelemek için harcadıkları zamanın azaltılması

3. Veri madenciliği teknikleri ile elde edilen verilerin analiz edilmesi ve varlıklar arasındaki gizli ilişkilerin çıkartılmasıdır.

Bu amaçlar doğrultusunda KBB hasta bilgi formlarını yapılandırılmış formata dönüştüren, hasta bilgilerine erişimi kolaylaştıran ve metinlerdeki varlıklar/kavramlar arasındaki ilişki örüntülerini ortaya çıkaran bir metin analiz aracı geliştirilmiştir.

GENEL BİLGİLER

2.1. Veri Madenciliği

2.1.1. Veri, Enformasyon ve Bilgi Kavramları

Günümüzde bilişim sistemlerinin hayatın hemen hemen her alanında aktif bir rol oynuyor olması veri, enformasyon, bilgi vb. birçok kavrama aşina olmamızı sağlamıştır. Günlük hayatımızda sıkça geçen bu kelimeler bilişim dünyasının yapı taşlarıdır. Genel olarak veri (Data), bilgi (Knowledge)-enformasyon (Information) ile eş anlamlı olarak kullanılmaktadır [6]. Veri, temel olarak varlığı bilinen, işlenmemiş, ham haldeki çeşitli sembol, harf, rakam ve işaretlerle temsil edilen kayıtlar veya gözlemler olarak adlandırılmaktadır. Bu kayıtlar ilişkilendirilmemiş, düzenlenmemiş yani anlamlandırılmamışlardır. Bazı durumlarda işlenerek farklı bir boyut kazanan bir veri, daha sonra başka bir amaç için bu haliyle kullanılmak üzere kayıt altına alındığında veri halini koruyacaktır. Enformasyon, veri kavramının tanımından yola çıkıldığında, adreslemedeki ikinci safhadır. Yani verilerin bilgi işlem yardımıyla ilişkilendirilmiş, düzenlenmiş, anlamlandırılmış, işlenmiş halidir. Bu haliyle enformasyon, potansiyel olarak içinde bilgi barından bir veri halindedir. Bilgi, bu süreçteki üçüncü aşamadır. Veri ve enformasyonun işlenmesiyle elde edilen anlamlı mesaj veren kavramlardır. Enformasyonun bilgiye dönüşmesi, bireyin onu algılaması, özümsemesi ve sonuç çıkarmasıyla gerçekleşir. Dolayısıyla bireyin algılama yeteneği, yaratıcılık, deneyim gibi kişisel nitelikleri de bu süreci doğrudan etkilemektedir [7].

2.1.2. Yapılandırılmış ve Yapılandırılmamış Veri

Dünyadaki verilerin yaklaşık olarak % 90'ı yapılandırılmamış formatta bulunmaktadır [3]. Yapılandırılmamış veri, bilgisayar tarafından kolayca anlaşılmayan ve veri yapısına sahip olmayan müzik, video dosyaları ve serbest formatta bulunan metinlerdir (e-mail, web sayfaları vb.) [8]. Yapılandırılmış veri ise veri tabanlarında depolanan, makine tarafından tanınabilen, herhangi bir parçasına sorgular aracılığı ile erişim sağlanabilen, kullanılabilen, iletilebilen ve güncellenebilen verilerdir. Yapılandırılmamış veri herhangi bir tanımlanabilir forma sahip değildir. Bu tip veriler üzerinde işlem yapılması, verilerin yönetilmesi veya bu verilere erişim zordur. Bu yüzden 1960'lı yıllardan itibaren veriyi makine tarafından anlaşılabilir hale dönüştürme ve veriler üzerinde işlem yapmayı kolaylaştırma amacıyla sistemler geliştirilmeye başlanmıştır. Çizelge 2. 1.'de yapılandırılmış veriye örnek gösterilmiştir. Ayrıca Şekil 2. 1.'de çalışmada kullanılan bir hastaya ait KBB hasta bilgi formu, yapılandırılmamış veriye örnek olarak verilmiştir.

Çizelge 2. 1. Yapılandırılmış Veri Örneği

Cinsiyet	Ağırlık	Sistolik Kan Basıncı	Hastalık Kodu
E	72	175	3
K	65	141	1
K	48	151	2
E	79	160	2
…	…	…	…
E	51	165	3

Şikayeti: Nefes darlığı,ses kısıklığı
Hikayesi: 5 aydır ses kısklığı olan ve son aylarda nefes darlığı oluşan hastaya kepez DH'de bx alınmış.patoloji scc?? gelmiş.Hastaya operasyon önerilmiş,ancak hasta reddetmiş.Ciddi solunum sıkıntısı olusan hasta acil trakeostomi açılmak üzere kliniğimize yatırıldı.
Özgeçmiş: 50 paket sigara yılı
Soygeçmiş: Özellik yok.
Fizik Muayene:
Orofareks: Doğal
Rinoskopi anterior: Doğal
Otoskopi: Doğal
İndirekt larengoskopi: Doğal
Boyun muayenesi: Sağ scm arkasında 1x1 cm LAP mevcut
Rinoskopi posterior : Doğal
Laboratuar İncelemeleri:
Hemogram - biyokimya: Normal
PA AC Grafisi: Doğal
Batın USG : Doğal
Tanı: Yassı epitel hücreli karsinom, iyi derecede diferansiye, larinks, biopsi
Yorum: Cerrahi sınırlarda tümör devam etmektedir.

Şekil 2. 1. Yapılandırılmamış Veri Örneği

Daha önce de değinildiği gibi veri madenciliği uygulamalarında çoğunlukla yapılandırılmış veriler kullanılmaktadır. Veriler herhangi bir teknik uygulanmadan önce özel yollarla hazırlanmalıdır. Çizelge 2. 1.'de de gösterildiği gibi veri madenciliği uygulamalarında iki tip veri türü kullanılmaktadır; 1-Sayısal, 2-Kategorik [9]. Ağırlık, Boy, Yaş, Sistolik Kan Basıncı vb. değişkenler sayısal, cinsiyet, eğitim, medeni hal vb. değişkenler ise kategorik veriye örnek verilebilir. En genel kullanılan kategorik özellikler 1 veya 0 ya da doğru-yanlış ile gösterilebilenlerdir. Diğer yaygın olan bir özellik ise kodlanmış olarak sunulanlardır [9]. Yukarıdaki örnekte "hastalık" değişkeni kodlanmış olarak girilmiştir.

Veri madenciliğinde elektronik tablo halinde sunulan veriler kullanılırken metin madenciliği uygulamaları metin formatındaki verileri kullanmaktadır. Metin madenciliğinin ana konularından biri metin verilerin sayısal veri haline dönüştürülüp elektronik tablo şeklinde sunulmasıdır. Böylelikle yapılandırılmamış formatta bulunan metinler veri madenciliği tekniklerinin uygulanabileceği yapılandırılmış formata dönüştürülebilmektedir.

2.1.3. Veri Madenciliği

Veri madenciliği, veri ambarlarında yararlı olma potansiyeline sahip, aralarında beklenmedik/bilinmedik ilişkilerin olduğu verilerin keşfedilerek hem anlaşılır hem de kullanılabilir bir biçime getirilmesine yönelik geliştirilmiş yöntemler topluluğudur. Diğer bir deyişle, veri madenciliği tek başına bir şey ifade etmeyen veriler içindeki gizli örüntüleri ve ilişkileri ortaya çıkarmak için istatistik, yapay zeka ve makine öğrenmesi gibi yöntemlerin ileri veri çözümleme araçlarıyla kullanılmasını kapsayan süreçler topluluğudur. Geleneksel sorgu (Query) ve raporlama araçlarının veri yığınları karşısında yetersiz kalması, saklı ve işlenmemiş bilgiye olan büyük ihtiyaç Veritabanlarında Bilgi Keşfi ve Veri Madenciliği gibi alanların keşfiyle anlaşılabilir ve yorumlanabilir hale gelmiştir [10].

Veri madenciliği 1960'lı yıllarda, bilgisayarların veri analiz problemlerini çözmek için kullanılmaya başlanmıştır. O dönemlerde bilgisayar yardımıyla yeterince uzun bir tarama yapıldığında, istenilen verilere ulaşmanın mümkün olacağı gerçeği kabullenilmiş ve bu işleme veri madenciliği yerine önceleri veri taraması (data dredging), veri yakalanması (data fishing) gibi isimler verilmiştir. 1990'lı yıllara gelindiğinde veri madenciliği ismi, bilgisayar mühendisleri tarafından ortaya atılmıştır [11].

Veri madenciliği ile büyük veri kümeleri üzerinde ilginç ve gizli kalmış örüntüleri tespit etmek mümkündür. Bu noktada veri madenciliği istatistikten ayrılır. İstatistikçiler bilinen faktörler arasındaki ilişkilerin güçlülüğünü araştırır. Veri madenciliğinde ise bilinmeyen faktörler arasında tahmin edilemeyecek ilişkilerin güçlülüğü araştırılır. Örneğin "Çocuk bezi alan müşterilerin %30'u bira da satın alır" gibi bir ilişkiyi istatistik yaklaşımları ile bulmak zordur. Çünkü "Çocuk bezi" ve "Bira", arasında ilişki olduğu düşünülecek ürünler değildir. "Çocuk bezi" ve "Bira" arasındaki ilişki önceden tahmin edilemese de veri madenciliği teknikleri ile keşfedilebilecek bir ilişkidir [12].

Veri madenciliği teknikleri pazarlama, bankacılık, tıp, mühendislik başta olmak üzere birçok alanda yaygın olarak kullanılmaktadır. Özellikle tıp alanında bulunan mevcut veri oldukça fazla ve hayati öneme sahiptir. Bu veriler kullanılarak belirli bir hastalığa sahip kişilerin ortak özellikleri, hastaların ön tanıları, tıbbi tedaviden sonra hastaların durumları, hastane maliyetleri, ölüm oranları ve salgın hastalıklar gibi tahminler yapılabilmektedir.

2.1.4. Veri Madenciliği Metotları

Veri madenciliğinde kullanılan metotlar, tahmin edici (Predictive) ve tanımlayıcı (Descriptive) olmak üzere iki ana başlık altında incelenmektedir. Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır. Tanımlayıcı modellerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır [13]. Sınıflama ve regresyon modelleri tahmin edici, kümeleme, birliktelik kuralları ve ardışık zamanlı örüntü modelleri ise tanımlayıcı modellerdir [1].

Bu çalışmada hasta bilgi formlarında bulunan varlıklar arasındaki ilişkilerin belirlenmesi için veri madenciliği tekniklerinden **Birliktelik Kuralları** (Association Rules) yöntemi kullanılmıştır. Veri madenciliğinde kullanılan ilk tekniklerden birisi de birliktelik kurallarıdır [14]. Birliktelik analizi, belirli bir veri kümesinde yüksek sıklıkta birlikte görülen özellik değerlerine ait ilişkisel kuralların keşfidir. Birliktelik kuralı, geçmiş verilerin analiz edilerek bu veriler içindeki birliktelik davranışlarının tespiti ile geleceğe yönelik çalışmalar yapılmasını destekleyen bir yaklaşımdır [15]. Birliktelik kuralının matematiksel modeli Agrawal, Imielinski ve Swami tarafından 1993 yılında sunulmuştur. Bu modelde, $I = \{i_1, i_2, .., i_m\}$ kümesine "ürünler" adı verilmektedir. D, veri bütünlüğündeki tüm hareketleri, T ise ürünlerin her bir hareketini simgeler. TID ise, her harekete ait olan tek belirteçtir [15].

Birliktelik kuralı şu şekilde tanımlanabilir;

$A_1, A_2, ..., A_m \quad => \quad B_1, B_2, ..., B_n$

Bu ifadede yer alan, A_m ve B_n, yapılan iş veya nesnelerdir. Bu kural, genellikle "$A_1, A_2, ..., A_m$" iş veya nesneleri meydana geldiğinde, sık olarak "$B_1, B_2, ..., B_n$" iş veya nesnelerinin aynı olay veya hareket içinde yer aldığını belirtir [16].

Birliktelik Kuralında, öğeler arasındaki bağıntı, destek ve güven kriterleri ile hesaplanır. Destek kriteri, veride öğeler arasındaki bağıntının ne kadar sık olduğunu, güven kriteri ise Y öğesinin hangi olasılıkla X öğesi ile beraber olacağını söyler. İki öğenin birlikteliğinin önemli olması için hem destek, hem de güven kriterinin olabildiğince yüksek olması gerekmektedir [17].

Birliktelik kuralı, kullanıcı tarafından minimum değeri belirlenmiş destek ve güven eşik değerlerini sağlayacak biçimde üretilir. A ve B öğe kümelerinin, birliktelik kuralı "A=>B" olarak gösterilirse, destek aşağıdaki gibi tanımlanır.

Destek (A=>B)= (A ve B'nin bulunduğu satır sayısı) / (toplam satır sayısı)

A=>B birliktelik kuralının güven değeri ise, A'yı içeren hareketlerin B'yi de içerme yüzdesidir. Örneğin, bir kural % 85 güvenirliğe sahip ise, A'yı içeren ürün kümelerinin % 85'i B'yi de içermektedir. İşe bağlı veri satırları verilmiş ise, (A=>B) güveni aşağıdaki gibi tanımlanır.

Güven (A=>B) = (A ve B'nin bulunduğu satır sayısı) / (A'nın bulunduğu satır sayısı)

Güven değerinin % 100 olması durumunda, kural bütün veri analizlerinde doğrudur ve bu kurallara "kesin" denir. Elde edilen kuralın daha önceden belirlenen minimum destek ve güven değerini sağlaması gerekmektedir. Birliktelik kurallarının kalitesi ile ilgili olarak geliştirilen diğer bir parametre ise "lift" değeridir. Lift değeri güven değerinin beklenen güven değerine bölünmesiyle elde edilmektedir. 0 ile sonsuz arasında bir değer almaktadır [18]. Beklenen güven değeri ve lift değeri aşağıda tanımlanmıştır.

Beklenen Güven (A=>B) = Destek (A)*Destek (B)/Destek (A)

Lift (A=>B) = Güven (A=>B)/ Beklenen Güven (A=>B)

Lift değerinin birden küçük olması A olayının meydana gelmesinin B olayı üzerinde negatif bir etki yaratacağı anlamına gelmektedir. Değerin 1'e yakın olması A olayının meydana gelmesinin B olayı üzerinde hiçbir etkisi olmadığını, yani birbirinden bağımsız olaylar olduğunu göstermektedir. 1'den büyük olması ise A olayının meydana gelmesinin B olayı üzerinde pozitif bir etkisi olduğunu belirtmektedir [18]. Kuralın kaliteli bir kural olması için lift değerinin 1'den büyük olması gerekmektedir.

Birliktelik kuralına ilişkin olarak geliştirilen bazı algoritmalar şunlardır; AIS- mining association rules between sets of items in large databases [14], SETM- SET-oriented Mining of association rules [19], Apriori [20], Partition [21], RARM - Rapid Association Rule Mining [22], CHARM- Closed Association Rule Mining [23]. Bu algoritmalar içerisinde, ilk olanı AIS, en bilineni ise Apriori algoritmasıdır [24].

Birliktelik kuralının en yaygın kullanıldığı örnek market sepet analizidir [25]. Market sepet analizi, müşterilerin yaptıkları alışverişlerdeki ürünler arasındaki birliktelikleri bularak müşterilerin satın alma alışkanlıklarını belirlemeye çalışır. Veriler metin formatında bulunduğunda market sepet analizindeki ürünler yerini kavramlara bırakmaktadır. Kavramların birlikte bulunma durumlarına bakılarak aralarındaki ilişki örüntüleri tespit edilmektedir. Birliktelik Kuralı analizi tıp alanında genel olarak genler arasındaki ilişkilerin belirlenmesinde [26], hastalıkların tahmininde [27], risk faktörlerinin belirlenmesinde vb. uygulamalarda kullanılmaktadır.

2.2. Metin Madenciliği

2.2.1. Metin Madenciliği Nedir?

İnternet kullanımının hızla artması ve kişisel bilgisayarların yaygınlaşması ile birlikte gittikçe büyüyen hacme sahip doküman yığınları oluşmaktadır. Pek çok araştırma alanında ve günlük hayatın içinde üretilen bilgiler ağırlıklı olarak metin belgeler şeklinde oluşturulur, bu belgeler taraflar arasında gönderilir, farklı birikimlere sahip kişiler tarafından güncellenir ve belirli amaçlar için değişik ortamlarda saklanır [28]. Giderek artan bu belge yığınları içinde önemli bilgiler kaybolup giderken, değerli bilgilere ulaşmak için dokümanların içeriğinin belirlenmesi ve buna uygun sorgulanabilmesi ihtiyacı kendini hissettirmektedir. Geleneksel leşmiş yöntemleri içeren bilgi erişim (information retrieval) sistemleri belge yığınlarından faydalı ve gerekli bilgileri bulmaya yardımcı olsalar da gerekli detay ve özel bilgilere, bu yöntemler ile ulaşmak zordur. Oysa pek çok açıdan belgeler içindeki bilgilere, ilişkilere ulaşmak son derece önemlidir. Örneğin bir hastalık için ilaç bulmaya çalışan bir araştırmacının, kendisinden önce yapılmış tüm çalışmaları olabildiğince hızlı bir şekilde incelemesi ve bu inceleme sürecinde belgelerin içeriğine, konusuna, içinde geçen kavramlara ve bu kavramların diğer belgelerde geçen farklı kavramlarla ilişkisine ulaşması gerekir [28]. Metin madenciliği metin formatındaki verileri kullanarak içerisindeki bilgileri gün ışığına çıkaran ve özellikle 2000'li yıllardan sonra ilginin giderek arttığı önemli bir alandır [4].

Metin madenciliği, belirli bir formatta olmayan, yazı tipindeki veriler içerisinde gizli olan nitelikli bilginin çıkarılması, düzensiz haldeki verinin formatlanması sürecidir. Metin Madenciliği, Metin Veri Madenciliği (Text Data Mining) ve Metin Veritabanlarından Bilgi Keşfi (Knowledge Discovery from Textual Databases) olarak da adlandırılır [29]. Metin madenciliği yeni bir terim olmasına rağmen, bilgi erişim sistemleri ve DDİ (Doğal Dil İşleme) ile ilgili yapılan araştırmalara bağlı olarak ortaya çıkmıştır. Bilgi erişim ile ilgili çalışmalar 1960'lı yıllarda başlamış, doğal dili anlamaya ve sayısallaştırmaya yönelik programlar geliştirilmeye çalışılmıştır. 1990'lı yıllara gelindiğinde ise metinlere erişim, metinlerden bilgi çıkarımı, metin kategorizasyonu ve metinleri yapısal hale getirmeye yönelik çalışmalar hızla artmaya başlamıştır.

Kostoff and DeMarco [30] bilim ve teknoloji metin madenciliğini "bilginin teknik literatürden çıkartılması" olarak tanımlamış ve bilgi erişim, bilgi işleme ve bilgi entegrasyonu olmak üzere üç bileşenden oluştuğunu belirtmişlerdir. Bilgi işlemeyi, erişilen belgelerdeki örüntülerin çıkartılması işlemi, bilgi entegrasyonunu ise erişilen ilgili belgelerin okunarak bilgi işleme aşamasından sonra çıkan sonuçlarla kombinasyonun sağlanması süreci olarak tanımlamışlardır. Losiewicz et al. [31] metin veri madenciliğini, metin koleksiyonlarından bilgiye erişmeyi, bireysel metinlerden bilgi çıkarmayı, veritabanlarından bilgi keşfini, organizasyonlarda bilgi yönetimini ve verinin ve bilginin görselleştirilmesi aşamalarını birleştiren bir mimari olarak tanımlamışlardır.

Literatürde yapılan tanımlamalara bakıldığında metin madenciliğinin metin içindeki kalıpları tanımlayarak bilinmeyen bilgiyi ortaya çıkaran ve var olan yapısı ile metinleri bilgiye dönüştüren anahtar bir süreç olduğu söylenebilir. Farklı dillerde binlerce doküman, 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 [32]. Bu metinlerin kısa sürede analiz edilmesi ve nitelikli bilgilere çok kısa sürede erişilmesi için metin madenciliği yöntemi günümüzde sıklıkla kullanılmaktadır.

2.2.2. Metin Madenciliği Adımları

Metin madenciliği genel olarak beş adımdan oluşmaktadır (Şekil 2. 2.). Bu adımlar;

- **Metin koleksiyonu oluşturma:** İlgilenilen konularda bilgi erişim sistemleri kullanılarak metin koleksiyonu oluşturma sürecidir. Bu süreç, günümüzde genel olarak internet üzerinden, özellikle Google vb. arama motorları kullanılarak gerçekleştirilmektedir. Tıp alanında ise metin toplama süreci yaygın olarak PubMed çevrim içi veritabanı üzerinden yapılmaktadır [33]. Çevrim içi veritabanlarının yanı sıra veritabanlarında ya da kişisel bilgisayarlarda bulunan metin türü veriler ile oluşturulan koleksiyonlar da metin madenciliğinde kullanılmaktadır.

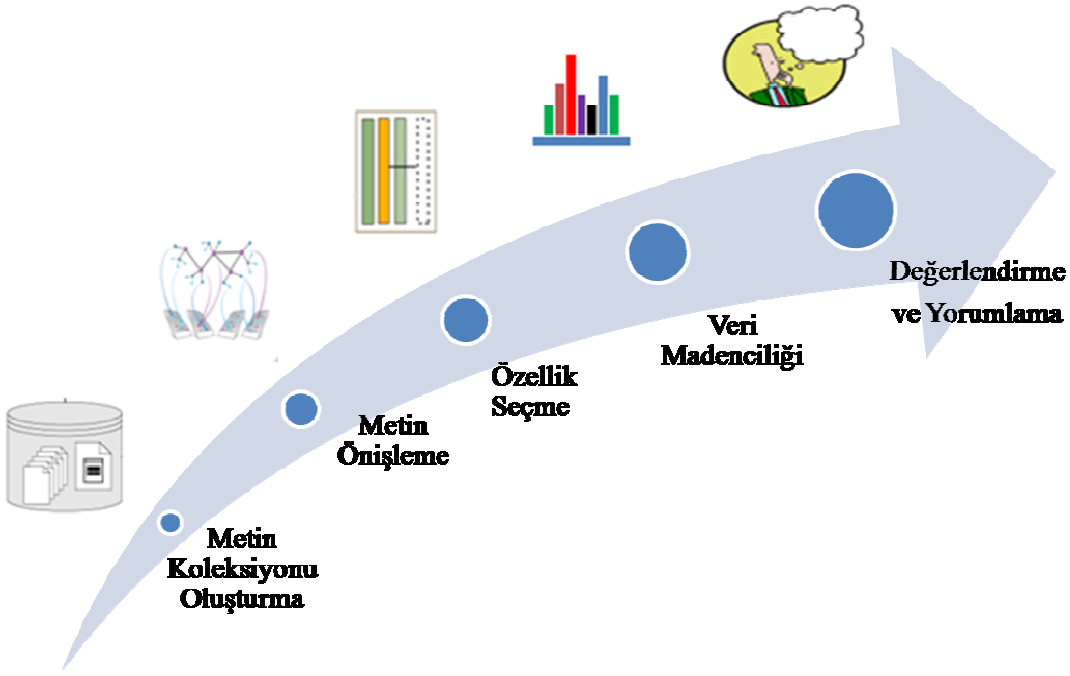
- **Metin önişleme:** Metni kelimelere ayırma, kelimelerin anlamsal değerlerini bulma (isim, sıfat, fiil, zarf, zamir vb.), kelimeleri köklerine ayırma ve gereksiz kelimeleri ayıklama, yazım kurallarına uygunluğunu tespit etmek ve var olan hataları

düzeltmek gibi metin belgelerin yapıtaşı olan kelimelerle ilgili işlemleri içeren süreçtir [33].

