İBRAHİM ŞAHİN

BİYOİSTATİSTİK ANABİLİM DALI YÜKSEK LİSANS TEZİ

2019

T.C.
BURSA ULUDAĞ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK ANABİLİM DALI

k-ORTALAMALAR KÜMELEME YÖNTEMİ İLE HİPOTİRODİZM TANISI KONMUŞ OLGULARIN BÜYÜK VERİ KULLANILARAK İNCELENMESİ

İBRAHİM ŞAHİN

YÜKSEK LİSANS TEZİ

BURSA-2019

T.C.
BURSA ULUDAĞ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK ANABİLİM DALI

k-ORTALAMALAR KÜMELEME YÖNTEMİ İLE HİPOTİRODİZM TANISI KONMUŞ OLGULARIN BÜYÜK VERİ KULLANILARAK İNCELENMESİ

İBRAHİM ŞAHİN

(YÜKSEK LİSANS TEZİ)

DANIŞMAN:
Prof.Dr. İlker ERCAN

BURSA-2019

T.C.
ULUDAĞ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

ETİK BEYANI

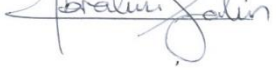
Yüksek Lisans tezi olarak sunduğum,

"k-ortalamalar Kümeleme Yöntemi ile Hipotirodizm Tanısı Konmuş Olguların Büyük Veri Kullanılarak İncelenmesi" adlı çalışmanın, proje safhasından sonuçlanmasına kadar geçen bütün süreçlerde bilimsel etik kurallarına uygun bir şekilde hazırlandığını ve yararlandığım eserlerin kaynaklar bölümünde gösterilenlerden oluştuğunu belirtir ve beyan ederim.

İBRAHİM ŞAHİN

Tarih ve İmza

08.07.2019



SAĞLIK BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜ'NE

Biyoistatistik Anabilim Dalı Yüksek Lisans öğrencisi İbrahim Şahin tarafından hazırlanan "k-Ortalamalar Kümeleme Yöntemi ile Hipotirodizm Tanısı Konmuş Olguların Büyük Veri Kullanılarak İncelenmesi" konulu Yüksek Lisans tezi 19/07/2019 günü, 11:00 – 12:00 saatleri arasında yapılan tez savunma sınavında jüri tarafından oy birliği/~~oy çokluğu~~ ile kabul edilmiştir.

	Adı-Soyadı	**İmza**
Tez Danışmanı	Prof. Dr. İlker Ercan	
Üye	Doç. Dr. Deniz Sığırlı	
Üye	Doç. Dr. Selcen Yüksel	
Üye		
Üye		

Bu tez Enstitü Yönetim Kurulu'nun ..... tarih ve ..... sayılı toplantısında alınan ..... numaralı kararı ile kabul edilmiştir.

Prof.Dr. Gülşah Çeçener
Enstitü Müdürü

TEZ KONTROL ve BEYAN FORMU

08/07/2019

Adı Soyadı: İbrahim SATLIN

Anabilim Dalı: Biyoistatistik

Tez Konusu: k-Ortalamalar Kümeleme Yöntemi ile Hipotiroidizm Tanısı Konmuş Olguların Büyük Veri Kullanılarak İncelenmesi

ÖZELLİKLER	UYGUNDUR	UYGUN DEĞİLDİR	AÇIKLAMA
Tezin Boyutları	☑	☐	
Dış Kapak Sayfası	☑	☐	
İç Kapak Sayfası	☑	☐	
Kabul Onay Sayfası	☑	☐	
Sayfa Düzeni	☑	☐	
İçindekiler Sayfası	☑	☐	
Yazı Karakteri	☑	☐	
Satır Aralıkları	☑	☐	
Başlıklar	☑	☐	
Sayfa Numaraları	☑	☐	
Eklerin Yerleştirilmesi	☑	☐	
Tabloların Yerleştirilmesi	☑	☐	
Kaynaklar	☑	☐	

DANIŞMAN ONAYI Prof.Dr

Unvanı Adı Soyadı: İlker ERCAN

İmza:

İÇİNDEKİLER

TÜRKÇE ÖZET

Büyük veriyi açıklanabilir hale getirmek geçmişte güç olmasından dolayı ve yakın geçmişte ise zaman ve maliyet bakımından kısıtlarının olması nedeniyle, büyük veri çalışmaları yaygın değildi. Günümüzde ise büyük veriyi analiz etmek hem donanımsal hem de yazılımsal gelişmeler ve her gün genişleyen veri havuzuna karşın mümkün hale gelmiştir.

Tez çalışmasında; konjenital hipotiroidizm, hipotroidizm, akut tiroidit tanısı almış olguların laboratuvar ve sosyo-demografik özelliklerine göre büyük veri kullanımı ile analizi yapılması amaçlanmıştır.

Tez çalışmasında, Bursa Uludağ Üniversitesi Sağlık Uygulama ve Araştırma Merkezi Hastanesi'nde, bilgi işlem veri tabanından tarama yapılarak ulaşılabilir olan 2005-2018 yılları arasında belirtilen tanıları olan hastaların, ilk tanıda aldığı değerler dikkate alınarak 21125 hasta analize dahil edilmiştir.

Veri setinde bulunan laboratuvar ölçüm ve demografik değişkenlere göre k-ortalamalar kümeleme metodu iki kümede şekillenecek şekilde analizler yapılmıştır. Büyük veri kullanımına ek olarak Cliff's Delta etki büyüklüğü katsayısı ile kümeler analiz edilmiştir.

Tanı koymada referans alınan Serbest T3 ve Serbest T4 laboratuvar değerleri çalışmamızdaki büyük veri analiz sonuçlarıyla uyumlu çıkarken, TSH laboratuvar ölçüm değerleri uyumsuz çıkmıştır. Büyük veri analizinin sonrasında ortaya çıkan bu farklılıklar, kontrollü çalışmalar ile planlanarak farklılıkların değerlendirilmesi ve araştırılması gerektiğini düşündürmektedir.

Anahtar Kelimeler: büyük veri, hipotiroid, k-ortalamalar, kümeleme analizi

İNGİLİZCE ÖZET

HYPOTHYROIDISM DIAGNOSED CASES USING BIG DATA ANALYSIS WITH k-MEANS CLUSTERING METHOD

In the past, it was difficult to make big data explainable. Until recently, it was not commonly used to cause huge costs and long-term period. Nowadays, it is possible to analyze big data with hardware and software developments and despite the expanding data repository.

