

T.C.
ÇUKUROVA ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK ANABİLİM DALI

**HASTANE BİLGİ YÖNETİM SİSTEM VERİLERİNE
VERİ MADENCİLİĞİ İLE İŞLERLİK KAZANDIRILMASI İÇİN
BİR ADIM :
K-ORTALAMA ALGORİTMASINDA UYGUN PARAMETRE
SEÇİMİ**

Meryem GÖRAL YILDIZLI
BİYOİSTATİSTİK DOKTORA PROGRAMI
DOKTORA TEZİ

DANIŞMANI
Prof.Dr. Z. Nazan ALPARSLAN

Bu tez Çukurova Üniversitesi Bilimsel Araştırma Projeleri tarafından TDK-2015-4639 nolu proje ile desteklenmiştir.

ADANA – 2018

TEŞEKKÜR

Bu tez çalışmamda beni yönlendiren, tezimin her aşamasında deneyim ve bilgilerini benimle paylaşarak bana ışık tutan danışman hocam Prof.Dr. Z. Nazan ALPARSLAN'a, tez izleme raporlarımda eleştiri ve önerilerle katkı sağlayan tez izleme jüri üyesi Prof.Dr. H. Refik BURGUT, Prof.Dr. Hamza EROL, Prof.Dr Necdet AYTAÇ hocalarıma, ayrıca biyoistatistik anabilim dalı öğretim üyesi arkadaşlarım Prof.Dr. Gülşah SEYDAOĞLU, Doç.Dr. Yaşar SERTDEMİR, Yrd.Doç. Dr. İlker ÜNAL ve araştırma görevlisi arkadaşlarıma bana verdikleri destekler için teşekkürlerimi sunarım.

Tez çalışmamı hazırladığım dönem boyunca gerektiğinde mesai saatleri konusunda esneklik sağlayarak çalışmamı destekleyen Çukurova Üniversitesi Balcalı hastanesi başhekimimi Prof. Dr. Yeşim TAŞOVA ve başhekim yardımcıları hocalarıma, idari görevlerimi benimle paylaşarak tez çalışmalarına zaman ayırmamı sağlayan mesai arkadaşlarım Ramazan ÇAKIR, Oğuzhan TİMUR ve Fatih YILDIZ'a, çalışmamın büyük bir kısmını mesai saatleri dışında yaparken anlayış gösteren ve destek olan aileme, çalışmam süresince tezime katkı sağlayan arkadaşlarıma, doktora tez proje kapsamında maddi destek sağlayan Çukurova Üniversitesi Bilimsel Araştırma Projeleri birimine, teşekkür ederim.

KABUL VE ONAY

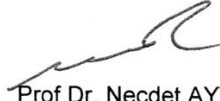
Biyoistatistik Anabilim Dalı
Doktora Programı Çerçevesinde yürütölmüş olan
“Hastane Bilgi Yönetim Sitem Verilerine
Veri Madenciliğı ile İşlerlik Kazandırılması İçin Bir Adım :
K-Ortalama Algoritmasında Uygun Parametre Seçimi ”
adlı çalışma, aşağıdaki jüri tarafından Doktora Tezi olarak kabul edilmiştir.

Tarihi: 10 / 01 / 2018

TEZ SINAV JÜRİSİ



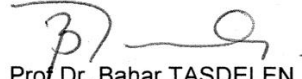
Prof Dr. Z.Nazan ALPARSLAN
Çukurova Üniversitesi
Başkan



Prof Dr. Necdet AYTAÇ
Çukurova Üniversitesi
Üye



Doç Dr. Gülhan ÖREKİCİ TEMEL
Mersin Üniversitesi
Üye



Prof Dr. Bahar TAŞDELEN
Mersin Üniversitesi
Üye



Doç Dr. Yaşar SERTDEMİR
Çukurova Üniversitesi
Üye

Yukarıdaki Tez, Yönetim Kurulunun / / tarih ve
edilmiştir.

sayılı kararı ile kabul

Prof. Dr. Behice DURGUN
Sağlık Bilimleri Enstitü Müdürü

T.C. ÇUKUROVA ÜNİVERSİTESİ SAĞLIK BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS / DOKTORA TEZ ÇALIŞMASI

ETİK BEYANI

Çukurova Üniversitesi Bilimsel Araştırma ve Yayın Etiği Yönergesini okuduğumu ve anladığımı ve Çukurova Üniversitesi Sağlık Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmada;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik kurallarına uygun olarak sunduğumu,
- Tez çalışmada yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde ve ortaya çıkan sonuçlarda herhangi bir değişiklik yapmadığımı,
- Tez olarak sunduğum bu çalışmanın özgün olduğunu,

bildirir, aksi bir durumda bu konuda hakkımda yapılacak tüm yasal işlemleri ve aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim. 10./01/2018


Meryem GÖRAL YILDIZLI

Kayıtlı olunan Program : Ç.Ü. Tıp Fakültesi Biyoistatistik Anabilim Dalı
Tezin Konusu : K-ortalama kümeleme algoritmasının dezavantajına çözüm olan yeni geliştirilen algortimanın oluşturulması ve Hastane Bilgi Yönetim Sistemi verileriyle uygulamasının yapılması.

Tezin Türü : Yüksek Lisans : ☐ Doktora: X ☐

Danışmanın Adı-Soyadı : Prof.Dr. Z.Nazan ALPARSLAN

Danışmanın İletişim Bilgileri

Telefon : 05326504801

E-Posta : nazan@cu.edu.tr

Öğrencinin İletişim Bilgileri

Telefon : 05369690176

E-Posta : meryem@cu.edu.tr

Adresi : Ç.Ü. Tıp Fakültesi
Biyoistatistik Anabilim Dalı

**Bu belgenin Lisansüstü eğitim tezleri savunmaya alınmadan önce öğrenci tarafından doldurulup imzalanarak Enstitü Müdürlüğüne teslim edilmesi gerekmektedir.*

İÇİNDEKİLER

TEŞEKKÜR.....	ii
İÇİNDEKİLER	v
ŞEKİLLER DİZİNİ.....	vii
ÇİZELGELER DİZİNİ	viii
SİMGELER ve KISALTMALAR DİZİNİ	ix
ÖZET.....	x
ABSTRACT.....	xi
1. GİRİŞ.....	1
1.1. Tıp Alanında Veri Madenciliği Yöntemlerinin Kullanıldığı Çalışmalar	2
1.2. K-Ortalama Kümeleme Yönteminin Küme Sayısı, Başlangıç Merkezlerini Belirlemek için Yapılan Çalışmalar	8
1.3. Çalışmanın Amacı	11
2. GENEL BİLGİLER.....	12
2.1. Sağlık Hizmetleri, HBYS ve Büyük Veri	12
2.1.1. Hastane Bilgi Yönetim Sistemi (HBYS).....	12
2.1.2. Büyük Veri (Big Data)	14
2.2. Büyük Veri Analizi	16
2.2.1. Veri Madenciliğinin Sağlık Alanında Kullanımı	17
2.2.2. Veri Madenciliği Süreçleri	18
2.2.3. Veri Madenciliği Yöntemleri	19
2.3. Kümeleme Yöntemleri	20
2.3.1. Kümeleme Yönteminde Kullanılan Tanımlamalar	20
2.3.2. Kümeleme Analizinin Uygulama Aşamaları.....	22
2.3.3. Kümeleme Yönteminin Matematiksel Tanımı	23
2.3.4. Kümeleme Teknikleri	23
2.3.5. Niteliksel Verilerde Küme Geçerliliğinin Ölçülmesi	24
2.4. K-ortalama Yöntemi ve Farklı Yaklaşımlar	25
2.4.1. K-Ortalama Algoritması	25
2.4.2. Canopy Algoritması.....	27
2.4.3. K-Ortalama ++ Algoritması	29

3. GEREÇ VE YÖNTEMLER.....	30
3.1. NAMGY Algoritmasının Ana Hatları	31
3.2. NAMGY Algoritması.....	31
3.3. NAMGY K-Ortalama Algoritmasında Başlangıç Merkezlerinin Belirlenmesi.....	32
3.4. NAMGY K-Ortalama Algoritmasında Küme Sayılarının Belirlenmesi.....	34
3.5. NAMGY Algoritması ve Akış Şeması	36
3.6. NAMGY algoritmasının K-ortalama Kümeleme Yöntemi İle Birleştirilmesi İçin Arayüz Programının Düzenlenmesi	41
4. BULGULAR	42
4.1. K-Ortalama Algoritmasının HBYS Verilerine Uygulanması.....	42
4.2. NAMGY Başlangıç Merkezlerini Belirleme Algoritmasının Uygulanması.....	44
4.3. NAMGY Algoritması ile Küme Sayısı Belirleme Uygulanması	47
4.4. NAMGY Birleştirilmiş Algoritmanın Uygulanması	49
4.5. NAMGY Birleştirilmiş Algoritmanın Küme Geçerlilikleri Uygulanması.....	54
4.6. NAMGY Algoritmasının Arayüz Programının Oluşturulması	56
5. TARTIŞMA.....	65
5.1. K-ortalama Algoritmasının HBYS Verilerine Uygulanması	65
5.2. Literatürdeki Başlangıç Merkezlerini Belirleme Algoritma Çalışmaları ile Karşılaştırmalar	66
5.3. Literatürdeki Uygun Küme Sayısını Belirleme Algoritma Çalışmaları İle Karşılaştırmalar	68
5.4. K-Ortalama++, Canopy Algoritmalarının NAMGY Algoritması ile Çalışma Prensipleri Açısından Değerlendirilmesi.....	69
6. SONUÇLAR VE ÖNERİLER	72
ÖZGEÇMİŞ	79
EKLER.....	80

ŞEKİLLER DİZİNİ

Şekil 2.1 : HBYS Bileşenleri

Şekil 2.2 : Büyük Veri Bileşenleri

Şekil 2.3 : Veri Madenciliğinin Yararlandığı Alanlar

Şekil 2.4 : Veri Madenciliği Süreçleri

Şekil 2.5 : Kümeleme Yönteminin Şematik Gösterimi

Şekil 2.6 : Canopy Algoritmasının Şematik Gösterimi

Şekil 3.1 : Başlangıç Merkezlerinin Şematik Gösterimi

Şekil 3.2 : NAMGY Algoritmasının Akış Şeması

Grafik 4.1 : Küme Sayıları ve SSE Değişim Grafiği

Grafik 4.2.: Vitamin B12 Veri Sonuçlarına Göre SSE ve Küme Sayıları Değişim Grafiği

Şekil 4.1 : Arayüz Programın Ana Menüsü

Şekil 4.2 : Veri dosyasını Kopyalama menüsü

Şekil 4.3 : Verilerin Veri Tabanı Tablosuna Aktaran Menü

Şekil 4.4 : Veri Değişkenlerinin Tiplerini Belirleme Menüsü

Şekil 4.5 : Kümeleme İşlemi Menüsü

Şekil 4.6 : Kümelemeye Ait Özet Bilgiler

Şekil 4.7 : Kümelenmiş Verilerin Listesi

Şekil 4.8 : Kümeleme Sonuçlarına Göre Küme Özellikleri

ÇİZELGELER DİZİNİ

Çizelge 1.1 : Farklı Hastalıklar İçin Kullanılan Veri Madenciliği Uygulamaları

Çizelge 2.1 : HBYS Bileşenleri

Çizelge 2.2 : Kappa İstatistiğinin Hesaplanması ve Çapraz Tablo (2x2)

Çizelge 3.2 : Küme Sayılarına Göre Küme İçindeki Grupların Kayıt Sayısı ve Toplam Karesel Hata

Çizelge 4.1 : K-ortalama , K-ortalama++, Canopy Algoritmalarına Göre Kümeleme Sonuçları

Çizelge 4.2 : Diyabet Veri Seti Bilgileri

Çizelge 4.3 : WEKA ve NAMGY Algoritmasının Diyabet Veri Seti İle Kümeleme Sonuçları

Çizelge 4.4 : WEKA Ve NAMGY Algoritmasının Çalışma Prensibi Sonuçları

Çizelge 4.5 : Orijinal Sınıflama ve NAMGY Algoritmasının Diyabet Veri Seti İle Kümeleme Sonuçları

Çizelge 4.6 : Farklı Küme Sayılarına Göre Kümeleme Sonuçları

Çizelge 4.7 : WEKA Ve NAMGY Arayüz Programın Vitamin B12 Veri Seti Uygulama Sonuçları

Çizelge 4.8 : Vitamin B12 Verisi Yüzdelik Dilimleri Sağlayan Küme Sayısı Ortak Kümeleme Sonuçları

Çizelge 4.9 : Weka Ve NAMGY Algoritmasının Çalışma Prensibi Açısından Karşılaştırılması

Çizelge 4.10 : WEKA ve NAMGY Algoritmasının Analiz Sürecindeki Farklılıkları

Çizelge 4.11 : İris Veri Seti Yapısı

Çizelge 4.12 : İris Veri Seti Çapraz Tablo

Çizelge 5.1 : Algoritmaların Küme Sayısı ve Merkezleri Belirleme Yaklaşımları

Çizelge 5.2 : NAMGY Algoritmasının Diğer Algoritmalarla Benzerlik ve Farklılıkları

SİMGELER ve KISALTMALAR DİZİNİ

AGNES : AGglomerative NEsting (Hiyerarşik Kümeleme)

ATC : Anatomic Therapeutic Chemical Classification (Yapısal Tedavi Edici Kimyasallar Sınıflaması)

Biruz : Merkezleri belirlemede kullanılan birim uzunluk

CSS : Cascading Style Sheets

EM : Expectation Maximization (Beklenti Maksimizasyon)

FCM : Fuzzy c-means (Bulanık C-Ortalama)

Fmr1 : Biyomedikal Rezonans Görüntülemenin Sınıflandırılması

GOU: Genel Ortalama Uzaklık

HBYS : Hastane Bilgi Yönetim Sistemi

(HI) : Hemoliz İndeksi

HTML: Hyper Text Markup Language

ICD-10 : International Classification of Diseases

KS : Toplam komşu nokta sayısı

KUO : Komşu noktalarına uzaklıklar ortalaması

maxuz : Birinci merkez noktası ile kendisine en uzak nokta arasındaki uzaklık

MST : Minimal Spanning Tree (En Küçük Örtün Ağaç)

NAMGY : Noktalar Arası Mesafe ve Gözlemlerin Yoğunluğu

PACS : Picture Archiving and. Communication System (Görüntü saklama ve Arşivleme Sistemleri)

SSE : Summed Squared Error (Toplam Karesel Hata)

SSE_i FARK % : SSE_i ile SSE_{i-1} arasındaki fark

SOM : Self-Organizing Maps (Özdüzenleyici Haritalar)

SVM : Support Vector Machines (Destek Vektör Makineleri)

TC : Türkiye Cumhuriyeti

TİG : Teşhis İlişkili Gruplar

Y_{ni} % : Küme içindeki i. grubun yığılımlı yüzde oranı

ÖZET

HASTANE BİLGİ YÖNETİM SİSTEM VERİLERİNE VERİ MADENCİLİĞİ İLE İŞLERLİK KAZANDIRILMASI İÇİN BİR ADIM : K-ORTALAMA ALGORİTMASINDA UYGUN PARAMETRE SEÇİMİ

Sağlık kurumları yöneticileri; finansal, tıbbi işlemleri, teknolojik gelişmeleri ve risk faktörleri gibi çok sayıda alt sistemleri de yönetebilmelidir. Yöneticilerin kuruluşlarını güvenli, verimli yönetebilmeleri, sorumluluklarını daha iyi yapabilmeleri, yönetsel becerilerini daha kolay ortaya koyabilmeleri için Hastane Bilgi Yönetim Sistemlerine (HBYS) ihtiyaç duyulmaktadır.

Sağlık alanında toplanan manuel veya tıbbi cihazlardan elde edilen farklı formattaki tıbbi verileri, bilişim ve iletişim teknolojilerdeki gelişmelere bağlı olarak dijital ortamlara aktarılmaktadır. Ülkemizde HBYS'lerin yapısal gelişiminde ilerlemeler olmasına rağmen, sistemde toplanan verilere hala yeteri kadar işlerlik kazandırılmış değildir. HBYS'ler sağlık yöneticilerine yardımcı olacak bazı raporlar içerse de, veriler arasında çoklu bağlantıları ortaya çıkarabilecek raporları kapsamamaktadır. Veri madenciliği yöntemlerinin sistemlere adapte edilmiş olmaması, veri-bilgi dönüşüm sürecini yönetebilecek deneyimli ve yeterli bilgi seviyesi olan sistem sorumlularının bulunmaması bu sistemler için önemli bir eksikliktir. HBYS'lerin uygun analiz yöntemlerle entegre olmuş, kolay kullanılabilen daha kapsamlı tasarlanmasına ihtiyaç vardır.

Tıp alanındaki bilimsel çalışmalarda veri madenciliği kümeleme yöntemlerinden K-ortalama algoritması yaygın olarak kullanılmaktadır. Ancak K-ortalama algoritmasının başlangıç parametreleri olan küme sayısının ve başlangıç merkezlerinin kullanıcı tarafından belirlenmesi gereği algoritmanın önemli dezavantajıdır.

Bu probleme çözüm olabileceği düşünülerek, yöntemin parametre değerlerine ihtiyaç duyulmadan çalışan mevcut algoritmayı geliştirmek amaçlanmıştır.

NAMGY (Noktalar Arası Mesafe ve Gözlemlerin Yoğunluğu) isimli algoritmayı içeren yazılım ile HBYS'lerdeki veriler kullanarak hastane yöneticilerin ihtiyaç duyduğu bilgiler daha kolay rapor edilebilecek ve atıl olarak biriken büyük miktardaki veriler işlerlik kazanmış olacaktır.

Ek olarak, çalışma NAMGY K-ortalama algoritması, HBYS verilerinin bilgiye dönüştürülmesinde daha kolay ve kullanıcı dostu yazılım ile sağlık kurumları yönetimine katkı sağlayıp, kurumlar için ekonomik açıdan bir değer oluşturacaktır.

Anahtar Sözcükler: HBYS, K-Ortalama kümeleme yöntemi, veri madenciliği, küme sayısı, başlangıç değerleri .

ABSTRACT

A STEP TOWARDS THE FUNCTIONALITY OF HOSPITAL INFORMATION MANAGEMENT SYSTEM DATA USING DATA MINING TECHNIQUES :

SELECTION OF APPROPRIATE PARAMETERS FOR K- MEANS

Administration of healthcare services require simultaneous handling many issues such as financial, medical, technological subsystems. A hospital management team has to gather many details and convert them into sufficient information to take rational administrative decisions. In this respect hospital information management systems (HIMS) combine information technology to support healthcare service in an easier and more efficient way.

HIMS is whole of integrated information systems that help to store and access information of many services and products. Even though it stores huge data, HIMS in Turkey are used for basic fundamental reports. Thus, while there is a high demand for well-processed information; perception of metadata to point out evident relationships for healthcare data is lacking and there is need to benefit from HIMS in a more elaborate way. By use of data mining techniques one can process multiple databases, summarize big data from different perspectives and convert data into necessary information.

K-means cluster method can be used to extract relationships from the big multivariate data in HIMS. The method mentioned require the user to provide the initial parameters (a number of clusters and initial seeds) and this requirement becomes a disadvantage for the method.

This study, attempts to solve this problem to propose a method for finding optimal number of cluster and initial seeds. The software which includes the algorithm named NAMGY (Neighborhood And Midpoint Gain Yield) designed for this purpose will be able to report the information that the hospital administrators need using HIMS data more easily and large amount of cumulative data will be available.

KEYWORDS : HIMS, K-means cluster algorithm, data mining, number of cluster, initial seeds

1. GİRİŞ

Sağlık hizmetleri , toplumun sağlık bakımını gereksinimlerine ve isteklerine göre her türlü tedavi edici, koruyucu, rehabilitasyon sağlık hizmetlerini sağlamak üzere ülke genelinde örgütlenmiş sistemlerdir (1).

Sağlık hizmetlerinde rekabetin artması, toplumun sağlık hizmetlerindeki beklentilerinin yüksek olması sağlık kuruluşlarının hizmetlerini daha verimli ve kaliteli sunmalarını zorunlu kılmıştır (2). Teknolojinin gelişmesine bağlı olarak tıbbi verilerin hızla artması ve çeşitlenmesi sonucu sağlık kurumlarının sağlık hizmetlerini sunabilmeleri için Hastane Bilgi Yönetim Sistemlerine (HBYS) ihtiyaç vardır. Günümüzde sağlık kurumları, HBYS'leri sadece faturalandırma işlevlerini içeren sistemler olarak kullanılmak yerine teknolojik gelişmelerin artmasıyla kağıtsız hastane (Dijital Hastane) özelliklerini taşıyan sistemler olarak tasarlanmaya başlanmıştır. Dijital hastane; çok çeşitli teknolojik cihazlarla entegre olmuş HBYS'leriyle sağlık kurumlarının işletilmesinde yapılan tüm işlemlerini dijital ortamda kağıtsız olarak yapılabilen hastanelerdir. TC. Sağlık Bakanlığı Stratejik Planı (2013-2017) raporunda; Sağlık Bakanlığına bağlı dijital hastane oranı % 90 olarak hedeflendiği belirtmiştir (3).

Sağlık kurumları, sağlık hizmetlerinin faturalandırma süreçlerini kontrol etmek amacıyla kullandıkları ilk sistemlerin bu amacı yeni nesil HBYS'lerinde de ilk sırada yer almaktadır. Ancak, sağlık kurumları HBYS'deki verileri değerlendirirerek, geçmişten çıkarımlar yapıp sağlık hizmetlerini daha iyi sunabilecek stratejiler geliştirebilir.

Ülkemizdeki HBYS'ler gelişmiş teknolojilerle entegre edilmesinde ileri bir düzeydedir. Ancak sistemlerin ana işlevi hala toplanan büyük boyuttaki verilerin depolanmasıdır. Halbuki, sağlık alanındaki gerek ülkenin sağlık politikasını belirlemede gerek üniversitelerde ve araştırma merkezlerinde yapılan bilimsel çalışmalarda HBYS'ler veri kaynağı olarak etkin bir şekilde kullanılabilir.

Sistemlerde toplanan ham veri birçok işleme katmanlarından sonra sağlık kurumlar için bir artı değer oluşturabilir. Ancak, HBYS'indeki karmaşık verilerin anlaşılabilir hale getirilme süreci bir çok alanda uzmanlık ve deneyim gerektiren işlemleri içerir. Farklı kaynaklardan üretilen, dinamik olarak artan heterojen yapıdaki verilerin analizinde geleneksel tekniklerin yetersiz kalması, verilere uygun tekniklerin HBYS'lerine adapta

edilmesini zorunlu kılmaktadır. Tıp alanında, büyük veri niteliğindeki verilerle yapılan çalışmalarda veri madenciliği yaklaşımlarının kullanılması, bilinmeyen örüntülerinin ortaya çıkarılmasını sağlayacak ve toplumun sağlığını ilgilendiren stratejileri daha doğru, hızlı ve efektif olarak tasarlamada etkili olacaktır (4). Ayrıca sağlık alanında kümeleme algoritmaları içeren çalışmaların yapılması, toplumun yaşam kalitesinin artmasını, hastalık risklerin belirlenmesi gibi sağlık alanındaki bulgular tıptaki gelişmelere katkı sağlayacaktır (5).

Kümeleme algoritmalarından biri olan K-ortalama kümeleme algoritması uygulama kolaylığı nedeniyle sağlık alanında yaygın olarak kullanılan bir algoritmadır. Ancak kümeleme sonuçlarında etkili olan küme sayısı ve başlangıç merkezlerinin parametre olarak belirlenmesi bu algoritmanın bir dezavantajıdır (6). Probleme çözüm olabilecek algoritmaların geliştirilmesi K-ortalama algoritmasının kullanımını artıracak, sağlık alanındaki çalışmalarda kümeleme sonuçlarının daha doğru yapılmasına katkı sağlayacaktır.

İlgili literatürlerin taranması tıp alanındaki bilimsel araştırmalarda veri madenciliği yöntemlerinin yaygın olarak kullanıldığını göstermektedir. Konuyla ilgili kaynaklar, tıp alanında veri madenciliğinin kullanılması ve K-ortalama algoritmasındaki farklı yaklaşımları içeren çalışmalar olarak iki ayrı başlık altında incelenebilir.

1.1. Tıp Alanında Veri Madenciliği Yöntemlerinin Kullanıldığı Çalışmalar

[Daghistani and Alshammari (7)] çalışmasında; Suudi Arabistan'daki nüfusun % 66 'sından fazlasının yaşadığı en büyük nüfuslu bölgedeki gerçek sağlık hizmeti verilerini içeren MNGHA veritabanı üzerinde yapılmıştır. Diyabet hastalığında risk kabul edilen 18 değişkenli veri setinde veri madenciliği tekniklerinden Özdüzenleyici Haritalar (Self-Organizing Maps (SOM)) ve Karar Ağacı teknikleri uygulanarak diyabet hastaları tespit edilmeye çalışılmıştır. Çalışma sonucuna göre oluşturulan veri madenciliği modelinin sağlık çalışanlarına, diyabet hastalarının belirlenmesinde yardımcı olabileceğini gösterdiği belirtilmiştir. Ayrıca, diyabet hastaları genellikle hastalığın ilerleyen aşamasına kadar veya komplikasyonların gelişimine kadar tespit

edilemediğinden, modellemenin, gelecekte diyabet için bir kontrol planı oluşturmak amacıyla kullanılabileceği belirtilmiştir.

[Patel and Patel (8)] çalışmasında; Veri Madenciliği tekniklerinden sınıflandırma, kümeleme, ilişkilendirme algoritmalarının hastalıkları önceden tahmin etmede, hastalık düzeylerinin ve sınıflandırılmalarını içeren çalışmaları vurgulanmıştır. HBYS'lere çeşitli tıbbi cihazların entegre olması büyük veri niteliğindeki biyomedikal verilerinde artış görüldüğü belirtilmiştir. Hasta hizmet sunumunda katkı sağlayacak bu verilerin analizlerinde veri madenciliği yöntemlerinin kullanılmasına ihtiyaç duyulmaktadır.

Tıbbi verilerin uygun veri madenciliği teknikleriyle analiz edilmesi ile insan sağlığı için önemli olan doğru tanının, erken teşhis edilen bir çok hastalığın önlenmesi ve tedavi süreçlerinde karar destek sistemleriyle önemli katkı sağlandığı belirtilmiştir.

Patel and Patel çalışmasında özetlenen çeşitli hastalıkların belirlenmesinde kullanılan verimadenciliği yöntemleri (Çizelge 1.1.) de verilmiştir.

Çizelge 1.1 Farklı Hastalıklar İçin Kullanılan Veri Madenciliği Uygulamaları

HASTALIK	VERİ MADENCİLİĞİ YÖNTEMLERİ
Konvansiyonel Patoloji Verileri	Yapay sinir ağlarını kullanarak hastalık eğilim oranlarının saptanması
Koroner kalp hastalığı	C4.5, C5 ve CART gibi Karar Ağacı Algoritmalarını kullanan IPDre3d modelleri oluşturarak hastalığın tespit edilmesi
Hastalık Sıklığı	Apriori algoritması kullanılarak coğrafi alanlarda hastalık sıklıklarının belirlenmesi
Cilt hastalıkları	Yapay Sinir ağı sınıflandırma yöntemleri ve karar ağaçları algoritması entegre edilmiş

	modelleme ile cilt hastalıklarının düzeyinin belirlenmesi
Diyabet	Genetik Algoritma yöntemlerini kullanarak medical verilerinin sınıflandırılması.
Biyomedikal Rezonans Görüntülemenin Sınıflandırılması (fMRI)	Sınıflandırma ve Kümeleme yöntemleri ile fMRI görüntülerinin sınıflandırılması.
Göğüs Hastalıkları	Yapay Sinir Ağı kullanarak oluşturulan model ile hastalığın sınıflandırılması.
Kronik Hastalıklar	Apriori algoritması ile hastalıkların tahmin edilmesi.
Meme kanseri	K-Ortalama algoritması tıbbi verilere uygulanarak hastalığın doğru olarak sınıflandırılması

[Niaksu 9]] tez çalışmasında; veri madenciliği yöntemlerinin tıp ve sağlık alanlarında uygulanmasına yönelik metodolojiyi geliştirilerek, sağlık alanında yapılan veri analizinin etkinliğinin artırıldığı ortaya konmuştur. Veri Madenciliği için endüstri standardı Çapraz Sektör Standart İşlemine dayanan yeni bir CRISP-MED-DM uygulama metodolojisi geliştirilerek, kardiyoloji ve onkoloji alanlarındaki tahmini modellemeyle yapılan uygulamaların sonuçların başarılı olduğu gösterilmiştir. Ayrıca, aort kapak stenoz derecesinin diagnostiklerinin yarı otomasyonunu sağlayan yeni bir kan akımı ekokardiyografi görüntü işleme tekniğinin geliştirildiği ifade edilmiştir.

Veri madenciliği yöntemlerinin tıbbi alanda uygulanmasıyla sağlık alanında yeni fırsatların sağlandığı belirtilmiştir.

[Gayathri and Jaisankar (10)] çalışmanın, tıbbi verileri içeren veri tabanlarındaki verilerin bilgiye dönüştürme sürecinde uygun veri madenciliği yöntemlerinin seçimini, gerekliliğini ve sağlık hizmetlerine etkisini belirlemek amacıyla yapıldığı vurgulanmıştır.