- **Özellik seçme:** Metin madenciliği uygulamalarında her zaman gürültülü ve önemsiz bilgi içeren metin koleksiyonlarıyla uğraşılma ihtiyacı bulunmaktadır. İlgili verilerin saptanması üzerine odaklanan özellik seçme, büyük miktarlardaki veriler üzerinde işlem yapılırken iş yükünü azaltmada yardımcı olmaktadır [34]. Özellik seçme aşamasında, ön işlemden geçen metinlerdeki önemli kelimeleri (varlıkları) belirleme (isimler, tamlamalar, bileşik kelimeler, kısaltmalar, sayılar, tarihler, para birimleri vb.) ve ilişkili olmayan özelliklerin çıkarılması (sadece birkaç dokümanda gözlemlenen özelliklerin çıkarılması, birçok dokumanda gözlemlenen özellikleri azaltma vb.) işlemleri yapılmaktadır.

- **Veri madenciliği:** Yapılandırılmış bir formata dönüştürülen metinlerin geleneksel veri madenciliği teknikleriyle (karar ağaçları, yapay sinir ağları, kümeleme, genetik algoritma vb.) analizi sürecidir. Hem veri madenciliğinde hem de metin madenciliğinde gizli bilgilere bakılmakta ve genel yapay zeka, makine öğrenme ve istatistik algoritmaları kullanılmaktadır. Veri madenciliğinde yapılandırılmış sayısal veri kullanılırken metin madenciliği yapılandırılmamış metinlerle ilgilidir. Veri madenciliğinde, veri ambarlarında çıkartılmış, dönüştürülmüş ve yüklenmiş durumda bulunan verileri kullanırken metin madenciliği kesin olmayan verileri modellemeye çalışmaktadır [4].

- **Değerlendirme ve Yorumlama:** Veri madenciliği yöntemleri ile verilerin analizinden elde edilen sonuçların değerlendirilip kullanıcıya uygun ve anlaşılır bir şekilde sunulması işlemidir.



Şekil 2. 2. Metin Madenciliği Süreci

2.2.3. Metin Madenciliği ile İlişkili Alanlar

İnternette yapılan taramalar ve literatür gözden geçirme sonucunda metin madenciliği ile ilişkili bazı alanlar ön plana çıkmıştır. Bunlardan en önemlileri aşağıda açıklanmıştır.

2.2.3.1. Doğal Dil İşleme

Dil yeteneği, insan beyninin nasıl çalıştığına ışık tutan insan türüne özgü tek özellik olduğu için dilbilim bilişsel bilimlerde önemli bir yer tutar. Dilin bilgisayar ortamında modeli oluşturulabilirse iletişim için oldukça yararlı bir araç elde edilmiş olur. DDİ, ana işlevi bir doğal dili çözümleme, anlama, yorumlama ve üretme olan bilgisayar sistemlerinin tasarımını ve gerçekleştirilmesini konu alan bir mühendislik alanıdır. DDİ, yapay zeka (bilgi gösterimi, planlama, akıl yürütme, vb.) biçimsel diller kuramı (dil çözümleme), kuramsal dilbilim ve bilgisayar destekli dilbilim, bilişsel psikoloji gibi çok değişik alanlarda geliştirilmiş kuram, yöntem ve teknolojileri bir araya getirir. 1950 ve 1960'larda yapay zekanın küçük bir alt alanı olarak görülen bu konu, araştırmacıların ve gerçekleştirilen uygulamaların elde ettiği başarılar sonunda artık bilgisayar bilimlerinin temel bir disiplini olarak kabul edilmektedir [35]. Örneğin çoğumuzun kullandığı sözcük işlemcilerde bulunan hatalı yazılmış sözcüğün bulunması ve düzeltilmesi özelliği bu tip uygulamaların en basitlerinden biridir [36]. DDİ alanındaki temel araştırmalar şunlar olmuştur:

- Doğal dillerin işlev ve yapısının daha iyi anlaşılması;
- Bilgisayarlar ile insanlar arasındaki arabirim olarak doğal dil kullanmak ve bu şekilde bilgisayarlar ile insanlar arasındaki iletişimi kolaylaştırmak;
- Bilgisayar ile dil çevirisi yapmak [35].

DDİ beş ana seviyede incelenebilir [37];

- Sesbilim
- Biçimbilim
- Sözdizimbilim
- Anlambilim
- Kullanımbilim

Sesbilim harflerin seslerini ve bunların dil içinde nasıl kullanıldığını inceler. Tüm dillerin bir alfabesi vardır ve her harfin sesi diğerlerinden farklıdır. Biçimbilim sözcük kurumudur. İki tür sözcük oluşturma yöntemi mevcuttur. Bunlar türetme ve ses değişmesidir. Sözdizimbilim sözcüklerin cümle oluşturmak için ne şekilde sıralanmaları gerektiğini inceler. Ancak günlük hayatta bazen olması gereken sözdiziminin dışında da cümleler kullanılabilir. Anlambilim dilin gerçek dünyayla iletişim kurmasını sağlar. Cümle yapısının anlaşılması ve bunun sonucunda eyleme geçilmesi bu aşamada olur. Günümüzde anlambilim sonuçlandırılmaya yaklaşılmış bir çalışma durumundadır. Kullanımbilim dilin duruma göre değişimini inceler. Bir sözcük tek başınayken ya da bir cümle içindeyken farklı anlamlar ifade edebilir [37].

DDİ, doğal dillerin kurallı yapısının çözümlenerek anlaşılması veya yeniden üretilmesi amacını taşır. Bu çözümlemenin insana getireceği kolaylıklar; yazılı dokümanların otomatik çevrilmesi, soru-cevap makineleri, otomatik konuşma ve

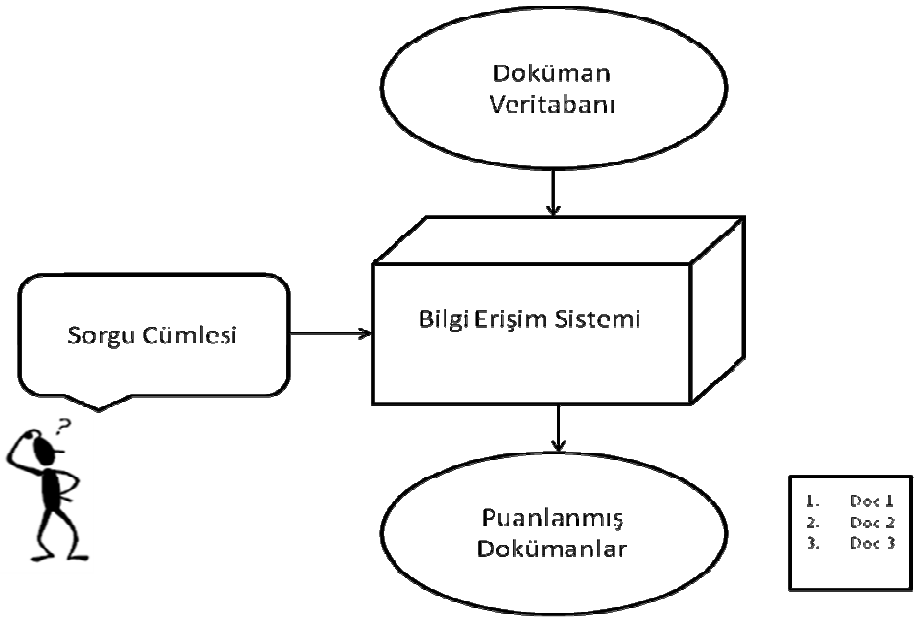
komut anlama, konuşma sentezi, konuşma üretme, otomatik metin özetleme, bilgi sağlama gibi birçok başlıkla özetlenebilir [38].

Günümüzde tıp literatüründe DDİ ile ilgili olarak yapılan birçok çalışma bulunmaktadır. PubMed'de DDİ ile ilgili taramalar yapıldığında 1757 adet çalışma bulunduğu gözlemlenmiştir. Coden et al. [39] çalışmasında kanser hastalığı karakteristiklerini serbest metin formatında bulunan patoloji raporlarından DDİ ve makine öğrenme yöntemlerini kullanarak otomatik olarak çıkartan MedTAS/P adlı bir sistem geliştirmişlerdir. Wang et al. [40] çalışmasında bir DDİ sistemi olan MedLee [41] programı kullanılarak hasta raporlarında bulunan klinik varlıkların (hastalık-semptom) çıkartılması ve keşfedilmesi için otomatikleştirilmiş metotlar geliştirilmiş ve bu varlıklar arasındaki ilişkilere bakılmıştır. Gysbers et al. [42] çalışmasında klinik raporlardaki ilaç yan etkileri ile ilgili terimlerin saptanması için internette halka açık olarak bulunan DDİ aracı caTIES [43] adlı sistemi kullanmışlardır. Goryachev et al. [44] çalışmasında taburcu özetleri ve ayakta tedavi gören hasta raporlarından hastanın aile öyküsünün belirlenmesi ve çıkartılması amacıyla geliştirilen basit bir DDİ algoritması anlatılmıştır. Uzuner et al. [45] çalışmasında hastaların taburcu raporlarından sigara içip içmediklerinin otomatik olarak saptanması için DDİ teknikleri kullanılmış ve sınırlı sayıda özelliğe bakılmıştır ("smok", "tobac", "cigar", SOCIAL HISTORY, vb.).

2.2.3.2. Bilgi Erişim Sistemleri

Bilgisayarların günlük hayatta önemli bir yere sahip olmasıyla birlikte kağıt tabanlı bilgi yerini sayısal ortama bırakmıştır. Kişiler arası bilgi aktarımı, taşınabilir disklerle ve ağ teknolojilerinin gelişmesiyle beraber internet aracılığı ile sayısal olarak daha hızlı bir şekilde yapılabilir hale gelmiştir. Böylelikle internette bilgi patlaması yaşanmış, istenilen bilgiye erişim zorlaşmıştır. Doğru bilgilere kısa sürede erişimi sağlayabilmek, kişilerin dağıtık sistemler üzerinden bilgi edinebilmesini ve büyüklüğü ölçülemez hale gelmiş dağarcıklardan faydalanılabilmesini sağlamak amacı ile bilgi erişim sistemleri geliştirilmiştir. Bilgi Erişimi, belgeler içindeki bilgileri arama, belgelerin kendilerini arama, belgeleri tanımlayan üst veriyi arama veya bilgi kütüphaneleri içinde arama üzerine kurulu bir bilim alanıdır [46]. Bir bilgi erişim sisteminin temel işlevi, kullanıcıların bilgi ihtiyaçlarını karşılaması muhtemel derlemdeki ilgili (relevant) belgelerin tümüne erişmek, ilgili olmayanları da ayıklamaktır [47]. Bir bilgi erişim sisteminin temel bileşenlerinin aşağıda belirtilen 3 ana parçadan oluştuğu söylenebilir.

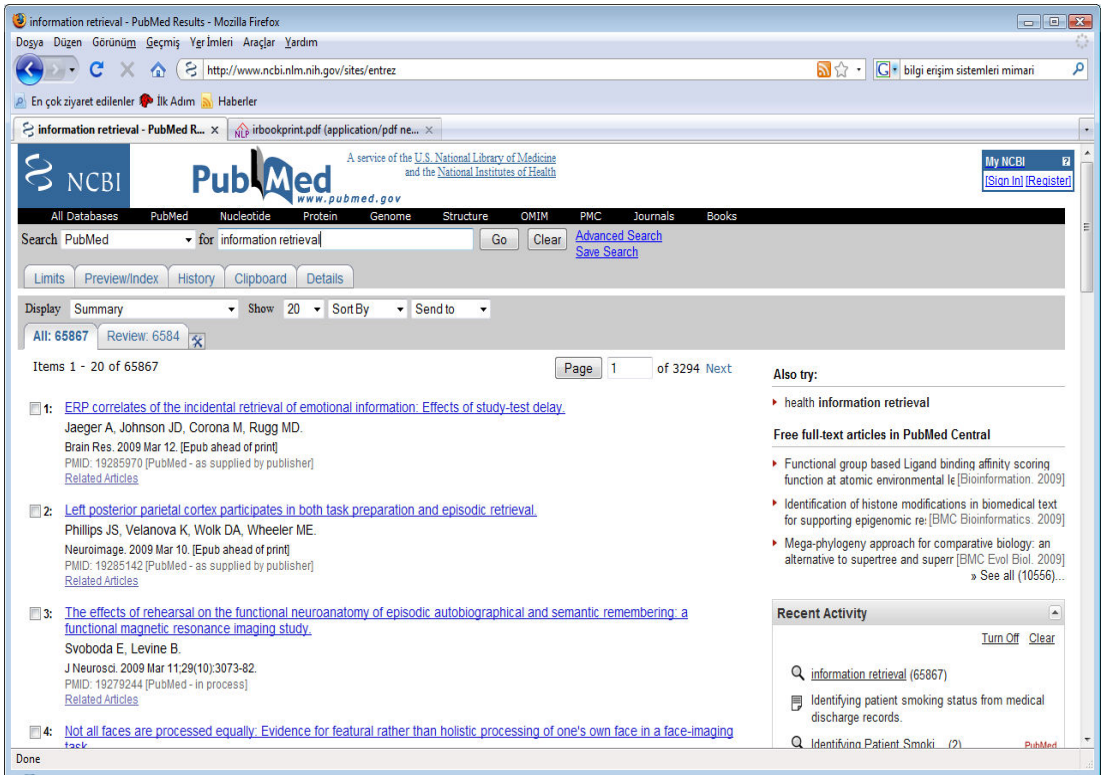
- Bir belge külliyatı ya da bu belgeleri temsil eden dizin terimlerini içeren tutanaklar
- Kullanıcı sorguları
- Kullanıcıların sorgularında yer alan terimler ile belge külliyatında yer alan belgelere atanan terimleri karşılaştırıp ilgili çakışan belgeleri sunan erişim kuralları [46]



Şekil 2. 3. Bilgi Erişim Sistemlerinin Genel Mimarisi [48]

Şekil 2. 3.'te görüldüğü üzere ihtiyaç duyulan bilgiye göre kullanıcı sorguları sisteme gönderilmekte, sistem sorguda kullanılan kelimeleri belge koleksiyonu dizini ile çakıştırmakta ve sonuç olarak belirlenen kurallara (sorgu kelimelerinin doküman içerisinde ve tüm dokümanlarda geçme sıklığı, doküman içerisinde birbirine uzaklığı) göre en yüksek skora sahip belgeler liste halinde kullanıcıya sunulmaktadır. Bilgi erişim sistemlerinin performansının değerlendirilmesinde duyarlılık (precision), anma (recall) ve bu iki ölçütün kombinasyonundan oluşan F-score ölçütleri kullanılmaktadır [4]. Duyarlılık, erişim çıktısındaki ilgili belge sayısının erişim çıktısındaki belge sayısına oranıdır. Anma ise, erişim çıktısındaki ilgili belge sayısının belgeler kümesinde ilgili belgeler sayısına oranıdır [33].

Bilgi erişim sistemleri özellikle akademik ve uzmanlaşmış alanlarda kullanılmaktadır. MEDLINE, 1960'ların sonunda geliştirilmiş, 1971'de uygulanmaya başlanmış tıp alanında yaygın olarak kullanılan, tıp literatüründe istenilen bilgiye erişimi sağlayan bilgi erişim sistemlerinden biridir [4]. Google, Altavista ve Yahoo gibi arama motorları ise genel amaçlı olarak geliştirilen sistemlere örnek olarak verilebilir. Şekil 2. 4.'te PubMed'de "Information Retrieval" anahtar kelimeleri kullanılarak yapılan sorgu sonucu elde edilen sonuç listesi örnek olarak verilmiştir.



Şekil 2. 4. Pubmed'de Yapılan Sorgu Sonucu

2.2.3.3. Bilgi Çıkarım Sistemleri

Bilgi çıkarım sistemleri yapılandırılmamış metinleri yapılandırılmış formata dönüştüren, metin madenciliği için geliştirilen etkili yaklaşımlardan biridir. Bilgi çıkarım sistemleri doğal dille yazılmış metinler kümesindeki her metinden önceden tanımlanmış varlık sınıfları ve ilişkiler ile ilgili bilgiyi çıkartarak bu bilgiyi bir şablon içerisine veya veritabanına yerleştirmektedir [49]. Bu sistemler doğal dille yazılmış dokümanlarda bulunan belirli veri parçalarıyla ilgilenmektedir. Yani yapılandırılmamış metinlerden yapılandırılmış bilgiyi çıkarmaya çalışmaktadır [50].

Sistemler genellikle yapılandırılmamış metinleri veritabanı tablosuna aktarılabilecek bir formata dönüştürmektedirler. Metinlerdeki kişi, yer veya organizasyon isimleri gibi faydalı bilgiler metinleri derin bir şekilde anlamaya çalışmadan çıkartılmaktadır [4]. NER (Named Entity Recognition) önceden belirlenmiş kategorilere göre metin içerisindeki kelimeleri bulmayı ve sınıflandırmayı amaçlayan ve bilgi çıkarımının ön koşulu olarak tanımlanan sistemlerdir (kişi isimleri, organizasyonlar, yerler vb.) [51]. Sistem sonuçlarının değerlendirilmesinde bilgi erişim sistemlerinde de olduğu gibi duyarlılık ve anma ölçütleri kullanılmaktadır. Fakat burada belgeler yerine, yapılan tahminler ölçüm değişkenleri olarak kullanılmaktadır. Duyarlılık, sistemin doğru yaptığı tahminlerin tüm tahminlere bölümü ile hesaplanmaktadır. Anma ise sistemin yaptığı doğru tahminlerin metinde bulunan bütün varlıkların sayısına bölünmesi ile elde edilmektedir.

Tıp alanında yapılandırılmamış metin verilerden bilgi çıkarma problemini çözmek için birçok sistem geliştirilmiştir. Johnson et al. [52] çalışmasında serbest

metin formatında bulunan radyoloji raporlarından bilgi çıkarmak ve yapılandırmak için tasarlanan RADA, bu sistemlerden biridir. RADA, DDİ tekniklerini kullanarak indeksleme ve görüntü veri tabanlarına erişim sağlama amacıyla serbest metin formatındaki açıklamaları yapılandırılmış bilgiye dönüştürmektedir. 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. Schadow and Mcdonald [53] serbest formatta bulunan cerrahi patoloji raporlarından numuneler ve numunelerle ilgili bulgular hakkındaki bilgileri çıkarmayı sağlayan bir sistem 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 metodu ile ilgili bilgiler elde edilmiştir. Biyomedikal alanda, bilgi çıkarma sistemleri ile ilgili var olan çoğu çalışmada özellikle gen ve proteinlerin tespit edilmesine odaklanılmıştır [51]. Tanabe et al. [54] 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.

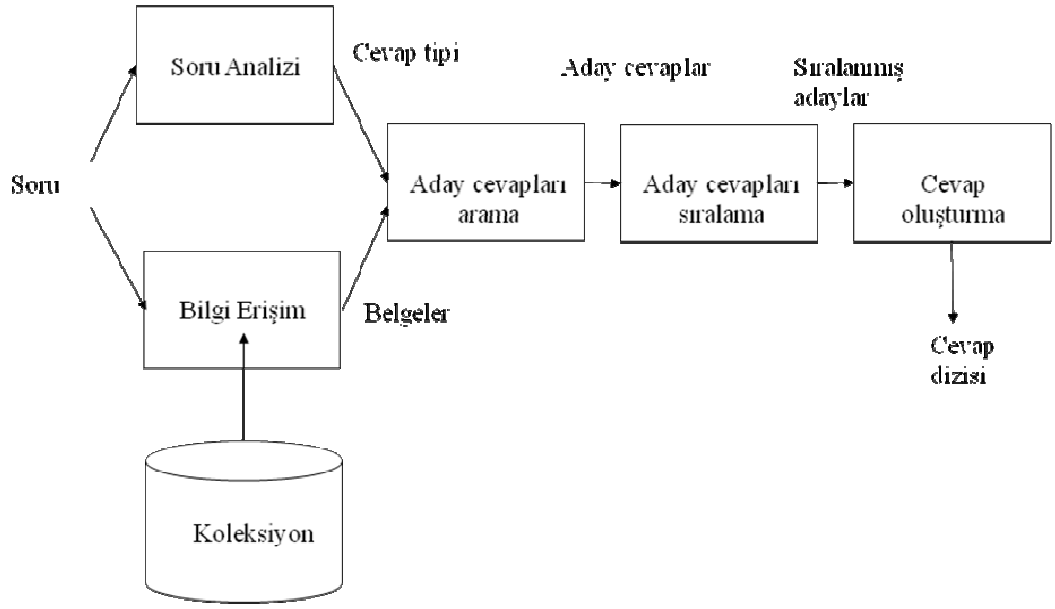
2.2.3.4. Soru Cevaplama Sistemleri

Soru cevaplama sistemleri, kullanıcısından sorusunu doğal dillerde kabul eden ve cevabı bulması için sayfa adresleri listesi yerine cevabın kendisini veren sistemlerdir. Bu sistemler bilgi erişim sistemi türlerinden biridir ve şu anki arama motoru teknolojilerinin bir adım ilerisinde gibi görülmektedir [51].

Şekil 2. 5.'de soru cevaplama sistemlerinin genel mimarisi gösterilmiştir. Sistem, öncelikli olarak doğal dille sorulan kullanıcı sorusunu almakta, analiz etmekte ve cevap tipleri (tarih, yer, isim, oran vb.) belirlemektedir. Aynı anda bir bilgi erişim sistemi kullanılarak soru cümlesindeki kelimeleri içeren belgelere erişilmektedir. Belirlenen cevap tipine uygun olarak erişilen belgelerden aday cevaplar aranmakta ve önceden belirlenen kurallara göre aday cevaplar puanlanıp sıralanmaktadır. Yüksek skorlu cevap kullanıcıya uygun bir arayüzle sunulmaktadır. Bilgi erişim sistemleri ile arasındaki fark; bilgi erişim sistemlerinde çıktı olarak doküman listesi sunulurken, soru cevaplama sistemlerinde cümlelerin veya kelimelerin cevap olarak sunulmasıdır.

Amasyalı [55], kullanıcının doğal dille sorduğu soruya cümle şeklinde cevap veren BayBilmiş adlı bir sistem geliştirmiştir. Bu sistem, bilgi erişim sistemi olarak Google arama motorunu kullanmaktadır. Sistem, cümlenin içerdiği sorgu kelimesi sayısı, cümle içindeki sorgu kelimelerinin birbirine yakınlığı ve cevap tipine uygunluk ölçütleri kullanarak aday cümlelere puan vermekte ve en yüksek puana sahip beş cümleyi kullanıcıya sunmaktadır. Ayrıca START [56], Ask.com [57], AnswerBus [58] vb. genel sorulara yönelik geliştirilmiş olan internet tabanlı sistemlere örnek verilebilir.