Aim of this thesis is to analyze congenital hypothyroidism, hypothyroidism and acute thyroiditis diagnosed cases' laboratory measurements and socio-demographic characteristics using big data.

In this thesis, were included 21125 patients to the study by taking into consideration the first diagnosis' laboratory measurement values which were between 2005 and 2018 accessible defined as above diagnosed cases from the database of Bursa Uludag University Health Application and Research Center.

The data were divided into two clusters according to laboratory measurement and demographic variables. In addition to the usage of big data, clusters were analyzed with the Effect Size of Cliff's Delta.

Reference values to diagnose for laboratory measurements which are Free T3 and Free T4 were compatible with results of the big data analysis in our study, although TSH laboratory values were incompatible. The differences after analyzing big data should be considered to evaluate and investigate with planned and controlled studies.

Keywords: big data, hypothyroid, k-Means Method, cluster analysis

1. GİRİŞ

Dijitalleşme ve teknoloji kullanımının ciddi artış göstermesiyle ve depolanarak oluşan bilgi yoğunluğunun incelenmesi gerekliliğiyle büyük veri çalışmalarına her alanda ihtiyaç duyulmaktadır. Son yıllarda önemli bir kavram haline gelen "Büyük Veri", verilerin anlamlı ve işlenebilir hale getirilmesi sürecidir. İlk kez 2005 yılında, Roger Mougalas tarafından günümüzdeki anlamıyla eşdeğer olarak tanımlamıştır (Sangeetha ve Sreeja, 2015).

Büyük veriyi tanımlama amaçlı (i) hacim, (ii) değer, (iii) hız, (iv) çeşitlilik ve (v) doğrulama olmak üzere beş bileşenden oluşmaktadır (Aksoy ve ark. , 2017). Hacim (Volume), büyük veriye ilişkin boyutu tanımlamaktadır. Verinin boyutunun büyüklüğü ve verinin nasıl depolanacağını belirtir. Değer (Value), veri ile ilgili yapılan çalışmaların sonucunda kurum için katacağı artı veya katma değeri ifade etmektedir. Anlamlandırılamayan ya da sonuçlandırılamayan veri değersiz olarak tanımlanmaktadır. Hız (Velocity), çok hızlı şekilde üretilmeye devam eden veriler benzer şekilde hızlı bir şekilde düzenlenmesi ve işlenmesi gerekliliği belirtilmiştir. Çeşitlilik (Variety), benzer ya da farklı kaynaklardan elde edilen farklı tipte çeşitli veri türlerinden oluşmasını tanımlar. Doğrulama (Verification), geçerli bir kaynağa, gizliliğe ve verilerin güvenliğe sahip olması gerekliliğini belirten tanımlamadır.

Sağlık alanında yapılan çalışmalarında büyük sayıda veri incelenmesi sayesinde, hastaların geçmiş ölçüm değerlerinin incelenmesi veya geleceğe yönelik tahminlerin yapılabilmesinde geleneksel ve modern yöntemler ile hastalıklara ilişkin daha erken tanılar konabilmesi ve tedaviye ilişkin çözümlerin üretilmesi mümkün olabilir.

Yapılan çalışmalarda sağlık alanında büyük veri hakkındaki yayınlar, 2008 yılında ilk ivmelenmeyi göstermektedir. Büyük veri kullanımının, 2012 yılı öncesine

kadar çok tercih edilmese de sonrasında yayımlanan akademik yayınlarda çok ciddi oranda artışın başladığını göstermektedir (Baro ve ark. , 2015 ; Chaorasiya ve Shrivastava, 2015).

Büyük veri çalışmaları, doğası gereği çok büyük sayılara sahip veri setlerinin incelenmesi ile ilgili olduğu için verilerin, çalışma amacına uygun olarak kümelenmesine ihtiyaç duyulmaktadır. Verilerin amaca yönelik kümelenmesi, araştırmada maliyet ve zaman açısından önemli tasarruflar sağlamakla birlikte belli örüntülerin ve yapıların görülmesine de imkân sağlamaktadır. Kümeleme analizi, verileri kümeleme için kullanışlı çok değişkenli analiz tekniğidir. Kümeleme analizinin amacı, kümelenmemiş verilerin benzerliklerine göre sınıflandırmak ve araştırmacıya kümelere ilişkin özetleyici bilgileri elde etmesinde yardımcı olmaktadır.

Verilerin kümelere ayrılmasında seçilecek yöntem için verinin boyutu ve büyüklüğü, oluşturulacak küme sayısı, zaman ve maliyet gibi birçok faktör vardır. k-ortalamalar kümeleme yöntemi, başarılı bir şekilde uygulanabilmesi için yöntemin başında belirlenen optimum sayıda küme sayısına bağlıdır. Hızlı ve büyük veriler için uygun bir yöntem olan k-ortalamalar kümeleme yöntemi, diğer kümeleme yöntemlerine benzer olarak, küme içi elemanların benzerliği yüksek, kümeler arası benzerlik ise düşük olması amaçlanmaktadır (Han ve ark. , 2012).

Sağlık alanında da büyük veri kullanılarak analizlere gereksinim duyulduğu görülmektedir. Büyük veri ile ilgili literatür incelendiğinde, sağlık alanında yapılan çalışmaların önem kazandığı görülmektedir (Grasso ve ark. , 2015; Jeong ve ark. , 2018; Kumari, 2018). Tez çalışmasında tiroid fonksiyon bozukluklarına ilişkin hastalıkların incelenmesi günden güne ülkemizde ve dünyada artarak devam etmektedir. Prevelans hızı %30 olarak yaygın olarak görülen hastalıklar arasındadır (Serin ve ark. , 2006). Tez çalışmasında da; konjenital hipotiroidizm, hipotiroidizm, akut tiroidit tanısı almış olguların laboratuvar ve sosyo-demografik özelliklerine göre büyük veri kullanımı ile analizi yapılması amaçlandı.