Veri madenciliği ile yapılan araştırmalarda, kapsamlı klinik veri tabanlarından güçlü klinik ve idari karar vermeyi destekleyebilecek bilgilerin oluşturulması ile yeni sağlık alanında bilgilerin ortaya çıkarılabileceğine, aksi durumda büyük veri niteliğini taşıyan veri tabanlarının atıl kalabileceğine dikkat çekilmiştir.

Dünyada sağlık alanında büyük değişimlerin oluşturulabilmesi için veri madenciliği yöntemlerini tıbbi veriler üzerinde daha yaygın kullanılması önerilmiştir.

[**Solmaz ve ark. (11)**] çalışmalarında; tıp alanındaki tedavilerde, insandan kaynaklı hataları azaltmak, teşhis ve tedavilerde hekime yardımcı olacak karar destek sistemi olarak kullanılabilecek uzman sistemlere örnek olması amacıyla tasarlandığı belirtilmiştir.

Sistem, tiroit hastalığına ait veriler üzerinde farklı sınıflama ve kümeleme yöntemleriyle test edildiği ifade edilmiştir.

Sınıflama algoritmalarından ; yapay sinir ağları, SVM (Support Vector Machines) Destek Vektör Makineleri, Lineer Diskriminant Analizi, kümeleme algoritmalarından; FCM (fuzzy c-means) ve K-ortalama kümeleme yöntemlerinin kullanıldığı ve tüm yöntemler için kabul edilebilir başarılı sonuçların alındığı belirtilmiştir.

[**Yurtayı ve ark. (12)**] yaptıkları çalışmada; veri madenciliği hiyerarşik kümeleme yöntemlerinden AGNES (AGglomerative NEsting) algoritması laboratuvardaki kan değerleri verileriyle çalıştırılarak dendogram grafiği ile okunabilirliği sağlanmış uzman doktor yardımıyla, hemogram test sonuçlarıyla karşılaştırılıp kansızlığa sebep olan etkenler tesbit edilmeye çalışıldığı bildirilmiştir.

Ayrıca veri madenciliği yöntemlerini kullanarak tıbbi alanlardaki teşhis ve tedavi süreçlerine destek sağlanmasının amaçlandığı belirtilmiştir.

[**Seğmen and Uyar (13)**] çalışmalarında; elektronik sağlık sistemlerinin bir parçası olarak kullanılan tanı karar destek sistemlerinin güvenilirliğinin sağlanması için model seçiminin önemi vurgulanmıştır. Kamu sağlık kurumlarındaki tıbbi veri tabanları üzerinde yaygın olarak kullanılan sınıflandırma algoritmaları uygulanmış, deneysel sonuçlara göre, tüm veri setleri için başarılı tek bir algoritma bulunmamasına karşın, Yapay Sinir Ağları ve Naive Bayes yöntemlerinin nispeten daha yüksek performans sağladığı belirtilmiştir.

[**Dolgun ve ark. (14)**] çalışmalarında; Veri madenciliği sınıflandırma yöntemleri uygulamalarında, bağımsız değişkenlerinin dağılımının, ölçüm türlerinin,

bağımlı değişkenle arasındaki ilişkisinin, bağımlı değişkenin yapısal farklılıklarının, örneklem büyüklüğünün model başarısına olan etkilerini belirlemek amacıyla sınıflandırma performans ölçüm kriterleriyle değerlendirildiği belirtilmiştir.

[İşler ve Narin (15)] çalışmalarında; Konjestif Kalp Yetmezliği hastalarını belirlemek amacıyla, kalp hızı değişkenliği veri seti, veri madenciliği WEKA programı K-ortalama yöntemiyle analiz edilmiş küme sayısının 4 olduğu durumda başarı oranının en yüksek olduğu belirtilmiştir.

[Turhan ve ark. (16)] Veri madenciliği tekniklerinden kümeleme analizi yöntemiyle Karadeniz Teknik Üniversitesi Tıp Fakültesi Farabi Hastanesi'nde yatan 9777 hastanın verileriyle, Potasyum (K) ve hemoliz indeksi (HI) arasındaki ilişki araştırılmış, ayrıca ; potasyum ve hemolizden etkilenen diğer testler için de düzeltme faktörü geliştirilmesinde kümeleme yönteminin, ön çalışma olarak kullanılabileceği vurgulanmıştır.

[Irmak ve ark. (17)] çalışmalarında; hastane veritabanı verilerini bazı ön işlemlerden geçirildikten sonra yapay sinir ağları teknikleri ve üstel düzeltme modelleri kullanılarak, gelecekteki hasta yoğunluklarının tahmin edilebilmesi için modeller oluşturulmuş, Winters Additive üstel yumuşatma modelinin gerçek değerleri en iyi tahmin yapan model olduğu belirtilmiştir.

[Güllüoğlu (18)] çalışmasında; sağlık sektöründe uzman kişiler veri madenciliği içeren bir çalışmanın planlanması hakkında bilgilendirmek, uzmanların karar verme süreçlerine destek vermek amacıyla tasarlanmış, aynı zamanda; sağlam kişilerin ileride yakalanabileceği kanser tipini ve yakalanma ihtimalini belirleyerek, riskli ya da sağlıklı olarak gruplandırmak örneği üzerinde uygulama gösterilmiştir.

[Alapont et al. (19)] çalışmalarında; hastane bilgi sistemlerindeki farklılıkların çok fazla olmasına karşın veri madenciliği modellerinde benzerlik olduğu yapılan ön çalışmalarla tespit edilmiş, HBYS verilerini bilgilere dönüştürecek otomatik veri madenciliği yöntemlerini kullanan uzmanlaşmış yazılımlar oluşturulup, bilgilerin hastane yönetiminde kullanılması gereği vurgulanmıştır.

[Sunay (20)] Çalışmada; veri madenciliği sınıflandırma ve kümeleme yöntemleri kullanılmış, kalpteki fokal ventriküler aritmi kaynağının yerinin tespiti, farklı iki veri seti üzerinde üç farklı kümeleme ve beş farklı sınıflandırma algoritmaları uygulanarak yapılmıştır. Aynı zamanda fokal aritmi kaynağı lokalizasyonuna yönelik alternatif

yaklaşımlar geliştirilerek, kalp haritalama yöntemlerinin dezavantajlarının giderildiğini belirten doktora tez çalışması yapılmıştır.

[**Danacı ve ark. (21)**] Çalışmalarında; meme kanserinde erken tanıyı belirlemek amacıyla Irvine California Üniversitesi veri madenciliği havuzundan 659 hastaya ait meme altı doku örnekleri veri seti üzerinde Xcyt örüntü tanıma sistemiyle veri madenciliği programları Weka kullanılarak yapılan çalışmada; veri madenciliği yöntemlerinden C4.5 karar ağacı algoritmasının hastalık tanı ve teşhisinde %97,43 doğruluk oranı ile meme kanseri teşhis edilebildiği ve hastalığın erken tanı tespitinde büyük rol oynayacağı vurgulanmıştır.

[**Aksoy (22)**] Tez çalışmasında; dekompresyon kırığı sınıflandırılması, dalgıçların acil durum ağı'nın dalış yaralanmaları bildirim form verileriyle farklı kümeleme algoritmaları (COBWEB, EM, K-ortalama) uygulanmıştır. Bu sınıflama klasik sınıflandırma yöntemleri ve kliniksel sonuçlar karşılaştırılarak kümeleme yöntemleri ile oluşturulan sınıflamaların klasik sınıflandırma sonuçları ile uyumlu olduğu belirtilmiştir.

[**Doğan ve Türkoğlu (23)**] çalışmasında; demir eksikliği anemisi teşhisinde, kan biyokimya parametre verileri kullanılarak veri madenciliği tekniklerinden karar ağaçları algoritması ile oluşturulan sonuçların hekimin belirttiği kliniksel sonuçlarla tamamen örtüştüğü saptanmış, karar destek sisteminin kullanılması önerilmiştir.

[**Gülten ve Doğan (24)**] çalışmasında; hiperlipidemi teşhisinde, kandaki biyokimya parametre değerleri ile veri madenciliği yöntemlerinden genetik algoritma kullanılarak, hekime yardımcı olacak karar destek sistemi oluşturulmuş ve bu sonuçların hekimin verdiği kararlarla tamamen örtüştüğü gösterilmiştir.

[**Dinçer ve Duru (25)**] çalışmalarında; gırtlak kanseri ameliyatı olan hastaların geçmiş verilerini kullanarak veri madenciliği K-ortalama kümeleme algoritmasını içeren yazılım geliştirmiş ve bu yazılım tıp doktorlarının tıbbi kararları vermelerine destek sağlayacağını belirtmişlerdir. Çalışmada geleceğe dair kliniksel bilgilerin elde edilmesinde, geçmiş tıbbi verilerin veri madenciliği yöntemleriyle bilgiye dönüştürülmesinin önemi vurgulanmıştır.

1.2. K-Ortalama Kümeleme Yönteminin Küme Sayısı, Başlangıç Merkezlerini Belirlemek için Yapılan Çalışmalar

K-ortalama yöntemi ile ilgili metodoloji çalışmaları küme sayısı belirleme ve başlangıç merkezleri belirleme çalışmaları olarak iki konuya odaklanmıştır. Küme sayısı çalışmaları nispeten az olup (27) , genelde araştırmaların başlangıç merkezlerini belirlemede yoğunlaştığı (26,28,29,30,31) görülmektedir.

Konuyla ilgili yapılan bazı çalışmalar aşağıda verilmiştir.

[Dalhatu ve Sim (26)] çalışmalarında: K-ortalama algoritmasında başlangıç merkezlerini belirlemede noktaların yoğunluğuna dayalı algoritma K-ortalama başlangıç merkezlerinin rasgele seçilmesi problemine çözüm olabileceği düşünülerek geliştirilmiştir.

Dalhatu ve Sim'in önerdiği algoritmada bazı eşitlikler kullanılmıştır.

- Veri setindeki her noktanın birbirine uzaklıkları (1.2.1) eşitliğine göre bulunur.

$$d(c,p) = \sqrt{\sum_{i=1}^n (c_i - p_i)^2}$$

$d(c,p)$: c merkez ile p nokta arasındaki uzaklık (1.2.1.)
 c : Küme merkezi
 p : Veri setindeki nokta

- Veri setindeki her noktanın yoğunluğu (1.2.2.) eşitliğine göre bulunur.

$$\rho_i = \frac{1}{d(c,p)+1}$$

ρ_i : i . noktanın yoğunluğu (1.2.2.)

- Noktalar arasındaki maksimum ve minimum uzaklık değerleri kullanılarak (1.2.3.) eşitliğine göre T eşik değeri belirlenir.

$$T = [\text{sum}(\text{Max}(\text{dist}) - \text{Min}(\text{dist})) * \Gamma(0.5+1)] / (n\sqrt{n})$$

(1.2.3.)

T : Eşik değeri

n : Veri setindeki kayıt sayısı

$\text{Max}(\text{dist})$: Noktanın diğer noktalar ile arasındaki en büyük uzaklık

$\text{Min}(\text{dist})$: Noktanın diğer noktalar ile arasındaki en küçük uzaklık

$\Gamma()$: Gama fonksiyonu

Dalhatu ve Sim'in önerdiği algoritmanın adımları;

1. Veri seti, boş matris (TMAT) ve küme sayısı (K) okutulur.
2. Noktalar arasındaki uzaklıklar bulunur.
3. Her noktanın yoğunluğu bulunur.
4. Eşik değeri (T) bulunur.
5. Yoğunluğu en büyük olan nokta merkez (C) olarak TMAT matrisine kopyalanır.
6. Veri setindeki diğer noktalar ile C merkez arasındaki uzaklıklar bulunur.
7. Veri setinde C dışındaki her noktanın (p) $d(C,p)$ uzaklıkları eşik değerinden (T) küçük ise TMAT matrisine kopyalanır.
8. Veri setindeki kalan noktalar ile TMAT matrisindeki C noktası hariç olmak üzere diğer noktalar arasındaki uzaklıkları bulunur.
9. Uzaklıkları T eşik değerinden küçük eşit olan veri setindeki noktaları TMAT matrisine kopyalanır.
10. Veri setindeki kalan noktalar ile TMAT matrisindeki noktalar arasındaki uzaklık (C hariç) T eşik değerinden küçük eşit kalmayana kadar 8-9 adımları yapılır.
11. TMAT içindeki tüm noktalar veri setinden silinir.
12. 5-11 arası adımları TMAT içindeki C ler küme sayısı (K) olana kadar uygulanır.
13. TMAT içine kopyalanan k tane nokta başlangıç merkezleri olarak mevcut K-ortalama algoritmasında kullanılır.

Önerilen algoritmanın uygulama sonuçlarına göre, K-ortalama algoritmasına göre daha iyi kümeleme performansı gösterdiği belirtilmiştir.

[Çolak ve ark. (27)] çalışmalarında ; K-ortalama algoritmasında küme sayısını otomatik bulunması için önerilen algoritma ;

1. Küme sayısı $k=1$, $SSEMIN_1=0$, ESIK değeri belirlenir,
2. K-ortalama algoritmasını çalıştır, $SSEMIN=ToplamHataKare$
3. $FARK=ABS(SSEMIN-SSEMIN_1)/ SSEMIN > ESIK$ ise $k=k+1$;
 $SSEMIN_1=SSEMIN$; 2. Adıma git.
4. İşlemi bitirilir.

Uygulama Matlab programında farklı sayısal ve görüntüsel veri setleri üzerinde çalıştırılarak elde edilen kümelemelerin performansının iyi olduğu belirtilmiştir.

[**Godara and Sharma (28)**] çalışmasında ; K-ortalama başlangıç merkezleri hiyerarşik yapı yöntemlerinden en küçük örten ağaç (Minimal Spanning Tree (MST)) sınıflandırma yöntemi ile bulunduktan sonra K-Ortalama algoritma uygulanıp kümeleme işlemi tamamlanıyor. Önerilen yöntemle kümeleme sonuçları performansının klasik K-Ortalama algoritmasına göre daha iyi olduğu gösterilmiştir.

[**Kedar and Sawant (29)**] çalışmalarında; K-Ortalama algoritmasının başlangıç merkezlerinin rasgele seçilme problemine çözüm olacak yeni bir algoritma önerilmiştir. Önerilen algoritma ile en uygun başlangıç merkezleri seçilerek küme sonuçlarını daha iyi ve iterasyon sayısının daha az olduğu belirtilmiştir.

Kedar ve Sawant'ın önerdikleri algoritma aşamaları;

- Veri kümesinin toplam kayıt sayısı N , veri kümesinin her i noktasının diğer $N-1$ tane noktalar arasındaki uzaklıkların toplamı hesaplanır.
- Tüm noktaları kendilerine ait uzaklıklar toplamına göre sıralanır.
- Kullanıcının girdiği küme sayısına (k) göre tüm veri kümesini eşit sayıda kayıtlar içermek üzere parçalara ayrılır.
- Oluşturulan k tane kümenin ilk noktası o kümenin merkezi olarak belirlenir.
- Başlangıç merkezleri belirlendikten sonra K-ortalama algoritması uygulanır.

[**Singh and Kaur (30)**] çalışmalarında ; K-ortalama başlangıç merkezlerinin seçimini sağlayan algoritma tasarlanmıştır.

Singh ve Kaur'un önerdiği algoritma adımları aşağıdaki gibidir.

1. Küme sayısı (k) okutulur.
2. Tüm veri setinin ortalaması birinci merkez olarak alınır.
3. Veri seti iki parçaya bölünür.
4. Her bir parçanın ortalaması ikinci ve üçüncü merkez olarak alınır.
5. Yukarıdaki işlemler küme sayısı k tane oluncaya kadar devam edilir.

Başlangıç merkezlerinin çok basit hesaplanması yöntemin en önemli avantajı olduğu belirtilmiştir.

Singh ve Kaur'un önerdiği algoritma kümeleme sonuçlarına göre; küme içinde noktalar ile küme merkezi arasındaki uzaklıklar toplamını küçülttüğü buna bağlı olarak kümelerin toplam karesel hatanın daha küçük olduğu, bu durumun, önerilen algoritmanın

rastgele başlangıç merkezlerini seçen orijinal K-ortalama algoritmasından daha verimli olduğunu gösterdiği belirtilmiştir.

[Agha and Ashour (31)] çalışmalarında; K-ortalama algoritmasında başlangıç merkezlerini belirlemek için özellikle çok boyutlu veri setler için elverişli ELAgha initialization olarak adlandırdıkları algoritmayı önermişlerdir.

Agha ve Ashour' un önerdiği algoritmanın adımları;

1. Öncelikle veri sınırları bulunur. Hücre genişliği, yüksekliği (X_w, Y_w) olmak üzere; her bir boyutun en büyük değeri ile en küçük değeri arasındaki fark k küme sayısına bölerek bulunur.
2. (X_w, Y_w) oranları kullanılarak sol üst ile sağ alt köşegen doğrusu üzerinde bulunan noktalar üzerinden rasgele seçilerek başlangıç küme merkezleri belirlenir.

Başlangıç merkezlerinin seçimini güdümlü rasgele teknik (guided random technique) algoritmasıyla elde edilen sonuçların, kümeleme kalitesinin özellikle karmaşık veri setlerinde geleneksel rasgele seçim yöntemine göre büyük bir farkla daha iyi olduğu belirtilmiştir.

1.3. Çalışmanın Amacı

Bu çalışmanın birincil amacı Hastane Bilgi Yönetim Sistemi (HBYS) verilerinin bilgiye dönüştürme süreçlerine destek sağlamaktır.

Bu bağlamda hedeflenen, K-ortalama kümeleme algoritmasının HBYS verileri üzerinde uygulanmasında, parametre değerlerini belirleme konusunda, bilgi ve deneyimi olmayan kullanıcılara destek sağlamak üzere: parametre değerlerinin algoritma tarafından belirlendiği uygun yöntemi geliştirmek ve algoritmanın uygulanabileceği arayüz yazılımını oluşturmaktır.

Böylelikle K-ortalama algoritmasının HBYS'de kullanım kolaylığı sağlanarak, bu sistemlerde toplanmış büyük miktardaki verilerin bilgiye dönüştürülme sürecine destek verip verilerin okunabilirliği artıracaktır. Sağlık kurumlarının verimli yönetilmesine ve sağlık hizmetlerinin efektif sunulmasına katkı sağlayarak kurumlar için katma değer oluşturacaktır.

2. GENEL BİLGİLER

2.1. Sağlık Hizmetleri, HBYS ve Büyük Veri

Sağlık sisteminin önemli bileşeni olan hastaneler, tıp merkezleri, aile sağlık merkezleri toplumun çeşitli sağlık hizmetlerini karşılaması ile önemli bir misyon yüklenen hizmet odaklı işletmelerdir.

Ülkemizde sağlık hizmetleri, kamu, üniversite, vakıf üniversitesi, özel olmak üzere farklı statüdeki sağlık kurumları tarafından verilmektedir. Sağlık kurumları yapıları gereği hizmet sunumlarında farklı alanlardaki paydaşlarla çalışması, kurumlarının yönetiminde eğitilmiş ve profesyonel ekiplerle çalışılmasını zorunlu kılmaktadır. Sağlık hizmetleri, insan odaklıdır ve toplum sağlığını ilgilendirir. Bu nedenle kurumların hizmetin sürekliliğinin sağlanması, bilgi teknolojilerinde dijitalleşme, tıbbi ekipmanların modernleşmesini sağlayacak şekilde düzenlemeler yapmaları gerekir (2).

2.1.1. Hastane Bilgi Yönetim Sistemi (HBYS)

Hastane Bilgi Yönetim Sistemi; elektronik tıbbi kayıtları, spesifik biyomedikal, idari, finansal kayıtları içeren veritabanlarıyla, tıbbi ve iletişim cihazlarla, çeşitli ağ, bilişim teknolojileriyle entegre olmuş bütünleşik sistemlerdir (32).

Sağlık kurumlarının varlıklarını sürdürebilmeleri için sağlık hizmetlerinin efektif, kaliteli ve verimli sunabilme gerekliliği HBYS'lerdeki beklentileri artırmaktadır (33).

HBYS'lerdeki hastaya ait doğru bilgilere ulaşma sürelerinin ve tıbbi hataların azalması, kurumun yönetilmesinde dinamikliliğin devamlılığına destek sağlayarak sağlık hizmeti sunum kalitesini artıracaktır (32).

HBYS'ler aynı zamanda ;

- Uluslararası kodlama standartları
- Ağ iletişim standartları
- Elektronik veri aktarımı standartları

- Elektronik sađlık kayıt standartları
- Mesajlaşma standartları

olmak üzere temel bilgi standartlarını içermelidir (34).

Hastane Bilgi Sisteminin işlevlerinden biri sađlık kurumlarının yönetimine operasyonel, taktiksel ve stratejik seviyelerde destek sađlamasıdır. Ayrıca hastanedeki faaliyetlerin aktif olarak yapılabilmesi için gerekli olan bilgileri toplamak, depolamak, işlemek, iletişimi sađlamak ve tüm yetkili kullanıcıların gereksinimlerini karşılamalıdır.

HBYS'lerinde oluşturulan Elektronik Tıbbi Kayıtlar yaygın olarak dünyada kullanılmaktadır. HBYS'lerin büyük miktardaki tıbbi verileri içermesi, sistemlere tıbbi araştırmalar için veri kaynağı olma niteliğı kazandırmıştır (35).

HBYS'leri donanımsal ve yazılımsal başlıklar altında gruplanan bileşenler (Şekil 2.1) de gösterilmiştir.

Sunucular, bilgisayarlar, tıbbi cihazlar, ađ iletişimi sađlayan aktif cihazlar, mobil cihazlardan oluşan teknolojiler donanım kısmını oluşturur.

Sistemlerin kullanılabilirliğini sürdürmek ve artırmak için çok farklı alanlarda işlem yapabilmesi için programlanmış modüller yazılım kısmını oluşturur.



Şekil 2.1 : HBYS Bileşenleri

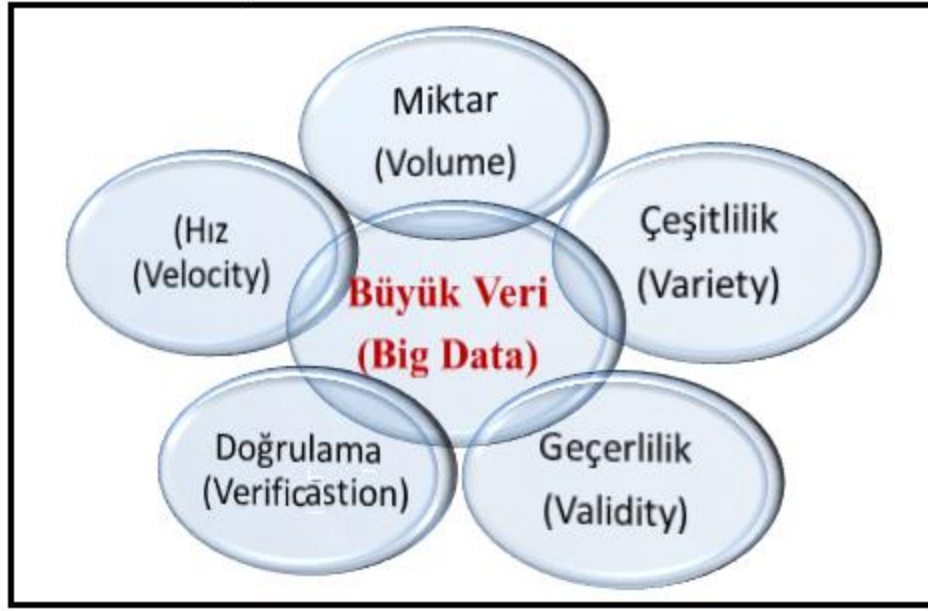
Çizelge 2.1. : HBYS Yazılımsal Modülleri

HASTANE BİLGİ YÖNETİM SİSTEM MODÜLLERİ		
TIBBİ İÇERİKLİ	FİNANSAL VE YÖNETİMSEL İÇERİKLİ	STRATEJİK KARAR DESTEK SİSTEMLERİ
- Hasta bilgi yönetim sistemi - Laboratuvar bilgi sistemi - Radyoloji bilgi sistemi ve PACS - Randevu sistemleri (sms, internet, web, çağrı merkezi vs.) - Klinik Karar Destek Sistemleri - Kliniklere özgün tıbbi modüller ve Karar destek sistemleri - Kan Bankası - Ameliyathane - Dış bağlantı uygulamaları (Medula, Mernis, Sağlık-Net, E-Nabız, E-Reçete, Armakom, Anlaşmalı kurum sistemleri, TİG)	- Tahakkuk ve Fatura İşlemleri - Doküman yönetim sistemi - Dijital arşiv - Kapalı sistem ilaç ve tıbbi malzeme takibi.(barkod ve RFID entegre) - Satın Alma, Stok , Demirbaş takibi - Tıbbi ve İdari Talep Süreçleri - İnsan Kaynakları Yönetimi - Hastane İdari Yönetim Süreçleri	- Gelir Gider Kaynakların Belirlenmesi - Programlama ve Planlama - Performans Yönetimi

2.1.2. Büyük Veri (Big Data)

HBYS’inde toplanan tıbbi verilerin çok yoğun, farklı alanlardan büyük bir hızla ve farklı formatlarda olması, sağlık alanında büyük veri sistemlerini kullanma zorunluluğunu getirmiştir (35).

Büyük Veri (Big Data) ; Miktar, Hız, Çeşitlilik, Doğruluk, ve Geçerlilik olmak üzere beş bileşen özelliklerini taşıyan veriler olup, bilgiye ulaşmanın gücü olarak kabul görmektedir (36). Büyük verinin bileşenleri (Şekil 2.2) de gösterilmiştir.



Şekil 2.2: Büyük Veri Bileşenleri

Büyük verinin bileşenleri aşağıdaki gibi açıklanabilir.

Miktar (Volume): Büyük veri olarak nitelendirilen veriler miktarları TeraByte ları aşan verilerdir. Büyüklüğe bir emsal olmak üzere, tüm internette toplanan verilerin 2019 da 1 Yottabyte (10^{12} TeraByte) ulaşılabilceği söylenebilir.

Hız (Velocity) : İnternet üzerinden giderek artan miktarda yapılandırılmamış verilerin akış hızı olarak açıklanabilir. Veri akış hızı bir saniyede üretilen verinin bit ya da byte sayısı ile gösterilir.

Çeşitlilik (Variety); Çeşitlilik resimler, videolar, sensör verileri, sağlık verileri, resmi verileri, web dosyaları, araştırmalar gibi farklı kaynaklardan yapılandırılmamış veri biçiminde oluşturulabilen verilerin bütünleşik olarak bulunabilmesidir.

Doğrulama (Verification) : Miktarı hızlı büyüyen verilerin akışı sırasında verilerin gürültülü, eksik veya ihlal olabileceği düşünülerek verilerin doğruluğunun kontrolü, güvenliğinin sağlanması, doğru kişiler tarafından görünebilirliğinin sağlanmasıdır.

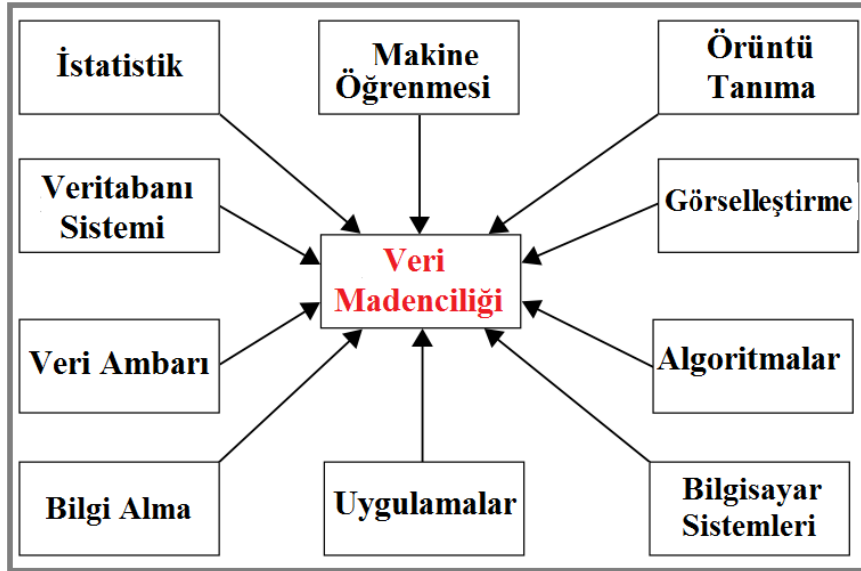
Geçerlilik (Validity) : Aynı bilginin farklı formatlarla oluşan verilerle desteklendiği araştırılıp verinin geçerliliğinin ve güncelliğinin sağlanmasıdır.

2.2. Büyük Veri Analizi

Dünyada her gün çok miktarda toplanan verilerin analizi önemli bir ihtiyaç haline gelmiştir. Büyük veri niteliğindeki bu verilerden bilinmeyen kalıpları keşfederek bilgiye ulaşmanın son yıllardaki en çok kullanılan tekniği veri madenciliğidir. Veri madenciliği, büyük hacimli veri ambarlarında toplanan verilerin üzerinde bilgisayar teknolojileriyle, uygulamaya dayalı çeşitli disiplin yöntemlerini kullanarak verilere işlerlik kazandırarak farklı nitelikteki kuruluşlara katma değer sağlar (37).