Tıbbi literatür, hasta bakımı ile ilgili en son bilgiler, tedavilerin yan etkileri, herhangi bir hastalığın semptomları vb. bilgileri içermesiyle hastalara uygulanacak tedavilere karar vermede klinisyenlere yardımcı olan önemli bir kaynaktır. Sadece hekimler değil diğer sağlık bakım uzmanları, hastalar, tıp öğrencileri ve araştırmacılar da sağlıkla ilgili konularda internet üzerinden aramalar yapabilmektedirler. Fakat şu anki sistemler, deneyimsiz araştırmacılar için yetersiz kalmakta ve sadece istedikleri cevabı sunamamaktadır. Bu sebeple tıp alanında da son yıllarda soru cevaplama sistemleri üzerine yoğun çalışmalar yapılmış ve çeşitli alanlarda sistemler geliştirilmiştir. Özellikle tıbbi literatür taramalarında kullanıcının doğal dille girdiği soruyu yapılandırarak ya da soru sormasını kolaylaştırarak kullanıcı emeğini en aza indirgeyen sistemler tasarlanmıştır. PICO (Problem/Population, Intervention, Comparison, and Outcome) bu alanda geliştirilen önemli sistemlerden biridir. Kullanıcı sorusunu formüle ederek iyi yapılandırılmış klinik sorgu oluşturulmasını sağlamaktadır [59]. Fontelo et al. [60] çalışmasında uzmanlaşmış kelime bilgisi gerektirmeyen, doğal dil kullanan herhangi bir kişiye MEDLINE/PubMed'de ilgili dokümanlara erişim sağlayan askMEDLINE adlı bir arama aracı geliştirilmiştir. Tıp alanında geliştirilen sistemlerden biri olan MedQA sistemi, biyomedikal alanda hekimlerin sorduğu belirli sorulara (tanım soruları) otomatik olarak kısa metin cevaplar üretmektedir [61, 62].



Şekil 2. 5. Soru Cevaplama Sistemlerinin Mimarisi

2.2.4. Tıpta Metin Madenciliği

Metin madenciliği tekniklerinin tıpta kullanımı son birkaç yılda büyük oranda artmıştır. Tıptaki verilerin genel olarak serbest metin formatında bulunması, hasta ile ilgili önemli bilgilerin gözden kaçmasına, bilgiye erişimin zorlaşmasına sebep olmaktadır. Yapılan klinik çalışmalar, araştırma raporları, hastane kayıtları, doktor notları, hasta formları ve faturalar tıptaki en önemli veri kaynaklarıdır. Bu verilerin çoğu serbest metin formatında bulunmaktadır [4]. Özellikle Elektronik Sağlık Kayıtları, Sağlık Bilgi Yönetiminin son yıllarda en önemli hedeflerinden birisiyken, böyle bir sistemin başarısının, klinik dokümantasyonun serbest metin formatında

yapılmasından dolayı sınırlanmış durumda olması bu tür sistemlere olan ihtiyacı ortaya çıkarmıştır.

Metin madenciliği, tıp alanında özellikle tıbbi araştırmalarda, semptomlarla hastalıklar ve ilaçlarla kimyasal maddeler arasında nedensel bağların bulunması, hasta kayıtlarının analiz edilmesi, gen-gen ve protein-protein ilişkilerinin tanımlanması, tanı ve tedavilerin geliştirilmesi, servis kalitesinin ve faydanın arttırılması ve maliyetlerin kontrol edilmesi gibi amaçlarla kullanılmaktadır [3]. Biyomedikal ve sağlık alanında, geliştirilen bilgisayar sistemlerinin, kullanılan dili anlayabilmesine olanak sağlayan, varlıklar arasındaki anlamsal ilişkilerin belirlenmesi üzerine geliştirilmiş kavramsal ontolojiler/bilgi tabanları bulunmaktadır. Genel olarak ontolojiler, kavramların sınıf-altsınıf ve parça-bütün ilişkilerine göre oluşturulmaktadır [28]. UMLS (Unified Medical Language System) yaygın olarak kullanılan, varlıklar arasındaki hiyerarşik ve anlamsal ilişkilerin tanımlandığı en önemli kavram dizinlerinden biridir. NLM (National Library of Medicine) tarafından geliştirilmektedir. Temel klinik kodlama ve referans sistemlerinin terminoloji, semantik ve formatları arasında bağlantılar kuran bir "metathesaurus" içeren bir sistemdir. Uzman bir "lexicon", bir "semantic" ağ ve bir enformasyon kaynakları haritalaması içeren UMLS, 1995 yılı itibarıyla 223.000 kavram ve 442.000 terim içermektedir ve bu sayı son yıllarda giderek artmaktadır [63]. Biyomedikal alanda ise genler ve proteinler arasındaki ilişkileri belirleyen GO (Gene Ontology) [64] yazılım geliştirmede sağladığı faydalarla yapılan çalışmalarda sıklıkla kullanılmaktadır. Bu tür sistemler özellikle metin madenciliği, doğal dil işleme vb. ile ilgili sistemlerin geliştirilmesinde büyük kolaylıklar sağlamaktadır.

Pubmed'de yapılan taramalar sonucu metin madenciliği ile ilgili toplamda 425 adet çalışma bulunmuştur. Yapılan çalışmalar incelendiğinde özellikle biyomedikal alanda gen-gen ve protein-protein ilişkilerinin saptanmasına ve diğer alanlarda ise hastalıklar ile semptomlar arasındaki nedensel bağların tespit edilmesine odaklanıldığı gözlemlenmiştir. Shatgay et al. [65] çalışmasında, DNA mikroarray deneylerinde genler arasındaki fonksiyonel ilişkilerin keşfedilmesi için biyomedikal yayınlar taranmış ve erişilen makalelerin özetlerinin içeriğine dayanarak ilişkiler bulunmaya çalışılmıştır. 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 [3]. 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 [66]. Swanson'ın sisteminde MEDLINE'da iki başlık kullanılarak arama yaptırılır (magnezyum ve migren) ve sonuçlar (başlıklar ve özetler), yaygın olarak bulunan önemli kelime ve kelime tamlamalarını 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 [67]. Lindsey and Gordon [68] Swanson'ın yaklaşımını, metin madenciliği olarak adlandırmadan genişletmişlerdir. 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. Garten and Altman [69], literatürden farmakogenomik kavramları otomatik olarak çıkartan ve gen-ilaç ilişkilerini sayısal bir forma dönüştüren Pharmspresso isimli bir metin analiz aracı geliştirmişlerdir. 178 adet gen adının geçtiği 45 makale kullanılarak yapılan değerlendirme çalışmasında, sistem genlerin %78'ini tanımlayabilmiştir. Sistemdeki en büyük eksiklik, sadece önceden tanımlanan tam metin makale koleksiyonlarında çalışıyor olmasıdır.

Biyomedikal alanda metin madenciliği ile ilgili özellikle NER alanında çalışmalar yapılmıştır. NER, metin koleksiyonlarında bulunan tüm isim varlıkların (ilaç isimleri, hastalık isimleri, simgeler) tanımlanması işlemidir. NER, metin içindeki varlıkların tanımlanmasına, varlıklar arasındaki ilişkilerin bulunmasına, anahtar kavramların tanımlanmasına ve bu kavramların uygun bir şekilde sunulmasına olanak sağlamaktadır [54]. Biyomedikal alanda NER ile ilgili yapılan çalışmalarda serbest metinlerdeki gen ve protein isimlerini otomatik olarak tanımaya odaklanılmıştır [54]. Hanisch çalışmasında gen ve protein isimlerinin yer aldığı ve kelimelerin anlamsal olarak sınıflandırıldığı geniş bir sözlük kullanmıştır [54]. Sistemin seçiciliği % 0.95, duyarlılığı % 0.90 olarak hesaplanmıştır. He et al. [70] PubMed'deki makale özetlerinden insan proteinleri arasındaki ilişkileri, birlikte bulunma durumlarına ve etkileşimde olduğu kelimelere bakarak çıkartan, internet tabanlı bir araç geliştirmişlerdir. Sistemin duyarlılığı %92, seçiciliği ise % 100 olarak bulunmuştur. Biyolojik varlıkların tüm tiplerinin tam olarak belirtildiği bir sözlük bulunmaması, isimlerinin çok kelimeli olabilmesi, aynı varlığın birden fazla isim alabilmesi vb. yaşanan problemler arasındadır [54].

Son yıllarda yabancı ülkelerde, veri madenciliğinin yanı sıra metin madenciliği de büyük bir ilgi görmekte ve bu alanda birçok sistem geliştirilmektedir. Fakat Türkiye'de bu konuya yeteri kadar önem verilmediği ve internet üzerinden yapılan aramalarda özellikle sağlıkta sadece birkaç çalışmanın yapıldığı gözlemlenmiştir. Bu yüzden geliştirilen sistemin faydalı olacağı düşünülmektedir.

GEREÇ VE YÖNTEM

Bu çalışmada, KBB uzman hekimleriyle yapılan görüşmeler sonucunda ellerinde bulunan ve hastalarla ilgili en kapsamlı veriye erişilebilecek belge olan hasta bilgi formlarının kullanılması kararlaştırılmıştır. Bu sebepten, öncelikli olarak KBB Anabilim Dalı'ndan Microsoft Office Word (doc uzantılı) formatında ve her biri yaklaşık olarak 40 Kb boyutunda, 2002-2007 yılları arasında gelen ameliyat geçiren hastalara ait 600 adet hasta bilgi formu alınmıştır. Daha hızlı işlem yapılabilmesi için bu belgeler Metin Belgesi (txt uzantılı) formatına dönüştürülmüştür. Bu 600 belgeden dokuz tanesinde birçok alan boş bırakıldığı için bu dokümanlar çalışmaya dahil edilmemiştir. Bu bölümde hasta bilgi formlarının hangi alanlardan oluştuğu, yazılımın geliştirilme aşamasında kullanılan sistemler ve teknikler, oluşturulan kelime listeleri ve süreç içerisinde izlenen adımlar anlatılmıştır.

3.1. Hasta Bilgi Formlarının İçerikleri

Hasta bilgi formlarından 100 tanesi örnek olarak seçilmiş ve içerikleri elle incelenmiştir. Yapılan bu inceleme sonucunda hasta bilgi formlarının beş ana bölümden oluştuğu gözlemlenmiştir. Bu ana bölümler ve alt bölümleri aşağıda sunulmuştur.

- Kişisel bilgiler
    - Ad-soyad
    - Yaş
    - Dosya No
    - Yapılan Ameliyat
    - Yatış Tarihi
    - Çıkış Tarihi
    - Mesleği
- Anamnez
    - Şikayet
    - Hikaye
    - Özgeçmiş
    - Soygeçmiş
    - Alkol
    - Sigara
- Fizik muayene
    - Orofarenks

- Rinoskopi Anterior
- Otoskopi
- Nazal Endoskopi
- Rinoskopi Posterior
- İndirekt Larengoskopi
- Baş-Boyun Muayenesi

- Laboratuvar incelemeleri;

  - Hemogram - Biyokimya
  - PA Akciğer Grafisi
  - Toraks BT
  - VLS Raporu
  - Batın USG
  - Tiroid USG
  - Temporal BT
  - Temporal CT
  - Temporal MRG
  - EMG
  - Boyun ve Maxillofasyal BT
  - Paranazal Sinüs Tomografisi

- Ameliyat Raporu;

  - Adı, Soyadı
  - Oda-Koğuş
  - Prot. No
  - Cinsiyeti
  - Yaşı
  - Kaçıncı Ameliyatı
  - Doktor
  - Ameliyat Tarihi
  - Ameliyat Öncesi Tanı
  - Ameliyat Sonu Tanı
  - Ameliyat Ekibi
  - Ameliyat Raporu
  - Klinik İzlem
  - Öneriler

Hasta bilgi formlarında bulunan bu alanlardan genel olarak hekimler tarafından doldurulmayan veya her hasta bilgi formunda bulunmayan alanlar çıkartılmış ve geriye kalan alanlar üzerinde işlemler yapılmıştır. Çalışmanın ilerleyen kısımlarında bu alanlarla ilgili daha ayrıntılı bilgi verilecektir.

3.2. Yazılım Geliştirme Sırasında Kullanılan Sistemler ve Teknikler

Yazılımın geliştirilmesi sırasında Visual Studio .NET platformu ve Visual C# .NET programlama dili kullanılmıştır. Visual Studio .NET bir yazılım geliştirme

ortamıdır. Ayrıca verilerin kaydedilmesi ve saklanmasında veritabanı ve veritabanı yönetim sistemi olarak SQL Server 2005 VTYS kullanılmıştır. SQL Server 2005 veritabanı yönetim sistemi çok kullanıcıyı destekleyen gelişmiş VTYS'ler arasında yer almaktadır [71].

Girdi metinlerinde bulunan yazım hatalarının tespit edilmesi, bu hataların düzeltilebilmesi ve metinlerin yapıtaşı olan kelimelerin gövdelerini elde etmek için açık kaynak kodlu bir doğal dil işleme yazılımı olan Zemberek [72] kullanılmıştır. Zemberek, Türkçe ve diğer Türkî diller için yazılmış, biçimbirimsel çözümleme, yazım denetimi, sözcük üretme gibi temel DDİ işlemlerini yapabilen açık kaynak kodlu DDİ kütüphanesidir. Zemberek altyapısının desteklediği DDİ işlemleri şunlardır:

- Yazım denetimi
- Biçimbirimsel çözümleme
- Gövdeleme
- Sözcük üretimi
- Sözcük önerme
- Sadece ASCII (American Standard Code for Information Interchange) karakterle yazılmış sözcükleri Türkçe karakterli hale çevirme
- Heceleme [73]

Yazılımın geliştirilmesi sırasında yazılım içinde kullanılan bazı sınıflar oluşturulmuştur. Bu sınıfların oluşturulmasındaki en büyük neden, yazılımın geliştirilmesi aşamasında programcılara kolaylıklar sağlamasıdır. Sınıflar kullanılarak sık kullanılan kodların tekrar tekrar programcı tarafından yazılması engellenmiştir. Oluşturulan sınıflardan ilki, Microsoft Office Excel sayfası formatında bulunan kelime listelerine yazılım tarafından erişimi ve bu dosyalardaki bilgileri okumasını sağlayan ExceldenVeriAl sınıfıdır. Ayrıca gerekli yerlerde elde edilen verilerin Microsoft Office Excel dosyası formatında çıktısını alabilmek için ExceleVeriAt sınıfı oluşturulmuştur. Diğer bir sınıf ise FrekansHesaplamadır. Bu sınıf, kullanıcıların ve sistem geliştiricilerin belli aşamalarda ihtiyaç duydukları kelime frekanslarına erişimini sağlayabilmek için geliştirilmiştir. Aşağıda örnek teşkil etmesi amacı ile ExceldenVeriAl sınıfı gösterilmiştir.

```
using System;
using System.Collections.Generic;
using System.Text;
using System.Data;
using System.Data.OleDb;
using System.Data.SqlClient;
namespace dosyadanokuma
{
    class ExceldenVeriAl
    {
```

```
        public DataSet GetData(string fileName)
        {
            OleDbConnection conn = new OleDbConnection();
            DataSet ds = new DataSet();
            OleDbDataAdapter da = new OleDbDataAdapter();
            conn = SetConnection(fileName);
       da.SelectCommand = new OleDbCommand("Select * from
      [Sayfa1$]", conn);
            da.Fill(ds,"filename");
            if (conn.State != ConnectionState.Closed)
               conn.Close();
            return ds;
        }
        private OleDbConnection SetConnection(string filePath)
        {
            string _cnString =
@"Provider=Microsoft.Jet.OLEDB.4.0;Data Source=" + filePath +
@";Extended Properties='Excel 8.0;HDR=Yes;IMEX=1'";
            return new OleDbConnection(_cnString);
        }
 }
```

Yukarıda gösterilen sınıf içerisinde yer alan kodlar, yazılım içinde gerekli olan yerlerde tekrar yazılmamakta ve sadece aşağıdaki kodlar yazılarak kullanılmaktadır. Böylelikle yazılımcı hem zaman kazanmakta hem de hata yapma olasılığını minimuma indirmektedir.

```
string pathOfExcelFile
=System.Windows.Forms.Application.StartupPath +
"\\duzeltme.xls";
ExceldenVeriAl obj = new ExceldenVeriAl();
DataSet ds = obj.GetData(pathOfExcelFile);
```

Ayrıca sınıflar dışında yazılım içerisinde sınıf yapısına benzer olan prosedürler kullanılmıştır. Prosedürler “void” ile deklare edilen kod bloklarından oluşmakta [74] ve hem sınıf içerisinde hem de yazılım içinde aynı kodun tekrarlı olarak yazılmasını engellemek amacıyla kullanılabilmektedir. Aşağıda, yazılım geliştirilirken kullanılan prosedürlerden “xmlkaydet” örnek olarak gösterilmiştir. Bu prosedürle birlikte tablodaki veriler XML formatına dönüştürülüp veritabanına kaydedilmektedir.

```
private void xmlkaydet()
{
     StringBuilder sb = new StringBuilder();
     StringWriter sw = new StringWriter(sb);
     dt.TableName = "Hastalar";
     dt.WriteXml(sw, XmlWriteMode.IgnoreSchema);
     SqlConnection mssqlcon = new SqlConnection();
     mssqlcon.ConnectionString = "Data Source=xxx; Initial
Catalog=yyy;User Id=vvv;Password=zzz";
     mssqlcon.Open();
     SqlCommand cmd = new SqlCommand("insert into xmldeneme
(XMLData) values (@P1)", mssqlcon);
     cmd.Parameters.Add("@P1", SqlDbType.Xml);
     cmd.Parameters["@P1"].Value = sb.ToString();
     cmd.ExecuteNonQuery();
     mssqlcon.Close();
     MessageBox.Show("İşlem Tamamlanmıştır", "Dikkat!",
MessageBoxButtons.OK, MessageBoxIcon.Information);
}
```

Yazılım geliştirilirken kullanıcı arayüzlerinin tasarlanmasında Visual Studio .NET platformunun sunduğu kontroller kullanılmıştır. “dataGridView”, “checkedListBox” ve “Button” kullanılan kontrollerden bazılarıdır.

3.3. Oluşturulan Kelime Listeleri

Hasta bilgi formlarından örnek olarak seçilen 100 adet belge incelendiğinde bu belgelerin “Şikayet”, “Yaş” gibi alanlardan ve bu alanların içerisinde bulunan hastalara ait bilgilerden oluştuğu gözlemlenmiştir. Yazılım geliştirilirken ilk amaç bu metinler üzerinde işlem kolaylığı sağlayabilmek ve daha kapsamlı analizler yapabilmek için bu formları veri tablosu haline dönüştürmekti. Bu yüzden alan isimleri sütuna ve içerikleri satırlara gelecek şekilde bir tablo oluşturabilmek için alan isimlerinin bulunduğu bir liste oluşturulmuştur. Listede geçen alan isimleri, yazılım tarafından metinlerde bulunarak alanlar ve içerikleri matris diziye atılmakta ve daha sonra oluşturulan bu iki boyutlu dizi veri tablosuna dönüştürülmektedir. Fakat bu alan isimleri her belgede aynı şekilde yazılmadığı için bunların standart bir forma dönüştürülmesi gerekliliği ortaya çıkmıştır. Bu yüzden alan isimlerinin tüm yazılış şekillerinin ve buna karşılık gelen standart formunun bulunduğu bir düzeltme listesi oluşturulmuştur. Yazılım çalışmaya başladığında öncelikli olarak bu düzeltme listesindeki kelimeleri metinler içerisinde bularak bunları listede karşılık gelen standart formlarına dönüştürmektedir. Daha sonra alan isimlerinin bulunduğu listeyi kullanarak metinleri alanlar ve içerikleri olarak ayırmakta ve tablo haline dönüştürüp kullanıcıya sunmaktadır. Çizelge 3. 1.’de düzeltme listesinden örnekler verilmiştir.

Çizelge 3. 1. Düzeltme Listesi

Yazım Türleri	Düzeltilen Kelime
Hikaye	Hikayesi
Hikayesi	Hikayesi
Orofareks	Orofarenks
Orofarenks	Orofarenks
Özgeçmiş	Özgeçmiş
Özgeçmişi	Özgeçmiş
Şikayet	Şikayeti
Şikayeti	Şikayeti
Yaş	Yaş
Yaşı	Yaş
Rinoskopi Anteror	Rinoskopi Anterior
Rinoskopianterior	Rinoskopi Anterior

Daha öncede bahsedildiği gibi Zemberek Kütüphanesi kullanılarak bir yazım denetimi modülü geliştirilmiştir. Tüm metinlerde yazım denetimi yapıldıktan sonra en sık rastlanan hata türleri aşağıda sunulmuştur;

- Yanlış yazılan karakter:
  Burun yerine *butun*
- Yer değiştiren komşu karakterler:
  Boyun yerine *bouyn*
- Düşen karakter:
  Ağızda yerine *ağızd*
- Eklenen karakter:
  Örnek yerine *örnerk*
- Bitişik yazılan kelimeler:
  Tarafta şişlik yerine *taraftaşişlik*

Yazım denetimi sonucunda Zemberek hatalı bulduğu kelimeler için yukarıda bahsedilen hata türlerine göre bir öneri listesi sunmaktadır. Çizelge 3. 2.'de önerilerin ve hata türünün belirtildiği bir liste örnek olarak gösterilmiştir. Yazım denetimi modülü, elde bulunan tüm belgeler programa yüklü iken çalıştırılmış, çıktı olarak elde edilen liste yazılım geliştiriciler tarafından incelenmiş ve doğru bulunan sonuçlardan uygun öneriler seçilerek düzeltme listesine eklenmiştir.

Çizelge 3. 2. Sık Karşılaşılan Yazım Hatası Türleri

Hatalı Kelime	Hata Türü	Öneriler
Ğüçlüğü	Yanlış yazılan karakter	/ Güçlüğü / güçlüğü
Ak8ntı	Yanlış yazılan karakter	/ Aktı / Akıntı
Altındaşişlik	Bitişik yazılan kelimeler	/ altında şişlik / Altında şişlik
Buruntıkanıklığı	Bitişik yazılan kelimeler	/ burun tıkanıklığı
Boyundaşişlik	Bitişik yazılan kelimeler	/ boyunda şişlik
Biopsi	Düşen karakter	/ Biyopsi
Boynda	Düşen karakter	/ Boyunda / Boyna / Boyda / Boyada
Dönmesiş	Eklenen karakter	/ Dönmesi / Dönmesin / Dönmesiz
Kulakata	Eklenen karakter	/ Kulakta / Kulaklata / Kulalata
Maliğn	Eklenen karakter	/ Malin
Şişilk	Yer değiştiren komşu karakterler	/ Şişlik / Şişilik / şiş ilk

Hasta bilgi formları incelendiğinde bazı alanların genellikle hekimler tarafından doldurulduğu bazı alanların ise boş bırakıldığı tespit edilmiştir. Bu sebeple "önemli alanlar" adı altında başka bir liste oluşturularak bu listedeki alanlar dışında kalan tüm alanların ve içeriklerinin yazılım tarafından silinmesi sağlanmıştır (Çizelge 3. 3.).