2. GENEL BİLGİLER

Büyük veri, veri hacmi büyük olan ve büyüklüğü nedeniyle de geleneksel yöntemlerle depolanması mümkün olmayan veri setlerini işleme düzenleme süreci olarak tanımlanmaktadır. Büyük veri tanımlamaları için metodolojide farklı şekilde açıklamalar ve tanımlamalar görülmektedir. Dumbill (2013) büyük veri tanımını, geleneksel veri tabanı sistemlerinin işleme kapasitesini aşan veri olarak yapmıştır. Ward ve Barker'a (2013) göre büyük veri, geleneksel araçlar ve proses teknikleri ile analiz edilemeyen veriler olarak belirtmiştir. Fisher ve arkadaşlarına (2012) göre büyük veri, baş edilemez ve işlenemez veriler olarak tanımlamıştır. Manyika ve ark. (2011) ise büyük veriyi tipik veri tabanı yazılım araçlarının depolama, yönetme ve analiz etme yeteneklerinin ötesinde olan veri setleri olarak tanımlamıştır.

Sürekli artan veri, zamanla bilgi yoğunluğu oluşturmakta ve kontrol edilmesi güç bir hal almaktadır. Büyük veri yöntemleri, benzeri durumlarda artarak devam eden veri akışını iyileştirme, amaca yönelik araştırmaya erişmede en az maliyetle ve kısa sürede erişimi sağlamaktadır.

Her alanda artış olduğu gibi sağlık alanında da benzer şekilde veri sayısı zamanla çok hızlı bir şekilde artmaktadır. Amerika'da yalnızca sağlık alanında 2011 yılında verinin 150 exabayta ulaştığı ve çok yakın zamanlarda da zettabayta ulaşılacağı belirtilmiştir (Raghupathi ve Raghupathi, 2014).

Sağlık hizmetlerinde birçok farklı kaynaktan çok büyük sayılarda veri üretilmektedir. Laboratuvar ölçümleri, metin ve görsel formatta olmak üzere klinik veriler, uygulama rehberleri olmak üzere klinik referanslar, yayınlar, genomik veriler gibi veri kaynakları ile veri tabanları çeşitlenmektedir.

Çalışmanın amacına yönelik karmaşık bir yapıya sahip olan büyük veriyi düzenli hale getirmek için kümeleme yöntemlerine ihtiyaç duyulur. Kümeleme yöntemleri verideki birimlerin benzerliklerine göre kümelerin oluşturulması amaçlanır.

Kümeleme yöntemlerinin aşamaları yöntem fark etmeksizin aşağıdaki şekilde sıralanabilir (Tatlıdil, 1992).

i. Kümeleme yöntemi uygulanacak veri seti için çalışmanın belirlenmesi,
ii. Değişkenlerin tanımlanması ve ölçümlerin uygulanması,
iii. Veri setine ilişkin benzerlik ölçümlerinin belirlenmesi,
iv. Çalışma için uygun kümeleme tekniğinin belirlenmesi,
v. Oluşan kümelere ilişkin anlamlılığın tartışılması

Kümeleme analizi uygulanırken dikkat edilmesi gereken bazı önemli hususlar vardır (Aldenderfer ve Blashfield, 1989).

i. Kümeleme yöntemlerinin çoğu basit işlemlere sahip olup genellikle de istatistiksel yönteme dayandırılmayabilmektedir.
ii. Kümeleme yöntemi uygulanırken farklı disiplin veya alanlardan türetilmiş yöntemler olabilmektedir. Kümeleme işlemi gerçekleştirilirken bu durum dikkate alınmalıdır.
iii. Çalışılacak veri setine farklı kümeleme yöntemleri uygulandığında farklı yorumlamalar ortaya çıkabilmektedir.

Kümeleme yöntemlerinin amacı; küme içi homojen, kümeler arası ise heterojen kümeler oluşturmaktır. Kümeleme yöntemleri, Hiyerarşik ve Hiyerarşik Olmayan Kümeleme yöntemleri olmak üzere ikiye ayrılır. Hiyerarşik Kümeleme yöntemlerinde küme sayısının belirlenmesi analiz sonuçları incelenilerek belirlenmektedir. Hiyerarşik olmayan kümeleme yöntemlerinde ise küme sayısı önceden belirlenir. Hiyerarşik olmayan kümeleme yöntemleri ayrıca hiyerarşik kümeleme yöntemlerine göre daha büyük veri setlerine uygulanabilir (Johnson ve Wichern, 1988). Hiyerarşik olmayan kümeleme yöntemlerinde, hiyerarşik kümeleme yöntemlerine göre küme sayısını belirleyebilmek için ön bilgilere ihtiyaç duyulmaktadır. Hiyerarşik olmayan kümeleme yöntemlerinde, değişkenler yerine

sadece birimleri dikkate alarak k adet küme oluşturulması amaçlanmaktadır. Hiyerarşik olmayan kümeleme yöntemlerinden k-ortalamalar kümeleme algoritması, k-Medoids kümeleme algoritması (PAM), CLARA ve CLARANS ayrıca büyük veride sınıflandırmada kullanılan Naive Bayesian Sınıflandırıcısı ve İkili Lojistik Regresyon Analizinden de bahsedilecektir.