Veri madenciliği, veri tabanı sistemleri, veri ambarı sistemleri, istatistiksel yöntemler, matematiksel algoritmalar, yapay sinir ağları, makine öğrenmesi, görselleştirme, görüntü işleme gibi çok disiplinli teknoloji ve alanları içerir. (Şekil 2.3) de gösterilmiştir.

Veri Madenciliği yöntemlerini içeren birçok ticari ve açık kaynak olmak üzere programlar mevcuttur. Açık kaynak kodlu veri madenciliği programlarından bazıları RapidMiner(YALE), WEKA ve R dir.



Şekil 2.3 : Veri Madenciliğinin Yararlandığı Alanlar

2.2.1. Veri Madenciliğinin Sağlık Alanında Kullanımı

Sağlık hizmeti sunumunda, hasta memnuniyetini karşılamayı amaçlayan hasta merkezli hizmet sunumu, dünyada ve ülkemizde yaygın olarak uygulanmaktadır. Bu dönüşüm, sağlık hizmet sunumunun hızlı, doğru ve zamanında verilmesinin gerekliliğini ön plana çıkarmıştır (40).

Toplumsal sağlık hizmetinden beklentilerinin yükselmesi, sağlık hizmetlerinde digitalleşme, tıbbi alanlarda toplanan verilerin hızla büyümesi ve sağlık hizmeti sunumunda veri madenciliği yöntemleriyle oluşturulan karar destek sistemlerini içeren HBYS'lerin kullanılması sağlık kurumu yöneticilerine ve sağlık personeline yardımcı olacaktır.

Tıp alanında, büyük veri niteliğindeki verilerle yapılan çalışmalarda veri madenciliği yaklaşımlarının kullanılması, bilinmeyen örüntülerinin ortaya çıkarılmasını sağlayarak, toplumun sağlığını ilgilendiren stratejileri daha doğru, hızlı ve efektif olarak planlanmasında etkili olacaktır.

Sağlık alanında yapılan veri madenciliği çalışmalarıyla aşağıda belirtilen önemli katkılar sağlanabilir.

- Tıbbi hataların azalmasını
- İlaç kullanım hata ve yan etkileri için erken uyarı sistemleri
- Risk sinyalleri belirlenerek hastalıkların önlenmesi
- Karar destek sistemi ile erken teşhis konulması
- Hasta odaklı tedavi ile sağlık kalitesinin artırılması
- Doğru tanı belirlemede bir çok faktörün beraber değerlendirilmesi
- Finansal değişkenlerin çoklu ilişki durumlarını değerlendirilerek sağlık harcamalarında ekonomik katkı sağlanması
- Hastanede yönetsel değişkenlerinin çok boyutlu analiz edilerek optimal değerlerle yol haritasının belirlenmesi

Veri madenciliğinin, sağlık sektöründe ve HBYS'lerde uzman kişiler tarafından kullanılması, sağlık kurumları kaynaklarını daha verimli, sağlık hizmetlerinin daha etkin sunulması ve bilgiye erişimde hızlı ve şeffaflık olması açısından önemlidir (4).

2.2.2. Veri Madenciliği Süreçleri

Veri analizi günümüzde karar destek sistemlerinde önemli rol oynar. Veriler üzerinde çeşitli işlemler yapılarak verilerin kullanılabilirliğinin sağlanması ve veri-bilgi dönüşümünü oluşturan aşamaların tümü veri madenciliği süreçlerini oluşturur. Bu süreçler (Şekil 2.4.) de gösterilmiştir (38).

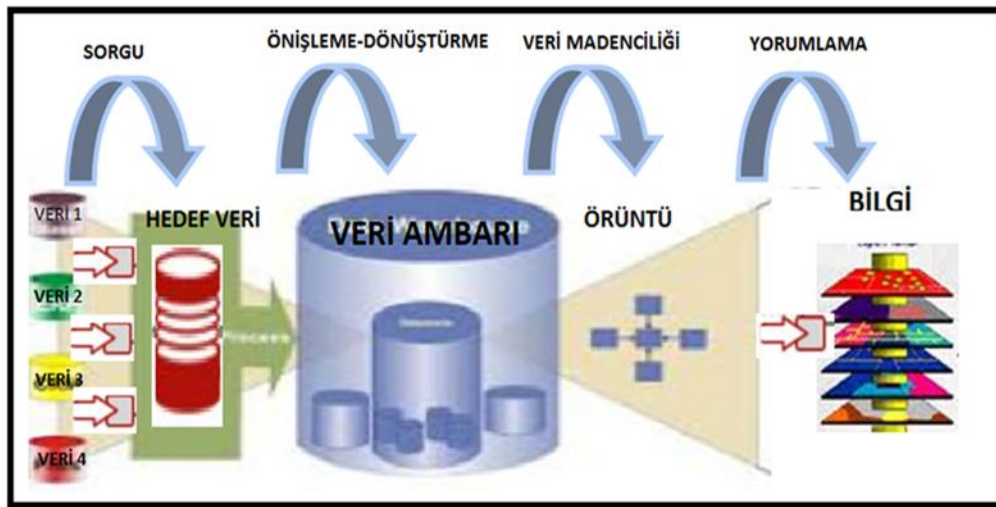
Verinin Seçilmesi : Veri madenciliği süreçlerinin uygulanacağı analizin amacına uygun olan veriler belirlenir.

Ön İşleme : Analiz sonuçlarının doğruluğu ve güvenilirliği verinin kalitesiyle bağlantılıdır. Sorunlu yanlış girilmiş verilerin belirlenmesi ve temizlenmesi, eksik girilmiş veri probleminin çözümü gibi işlemler bu aşamada tamamlanır.

Dönüştürme : Farklı biçimlerdeki verilerin uygulanacak veri madenciliği yöntemine uygun olarak verilerin dönüştürülmesi yapılır.

Veri Madenciliği: Hedeflenen amaçlara uygun veri madenciliği yönteminin belirlenip uygulanarak örüntülerin ortaya çıkarılması sağlanır.

Yorumlama : Veri madenciliği aşamasında elde edilen sonuçlar yorumlanarak bilgiye dönüştürülür.



Şekil 2.4 : Veri Madenciliği Süreçleri

2.2.3. Veri Madenciliği Yöntemleri

Veri madenciliği veriyi bilgiye dönüştürme sürecinde; veri tanımlanması, ilişkilendirme, korelasyon analizi, sınıflandırma, regresyon, kümeleme, dizi analizi ve trend analizi gibi çeşitli metodlar kullanır (37).

Veri madenciliği işlevselliğine göre farklı alanlarda kullanılan bir çok algoritmaları içerir. Algoritmalarda kullanılan temel yaklaşımlar aşağıdaki gibidir (39).

- **Tanımlama (Description)**

Araştırmacılar basitçe kalıpları tanımlamanın yollarını veri içinde yatan eğilimleri bulmaya çalışırlar. Veri madenciliği modelleri mümkün olduğunca verilerle ilgili net kalıpları yorumlanabilecek şekilde şeffaf olarak tanımlamalıdır.

- **Tahmin (Estimation)**

İstatistiksel analiz alanı çok yaygın kullanılan tahmin yöntemlerini içerir. Veri madenciliğinde sayısal hedef değişkeninin değerini bir dizi sayısal ve / veya kategorik tahmini değişken değerleriyle açıklayabilecek modelleri oluşturmak için farklı yöntemler kullanılabilir.

- **Öngörü (Prediction)**

Veri madenciliğinde geçmişteki bilgileri kullanarak gelecekte konuyla ilgili bilgi keşfini sağlayacak yöntemleri içerir.

- **Sınıflama (Classification)**

Verilerin bir kısmı eğitim seti olarak kullandıktan sonra, analize alınan değişkenler arasında ilişkileri tespit edip diğer veri setinin belirlenmiş sınıflara atamayı sağlayan metotlar kullanılır.

- **Kümeleme (Clustering)**

Kümeleme, sınıflandırmadan farklıdır. Kümeleme için hedef değişken yoktur. Tüm veri setini nispeten homojen alt gruplara veya kümelere bölmek için, kayıtların benzerliğini kullanır. Küme içi verilerin benzerliğinin yüksek, kümeler arası benzerliğin düşük olması ilkesine dayanır.

- **Birliktelik (Association)**

Veri setinde; Eş zamanlı gerçekleşen işlemler arasındaki ilişkiyi bir kurala bağlı olarak açıklayıp, önceden bilinmeyen bilgileri belirlemek amacıyla kullanır.

2.3. Kümeleme Yöntemleri

Kümeleme analizi, büyük miktardaki verileri birbirlerine uzaklıkları/benzerliklerine göre daha az boyuta indirerek kümelere ayırıp anlamlı bilgileri oluşturmak olarak özetlenebilir.

Kümeleme analizi, aynı gruptaki veriler arasında maksimum benzerlik, farklı gruplar arasındaki verilerde minimum benzerlik yaklaşımını kullanır (6).

2.3.1. Kümeleme Yönteminde Kullanılan Tanımlamalar

Kümeleme işleminde kullanılan bazı küme tanımları aşağıdaki (2.1) eşitliklere göre hesaplanır.

$$\text{Merkez : } X_0 = \frac{\sum_{i=1}^n X_{mi}}{N}$$

$$\text{Yarıçap: } R = \sqrt{\frac{\sum_{i=1}^N (X_{mi} - X_0)^2}{N}} \quad (2.1)$$

$$\text{Çap: } D = \sqrt{\frac{\sum_{i=1}^N \sum_{j=1}^N (X_{mi} - X_{mj})^2}{N(N-1)}}$$

X_0 : Küme ortalamasını gösterir.

R : Küme elemanlarının merkeze olan ortalama uzaklığını gösterir.

D : Küme içindeki iki nokta arasındaki ortalama uzaklıktır.

N : Toplam kayıt sayısı

n : Veri setindeki boyut sayısı

m : m. nokta

Kümeleme analizinde nesnelerin birbirleriyle benzerliklerini ölçmek için kullanılan uzaklık ve benzerlik ölçüleri tanımlanmıştır.

Uzaklık; iki nesne arasındaki konum farklılığının sayısal olarak belirlenmesidir.

Benzerlik; nesneler arasındaki yakınlık derecesinin belirlenmesidir.

Verilerin ölçüm biçimine göre farklılık gösteren bir çok uzaklık ve benzerlik ölçüleri tanımlanmıştır.

Sıkça Kullanılan Uzaklık Ölçüleri:

- **Öklit Uzaklık Ölçüsü**

Öklit uzaklık ölçüsü kümeleme algoritmalarında uzaklık ölçüsü olarak yaygınlıkla kullanılmaktadır. Bu çalışmanın kümeleme işleminde de kullanılan Öklit uzaklık ölçüsü (2.2) eşitliğine göre hesaplanır.

$$d(X_m, X_j) = \sqrt{\sum_{i=1}^n (X_{mi} - X_{ji})^2} \quad (2.2)$$

$d(X_m, X_j)$: X_m noktası ile X_j noktası arasındaki öklit uzaklığı

N : Toplam kayıt sayısı

n : Veri setindeki boyut sayısı

- **Manhattan City-Block Uzaklık Ölçüsü**

Bu uzaklık ölçütü, noktaların koordinatları arasındaki farkın mutlak değerlerinin toplamı alınarak (2.3) eşitliğinde belirtildiği gibi hesaplanır (38).

$$\text{Manhattan Uzaklığı}_{i,j} = \sum_{k=1}^p (|x_{ik} - x_{jk}|) \quad (2.3)$$

P : Noktaların boyut sayısı

i, j : i . Nokta ile j . nokta arasındaki uzaklık

- **Minkowski Uzaklık Ölçüsü**

Bu uzaklık ölçütü çok boyutlu hesaplamalar için genel bir formüldür, derece (boyut) (m) nin alacağı değerlere göre farklı hesaplamalar (2.4) eşitliğine göre yapılır (38).

$$\text{Minkowski Uzaklığı}_{i,j} = \sqrt[m]{\sum_{k=1}^p (|x_{ik} - x_{jk}|)^m} \quad (2.4)$$

P : Noktaların boyut sayısı

i, j : i . Noktası ile j . noktası arasındaki uzaklık

m : hesaplanması istenen derece (boyut)

Sıkça Kullanılan Benzerlik Ölçüleri:

- **Kosinüs Benzerlik Ölçüsü**

İki vektör arasındaki açı kullanarak vektörlerin benzerlikleri (2.5) eşitliğine göre belirlenir (38).

$$\text{Kosinüs Benzerliği} = \cos(t) = \frac{\sum_{i=1}^n A_i * B_i}{\sqrt{\sum_{i=1}^n A_i^2} * \sqrt{\sum_{i=1}^n B_i^2}} \quad (2.5)$$

A_i : A Vektörün i . boyutu

B_i : B Vektör i .boyutu

n : Boyut sayısı

t : A vektörü ile B vektörü arasındaki açı

- **Pearson Benzerlik Ölçüsü**

İki nesnenin öznitelikleri arasındaki ilişkinin gücünü ve yönünü belirlemek için yaygın olarak kullanılır (2.6) eşitliğine göre hesaplanır (38).

$$\text{Pearson Benzerliği}_{x,y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_X S_Y} \quad (2.6)$$

x,y : Veri dizileri

S_x, S_y : Veri dizilerin standart sapması

$\bar{X}, \bar{Y}$: Veri dizilerin aritmetik ortalaması

n : Kayıt sayısı

2.3.2. Kümeleme Analizinin Uygulama Aşamaları

Kümeleme analizi işlem aşamaları aşağıdaki gibi incelenebilir.

1. Analizin amacına uygun algoritmanın, değişkenlerin ve veri matrisinin belirlenmesi.
2. Birimlerin birbirleriyle olan benzerlik ya da uzaklıklarına uygun uzaklık ölçütü yöntemlerini kullanarak benzerlik/uzaklık matrisinin oluşturulması.
3. Verilere uygun kümeleme algoritması ile benzerlik/uzaklık matrisini kullanarak verilerin uygun sayıda kümelere ayrılması.

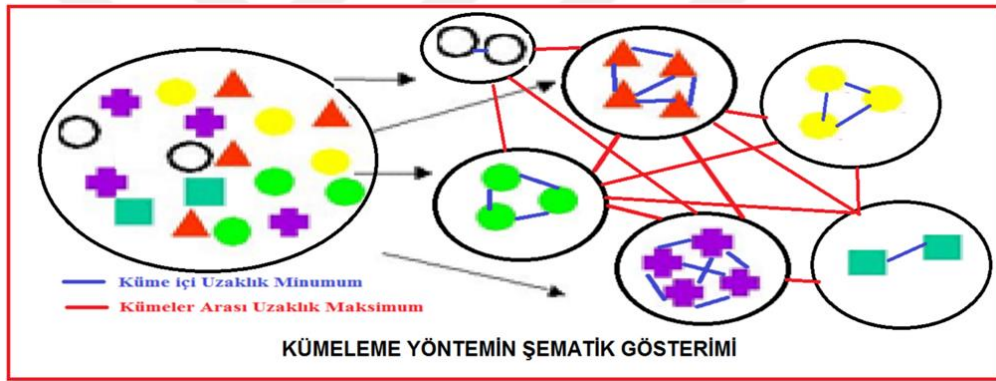
4. Oluşturulan kümeleri yorumlanarak, sonuçların tablo veya grafiklerle sunulması.

2.3.3. Kümeleme Yönteminin Matematiksel Tanımı

$A = \{X_1, X_2, \dots, X_m\}$, $n=1, 2, \dots, m$ veri seti, X_n bir kayıt olma üzere ;

$X = \{x_1, x_2, \dots, x_m\}$, $i=1, 2, \dots, m$, her bir x_i bir nitelik belirtsin.

Kümeleme, A veri setini, j adet K kümelerine bölmek, her biri için $K_j \subseteq A$ koşulunun sağlanmasıdır **(6)**. Şematik olarak (Şekil 2.5.) de gösterilmiştir.



Şekil 2.5: Kümeleme Yönteminin Şematik Gösterimi

2.3.4. Kümeleme Teknikleri

Kümeleme algoritmaları, kümelemenin oluşturuluş şekline, veri türüne, yapılacak araştırmanın amacına göre farklılık gösterirler (6).

Kümeleme algoritmaları;

- Hiyerarşik Kümeleme Yöntemleri
- Hiyerarşik Olmayan Kümeleme Yöntemleri

olmak üzere iki farklı grupta oluşturulmuştur.

Tıp alanındaki çalışmalarda, kullanım kolaylığı ve büyük miktardaki verilere uygunluğu nedeniyle hiyerarşik olmayan bölümlemeli kümeleme yöntemi K-ortalama algoritması yaygın olarak kullanılmaktadır (6).

K-ortalama algoritmasını diğer kümeleme algoritmaları ile karşılaştıran çalışmalarda K-ortalama algoritmasının ;

- Diğer yöntemlere göre daha kolay olduğu (41),
- Özellikle büyük boyutlu ve miktardaki heterojen ölçümlü verilerde daha iyi sonuçlar verdiği (41),
- Hiyerarşik yöntemlere kıyasla gürültülü verilere daha az duyarlı, büyük boyutlu veri kümelerini daha basit bir görüntüde oluşturduğu için yorumlaması daha anlaşılır olduğu (41),
- Grid temelli algoritmalara göre daha düşük bilgisayar belleği gerektirdiği (41),
- Yoğunluğa dayalı kümeleme algoritmalarına göre daha az donanımlı bilgisayarlarla uygulanabildiği (42),
- Geniş varyanslı verilerde, yoğunluğa dayalı kümeleme algoritmalarına göre daha iyi sonuçlar verdiği (42)

belirtilmektedir.

Veri Madenciliği Kümeleme analizinde kullanılan K-Ortalama algoritması, çalışmamızın amacına uygun bir yöntem olarak tercih edilmiştir.

2.3.5. Niteliksel Verilerde Küme Geçerliliğinin Ölçülmesi

Niteliksel verilerde güvenilirlik, sınıflamadaki tutarlılık (uyum) olarak tanımlanır. Kappa Katsayısı şansa bağlı olarak ortaya çıkan “şans uyuşmalarını” düzelterek güvenilirlik hakkında bilgi veren bir katsayıdır (43).

Çizelge 2.2 : Kappa İstatistiğinin Hesaplanması ve Çapraz Tablo (2x2)

Yöntem1	Yöntem2		Toplam
	1	2	
1	A	B	A+B
2	X	Y	X+Y
Toplam	A+X	B+Y	A+B+X+Y

Çizelge 2.2. deki A,B,X ve Y sıklıklarından yararlanarak iki yöntemin uyumlu olanların oranı (pa) ve şansa bağlı uyum oranı (pc) olmak üzere;

$$pa = (A+Y) / (A+B+X+Y)$$

$$pc = [(A+B) (A+X) + (X+Y) (B+Y)] / (A+B+X+Y)^2$$

$$k = (pa - pc) / (1-pc) \quad (k : Kappa Katsayısı)$$

şeklinde hesaplanır. Elde edilen kappa değeri aşağıdaki gibi yorumlanır.

Kappa Katsayısı	Uyum Düzeyi
0.81 – 1.00	Çok İyi
0.61 – 0.80	İyi
0.41 – 0.60	Orta
0.21 – 0.40	Düşük
0.00 – 0.20	Zayıf
< 0.00	Çok Zayıf

2.4. K-ortalama Yöntemi ve Farklı Yaklaşımlar

2.4.1.K-Ortalama Algoritması

K-ortalama yöntemi, en uygun küme sonucuna ulaşmaya kadar tekrarlanan ve sürekli olarak kümelerin yenilendiği döngüsel bir algoritmadır. K-ortalama algoritması gözetimsiz öğrenme yöntemi kapsamındaki algoritmalarının temelini oluşturan yöntem olup yaygın olarak kullanılmaktadır (6).

K-Ortalama algoritmasının adımları;

1. K tane küme merkezi rastgele seçilir.
2. Her noktanın seçilen merkez noktalara olan uzaklığı/benzerliği hesaplanır, veri hesaplanan uzaklıklara göre seçilen k tane merkezden kendine en yakın olan kümeye atanır.
3. Oluşan kümelerin ortalama değeri yeni küme merkezi olarak belirlenir.
4. Merkez noktalar değişmeyinceye kadar 2. ve 3. adımlar tekrarlanır (44).

K-Ortalama algoritmasında **küme sayısı (k)** ve **başlangıç küme merkezleri** kümeleme sonuçlarını ve performansını etkilediği için, bu değerlerin seçiminde dikkatli davranılması gerekmektedir.

K-Ortalama kümeleme yönteminin değerlendirilmesinde genel olarak Küme İçi Hata Kareler Toplamı (**Toplam Karesel Hata - Summed Squared Error**) (SSE) değeri kullanılır.

SSE değerinin ve küme sayısının küçük olduğu kümelemeler en uygun kümeleme sonucu olarak kabul edilir, SSE (2.2) eşitliğine göre bulunur (37).

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} dist^2(m_i, x) \quad (2.2)$$

x : C_i kümesinde bulunan bir nesne

m_i : C_i kümesinin merkez noktası

K-Ortalama kümeleme yönteminde küme sayısının (k) belirlenmesinde (2.3) eşitliği kullanılmaktadır. Ancak veri setindeki kayıt sayısının büyük olduğu durumda çok tercih edilen bir yöntem değildir (27).

$$k = \sqrt{\frac{N}{2}} \quad N: \text{Toplam kayıt sayısı} \quad (2.3)$$

K-ortalama yönteminin dezavantajları

K- ortalama algoritmasında;

- küme sayısı (k)
- başlangıç merkezleri

parametre olarak kullanıcı tarafından belirlenir.

Bu parametreler kümeleme performansını etkilediği için uygun değerlerin belirlenmesi kümeleme sonuçlarını belirlemede önemli bir faktördür.

Uygun parametre değerlerini belirlemek için algoritma farklı küme sayısı ve rasgele seçilen başlangıç merkezleri ile defalarca çalıştırılarak, uygun küme sonuçları elde edilebilir.

Uygun kümeleme sonuçlarına ;

- Parametre değerinin uzman kişiler tarafından belirlenmesi
- Algoritmanın tekrarlı olarak çalıştırılması (analiz süresini uzatmaktadır).
- Seçilen parametre değerlerine göre farklı kümeleme sonuçlarından geçerli kümeleme sonuçlarının bulunması

ile ulaşmak mümkündür.

Ancak özellikle büyük veri (Big Data) niteliğindeki verilerin analizinde ve kümeleme sonuçlarının tutarlılığında problem oluşturabilir.

Algoritmanın kullanıldığı çalışmalarda parametre değerlerinin kullanıcı tarafından belirlenmesi K-ortalama algoritmasının dezavantajıdır (6).

Veri Madeniliği Weka programında K-ortalama algoritmasında başlangıç merkezlerini belirlemek için K-ortalama++, K-ortalama canopy algoritmaları seçenek olarak sunulmuştur. Weka programındaki K-ortalama algoritmasının kullanılması tıp alanındaki bilimsel çalışmalarda kolaylık sağlaması nedeniyle tercih edilebilir.

2.4.2. Canopy Algoritması

Canopy kümeleme algoritması K-ortalama algoritmasının dezavantajlarına çözüm olabilecek ve destek olarak kullanılan etkin algoritmalarından biridir. Algoritma 2000 yılı ACM SIGKDD konferansında McCallum, Nigam ve Ungar tarafından tanıtılmıştır (38). Algoritma iki aşamalı olarak tasarlanmıştır.

Canopy Algoritma Aşamaları;

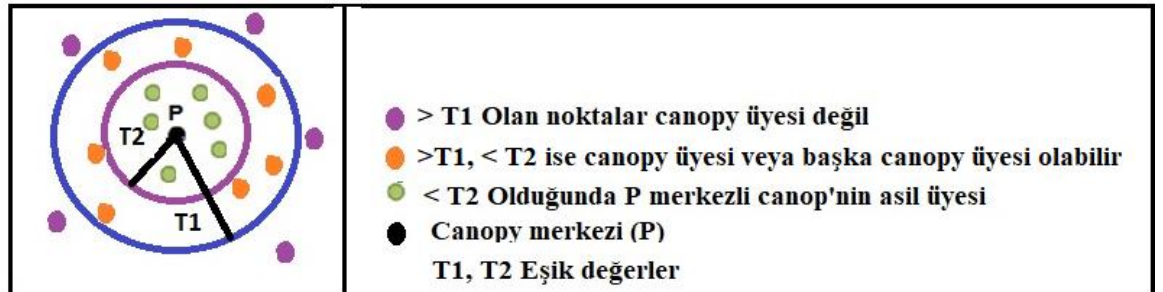
1. Veri dizileri basit uzaklık metrikleri kullanarak kaba olarak kümeler oluşturulur.
2. K-ortalama gibi bir algoritma kullanarak oluşturulan kaba kümelerin iyileştirilmesi sağlanır.

Canopy algoritması kullanıcı tarafından belirlenen iki uzaklık eşik değeri $T1$ ve $T2$ ($T1 > T2$) parametresine bağlıdır. Bu parametrenin çok büyük veya çok küçük seçilmesi durumunda taslak kümelerin oluşturulması aşaması başarısız olur, dolayısıyla parametre için uygun değerler seçilmesi gerekir (45).

Aşama 1 :

1. Veri dizisinde bir nokta merkez olmak üzere tesadüfi olarak seçilir.
2. Veri dizisinde bulunan tüm noktaların kaba uzaklık metriği kullanarak merkez seçilen noktaya uzaklıkları bulunur.
3. Merkezle noktalar arasında bulunan uzaklıklar $T2$ noktasından küçük ise bu noktalar ilgili canopy' nin asil üyesi olur ve veri dizisinden silinir, bu noktalar için süreç biter.
4. Noktalar $T2$ değerinden büyük, ancak $T1$ değerinden küçük ise canopy' nin bir üyesi olarak kabul edilir. Bu noktaların hala kendi canopy'lerini kurma olasılıkları devam ettiğinden, noktalar veri dizisinden silinmez.
5. Süreç veri dizisi boşalıncaya kadar devam eder.

Algoritmanın şematik gösterimi (Şekil 2.6.) da gösterilmiştir.



Şekil 2.6. : Canopy Algoritmasının Şematik Gösterimi

2.4.3. K-Ortalama ++ Algoritması

K-ortalama++ algoritması K-ortalama algoritmasının önemli sorunu olan başlangıç noktalarının belirlenmesine çözüm olabileceği düşünülerek 2007 yılında David Arthur ve Sergei Vassilvitskii tarafından geliştirilmiş bir algoritmadır. Algoritma K-ortalama tabanlıdır, ilk aşamada başlangıç merkezlerinin seçimi yapılır, ikinci aşamada standart K-ortalama algoritması uygulanır (44).

K-Ortalama++ Algoritma Aşamaları ;

1. Kümelenecek X veri setinden tesadüfî olarak $i=1$ ise C_i başlangıç merkezi seçilir.
2. Her x veri noktası için seçilmiş C_i küme merkezine olan uzaklıkları ($d_{C_i x}$) bulunur. $i=1$ ise geçerli uzaklık $dx = d_{C_i x}$, $i > 1$ ise $dx = \min(d_{C_{i-1} x}, d_{C_i x})$ olarak alınır ve (dx) ve uzaklıkların toplamı ($\sum dx$) bulunur.
3. Her nokta için ($dx / \sum dx$) oranları ve bu oranların kümülatif toplamaları hesaplanır.
4. $C_i = x \in X$ C_i . ($i=2,3 \dots k$) $i=i+1$ i. merkez olmak üzere oranların kümülatif toplamaları sütunundan rasgele bir nokta C_i i. merkez olarak seçilir.
5. $i \leq k$ ise **2. adıma** gidilir.
6. k tane başlangıç küme merkezi seçildikten sonra K-ortalama algoritması uygulanır.

3. GEREÇ VE YÖNTEMLER

Tez çalışmasında geliştirilen NAMGY algoritmasının uygululamaları aşağıda tanımlanan farklı veri setleri üzerinde yapılmıştır. Veri setleri Çukurova Üniversitesi Balcalı Hastanesi HBYS' inden oluşturulmuştur. Verileri [0 1] aralığında standartlaştırma ve temizleme ön işlemleri yapılarak analiz edilmiştir. Verilerin kullanılabilmesi için etik kurul onayları alınmıştır. Gerekli belgeler EK-1'de verilmiştir.

E-reçete veri seti : 2014 yılında polikliniklere başvuran ve e-reçete yazılan 365,065 hastanın cinsiyet, yaş, muayeneye geldiği bölüm, ICD-10 tanı kodları, ilaçların ATC kodları içeren veri seti.

Diyabet veri seti : Açık kaynak kodlu Weka programında tanımlanmış olan Hamile kalma sayısı, Plazma glikoz konsantrasyonu, diastolik tansiyon, triceps deri kat kalınlığı, 2 saatlik serum insülini, vücut kitle indeksi, ailede diyabet öyküsü, yaş gibi bilgileri içeren veri seti.