Çizelge 3. 3. Önemli Alanlar

Alan Adı
Dosya No
Ad-Soyad
Yaş
Şikayeti
Cinsiyeti
Özgeçmiş
Rinoskopianterior
Orofarenks
Otoskopi
Boyunmuayenesi
Rinoskopiposterior
Ameliyatöncesitanı
Ameliyatsonutanı

Ayrıca girdi metinlerinin boyutunu ve işlem yapılırken gereksiz kelimeler için harcanan zamanı azaltmak için internetten elde edilmiş "Türkçede sık kullanılan kelimeler" listesi [75, 76] kullanılarak girdi metinleri içerisinden bu kelimelerin çıkartılması sağlanmıştır. Bu liste toplamda 275 kelime içermektedir. Örnek bir liste Çizelge 3. 4.'te sunulmuştur.

Çizelge 3. 4. Sık Kullanılan Kelimeler

Sık Kullanılan Kelimeler
acaba
ama
amacıyla
ancak
aslında
bazı
belki
biri
birkaç
bir şey
biz
bu

Bir sonraki aşama önemli/anahtar kelimelerin belirlenmesi ve bu kelimelerin yazılım tarafından metinlerde etiketlenmesidir. Bu amaçla her alana özgü anahtar kelime listeleri oluşturulmuştur. Bu listeler oluşturulurken anahtar kelimelerin belirlenmesinde kelime grupları frekansları kullanılmıştır. Bu frekansların hesaplanması için yazılım içerisinde özel bir modül geliştirilmiştir. Bu modül sadece yazılımcılar tarafından, listeleri oluşturabilmek amacıyla kullanılmıştır. Metinler yazılım tarafından okunup ham haliyle veri tablosu haline dönüştürüldükten sonra tüm alanlardaki kelimeler tekli, ikili ve üçlü olmak üzere kelime ve kelime gruplarına ayrılmış ve frekansları hesaplanmıştır. İşlemin sonunda elde edilen sonuçlar yazılım tarafından Microsoft Office Excel'e aktarılmış ve elle yapılan incelemeler sonucunda her alana özgü listeler oluşturulmuştur. Kelimelerin tekli, ikili vb. kelime ve kelime gruplarına ya da harf ve harf gruplarına ayrılarak dizilişlerine bakılması ve örüntülerin çıkartılması literatürde "N-gram" yöntemi olarak geçmektedir [77-79]. Yöntem uygulanırken, ikili kelime gruplarında kelimenin kendisinden bir önce geçen kelimeyle (n-1, n), üçlü kelime gruplarında ise kelimenin kendisinden bir ve iki önce geçen kelimelerle birlikte bulunma frekansları hesaplanmıştır. Bu frekanslara bakılarak her alana özgü anahtar kelime listeleri oluşturulmuştur. Böylelikle sadece tek başına bir anlam ifade eden kelimeler değil, ikili ya da üçlü kelime grupları şeklinde anlam ifade eden kelimeler de oluşturulan listelere eklenmiştir. Modül 100 metin için çalıştırıldığında Çizelge 3. 5.'te verilen sonuçlar elde edilmektedir. "Şikayeti" alanı örnek olarak verilmiştir.

Çizelge 3. 5. Kelime ve Kelime Grupları Frekans Sonuçları

Tekli Kelimeler		İkili Kelimeler		Üçlü Kelimeler	
Kelime	**Frekans**	**Kelime**	**Frekans**	**Kelime**	**Frekans**
Ses	15	Ses Kısıklık	15	Ağız Açık Uyuma	4
Sol	14	Boyun Şişlik	8	Sağ Kulak Akıntı	3
Sağ	11	İşitme Azlık	5	Horlama İşitme Azlık	3
Kulak	10	Burun Tıkanıklık	4	Sık Boğaz Enfeksiyon	2
Boyun	10	Nefes Darlık	4	Boyun Kitle Akıntı	2
Kısıklık	6	Solunum Sıkıntı	4	Kulak Arka Şişlik	2
Akıntı	4	Kulak Akıntı	3	Kulak İşitme Kayıp	2

Çizelge 3. 6.'da oluşturulan anahtar kelime listelerinden bazı örnekler verilmiştir.

Çizelge 3. 6. Anahtar Kelime Listesi

Lokasyon	Şikayet	Ameliyat Öncesi Tanı
Ağız	Ağrı	Kitle
Baş	Akıntı	Deviasyon
Bel	Ateş	Hipertrofi
Boğaz	Bulantı	Vejetasyon
Burun	Çıkarma	Fistül
Çene	Darlık	Otitis
Damak	Dolgunluk	Tümör
Dil	Enfeksiyon	Osas*
Dudak	Felç	Tonsilit
Geniz	Horlama	Sinonazalpolipozis
Kan	Kanama	Karsinom
Kulak	Kısıklık	Schwannom
Larenks	Kusma	Teratom
Özafagus	Öksürük	Sinüzit
Tonsil	Şişlik	Lipom
Yanak	Tıkanıklık	Lap
Yüz	Yara	Periferikfasialparaliz
Cilt	Kızarıklık	Osteom
Göz	Çiftgörme	Travmatikfasialparaliz

*Obstrüktif Uyku Apne Sendromu

"Özgeçmiş" alanında hastanın sigara-alkol kullanıp kullanmadığı, kronik hastalıkları ve geçirdiği ameliyatlar ile ilgili bilgiler bulunmaktadır. Hastanın sigara ve alkol kullanıp kullanmadığını belirlemek için "Özgeçmiş" alanında bazı anahtar kelimeleri içerip içermediğine bakılmıştır. Bu kelimeler belirlenirken, sigara ve alkol kullanan hastaları tanımlamak için hangi terimlerin ne kadar sıklıkla kullanıldığına bakılmıştır. Sigara kullanan hastaları tanımlamak için genellikle "Paket", "Adet", "Sigara" vb., alkol kullanan hastalar için ise "Alkol", "Kadeh" vb. kelimelerin

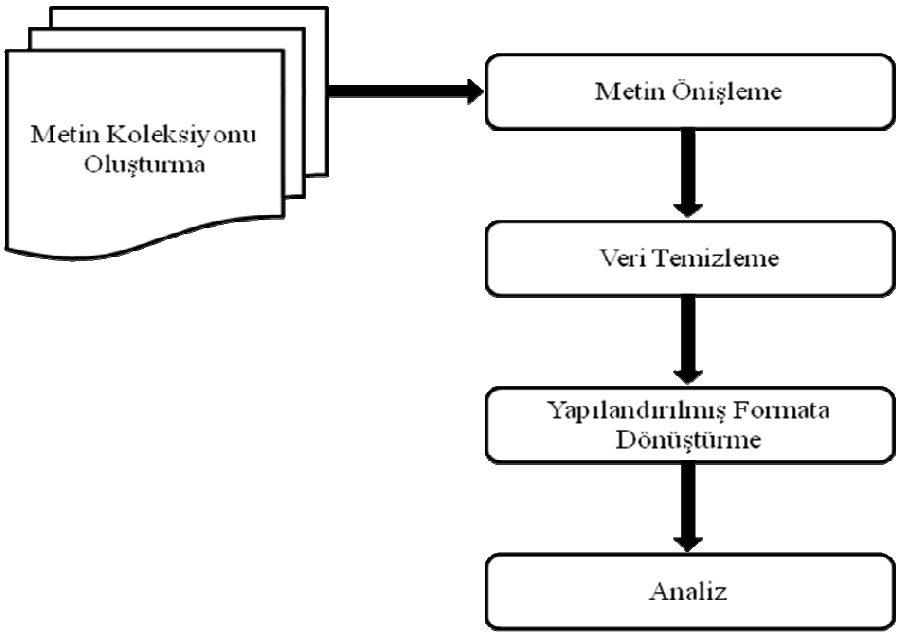
kullanıldığı gözlemlenmiştir. Eğer hastanın özgeçmiş bilgisinde bu kelimeler geçiyorsa yazılım bu kelimeleri "Sigara-Var" veya "Alkol-Var" olarak değiştirmekte, eğer bu kelimelerden herhangi birini içermiyorsa "Özellik-Yok" yazmaktadır. Hastanın geçirdiği ameliyatlar ve kronik hastalıklarının belirlenebilmesi için diğer alanlarda olduğu gibi hem geçirdiği ameliyatlara hem de kronik hastalıklara ait kelime ve kelime gruplarını içeren iki adet anahtar kelime listesi oluşturulmuştur.

3.4. Geliştirilen Yazılım Tarafından İzlenen Adımlar

Yazılımda, hem programcının yazılımı geliştirirken kullanması gereken hem de kullanıcı için tasarlanmış olan özellikler bulunmaktadır. Şekil 3. 1.'de yazılım geliştirme sürecinde izlenen adımlar gösterilmiştir. Geliştirilen yazılım tarafından yapılan işlemler temel olarak dört ana adımdan oluşmaktadır.

3.4.1. Metin Önişleme

Bu adımda girdi metinleri bazı önişlemlere tabi tutulmaktadır. İlk olarak, metinlerde bulunan birden fazla boşluklar ve noktalama işaretleri kaldırılmaktadır. Daha sonra metinler boşluk kullanılarak kelimelere ayrılmaktadır. Böylelikle her bir kelime ayrı ayrı incelenebilmektedir. Metinleri veri tablosu haline dönüştürebilmek için sütun isimlerinin tanımlanması gerekmektedir. Yazılım bu işlemi alan isimlerinin bulunduğu listeyi kullanarak yapmaktadır. Sütun isimleri olarak belirlenen şikayet, hikaye vb. alanlar metinler içerisinde bulunarak sütun adı olarak belirlenmekte ve bu alanlara ait içerikler de satırlara gelecek şekilde veri tablosuna aktarılmaktadır. Bu işlem yapılmadan önce tüm metinler elle incelenmiş ve daha verimli bir sonuç alabilmek için alan isimlerinde bulunan farklı yazılış şekilleri, oluşturulan liste kullanılarak standart formata dönüştürülmüştür.



Şekil 3. 1. Süreç İçerisinde Gerçekleştirilen Adımlar

Zemberek adlı açık kaynak kodlu doğal dil işleme yazılımı kullanılarak metinler içerisindeki yazım hataları düzeltilmiş ve tüm kelimeler üzerinde gövdeleme işlemi yapılarak gövde haline dönüştürülmüştür. Türkçede kök sözcüğe yapım ekleri

getirilerek türetilen yeni sözcüklere gövde denir. Bir sözcüğün eklenmiş çekim eklerinden arındırılarak gövde veya kökün bulunması işlemine *gövdeleme* adı verilir [73]. Aşağıda, metinlerde uygulanan gövdeleme işlemine örnek verilmiştir.

Örneğin; Dilde=Dil, Kandaki=Kan

Böylelikle program aynı kökten gelen fakat çekim eki almış kelimeleri farklı bir kelime olarak algılamamaktadır. Bu işlemle birlikte, metinlerin boyutu azalmakta, gereksiz yere işlem süresi uzatılmamakta, aynı kelimeler yazılım tarafından farklı kelime olarak algılanmamakta ve daha doğru frekans sonuçları elde edilebilmektedir.

3.4.2. Veri Temizleme

İkinci adımda girdi metinleri içerisinde bazı temizleme işlemlerinden geçmekte ve sadece ihtiyaç duyulan verilere indirgenmektedir. Bu temizleme işlemlerinden birincisi sık kullanılan gereksiz kelimelerin metinlerden çıkartılmasıdır. Bu işlem, hiçbir anlam ifade etmeyen gereksiz kelimeler üzerinde işlem yapılmasını önlemekte, metinlerin boyutunu azaltmakta ve işlemler yapılırken harcanan zamandan tasarruf sağlamaktadır. İkinci işlemde, genellikle bütün hasta bilgi formlarında bulunan alanların isimlerinin bulunduğu Önemli Alanlar listesi kullanılarak bu listede adı geçmeyen sütunlar program tarafından silinmektedir. Üçüncü olarak da her alana özgü oluşturulan anahtar kelime listeleri kullanılarak bu listelerde bulunan kelimeler veya kelime grupları metin içerisinde etiketlenmekte ve bu kelime ve kelime grupları dışındaki kelimeler metinlerden çıkartılmaktadır. Bu işlem sırasında anahtar kelimeler hem kod hem de orijinal hallerinde etiketlenebilmektedir.

3.4.3. Yapılandırılmış Formata Dönüştürme

Dördüncü adımda tüm alanlardaki anahtar kelimeler etiketlendikten sonra elde edilen veriler veri tabanına XML (eXtensible Markup Language) formatında atılmaktadır. XML, hem insanlar hem bilgi işlem sistemleri tarafından kolayca okunabilecek dokümanlar oluşturmaya yarayan, W3C (World Wide Web Consortium) tarafından tanımlanmış bir standarttır [80]. Kişilerin kendi sistemlerini oluşturabilecekleri, kendi etiketlerini tanımlayarak çok daha rahat ve etkin programlama yapabilecekleri ve bu belirlenen etiketleri kendi yapıları içerisinde standardize edebilecekleri esnek, genişleyebilir ve kolay uygulanabilir bir meta dildir [81]. Bu özelliği ile veri saklamanın yanında farklı sistemler arasında veri alışverişi yapmaya yarayan bir ara format görevi de görür. XML dokümanları ağaç veri yapısında olurlar. Bağımsız imler yapıyı oluştururken, içerik ya imin özelliği olarak ya da iki im arasında gösterilir Yapıyla ilgili ayrıntılar DTD (Document Type Definition) ya da XML Schema adı verilen harici dokümanlar ile tanımlanır. Aşağıdaki örnek, bir XML dokümanında verinin nasıl belirtildiğini göstermektedir [80].

```
<kullanicilar>
  <kullanici id="1">
      <ad>A</ad>
      <soyad>B</soyad>
  </kullanici>
  <kullanici id="2">
      <ad>C</ad>
      <soyad>D</soyad>
  </kullanici>
  <kullanici id="5">
      <ad>E</ad>
      <soyad>F</soyad>
  </kullanici>
  <kullanici id="8">
      <ad>G</ad>
      <soyad>H</soyad>
  </kullanici>
</kullanicilar>
```

Çok farklı tipteki verileri orijinal formatlarında tek bir havuzda tutabilen XML, bilgiye hızlı, kolay ve ortamdan bağımsız olarak erişebilme imkânı sunar. Günlük yaşantımızda kullanmakta olduğumuz verilerin % 90'ını oluşturan ve "yapılandırılmamış" olma özellikleri nedeniyle kendi bulundukları medya dışında veri özelliklerini koruyamayan (kelime işlem, elektronik tablo çıktıları, PDF dokümanları, ses, resim vb.) farklı tipteki verilerin, uyuma gerek duymadan hiyerarşik bir yapıda kullanılabilmelerine olanak vermekte ve bu verilerin hızlı bir şekilde sorgulanabilmelerini sağlamaktadır. Öncelikle veri transferinin kolaylaşmasını ve verinin içerik bilgisiyle saklanabilmesini hedefleyen XML, içerik ve sunum bilgilerini birbirinden ayırır. XML ile ilgili yapılan tanımlamalardan aşağıdaki çıkarımları elde edebiliriz [82].

- XML bir belgenin yapısını ve görünümünü tanımlamak için kullanılan uluslararası bir standarttır.
- XML yapılandırılmış belge ve verilerin evrensel formatıdır.
- XML metin tabanlı işaretleme dilidir ve veri alış verişinde kullanılan bir standarttır.
- XML bilginin yapısını tanımlamak için kullanılan bir teknolojidir.
- XML bilgiyi tanımlayan ve Web'te bilgi alış verişi için kullanılan standart bir biçimdir.
- XML markup dillerini tanımlayan bir meta dilidir.
- XML verinin yapılandırılması ve tanımlanması için kullanılan bir teknolojidir.
- XML herhangi bir verinin biçimlenmesi, tanımlanması için kullanılan bir teknolojidir [82].

Yukarıdaki tanımlar incelendiğinde bazı tanımlarda sadece XML teknolojisinin tanımlandığı bazılarında ise dil olarak XML'in tanımlandığı görülmektedir. Buna göre aşağıdaki sonuçlar çıkartılabilir.

- XML hem bir teknolojidir hem de bir dildir.
- XML dil olarak markup dil'leri (bir belgedeki verileri işaretlemeye yarayan diller) oluşturmaya yarar.
- XML verileri tanımlamak için kullanılan bir teknolojidir.
- XML verileri tanımlama açısından bir standart oluşturmak için oluşturulmuştur.
- XML verileri standart bir şekilde tanımladığından Web'te veya herhangi iki program arasında veri alışverişini kolaylaştırmaktadır [81].

Son yıllarda XML işaretleme standardı büyük bir ilgi görmüş ve özellikle gelişmiş ülkelerde birçok kurumda olduğu gibi sağlık kurum ve kuruluşları arası veri iletişiminde de yaygın olarak kullanılmaya başlanmıştır. XML'in veri iletişiminde bu kadar önemli role sahip bir teknoloji olmasından dolayı, bu çalışmada tüm metinlerin XML formatına dönüştürülerek veritabanına kaydedilmesi kararlaştırılmıştır. Hasta bilgi formları ilk olarak veri tablosuna dönüştürülmekte, daha sonra istenildiği takdirde XML olarak veritabanına kaydedilebilmektedir. Bulgular bölümünde elde edilen bir XML çıktısı örnek olarak verilmiştir.

3.4.4. Elde Edilen Verilerin Analiz Edilmesi

Beşinci ve son adımda ise metinlerin içerisindeki anahtar kelime ve kelime gruplarının EK-2'de verilen örnek tablodaki gibi kodlanmış hale dönüştürüldükten sonra elde edilen kodlanmış verilerin veri madenciliği yöntemlerinden Birliktelik Kuralı ile analiz edilmesi işlemi yapılmaktadır. Bu işlem için yazılım içerisinde, yöntemin temel algoritmasını kullanarak kuralların destek, güven ve lift değerlerini hesaplayan özel bir form tasarlanmıştır. Yöntem, sıklıkla girilmiş olan ve analiz için elverişli veriyi içeren "Şikayet", "Yaş", "Cinsiyet", "Özgeçmiş" ve "Ameliyat Öncesi Tanı" alanlarındaki veriler üzerinde uygulanmış ve belirlenen minimum destek ve güven değerlerine göre kurallar elde edilmiştir. Bunun dışında, alanlar içerisinde geçen anahtar kelime ve kelime gruplarının frekansları da hesaplanmıştır. Analiz sonuçları ile ilgili ayrıntılı bilgi Bulgular bölümünde sunulmuştur.

BULGULAR

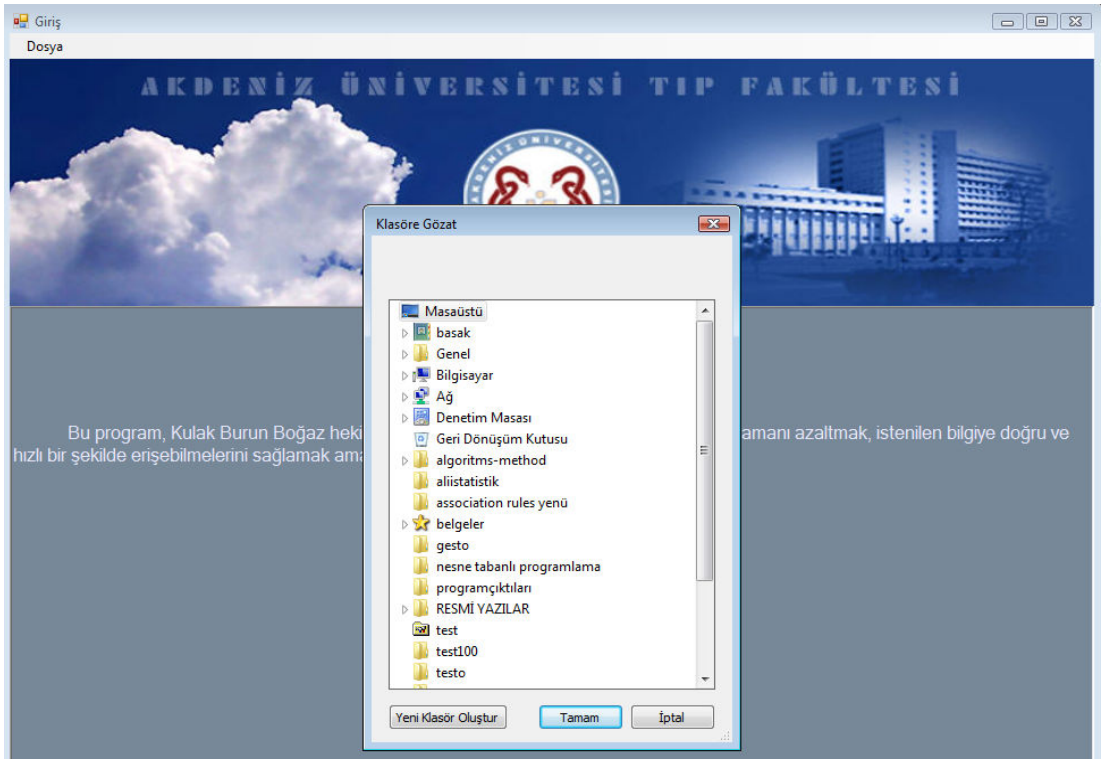
Bu bölümde KBB hekimlerine hasta bilgi formlarından hasta ile ilgili istedikleri verilere erişimlerini ve elde edilen bu verilerin analiz edilmesini sağlamak amacıyla geliştirilen yazılım anlatılmıştır. Bölüm, iki kısımdan oluşmaktadır. İlk kısımda yazılımla ilgili, ikinci kısımda ise analizden elde edilen sonuçlarla ilgili bilgiler verilmiştir.