2.1 Bölümlemeli Kümeleme Yöntemleri

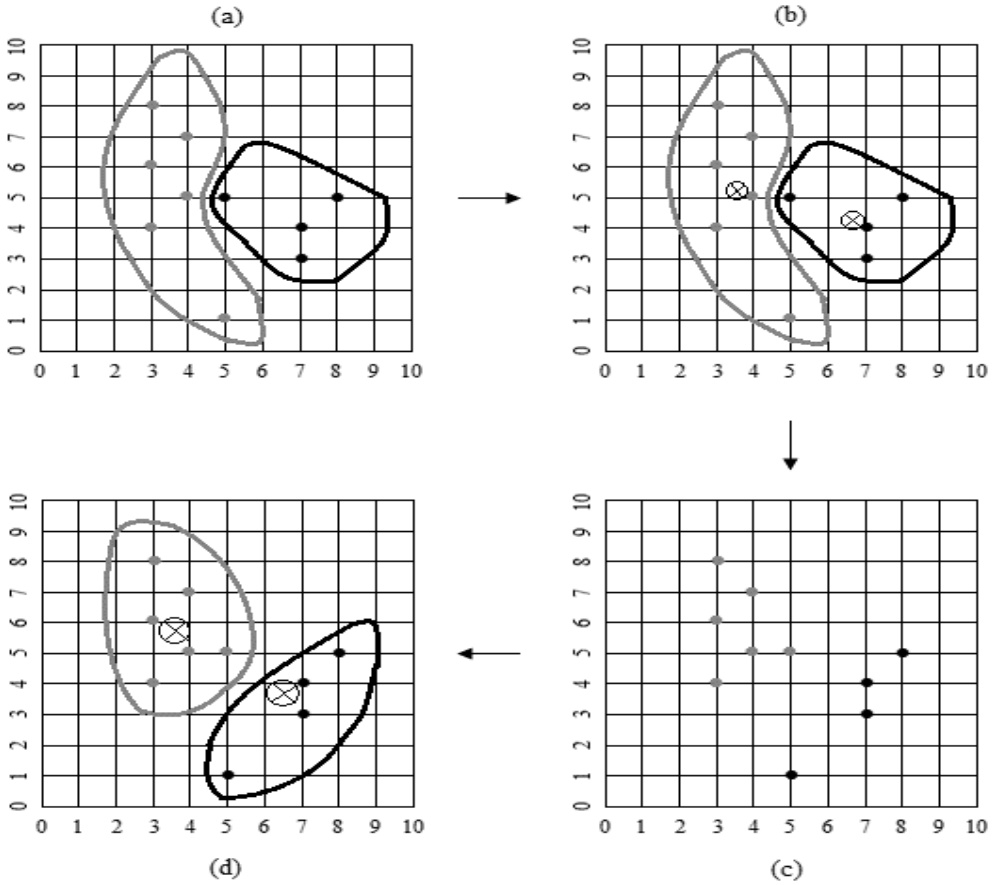
n adet gözlemden oluşan veri setini, çalışmanın başında belirlenen k adet kümeye bölmesi amaçlanır. Bölümlemeli kümeleme yöntemlerinde merkez noktası, kümenin özelliklerini yansıtmaktadır. Veri setine uygulanması kolay ve hızlı bir şekilde küme çözümlemesine ulaşılmaktadır.

2.1.1 k-ortalamalar Kümeleme Yöntemi

k-ortalamalar kümeleme yöntemi, J.B. MacQueen tarafından 1967 yılında geliştirilen, kümelerin merkezlerinin yardımıyla verileri karakterize eden hiyerarşik olmayan bölümlemeli (Partitional) kümeleme algoritmasıdır (MacQueen, 1967). k-ortalamalar kümeleme yöntemiyle veri yalnızca bir kümeye ait olmasından dolayı keskin bir kümeleme algoritmasıdır. k-ortalamalar kümeleme yöntemi sürekli verilerde uygulanır (Berkhin, 2006).

k-ortalamalar yönteminde n adet gözlemin her biri en yakın ortalamaya sahip k adet kümeye bölünür. Merkez noktanın kümeyi yansıtması hedeflenir. k-ortalamalar kümeleme yöntemi, diğer kümeleme yöntemlerine benzer olarak, küme içi elemanların benzerliği yüksek, kümeler arası benzerlik ise düşük olması amaçlanır (Han ve ark. , 2012).

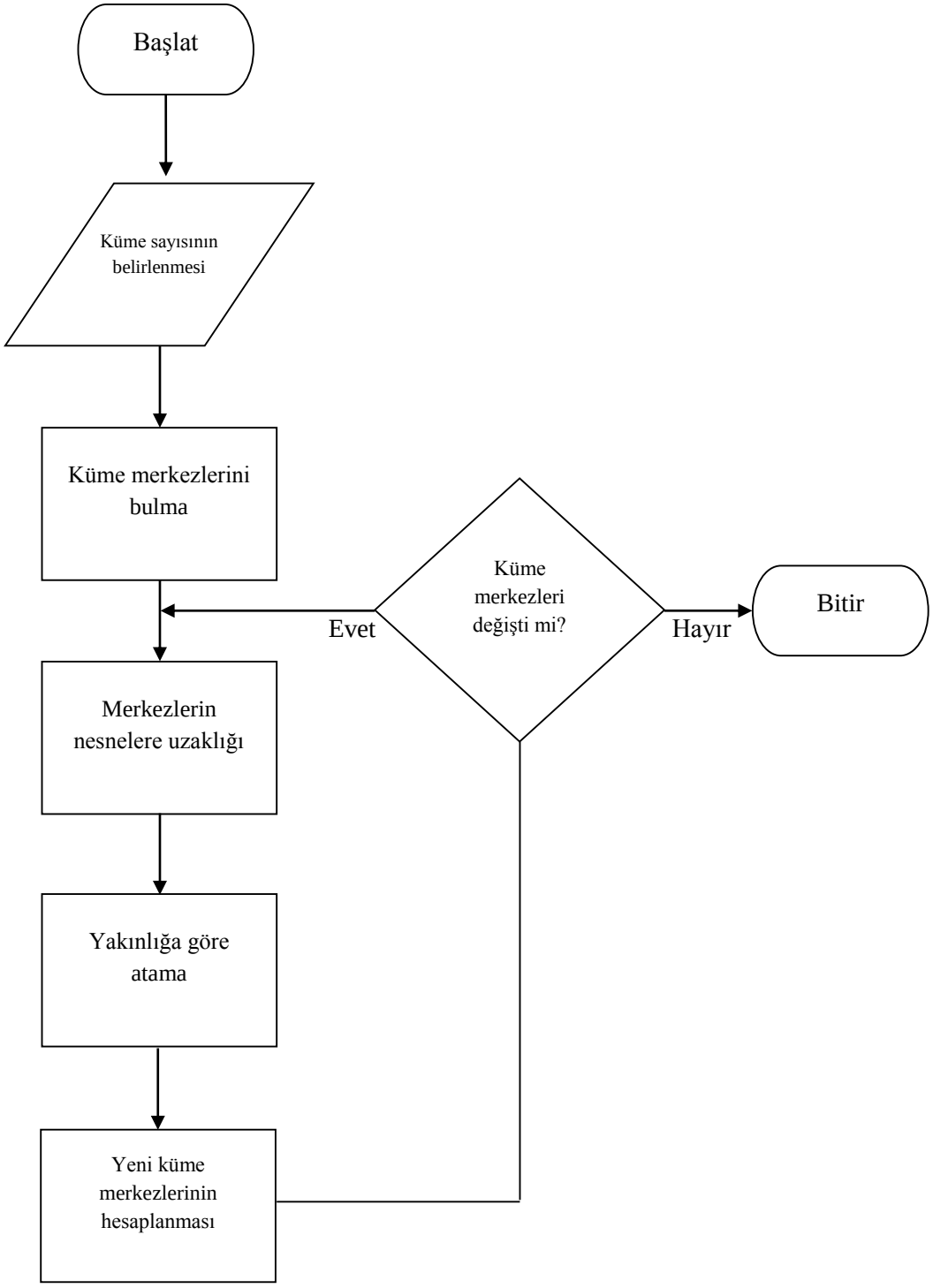
k-ortalamalar kümeleme yönteminin adımları Şekil-2.1'de gösterilmiştir.