Cerrahi müdahale veri seti : Çukurova Üniversitesi Balcalı Hastanesi HBYS'inden 2010 ve 2013 yılları arasında ameliyat olan 59869 hastanın yatış sürelerini etkileyen yaş, cinsiyet, yapılan cerrahi müdahale hakkındaki bilgiler, girişimsel işlem sayısı, hastalık tanısı, antibiyotik kullanımı bilgileri içeren veri seti.

VitaminB12 veri seti : Çukurova Üniversitesi Balcalı Hastanesi HBYS'inden 2016 yılında çocuk sağlığı anabilim dalına gelen 4457 hastanın vitaminB12 değerlerini etkileyen faktörleri belirlemek amacıyla cinsiyet, yaş, bilimdalı, vitamin B12 değeri ni içeren veri seti.

İris veri seti : Literatürde geniş bir şekilde kullanılan WEKA veri datası içindeki çanak uzunluğu, çanak genişliği, yaprak boyu, yaprak genişliği, çiçek grubu gibi üç çeşit çiçek bilgilerini içeren veri seti.

3.1. NAMGY Algoritmasının Ana Hatları

Çalışmanın amacına uygun olarak;

K-ortalama algoritmasında kümeleme sonuçlarında etkili olan küme sayısı (k) ve başlangıç merkezleri parametre değerleri olarak kullanıcı tarafından belirlenmektedir. Veri setindeki noktaların özelliklerine (noktalar arası uzaklık ve yayılımlar) göre optimal kümeleme sonuçları verecek uygun parametre değerleri belirleyen NAMGY (Noktalar Arası Mesafe ve Gözlemlerin Yoğunluğu) isimli algoritma geliştirilmiştir. Bu algoritmayla mevcut K-ortalama algoritmasının problemlerine çözüm bulmaya çalışılmıştır.

Bu amaçla belirlenen NAMGY algoritmasının genel işleyişi aşağıdaki gibidir.

- 1. Min-Mak Küme Sayısının Belirlenmesi :** En küçük küme sayısı $k=2$, her bir kümenin kayıt sayıları %5 den küçük olmayacağı düşünülerek en büyük küme sayısı (K) belirlenir.
- 2. Başlangıç Merkezlerinin Belirlenmesi :** Birinci adımdaki her k için merkez belirleme algoritması uygulanarak başlangıç merkezleri oluşturulur.
- 3. Standart K-Ortalama Algoritmasının Uygulanması:** Belirlenen en büyük küme sayısı kadar her küme sayısı (k) ve başlangıç merkezleri kullanılarak standart K-ortalama algoritması döngüsel olarak uygulanır.
- 4. Kümelemeyi Bitirme Kriterlerinin Saptanması :** Kümeleme işleminin sonlandırılması belirlenen kriterlere göre yapılır ve uygun küme sonucu elde edilir.

3.2. NAMGY Algoritması

Tıp alanındaki çalışmalarda;

- Kullanım kolaylığı
- Büyük miktardaki verilere uygunluğu

nedeniyle hiyerarşik olmayan bölümlmeli kümeleme yöntemi olan K-ortalama çok kullanılan algoritmadır.

K-ortalama algoritmasının başlangıç merkezlerinin ve küme sayısının (k) parametre olarak verilmesi kullanım dezavantajıdır. Ancak; parametreleri tanımlamak amaçlı bir çok algoritma geliştirilmesine rağmen yeterli bir çözüm bulunamamıştır. K-ortalama algoritmasında başlangıç merkezleri kümeleme sonuçlarını belirlemede etkili bir parametredir (46). Bu merkezlerin rasgele (random) olarak seçilmesi algoritma için çözülmesi gereken önemli problemdir.

Problemin çözümünde etkili olacağı düşünülerek, tez çalışmasında geliştirilen NAMGY algoritmasının bölümleri;

- K Tane Başlangıç Merkezlerin Belirlenmesi İçin Geliştirilen Yöntem
- En Uygun k Küme Sayısının Belirlenmesi İçin Geliştirilen Yöntem

başlıkları altında tanımlanabilir.

3.3. NAMGY K-Ortalama Algoritmasında Başlangıç Merkezlerinin Belirlenmesi

Küme sayısı belirlendiğinde başlangıç merkezlerini belirleyen algoritmanın adımları;

1. Veri kümesindeki her noktanın diğer noktalara olan uzaklıkları bulunur.
2. Veri kümesindeki noktalar arasındaki uzaklıklardan Genel Ortalama Uzaklık (GOU) (3.1) eşitliğindeki gibi hesaplanır.

$$\text{GOU} = \left[\sum_{i=1}^N \sum_{j=1}^N d(X_i, X_j) \right] / ((N * N) - N) \quad i \neq j \quad (3.1)$$

GOU: Genel Ortalama Uzaklık

d(X_i, X_j) : X_i noktası ile X_j noktası arasındaki uzaklık

N : Toplam kayıt sayısı

3. Her noktanın diğer noktalar arasındaki uzaklığı GOU dan küçük yada eşit ise nokta komşu kabul edilir, her noktanın komşuları (3.2) eşitliğin sağlayan noktalar olarak belirlenir, toplam komşu sayıları bulunur.

$$d(X_i, X_j) \leq \text{GOU} \quad \text{ise } X_i \text{ ile } X_j \text{ noktaları komşu} \quad (3.2)$$

4. Her nokta için kendi komşu noktalarına olan uzaklıkların ortalaması (3.3) eşitliğine göre hesaplanır.

$$X_i KUO = [\sum_{j=1}^{KS_i} d(X_i, X_j)] / KS_i \quad i = 1, 2, \dots, N \quad (3.3)$$

$X_i KUO$: X_i 'nin komşu noktalarına uzaklıklar ortalaması

$d(X_i, X_j)$: X_i noktası ile kendisine komşu olan X_j noktası arasındaki uzaklık

KS_i : X_i 'nin toplam komşu nokta sayısı

N : Toplam kayıt sayısı

5. M_1 ; Birinci merkez, komşu nokta sayısı büyük, komşu noktalarına uzaklıklar ortalaması küçük olan noktalardan seçilir.

Bu noktalar bulunurken komşu nokta sayısı (KS_i), komşu noktalarına uzaklıklar ortalaması ($X_i KUO$) ise $KS_i / X_i KUO$ oranı ve komşu sayısı en büyük değerler kullanılır.

6. Diğer merkezler; birinci merkeze olan uzaklıklarına ve komşu sayılarına göre (3.4) eşitlikliğindeki belirtilen koşulları sağlayan noktalardan;

$$(j-1)*Biruz \leq d(M_1, M_j) < j*Biruz$$

$$Biruz = \max_{uz} / (k) \quad (3.4)$$

$\max_{uz}$: M_1 noktası ile kendisine en uzak nokta arasındaki uzaklık

$Biruz$: Diğer merkezleri belirlemede kullanılan birim uzunluk

$d(M_1, M_j)$: M_1 merkez ile M_j . merkez arasındaki uzaklık

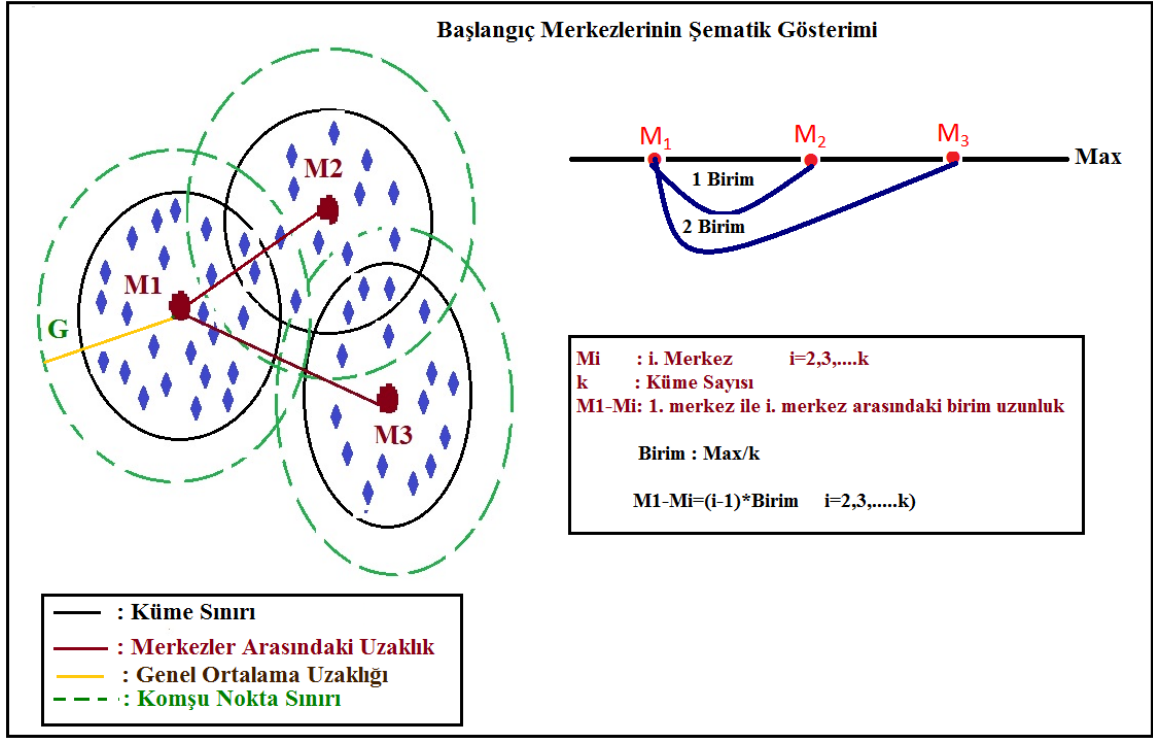
k : Küme Sayısı (verilen yada hesaplanmış olan)

j : Kümelemedeki merkez sırası $j=2, 3, \dots, k$

M_j :j. merkez

komşu nokta sayısı (KS_i), komşu noktalarına uzaklıklar ortalaması ($X_i KUO$) ise $KS_i / X_i KUO$ oranı ve komşu sayısı en büyük değeri sağlayan nokta M_j (j. Merkez) olarak seçilir.

Başlangıç merkezlerini belirleme yöntemi şematik olarak Şekil 3.1. de gösterilmiştir.



Şekil 3.1 : Başlangıç Merkezlerinin Şematik Gösterimi

3.4. NAMGY K-Ortalama Algoritmasında Küme Sayılarının Belirlenmesi

Algoritmanın adımları:

1. Kümeleme işleminde kullanmak üzere en küçük küme sayısı ve en büyük küme sayısı (3.5) eşitliğindeki gibi alınır.

En küçük küme sayısı =2

En büyük küme sayısı = 20 (3.5)

2. K-ortalama algoritması belirlenen başlangıç merkezleri ile yalnızca küme sayıları değiştirilerek çalıştırılır.
3. Küme sayısı k olmak üzere, farklı k değerleri için algoritma çalıştırılarak her k için oluşan grupların kayıt sayıları ve toplam karesel hata (SSE) (2.2) eşitliğine göre hesaplanır.
4. Her bir k için belirlenen gruplar kayıt sayılarına göre büyükten küçüğe doğru sıralanır ve yığılımlı yüzde oranları ($Y_{ni} \%$) (3.6) eşitliğine göre bulunur. Bu

oranlara göre %25, % 50, %75, %90 yüzdeliklerine kaçınıcı kümelerde ulaşılabiliyor bilgisi elde edilir.

$$Yn_i \% = (n_1 + n_2 + n_3 + \dots + n_i) / N * 100 \quad (3.6)$$

Yn_i : Yığılımlı Yüzde Oran

k : Küme sayısı

n_i : i . kümenin (grup) kayıt sayısı ($i=1,2,\dots,k$)

N : Toplam kayıt sayısı

5. Her k için SSE değerlerinin farklılıklarını gösteren $SSE_i FARK\%$ değerleri (3.7) eşitliğine göre hesaplanır.

$$SSE_i FARK \% = (SSE_i - SSE_{i-1}) * 100 / SSE_i \quad (3.7)$$

SSE_i : i . küme sayısına göre SSE

$SSE_i FARK$: Ardışık küme sayıları arasındaki SSE farkı

6. Daha küçük küme sayısı ile küçük SSE ilkesine göre aşağıda belirtilen koşulları sağlayan k uygun küme sayısı olarak kabul edilebilir.

1. $SSE_i FARK\%$ değerinin 10 dan büyük yada eşit olmalıdır.
 2. Kayıtların % 90 nını içeren grupların sayısı toplamı ile küme sayısı arasındaki farkın 2 den küçük eşit olan kayıtlar belirlenir.
- Yukardaki 1. ve 2. koşulu sağlayan küme sayılarından, $Max(SSE_k FARK\%)$ 'ın küme sayısına eşit yada küçük olan en büyük küme sayısı (k) uygun küme sayısı olarak alınabilir.
 - 1.ve 2. koşul sağlanmıyorsa $Max(SSE_k FARK\%)$ ve $min(k - \%90_{ks})$ sağlayan k lardan $uygunK = min(k)$ olarak seçilir.

Kümeler ile ilgili bilgiler Çizelge 3.2' de gösterilmiştir.

Çizelge 3.2 : Küme Sayılarına Göre Küme İçindeki Grupların Kayıt Sayısı ve Toplam Karesel Hata

Küme Sayısı k	Grupların Kayıt Sayısı n	Toplam Karesel Hata SSE	SSEFARK %	Kayıtarın 25%'ini içeren grup sayısı	Kayıtarın 50%'ini içeren grup sayısı	Kayıtarın 75%'ini içeren grup sayısı	Kayıtarın 90%'ini içeren grup sayısı
k_1	$n_{k11}, n_{k12}, \dots, n_{k1k}$	SSE_{k1}		$25\%_{k1}$	$50\%_{k1}$	$75\%_{k1}$	$90\%_{k1}$
k_2	$n_{k21}, n_{k22}, \dots, n_{k2k}$	SSE_{k2}	$SSE_{k2}FARK \%$	$25\%_{k2}$	$50\%_{k2}$	$75\%_{k2}$	$90\%_{k2}$
k_3	$n_{k31}, n_{k32}, \dots, n_{k3k}$	SSE_{k3}	$SSE_{k3}FARK \%$	$25\%_{k3}$	$50\%_{k3}$	$75\%_{k3}$	$90\%_{k3}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
k_k	$n_{kk1}, n_{kk2}, \dots, n_{kkk}$	SSE_k	$SSE_{kk}FARK \%$	$25\%_{kk}$	$50\%_{kk}$	$75\%_{kk}$	$90\%_{kk}$

3.5. NAMGY Algoritması ve Akış Şeması

Bu aşamada geliştirilen iki yöntem birleştirilirken önce olası bütün k (küme sayısı) lar için başlangıç merkezleri bulunup, sonrada uygun küme sayısı saptaması yapılmıştır.

Algoritmanın 1-13 arasındaki adımlar küme merkezlerini oluşturur, 14-20 arasındaki adımlarda uygun küme sayıları belirlenerek kümeleme işlemleri tamamlanır.

1. Veri seti okutulur.
2. Veri kümesindeki her noktanın diğer noktalarına olan uzaklıklar bulunur.
3. Veri kümesindeki noktalar arasındaki Genel Ortalama Uzaklık (GOU) (3.1) eşitliğine göre hesaplanır.
4. Her noktanın, GOU dan küçük eşit uzaklıktaki noktaların (komşu nokta) sayıları (KS) bulunur.
5. Her noktanın kendi komşu noktalara olan uzaklıkların ortalaması hesaplanır.
6. En küçük küme sayısı $k=2$,
En büyük küme sayısı $K= 20$
olmak üzere küme sayısı k olarak alınır.
7. Birinci merkezi (M_1); komşu nokta sayısı büyük, komşu noktalarına uzaklıklar ortalaması küçük olan noktalardan seçilir.

Bu noktaları bulunurken komşu nokta sayısı (KS), komşu noktalarına uzaklıklar ortalaması (XKUO) ise $KS / XKUO$ oranı ve komşu sayısı en büyük değerler kullanılır.

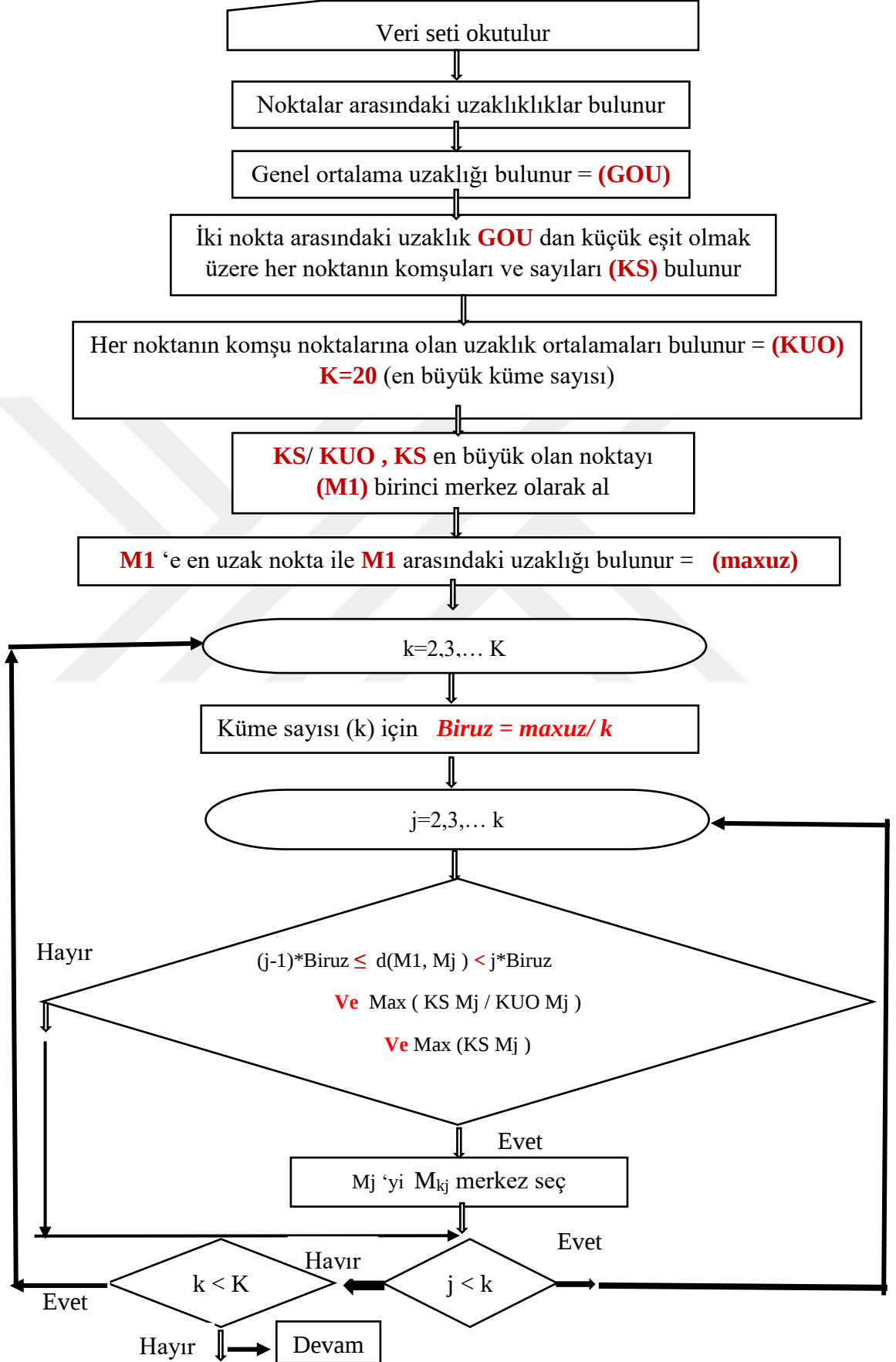
8. Birinci merkez (M_1) ile kendine en uzak nokta arasındaki uzaklık (**maxuz**) bulunur.
9. ilk değer $k=2, j=2, d(M_1, M_{kj}) : M_1$ ile M_{kj} noktalar arasındaki uzaklık hesaplanır.
10. Küme sayısı (k) için **Biruz** = **maxuz** / **k**
11. Diğer merkez seçimi için;

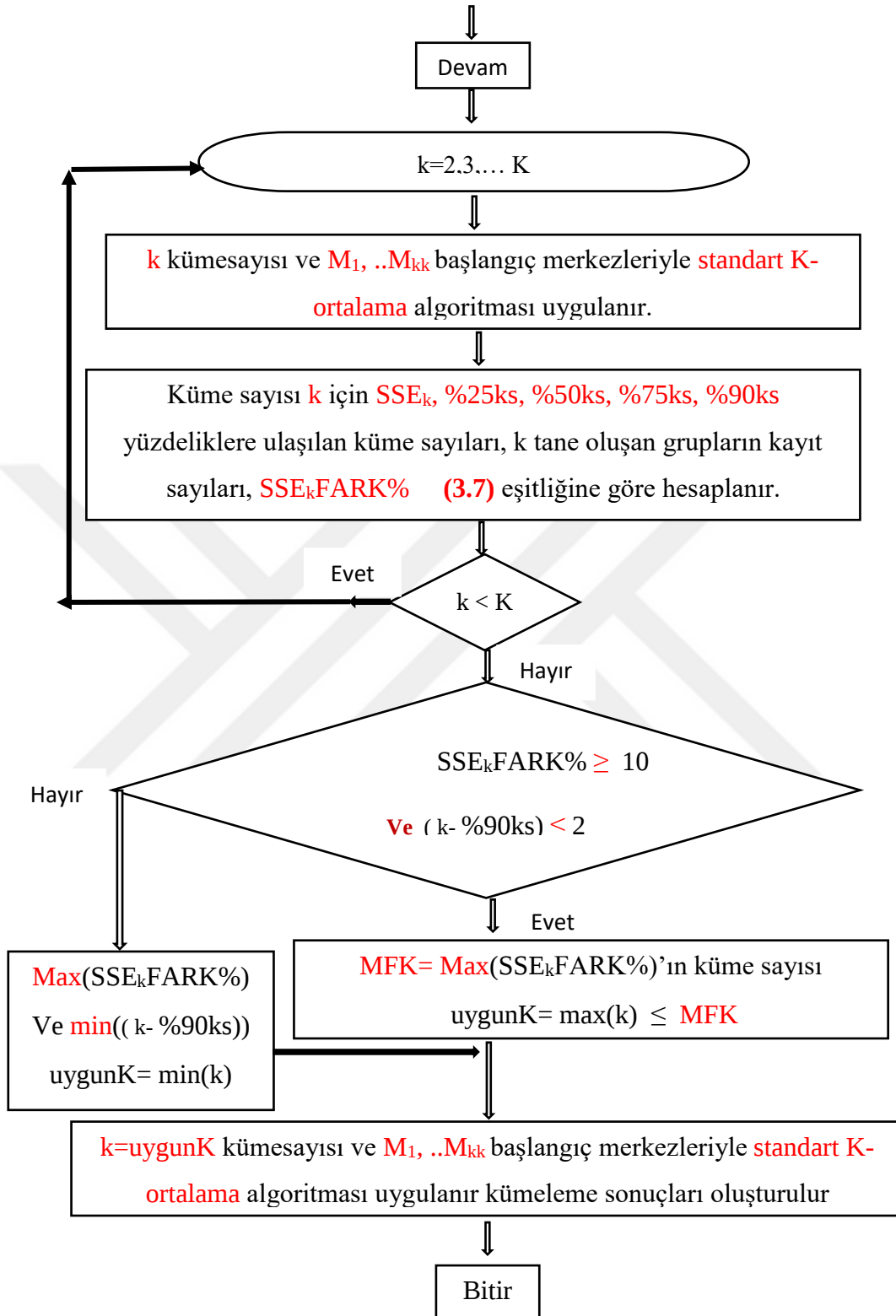
$$(j-1)*\text{Biruz} \leq d(M_1, M_j) < j*\text{Biruz} \quad (3.8)$$
eşitliğindeki koşulunu sağlayan noktalardan;
komşu nokta sayısı (KS_i), komşu noktalarına uzaklıklar ortalaması (X_iKUO) ise
 KS_i / X_iKUO oranı ve komşu sayısı en büyük değeri sağlayan nokta M_j 'yi
 kj . merkez M_{kj} olarak seçilir.
12. Merkez seçimi $j < k$ ise $j=j+1$, **11. Adıma** git.
13. $k < K$ ise $k=k+1, j=2$ al **10. Adıma** git.
($k=2,.. K$, olmak üzere her bir k için $M_1, M_{k2},, M_{kk}$ tane belirlenen başlangıç merkezleri seçilmiş olur.)
14. İlk önce küme sayısı $k=2$ ve $J=2$ için seçilmiş olan başlangıç merkezleri M_1, M_{22} alınır.
15. k küme sayısı ve $M_1, ..M_{kk}$ başlangıç merkezleriyle standart K-ortalama algoritması uygulanır.
16. Küme sayısı (k) için $SSE_k, \%25_{ks}, \%50_{ks}, \%75_{ks}, \%90_{ks}$ yüzdeliklere ulaşılan küme sayıları, k tane grupların kayıt sayıları, $SSE_k \text{FARK}\%$ **(3.7)** eşitliğine göre hesaplanır.
17. Küme sayısı $k < K$ ise $k=k+1$ al **15. Adıma** git .
18. $SSE_k \text{FARK}\% \geq 10, (k - \%90_{ks}) \leq 2$ koşulları sağlayan kayıtlardan,
 $\text{maxfarkksay} = \text{en büyük } SSE_k \text{FARK}\%$ 'ın küme sayısı ise
 $\text{uygunK} = \max(k) \leq \text{maxfarkksay}$ al, **20. Adıma** git
19. $\text{Max}(SSE_k \text{FARK}\%)$ ve $\min(k - \%90_{ks})$ sağlayan $\text{uygunK} = \min(k)$

20. uygunK=k küme sayısı ve $M_1, ..M_{kk}$ başlangıç merkezleri ile standart K-ortalama algoritması uygulanıp, kümeleme sonuçları elde edilir.



NAMGY ALGORİTMANIN AKIŞ ŞEMASI





Şekil 3.2 : NAMGY Algoritmasının Akış Şeması

3.6. NAMGY algoritmasının K-ortalama Kümeleme Yöntemi İle Birleştirilmesi İçin Arayüz Programının Düzenlenmesi

NAMGY algoritmasının değerlendirilmesinde, veri madenciliğinde yaygın kullanılan açık kaynak kodlu, Waikato üniversitesi tarafından geliştirilen WEKA (47) programı referans olarak kullanılmıştır.

K-Ortalama kümeleme yöntemine küme sayılarının ve başlangıç merkezlerinin belirlenmesi amacıyla NAMGY algoritması eklenerek yeni ara yüz programı tasarlanmıştır. Ara yüz programların yazılımı HTML, CSS, JAVASCRIPT programlama dilleri kullanılarak tasarlanmıştır.

NAMGY algoritmasının değerlendirilmesi, oluşturulan arayüz programı ile değişik veri setleri üzerinde çalıştırılıp uygulamalar yapılmıştır.

4. BULGULAR

Tezin amacında belirtilen çalışmaların tamamlanması sürecinde yapılan uygulamalar sonucunda elde edilen bulgular alt başlıklarla incelenebilir.

Bu bağlamda; K-ortalama algoritmasına eklemeler yapılarak NAMGY algoritmasının oluşturulma ve uygulamalarında HBYS verileri kullanarak algoritmanın geçerliliğini gösteren çalışmalar yapılmıştır.

Bu amaçla yapılan örnek çalışmalar aşağıdaki gibidir.

4.1. K-Ortalama Algoritmasının HBYS Verilerine Uygulanması

İlk aşamada mevcut olan K-Ortalama algoritmalarının HBYS verilerine uygulanabilirliği değerlendirilmiştir

Bu amaçla; mevcut K-Ortalama, K-Ortalama++ ve K-Ortalama Canopy algoritmaları, HBYS'den alınan veriler üzerinde çalıştırılarak sonuçlar elde edilmiştir.

Çalışma, açık kaynak kodlu veri madenciliği programı WEKA kullanılarak yapılmıştır.

2014 yılında polikliniklere başvuran ve e-reçete yazılan 365,065 hastanın aşağıdaki bilgileri veri seti olarak kullanılmıştır.

- Cinsiyet,
- Yaş,
- Muayeneye geldiği bölüm
- ICD-10 tanı kodları (International Classification of Diseases)
- İlaçların ATC (Anatomic Therapeutic Chemical Classification-Yapısal Tedavi Edici Kimyasallar Sınıflaması)

Küme merkezlerinin belirleme şekilleri aşağıdaki gibidir :

K-ortalama (Random) : Rasgele seçilmiştir.

K-ortalama (Plus) : K-ortalama++ algoritmasına göre seçilmiştir.

K-ortalama (Canopy) : Canopy algoritmasına göre seçilmiştir.