4.1. Sistem Çıktıları

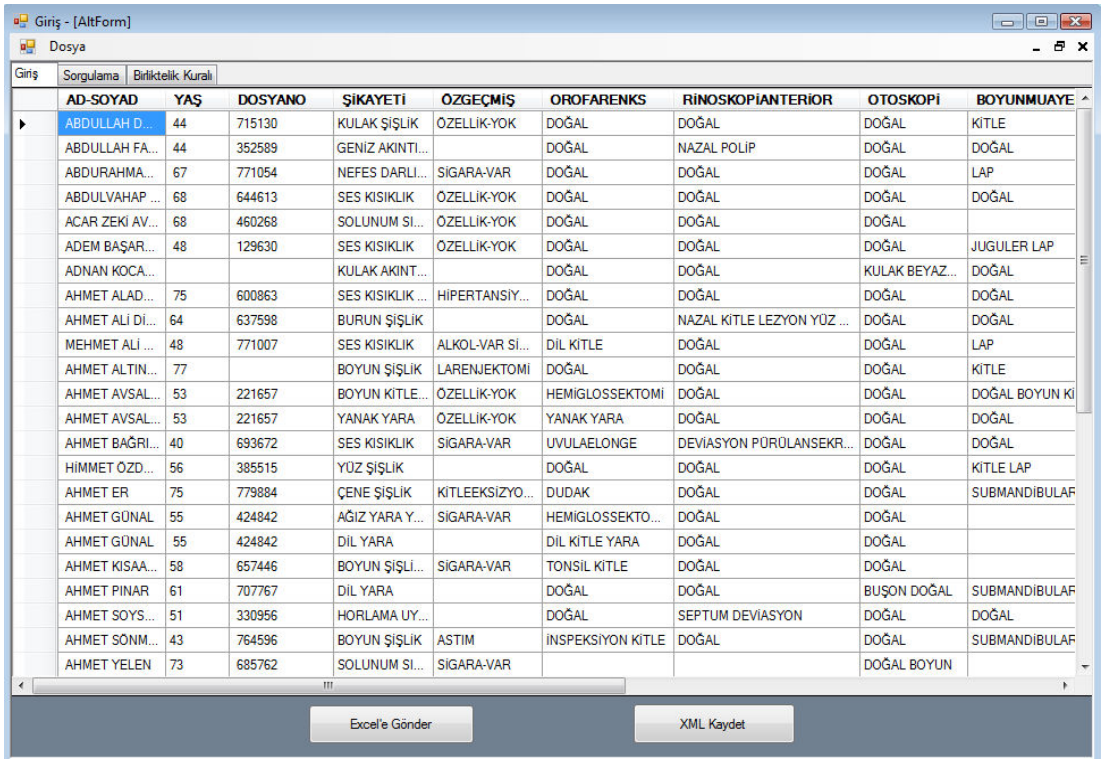
Yazılım ilk çalışmaya başladığı anda kullanıcı, Şekil 4. 1.'deki gibi bir ekran görüntüsüyle karşı karşıya gelmektedir. Kullanıcılar, Şekil 4. 2'de gösterildiği üzere Giriş ekranında bulunan "Dosya" menüsünden "Yeni" sekmesine tıkladıklarında işlem yapmak istedikleri metinlerin bulunduğu klasörün seçilebileceği "Klasöre Gözat" diyalogu açılmaktadır (Şekil 4. 3.). Açılmak istenen klasör seçilerek "Tamam" butonuna tıklanılmakta ve klasörde bulunan tüm metinler yazılım tarafından okunup yüklenmektedir. Yazılım hem metin belgesini (txt uzantılı) hem de Word belgesini (doc veya docx uzantılı) okuyup açabilmektedir. Bu işlem sonucunda, Şekil 4. 4.'deki gibi metinler, Şikayet, Yaş vb. alanlar sütuna, her alanın içerikleri satırlara gelecek şekilde veri tablosu haline dönüştürülüp kullanıcıya sunulmaktadır.

Şekil 4. 1. Giriş Ekranı

Şekil 4. 2. Yeni Dosya Ekleme



Şekil 4. 3. Klasöre Gözat Diyalogu



AD-SOYAD	YAŞ	DOSYANO	ŞİKAYETİ	ÖZGEÇMİŞ	OROFARENKS	RİNOSKOPİANTERİOR	OTOSKOPİ	BOYUNMUAYE
ABDULLAH D...	44	715130	KULAK ŞİŞLİK	ÖZELLİK-YOK	DOĞAL	DOĞAL	DOĞAL	KİTLE
ABDULLAH FA...	44	352589	GENİZ AKINTI...		DOĞAL	NAZAL POLİP	DOĞAL	DOĞAL
ABDURAHMA...	67	771054	NEFES DARLI...	SİGARA-VAR	DOĞAL	DOĞAL	DOĞAL	LAP
ABDULVAHAP ...	68	644613	SES KISIKLIK	ÖZELLİK-YOK	DOĞAL	DOĞAL	DOĞAL	DOĞAL
ACAR ZEKİ AV...	68	460268	SOLUNUM SI...	ÖZELLİK-YOK	DOĞAL	DOĞAL	DOĞAL	
ADEM BAŞAR...	48	129630	SES KISIKLIK	ÖZELLİK-YOK	DOĞAL	DOĞAL	DOĞAL	JUGULER LAP
ADNAN KOCA...			KULAK AKINT...		DOĞAL	DOĞAL	KULAK BEYAZ...	DOĞAL
AHMET ALAD...	75	600863	SES KISIKLIK ...	HİPERTANSİY...	DOĞAL	DOĞAL	DOĞAL	DOĞAL
AHMET ALİ Dİ...	64	637598	BURUN ŞİŞLİK		DOĞAL	NAZAL KİTLE LEZYON YÜZ ...	DOĞAL	DOĞAL
MEHMET ALİ ...	48	771007	SES KISIKLIK	ALKOL-VAR Sİ...	DİL KİTLE	DOĞAL	DOĞAL	LAP
AHMET ALTIN...	77		BOYUN ŞİŞLİK	LARENJEKTOMİ	DOĞAL	DOĞAL	DOĞAL	KİTLE
AHMET AVSAL...	53	221657	BOYUN KİTLE...	ÖZELLİK-YOK	HEMİGLOSSEKTOMİ	DOĞAL	DOĞAL	DOĞAL BOYUN Kİ
AHMET AVSAL...	53	221657	YANAK YARA	ÖZELLİK-YOK	YANAK YARA	DOĞAL	DOĞAL	DOĞAL
AHMET BAĞRI...	40	693672	SES KISIKLIK	SİGARA-VAR	UVULAELONGE	DEVİASYON PÜRÜLANSEKR...	DOĞAL	DOĞAL
HİMMET ÖZD...	56	385515	YÜZ ŞİŞLİK		DOĞAL	DOĞAL	DOĞAL	KİTLE LAP
AHMET ER	75	779884	ÇENE ŞİŞLİK	KİTLEEKSİZYO...	DUDAK	DOĞAL	DOĞAL	SUBMANDİBULAR
AHMET GÜNAL	55	424842	AĞIZ YARA Y...	SİGARA-VAR	HEMİGLOSSEKTO...	DOĞAL	DOĞAL	
AHMET GÜNAL	55	424842	DİL YARA		DİL KİTLE YARA	DOĞAL	DOĞAL	
AHMET KISAA...	58	657446	BOYUN ŞİŞLİ...	SİGARA-VAR	TONSİL KİTLE	DOĞAL	DOĞAL	
AHMET PINAR	61	707767	DİL YARA		DOĞAL	DOĞAL	BUŞON DOĞAL	SUBMANDİBULAR
AHMET SOYS...	51	330956	HORLAMA UY...		DOĞAL	SEPTUM DEVİASYON	DOĞAL	DOĞAL
AHMET SÖNM...	43	764596	BOYUN ŞİŞLİK	ASTIM	İNSPEKSİYON KİTLE	DOĞAL	DOĞAL	SUBMANDİBULAR
AHMET YELEN	73	685762	SOLUNUM SI...	SİGARA-VAR			DOĞAL BOYUN	

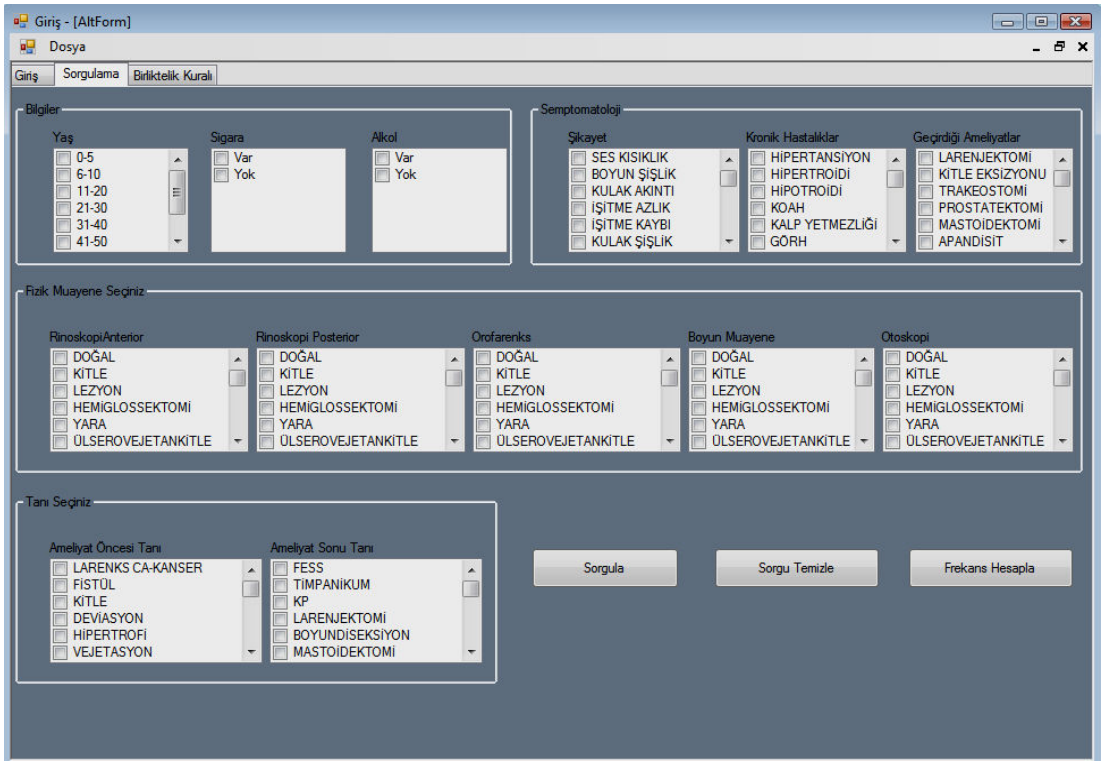
Şekil 4. 4. Veri Tablosu

Metinler veri tablosu haline dönüştürülmeden önce bazı işlemlerden geçmektedir. Daha önceki bölümlerde de belirtildiği üzere öncelikli olarak metinlerden noktalama işaretleri ve birden fazla boşluklar çıkartılmaktadır. Metinlerdeki tüm harfler büyük harfe dönüştürülmektedir. Daha sonra "Sık Kullanılan Kelimeler" listesindeki kelimeler metinlerden silinmektedir. Yazım hatalarının düzeltilmesi ve alan isimlerinin standart bir formata dönüştürülmesi için "Düzeltme" listesi kullanılmaktadır. Ayrıca Zemberek kütüphanesi kullanılarak oluşturulan modül arka planda çalışarak "gövdeleme" işlemi yapmakta, böylelikle kelimeler çekim eklerinden arındırılmaktadır. "Anahtar Kelime" listeleri kullanılarak bu listelerdeki kelimeler metinlerde etiketlenmekte, geri kalan kelimeler silinmektedir.

Metinler, veri tablosu haline dönüştürüldükten sonra Şekil 4. 4.'deki "Excel'e Gönder" butonuna tıklanarak Microsoft Office Excel'e ya da "XML Kaydet" butonuna tıklanarak Şekil 4. 5.'de verilen örnek çıktıda görüldüğü üzere XML olarak veritabanına gönderebilmektedir. İstenildiği takdirde metinlerdeki veriler EK-2'deki gibi kodlanmış olarak ya da EK-1'deki gibi orijinal formlarında kaydedilebilmektedir. Yazılımın 591 adet hasta bilgi formunu yapılandırılmış formata dönüştürme süresi, 1,83 GHz çift çekirdek işlemci ve 2 GB RAM'e sahip bir bilgisayarda yaklaşık 2-3 dakikadır.

```
<Hastalar>
     <Hasta DosyaNo="xxxxx">
       <AD-SOYAD>XXXXXXXXXXXXXXX</AD-SOYAD>
       <YAŞ>44</YAŞ>
       <ŞİKAYETİ>KULAK ŞİŞLİK</ŞİKAYETİ>
       <ÖZGEÇMİŞ>ÖZELLİK-YOK</ÖZGEÇMİŞ>
       <OROFARENKS>DOĞAL</OROFARENKS>
       <RİNOSKOPİANTERİOR>DOĞAL</RİNOSKOPİANTERİOR>
       <OTOSKOPİ>DOĞAL</OTOSKOPİ>
       <BOYUNMUAYENESİ>KİTLE</BOYUNMUAYENESİ>
       <CİNSİYETİ>E</CİNSİYETİ>
       <AMELİYATÖNCESİTANI>PAROTİS KİTLE</AMELİYATÖNCESİTANI>
       <AMELİYATSONUTANI>PAROTİDEKTOMİ
     LİPOMEKSİZYON</AMELİYATSONUTANI>
       <RİNOSKOPİPOSTERİOR />
     </Hasta>
     <Hasta DosyaNo="xxxxx">
       <AD-SOYAD>XXXXXXXXXXXXXXX</AD-SOYAD>
       <YAŞ>44</YAŞ>
       <ŞİKAYETİ>GENİZ AKINTI BURUN TIKANIKLIK</ŞİKAYETİ>
       <ÖZGEÇMİŞ />
       <OROFARENKS>DOĞAL</OROFARENKS>
       <RİNOSKOPİANTERİOR>NAZAL POLİP</RİNOSKOPİANTERİOR>
       <OTOSKOPİ>DOĞAL</OTOSKOPİ>
       <BOYUNMUAYENESİ>DOĞAL</BOYUNMUAYENESİ>
       <CİNSİYETİ />
       <AMELİYATÖNCESİTANI>SEPTUM DEVİASYON KONKA HİPERTROFİ
     SİNONAZALPOLİPOZİS</AMELİYATÖNCESİTANI>
       <AMELİYATSONUTANI>FESS SEPTUMPLASTİ
     KP</AMELİYATSONUTANI>
       <RİNOSKOPİPOSTERİOR>DOĞAL</RİNOSKOPİPOSTERİOR>
     </Hasta>
</Hastalar>
```

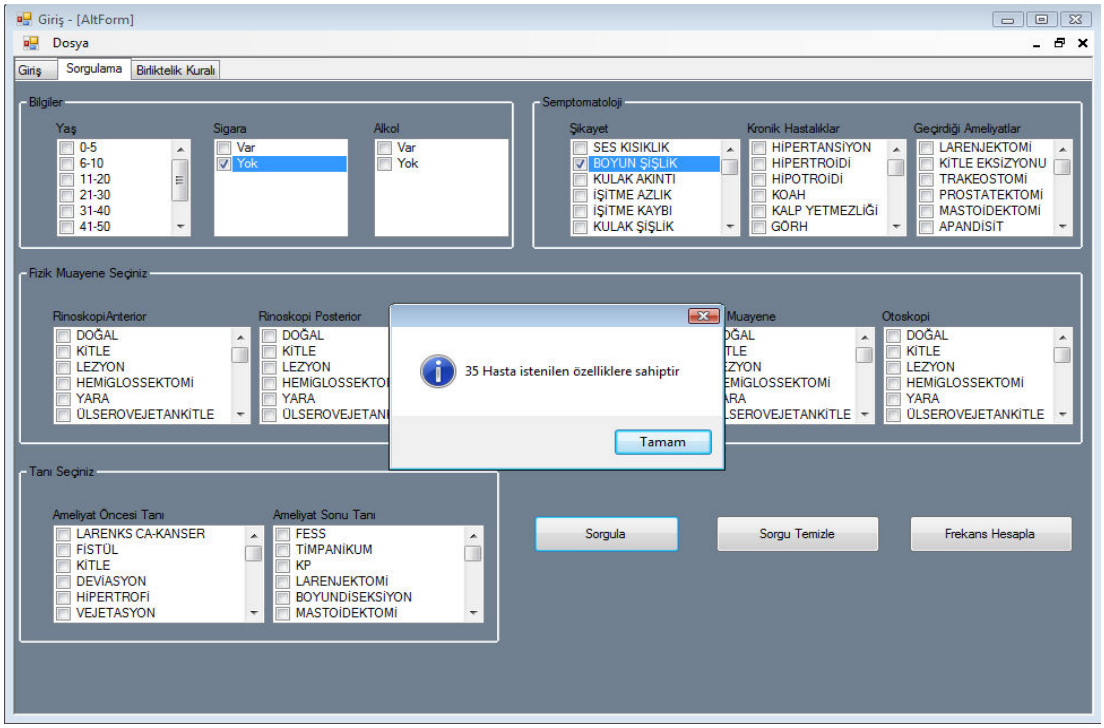
Şekil 4. 5. XML Formatında Hasta Kaydı Örneği



Şekil 4. 6. Metin Sorgulama Ekranı

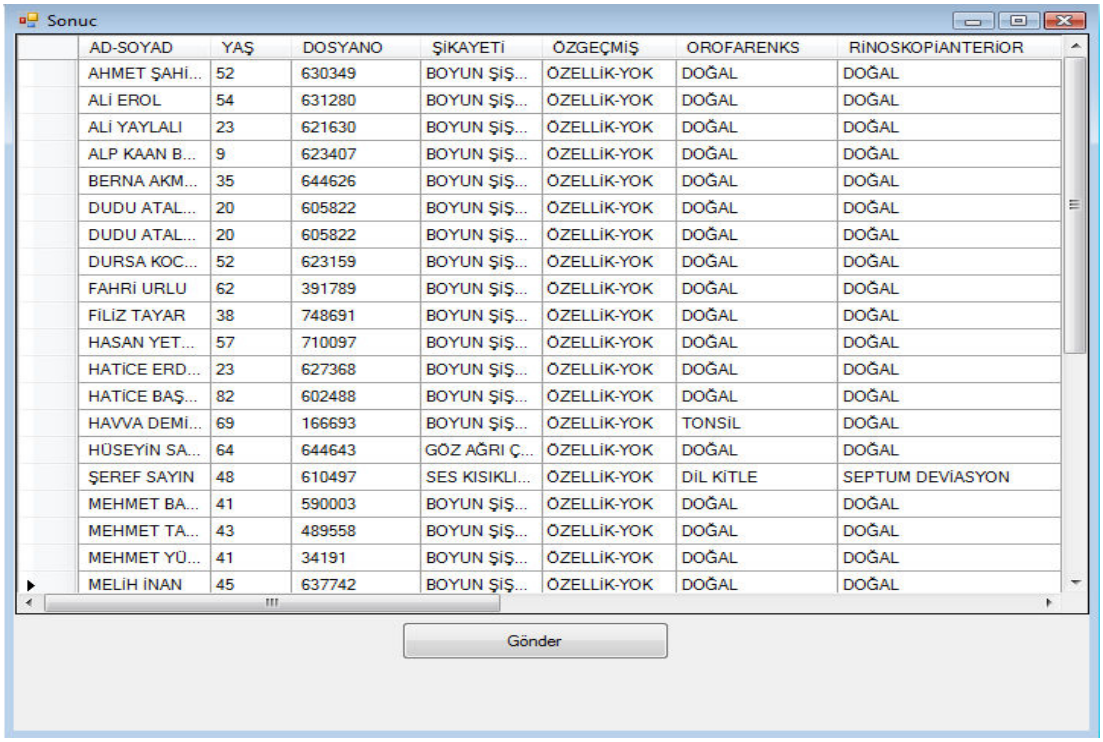
Geliştirilen yazılımda bulunan diğer bir ekran ise "Sorgulama" sekmesidir. Bu sekmeye tıklanarak hekimlerin yüklenen metinleri istedikleri hasta özelliklerine göre veri tablosundan sorgulayabilecekleri Şekil 4. 6.'daki "Sorgulama" ekranı açılmaktadır. Bu ekrandaki özellikler, Yaş, Sigara ve Alkol alanı dışında Microsoft Office Excel dosyalarından alınmaktadır. Sigara, Alkol, Geçirdiği Ameliyatlar ve Kronik Hastalıklarla ilgili bilgiler metinlerde "Özgeçmiş" alanında bulunmaktadır. Bu yüzden bu dört alandaki kelimeler "Özgeçmiş" alanında aratılmaktadır. Kullanıcı bu alanlarda istediği özelliklere tıklayarak sadece bu özelliklere sahip hastalara erişebilmektedir.

Kullanıcılar, istedikleri özellikleri seçip sorgula butonuna bastıklarında öncelikli olarak istenilen özelliklerde kaç adet hastanın bulunduğunu bildiren Şekil 4. 7.'deki gibi bir mesaj kutusu ile karşılaşmaktadır.



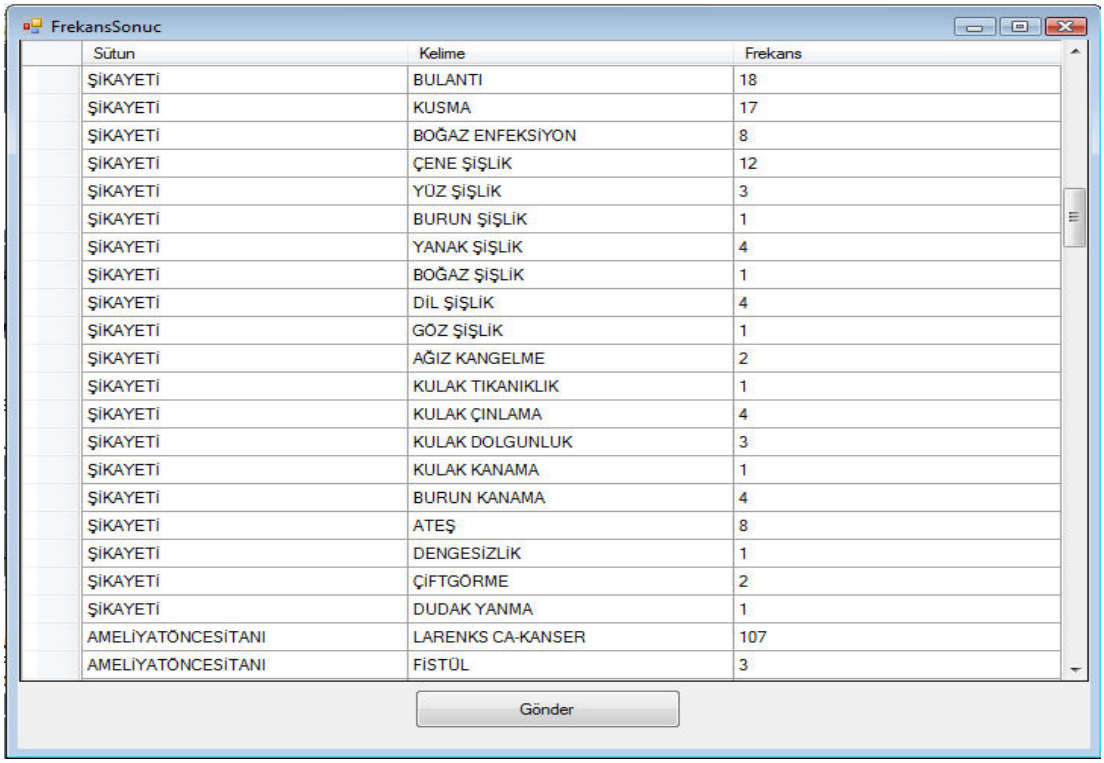
Şekil 4. 7. Mesaj Kutusu

Açılan bu mesaj kutusunda "Tamam" butonuna tıklanıldığı zaman erişilen hasta bilgileri "Sonuç" formunda kullanıcıya sunulmaktadır (Şekil 4. 8.). Bu formda bulunan "Gönder" butonuna tıklanıldığı zaman kullanıcı bu sonuçları Microsoft Office Excel'e gönderebilmektedir.