Şekil-2.1 k-ortalamalar Kümeleme Yöntemi Adımları (Sarıman, 2011)

Şekil-2.1 incelendiğinde; (a) ilk aşamada, küme merkezlerini oluşturacak kadar küme sayısı belirlenir. (b) İkinci ve (c) üçüncü aşamada ise küme merkezleri belirlendikten sonra her bir nokta küme merkezlerine bağlanır. Dördüncü aşamada (d), oluşan kümelerin merkezleri yeniden hesaplanarak yeni hesaplanan merkezler kümelerde konumlanır. Noktalar için yeniden küme merkezlerine uzaklıkları yeniden hesaplanır. Bu döngü, noktaların ait olduğu kümelerin merkezleri değişmediği duruma gelene kadar tekrarlanır. Küme merkezlerinin değişmediği durumda ise her bir nokta, bir kümeye ait olur. Yeni küme belirlendikten sonra eski merkez nokta sıradan küme elemanı olur (Berry ve Linoff, 2004).

k-ortalamalar kümeleme yönteminin akış diyagramına aktarılmış hali Şekil 2.2’ de gösterilmiştir.



Şekil-2.2 k-ortalamalar kümeleme yöntemine ilişkin akış diyagramı

Kümeleme süresince bölümlemeli kümeleme yöntemlerine özgü olarak hata kareler kriteri kullanılır. Varsayalım ki, $x_j \in \Re^d$, j=1,…n olsun ve veri setinin K adet kümeye bölünmesi amaçlansın. C={C_1,…,C_k). C_i'ler kümeleri ifade etmektedir. Hata kareler kriteri aşağıdaki şekilde tanımlanır.

$$J = (\Gamma, M) = \sum_{i=1}^{K} \sum_{j=1}^{N} \gamma_{ij} ||x_j - m_i||^2 \qquad (2.1)$$

Burada,

Γ= Bölümlenmiş matris

γ_{ij} , Parçalı fonksiyonu;

$$\gamma_{ij} = \begin{cases} 1, & eğer\ x_j \in \text{i. küme} \\ 0, & d.d. \end{cases} \quad ,\sum_{i=1}^{K} \gamma_{ij} = 1\ \forall j \qquad (2.2)$$

M= Merkezi matris, [m_1,…,m_k],

m_i= i. kümenin örneklem ortalaması $(\frac{1}{N_i}) \sum_{j=1}^{N} \gamma_{ij}\, x_j$

N_i = i. kümeye ait eleman sayısı

Hata kareler kriteri, k adet küme oluşturulurken hata kareler azaltılarak belirlenir. Kümenin benzerliği, kümede yer alan verilerin ortalama değeri ile ölçülüp, elde edilen değer kümenin ağırlık merkezini oluşturur (Xu ve Wunsch, 2005).

Gerçekleştirilmesi kolay ve karmaşıklığı diğer yöntemlere göre az olması, k-ortalamalar kümeleme yönteminin avantajları olarak belirtilebilir. Bölümleyici kümeleme yöntemlerinin ortak sorunları olarak k-ortalamalar kümeleme yönteminde de ilk aşamada oluşturulacak küme sayısını belirlemenin zorluğu ve belirtilen döngünün hangi aşamada duracağının çalışmanın başlangıcında bilinmemesi sorunlar oluşturabilmektedir (Han ve ark. , 2012 ; Pena ve ark. , 1999).

Bölümleyici kümeleme yöntemleri k sayısı ne olursa olsun geçerli veya geçersiz bir kümeleme sonucu verecektir. Doğru ve başarılı küme işlemi k'nın optimum sayıda olması ile ilişkilidir. Sonuç olarak belirtilen sebeplerden ötürü k'nın isabetli şekilde seçimi, değişkenler arttıkça daha da zorlaşabileceği için bölümleyici kümeleme yöntemlerinin önemli dezavantajları olarak belirtilmektedir (Ray ve Turi, 1999).

2.1.2 k-Medoids Kümeleme Yöntemi

PAM (Partitioning Around Medoids) olarak da bilinen k-Medoids kümeleme yöntemi, veriye ait öznitelikleri dikkate alarak k-ortalamalar kümeleme yöntemine benzer olarak küme sayısının başlangıçta belirlenmesini ve veriye uygun küme sayısı için denemeler yapılmasını gerektiren bir metottur. Kaufman ve Rousseeuw tarafından 1987 yılında geliştirilen k-Medoids kümeleme yöntemi, küme içinde temsilci nesne benzerliğinin yüksek, kümeler arasındaki temsilci nesnelere ilişkin benzerliği düşük olduğu kümeler bulmaktır. "medoid" kümenin merkezine yakın konumlanmış nokta olarak ifade edilir (Kaufman ve Rousseeuw, 1990). k-Medoids algoritmasında her yeni nesne kümeye katıldığında kümenin gelişmesine en fazla yarar sağlayacak nokta belirlenince bulunan nokta yeni merkez, eski merkez ise küme elemanı olacak şekilde yer değiştirme (swap) işlemi yapar.

Bölümleme metotlarının temel amacı farklı nesnenin arasındaki uzaklığı minimize eder. k-medoids yönteminde mutlak hata kriteri kullanılmaktadır. Uzaklık olarak Öklid ve Manhattan uzaklıkları k-medoids yönteminde kullanılabilmektedir (Hasan ve ark. , 2015).