Kümelerin Niceliksel Sonuçları

Çizelge 4.1. : K-ortalama , K-ortalama++, Canopy Algoritmalarına Göre Kümeleme Sonuçları

Algoritmaların Kayıt Sayılarına Göre Kümelere Dağılımı								
K-Ortalama (Random)			K-Ortalama (Plus)			K-Ortalama (Canopy)		
Küme Adı	Kayıt Sayısı	ORAN (%)	Küme Adı	Kayıt Sayısı	ORAN (%)	Küme Adı	Kayıt Sayısı	ORAN (%)
Küme0	73367	20.10	Küme0	86066	23.58	Küme0	74505	20.41
Küme5	52698	14.44	Küme7	49998	13.70	Küme1	62251	17.05
Küme2	48925	13.40	Küme1	47854	13.11	Küme2	51513	14.11
Küme1	46817	12.82	Küme6	38969	10.67	Küme7	38417	10.52
Küme3	36259	9.93	Küme2	35710	9.78	Küme4	37165	10.18
Küme7	31679	8.68	Küme5	27289	7.48	Küme5	28190	7.72
Küme6	27281	7.47	Küme3	24685	6.76	Küme6	21440	5.87
Küme4	24403	6.68	Küme8	20384	5.58	Küme9	19943	5.46
Küme9	12881	3.53	Küme4	18850	5.16	Küme8	16223	4.44
Küme8	10755	2.95	Küme9	15260	4.18	Küme3	15418	4.22
Toplam	365065	100.00	Toplam	365065	100.00	Toplam	365065	100.00
İterasyon Sayısı	9		İterasyon Sayısı	4		İterasyon Sayısı	4	
SSE	1020633		SSE	1024436		SSE	1031882	

Algoritmaların İşleyiş Özelliklerine Göre Değerlendirilmesi:

Üç algoritma işleyiş özellikleri yönünden bazı durumlarda farklılaşmaktadır.

- K-ortalama (plus) ve K-ortalama(Canopy) algoritmalarının son iki küme kayıt sayıları oranı K-ortalama(Random) algoritmasına göre daha büyüktür.
- K-ortalama(plus) ve K-ortalama(Canopy) algoritmaları tekrar (iterasyon) sayılarının K-ortalama(Random) algoritmasına göre daha küçük olması özellikle büyük veriler için önemli bir zaman kazancı sağlamaktadır.

- İterasyon sayısı küçük olan iki algoritmada K-ortalama(plus) algoritmasının SSE değeri K-ortalama(Canopy) algoritmasının değerinden daha küçüktür.

Kümeleme Sonuçlarının Genel Niteliksel Bulguları :

K-ortalama algoritmasının e-reçete veri seti ile uygulanmasında EK-2' de gösterilen kümeleme sonuçlarına göre aşağıdaki bilgiler özetlenebilir.

- Bir çok bölüm ve farklı tanıli hastalara yazılan ilaçların kümelenmesinde cinsiyet ve yaş değişkeninin etkili olduğu belirlenmiştir.
- Antineoplastik ve İmmünoloji grubundaki ilaçlar, daha çok kadın olmak üzere yaşlı ve orta yaş grubundaki onkoloji tanısı almış hastalara yazılmıştır.
- Sinir sistemi grubundaki ilaçlar genç, orta yaş grubundaki kadınlarda, mental ve davranışsal bozukluklar tanısı konmuş hastalara, erkeklerde ise genç, çocuk grubundaki mental ve davranışsal bozukluklar ve neoplazmlar tanısı konmuş hastalara yazılmıştır.
- Kalp Damar Sistemi grubu ilaçları, çoğunluğu erkek olmak üzere, kadın, yaşlı ve orta yaş grubunda, dolaşım sistemi hastalık tanısı konan kardiyoloji hastalarına yazılmıştır.

Uygulamanın sonuçlarına göre aşağıdaki çıkarımlar yapılabilir.

- K-ortalama algoritmasının büyük veri tanımlamasına uyan HBYS verilerinde çalışılabilir.
- K-ortalama algoritmasının dezavantajına çözüm olarak oluşturulan K-ortalma++, Canopy algoritmalarının söz konusu amaç doğrultusunda değerlendirildiğinde iki algoritmanın da çözüm sağladığı görülmektedir.
- Bu durumda başlangıç merkezlerini ve uygun küme sayısını (k) belirleme kapsamında daha kullanışlı algoritma geliştirilebilir.

4.2. NAMGY Başlangıç Merkezlerini Belirleme Algoritmasının Uygulaması

Mevcut K-Ortalama ve NAMGY Algoritmasının Weka programında verilen diyabet veri setinin üzerinde çalıştırılarak başlangıç merkezlerini belirleme açısından çalışma prensipleri karşılaştırılmıştır. Veri seti bilgileri (Çizelge 4.2) de verilmektedir.

Çizelge 4.2 : Diyabet Veri Seti Bilgileri

Diyabet Veri Seti		Gruplar	Kayıt Sayısı
Değişken Adı	Tipi	1- Test Pozitif	268
Hamile kalma sayısı	Numerik		
Plazma glikoz konsantrasyonu	Numerik		
Diastolik tansiyon (mm Hg)	Numerik	2- Test Negatif	500
Triceps deri kat kalınlığı (mm)	Numerik		
2 saatlik serum insülini (μ U / ml)	Numerik		
Vücut kitle indeksi	Numerik	Toplam Kayıt Sayısı	768
Ailede diyabet öyküsü	Numerik		
Yaş	Numerik		

K-Ortalama Algoritmasında kümeleme sonuçlarının değerlendirmesinde Toplam Karesel Hatanın (SSE) küçük olması istenilen durumdur.

Bu amaçla; WEKA programında 100 farklı başlangıç değerleri (SEED) ile program çalıştırılarak, en az küme sayısı ile en küçük SSE değeri bulunmuş. Sonuçlar NAMGY algoritmasının bulguları ile karşılaştırılmıştır.

Uygulamanın kümeleme sonuçları (Çizelge 4.3.) de gösterilmiştir. Bu sonuçlara göre iki algoritmanın uyumluluğu tam (Kappa katsayısı 1) bulunmuştur.

Çizelge 4.3 : WEKA ve NAMGY Algoritmasının Diyabet Veri Seti İle Kümeleme Sonuçları

Weka- NAMGY Sonuçları			
	NAMGY		
WEKA	1	2	Toplam
1	253	0	253
2	0	515	515
Toplam	253	515	768

Algoritmaların çalışma prensibi açısından farklılıkları (Çizelge 4.4.) de gösterilmiştir.

Farklı çalışma prensipleri içeren iki algoritmanın aynı kümeleme sonuçları verdiği görülmektedir.

Çizelge 4.4 : WEKA ve NAMGY Algoritmasının Çalışma Prensibi Sonuçları

PROGRAMLARIN ÇALIŞMA PRENSİBİ SONUÇLARI		
	WEKA	NAMGY
SSE	121.257902	121.257847
İterasyon Sayısı	7	5
Algoritmayı Çalıştırma Sayısı	100	1

Weka ile oluşturulan kümeleme sonuçları algoritmanın çok kez çalıştırılıp ve gözlemsel değerlendirilerek oluşturulurken, NAMGY Algoritması ile bir kez çalıştırılarak uygun başlangıç merkezleri ile aynı kümeleme sonucu elde edildiği gösterilmiştir.

Bu durumun özellikle büyük veri niteliğindeki verilerde zaman açısından önemli katkı sağlanacağı söylenebilir. Aynı zamanda SEED değerlerini kullanıcı tarafından belirleme yükünü ortadan kaldırdığı için NAMGY algoritmasının kullanıcı dostu olması bir avantajdır.

Ancak iki algoritma kendi aralarında tam uyumlu olsa da diyabet veri seti orijinal sınıflamasıyla uyum daha düşük bulunmuştur (Kappa katsayısı= 0.96). Karşılaştırma sonuçları (Çizelge 4.5) de gösterilmiştir.

Çizelge 4.5 : Orijinal Sınıflama Ve NAMGY Algoritmasının Diyabet Veri Seti İle Kümeleme Sonuçları

Orijinal Sınıflama - NAMGY Algoritma Sonuçları			
Orijinal Sınıflama	NAMGY		Toplam
	1	2	
1	253	15	268
2	0	500	500
Toplam	253	515	768

4.3. NAMGY Algoritması ile Küme Sayısı Belirleme Uygulaması

K-Ortalama algoritmasında küme sayısının belirlenmesinde NAMGY algoritmasında oluşturulan yöntemin geçerliliğini belirlemek amacıyla yapılan çalışma aşağıda özetlenmiştir.

Çukurova Üniversitesi Balcalı Hastanesi HBYS’inden 2010 ve 2013 yılları arasında ameliyat olan 59869 hastanın yatış sürelerini etkileyebileceği düşünülen;

- Yaş,
- Cinsiyet,
- Yapılan cerrahi müdahale hakkındaki bilgiler (Tipi ve Sayısı),
- Girişimsel işlem sayısı,
- Hastalık tanısı,
- Antibiyotik kullanımı bilgileri

değişkenleri çalışmaya alınmıştır.

Analiz aşağıdaki gibi iki aşamada planlanmıştır.

1. Veri madenciliği WEKA programı K-ortalama yöntemi hazırlanan veri setine uygulanmıştır. Uygun küme sayısını (k) belirlemek için farklı k ve seed değerleri ile algoritma tekrar çalıştırılarak hesaplanan toplam karesel hatanın (SSE) ve küme sayısının (k) küçük olduğu durumda kümeleme sonlandırılmıştır.

- Çalışmada K-ortalama algoritmasında uygun küme sayısının 7 olduğu algoritmanın kullanıcı tarafından bir çok kez çalıştırılıp gözlemsel değerlendirmeye tespit edilmiştir.

2. Veriye NAMGY algoritmasındaki uygun küme sayısı kriterlerine göre uygun küme sayısının 7 olduğu belirlenmiştir.

Çizelge 4.6. farklı yüzdelerle (25,50,75,90) kaç küme ile ulaşıldığını ve küme sayısı artarken oluşan SSE ve SSE farkının yüzdelerini göstermektedir.

Çizelge 4.6 : Farklı Küme Sayılarına Göre Kümeleme Sonuçları

YÜZDELİKLERE ULAŞILAN KÜME SAYILARI						
Küme Sayısı k	%25	%50	%75	%90	SSE	SSEFARK %
6	1	2	3	4	166980	
7	1	3	4	5	149837	10
8	1	3	5	6	148823	1
10	2	4	8	8	148405	0
12	3	5	7	9	142303	4
14	3	5	8	11	139301	2
15	3	6	9	12	135500	3
16	3	6	10	13	135006	0
18	3	6	11	14	134049	1
20	3	6	12	16	132497	1

Küme sayıları artarken SSE değerlerin azaldığını, bu azalmanın küme sayısı 7 den sonra stabilleştiği Grafik 4.1. ile desteklenmektedir.



Grafik 4.1 : Küme Sayıları ve SSE Değişim Grafiği

4.4. NAMGY Birleştirilmiş Algoritmanın Uygulaması

Çukurova Üniversitesi Balcalı Hastanesi HBYS'inden 2016 yılında çocuk sağlığı anabilimdalına gelen 4457 hastanın vitaminB12 değerlerini etkileyen faktörleri belirlemek amacıyla NAMGY Algoritması ile kümeleme işlemleri yapılmıştır.

Bu uygulamada Weka ve NAMGY algoritmasını içeren arayüz programı ile;

- Cinsiyet
- Yaş
- Bilimdalı
- Vitamin B12 değeri

veri seti üzerinde çalıştırılmış kümeleme sonuçları elde edilmiştir.

Bu çalışma iki aşamada gerçekleştirilmiştir.

1. Veri madenciliği WEKA programının uygulama aşamaları;

- $k = 2, 3, 4, 5, 6, 7$ küme sayıları parametre değeri alınarak program çalıştırıldı.

- Her bir k için ikinci parametre olarak 30 farklı başlangıç değeri (SEED) alınarak program $6 \times 30 = 180$ kez program çalıştırılıp, kümeleme sonuçları elde edilmiştir.
 - Kümeleme sonuçları gözlemsel olarak değerlendirilerek uygun kümeleme sonuçları belirlenmiştir.
3. NAMGY algoritmasının arayüz programı ile aynı veri seti kullanarak kümeleme sonuçları elde edilmiştir.

Her iki programın uygulamasından elde edilen kümeleme sonuçlarının aynı olduğu (Çizelge 4.7) de gösterilmiştir.

Çizelge 4.7 : WEKA ve NAMGY Arayüz Programın Vitamin B12 Veri Seti Uygulama Sonuçları

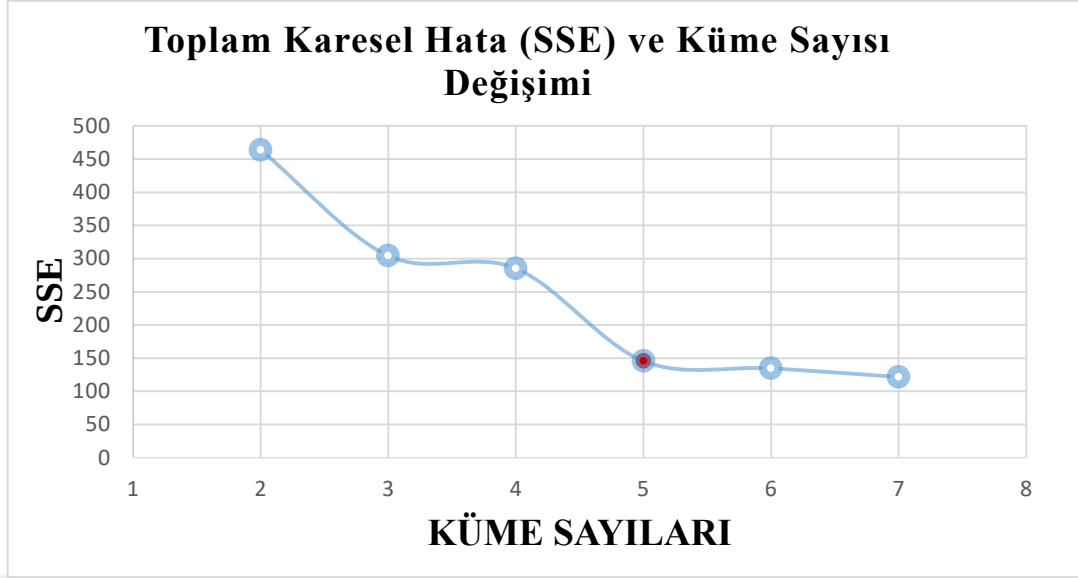
Kümelerin Niceliksel Sonuçları				
Küme Sayısı	Küme Grup No	Küme Grup Kayıt Sayısı	Küme Grup Kayıt Sayısı %	Küme Grup Kayıt Sayısı Yığılımlı %
2	1	2357	53	53
2	2	2100	47	100
3	1	2100	47	47
3	2	1285	29	76
3	3	1072	24	100
4	1	2100	47	47
4	2	1249	28	75
4	3	914	21	95
4	4	194	5	100
5	1	1266	28	28
5	2	1139	26	54
5	3	961	22	76
5	4	880	20	95
5	5	211	5	100
6	1	1146	26	26
6	2	1139	26	51
6	3	961	22	73
6	4	747	17	90
6	5	303	7	96
6	6	161	4	100
7	1	1139	26	26
7	2	961	22	47
7	3	913	20	68
7	4	690	15	83
7	5	280	6	89
7	6	293	7	96
7	7	181	4	100
Toplam Kayıt Sayısı : 4457				

Çizelge 4.8. de NAMGY ve WEKA algoritmalarının vitaminB12 veri setine uygulama sonuçlarına göre küme sayılarına göre farklı yüzdeliklere (25,50,75,90) kaç küme ile ulaşıldığını ve küme sayısı artarken oluşan SSE ve SSE farkının yüzdelerini göstermektedir.

Çizelge 4.8 : Vitamin B12 Verisi Yüzdelik Dilimleri Sağlayan Küme Sayısı Ortak Kümeleme Sonuçları

Küme Sayısı	Toplam Kayıtların Yüzdeliklerine Ulaşılan Küme Sayıları				SSE	SSEFARK %
	25%	50%	75%	90%		
2	1	1	2	2	463.8563	
3	1	2	2	3	304.3152	34.39
4	1	2	2	3	285.2113	6.28
5	1	2	3	4	145.5895	48.95
6	1	2	4	5	134.7872	7.42
7	1	3	4	6	121.7085	9.70

Küme sayıları artarken SSE değerlerin azaldığını, bu azalmanın küme sayısı 5 den sonra daha küçük oranlarda olduğu Grafik 4.2. de görülmektedir.



Grafik 4.2.: Vitamin B12 Veri Sonuçlarına Göre SSE ve Küme Sayıları Değişim Grafiği

İki algoritmanın çalışma prensibi değerlendirildiğinde NAMGY algoritmasının WEKA' ya göre avantajları (Çizelge 4.9) da gösterilmiştir.

Çizelge 4.9 :Weka ve NAMGY Algoritmasının Çalışma Prensibi Açısından Karşılaştırılması

İki Programın Çalışma Prensipleri Açısından Karşılaştırılması		
İşlevsellik	NAMGY	Weka
Küme Sayısı	Uygun küme sayısını program belirliyor	k=2,3,4,5,6,7 olarak parametre
Başlangıç Merkezleri	Program belirliyor	30 farklı başlangıç değeri parametre olarak veriliyor
Programı Çalıştırma Sayısı	1 kez	6*30=180 kez
Süre (dakika)	Toplam 15 dakika	Her k ve 30 seed değeri 10 dakika Toplam 10*6=60 dakika
Kullanıcı Dostu	Kullanıcı deneyimi gerektirmez	Kullanıcı deneyim ve alan bilgisi gerektirir.

- Algoritmalar, çalışma prensibi açısından değerlendirildiğinde WEKA programında nitelikli kümeleme sonuçlarının bir çok küme sayısı (k) ve başlangıç değerleri (SEED) verilerek belirlenirken NAMGY algoritmasını içeren arayüz programıyla aynı kümeleme sonuçlarını kullanıcı bilgisine gerek duymadan daha kısa zamanda belirlenmektedir.

İki algoritmanın kümeleme analiz sonuçlarını oluşturma sürecindeki farklılıkların değerlendirilmesi (Çizelge 4.10) da gösterilmiştir.

Çizelge 4.10 :WEKA ve NAMGY Algoritmasının Analiz Sürecindeki Farklılıkları

Algoritmaların Karşılaştırma Sonuçları					
Küme Sayısı	SSE	İterasyon		Program Çalıştırma Sayısı	
		WEKA	NAMGY	WEKA	NAMGY
2	463.8563	4	1	10	1
3	304.3152	7	1	10	1
4	285.2113	27	6	10	1
5	145.5895	19	15	10	1
6	134.7872	16	9	10	1
7	121.7085	45	21	10	1

Çalışmayla ilgili genel değerlendirmeler aşağıda belirtilmiştir.

- İki algoritmanın kümeleme sonuçlarında aynı küme sayıları ile aynı toplam karesel hata (SSE) elde edilmiştir.
- Ancak; analiz sonuçları aynı olmasına rağmen, analiz sürecinde farklılıklar gözlenmiştir.
- Optimal kümeleme sonucunu belirleme kriterlerinden biri olan en küçük SSE değeri WEKA programı ile analiz sürecinde, her bir küme sayısı (k) için 10 farklı seed değeri rasgele seçilerek algoritma çok kez çalıştırılırken, NAMGY algoritmasını içeren arayüz programı ile algoritma bir kez çalıştırılarak aynı sonuca ulaşılmıştır.

- K-ortalama algoritması döngüsel bir yapıda olması; her bir k için kümelemeyi bitirme işleminde, başlangıç değerlerine göre algoritmanın uygulanma tekrarı (İterasyon) farklılıklar göstermektedir.
- WEKA programında bütün küme sayılarında başlangıç değerlerinin seçimine bağlı olarak iterasyon sayıları daha büyük bulunmuştur. NAMGY algoritmasını içeren arayüz programında uygun başlangıç merkezlerinin noktaların yayılımına ve birbirine uzaklıklarına bağlı olarak program tarafından tesbit edilmesi iterasyon sayısını küçültmüştür.

WEKA programı ile uygun küme sayısının belirlenmesi: kayıtların küme içindeki dağılımlarına ve SSE değişimine bakıldığında gözlemsel olarak uygun küme sayısının (k) 5 olduğu söylenebilir.

NAMGY algoritmasını içeren arayüz programında da uygun küme sayısı algoritmada belirlenen kriterlere göre program tarafından 5 olarak belirlenmiştir.

Sonuç olarak, bulgularının ortak olmasına rağmen iki program çalışma prensibi açısından farklı bulunmuştur.

4.5. NAMGY Birleştirilmiş Algoritmanın Küme Geçerlilikleri Uygulaması

Kümeleme sonucunun performansını göstermek için, literatürde geniş bir şekilde kullanılan WEKA veri datası içindeki ‘iris veri seti’(Fisher 1936) referans veri seti olarak kullanılmıştır.

Sınıflandırılmış ‘iris veri seti’ üzerinde, yazılım uygulanarak, verilerin kümelere ayrılması sağlanmıştır.

Kümeleme sonuçlarının orijinal sınıflanmış veriler ile uyumluluğu Kappa istatistiği yöntemi ile test edilmiştir.

Veri seti (Çizelge 4.11) de gösterilmiştir.

Çizelge 4.11 : İris Veri Seti Yapısı

İRİS VERİ SETİ YAPISI		GRUPLAR	
Değişken Adı	Değişken tipi	Grup Adı	Kayıt Sayısı
Çanak uzunluğu	Nümerik	1- Iris-setosa	50
Çanak genişliği	Nümerik	2- Iris-versicolor	50
Yaprak boyu	Nümerik	3- Iris-virginica	50
Yaprak genişliği	Nümerik	Toplam Kayıt Sayısı	150
Grup	Nümerik		

Referans veriler üzerinde NAMGY Algoritmasının yazılımı uygulanıp, çiçeklerin gruplamaları yapılmış, gruplamalar (Çizelge 4.12) de gösterilmiştir.

Çizelge 4.12 : İris Veri Seti Çapraz Tablo

Orijinal Algoritma Sonuçları Ve NAMGY Algoritmasının Sonuçla Çapraz Tablo				
Orijinal Gruplar	NAMGY Grupları			TOPLAM
	1-Iris Setosa	2-Iris Versicolor	3- Iris Vircinica	
1-Iris Setosa	50	0	0	50
2-Iris Versicolor	0	40	10	50
3- Iris Vircinica	0	8	42	50
TOPLAM	50	48	52	150

İris veri seti için kappa istatistiği hesaplanmış ve kappa istatistiği 0.92 olarak bulunmuştur.

Sonuç olarak NAMGY algoritması ile yapılan bu analizde ;

- K-ortalama kümeleme sonuçlarının değerlendirilmesinde kullanılan SSE değeri küçük,
- Orijinal sınıflanmış veriler ile uyumluluk oranı çok iyi düzeyde (% 92) bulunmuştur.

4.6. NAMGY Algoritmasının Arayüz Programının Oluşturulması

Mevcut K-ortalama algoritmasında uygun küme sayısının ve başlangıç merkezlerinin belirlenmesi problemine çözüm olabilmesi için NAMGY Algoritması farklı veri setleri üzerinde uygulanarak kümeleme sonuçlarının performans değerlendirmeleri yapıldıktan sonra arayüz programı tasarlandı.

Algoritmanın program yazılımı tamamlanarak HBYS' de kullanılabilecek bir modül oluşturuldu. Modülün HBYS'de kullanılması sistemdeki verilerin sağlık hizmet sunumlarına katkı sağlayarak raporların hazırlanmasında kolaylık getireceği düşünülmektedir.

Yazılımsal olarak tamamlanan modülün mimari yapısı ;

- Sunucu ve veri tabanı tarafında yapılan işlemler,
- Kullanıcıların veri tabanında oluşturulan verileri ve hesaplamaları görebileceği arayüz programı

olmak üzere iki katmandan oluşmaktadır.

Arayüz programında NAMGY Algoritması yazılımı kullanım aşamalarını içeren ekran tasarımları ve işlevleri (Şekil 4.1 – 4.8) arasında gösterilmiştir.

Kümeleme analizi yapılacak veri setinin okutulması, sisteme aktarılması, analiz için gerekli hazırlık adımları, kümeleme işleminin yapılması ve küme sonuçlarının raporlanması gibi adımlar programın ana menüsünü oluşturur. Ana menü görüntüleri (Şekil 4.1) de gösterilmiştir.

K-ORTALAMA KÜMELEME ANALİZİ

1- Veri Dosyasını Yükle

Dosya Seç

Dosya seçilmedi

Veriyi Dosyasını Yükle

2- Verileri Aktar

Gönder

3- Veri Alanların Kolon Tiplerini Belirle

Gönder

4- Kümelemeyi Yap

Gönder

5- Kümeleme Özet Bilgiler

Gönder

6- Verilerin Kümelerini Liste

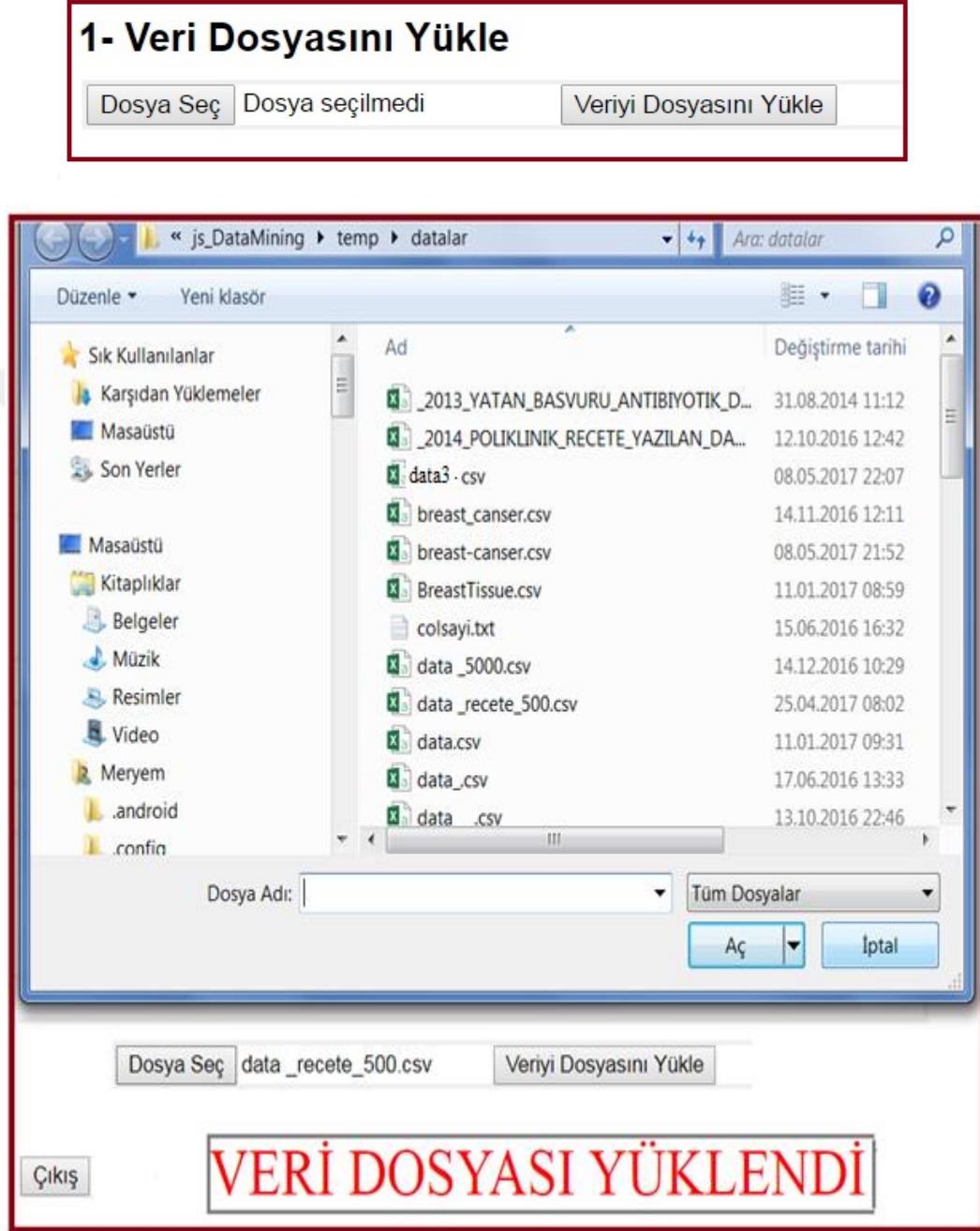
Gönder

7- Kümelenmiş Verilerin Sonuçları

Gönder

Şekil 4.1 : Arayüz Programın Ana Menüsü

Kümelenecek veri dosyasını kopyalama işlemini içeren ekran çıktıları (Şekil 4.2.) de verilmiştir.



Şekil 4.2 : Veri dosyasını Kopyalama menüsü

Kopyalanan dosyadaki verileri veri tabanı tablosuna aktaran menü ve çıktısı (Şekil 4.3.) de verilmektedir.

2- Verileri Aktar

Şekil 4.3 : Verilerin Veri Tabanı Tablosuna Aktaran Menü

Veri setindeki deęişken tiplerinin kategorik yada sayısal niteliklerinin belirlenmesi menüsü ve çıktısı (Şekil 4.4) de verilmektedir.