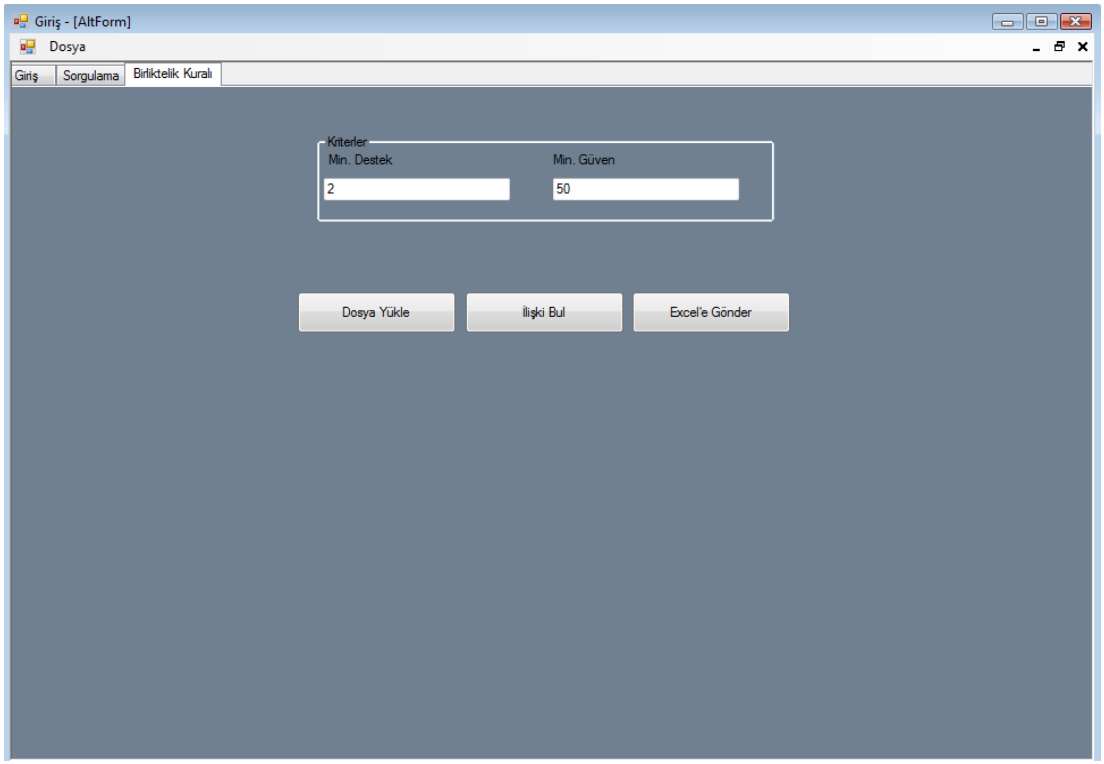
Şekil 4. 8. Sorgu Sonuç Formu

Ayrıca “Sorgulama” ekranında bulunan “Frekans Hesapla” butonuna tıklanıldığı zaman kullanıcıya, her alan içinde geçen kelimelerin frekansları “Frekans Sonuç” formuyla sunulmaktadır (Şekil 4. 9.). Formda bulunan tablo üç sütun içermektedir. İlk sütunda kelime veya kelime grubunun hangi alanda geçtiği, ikinci sütunda kelime veya kelime grupları ve üçüncü sütunda ise frekanslar verilmektedir. Burada da “Gönder” butonuna tıklanıldığı zaman elde edilen sonuçlar Microsoft Office Excel’e gönderilebilmektedir.

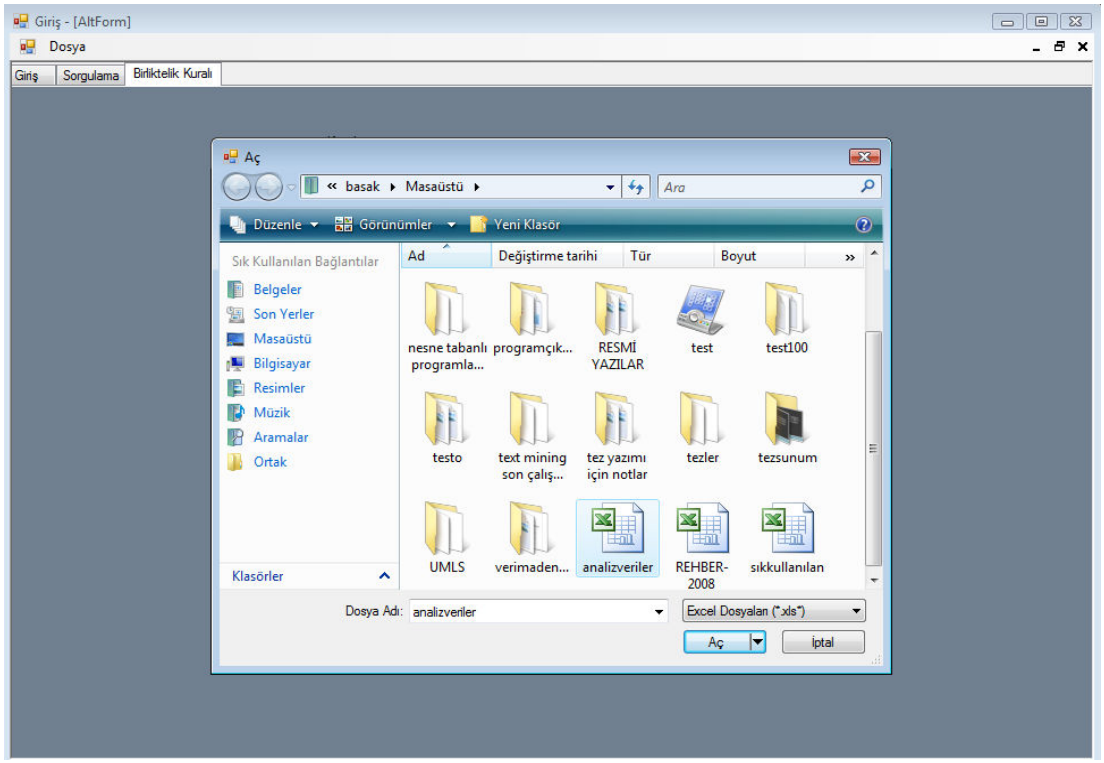


Şekil 4. 9. Frekans Sonuç Formu

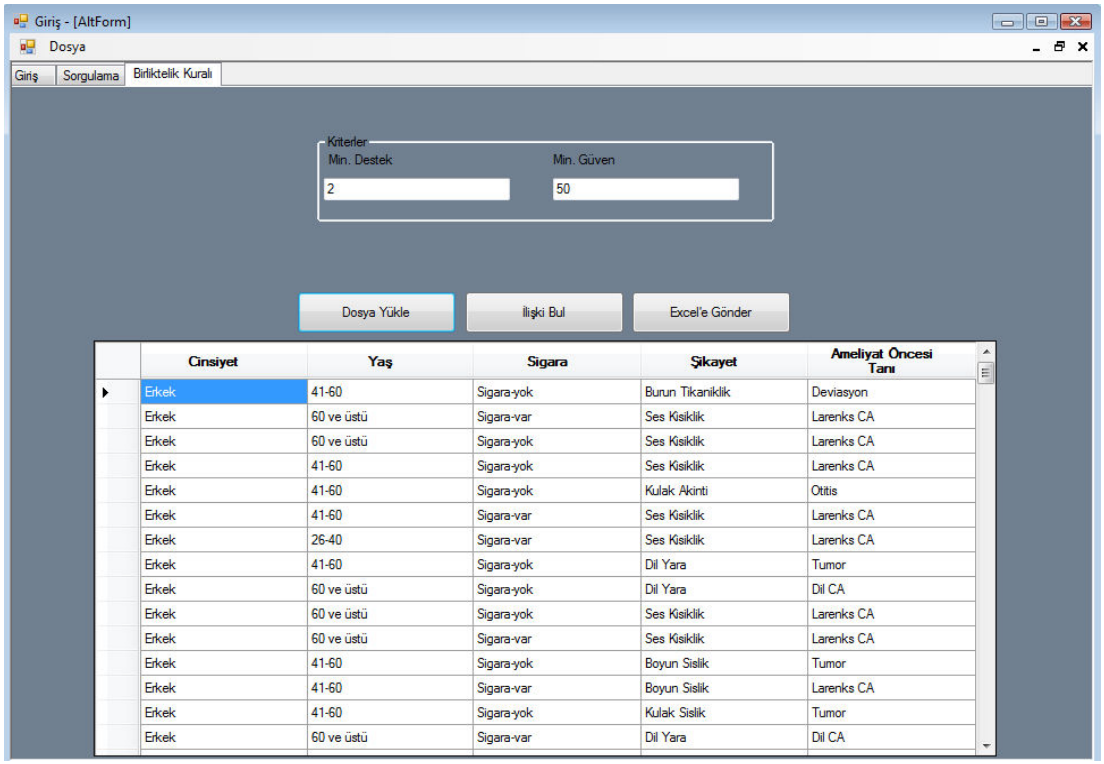
Yazılım içerisinde bulunan diğer bir özellik ise “Birliktelik Kuralı” sekmesidir. Kullanıcı bu sekmeye tıkladığı zaman Şekil 4. 10.’daki ekranla karşılaşmaktadır. Bu ekrana gelindiği zaman öncelikle kullanıcının, yöntemin uygulanacağı verileri yüklemesi gerekmektedir. Bu işlem, ekranda bulunan “Dosya Yükle” butonuna tıklanarak yapılmaktadır. Şekil 4. 11.’deki gibi analiz edilecek veriler “Dosya Aç” diyalogu kullanılarak seçilmekte ve Şekil 4. 12.’deki gibi bir veri tablosuyla kullanıcıya sunulmaktadır. “İlişki Bul” butonuna tıklanıldığı zaman Birliktelik Kuralı algoritması verilere uygulanmakta ve kullanıcı tarafından belirlenen minimum destek ve güven değerine göre kurallar elde edilmektedir (Şekil 4. 13.). Birliktelik Kuralı, uygulanan algoritma ve sonuçların yorumlanması ile ilgili bilgi “Hasta Bilgi Formlarının Analizinden Elde Edilen Sonuçlar” bölümünde ayrıntılı olarak verilmiştir.



Şekil 4. 10. Birliktelik Kuralı Uygulama Formu

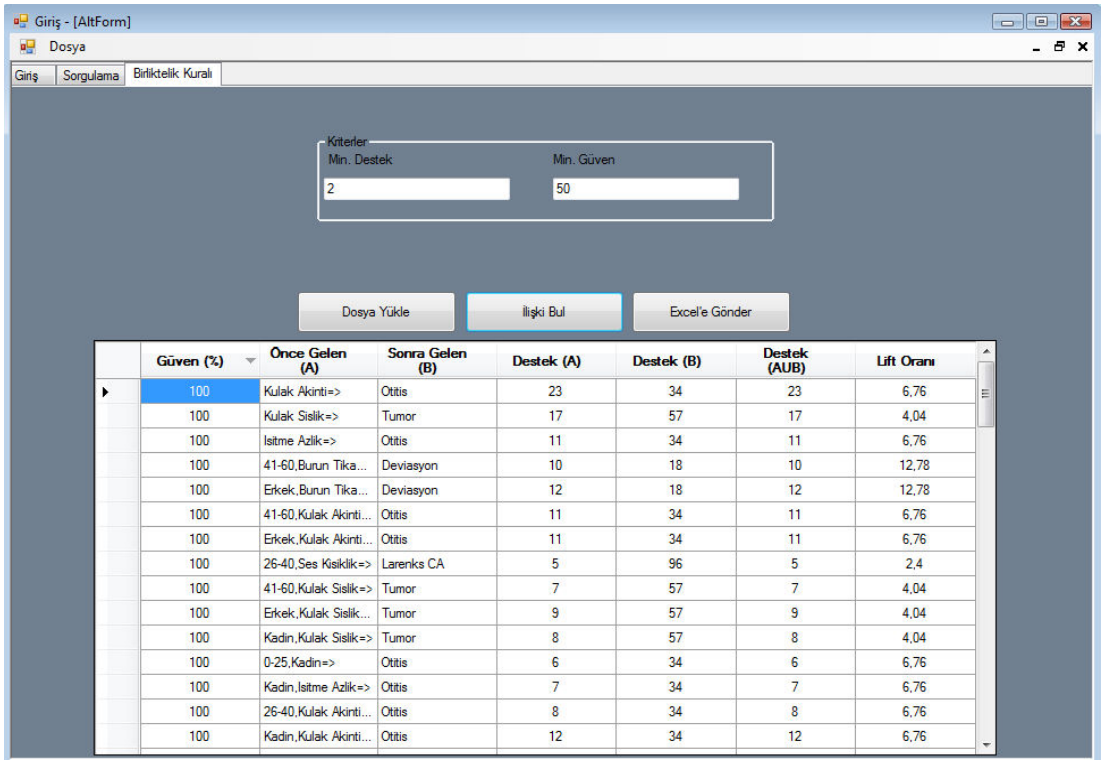


Şekil 4. 11. Dosya Aç Penceresi



Cinsiyet	Yaş	Sigara	Şikayet	Ameliyat Öncesi Tanı
Erkek	41-60	Sigara-yok	Burun Tikaniklik	Deviasyon
Erkek	60 ve üstü	Sigara-var	Ses Kisiklik	Larenks CA
Erkek	60 ve üstü	Sigara-yok	Ses Kisiklik	Larenks CA
Erkek	41-60	Sigara-yok	Ses Kisiklik	Larenks CA
Erkek	41-60	Sigara-yok	Kulak Akinti	Otitis
Erkek	41-60	Sigara-var	Ses Kisiklik	Larenks CA
Erkek	26-40	Sigara-var	Ses Kisiklik	Larenks CA
Erkek	41-60	Sigara-yok	Dil Yara	Tumor
Erkek	60 ve üstü	Sigara-yok	Dil Yara	Dil CA
Erkek	60 ve üstü	Sigara-yok	Ses Kisiklik	Larenks CA
Erkek	60 ve üstü	Sigara-var	Ses Kisiklik	Larenks CA
Erkek	41-60	Sigara-yok	Boyun Sislik	Tumor
Erkek	41-60	Sigara-var	Boyun Sislik	Larenks CA
Erkek	41-60	Sigara-yok	Kulak Sislik	Tumor
Erkek	60 ve üstü	Sigara-var	Dil Yara	Dil CA

Şekil 4. 12. Yüklenen Analiz Verileri



Güven (%)	Önce Gelen (A)	Sonra Gelen (B)	Destek (A)	Destek (B)	Destek (AUB)	Lift Oranı
100	Kulak Akinti=>	Otitis	23	34	23	6,76
100	Kulak Sislik=>	Tumor	17	57	17	4,04
100	Isitme Azlik=>	Otitis	11	34	11	6,76
100	41-60,Burun Tika...	Deviasyon	10	18	10	12,78
100	Erkek,Burun Tika...	Deviasyon	12	18	12	12,78
100	41-60,Kulak Akinti...	Otitis	11	34	11	6,76
100	Erkek,Kulak Akinti...	Otitis	11	34	11	6,76
100	26-40,Ses Kisiklik=>	Larenks CA	5	96	5	2,4
100	41-60,Kulak Sislik=>	Tumor	7	57	7	4,04
100	Erkek,Kulak Sislik...	Tumor	9	57	9	4,04
100	Kadin,Kulak Sislik=>	Tumor	8	57	8	4,04
100	0-25,Kadin=>	Otitis	6	34	6	6,76
100	Kadin,Isitme Azlik=>	Otitis	7	34	7	6,76
100	26-40,Kulak Akinti...	Otitis	8	34	8	6,76
100	Kadin,Kulak Akinti...	Otitis	12	34	12	6,76

Şekil 4. 13. Analiz Sonucu

4.2. Hasta Bilgi Formlarının Analizden Elde Edilen Sonuçlar

KBB polikliniğinden alınan 600 adet hasta bilgi formundan 591 tanesi (genel olarak tüm alanları dolu olan) yapılandırılmış formata dönüştürülmüş ve analiz edilmiştir. Her bir hasta bilgi formunda ortalama 410 kelime bulunmaktadır. Analiz aşamasında hem SPSS paket programı hem de geliştirilen Birliktelik Kuralı modülü kullanılmıştır. SPSS paket programı, hastaların demografik özellikleri ile ilgili parametrelerin saptanması aşamasında tercih edilmiştir.

591 metinden elde edilen sonuçlara göre hastaların % 34,1'i (200) kadın, % 65,9'u (387) erkektir. Dört belgede bu alan boş bırakılmıştır. Kadınların yaş ortalaması 44±16,9, erkeklerinki ise 51,7±16,3'tür. Kadın ve erkeklerin yaş ortalamaları arasında farka bakıldığında erkeklerin yaş ortalamasının daha fazla olduğu istatistiksel olarak anlamlı bulunmuştur (MannWhitney-U, $p<0,05$).

Hastaların sigara-alkol içip içmedikleri, kronik hastalıkları veya geçirdiği ameliyatlarla ilgili bilgiler "Özgeçmiş" alanında yer almaktadır. Metinler incelendiğinde 398 metinde bu alanın doldurulduğu bu hastalardan % 33'ünün (130) düzenli olarak sigara içtiği, % 7'sinin (28) düzenli olarak alkol aldığı ve % 6'sının (25) ise her ikisini birden kullandığı tespit edilmiştir. Hastanın kronik hastalıkları ve geçirdiği ameliyatları ile ilgili bilgilerin girildiği metin sayısı az olduğu için analize dahil edilmemiştir.

Çizelge 4. 1.'de hasta bilgi formlarının analizi sonucunda elde edilen frekanslara örnekler verilmektedir. "Şikayeti" alanında en sık geçen problemler "Ses Kısıklık", "Boyun Şişlik" ve "İşitme Azlık" olarak bulunmuştur. "Ameliyat Öncesi Tanı" alanında "Larenks CA", "Kitle" ve "Otitis", "Ameliyat Sonu Tanı" alanında ise "Boyun Diseksiyon", "Larenjektomi" ve "Parotidektomi" en sık rastlanan kavramlardır.

Bu çalışmanın en büyük hedeflerinden biri metinleri veri madenciliği yöntemlerinin uygulanabileceği bir formata dönüştürmektir. Bu amaçla ilk olarak metinler veri tablosuna dönüştürülmüş ve içeriklerinin yapılandırılması ve bilgisayar tarafından önemli kelime ve kelime gruplarının tanınmasının sağlanması için metinler içerisindeki önemli/anahtar kelimeleri içeren her alana özgü kelime listeleri oluşturulmuştur. Bu listelerde bulunan kelimeler yazılım tarafından metinlerde etiketlenerek kod haline dönüştürülebilmekte ve Microsoft Office Excel çıktısı olarak alınabilmektedir. Çalışmada, metinlerde bulunan varlıklar arasındaki ilişkilerin belirlenebilmesi ve elde edilen kodlanmış veriler üzerinde veri madenciliği yöntemlerinin uygulanabilirliğinin gösterilmesi için Birliktelik Kuralı yöntemi kullanılmıştır. Analizde genel olarak tüm hastalar için girilmiş olan "Yaşı", "Özgeçmiş", "Cinsiyet", "Şikayeti" ve "Ameliyat Öncesi Tanı" alanlarındaki veriler kullanılmıştır. Toplamda 230 hastaya ait veri analize dahil edilmiştir. Özgeçmiş alanından sadece hastanın sigara kullanımı ile ilgili bilgiler alınmıştır. Hastaların yaşları da 0-25, 26-40, 41-60 ve 61 ve üstü olmak üzere dört kategoriye ayrılmıştır. Hasta verilerine ait frekans ve yüzdeler Çizelge 4. 2.'de gösterilmiştir.

Çizelge 4. 1. Frekans Tablosu

Şikayet		Ameliyat Öncesi Tanı		Ameliyat Sonu Tanı	
Kelime	**Frekans**	**Kelime**	**Frekans**	**Kelime**	**Frekans**
Ses Kısıklık	124	Larenks CA	107	Boyun Diseksiyon	145
Boyun Şişlik	75	Kitle	53	Larenjektomi	83
İşitme Azlık	53	Otitis	36	Parotidektomi	44
Yutma Güçlük	39	Dil CA	29	Septumplasti	32
Burun Tıkanıklık	37	Tümör	25	Tümör Rezeksiyon	24
Baş Dönme	34	Septum Deviasyon	19	Hemiglossektomi	18
Kulak Akıntı	31	Hipertrofi	12	Mastoidektomi	16
Kulak Şişlik	28	Vejetasyon	10	Tümör Eksizyon	13
Dil Yara	25	Sinüzit	9	Timpanoplasti	13
Nefes Darlık	24	Dudak CA	7	Biyopsi	13
Kitle	24	Tonsilit	6	FESS	11
Boğaz Ağrı	19	OSAS	5	Adenoidektomi	8
Bulantı	18	Lipom	4	Tonsillektomi	7
Kusma	17	Larenjektomi	4	Kordektomi	5
İşitme Kaybı	14	Fistül	3	KP	4
Horlama	14	Sinonazalpolipozis	3	UPPP	4
Yüz Şekil Bozukluk	12	Damak CA	3	Septokonkaplasti	4
Çene Şişlik	12	Hipofarenks CA	3	Mandibulatomi	3
Solunum Sıkıntı	11	Nazofarenks CA	3	Laringofissür	3
Uğultu	11	Konka Büllöza	3	Tiroidektomi	3
Ağız Açık Uyuma	10	Periferik Fasial Paraliz	2	Bronkoskopi	2

Çizelge 4. 2. Analizde Kullanılan Hasta Verilerinin Frekans ve Yüzdeleri

		N	%
Ameliyat Öncesi Tanı	Larenks-CA	96	41,7
	Septum Deviasyon	18	7,8
	Tümör	57	24,8
	Otitis	34	14,8
	Dil-CA	25	10,9
Yaş	0-25	15	6,5
	26-40	29	12,6
	41-60	114	49,6
	61 ve üstü	72	31,3
Sigara	Var	65	28,3
	Yok	165	71,7
Şikayet	Ses Kısıklık	94	40,9
	Boyun Şişlik	27	11,7
	Kulak Akıntı	23	10,0
	İşitme Azlık	11	4,8
	Kulak Şişlik	17	7,4
	Burun Tıkanıklık	19	8,3
	Nefes Darlık	7	3,0
	Kitle	11	4,8
	Dil Yara	21	9,1
Cinsiyet	Kadın	65	28,3
	Erkek	165	71,7
Toplam		230	100,0

Birliktelik Kuralı analizi ile yüksek sıklıkta birlikte görülen kelime ve kelime grupları bulunmakta ve belirlenen minimum destek ve güven değerine göre kurallar üretilmektedir. Ayrıca kuralın kalitesiyle ilgili başka bir parametre de Lift oranıdır. Lift oranının 1'den büyük olması, hastalara konulan tanıların birlikte bulunduğu diğer değişkenlere bağımlı olduğunu, 1'e eşit veya yakın olması değişkenlerin birbirinden bağımsız olduğunu göstermektedir. Çizelge 4. 3.'de, % 4 minimum destek, % 50 minimum güven değeri ve 2'den büyük lift değerine göre "Yaşı", "Özgeçmiş", "Cinsiyet", "Şikayeti" ve "Ameliyat Öncesi Tanı" alanları kullanılarak toplamda 230 hasta için yapılan analizden çıkartılan kurallar gösterilmiştir.