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, o_i) \qquad (2.3)$$

Burada,

E= Veri setinde bütün nesnelerin mutlak hata toplamı

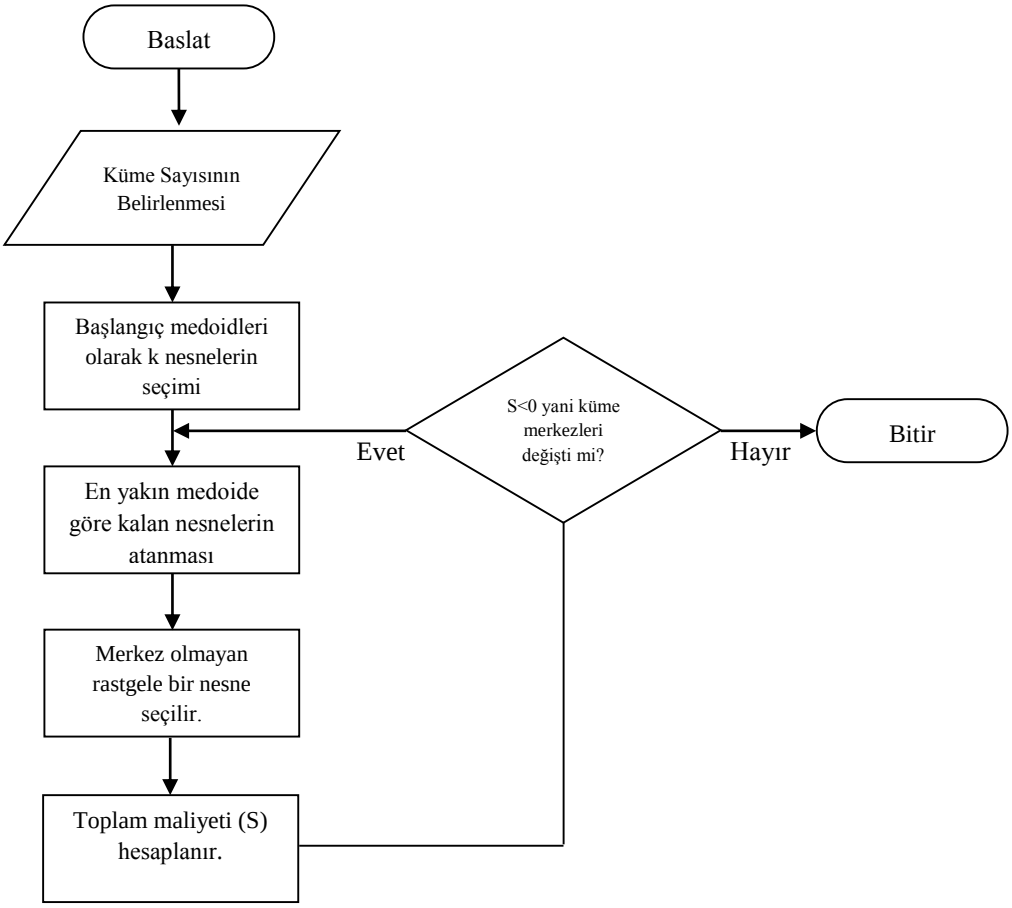
C_i= i. küme sayısı

p= C_i kümesindeki nesneye ait nokta

o_i= C_i kümesinin orta noktası

k- Medoids kümeleme yönteminde öncelikle çalışmanın amacına yönelik veri setinin bölüneceği küme sayısı k belirlenip, k adet temsilci nesneler belirtilen kümelere atanır. Kalan nesneler de en yakın kümeye atanır. Merkez olmayan (Medoid olmayan) rastgele bir nesne seçilir. ($o_{rastgele}$) Yer değiştiren temsilciler (o_j) ile $o_{rastgele}$'nin toplam maliyeti (S) hesaplanır. Eğer $S < 0$ ise o_j ile $o_{rastgele}$ yer değiştirir. Bu döngü, kümedeki nesneler değişmediği duruma kadar devam eder (Han ve ark. 2012).

k-Medoids kümeleme yöntemi, akış diyagramı aşağıdaki şekilde ifade edilebilir.



Şekil-2.3 k-Medoids kümeleme yöntemi akış diyagramı

k-Medoids kümeleme yöntemi çok büyük veri ve çok boyutlu kümelerde zaman ve ciddi bellek sorunlarına sahip olmaktadır (Işık ve Çamurcu, 2007).

2.1.3 CLARA Kümeleme Algoritması (Clustering Large Applications)

CLARA kümeleme algoritması, k-medoids kümeleme algoritmasında yaşanan depolama ve zaman sorunlarından dolayı Kaufman ve Rousseeuew tarafından k-medoids kümeleme temelli geliştirdiği algoritmadır (Kaufman ve Rousseeuw, 1990).

CLARA kümeleme algoritması, veri setinin tamamını tarayarak kümeye ilişkin temsilci noktalar yerine, rastgele bir örnek küme üzerinde PAM algoritmasını uygular. PAM algoritması uygulandıktan sonra her bir kümeye ilişkin temsilci atanır. Kümeyi oluşturan veri setinden bir örnek küme daha belirlenir. Temsilcilerin rastgele seçilmesi yerine bir önceki durumda belirlenen temsilciler kullanılır. Bu durum, hem temsilci değişimini azaltacak hem de hızlı bir şekilde sonuçlara ulaşılmasını sağlayacaktır.

Küme seçim işlemlerinin 5 defa yinelenmesi gerektiği ve her zaman 40 + 2k adet örnek seçilmesinin en iyi sonucu verdiği Kaufman ve Rousseeuew (1990) tarafından belirtilmiştir.

2.1.4 CLARANS Kümeleme Algoritması (Clustering Large Applications based on Randomized Search)

Raymond ve Jiawei tarafından geliştirilen CLARANS Kümeleme Algoritması, PAM ve CLARA algoritmalarının geliştirilmiş versiyonu olup ayrıca PAM ve CLARA'ya göre etkinliği ve verimliği daha yüksektir (Raymond ve Han, 1994).