3- Veri Alanların Kolon Tiplerini Belirle

Gönder

Kolon Tiplerini Belirle

KOLON NO	KOLON ADI	KOLON TİPİ
1	CIN_KOD	KATEGORİK ▼
2	YAS_KOD	NÜMERİK ▼
3	BOL_KOD	KATEGORİK ▼
4	TAN1_KOD	KATEGORİK ▼
5	TAN2_KOD	KATEGORİK ▼
6	ATC1_KOD	KATEGORİK ▼

Kayıt Yap

Çıkış

Şekil 4.4 : Veri Deęişkenlerinin Tiplerini Belirleme Menüsü

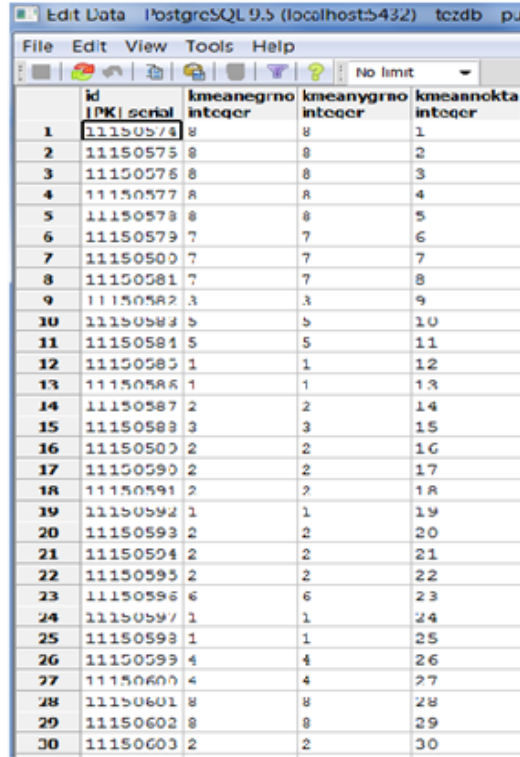
K-ortalama kümeleme algoritmasını tablolara aktarılan veriler üzerinde uygulanarak kümeleme işlemi yapılır. (Şekil 4.5) de gösterilmiştir.

4- Kümelemeyi Yap

Gönder

KÜMELEME YAPILIYOR BEKLEYİNİZ

Çıkış



	id [PK] serial	kmeanegrho integer	kmeanygrno integer	kmeannokta integer
1	11150574	8	8	1
2	11150575	8	8	2
3	11150576	8	8	3
4	11150577	8	8	4
5	11150578	8	8	5
6	11150579	7	7	6
7	11150580	7	7	7
8	11150581	7	7	8
9	11150582	3	3	9
10	11150583	5	5	10
11	11150584	5	5	11
12	11150585	1	1	12
13	11150586	1	1	13
14	11150587	2	2	14
15	11150588	3	3	15
16	11150589	2	2	16
17	11150590	2	2	17
18	11150591	2	2	18
19	11150592	1	1	19
20	11150593	2	2	20
21	11150594	2	2	21
22	11150595	2	2	22
23	11150596	6	6	23
24	11150597	1	1	24
25	11150598	1	1	25
26	11150599	4	4	26
27	11150600	4	4	27
28	11150601	8	8	28
29	11150602	8	8	29
30	11150603	2	2	30

Şekil 4.5 : Kümeleme İşlemi Menüsü

Veri kümelerine ait özet bilgilerin listelendiği menü ve ekran çıktıları (Şekil 4.6.)
daki gibidir.

5- Kümeleme Özet Bilgiler

Küme Sayısı	Küme Kayıt Sayısı	Küme Adı	Başlangıç Merkezleri	Küme SSE	Küme Grup Kayıt Sayısı	Küme İterasyon Sayısı
2	499	1	49	155.144374	121	5
2	499	2	398	155.144374	378	5
3	499	1	49	132.903046	121	9
3	499	2	398	132.903046	246	9
3	499	3	354	132.903046	132	9
4	499	1	49	105.026427	96	6
4	499	2	487	105.026427	121	6
4	499	3	398	105.026427	179	6
4	499	4	438	105.026427	103	6
5	499	1	49	93.939447	68	12
5	499	2	320	93.939447	57	12
5	499	3	398	93.939447	185	12
5	499	4	354	93.939447	68	12
5	499	5	120	93.939447	121	12
6	499	1	49	87.354928	68	10
6	499	2	69	87.354928	60	10
6	499	3	487	87.354928	55	10
6	499	4	398	87.354928	148	10
6	499	5	438	87.354928	47	10
6	499	6	124	87.354928	121	10
7	499	1	49	76.791523	65	8
7	499	2	60	76.791523	56	8
7	499	3	320	76.791523	48	8
7	499	4	398	76.791523	158	8
7	499	5	354	76.791523	51	8
7	499	6	120	76.791523	66	8

Şekil 4.6 : Kümelemeye Ait Özet Bilgiler

K-ortalama kümeleme yöntemine göre ham verilerin ait oldukları kümeleri listeleyen menü ve ekran çıktısı (Şekil 4.7.) verilmektedir.

6- Verilerin Kümelerini Listele

Gönder

Kümelenmiş Veriler

Çıkış

kayitno	kumeno	cin_kod	yas_kod	bol_kod	tan1_kod	tan2_kod	atc1_kod
12	1	2	3	30	584	923	12
13	1	2	3	30	584	923	12
19	1	2	2	11	657	923	10
24	1	2	5	44	663	674	8
25	1	2	5	44	663	674	10
34	1	2	7	44	658	923	8
43	1	2	7	30	579	923	4
44	1	2	7	30	579	923	9
45	1	2	7	28	584	923	9
73	1	2	5	30	626	652	12
82	1	2	5	38	861	923	12
83	1	2	5	38	861	923	12
84	1	2	7	30	613	923	4
85	1	2	7	30	613	923	12
86	1	2	7	30	613	923	13
87	1	2	7	30	613	923	4

Şekil 4.7 : Kümelenmiş Verilerin Listesi

Kümeleri temsil eden özellikler belirlenip listelenip, istenirse excel dosyası olarak oluşturulabildiğini gösteren ekran çıktıları (Şekil 4.8.) de gösterilmiştir.

7- Kümelenmiş Verilerin Değerlendirilmesi

Gönder

Kümelerin Değerlendirilmesi

Excele Aktar

Çıkış						
küme_adı	kayıt_sayısı	kolon_sırası	kolon_adı	değer	sayı_ortalama	kolon_tipi
1	65	1	CIN_KOD	2	65	KATEGORİK
1	65	2	YAS_KOD		5.769231	NÜMERİK
1	65	3	BOL_KOD	30	15	KATEGORİK
1	65	3	BOL_KOD	34	9	KATEGORİK
1	65	3	BOL_KOD	44	8	KATEGORİK
1	65	3	BOL_KOD	38	5	KATEGORİK
1	65	3	BOL_KOD	47	5	KATEGORİK
1	65	3	BOL_KOD	25	5	KATEGORİK
1	65	3	BOL_KOD	28	4	KATEGORİK
1	65	3	BOL_KOD	18	4	KATEGORİK
1	65	3	BOL_KOD	35	3	KATEGORİK
1	65	3	BOL_KOD	17	3	KATEGORİK
1	65	3	BOL_KOD	29	1	KATEGORİK
1	65	3	BOL_KOD	11	1	KATEGORİK
1	65	3	BOL_KOD	19	1	KATEGORİK
1	65	3	BOL_KOD	54	1	KATEGORİK
1	65	4	TAN1_KOD	613	12	KATEGORİK
1	65	4	TAN1_KOD	584	6	KATEGORİK

Şekil 4.8 : Kümeleme Sonuçlarına Göre Küme Özellikleri

5. TARTIŞMA

5.1. K-ortalama Algoritmasının HBYS Verilerine Uygulanması

Literatürde yapılan çalışmalarda görüldüğü üzere K-ortalama algoritması sağlık alanında bir çok bilimsel çalışmada kullanılmıştır.

Sağlık alanında yapılan uygulamalara benzer olarak tez çalışmamızda K-ortalama kümeleme yöntemiyle Çukurova Üniversitesi Balcalı Hastanesi HBYS veri setleri kullanılarak farklı amaçlı alt çalışmalar yapılarak sistem verilerine işlerlik kazandırılması amaçlanmıştır.

- 2010 ve 2013 yılları arasında ameliyat olan 59869 hastanın yatış sürelerini etkileyen faktörleri belirlemek amacıyla yapılan çalışmada hasta yatış sürelerini azaltılabilecek önlemlerin alınmasına destek sağlayacak öneriler sunulmuştur.

- 2014 yılında polikliniklere başvuran ve e-reçete yazılan 365,065 hastaların cinsiyet, yaş, muayeneye geldiği bölüm, ICD-10 tanı kodları (International Classification of Diseases) , ilaçların ATC (Anatomic Therapeutic Chemical Classification-Yapısal Tedavi Edici Kimyasallar Sınıflaması) bilgileri kullanılarak, ilaçların hastalık tanılarına ve demografik özelliklerine göre gruplamaları tesbit edilmiş, ilaç kullanım yoğunluğunun hastalıklara, cinsiyet ve yaş özelliklerine göre analiz edilmiştir.

İlaç üretim miktarları ve maliyetlerini kapsayan ileriye dönük stratejik planlamalarında, hastanın profili ve ilaç sınıflarıyla ilişkisi belirlenen bilgileri kullanılarak daha etkin kararlar alınmasına katkı sağlanabilir.

- Hastaneye yatışı yapılan hastalarda antibiyotik tutarlarının tedavi maliyetlere oranlarını belirlemek amacıyla 2013 yılında hastaneye yatışı yapılan 36788 hasta bilgileri kullanılmıştır. Hastanın; yattığı bölüm, yaş, cinsiyet, hastaneye yatış süresi, antibiyotik tutarları, fatura tutarı bilgileri analiz değişkenleri olarak alınmıştır (48).

Çalışma sonucuna göre;

- ✓ Hastaların tedavi sürecine ve tedavi maliyetlerine antibiyotik kullanımının önemli etkisi olduğu görülmüştür.
- ✓ Akılcı antibiyotik kullanımına destek verilerek, hastane enfeksiyon oranlarının azaltılmasına bağlı olarak hasta yatış sürelerinin kısaltılması tedavi masraflarının düşürülmesinde önemli bir fayda sağlayabilir.

- ✓ Özellikle enfeksiyon hastalıkları uzmanı onayı ile kullanılan antibiyotikler akılcı antibiyotik kullanım oranını artırmıştır.
- ✓ Buna bağlı olarak diğer antibiyotiklerde de aynı uygulamanın yapılması antibiyotik kullanım oranlarını düşürerek tedavi maliyetlerini azaltmada önemli katkı sağlanabileceği söylenebilir.

Uygulamalar HBYS verilerine uyarlanacak bir K-ortalama algoritması hastane süreçlerine katkı sağlayacağını göstermektedir.

5.2. Literatürdeki Başlangıç Merkezlerini Belirleme Algoritma Çalışmaları ile Karşılaştırmalar

Giriş bölümünde detayları verilen literatürdeki çalışmaların NAMGY algoritmasıyla benzerlik ve farklılıkları aşağıdaki gibidir.

[Dalhatu ve Sim (26) Çalışmalarında önerilen algoritma merkezleri belirlemede ;

- Veri setindeki noktalardan yoğunluğu büyük olan (noktaların birbirine yakın olması) noktayı merkez olarak belirlemesi, NAMGY algoritması merkez seçimindeki kriterlerden biri olan komşu noktaların ortalama uzaklığının küçük olması (noktaların birbirine yakın olması) ile benzerlik göstermektedir.

Ancak, NAMGY algoritmasında merkez seçiminde uzaklıklarla beraber ikinci kriter olarak merkezin, komşu nokta sayısının büyük olan noktalardan seçilmesi, küme içindeki noktalar arasındaki uzaklık toplamalarının küçülmesini sağlamaktadır.

- İki algoritmanında merkezleri belirlemede bir eşik değerinin belirlenmesi ve bu değer hesaplanmasında veri setindeki noktalar arasındaki uzaklıkların kullanılması algoritmaların benzerliğini göstermektedir.

Eşik değerin hesaplanmasında; NAMGY algoritmasında noktalar arasındaki uzaklıkların ortalaması gözönüne alınmaktadır. Oysa Dalhatu ve Sim'in algoritmasında noktanın diğer noktalar arasındaki uzaklıklarından yalnızca en küçüğünü ve en büyüğünü dikkate alındığından, uç noktaların çok olduğu veri setlerinde daha büyük eşik değeri bulunmasına neden olabilir.

Büyük eşik değeri ile bulunan merkezler;

- Gerçekten noktaların merkezini temsil etmediği için kümeleme işleminde algoritmanın tekrarlanma (iterasyon) sayının artmasına sebep olabilir.

- Kümeleme sonucunda birbirinden uzak olan noktaların aynı kümeye dahil olma olasılığını artırabilir.

Her iki durum da optimal kümeleme sonuçlarının oluşturulması bir dezavantaj olabilir.

[**Kedar and Sawant (29)**] çalışmalarında önerilen algoritma ile en uygun başlangıç merkezleri seçilerek küme sonuçlarını daha iyi ve algoritmanın uygulanma sayısının (iterasyon) daha az olduğu belirtilmiştir.

Algoritmalar karşılaştırıldığında;

- Kedar and Sawant algoritmasında merkezleri belirlerken, noktalar arasındaki uzaklıkların dikkate alınması NAMGY algoritmasıyla benzerlik göstermektedir.
- Kedar and Sawant algoritmasında; tüm noktalar kendine ait uzaklıklar toplamına göre sıralanıp sistematik olarak k parçaya bölünüp ve ilk noktanın alınması, heterojen veri setlerinde merkez olarak seçilen noktanın aykırı nokta olması durumunda, kümeleme işleminde iterasyon sayısını artırabilir, küme içi uzaklıklar toplamını büyütebilir.
- NAMGY algoritmasında merkezlerin belirlenmesinde Kedar and Sawant algoritmasındaki kritere ek olarak noktaların komşu noktalarla ilgili niteliksel verilerin kullanılması heterojen veri setleri için geçerli bir yöntem olduğu söylenebilir.

[**Singh and Kaur (30)**] çalışmalarında önerilen algoritmalarına göre karşılaştırma sonuçları ;

- Singh and Kaur algoritmasında noktalar arasındaki uzaklıklar kullanılmadığı için merkezlerin basit hesaplamalarla bulunması algoritmanın avantajı olabilir.
- Singh and Kaur algoritması ile veri setini parçalara ayırırken hiçbir metod kullanılmadan grupların rasgele oluşturulması durumunda, niteliklerin ve sıralamaların farklı olduğu veri setlerinde bulunan ortalamalar uygun merkezler olmayabilir. Kümeleme işlemleri uygun başlangıç merkezleriyle yapılmadığı durumlarda; iterasyon sayılarının artmasına ve küme sonuçlarının optimal olmamasına sebep olabilir.

- NAMGY algoritmasında başlangıç merkezlerin seçimi noktaların yayılımına ve merkezlerin birbirine uzaklıklarına göre belirlendiği için veri setindeki noktaların sıralanma farklılıklarından etkilenmemesi algoritmamızın farklı nitelikteki veri setleri için de uygun olması bir avantajdır.

[**Agha and Ashour (31)**] çalışmalarında ElAgha initialization olarak adlandırdıkları algoritmaya göre karşılaştırma sonuçları;

- ElAgha algoritmasında özellikle çok boyutlu veriler için başlangıç merkezlerinin yine noktaların boyutlardaki en küçük ve en büyük değerleri ile tanımlanan köşegen üzerinden seçilmesi, noktaların birbirine uzaklıklarını kullanması açısından NAMGY algoritmasıyla benzerlik göstermektedir.
- ElAgha algoritmasında merkezlerin belirlenen bölge üzerinden rasgele seçilmesi yani; seçilen merkezlerin kendi yakınındaki noktalarla arasındaki uzaklıklardan ve yakın olan noktaların sayılarından bağımsız yapılması, NAMGY algoritmasıyla farklılık göstermektedir.

5.3. Literatürdeki Uygun Küme Sayısını Belirleme Algoritma Çalışmaları İle Karşılaştırmalar

[**Çolak ve ark. (27)**] çalışmalarında ; K-ortalama algoritmasında küme sayısını otomatik bulunması için önerilen algoritmanın karşılaştırma sonucuna göre;

- NAMGY algoritmasında uygun küme sayısını belirlemede toplam hata karelerinin farkı kümelemeyi bitirme koşulu olarak kullanılması Çolak ve ark. algoritması ile benzerlik göstermektedir.
- Ancak **Çolak ve ark.** algoritmasında kullanılan sınır değerinin kullanıcı tarafından belirlenmesi kritik bir farklılıktır. Zira sınır değerinin büyük seçildiği durumunda kümeleme erken bitirilebilir ve efektif kümeleme sonuçları oluşmayabilir. Sınır değer çok küçük seçilirse, veri setinin yapısına bağlı olarak sonsuz döngüye girilebilir ya da küme gruplarına düşen kayıt sayılarının az olması gereksiz olarak küme sayısını büyütmüş olabilir.

- NAMGY algoritmasında kümelemeyi sonlandırma koşulunun toplam hata karelerinin farkına ek olarak kayıtların gruplar içindeki dağılımına da bakılıyor olması kümelemeyi bitirmek için kullanılan ikinci kriteri oluşturmaktadır.
- Kanatımızca iyi kümeleme sonuçlarının ölçütü küçük toplam hata kareleri'ni en az küme sayısı ile elde etmektir. NAMGY algoritmasında her iki kriteri de kümelemeyi bitirme koşulu olarak kullanılmaktadır.

5.4. K-Ortalama++, Canopy Algoritmalarının NAMGY Algoritması ile Çalışma Prensipli Açısından Değerlendirilmesi

NAMGY algoritması ile K-ortalama, K-ortalama++, K-ortalama Canopy algoritmalarının çalışma sistemlerini karşılaştırıldığında;

- Üç algoritmanın Parametre değerlerine göre yaklaşımlarına bakıldığında, NAMGY algoritmasında parametre değerinin kullanıcı tarafından girilmemesi, kayıtların yayılımına göre uygun parametrelerin algoritma tarafından saptanması bir avantajdır.
- NAMGY algoritmasında ve K-ortalama++ algoritmaları başlangıç merkezlerini belirleme anlamında benzerlik göstermektedir. K-ortalama++ önce belirli metodla merkezleri belirlemesi NAMGY algoritmasıyla benzerliğidir. Ancak K-ortalama++ algoritmalarının merkezlerin içindeki seçimi rasgele yapması NAMGY algoritmasında bu işlemin kurala bağlanması farklılıktır.
- NAMGY Algoritmasında ve K-ortalama Canopy algoritmaları başlangıç merkezlerini belirlemede uzaklıkları kullanması benzerliğidir. Ancak Canopy algoritmasında söz konusu uzaklıklar (T_1 , T_2) kullanıcı tarafından belirlenmesine karşın NAMGY algoritmasında verilere dayanarak saptanması farklılığıdır.

NAMGY algoritmasındada parametre değerlerinin

- Küme sayısı
- Başlangıç merkezleri

Verilerinin noktaların yayılım özelliklerine göre bulunması bir kullanım avantajıdır, bu durum algoritmanın daha kullanışlı olduğunu göstermektedir.

Anlatılan benzerlik ve farklılıklar (Çizelge 5.1) ve (Çizelge 5.2) de şematik olarak gösterilmiştir.

Çizelge 5.1 : Algoritmaların Küme Sayısı ve Merkezleri Belirleme Yaklaşımları

Algoritmaların Küme Sayısı ve Küme Merkezlerini Belirleme Yaklaşımları				
Algoritma	Başlangıç Değeri (SEED)	Başlangıç Küme Merkezleri	Küme Sayısı	Parametre
K-Ortalama	Parametreye Bağlı	Rasgele	Parametreye Bağlı	1-Küme Sayısı 2-SEED değeri
K-Ortalama++	Parametreye Bağlı	Algoritma + Rasgele	Parametreye Bağlı	1-Küme Sayısı 2-SEED değeri
K-Ortalama Canopy	Parametreye Bağlı	Parametreye Bağlı	Parametreye Bağlı	1-Küme Sayısı 2-SEED değeri 3-T1 VE T2
NAMGY	Algoritma buluyor	Algoritma buluyor	Algoritma buluyor	Yok

Çizelge 5.2 : NAMGY Algoritmasının Diğer Algoritmalarla Benzerlik ve Farklılıkları

NAMGY Algoritma ile Benzerlik/Farklılık		
	NAMGY	
Algoritma	Benzerlik	Farklılık
K-Ortalama++	Merkezleri Belirlemede iki algoritma da yöntem kullanıyor.	Merkezleri Belirlemede K-Ortalama++ Rasgele seçim kullanıyor. NAMGY algoritmasında verilerin özelliğine göre bulunuyor
K-Ortalama Canopy	Merkezleri belirlemede her iki algoritma da belli bir uzaklık kullanıyor	Merkez seçimleri için belirlenen uzaklıkları K-Ortalama Canopy algoritmasında kullanıcı belirliyor. NAMGY algoritmasında algoritma belirliyor.

NAMGY algoritmasının avantajları aşağıdaki gibi özetlenebilir;

- Parametrelerin kullanıcı tarafından belirlenmemesi
- Başlangıç merkezlerinin verilerin özelliklerine göre belirlenmesi



6. SONUÇLAR VE ÖNERİLER

Sağlık alanında, veri madenciliği yöntemleri kullanılarak büyük veri niteliğindeki tıbbi verilerin sağlık profesyonellerine destek sağlayacak bilgiye dönüşmesini sağlayan bilimsel çalışmalar yapılmaktadır.

Bu çalışmada da, HBYS verilerinin sağlık kurumlarının yönetilmesine katkı sağlayacak bilgilere dönüştürülmesi için;

veri madenciliği K-ortalama kümeleme algoritmasına ekleme yapılarak,

- Küme sayısı,
- Başlangıç noktaları,

parametrelerinin kullanıcı tarafından verilmesine gereksinim duyulmadan en uygun kümeleme sonucunu belirleyen yöntemler geliştirmek amaçlanarak farklı başlıklar altında uygulamalar yapılmıştır.

Yapılan çalışmalarla:

A. Veri madenciliği yöntemlerinin büyük veri niteliğindeki HBYS verilerinde kullanımının uygun olduğu görülmüştür.

B. Sağlık alanındaki çalışmalarda yoğun olarak kullanılan K-ortalama kümeleme algoritmasının büyük hacimli verilere de uygulanabilir olması, HBYS verileriyle kullanımında tercih edilebileceğini göstermiştir.

C. Veri Madenciliği programı olarak yaygın kullanılan Weka ile Çukurova Üniversitesi Balcalı Hastanesi HBYS veri setleri üzerindeki uygulamalarla, sistem verileri bilgiye dönüştürülerek sağlık hizmet sunucularına katkı sağlanmıştır.

D. K-ortalama algoritmasının küme sayısını ve başlangıç merkezlerinin parametre ile belirlenmesinin algoritmanın dezavantajı olduğu algoritmayla yapılan uygulamalarla da gözlemlenmiştir.

E. D' maddesinde belirtilen probleme çözüm olması amacıyla mevcut algoritmaya eklemeler yapılarak NAMGY algoritması geliştirilmiş, uygun küme sayısını ve başlangıç

merkezlerini belirleyen algoritma, standart K-ortalama algoritması ile beraber kullanılmıştır.

Sonuç olarak:

1. K-ortalama algoritmasının küme sayısını ve başlangıç değerlerini belirlemek amacıyla farklı çalışmalarda algoritmalar önerilmiştir.

- ✓ Yaptığımız literatür taramasında her iki probleme aynı algoritma içinde çözüm olacak çalışmaya rastlanmamıştır.
- ✓ NAMGY algoritmasında her iki probleme çözüm olacak yöntemleri içermektedir.

2. Uygun küme sayısını belirlemek için literatürde önerilen algoritmalarda, kümeleme sonuçlarına etkili olan eşik değerleri, kullanıcı tarafından belirlenirken, NAMGY algoritmasında veri yapısına uygun olarak algoritma tarafından belirlenmesi algoritmanın oluşturduğu kümeleme sonuçlarının daha optimal olacağını göstermektedir.

3. Başlangıç merkezlerinin belirlenmesinde noktalar arasındaki uzaklıkların kullanılması algoritmaların ortak yaklaşımlarıdır. NAMGY algoritmasında; aynı zamanda merkeze yakın olan noktaların toplam sayısının da kullanılması, merkezin noktaların daha yoğun olduğu bölgeden alınmasını sağlayacağından algoritmanın uygulama sayısını (iterasyon) azaltmasında etkili olacaktır.

4. Başlangıç merkezleri belirleme problemi için geliştirilen K-ortalama++, Canopy algoritmaları ile NAMGY algoritması aynı veri seti üzerinde çalıştırılarak, üç algoritmanın probleme çözüm yaklaşımları incelendiğinde:

- ✓ K-ortalama ++ ve Canopy algoritmaları kullanıcının belirlemesi gereken başlangıç noktası (seed) ile çalışır, buna ek olarak Canopy algoritması eşik değerleri parametre değeri olarak kullanır. NAMGY algoritmasında başlangıç merkezleri ve küme sayısının, noktaların yayılımına göre algoritma tarafından bulunması algoritmanın bir avantajıdır.
- ✓ NAMGY algoritmasında alan deneyimi gerektirmeden kullanılabilirliği nedeni ile ‘kullanıcı dostu’ olarak nitelendirilebilir.

4. NAMGY algoritmasını içeren arayüz programı HBYS'lerinde bir modül olarak kullanılabilmesi, büyük veri niteliğindeki HBYS'indeki verilere işlerlik kazandırılmasında destek sağlayacaktır.

Tez çalışmasının sonuçları değerlendirildiğinde;

- ✓ HBYS verilerinin bilgiye dönüşümünü sağlayan kullanıcı dostu arayüz programı oluşturularak büyük veri niteliğindeki sistem verilerine işlerlik kazandırılmıştır.
- ✓ Yaygın olarak kullanılan K-ortalama algoritmasında kümeleme sonuçlarını etkileyen küme sayısı ve başlangıç merkezlerinin seçiminde veri setinin özelliklerini kullanarak algoritmanın otomatik belirlemesi optimal kümeleme sonuçları elde etmeyi sağlamıştır.
- ✓ Tez çalışmamızın başında belirlenen hedeflere ulaşıldığı görülmektedir.

İleriye dönük çalışmalarımızda,

HBYS verileri ile yapılan çalışmalarda; sayısal veri setleri üzerinde uygulamalar yapılmıştır. Oysa büyük veri özelliğinde olan tıbbi veriler radyolojik görüntüler gibi farklı formatta verileride içermektedir.

- ✓ HBYS verilerine işlerlik kazandırmak amacıyla ; hazırlanmış olan arayüz programı HBYS'lerin alt yapılarına uygun tanımlamalar yapılarak, sistem içinde modül olarak kullanılabilir.
- ✓ Oluşturulmuş olan arayüz programına, görüntü işleme gibi modüller ekleyerek NAMGY algoritması farklı formattaki veriler üzerinde çalışması sağlanacaktır.
- ✓ Veri madenciliğinin Makine Öğrenmesi, Örüntü Tanıma, Birliktelik Kuralları gibi sağlık alanında çok kullanılan algoritmalarıda arayüz programına eklenerek HBYS'inde kullanılabilen daha kapsamlı veri madenciliği modülleri oluşturulabilir.

7. KAYNAKLAR

1. **Akdur R.**, Türkiye’de Sağlık Hizmetleri ve Avrupa Topluluğu Ükeleri İle Kıyaslanması, 1.Baskı, Ankara: A.Ü.T.F. Halk Sağlığı Anabilim Dalı, **1999**.
2. **Kaya S, Tengilimoğlu D, Işık O, Akbolat M, Yılmaz A**, Sağlık Kurumlarında Kalite Yönetimi, 1.Baskı, Eskişehir:Anadolu Üniversitesi, **2013**.
3. **TC. Sağlık Bakanlığı**, ‘Sağlık İstatistikleri Yıllığı 2015’; Ankara; **2016**
4. **Koyuncuğil A S, Özgülbaş N**. Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları. Bilişim Teknolojileri Dergisi. **2009**; 2/2: 21-32.
5. **Pasin Ö, Ankaralı H**. Sağlık Alanında Yapılan Araştırmalarda Kümeleme Algoritmalarının kullanımı: Bir Uygulama. 17. Ulusal Biyoistatistik Kongresi. Girne, KKTC, 5-9 Kasım **2015**.
6. **Silahtaroglu G**. Kavram Ve Algoritmalarıyla Temel Veri Madenciliği , 1.Basım , İstanbul : Papatya Yayıncılık, 2008.
7. **Daghistani.T, Alshammari. R**. Diagnosis of Diabetes by Applying Data Mining Classification Techniques. (IJACSA) International Journal of Advanced Computer Science and Applications. 2016; 7/7;329-332
8. **Patel.S, Patel.H**, Survey Of Data Mining Techniques Used In Healthcare Domain. International Journal of Information Sciences and Techniques (IJIST). **2016**; 6/1/2; 53-60.
9. **Niaksu.O**. Development And Application Of Data Mining Methods In Medical Diagnostics And Helathcare Management. PhD Thesis. University of Vilnius, Lithuanian, **2015**.
10. **Gayathri. P, Jaisankar.N**. Medical Data Mining Applications and its Uses .International Journal of Engineering Technology Science and Research. **2015**;2/6: ISSN 2394 – 3386
11. **Solmaz. R, Günay.M, Alkan. A**. Uzman Sistemlerin Tiroit Teşhisinde Kullanılması. Akademik Bilişim 2013 – XV. Akademik Bilişim Konferansı Bildirileri. Antalya **2013**: 919-922.
12. **Yurtayı, Y. Salmanı, Y. Ve arkadaşları**. Kansızlık Tanısına İlişkin Bir Veri Madenciliği Uygulaması. ISITES2013 International Symposium On Innovate Technologies In Engineering Science, Sakarya Üniversitesi, **2013** Haziran; 7-9.