Çizelge 4. 3. Analiz Sonucunda Elde Edilen Kurallar

Kural No	Birliktelik Kuralları (A=>B)	Destek (%)	Güven (%)	Lift Oranı
1	41-60, Burun Tıkanıklık=>Septum Deviasyon	4,3	100	12,78
2	Burun Tıkanıklık, Erkek=>Septum Deviasyon	5,2	100	12,78
3	Burun Tıkanıklık=>Septum Deviasyon	7,8	94,74	12,11
4	Dil Yara, Kadın=> Dil CA	4,8	91,67	8,43
5	Dil Yara=> Dil CA	7,8	85,71	7,89
6	Kadın, Kulak Akıntı=> Otitis	5,2	100	6,76
7	İşitme Azlık=> Otitis	4,8	100	6,76
8	Kulak Akıntı=> Otitis	10	100	6,76
9	Erkek, Kulak Akıntı=> Otitis	4,8	100	6,76
10	41-60, Kulak Akıntı=> Otitis	4,8	100	6,76
11	Kulak Şişlik=> Tümör	7,4	100	4,04
12	41-60, Boyun Şişlik=> Tümör	6,1	87,5	3,53
13	Boyun Şişlik=> Tümör	9,1	77,78	3,14
14	41-60, Erkek, Ses Kısıklık=> Larenks CA	19,1	95,65	2,29
15	41-60, Ses Kısıklık=> Larenks CA	20	93,88	2,25
16	41-60, Ses Kısıklık, Sigara-var=> Larenks CA	12,2	93,33	2,24
17	Erkek, Ses Kısıklık=> Larenks CA	35,7	93,18	2,23
18	41-60, Erkek, Ses Kısıklık, Sigara-var=> Larenks CA	11,3	92,86	2,22
19	Ses Kısıklık=> Larenks CA	37,4	91,49	2,19
20	Ses Kısıklık, Sigara-var=> Larenks CA	20	90,2	2,16
21	Erkek, Ses Kısıklık, Sigara-var=> Larenks CA	18,7	89,58	2,15
22	60 ve üstü, Erkek, Ses Kısıklık=> Larenks CA	14,8	89,47	2,14
23	60 ve üstü, Ses Kısıklık=> Larenks CA	15,2	87,5	2,10
24	41-60, Erkek, Sigara-var=> Larenks CA	13	85,71	2,05
25	41-60, Sigara-var=> Larenks CA	13,9	84,21	2,02
26	60 ve üstü, Ses Kısıklık, Sigara-var=> Larenks CA	6,5	83,33	2,00

Analiz sonucunda elde edilen kuralların sol tarafında bulunan ve "Yaşı", "Özgeçmiş", "Cinsiyet" ve "Şikayeti" alanlarına ait verileri içeren kısım "kuralın gövdesi" olarak adlandırılmaktadır. Sağ tarafta bulunan ve "Ameliyat Öncesi Tanı" alanına ait verileri içeren kısım ise "kuralın başı" olarak adlandırılmaktadır [18]. Kurallar yorumlanırken hem destek değerinin hem de güven değerinin yüksek

olmasına dikkat edilmelidir. Kuralların nasıl yorumlandığına dair aşağıda üç örnek verilmiştir.

- **Kural 1:** Yaşı 41-60 arasında olan ve Burun Tıkanıklığı şikayeti olan hastaların tamamına Septum Deviasyonu tanısı konulmuştur (Güven=%100).

- **Kural 12:** Yaşı 41-60 arasında olan ve Boyun Şişlik şikayeti olan hastaların %87,5'ine Tümör tanısı konulmuştur (Güven=%87,5).

- **Kural 15:** Yaşı 41-60 arasında olan ve Ses Kısıklığı şikayeti olan hastaların %93,88'ine Larenks CA tanısı konulmuştur (Güven=%93,88).

TARTIŞMA

Sağlık alanında bilgi sistemlerinin yaygınlaşması ile birlikte hasta ile ilgili veriler klinik veri tabanlarında tutulmakta ve istenildiği zaman bu verilere sağlık bakım uzmanları tarafından erişilebilmektedir. Klinik veritabanlarında bazı veriler yapılandırılmış formatta (yaş, sistolik arter basıncı gibi sayısal veya ICD-10 gibi kategorik veriler), bazıları ise yapılandırılmamış formatta (metin, video vb.) saklanmaktadır. Bunun yanı sıra birçok hastanede bilgi sistemleri yönetsel ve mali amaçlı kullanılmakta ve hasta bilgilerinin büyük çoğunluğu kağıt tabanlı ya da elektronik ortamda hasta bilgi formlarında tutulmaktadır. Bunun yanı sıra çevrimiçi veritabanları ve internet de sağlık alanındaki bilgilerin tutulduğu ve hekimlerin sıklıkla faydalandığı önemli kaynaklardan biridir. Görüldüğü üzere hekimler klinik araştırmalarda ya da tıbbi karar vermede hem yapılandırılmış hem de yapılandırılmamış formatta bulunan bu verileri kullanmaktadır. Veri madenciliği yöntemleri yapılandırılmış verilerin analizinde ve varlıklar arasındaki gizli ilişkilerin çıkartılmasında kullanılmakta ve hem yönetimsel hem de klinik karar vermede büyük bir rol oynamaktadır. Ne yazık ki yapılandırılmamış verilerin analizinde veri madenciliği yöntemleri yetersiz kalmakta [9] ve bu türdeki veriler içerisinde birçok bilgi kaybolmaktadır. Son yıllarda büyük ilgi gören metin madenciliği, daha önce de bahsedildiği gibi serbest formatta bulunan metinlerin formatlanması, metinlerden bilinmeyen/beklenmeyen ilişkilerin çıkartılması ve anlaşılır bir şekilde sunulmasını sağlamaktadır. Metin madenciliği diğer alanlarda olduğu gibi sağlık alanında da sıklıkla kullanılmakta ve metinler içerisinde bilgilerin kaybolmasını önleyerek sağlık bakım uzmanlarına büyük faydalar sağlamaktadır.

Bu çalışmada, yapılandırılmamış formatta bulunan KBB hasta bilgi formlarının yapılandırılmış hale dönüştürme, hekimlerin hasta ile ilgili ihtiyaç duydukları bilgilere erişimini kolaylaştırma ve karar verirken ya da araştırma yaparken hasta bilgi formlarını incelemek için harcadıkları zamanı azaltma, veri madenciliği teknikleri ile elde edilen verileri analiz etme ve varlıklar arasındaki gizli ilişkileri çıkartma amaçlanmıştır. Bu amaçlar doğrultusunda Visual Studio platformu ve C# programlama dili kullanılarak bir yazılım geliştirilmiştir. Visual Studio platformu yazılım geliştirme sırasında görsel tasarımın oluşturulmasında sunduğu araçlarla ve kod yazarken sağladığı kolaylıklarla (Intellisense özelliği) yazılımcılar tarafından sıklıkla tercih edilmektedir. Ayrıca C# programlama dilinin nesnel bir dil olma özelliği ile sınıf mantığı kullanılarak az kodla profesyonel projeler geliştirilebilmektedir [74].

Yazılım geliştirilirken ilk amaç metinlerin, veri madenciliği yöntemlerinin kolaylıkla uygulanabileceği bir formata dönüştürülmesiydi. Çünkü geleneksel veri madenciliği yüksek derecede yapılandırılmış veriler üzerinde uygulanmaktadır. Bu sebeple ilk olarak metin formatında bulunan hasta bilgi formları veri tablosuna

dönüştürülmeye ve böylelikle metinler üzerinde daha kolay işlem yapılabilir hale getirilmeye çalışılmıştır. Fakat bu aşamada, kullanılan hasta bilgi formlarının hekimler tarafından özenle doldurulmamış olması ve serbest formatta bulunan metinler üzerinde çalışılmasından kaynaklanan bazı problemler yaşanmıştır. Özellikle yazım hatalarının fazla olması, formların genel olarak standart bir yapıya sahip olmaması (bir formda bulunan bir alanın başka bir formda bulunmaması) ve alan isimlerinin farklı şekillerde yazılmış olması karşılaşılan problemlerin başında gelmektedir ve yazılım geliştirilirken zaman kaybına yol açmıştır. Bu problemlerin nasıl çözüleceğine yönelik yapılan araştırmalar sonucunda elde edilen bilgiler kullanılarak bazı işlemler yapılmıştır. İlk olarak, tüm alan isimlerinin tek bir formata dönüştürülebilmesi için düzeltme listesi oluşturulmuştur. Daha sonra tüm formlarda bulunan, sıklıkla girilen ve önem derecesi yüksek olan alanlar uzman hekimin de görüşleri alınarak belirlenmiş ve önemli alanlar listesi oluşturulmuştur. Bu listede bulunmayan alanlar yazılım tarafından otomatik olarak silinmektedir. Yazım hatalarının tespitinde Zemberek Kütüphanesi kullanılmış ve metinler içerisindeki hataların düzeltilmesi için hatalı bulunan kelimeler düzeltme listesine eklenmiştir. Zemberek kütüphanesinin tercih edilmesindeki en büyük etken Türkiye'de iyi bilinen ve birçok çalışmada kullanılmış ve olumlu sonuçlar vermiş bir sistem olmasıdır [83-87]. Ayrıca açık kaynak kodlu bir yazılım olduğu için ihtiyaca göre değişiklikler yapılabilmekte ve sözlüğe yeni kelimeler eklenebilmektedir. Önemli problemlerden biri olan metin boyutlarının azaltılması için internette bulunan listelerin birleştirilmesi ile oluşturulan sık kullanılan kelimeler listesi kullanılmıştır. Tüm bu işlemler yapıldıktan sonra yazılım tüm metinleri veri tablosu haline dönüştürmekte ve istenildiği takdirde Microsoft Office Excel'e gönderilebilmekte veya XML olarak veritabanına kaydedilebilmektedir. XML'in tercih edilmesindeki en büyük neden hem yapılandırılmış belge ve verilerin evrensel formatı olması hem de verileri standart bir formata dönüştürerek farklı sistemler arasında veri iletişimine olanak sağlamasıdır. Ayrıca XML standardı, literatürde metin madenciliği ile ilgili yapılan birçok çalışmada sıklıkla tercih edilmiş bir teknolojidir [53, 88-90].

Çalışmanın ikinci hedefi, metinlerde bulunan önemli verilerin gözden kaçmasını önlemek ve bu verileri veri madenciliği yöntemleri ile analiz ederek varlıklar arasındaki gizli ilişkileri çıkartmaktır. Öncelikli olarak, metinlerde önemli olan anahtar kelimelerin saptanıp kod haline dönüştürülebilmesi için her alana özgü kelime listeleri oluşturulmuştur. Kelime listeleri oluşturulurken yazılım tarafından tüm metinlerden çıkartılan tekli, ikili, üçlü kelime ve kelime gruplarının alanlara göre frekanslarına bakılmıştır. Yazılım her alandaki kelime ve kelime gruplarını ayırdıktan sonra liste halinde frekanslarıyla birlikte Microsoft Office Excel'e göndermektedir. Bu listedeki frekanslara göre elle anahtar kelime listeleri oluşturulmuştur. Kelime listeleri Microsoft Office Excel'de oluşturulduğu için ihtiyaca göre eklemeler ya da çıkarmalar yapılabilmektedir. Buradaki en büyük problem, kelime listelerinin sadece KBB alanına ait kavramları içermesi ve yazılımın KBB hasta bilgi formlarının içeriğine göre hazırlanmış olmasından dolayı, sistemin başka alanlara ait ve farklı yapıdaki dokümanlarda, var olan özellikleri ile kullanılamamasıdır. Bunun sağlanabilmesi için kelime listelerinin ilgili alana ve dokümana göre tekrardan oluşturulması ve buna ek olarak yazılımda da bazı değişikliklerin yapılması gerekmektedir. İlerleyen zamanlarda, bu kısıtlılığın giderilebilmesi için literatürde yapılan birçok çalışmada [50, 91-94] olduğu gibi

kelime listelerinin olasılıksal modellerin kullanımıyla otomatik olarak oluşturulması ve insan emeğinin en aza indirgenmesi planlanmaktadır. Oluşturulan kelime listelerinin kullanımıyla kelime ve kelime grupları, EK-2'de gösterildiği gibi kodlanmış bir şekilde kaydedilerek analiz edilebilir hale getirilmiştir. Bu şekilde elde edilen veriler üzerinde veri madenciliği yöntemleri uygulanarak ilişki örüntüleri çıkarılabilmektedir.

Bu çalışmada veri madenciliği yöntemlerinden Birliktelik kuralı kullanılmış ve elde edilen kurallar Bulgular bölümünde sunulmuştur. Birliktelik Kuralı veri madenciliğinde ilk geliştirilen tekniklerden biridir. Bu yöntemle kavramlar/varlıkların birlikte bulunma durumlarına bakılarak aralarındaki yüksek sıklıkta görülen ilişki örüntüleri tespit edilmekte ve kurallar çıkartılmaktadır [95]. Ayrıca Birliktelik Kuralları hem analistler hem de normal kullanıcılar tarafından kolayca anlaşılıp yorumlanabildiği için çalışmalarda sıklıkla tercih edilmektedir [96]. Analiz sürecinde yaşanan en büyük problem, doktorların birçok formda tüm alanları doldurmamış olması ve bu yüzden analize dahil edilen özellik ve hasta sayısında düşüş yaşanmasıdır. Çalışmada, 591 hasta bilgi formundan elde edilen veriler incelendikten sonra analiz için elverişli alanlar (yaş, cinsiyet, şikayet, özgeçmiş, ameliyat öncesi tanı) belirlenmiş ve bu alanların tümüne sahip olan 230 hasta verisi Birliktelik kuralı analizinde kullanılmıştır. Analiz sonucunda minimum destek, güven ve lift değerine göre 26 kural elde edilmiştir. Bu tür kuralların, gelecekte geliştirilecek olan karar destek sistemlerine veya klinik rehberlere fayda sağlayacağı düşünülmektedir. Şerban et al. [97] çalışmasında kanserli hastalara ait tıbbi veriler üzerinde Birliktelik kuralı analizi uygulanmış ve elde edilen kurallar kullanılarak hastaların semptomlarına göre kanserli olup olmadıklarını tahmin eden küçük ölçekli bir sistem geliştirilmiştir. Ordonez'in çalışmasında [27] hastanın kronik hastalıkları, yaşı, cinsiyeti, sigara içip içmediği, kan basıncı vb. verileri içeren ve kalp rahatsızlığı olan hastalara ait 655 adet hasta kaydı kullanılarak Birliktelik Kuralı analizi yapılmış ve risk faktörlerine göre kalp hastalığının olup olmadığı tahmin edilmeye çalışılmıştır. Mahgoub et al. [96] çalışmasında kuş gribi ile ilişkili birçok kaynaktan (Reuters, BBC, Medical News Today, Yahoo vb.) toplanan örnek 100 adet internet sayfası XML formatına dönüştürülmüş ve anahtar kelimeler arasındaki ilişkilere bakılarak hastalıkla ilişkili özellikler (hastanın durumu, lokasyon vb.) EART adlı metin madenciliği sistemi ile çıkartılmaya çalışılmıştır. Literatürde KBB bölümüne özgü yapılan benzer bir çalışma bulunmamaktadır.

Çalışmanın üçüncü hedefi ise hekimlerin karar verirken veya araştırma yaparken hasta ile ilgili bilgilere kolaylıkla erişebilmelerini sağlamak ve hasta bilgi formlarını incelemek için harcadıkları zamanı azaltmaktır. Bu nedenle hekimlerin istedikleri özellikteki hasta bilgilerine erişimlerini sağlayacak bir sorgu formu tasarlanmıştır. Hekimler, sorgu formunda bulunan yaş, cinsiyet, şikayet gibi hasta özelliklerini seçerek istedikleri hastalara erişebilmekte ve bu hastalara ait klinik bilgileri Microsoft Office Excel sayfası olarak görüntüleyebilmektedirler. Hekimlerin ihtiyaç duydukları hasta bilgilerine erişebilmek için elde bulunan tüm belgeleri inceledikleri göz önünde bulundurulduğunda, geliştirilen yazılımın bu süreci kolaylaştırarak harcanan zamanı büyük bir oranda azaltacağı söylenebilir. Tasarlanan sorgu formunun arayüzü şu anda sadece içerik olarak tamamlanmıştır. Yazılımdaki

eksiklikler giderildikten sonra kullanıcıların da görüşlerini alarak daha kullanışlı bir arayüzün tasarlanması planlanmaktadır.

Sonuç olarak, metin madenciliği ve kullanılan tekniklerle ile ilgili olarak yapılan uzun soluklu bir araştırma sonucunda elde edilen bilgiler ve belirlenen hedefler doğrultusunda KBB hasta bilgi formlarının analizi için bir yazılım geliştirilmiş ve süreç içerisinde metinlerle çalışmanın getirdiği birçok deneyim kazanılmıştır. Bu deneyimlerden yola çıkarak yaşanan problemlerin en aza indirgenmesi ve yazılımdaki eksikliklerin giderilerek daha profesyonel hale getirilmesi düşünülmektedir.

SONUÇ

Elektronik bir devrimin yaşandığı günümüzde her alanda olduğu gibi tıp alanında da potansiyel olarak depolanan veri hacmi hızla artmaktadır. Bununla birlikte bu verilerden faydalı bilgiler elde etmek giderek zorlaşmaktadır. Tıbbi veri tabanlarında yeni bilgilerin keşfedilmesi, veritabanlarında depolanan verinin etkili bir şekilde kullanılması açısından çok önemlidir. Bu sebepten veri madenciliği tekniklerinin özellikle tıpta kullanımı ve uygulamaları her geçen yıl artmaktadır. Veri madenciliği, daha önce belki de birçok klinik araştırma gerektiren, hem ekonomik hem de insan sağlığı açısından sakıncaları olan tıbbi araştırmaların yerini kısmen de olsa doldurmakta ve tıbbi araştırmalar için yeni bir ufuk sağlamaktadır. Fakat veri madenciliği yöntemleri sadece yapılandırılmış verilerde kullanılabilmekte ve serbest metin formatı gibi yapılandırılmamış verilerde yetersiz kalmaktadır. Metin madenciliği, bu probleme çözüm olarak geliştirilen ve son yıllarda büyük ilgili gören bir alandır.

Metin madenciliği, araştırmacılara literatür gözden geçirmede, işletmelerde serbest metin formatında bulunan bilgilere erişmede yeni çözümler sunmakta ve harcanılan süreyi büyük oranda azaltmaktadır. Tıptaki verilerin çoğu yarı yapılandırılmış veya yapılandırılmamış formatta bulunduğu için hem klinik araştırmalarda hem de karar verme sürecinde hekimlere yardımcı olacak, metin yığınları içerisinden istenilen bilgiye erişmeyi kolaylaştıracak, harcanan süreyi azaltacak ve hasta bakım kalitesini arttıracak bu tür sistemlerin önemi büyüktür.

Bu çalışmada, KBB hasta bilgi formlarını yapılandırılmış formata dönüştüren ve formlardaki verileri kullanarak kavramlar arasındaki ilişki örüntülerini ortaya çıkartan bir yazılım geliştirilmiştir. Yazılımla birlikte hekimlere hasta bilgilerine erişmede yardımcı olmak ve klinik araştırmalarda fayda sağlamak amaçlanmaktadır. İlerleyen zamanlarda kazanılan deneyimlerle birlikte yazılımdaki eksikliklerin giderilerek insan emeğini en aza indirgeyen ve diğer anabilim dallarına da rahatlıkla adapte edilebilecek daha kapsamlı bir yazılımın geliştirilmesi planlanmaktadır. Metin madenciliği, Türkiye'de son birkaç yılda ilgi gören ve bilgi eksikliği, deneyimsizlik ve altyapı yetersizliğinden dolayı özellikle sağlık alanında yeterli sayıda çalışmanın yapılamadığı bir alandır. Çalışmanın en büyük hedeflerinden biri de bu alanda araştırma yapmak isteyen kişilere yol gösterici olmak, konu ile ilgili temel bilgileri vermek, süreç içerisinde ne tür problemlerle karşılaşıldığını ve bu problemlerin nasıl çözümlendiğini göstermektir. Bu yüzden, çalışmanın Türkiye'de konu ile ilgili yapılacak olan çalışmalar için iyi bir kaynak olacağı ve yazılım geliştiricileri süreç içerisinde kullanılması gereken teknikler konusunda bilgilendirerek bu kişilere fayda sağlayacağı düşünülmektedir.

KAYNAKLAR

1. Doğan Ş. Veri Madenciliği Kullanılarak Biyokimya Verilerinden Hastalık Teşhisi. Yüksek Lisans Tezi, 2007, Elazığ.

2. Fayyad U, Piatetsky-Shapiro G, Smyth P. From Data Mining to Discovery Knowledge in Databases. Al Magazine 1996; 3(17): 37-54.

3. Bhatt C, Mining the Medical Literature, http://ai.stanford.edu/~serafim/CS374_2004/Lecture%20Notes/lecture6.pdf, 11.12.2006.

4. Konchady M. Text Mining Application Programming. 1st ed. Charles River Media, Boston, 2006.

5. Wikipedia, Otolaryngology, http://en.wikipedia.org/wiki/Otolaryngology, 02.10.2008.

6. Döşlü A. Veri Madenciliğinde Market Sepet Analizi ve Birliktelik Kurallarının Belirlenmesi. Yüksek Lisans Tezi, 2008, İstanbul.

7. Öğüt S, http://www.sertacogut.com/papers/Sertac_Ogut_-_Veri_Madenciligi_Kavrami_ve_Gelisim_Sureci.pdf, 01.06.2008.

8. Wikipedia, Unstructured Data, http://en.wikipedia.org/wiki/Unstructured_data 30.09.2007.

9. Weiss SM, Indurkhya N, Zhang T, Damerau FJ. Text Mining; Predictive Methods for Analyzing Unstructured Information. Springer Science+Business Media, Newyork, 2005.

10. Dolgun MO, Özdemir TG, Deliloğlu S. Öğrenci Seçme Sınavında (ÖSS) Öğrencilerin Tercih Profillerinin Veri Madenciliği Yöntemleriyle Tespiti. Bilişim'07 Kongresi, Ankara, 2007.

11. Köktürk F, Ankaralı H, Sümbüloğlu V. Veri Madenciliği Yöntemlerine Genel Bakış. Türkiye Klinikleri J Biostat 2009; 1(1): 20-25.

12. Tuğ E. Genetik Algoritmalar ile Tıbbi Veri Madenciliği, Yüksek Lisans Tezi, 2005, Konya.

13. Akpınar H. Veri tabanlarında Bilgi Keşfi ve Veri Madenciliği. İ.Ü. İşletme Fakültesi Dergisi 2000; 29: 1-22.

14. Agrawal R, Imielinski T, Swami A. Mining Association Rules Between Sets of Items in Large Databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data 1993, Washington, 207-216.

15. Özçakır FC, Çamurcu AY. Birliktelik Kuralı Yöntemi için Bir Veri Madenciliği Yazılımı Tasarımı ve Uygulaması. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi 2007; 6(12): 21-37.

16. Zhu H. On-Line Analytical Mining of Association Rules. MSc. Thesis, 1998, Ottawa.

17. Karabatak M, İnce MC, Apriori Algoritması ile Öğrenci Başarısı Analizi, http://www.emo.org.tr/ekler/24f4c5eef7ec01c_ek.pdf, 29.01.2009.