CLARANS kümeleme algoritması, n adet gözleme sahip olan veri seti için temsilciler aracılığıyla diyagram ve alt veri yapıları (Subgraph) yardımı ile verinin k

adet kümeye bölünmesini sağlamaktadır. Diyagram, $G_{n,k}$ ile ifade edilir. Diyagramdaki düğümler $\{o_{m1}, o_{m2}, \dots, o_{mk}\}$ olarak k adet temsilciyi ifade eder. İki düğümün kümeleri eğer yalnızca bir nesneye göre farklılık gösteriyorsa bu iki düğüm komşu olarak ifade edilir. İki düğümün $S_1 = \{O_{m1}, O_{m2}, \dots, O_{mk}\}$ ve $S_1 = \{o_{w1}, o_{w2}, \dots, o_{wk}\}$ komşu olabilmesi için $|S_1 \cap S_2| = k - 1$ olması gerekmektedir. Bahsedilen durumun sonucu olarak her bir düğüm $k(k-1)$ tane komşuya sahiptir. Bir düğüm k adet medoid ya da temsilci içerdiği için her bir düğüm bir kümelemeyi tanımlar. Eğer O_i, O_h temsilcileri iki S_1 ve S_2 komşuluklarının kesişimlerinin farklılıkları olduğu durumda $(O_i, O_h \notin S_1 \cap S_2)$, iki komşuluğun arasındaki toplam maliyet değişimi TC_{ih} aşağıda belirtilen (2.4) formülü ile hesaplanabilir (Raymond ve Han, 1994).

$$TC_{ih} = \sum_j C_{jih} \quad , \qquad (2.4)$$

k-Medoids kümeleme algoritmasında her adımda incelenen düğümün tüm komşularının taranması gerekmektedir. Bu döngü, benzer şekilde en uygun kümeler bulunacak ana kadar devam etmesi gerektiğinden $n = 1000$ ve $k = 10$ veya daha yüksek değeri için $k(n-k)$ adet bir düğümün komşulukları incelenmesi gerektiğinden zaman kaybına neden olmaktadır. k-Medoids kümeleme yöntemi CLARA kümeleme yöntemine göre büyük veri setleri için etkinliği ve verimliliğini, hem zaman hem de maliyet olarak arttırmaktadır (Raymond ve Han, 1994).

2.2 Büyük Veri Analizinde Kullanılan Diğer Sınıflandırıcılar

Büyük veriyi sınıflandırma amaçlı istatistiksel yöntemler olarak Naive Bayesian Sınıflandırması ve İkili Lojistik Regresyon analizi de kullanılmaktadır.

2.2.1 Naive Bayesian Sınıflandırması

İstatistiksel bir metot olan Naive Bayesian Sınıflandırmasının temeli, Bayes'in teoremine dayanan ve bir verinin bir sınıfa ait olma olasılığı tahmin eden sınıflandırma yöntemi olarak kullanılmaktadır. (Han ve ark. , 2012 ; Kılınç ve ark. , 2016 ; Olgun ve Özdemir, 2012).

Naive Bayesian yöntemi, sınıf-koşullu bağımsızlık (class-conditional independence) varsayımı olarak tanımlanan, x öznitelik değerinin belirli bir sınıf üzerindeki etkisinin diğer x öznitelik değerlerinden bağımsız olduğunu belirtir. Varsayımlara ilişkin hesaplamaları basitleştirdiğinden dolayı basit (naive) tanımını almıştır (Han ve ark. , 2012).

Bayes teoremi temelli olduğu için $P(C_i|X)$ koşullu olasılığı maksimize edilir.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \qquad (2.5)$$

C_i: Verilere ilişkin sınıflar i=1,2,…,n

X: (x_1, x_2,…,x_n) vektörü

Eğer tüm sınıflar için P(X) sabit bir değer ise, maksimize edilen koşullu olasılık,

$$P(C_i|X) = P(X|C_i)P(C_i) \qquad (2.6)$$

eşitliğidir (Han ve ark. , 2012).

Naive Bayes sınıflandırıcısı için optimum olasılıkla hangi sınıfa ait ise karar sınıfı olarak belirlenir. Çalışmalarda, girdi değişkenlerinin tamamının birbirinden bağımsız olması gerçek bir veride çoğu zaman mümkün olmayışı Naive Bayes sınıflandırıcısının zayıf yönü olarak gösterilebilir.

2.2.2 İkili Lojistik Regresyon

Bir ya da daha fazla bağımsız değişken ile ikili cevap değişken arasındaki bağıntıyı açıklayan ikili lojistik regresyon analizi aynı zamanda verilerin sınıflandırılması için veri madenciliği ve makine öğreniminde de kullanılmaktadır (Hosmer ve Lemeshow, 2000 ; Musa, 2014).

Lojistik regresyon modeli,

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}} \qquad (2.7)$$

şeklinde tanımlanır (Hosmer ve Lemeshow, 2000).

Burada,

$\pi(x) = P(Y = 1|x)$: İncelenen duruma eşit olma olasılığı

β_0: Denklem sabiti

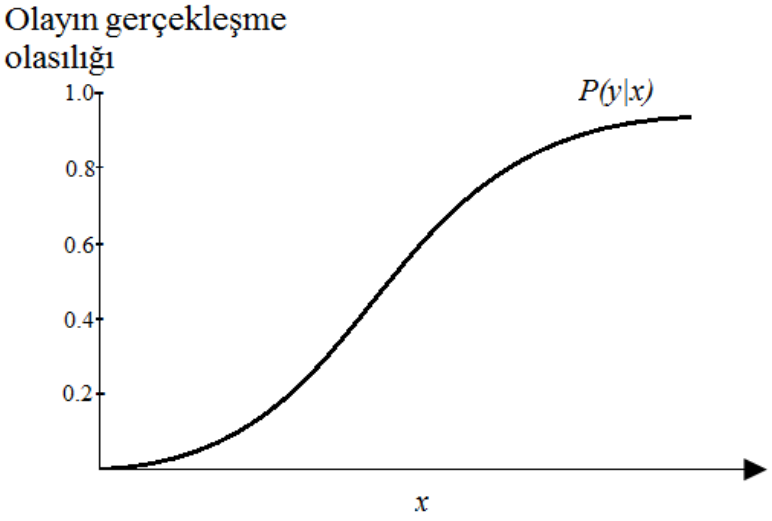
β_i: Bağımsız değişkenlere ait katsayılar $i = 1,2,\ldots,p$

x_i: Bağımsız değişkenler $i = 1,2,\ldots,p$

Lojistik regresyon analizinde, bağımlı değişkenin 1 değerine eşit olması olasılığı (İncelenen durum 1 olması) araştırılmaktadır. Hesaplanan sonuç olasılığa ilişkin değer olduğu için 0 ve 1 arasında değer alır (Alpar, 2011).