13. **Seğmen, E. Uyar, A.** Tıbbi Teşhis Karar Destek Sistemlerinde Sınıflandırıcı Modellerin Performans Analizi. IEEE Sinyal İşleme ve İletişim Uygulamaları (SİU2013) Kurultayı, Uluslararası Kıbrıs Üniversitesi, **2013** Nisan; 1-4.
14. **Dolgun, M.Ö. Dolgun, N.A. Saraçbaşı, O.** Veri Madenciliği Sınıflama Yöntemlerinin Başarılarının; Örneklem Büyüklüğü, Bağımlı Değişken Prevelansı ve Bağımsız Değişkenler Arası İlişki Miktarına Göre Karşılaştırılması. Uluslararası Katılımlı XIV. Ulusal Biyoistatistik Kongresi, Kayseri, **2012** Eylül; 47.
15. **İşler, Y. Narin, A.** WEKA Yazılımında K-Ortalama Algoritması Kullanılarak Konjestif Kalp Yetmezliği Hastalarının Teşhisi, SDU Teknik Bilimler Dergisi, **2012**; 2,4, 21-29.
16. **Turhan, K. Engin, Y.Z. Kamburoğlu, S. Kurt, B. Örem, A.** Potasyum (K) ve Hemoliz İndeksi (HI) Arasındaki İlişkinin Kümeleme Analizi Yöntemi ile İrdelenmesi.IX. Ulusal Tıp Bilişim Kongresi, Biyoistatistik ve Tıp Bilişimi AD, Tıbbi Biyokimya AD, Karadeniz Teknik Üniversitesi, Trabzon. **2012**; 106
17. **Irmak, S. Köksal, C.D. Asilkan, Ö.** Hastanelerin Gelecekteki Hasta Yoğunluklarının Veri Madenciliği Yöntemleri İle Tahmin Edilmesi. Uluslararası Alanya İşletme Fakültesi Dergisi, **2012**; 4,1,101-114
18. **Güllüoğlu, S.S.** Tıp ve Sağlık Hizmetlerinde Veri Madenciliği Çalışmaları: Kanser Teşhisine Yönelik Bir Ön Çalışma. AJIT-e: Online Academic Journal of Information Technology, **2011**; 2,5.
19. **Alapont, J. Bella-Sanjuan, A. Ferri, C. Hernández-Orallo, J. Llopis-Llopis2, J.D. Ramírez-Quintana, M. J.** Specialised Tools for Automating Data Mining for Hospital Management. Dimension Informatica, Av. Cataluna, **2011**; 11, 46020
20. **Sunay, A.S.** Kalpteki Fokal Ventriküler Aritmi Kaynağının Yerinin Tespitinde Kümeleme ve Sınıflandırma uygulamaları. Doktora Tezi, Başkent Üniversitesi Fen Bilimleri Enstitüsü, Ankara, **2010**.
21. **Danacı, M. Çelik, M. Akkaya, A.E.** Veri Madenciliği Yöntemi Kullanılarak Meme Kanseri Hücrelerinin Tahmini ve Teşhisi. Akıllı sistemlerde Yenilikler ve Uygulamaları Sempozyumu, Kayseri, **2010** Haziran; 9-12.
22. **Aksoy, B.** Cluster Analysis Of Decompression Illness. Master 's Thesis, Galatasaray University, Institute of Science and Engineering. İstanbul, **2009**.
23. **Doğan, Ş. Türkoğlu, İ.** Iron-Deficiency Anemia Detection From Hematology Parameters By Using Decision Trees. International Journal of Science&Technology, **2008**; 3,1, 85-92.
24. **Gülten, A. Doğan, Ş.** Genetik Algoritmalar Yönteminin Biyomedikal Verileri Üzerindeki Uygulamaları. Doğu Anadolu Bölgesi Araştırması, Fırat Üniversitesi, Mühendislik Fakültesi, Elektrik – Elektronik Mühendisliği Bölümü, ELAZIĞ,**2008**.

25. **Dinçer, E. Duru, N.** Gırtlak Kanseri Ameliyat Verilerinin K-means Yöntemiyle Analizi. TBV Bilgisayar Bilimleri ve Mühendislik Dergisi, **2006**; Aralık;29-49.
26. **Dalhatu. K, Sim. A T H.** Density base k-Mean's Cluster Centroid Initialization Algorithm. International Journal of Computer Applications. **2016**; 137/11: 49-51.
27. **Çolak.B, Durdağ.Z, Erdoğan.P.** K-Means Algoritması İle Otomatik Kümeleme. El-Cezerî Fen ve Mühendislik Dergisi. **2016**; 3/2;315-323
28. **GODARA, A. SHARMA, V.** Improvement Of Initial Centroids in K Means Clustering Algorithm. IJARIE-ISSN(O)-2395-4396. Volume-2 Issue-2, **2016**
29. **Kedar.B, Sawant.K B.** Efficient Determination of Clusters in K-Mean Algorithm Using Neighborhood Distance. International Journal of Emerging Engineering Research and Technology. **2015**;2349-4409;22-27.
30. **SINGH, H. KAUR, K.** New Method for Finding Initial Cluster Centroids in K-means Algorithm. International Journal of Computer Applications (0975 – 8887). **2013**; 74/6: 27-30
31. **AGHA,M. ASHOUR,W.** Efficient and Fast Initialization Algorithm for K-means Clustering. I.J. Intelligent Systems and Applications.**2012**; 21-31.
32. **Ngafeeson, M N,** Healthcare Information Systems: Opportunities and Challenges, 3rd Ed, USA: Northern Michigan University,**2014**.
33. **Balaraman P, Kosalram K. E** –Hospital Management & Hospital Information Systems – Changing Trends. I.J. Information Engineering and Electronic Business. **2013**; 1: 50-58
34. **Wager A K, Lee, F W. Glaser, J P,** Health Care Information System, 2nd Ed, USA: John Wiley & Sons, Published by Jossey-Bass, **2009**.
35. **Li J S, Zhang Y F, Tian Y,** Medical Big Data Analysis in Hospital Information System . InTech. **2016** . Erişim: <http://dx.doi.org/10.5772/63754> Erişim tarih: 29.06.2017.
36. **Karina K.** Big Data And Its Opportunitities For The Engineering Ompanies. 1nd Ed, Finland : Savonia University of Applied Sciences, **2015**.
37. **Han J,** Data Mining Concepts and Techniques, 3nd Ed, USA: Morgan Kaufmann Publishers is an imprint of Elsevier, **2012**.
38. **Akpınar H.** Data Veri Madenciliği Veri Analizii. 1.Basım, İstanbul : Papatya Yayıncılık, 2014.

39. **Larose D T, Larose C D**, Data Mining And Predictive Analytics, 2nd Ed, USA: John Wiley & Sons, Inc, **2015**.
40. **Gökçe O, Bulduklu Y**. Sağlık Hizmetlerinin Sunumunda “Müşteri Hasta” ve Etik Sorunlar. Sosyal Ekonomik Araştırmalar Dergisi. **2015**; 15/29:39-61
41. **Tomar D, Agarwal S**. A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology, **2013**; 5/5: 241-266.
42. **Kameshwaran K, Malarvizhi K**. Survey on Clustering Techniques in Data Mining. International Journal of Computer Science and Information Technologies, **2014**; 5/2: 2272-2276.
43. **Alpar R**, Uygulamalı İstatistik ve Geçerlilik-Güvenirlilik. 4. Baskı, Ankara: Detay Yayıncılık, **2016**.
44. **Arthur D, Vassilvitskii S**. k-means++: The Advantages of Careful Seeding. Proceeding SODA '07 Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, **2007**; 1027-1035
45. **Kumar A, Ingle Y S, Pande A, Dhule P**. Canopy Clustering: A Review on Pre-Clustering Approach to K-Means Clustering. International Journal of Innovations & Advancement in Computer Science, **2014**; 3/5: 24-28.
46. **Yıldızlı G.M, Alparslan N**. K-Ortalama Kümeleme Yönteminde Başlangıç Merkezlerinin Kümeleme Sonuçlarına Etkisi. XIX. Ulusal ve II. Uluslararası Biyoistatistik Kongresi. Antalya, 25-28 Ekim 2017
47. **The University of Waikato**, WEKA Version 3.7.12, New Zealand: **2014**
48. **Yıldızlı G.M, Alparslan N, Kuşçu F**. Proportion of Cost of Antibiotic to Total Expenditures for Hospitalized Patients of Cukurova University Hospital in 2013. The 8th conference of the Eastern Mediterranean Region of the International Biometric Society (EMR-IBS). Nevşehir , 11-15 May 2015

ÖZGEÇMİŞ

1962 yılında Adana’da doğdum. İlk, orta ve lise eğitimimi Adana’da tamamladım. 1982-1986 yılları arasında Ankara Hacettepe Üniversitesi Fen Fakültesi Matematik bölümünde lisans eğitimimi tamamladım.

1986 yılında özel bir lisede matematik öğretmenliği yaptım. 1987- 1999 yılları arasında Çukurova Üniversitesi Bilgisayar Bilimleri Araştırma Uygulama Merkezinde çalıştım. Bu sürede üniversitenin farklı birimlerine ihtiyaçları doğrultusunda programlar yazdım, lisans ve yüksek lisans öğrencilerine bilgisayar dersleri verdim, merkezin açtığı kurslara eğitimci olarak katıldım.

1996 yılında Çukurova Üniversitesi Tıp Fakültesi Biyoistatistik anabilim dalında yüksek lisansımı tamamladım. 1998 yılında başladığım doktora eğitimime 2003 yılında ara verip, 2011 yılında tekrar başlayarak doktora eğitimime devam ettim.

2000 yılında Çukurova Üniversitesi Balcalı hastanesi bilgi işlem birimine uzman olarak çalışmaya başladım. Hastane Bilgi Yönetim Sistem sorumlusu olarak halen görev yapmaktayım.

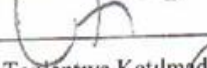
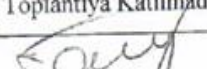
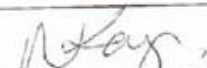
EKLER

EK – 1 : Gereçler ve Yöntemler bölümünde tanımlanan veri setlerinin kullanımı için alınan etik kurul kararı.

T.C. ÇUKUROVA ÜNİVERSİTESİ TIP FAKÜLTESİ GİRİŞİMSSEL OLMAYAN KLİNİK ARAŞTIRMALAR ETİK KURULU

Toplantı Sayısı	Tarih
72	28 Aralık 2017

KARAR NO 10- Biyoistatistik Anabilim Dalı'nda, Z. Nazan Alparslan'ın sorumlu araştırmacı, Uzman Meryem Göral Yıldızlı'nın yardımcı araştırmacı olarak yürütmesi öngörülen, "K - Ortalama, K - Ortalama++, K - Ortalama Canopy ve NAMGY Algoritmalarının Çalışma Prensipleri Açısından Karşılaştırılması ve Uygun Parametre Değerlerinin Belirlenmesini Sağlayan Yöntemin Uygulaması" başlıklı proje araştırma etiği yönünden değerlendirildi. Toplantıya katılan üyelerin oybirliğiyle uygun olduğuna karar verildi.

BAŞKAN	Doç Dr Selim Kadioğlu Tıp Tarihi ve Etik Anabilim Dalı	
ÜYELER	Prof Dr Davut Alptekin Tıbbi Biyoloji Anabilim Dalı	
	Prof Dr Dinçer Yıldızdaş Çocuk Sağlığı ve Hastalıkları Anabilim Dalı	
	Prof Dr Gülşah Seydaoğlu Biyoistatistik Anabilim Dalı	
	Prof Dr Gürhan Sakman Genel Cerrahi Anabilim Dalı	
	Prof Dr Murat Gündüz Anesteziyoloji ve Reanimasyon Anabilim Dalı	Toplantıya Katılmadı
	Doç Dr Ezgi Özyılmaz Saraç Göğüs Hastalıkları Anabilim Dalı	
	Av. Zehra Bulut Hukukçu Üye	Toplantıya Katılmadı
	Dr Neşe Kayrın Kurum Dışı Üye	

Çukurova Üniversitesi Tıp Fakültesi Dekanlık Binası, Balcalı 01330 Adana
Telefon: 0322 338 60 60 dahili 3465, Faks: 0322 338 67 22

EK - 2 : K-ortalama (Random), K-ortalama (Plus), K-ortalama (Canopy)

Algoritmalarının Kümeleme Sonuçları.

K-ORTALAMA(RANDOM) KÜMELEME SONUÇLARI											
RANDOM	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme0	73367 (% 20)	8752 (%12)	64615 (%88)	45-59	43065 (%59)	PSİKİYATRİ	18035 (%25)	MENTAL VE DAVRANIŞSAL BOZUKLUKLARI (F00-F99)	22907 (%31)	Sinir Sistemi	30797 (%42)
				30-44	9837 (%13)	NÖROLOJİ	7941 (%11)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00-M99)	7968 (%11)	Sindirim Sistemi ve Metabolizma	7441 (%10)
				15-29	6779 (%9)	FİZİKSEL TIP- REHABİLİTASYON	4275 (%6)	DOLAŞIM SİSTEMİ HASTALIKLARI (I00-I99)	5567 (%8)	Kas İskelet Sistemi	7165 (%10)
				70-79	5343 (%7)	DAHİLİYE ROMATOLOJİ- İMMUNOLOJİ	3743 (%5)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	5120 (%7)	Kan ve Kan Yapıcı Organlar	5218 (%7)
						DAHİLİYE GASTROENTEROLOJİ	3035 (%4)	SEMPTOMLAR, BELİRTİLER VE ANORMAL KLİNİK VE LABRATUVAR BULGULARI, BAŞKA YERDE SINIFLANMAMIŞ(R00-R99)	4518 (%6)	Kalp Damar Sistemi	4790 (%7)
						DAHİLİYE HEMATOLOJİ	3014 (%4)	SİNİR SİSTEMİ HASTALIKLARI (G00-G99)	4401 (%6)	Çeşitli İlaçlar	4497 (%6)
						GÖZ HASTALIKLARI	2868 (%4)	GENİTOÜRİNER SİSTEM HASTALIKLARI (N00-N99)	4156 (%6)		
						KARDİYOLOJİ	2775 (%4)	NEOPLAZMLAR (C00-D48)	3978 (%5)		
								KAN VE KAN YAPICI ORGANLARIN HASTALIKLARI VE İMMÜN SİSTEMİN BAZI BOZUKLUKLARI (D50-D89)	3743 (%5)		
RANDOM	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme5	52698 (%14)	48607 (%92)	4091 (%8)	45-59	31339 (%59)	DAHİLİYE ONKOLOJİ	26835 (%51)	NEOPLAZMLAR (C00-D48)	28397 (%54)	Sindirim Sistemi ve Metabolizma	24460 (%46)
				60-69	4706 (%9)	DAHİLİYE GASTROENTEROLOJİ	2662 (%5)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	5539 (%11)	Antineoplastik ve İmmünomodülatör Ajanlar	9037 (%17)
				70-79	3944 (%7)	GÖZ HASTALIKLARI	1922 (%4)	ENDOKRİN, BESLENME VE METABOLİZMA HASTALIKLARI (E00-E90)	2815 (%5)	Çeşitli İlaçlar	4081 (%8)
				15-29	3250 (%6)	DAHİLİYE HEMATOLOJİ	1865 (%4)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00-M99)	2331 (%4)	Kan ve Kan Yapıcı Organlar	2998 (%6)
				05-14	3101 (%6)	ÇOCUK SAĞLIĞI	1802 (%3)	SEMPTOMLAR, BELİRTİLER VE ANORMAL KLİNİK VE LABRATUVAR BULGULARI, BAŞKA YERDE SINIFLANMAMIŞ(R00-R99)	2157 (%4)	Enfeksiyona Karşı Kullanılan (Antienfektif) İlaçlar	2568 (%5)
						DAHİLİYE ENDOKRİNOLOJİ	1740 (%3)			Kas İskelet Sistemi	2541 (%5)

RANDOM	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme2	48925 (%13)	44575 (%91)	4350 (%9)	60-69	21204 (%43)	ÜROLOJİ	7048 (%14)	DOLAŞIM SİSTEMİ HASTALIKLARI (I00-I99)	12735 (%26)	Kalp Damar Sistemi	12488 (%26)
				15-29	7144 (%15)	KARDİYOLOJİ	6912 (%14)	GENİTOÜRİNER SİSTEM HASTALIKLARI (N00-N99)	6620 (%14)	Kan ve Kan Yapıcı Organlar	4967 (%10)
				70-79	5947 (%12)	GÖZ HASTALIKLARI	3598 (%7)	NEOPLAZMLAR (C00-D48)	6164 (%13)	Çeşitli İlaçlar	4746 (%10)
				45-59	5706 (%12)	DAHİLİYE HEMATOLOJİ	2658 (%5)	GÖZ VE ADNEKSLERİNİN HASTALIKLARI (H00-H59)	3456 (%7)	Kas İskelet Sistemi	4503 (%9)
				0-4	4142 (%8)	NÖROLOJİ	2334 (%5)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00- M99)	2817 (%6)	Duyu Organları	3781 (%8)
						DAHİLİYE NEFROLOJİ	2311 (%5)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	2665 (%5)	Enfeksiyona Karşı Kullanılan (Antienfektif) İlaçlar	3362 (%7)
						KULAK-BURUN- BOĞAZ	1961 (%4)	SEMPTOMLAR, BELİRTİLER VE ANORMAL KLİNİK VE LABRATUVAR BULGULARI, BAŞKA YERDE SINIFLANMAMIŞ(R00-R99)	2654 (%5)	Sindirim Sistemi ve Metabolizma	3087 (%6)
						DAHİLİYE ENDOKRİNOLOJİ	1887 (%4)	KAN VE KAN YAPICI ORGANLARIN HASTALIKLARI VE İMMÜN SİSTEMİN BAZI BOZUKLUKLARI (D50-D89)	2616 (%5)	Solunum Sistemi	3028 (%6)
						DAHİLİYE ONKOLOJİ	1868 (%4)	SOLUNUM SİSTEMİ HASTALIKLARI (J00-J99)	2492 (%5)		
RANDOM	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme1	46817 (%13)	12206 (%26)	34611 (%74)	30-44	31072 (%66)	KADIN HASTALIKLARI VE DOĞUM	10942 (%23)	GENİTOÜRİNER SİSTEM HASTALIKLARI (N00-N99)	10311 (%22)	Enfeksiyona Karşı Kullanılan (Antienfektif) İlaçlar	19113 (%41)
				15-29	7298 (%16)	DAHİLİYE HEMATOLOJİ	3283 (%7)	KAN VE KAN YAPICI ORGANLARIN HASTALIKLARI VE İMMÜN SİSTEMİN BAZI BOZUKLUKLARI (D50-D89)	4706 (%10)	Kan ve Kan Yapıcı Organlar	4995 (%11)
				0-4	2667 (%6)	ENFEKSİYON HASTALIKLARI VE KLİNİK MİKROBİYOLOJİ	2886 (%6)	SOLUNUM SİSTEMİ HASTALIKLARI (J00-J99)	4245 (%9)	Kas İskelet Sistemi	4722 (%10)
						KULAK-BURUN- BOĞAZ	2865 (%6)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00- M99)	3760 (%8)	Ürogenital Sistem ve Cinsiyet Hormonları	4395 (%9)
						DAHİLİYE GASTROENTEROLOJİ	1725 (%4)	SAĞLIK SERVİSLERİYLE TEMAS VE SAĞLIK DURUMUNU ETKİLEYEN FAKTÖRLER (Z00- Z99)	3346 (%7)	Çeşitli İlaçlar	3692 (%8)
						NÖROLOJİ	1713 (%4)	SEMPTOMLAR, BELİRTİLER VE ANORMAL KLİNİK VE LABRATUVAR BULGULARI, BAŞKA YERDE SINIFLANMAMIŞ(R00-R99)	3293 (%7)	Solunum Sistemi	2248 (%5)
						FİZİKSEL TIP- REHABİLİTASYON	1585 (%3)	NEOPLAZMLAR (C00-D48)	2812 (%6)		
						GÖZ HASTALIKLARI	1504 (%3)	BAZI ENFEKSİYÖZ VE PARAZİTER HASTALIKLAR (A00- B99)	2641 (%6)		
						ÜROLOJİ	1480 (%3)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	2572 (%5)		
						DAHİLİYE ENDOKRİNOLOJİ	1440 (%3)				
						DAHİLİYE NEFROLOJİ	1418 (%3)				
						GENEL CERRAHİ	1345 (%3)				
						GÖĞÜS HASTALIKLARI	1339 (%3)				

RANDOM	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme3	36259 (%10)	5627 (%16)	30632 (%84)	30-44	15500 (%43)	DAHİLİYE ROMATOLOJİ- İMMUNOLOJİ	18867 (%52)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00- M99)	16783 (%46)	Sindirim Sistemi ve Metabolizma	23541 (%65)
				45-59	5526 (%15)	DAHİLİYE GASTROENTEROLOJİ	3035 (%8)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	7067 (%19)	Kas İskelet Sistemi	4609 (%13)
				60-69	4907 (%14)	DAHİLİYE ONKOLOJİ	1814 (%5)	NEOPLAZMLAR (C00-D48)	2842 (%8)	Antineoplastik ve İmmünomodülatör Ajanlar	2142 (%6)
				70-79	2845 (%8)	ÇOCUK SAĞLIĞI	1516 (%4)	ENDOKRİN, BESLENME VE METABOLİZMA HASTALIKLARI (E00-E90)	2010 (%6)		
				05-14	2731 (%8)	FİZİKSEL TIP- REHABİLİTASYON	1278 (%4)	KAN VE KAN YAPICI ORGANLARIN HASTALIKLARI VE İMMÜN SİSTEMİN BAZI BOZUKLUKLARI (D50-D89)	1988 (%5)		
				15-29	2338 (%6)	ÇOCUK GASTROENTEROLOJİ HEPATOLOJİ VE BESLENME	1110 (%3)				
						DAHİLİYE HEMATOLOJİ	985 (%3)				
RANDOM	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme7	31679 (%9)	25773 (%81)	5906 (%19)	05-14	22927 (%72)	ÇOCUK PSİKİYATRİ	18927 (%60)	MENTAL VE DAVRANIŞSAL BOZUKLUKLARI (F00-F99)	18778 (%59)	Sinir Sistemi	19368 (%61)
				15-29	5674 (%18)	ÇOCUK ONKOLOJİ	2279 (%7)	NEOPLAZMLAR (C00-D48)	2326 (%7)	Kan ve Kan Yapıcı Organlar	2865 (%9)
						NÖROLOJİ	1210 (%4)	KAN VE KAN YAPICI ORGANLARIN HASTALIKLARI VE	1798 (%6)	Enfeksiyona Karşı Kullanılan	2609 (%8)
						KULAK-BURUN- BOĞAZ	1153 (%4)	SİNİR SİSTEMİ HASTALIKLARI (G00-G99)	1742 (%5)	Antineoplastik ve İmmünomodülatör Ajanlar	1616 (%5)
						ÇOCUK SAĞLIĞI	1031 (%3)				
RANDOM	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme6	27281 (%7)	2019 (%7)	25262 (%93)	60-69	13780 (%51)	DAHİLİYE ONKOLOJİ	18948 (%69)	NEOPLAZMLAR (C00-D48)	20071 (%74)	Antineoplastik ve İmmünomodülatör Ajanlar	14742 (%54)
				45-59	6167 (%23)	GÖZ HASTALIKLARI	1415 (%5)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00-M99)	1629 (%6)	Kas İskelet Sistemi	2917 (%11)
				30-44	2881 (%11)	ÇOCUK ONKOLOJİ	1093 (%4)	GÖZ VE ADNEKSLERİNİN HASTALIKLARI (H00-H59)	1421 (%5)	Çeşitli İlaçlar	1802 (%7)
				70-79	2008 (%7)					Kan ve Kan Yapıcı Organlar	1665 (%6)
										Sindirim Sistemi ve Metabolizma	1563 (%6)
RANDOM	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme4	24403 (%7)	5830 (%24)	18573 (%76)	15-29	14454 (%59)	DERMATOLOJİ	15054 (%62)	DERİ VE DERİALTI DOKUNUN HASTALIKLARI(L00-L99)	13430 (%55)	Dermatolojik İlaçlar	13199 (%54)
				45-59	2927 (%12)	DAHİLİYE HEMATOLOJİ	1186 (%5)	KAN VE KAN YAPICI ORGANLARIN HASTALIKLARI VE İMMÜN SİSTEMİN BAZI BOZUKLUKLARI (D50-D89)	2211 (%9)	Solunum Sistemi	2680 (%11)
				30-44	2914 (%12)	KULAK-BURUN- BOĞAZ	1035 (%4)	BAZI ENFEKSİYÖZ VE PARAZİTER HASTALIKLAR (A00-B99)	1627 (%7)	Kan ve Kan Yapıcı Organlar	2314 (%9)
				05-14	1531 (%6)					Kas İskelet Sistemi	1380 (%6)

RANDOM	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme9	12881 (%4)	3657 (%28)	9224 (%72)	05-14	8332 (%65)	ÇOCUK ALLERJİ- İMMUNOLOJİ	5955 (%46)	SOLUNUM SİSTEMİ HASTALIKLARI (J00-J99)	8110 (%63)	Solumun Sistemi	7273 (%56)
				0-4	2042 (%16)	GÖĞÜS HASTALIKLARI	1557 (%12)	ENDOKRİN, BESLENME VE METABOLİZMA HASTALIKLARI (E00-E90)	858 (%7)	Enfeksiyona Karşı Kullanılan (Antienfektif) İlaçlar	1057 (%8)
				15-29	836 (%6)	ÇOCUK SAĞLIĞI	964 (%7)	GENİTOÜRİNER SİSTEM HASTALIKLARI (N00-N99)	684 (%5)	Çeşitli İlaçlar	944 (%7)
						KULAK-BURUN- BOĞAZ	774 (%6)	KAN VE KAN YAPICI ORGANLARIN HASTALIKLARI VE İMMÜN SİSTEMİN BAZI BOZUKLUKLARI (D50-D89)	659 (%5)	Kan ve Kan Yapıcı Organlar	892 (%7)
						ÇOCUK NEFROLOJİ	736 (%6)			Endokrin Sistem	649 (%5)
										Kas İskelet Sistemi	647 (%5)
RANDOM	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme8	10755 (%3)	1273 (%12)	9482 (%88)	15-29	4956 (%46)	DAHİLİYE ENDOKRİNOLOJİ	7164 (%67)	ENDOKRİN, BESLENME VE METABOLİZMA HASTALIKLARI (E00-E90)	6227 (%58)	Sindirim Sistemi ve Metabolizma	7019 (%65)
				45-59	2430 (%23)	DAHİLİYE GASTROENTEROLOJİ	698 (%6)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	1433 (%13)	Çeşitli İlaçlar	1384 (%13)
				60-69	1499 (%14)			KAN VE KAN YAPICI ORGANLARIN HASTALIKLARI VE İMMÜN SİSTEMİN BAZI BOZUKLUKLARI (D50-D89)	635 (%6)	Kalp Damar Sistemi	623 (%6)
				30-44	1031 (%10)					Endokrin Sistem	580 (%5)
				70-79	698 (%6)						
K-ORTALAMA(PLUS) KÜMELEME SONUÇLARI											
PLUS	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme0	86066 (%24)	9896 (%11)	76170 (%89)	45-59	46834 (%54)	DAHİLİYE ONKOLOJİ	32332 (%38)	NEOPLAZMLAR (C00-D48)	36461 (%42)	Antineoplastik ve İmmünomodülatör Ajanlar	26573 (%31)
				60-69	15825 (%18)	DAHİLİYE ENDOKRİNOLOJİ	5142 (%6)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00- M99)	9191 (%11)	Sindirim Sistemi ve Metabolizma	12866 (%15)
				70-79	7552 (%9)	GÖZ HASTALIKLARI	4306 (%5)	ENDOKRİN, BESLENME VE METABOLİZMA HASTALIKLARI (E00-E90)	5935 (%7)	Kas İskelet Sistemi	9714 (%11)
				30-44	6057 (%7)	DAHİLİYE ROMATOLOJİ- İMMUNOLOJİ	4274 (%5)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	5869 (%7)	Çeşitli İlaçlar	7540 (%9)
						FİZİKSEL TIP- REHABİLİTASYON	3916 (%5)	GENİTOÜRİNER SİSTEM HASTALIKLARI (N00-N99)	5476 (%6)	Enfeksiyona Karşı Kullanılan (Antienfektif) İlaçlar	6907 (%8)
						DAHİLİYE GASTROENTEROLOJİ	3297 (%4)	DOLAŞIM SİSTEMİ HASTALIKLARI (I00-I99)	4419 (%5)	Duyu Organları	4402 (%5)
						DAHİLİYE HEMATOLOJİ	2649 (%3)	GÖZ VE ADNEKSLERİNİN HASTALIKLARI (H00-H59)	4226 (%5)	Kalp Damar Sistemi	4256 (%5)
						KADIN HASTALIKLARI VE DOĞUM	2420 (%3)			Solumun Sistemi	3911 (%5)
						ORTOPEDİ VE TRAVMATOLOJİ	2309 (%3)				
						ÇOCUK ONKOLOJİ	2220 (%3)				