18. DB2 Universal Database, http://publib.boulder.ibm.com/infocenter/db2luw/v8/index.jsp?topic=/com.ibm.im.model.doc/c_lift_in_an_association_rule.html, 03.03.2009.

19. Houtsma M, Swami A. Set-Oriented Mining for Association Rules in Relational Databases. Proceedings of the 11th IEEE International Conference on Data Engineering 1995, Taipei, 25-34.

20. Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules. In Proceedings of the 20th International Conference on Very Large Databases 1994, Santiago, 487-489.

21. Savesere A, Omiecinski E, Navathe S. An Eficient Algorithm for Mining Association Rules in Large Databases. In Proceedings of 20th International Conference on VLDB 1995, 432-444.

22. Das A, Ng WK, Woon YK. Rapid Association Rule Mining. In Proceedings of the Tenth International Conference on Information and Knowledge Management 2001, ACM Press, Atlanta, 487-499.

23. Zaki MJ, Hsiao CJ. CHARM: An Efficient Algorithm for Closed Itemset Mining. In 2nd SIAM International Conference on Data Mining 2002, Arlington, 457-473.

24. Agrawal R, Srikant R, Mining Sequential Patterns, 11th International Conference on Data Engineering 1995, Taipei, Taiwan, 3-14.

25. Kononenko I, Kukar M, Machine Learning and Data Mining; Introduction to Principles and Algorithms, Horwood Publishing, 2007, Chichester.

26. Nam H, Lee K, Lee D. Identification of Temporal Association Rules from Time-Series Microarray Data Sets. BMC Bioinformatics 2009; 10(3): 6.

27. Ordonez C. Association Rule Discovery With the Train and Test Approach for Heart Disease Prediction. IEEE Transactions On Information Technology In Biomedicine 2006 April; 10(2).

28. Güven A. Türkçe Belgelerin Anlam Tabanlı Yöntemlerle Madenciliği. Doktora Tezi, 2007, İstanbul.

29. Sehgal AK, Text Mining: The Search for Novelty in Text, http://www.cs.uiowa.edu/~sehgal/Papers/comp04.pdf , 12.02.2007.

30. Kostoff RN, DeMarco RA, Information extraction from scientific literature with text mining, http://www.onr.navy.mil/sci_tech/special/technowatch/kdocs/anchem2/txt, 22.02.2008.

31. Losiewicz P, Oard DW, Kostoff RN. Textual Data Mining to Support Science and Technology Management. Journal of Intelligent Information Systems 2000; 15(2): 99-119.

32. Cerrito P, 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, 20.12.2006.

33. Mathiak B, Eckstein S, Five Steps to Text Mining in Biomedical Literature, http://www2.informatik.hu-berlin.de/Forschung_Lehre/wm/ws04/7.pdf, 17.01.2007.

34. Do TD, Hui SC, Fong ACM. Associative Feature Selection for Text Mining. International Journal of Information Technology 2006; 12(4): 59-68.

35. Çakıroğlu Ü, Türkçe İçin Doğal Dil İşleme Çalışmaları – 1 Biçimbilimsel ve Morfolojik Analize Çözüm Önerileri, http://www.teknoturk.org/docking/yazilar/tt000135-yazi.htm, 25.02.2008.

36. Oflazer K, Bozşahin HC, Türkçe Doğal Dil İşleme, http://turkoloji.cu.edu.tr/DILBILIM/turkce_dogal_dil_isleme.pdf, 25.02.2008.

37. Maden İ, Demir Ş, Özcan E, Türkçe' den SQL Sorgularına Çeviri Yapan Bir Doğal Dil İşleme Uygulaması (NALAN-TS), http://cse.yeditepe.edu.tr/~eozcan/research/papers/TBD20_2.pdf, 25.02.2008.

38. Wikipedia, Doğal Dil İşleme, http://tr.wikipedia.org/wiki/Do%C4%9Fal_dil_i%C5%9Fleme, 25.02.2008.

39. Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, Cooper J, Guan W, de Groen PC. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. J Biomed Inform 2009.

40. Wang X, Friedman C, Chused A, Markatou M, Elhadad N. Automated knowledge acquisition from clinical narrative reports. AMIA Annu Symp Proc. 2008 Nov 6: 783-7.

41. MedLEE, http://lucid.cpmc.columbia.edu/medlee/, 16.03.2009.

42. Gysbers M, Reichley R, Kilbridge PM, Noirot L, Nagarajan R, Dunagan WC, Bailey TC. Natural language processing to identify adverse drug events. AMIA Annu Symp Proc. 2008 Nov 6: 961.

43. caTIES, https://cabig.nci.nih.gov/tools/caties, 16.03.2009.

44. Goryachev S, Kim H, Zeng-Treitler Q. Identification and extraction of family history information from clinical reports. AMIA Annu Symp Proc. 2008 Nov 6: 247-51.

45. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying Patient Smoking Status from Medical Discharge Records. J Am Med Inform Assoc. 2007; 15(1): 14-24.

46. Onur H. Dizinleme Amacı ile Kullanılabilecek Yöntemlerin Kıyaslanması ve Arama Sistemi Geliştirilmesi. Yüksek Lisans Tezi, 2007, Ankara.

47. Bilgi Erişim Sistemleri, http://yunus.hacettepe.edu.tr/~tonta/yayinlar/kitap/bolum-2.pdf, 20.03.2009.

48. Bilgiye Erişim Sistemleri, http://www.ce.yildiz.edu.tr/mygetfile.php?id=994, 10.01.2008.

49. Gaizauskas R, An Information Extraction Perspective on Text Mining: Tasks, Technologies and Prototype Applications, http://www.itri.brighton.ac.uk/projects/euromap/Text%20Mining%20Event/Rob_Gaizauskas.pdf, 10.10.2008.

50. Mooney RC, Bunescu R. Mining Knowledge from Text Using Information Extraction. ACM SIGKDD Explorations Newsletter 2005; 7(1): 3-10.

51. Erhardt RA, Schneider R, Blaschke C. Status of Text Mining Techniques Applied to Biomedical Text. Drug Dicovery Today 2006; 11(7-8): 315-25.

52. Johnson DB, Taira RK, Cardenas AF, Aberle DR. Extracting Information from Free Text Radiology Reports. Int J Digit Libr. 1997; 1(3): 297-308.

53. Schadow G, Mcdonald CJ. Extracting Structured Information from Free Text Pathology Reports. AMIA Annu Symp Proc. 2003: 584–8.

54. Cohen AM, Hersh WR. A Survey of Current Work in Biomedical Text Mining. Briefings in Bioinformatics 2005; 6(1): 57–71.

55. Amasyalı MF. Bir Doğal Dil İşleme Uygulaması: Soru Cevaplama Sistemi. Yüksek Lisans Tezi, 2003, İstanbul.

56. START, http://start.csail.mit.edu/, 12.12.2008.

57. Ask.com, http://www.ask.com/, 12.12.2008.

58. AnswerBus, http://www.answerbus.com/index.shtml, 12.12.2008.

59. Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as a Knowledge Representation for Clinical Questions. AMIA Annu Symp Proc. 2006: 359–363.

60. Fontelo P, Liu F, Ackerman M. askMEDLINE: a free-text, natural language query tool for MEDLINE/PubMed. BMC Medical Informatics and Decision Making 2005; 5: 5.

61. Lee M, Cimino J, Zhu HR, Sable C, Shanker V, Ely J, Yu H. Beyond Information Retrieval—Medical Question Answering. AMIA Annu Symp Proc. 2006: 469–473.

62. Yu H, Kaufman D. A Cognitive Evaluation Of Four Online Search Engines For Answering Definitional Questions Posed By Physicians. Pacific Symposium on Biocomputing 2007; 12: 328-339.

63. Tıp Bilişiminde Standartlar, http://cioturk.info/6/6.2.1.3%20Standartlar.htm, 05.02.2009.

64. Smith B, Williams J, Schulze-Kremer S. The Ontology of the Gene Ontology. AMIA Annu Symp Proc. 2003: 609-13.

65. Shatkay H, Edwards S, W. Wilbur J, Boguski M. 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 2000: 317-328.

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

67. Gürsakal N, Sözcük ve Sayı, www.20.uludag.edu.tr/~gursakal/down/say.ppt, 05.11.2006.

68. Lindsay RK, Gordon MD. Literature-based discovery by lexical statistics. Journal of the American Society for Information Science 1999; 50: 574-587.

69. Garten Y, Altman RB. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. BMC Bioinformatics. 2009; 10(2): 6.

70. He M, Wang Y, Li W. PPI finder: a mining tool for human protein-protein interactions. PLoS ONE. 2009; 4(2): 4554.

71. Samur MK. Nesne Tabanlı Yaklaşım ile Bir Klinik Bilgi Sistemi Örneği; Pediatrik Endokrinoloji. Yüksek Lisans Tezi, 2008, Antalya.

72. Zemberek; Project Home Page, https://zemberek.dev.java.net/, 15.12.2007.

73. Tülek M. Türkçe İçin Metin Özetleme. Yüksek Lisans Tezi, 2007, İstanbul.

74. Demirli N, İnan Y. Visual Studio C#.Net 2005. Palme Yayıncılık, 2006, Ankara.

75. Turkish Stopwords, http://www.ranks.nl/stopwords/turkish.html, 06.01.2007.

76. Can F, Kocberber S, Balcik E, Kaynak C, Ocalan HC, Vursavas OM. Information Retrieval on Turkish Texts. Journal Of The Amerıcan Socıety For Informatıon Scıence and Technology 2008; 59(3): 407–421.

77. Bruijn LM, Hasman A, Arends JW. Supporting the classification of pathology reports: comparing two information retrieval methods. Computer Methods and Programs in Biomedicine 2000; 62: 109–113.

78. Heja G, Surjan G. Using n-gram method in the decomposition of compound medical diagnoses. International Journal of Medical Informatics 2003; 70: 229-236.

79. Çebi Y, Dalkılıç G. Turkish Word N-gram Analyzing Algorithms for a Large Scale Turkish Corpus –TurCo. Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04).

80. Wikipedia, Genişletilebilir İşaretleme Dili, http://tr.wikipedia.org/wiki/XML, 24.02.2008.

81. T.C. Milli Eğitim Bakanlığı. Bilişim Teknolojileri; Nesne Tabanlı Programlama 5. 2007, Ankara.

82. RSSnedir?com, XML Nedir?, http://www.rssnedir.com/xml_nedir.php, 24.02.2008.

83. Taş T, Görür A. Author Identification for Turkish Texts. Journal of Arts and Sciences 2007; 7: 151-161.

84. Özbek G, Jonathan S, TURKALATOR; A Suite of Tools for Augmenting English-to-Turkish Statistical Machine Translation, http://infolab.stanford.edu/~jonsid/turkalator.pdf, 02.02.2007.

85. Güngör O, Güngör T,Türkçe için Bilgisayarla İşlenebilir Sözlük Kullanarak Kavramlar Arasındaki Anlamsal İlişkilerin Belirlenmesi, http://www.cmpe.boun.edu.tr/~gungort/papers/Turkce%20icin%20Bilgisayarl

a%20Islenebilir%20Sozluk%20Kullanarak%20Kavramlar%20Arasindaki%20Anlamsal%20Iliskilerin%20Belirlenmesi.pdf, 02.02.2007.

86. Güngör O, Güngör T. Türkçe Bir Sözlükteki Tanımlardan Kavramlar Arasındaki Üst-Kavram İlişkilerinin Çıkarılması. Akademik Bilişim 2007, Kütahya.

87. Yıldız HK, Gençtav M, Usta N, Diri B, Amasyalı MF, Metin Sınıflandırmada Yeni Özellik Çıkarımı, http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04298870, 02.02.2007.

88. Taira RK, Soderland SG, Jakobovits RM. Automatic Structuring of Radiology Free-Text Reports. Radiology 2001; 21: 237–245.

89. Corney DPA, Buxton BF, Langdon WB, Jones DT. BioRAT: extracting biological information from full-length papers. Bioinformatics 2004; 20(17): 3206-3213.

90. Sotelsek-Margalef A, Villena-Román J. MIDAS: An Information-Extraction Approach to Medical Text Classification. Procesamiento del lenguaje Natural 2008; 41: 97-104.

91. Tür G, Hakkanı-Tür D, Oflazer K. A Statistical Information Extraction System For Turkish. Natural Language Engineering 2003; 9 (2): 181-210.

92. Harkema H, Roberts I, Gaizauskas R, Hepple M. Information Extraction From Clinical Records. 4th UK e-Science All Hands Meeting 2005.

93. Long W. Extracting Diagnoses from Discharge Summaries. AMIA 2005 Symposium Proceedings: 470-474.

94. Bikel DM, Schwartz R, Weischedel RM, An Algorithm That Learns What's in A Name, http://www.cis.upenn.edu/~dbikel/papers/algthatlearns.doc.pdf, 15.03.2009.

95. Feldman R, Sanger J. The Text Mining Handbook; Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, Newyork, 2007.

96. Mahgoub H, Rösner D, Ismail N, Torkey F. A Text Mining Technique Using Association Rules Extraction. International Journal of Computational Intelligence 2008; 4: 1.

97. Şerban G, Czibula IG, Campan A. A Programming Interface For Medical Diagnosis Prediction. Studia Unıv. Babes Bolyai Informatica; 1(1): 21-30.

98. Abulaish M, Dey L, Biological Relation Extraction and Query Answering from MEDLINE Abstracts Using Ontology-Based Text Mining, Data & Knowledge Engineering 2007; 61: 228–262

ÖZGEÇMİŞ

Başak Oğuz, 3 Haziran 1983 yılında Antalya'da doğdu. İlk ve orta öğrenimini Antalya'da tamamladı. 2001 yılında lisans eğitimine başladığı İstanbul Üniversitesi İktisat Fakültesi İngilizce İktisat bölümünden 2006 yılında mezun oldu. 2006 yılı Eylül döneminde Akdeniz Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik ve Tıp Bilişimi Anabilim Dalı'nda yüksek lisans eğitimine başladı. Halen burada araştırma görevlisi olarak çalışmaktadır. Yabancı dili İngilizcedir.

EKLER

Microsoft Office Excel'de Çıktı Olarak Elde Edilen Örnek Veri Tablosu

AD-SOYAD	YAŞ	ŞİKAYETİ	ÖZGEÇMİŞ	OROFAREKS	RİNOSKOPİANTERİOR	OTOSKOPİ	BOYUNMUAYENESİ	RİNOSKOPİPOSTERİOR	AMELİYATÖNCESİTANI	AMELİYATSONUTANI
xxxxxxxx	67	NEFES DARLIK SI	SİGARA-VAR	DOĞAL	DOĞAL	DOĞAL	LAP	DOĞAL	LARENKS CA-KANSER	LARENJEKTOMİ BOYUN
xxxxxxxx	68	SES KISIKLIK	ÖZELLİK-YOK	DOĞAL	DOĞAL	DOĞAL	DOĞAL	DOĞAL	LARENKS CA-KANSER	LARENJEKTOMİ BOYUN
xxxxxxxx	68	SOLUNUM SIKIN	ÖZELLİK-YOK	DOĞAL	DOĞAL	DOĞAL		DOĞAL		
xxxxxxxx	48	SES KISIKLIK	ÖZELLİK-YOK	DOĞAL	DOĞAL	DOĞAL	JUGULER LAP		ÖZAFAGUS FİSTÜL	FLEPÇEVİRİMİ
xxxxxxxx	75	SES KISIKLIK NEI	HİPERTANSİYON	DOĞAL	DOĞAL	DOĞAL	DOĞAL	DOĞAL		
xxxxxxxx	48	SES KISIKLIK	ALKOL-VAR SİGA	DİL KİTLE	DOĞAL	DOĞAL	LAP		LARENKS CA-KANSER	LARENJEKTOMİ BOYUN
xxxxxxxx	77	BOYUN ŞİŞLİK	LARENJEKTOMİ	DOĞAL	DOĞAL	DOĞAL	KİTLE			
xxxxxxxx	53	BOYUN KİTLE AI	ÖZELLİK-YOK	HEMİGLOSSEKTO	DOĞAL	DOĞAL	DOĞAL BOYUN KİTL	DOĞAL	BUKKAL MUKOZA TÜMÖR	TÜMÖRREZEKSİYON
xxxxxxxx	53	YANAK YARA	ÖZELLİK-YOK	YANAK YARA	DOĞAL	DOĞAL	DOĞAL	DOĞAL	BUKKAL CA-KANSER	TÜMÖRREZEKSİYON BO
xxxxxxxx	40	SES KISIKLIK	SİGARA-VAR	UVULAELONGE	DEVİASYON PÜRÜLAN	DOĞAL	DOĞAL		LARENKS CA-KANSER	LARENJEKTOMİ LAREN
xxxxxxxx	75	ÇENE ŞİŞLİK	KİTLEEKSİZYON	DUDAK	DOĞAL	DOĞAL	SUBMANDİBULAR KİTLE			
xxxxxxxx	43	BOYUN ŞİŞLİK	ASTIM	İNSPEKSİYON Kİ	DOĞAL	DOĞAL	SUBMANDİBULAR KİTLE SERVİKAL SUPRAKL		AVİKULER	
xxxxxxxx	66	SES KISIKLIK NEI	TRAKEOSTOMİ	DOĞAL	SEPTUM DEVİASYON K	DOĞAL	DOĞAL	DOĞAL	LARENKS CA-KANSER	LARENJEKTOMİ BOYUN
xxxxxxxx	72	SES KISIKLIK	ALKOL-VAR SİGA	DOĞAL	DOĞAL	DOĞAL	KİTLE LAP	DOĞAL	LARENKS CA-KANSER ÖZA	BOYUNDİSEKSİYON ÖZ
xxxxxxxx	52	BOYUN ŞİŞLİK	ÖZELLİK-YOK	DOĞAL	DOĞAL	DOĞAL	BOYUN SERVİKAL	DOĞAL	PAROTİS KİTLE	PAROTİDEKTOMİ
xxxxxxxx	52	KULAK ŞİŞLİK	ÖZELLİK-YOK	DOĞAL	DOĞAL	DOĞAL	BOYUN KULAK SERI	DOĞAL	PAROTİS KİTLE	PAROTİDEKTOMİ
xxxxxxxx	65	DİL YARA	ALKOL-VAR SİGA	DİL KİTLE	DOĞAL	DOĞAL	DOĞAL		DİL CA-KANSER	HEMİGLOSSEKTOMİ BO
xxxxxxxx	49	KULAK ŞİŞLİK	ÖZELLİK-YOK	DOĞAL	DOĞAL	DOĞAL	BOYUN KULAK KİTL	DOĞAL		
xxxxxxxx	33	SES KISIKLIK NEI	ÖZELLİK-YOK	DOĞAL	DOĞAL	DOĞAL	DOĞAL		LARENKS CA-KANSER	LARENJEKTOMİ BOYUN
xxxxxxxx	38	YANAK ŞİŞLİK	SİGARA-VAR	ÇENE GİNGİVA K	DOĞAL	DOĞAL		DOĞAL BOYUN LAP		
xxxxxxxx	54	BOYUN ŞİŞLİK	ÖZELLİK-YOK	DOĞAL	DOĞAL	DOĞAL	SERVİKAL SERTFİXE	DOĞAL	LARENKS CA-KANSER	BOYUNDİSEKSİYON BO
xxxxxxxx	43	YUTMA GÜÇLÜK	ALKOL-VAR SİGA	TONSİL YARA SU	SEPTUM DEVİASYON	DOĞAL	DOĞAL	DOĞAL	TONSİL CA-KANSER	TONSİLLEKTOMİ BOYU
xxxxxxxx	52	SES KISIKLIK	SİGARA-VAR SİG	DOĞAL	DOĞAL	DOĞAL	DOĞAL		LARENKS CA-KANSER	LARENJEKTOMİ BOYUN
xxxxxxxx	43	KULAK İŞİTME K	ÖZELLİK-YOK	DOĞAL	DOĞAL	DOĞAL	DOĞAL	DOĞAL		
xxxxxxxx	21	KULAK KİTLE		YANAK MUKOZ	BURUN		EKSİZYON CİLT ZİGOMA PAROTİS SUBMANDİBULAR LAP			
xxxxxxxx	71	SES KISIKLIK	PROSTATEKTOM	DOĞAL	DEVİASYON	DOĞAL	DOĞAL		LARENKS CA-KANSER	LARİNGOFİSSÜR KORDI

EK-1

Kodlanmış Olarak Elde Edilen Örnek Veri Tablosu

CİNSİYETİ	YAŞ	ŞİKAYETİ	ÖZGEÇMİŞ	OROFAREKS	RİNOSKOPİANTERİOR	OTOSKOPİ	BOYUNMUAYENESİ	RİNOSKOPİPOSTERİOR	AMELİYATÖNCESİTANI	AMELİYATSONUTANI
1.0	44.0	6.0		1.0	18.0	1.0	1.0	1.0	4.0	1.0
1.0	67.0	1.0	2.0	1.0	1.0	1.0	1.0	26.0	1.0	4.0
1.0	68.0	1.0	4.0	1.0	1.0	1.0	1.0	1.0	1.0	4.0
1.0	48.0	1.0	4.0	1.0	1.0	1.0		26.0	1.0	4.0
1.0	57.0	3.0		1.0	1.0	19.0		1.0	12.0	6.0
1.0	48.0	1.0	1.0	2.0	1.0	1.0		26.0	1.0	4.0
1.0	40.0	1.0	2.0	7.0	19.0	1.0		1.0	1.0	4.0
1.0	55.0	13.0		2.0	1.0	1.0	1.0		9.0	5.0
1.0	61.0	13.0		1.0	1.0	1.0		2.0	25.0	5.0
1.0	66.0	1.0		1.0	12.0	1.0	1.0	1.0	1.0	4.0
1.0	72.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	5.0
1.0	52.0	2.0	4.0	1.0	1.0	1.0	1.0		3.0	9.0
1.0	57.0	2.0	1.0		1.0	1.0		26.0	1.0	4.0
1.0	52.0	5.0	4.0	1.0	1.0	1.0	1.0	2.0	3.0	9.0
1.0	65.0	13.0	1.0	2.0	1.0	1.0		1.0	25.0	5.0
0.0	31.0	5.0			1.0	1.0	1.0	2.0	3.0	5.0
1.0	33.0	1.0	4.0	1.0	1.0	1.0		1.0	1.0	4.0
1.0	54.0	2.0	4.0	1.0	1.0	1.0	1.0	2.0	1.0	5.0
1.0	52.0	1.0	2.0	1.0	1.0	1.0		1.0	1.0	4.0
1.0	71.0	1.0	2.0	1.0	28.0	1.0		1.0	1.0	5.0
1.0	66.0	1.0	4.0	1.0	1.0	1.0	1.0	1.0	1.0	4.0
1.0	77.0	1.0	4.0	1.0	1.0	1.0	1.0	1.0	3.0	40.0
1.0	22.0	2.0	4.0	1.0	28.0	1.0		2.0	3.0	41.0
1.0	23.0	2.0	4.0	1.0	1.0	1.0	1.0	32.0	9.0	5.0
1.0	57.0	3.0	4.0	1.0	1.0	1.0	1.0	1.0	12.0	6.0
1.0	69.0	10.0		1.0	1.0	1.0	1.0	1.0	1.0	4.0

EK-2