PLUS	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme7	49998 (%14)	1326 (%3)	48672 (%97)	30-44	30020 (%60)	DAHİLİYE ROMATOLOJİ- İMMUNOLOJİ	17022 (%34)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00- M99)	16727 (%33)	Sindirim Sistemi ve Metabolizma	21265 (%43)
				60-69	5855 (%12)	KADIN HASTALIKLARI VE DOĞUM	6132 (%12)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	7286 (%15)	Kas İskelet Sistemi	6796 (%14)
				45-59	4676 (%9)	DAHİLİYE GASTROENTEROLOJİ	3598 (%7)	GENİTOÜRİNER SİSTEM HASTALIKLARI (N00-N99)	5604 (%11)	Enfeksiyona Karşı Kullanılan (Antienfektif) İlaçlar	5394 (%11)
				15-29	4572 (%9)	DAHİLİYE ENDOKRİNOLOJİ	2985 (%6)	ENDOKRİN, BESLENME VE METABOLİZMA HASTALIKLARI (E00-E90)	3955 (%8)	Kan ve Kan Yapıcı Organlar	3028 (%6)
						FİZİKSEL TIP- REHABİLİTASYON	2113 (%4)	SEMPTOMLAR, BELİRTİLER VE ANORMAL KLİNİK VE LABRATUVAR BULGULARI, BAŞKA YERDE SİNİFLANMAMIŞ(R00-R99)	2495 (%5)	Ürogenital Sistem ve Cinsiyet Hormonları	2664 (%5)
						KULAK-BURUN- BOĞAZ	1609 (%3)	KAN VE KAN YAPICI ORGANLARIN HASTALIKLARI VE İMMÜN SİSTEMİN BAZI BOZUKLUKLARI (D50-D89)	2423 (%5)	Çeşitli İlaçlar	2568 (%5)
						NÖROLOJİ	1543 (%3)				
						GENEL CERRAHİ	1436 (%3)				
PLUS	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme1	47854 (%13)	41825 (%87)	6029 (%13)	15-29	18327 (%38)	ÇOCUK ALLERJİ- İMMUNOLOJİ	6443 (%13)	SOLUNUM SİSTEMİ HASTALIKLARI (J00-J99)	10928 (%23)	Enfeksiyona Karşı Kullanılan (Antienfektif) İlaçlar	18154 (%38)
				60-69	6340 (%13)	ÜROLOJİ	4391 (%9)	GENİTOÜRİNER SİSTEM HASTALIKLARI (N00-N99)	5472 (%11)	Solunum Sistemi	7109 (%15)
				0-4	6182 (%13)	GÖZ HASTALIKLARI	3306 (%7)	NEOPLAZMLAR (C00-D48)	3757 (%8)	Kas İskelet Sistemi	3613 (%8)
				05-14	5956 (%12)	KULAK-BURUN- BOĞAZ	2991 (%6)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00- M99)	3512 (%7)	Çeşitli İlaçlar	3608 (%8)
				70-79	4079 (%9)	ENFEKSİYON HASTALIKLARI VE KLİNİK MİKROBİYOLOJİ	2685 (%6)	GÖZ VE ADNEKSLERİNİN HASTALIKLARI (H00-H59)	3249 (%7)	Duyu Organları	3450 (%7)
				45-59	3202 (%7)	DERMATOLOJİ	2419 (%5)	SEMPTOMLAR, BELİRTİLER VE ANORMAL KLİNİK VE LABRATUVAR BULGULARI, BAŞKA YERDE SİNİFLANMAMIŞ(R00-R99)	3071 (%6)	Sindirim Sistemi ve Metabolizma	2699 (%6)
						GÖĞÜS HASTALIKLARI	2104 (%4)	BAZI ENFEKSİYÖZ VE PARAZİTER HASTALIKLAR (A00- B99)	2826 (%6)		
						DAHİLİYE GASTROENTEROLOJİ	2010 (%4)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	2753 (%6)		
						ÇOCUK SAĞLIĞI	1606 (%3)	DERİ VE DERİALTI DOKUNUN HASTALIKLARI(L00-L99)	2299 (%5)		
						ÇOCUK ONKOLOJİ	1286 (%3)	KAN VE KAN YAPICI ORGANLARIN HASTALIKLARI VE İMMÜN SİSTEMİN BAZI BOZUKLUKLARI (D50-D89)	2166 (%5)		
						NÖROLOJİ	1279 (%3)				
						DAHİLİYE ROMATOLOJİ- İMMUNOLOJİ	1242 (%3)				

PLUS	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme6	38969 (%11)	38354 (%98)	615 (%2)	45-59	23817 (%61)	DAHİLİYE ONKOLOJİ	18257 (%47)	NEOPLAZMLAR (C00-D48)	19525 (%50)	Sindirim Sistemi ve Metabolizma	18022 (%46)
				60-69	7430 (%19)	DAHİLİYE ENDOKRİNOLOJİ	2130 (%5)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	3817 (%10)	Çeşitli İlaçlar	4272 (%11)
				70-79	3027 (%8)	DAHİLİYE GASTROENTEROLOJİ	2052 (%5)	ENDOKRİN, BESLENME VE METABOLİZMA HASTALIKLARI (E00-E90)	2572 (%7)	Kas İskelet Sistemi	3075 (%8)
				0-4	2275 (%6)	GÖZ HASTALIKLARI	1866 (%5)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00- M99)	2417 (%6)	Kan ve Kan Yapıcı Organlar	2794 (%7)
						ÜROLOJİ	1503 (%4)	GÖZ VE ADNEKSLERİNİN HASTALIKLARI (H00-H59)	1869 (%5)	Antineoplastik ve İmmünomodülatör Ajanlar	2566 (%7)
						DAHİLİYE HEMATOLOJİ	1305 (%3)	GENİTOÜRİNER SİSTEM HASTALIKLARI (N00-N99)	1839 (%5)	Endokrin Sistem	2128 (%5)
						DAHİLİYE ROMATOLOJİ- İMMUNOLOJİ	1212 (%3)			Duyu Organları	1930 (%5)
PLUS	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme2	35710 (%10)	28558 (%80)	7152 (%20)	05-14	24998 (%70)	ÇOCUK PSİKİYATRİ	19632 (%55)	MENTAL VE DAVRANIŞSAL BOZUKLUKLARI (F00-F99)	19357 (%54)	Sinir Sistemi	21811 (%61)
				15-29	5228 (%15)	ÇOCUK ONKOLOJİ	2707 (%8)	NEOPLAZMLAR (C00-D48)	3138 (%9)	Sindirim Sistemi ve Metabolizma	4098 (%11)
						ÇOCUK SAĞLIĞI	1687 (%5)	SİNİR SİSTEMİ HASTALIKLARI (G00-G99)	2257 (%6)	Kan ve Kan Yapıcı Organlar	2821 (%8)
						NÖROLOJİ	1501 (%4)	KAN VE KAN YAPICI ORGANLARIN HASTALIKLARI VE İMMÜN SİSTEMİN BAZI BOZUKLUKLARI (D50-D89)	1976 (%6)		
								SEMPATOMLAR, BELİRTİLER VE ANORMAL KLİNİK VE LABRATUVAR BULGULARI, BAŞKA YERDE SINIFLANMAMIŞ(R00-R99)	1758 (%5)		
PLUS	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme5	27289 (%7)	19699 (%72)	7590 (%28)	30-44	22584 (%83)	PSİKİYATRİ	10808 (%40)	MENTAL VE DAVRANIŞSAL BOZUKLUKLARI (F00-F99)	12664 (%46)	Sinir Sistemi	13576 (%50)
				15-29	1928 (%7)	NÖROLOJİ	3154 (%12)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00- M99)	2208 (%8)	Sindirim Sistemi ve Metabolizma	3464 (%13)
				45-59	1830 (%7)	DAHİLİYE GASTROENTEROLOJİ	1077 (%4)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	1845 (%7)	Kas İskelet Sistemi	2336 (%9)
						FİZİKSEL TIP- REHABİLİTASYON	1044 (%4)	SİNİR SİSTEMİ HASTALIKLARI (G00-G99)	1368 (%5)	Çeşitli İlaçlar	1599 (%6)
						KULAK-BURUN- BOĞAZ	973 (%4)	SEMPATOMLAR, BELİRTİLER VE ANORMAL KLİNİK VE LABRATUVAR BULGULARI, BAŞKA YERDE SINIFLANMAMIŞ(R00-R99)	1352 (%5)		
						DAHİLİYE HEMATOLOJİ	957 (%4)	NEOPLAZMLAR (C00-D48)	1255 (%5)		

PLUS	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme3	24685 (%7)	4138 (%17)	20547 (%83)	05-14	10356 (%42)	DERMATOLOJİ	14320 (%58)	DERİ VE DERİALTI DOKUNUN HASTALIKLARI(L00-L99)	12778 (%52)	Dermatolojik İlaçlar	12607 (%51)
				15-29	4172 (%17)	ÇOCUK SAĞLIĞI	1858 (%8)	BAZI ENFEKSİYÖZ VE PARAZİTER HASTALIKLAR (A00-B99)	1637 (%7)	Sindirim Sistemi ve Metabolizma	3019 (%12)
				30-44	4043 (%16)	ÇOCUK NEFROLOJİ	1476 (%6)	GENİTOÜRİNER SİSTEM HASTALIKLARI (N00-N99)	1565 (%6)	Solunum Sistemi	2665 (%11)
				45-59	2928 (%12)	ÇOCUK ONKOLOJİ	979 (%4)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	1525 (%6)	Enfeksiyona Karşı Kullanılan (Antienfektif) İlaçlar	2100 (%9)
				60-69	1701 (%7)	ÇOCUK GASTROENTEROLOJİ HEPATOLOJİ VE BESLENME	931 (%4)	ENDOKRİN, BESLENME VE METABOLİZMA HASTALIKLARI (E00-E90)	1335 (%5)		
								SEMPTOMLAR, BELİRTİLER VE ANORMAL KLİNİK VE LABRATUVAR BULGULARI, BAŞKA YERDE SINIFLANMAMIŞ(R00-R99)	1118 (%5)		
PLUS	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme8	20384 (%6)	1367 (%7)	19017 (%93)	15-29	12831 (%63)	DAHİLİYE HEMATOLOJİ	6558 (%32)	KAN VE KAN YAPICI ORGANLARIN HASTALIKLARI VE İMMÜN SİSTEMİN BAZI BOZUKLUKLARI (D50-D89)	8723 (%43)	Kan ve Kan Yapıcı Organlar	9611 (%47)
				30-44	2087 (%10)	KADIN HASTALIKLARI VE DOĞUM	2642 (%12)	GENİTOÜRİNER SİSTEM HASTALIKLARI (N00-N99)	2492 (%12)	Ürogenital Sistem ve Cinsiyet Hormonları	1718 (%8)
				60-69	1579 (%8)	NÖROLOJİ	1467 (%7)	NEOPLAZMLAR (C00-D48)	1031 (%5)	Kas İskelet Sistemi	1559 (%8)
				45-59	1525 (%7)	KULAK-BURUN-BOĞAZ	1017 (%5)	SOLUNUM SİSTEMİ HASTALIKLARI (J00-J99)	990 (%5)	Çeşitli İlaçlar	1362 (%7)
				70-79	1158 (%6)	DAHİLİYE ENDOKRİNOLOJİ	811 (%4)	SEMPTOMLAR, BELİRTİLER VE ANORMAL KLİNİK VE LABRATUVAR BULGULARI, BAŞKA YERDE SINIFLANMAMIŞ(R00-R99)	949 (%5)	Sinir Sistemi	1186 (%6)
						DAHİLİYE GENEL	600 (%3)			Solunum Sistemi	1174 (%6)
						GÖZ HASTALIKLARI	547 (%3)				
						ÇOCUK HEMATOLOJİ	523 (%3)				
						DAHİLİYE NEFROLOJİ	516 (%3)				
PLUS	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme4	18850 (%5)	13156 (%70)	5694 (%30)	60-69	9838 (%52)	KARDİYOLOJİ	9002 (%48)	DOLAŞIM SİSTEMİ HASTALIKLARI (I00-I99)	12839 (%68)	Kalp Damar Sistemi	10652 (%57)
				45-59	4099 (%22)	DAHİLİYE ENDOKRİNOLOJİ	1592 (%8)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	1386 (%7)	Kan ve Kan Yapıcı Organlar	2274 (%12)
				70-79	2226 (%12)	NÖROLOJİ	1109 (%6)	ENDOKRİN, BESLENME VE METABOLİZMA HASTALIKLARI (E00-E90)	995 (%5)	Sindirim Sistemi ve Metabolizma	2060 (%11)
				30-44	1474 (%8)	DAHİLİYE NEFROLOJİ	1087 (%6)				
						DAHİLİYE ROMATOLOJİ-İMMUNOLOJİ	746 (%4)				

PLUS	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme9	15260 (%4)	0	(%0)	15260 (%100)	45-59 7187 (%47)	PSİKİYATRİ	7103 (%47)	MENTAL VE DAVRANIŞSAL BOZUKLUKLARI (F00-F99)	9069 (%59)	Sinir Sistemi	13482 (%88)
					15-29 3150 (%21)	NÖROLOJİ	3038 (%20)	SİNİR SİSTEMİ HASTALIKLARI (G00-G99)	1939 (%13)		
					60-69 2676 (%18)	AĞRI	1195 (%8)	SEMPATOMLAR, BELİRTİLER VE ANORMAL KLİNİK VE LABRATUVAR BULGULARI, BAŞKA YERDE SINIFLANMAMIŞ(R00-R99)	1153 (%8)		
					70-79 1377 (%9)	FİZİKSEL TIP- REHABİLİTASYON	824 (%5)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00- M99)	988 (%6)		

K-ORTALAMA(CANOPY) KÜMELEME SONUÇLARI											
CANOPY	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme0	74505 (%20)	62106 (%83)	12399 (%17)	5_14 (%42)	31310 (%42)	ÇOCUK PSİKİYATRİ	19722 (%26)	MENTAL VE DAVRANIŞSAL BOZUKLUKLARI (F00-F99)	21624 (%29)	Sinir Sistemi	25638 (%34)
				15-29 (%20)	15064 (%20)	ÇOCUK ONKOLOJİ	4895 (%7)	NEOPLAZMLAR (C00-D48)	7181 (%10)	Enfeksiyona Karşı Kullanılan (Antienfektif) İlaçlar	10714 (%14)
				30-44 (%16)	11851 (%16)	DERMATOLOJİ	4824 (%6)	KAN VE KAN YAPICI ORGANLARIN HASTALIKLARI VE İMMÜN SİSTEMİN BAZI BOZUKLUKLARI (D50-D89)	5718 (%8)	Kan ve Kan Yapıcı Organlar	6934 (%9)
				0-4 (%8)	5889 (%8)	NÖROLOJİ	3928 (%5)	GENİTOÜRİNER SİSTEM HASTALIKLARI (N00-N99)	5703 (%8)	Çeşitli İlaçlar	5809 (%8)
				70-79 (%6)	4700 (%6)	ÇOCUK SAĞLIĞI	3231 (%4)	SEMPATOMLAR, BELİRTİLER VE ANORMAL KLİNİK VE LABRATUVAR BULGULARI, BAŞKA YERDE SINIFLANMAMIŞ(R00-R99)	4837 (%6)	Dermatolojik İlaçlar	5021 (%7)
						ÇOCUK NEFROLOJİ	3129 (%4)	DERİ VE DERİALTI DOKUNUN HASTALIKLARI(L00-L99)	4636 (%6)	Kas İskelet Sistemi	4430 (%6)
						ÜROLOJİ	2993 (%4)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00- M99)	4058 (%5)	Antineoplastik ve İmmünomodülatör Ajanlar	4362 (%6)
						DAHİLİYE HEMATOLOJİ	2569 (%3)	ENDOKRİN, BESLENME VE METABOLİZMA HASTALIKLARI (E00-E90)	3935 (%5)	Sindirim Sistemi ve Metabolizma	4134 (%6)
						ÇOCUK HEMATOLOJİ	2410 (%3)	SİNİR SİSTEMİ HASTALIKLARI (G00-G99)	3810 (%5)		
						KULAK-BURUN- BOĞAZ	2241 (%3)				

CANOPY	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme1	62251 (%17)	3199 (%5)	59052 (%95)	30-44	39612 (%64)	KADIN HASTALIKLARI VE DOĞUM	12492 (%20)	GENİTOÜRİNER SİSTEM HASTALIKLARI (N00-N99)	10099 (%16)	Sindirim Sistemi ve Metabolizma	21292 (%34)
				15-29	7033 (%11)	DAHİLİYE ROMATOLOJİ- İMMUNOLOJİ	4002 (%6)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	8238 (%13)	Enfeksiyona Karşı Kullanılan (Antienfektif) İlaçlar	7662 (%12)
				0-4	5222 (%8)	DAHİLİYE GASTROENTEROLOJİ	3975 (%6)	KAN VE KAN YAPICI ORGANLARIN HASTALIKLARI VE İMMÜN SİSTEMİN BAZI BOZUKLUKLARI (D50-D89)	5924 (%10)	Kan ve Kan Yapıcı Organlar	5919 (%10)
				70-79	4295 (%7)	DERMATOLOJİ	3473 (%6)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00- M99)	5893 (%9)	Ürogenital Sistem ve Cinsiyet Hormonları	4884 (%8)
						DAHİLİYE HEMATOLOJİ	3428 (%6)	NEOPLAZMLAR (C00-D48)	4847 (%8)	Çeşitli İlaçlar	4315 (%7)
						NÖROLOJİ	2970 (%5)	SEMPTOMLAR, BELİRTİLER VE ANORMAL KLİNİK VE LABRATUVAR BULGULARI, BAŞKA YERDE SINIFLANMAMIŞ(R00-R99)	4312 (%7)		
						ÇOCUK SAĞLIĞI	2392 (%4)	ENDOKRİN, BESLENME VE METABOLİZMA HASTALIKLARI (E00-E90)	4244 (%7)		
						KULAK-BURUN- BOĞAZ	2290 (%4)	SAĞLIK SERVİSLERİYLE TEMAS VE SAĞLIK DURUMUNU ETKİLEYEN FAKTÖRLER (Z00- Z99)	3429 (%6)		
						DAHİLİYE ENDOKRİNOLOJİ	2290 (%4)				
						FİZİKSEL TIP- REHABİLİTASYON	2246 (%4)				
						DAHİLİYE ONKOLOJİ	1875 (%3)				
						GENEL CERRAHİ	1689 (%3)				
						ÇOCUK GASTROENTEROLOJİ HEPATOLOJİ VE BESLENME	1564 (%3)				
CANOPY	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme2	51513 (%14)	9752 (%19)	41761 (%81)	45-59	34356 (%67)	DAHİLİYE ONKOLOJİ	28886 (%56)	NEOPLAZMLAR (C00-D48)	30641 (%59)	Antineoplastik ve İmmünomodülatör Ajanlar	25286 (%49)
				60-69	5351 (%10)	DAHİLİYE HEMATOLOJİ	2793 (%5)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00- M99)	2881 (%6)	Çeşitli İlaçlar	3359 (%7)
				30-44	4050 (%8)	DERMATOLOJİ	2369 (%5)	GENİTOÜRİNER SİSTEM HASTALIKLARI (N00-N99)	2327 (%5)	Sinir Sistemi	3260 (%6)
				70-79	3243 (%6)	NÖROLOJİ	1716 (%3)	KAN VE KAN YAPICI ORGANLARIN HASTALIKLARI VE İMMÜN SİSTEMİN BAZI BOZUKLUKLARI (D50-D89)	2114 (%4)	Kan ve Kan Yapıcı Organlar	3238 (%6)
						DAHİLİYE ROMATOLOJİ- İMMUNOLOJİ	1618 (%3)	DERİ VE DERİALTI DOKUNUN HASTALIKLARI(L00-L99)	2061 (%4)	Sindirim Sistemi ve Metabolizma	3116 (%6)
						ÇOCUK ONKOLOJİ	1313 (%3)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	1957 (%4)	Enfeksiyona Karşı Kullanılan (Antienfektif) İlaçlar	3036 (%6)

CANOPY	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme7	38417 (%11)	38405 (%100)	12 (%0)	45-59	26499 (%69)	DAHİLİYE ONKOLOJİ	10433 (%27)	NEOPLAZMLAR (C00-D48)	11671 (%30)	Sindirim Sistemi ve Metabolizma	16311 (%42)
				15-29	3190 (%8)	KARDİYOLOJİ	3409 (%9)	DOLAŞIM SİSTEMİ HASTALIKLARI (I00-I99)	6408 (%17)	Kalp Damar Sistemi	4370 (%11)
				70-79	3115 (%8)	DAHİLİYE GASTROENTEROLOJİ	2698 (%7)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	4587 (%12)	Çeşitli İlaçlar	3983 (%10)
				0-4	2293 (%6)	DAHİLİYE ENDOKRİNOLOJİ	2504 (%7)	ENDOKRİN, BESLENME VE METABOLİZMA HASTALIKLARI (E00-E90)	2719 (%7)	Enfeksiyona Karşı Kullanılan (Antienfektif) İlaçlar	3307 (%9)
						ÜROLOJİ	2143 (%6)	GENİTOÜRİNER SİSTEM HASTALIKLARI (N00-N99)	2350 (%6)	Kan ve Kan Yapıcı Organlar	3165 (%8)
						DAHİLİYE HEMATOLOJİ	1458 (%4)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00-M99)	2349 (%6)	Kas İskelet Sistemi	2599 (%7)
						DAHİLİYE NEFROLOJİ	1353 (%4)				
						DAHİLİYE ROMATOLOJİ-İMMUNOLOJİ	1247 (%3)				
						DERMATOLOJİ	1236 (%3)				
						NÖROLOJİ	1053 (%3)				
						ENFEKSİYON HASTALIKLARI VE KLİNİK MİKROBİYOLOJİ	1047 (%3)				
						ÇOCUK SAĞLIĞI	986 (%3)				
						GENEL CERRAHİ	972 (%3)				
CANOPY	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme4	37165 (%10)	4608 (%12)	32557 (%88)	15-29	20699 (%56)	PSİKİYATRİ	15810 (%43)	MENTAL VE DAVRANIŞSAL BOZUKLUKLARI (F00-F99)	18660 (%50)	Sinir Sistemi	20793 (%56)
				30-44	6009 (%16)	NÖROLOJİ	3771 (%10)	DERİ VE DERİALTI DOKUNUN HASTALIKLARI (L00-L99)	3542 (%10)	Dermatolojik İlaçlar	3112 (%8)
				45-59	4451 (%12)	DERMATOLOJİ	3741 (%10)	KAN VE KAN YAPICI ORGANLARIN HASTALIKLARI VE İMMÜN SİSTEMİN BAZI BOZUKLUKLARI (D50-D89)	2685 (%7)	Kan ve Kan Yapıcı Organlar	2689 (%7)
				60-69	3438 (%9)	DAHİLİYE HEMATOLOJİ	1374 (%4)	SİNİR SİSTEMİ HASTALIKLARI (G00-G99)	1767 (%5)	Enfeksiyona Karşı Kullanılan (Antienfektif) İlaçlar	2502 (%7)
						KULAK-BURUN-BOĞAZ	1359 (%4)	SEMPATOMLAR, BELİRTİLER VE ANORMAL KLİNİK VE LABRATUVAR BULGULARI, BAŞKA YERDE SINIFLANMAMIŞ (R00-R99)	1766 (%5)		
CANOPY	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme5	28190 (%8)	9467 (%34)	18723 (%66)	60-69	25873 (%92)	DAHİLİYE ONKOLOJİ	11161 (%40)	NEOPLAZMLAR (C00-D48)	11587 (%41)	Sindirim Sistemi ve Metabolizma	12145 (%43)
						DAHİLİYE ENDOKRİNOLOJİ	1937 (%7)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00-M99)	2752 (%10)	Kas İskelet Sistemi	2464 (%9)
						DAHİLİYE GASTROENTEROLOJİ	1595 (%6)	DOLAŞIM SİSTEMİ HASTALIKLARI (I00-I99)	2702 (%10)	Çeşitli İlaçlar	2459 (%9)
						DAHİLİYE ROMATOLOJİ-İMMUNOLOJİ	1583 (%6)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	2649 (%9)	Kalp Damar Sistemi	2265 (%8)
						KARDİYOLOJİ	1439 (%5)	ENDOKRİN, BESLENME VE METABOLİZMA HASTALIKLARI (E00-E90)	2401 (%9)	Kan ve Kan Yapıcı Organlar	2216 (%8)
						FİZİKSEL TIP-REHABİLİTASYON	1209 (%4)			Enfeksiyona Karşı Kullanılan (Antienfektif) İlaçlar	1906 (%7)
						DAHİLİYE HEMATOLOJİ	1111 (%4)				

CANOPY	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme6	21440 (%6)	17378 (%81)	4062 (%19)	60-69	13127 (%61)	GÖZ HASTALIKLARI	10620 (%50)	GÖZ VE ADNEKSLERİNİN HASTALIKLARI (H00-H59)	10455 (%49)	Duyu Organları	10926 (%51)
				45-59	2872 (%13)	ÜROLOJİ	1815 (%8)	DOLAŞIM SİSTEMİ HASTALIKLARI (I00-I99)	2719 (%13)	Kalp Damar Sistemi	2318 (%11)
				70-79	1736 (%8)	KARDİYOLOJİ	1668 (%8)	GENİTOÜRİNER SİSTEM HASTALIKLARI (N00-N99)	1658 (%8)	Enfeksiyona Karşı Kullanılan (Antienfektif) İlaçlar	1869 (%9)
						DAHİLİYE HEMATOLOJİ	762 (%4)	NEOPLAZMLAR (C00-D48)	1188 (%6)	Çeşitli İlaçlar	1418 (%7)
						KULAK-BURUN-BOĞAZ	757 (%4)			Kan ve Kan Yapıcı Organlar	1322 (%6)
CANOPY	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme9	19943 (%5)	2006 (%10)	17937 (%90)	45-59	11057 (%55)	DAHİLİYE ROMATOLOJİ-İMMUNOLOJİ	13954 (%70)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00-M99)	15242 (%76)	Kas İskelet Sistemi	11720 (%59)
				30-44	3361 (%17)	FİZİKSEL TIP-REHABİLİTASYON	1744 (%9)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	1146 (%6)	Enfeksiyona Karşı Kullanılan (Antienfektif) İlaçlar	1620 (%8)
				70-79	1531 (%8)	ORTOPEDİ VE TRAVMATOLOJİ	1287 (%6)			Sindirim Sistemi ve Metabolizma	1467 (%7)
				60-69	1487 (%7)					Kan ve Kan Yapıcı Organlar	1223 (%6)
				15-29	1470 (%7)						
CANOPY	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme8	16223 (%4)	0 (%0)	16223 (%100)	45-59	13060 (%81)	DAHİLİYE ENDOKRİNOLOJİ	5469 (%34)	DOLAŞIM SİSTEMİ HASTALIKLARI (I00-I99)	5445 (%34)	Sindirim Sistemi ve Metabolizma	9116 (%56)
				70-79	1816 (%11)	DAHİLİYE GASTROENTEROLOJİ	1944 (%12)	ENDOKRİN, BESLENME VE METABOLİZMA HASTALIKLARI (E00-E90)	3978 (%25)	Kalp Damar Sistemi	3083 (%19)
						KARDİYOLOJİ	1825 (%11)	SİNDİRİM SİSTEMİ HASTALIKLARI (K00-K93)	2860 (%18)	Çeşitli İlaçlar	906 (%6)
						DAHİLİYE ROMATOLOJİ-İMMUNOLOJİ	1490 (%9)	KAS İSKELET SİSTEMİ VE BAĞ DOKUSU HASTALIKLARI (M00-M99)	1195 (%7)		
						FİZİKSEL TIP-REHABİLİTASYON	751 (%5)				
CANOPY	Kayıt	Cinsiyet		Yaş Grubu		Bölüm		Tanı 1		ATC	
	Sayısı	E	K	Grup	Sayı	Adı	Sayı	ADI	Sayı	ADI	Sayı
Küme3	15418 (%4)	11398 (%74)	4020 (%26)	5_14	6327 (%41)	ÇOCUK ALLERJİ-İMMUNOLOJİ	6558 (%43)	SOLUNUM SİSTEMİ HASTALIKLARI (J00-J99)	10231 (%66)	Solunum Sistemi	12275 (%80)
				0-4	2511 (%16)	KULAK-BURUN-BOĞAZ	2391 (%16)	DERİ VE DERİALTI DOKUNUN HASTALIKLARI (L00-L99)	1108 (%7)	Enfeksiyona Karşı Kullanılan (Antienfektif) İlaçlar	1400 (%9)
				15-29	2097 (%14)	GÖĞÜS HASTALIKLARI	2037 (%13)				
				45-59	1360 (%9)	DERMATOLOJİ	1135 (%7)				
				30-44	1237 (%8)						
				60-69	1126 (%7)						