2.2.2.1 Lojit Model

Olasılık π ile x bağımsız değişkenleri arasındaki ilişki, sıklıkla lojistik yanıt fonksiyonu ile gösterilebilir.



Şekil-2.4 Lojit Model

Şekil-2.4 incelendiğinde, başlangıçta x değerlerinin artışı ile olasılıkta artar, sonra artış hızlanır ve sonunda da durağanlaşır. Koordinat düzlemi için y ekseni bir olayın olma olasılığını tanımladığı için hiçbir zaman 1'in üzerine çıkamaz.

Lojit model;

$$\pi_i = P(Y_i = 1) \quad , \quad i = 1,2,\dots,n \qquad (2.8)$$

$$P(Y = 1) = \frac{e^{\beta_0+\beta_1 X_1}}{1 + e^{\beta_0+\beta_1 X_1}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1)}} \qquad (2.9)$$

olarak tanımlandığından, eşitliğin her iki tarafının ln alınması durumunda odds oranının doğal logaritması olacaktır.

$$lojit\ \pi(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \qquad (2.10)$$

$$\frac{\pi(x)}{1-\pi(x)} = e^{\beta_0+\beta_1 X_1} \qquad (2.11)$$

Odds oranının doğal logaritması alındığında ise doğrusal modele ulaşılır.

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \ln\left(e^{\beta_0+\beta_1 X_1}\right) = \beta_0 + \beta_1 X_1 \qquad (2.12)$$

Doğrusal regresyonda artıklar 0 ortalama ve sabit varyansla normal dağılırken, lojistik regresyonda artıklar sıfır ortalama ve π(1- π) varyansla binom dağılır (Alpar, 2011).

2.2.2.2 Lojit Modellerde Parametre Tahmini

Regresyon modellerinde katsayıların tahmin edilmesi gerekmektedir. Bilinmeyen katsayıların tahmininde kullanılan birden çok yöntem vardır.

2.2.2.2.1 En Çok Olabilirlik Yöntemi

Doğrusal regresyonda bilinmeyen parametrelere ilişkin katsayı tahmininde hata kareler yöntemi kullanılır. Hata kareler yöntemi, çalışmaya ilişkin bağımlı değişkenin gözlenen ve beklenen değerleri arasındaki farkın standart sapmaların karelerinin toplamını minimize eden β katsayılarını seçmektir. Bağımlı değişken, ikili (binary) cevap içeren değişken olduğunda hata kareler yöntemi uygulandıktan sonra kestiriciler aynı özelliklere sahip değildir. Lojistik regresyon analizinde belirtilen sebepten dolayı en çok olabilirlik yöntemi kullanılmaktadır (Hosmer ve Lemeshow, 2000).

En çok olabilirlik yöntemi uygulanabilmesi için öncelikle olabilirlik fonksiyonunun oluşturulması gerekmektedir. Oluşan bu fonksiyon, parametrelere ilişkin değerlerin olabilirliğini belirtmektedir. En çok olabilirlik tahmin edicileri ise olabilirlik fonksiyonunu maksimize ederek gözlenen veri setine en yakın tahminler elde edilebilmektedir (Hosmer ve Lemeshow, 2000).

İkili cevap içeren bağımlı değişken için, lojistik regresyonda 1 değerine koşullu olarak eşit olması olasılığı $P(Y = 1|x)$ (İncelenen durum 1 olması)

incelenmektedir. $1-\pi_x$ ise 0 değerine koşullu olarak eşit olma olasılığını $(Y=1|x)$ verecektir. Bağımlı ve bağımsız değişkene ilişkin olabilirlik fonksiyonu,

$$l(\beta)=\prod_{i=1}^{n}\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \qquad (2.13)$$

En çok olabilirlik yönteminin temeli (2.13)'deki eşitliğindeki β'nın tahmininin maksimize edilmesidir. Eşitliğe (2.13) ilişkin hesaplamaların zorluğundan logaritması ile çalışmak daha kolay olacaktır. Logaritması alınan yeni eşitlik log olabilirlik olarak tanımlanacaktır (Hosmer ve Lemeshow, 2000).

$$l(\beta)=ln[l(\beta)]=\sum_{i=1}^{n}\{y_i ln[\pi(x_i)]+(1-y_i)ln[1-\pi(x_i)]\} \qquad (2.14)$$

$l(\beta)$ değerini maksimum yapan katsayıları bulabilmek için $l(\beta)$'nın β'lara göre türevleri alınıp 0'a eşitlenir.

$$\sum[y_i-\pi(x_i)]=0 \qquad (2.15)$$

$$\sum x_i[y_i-\pi(x_i)]=0 \qquad (2.16)$$

Olabilirlik fonksiyonuna ilişkin denklemler (2.15) ve (2.16)'de gösterilmiştir. Denklemler doğrusal olmadığı için çözümünde özel metotlar gerektirir. Denklemlerin sonucunda elde edilen β değerlerine, en çok olabilirlik tahmin edicisi adı verilip $\hat{\beta}$ ile gösterilir. $\hat{\pi}(x_i)$, $\pi(x_i)$'nin en çok olabilirlik tahmini olarak gösterilir. Belirtilen

durum bağımlı değişkenin 1 değerine koşullu olarak eşit olmasının olasılığının tahmininde yardımcı olur. Regresyondan katsayılar, tahmin edilen ya da uyum gösteren (fitted) değeri tahmin etmede denklem (2.15) ile hesaplanır (Hosmer ve Lemeshow, 2000).

Sonuç olarak modelin uyum iyiliği de,

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i) \qquad (2.17)$$

denklem (2.17)'de görüleceği üzere, gözlenen değerlerin toplamının beklenen tahminlerin toplamına eşit olması olarak ifade edilir.

3. GEREÇ VE YÖNTEM

Çalışmada kullanılan veriler, Bursa Uludağ Üniversitesi Sağlık Uygulama ve Araştırma Merkezi Hastanesi'nde, bilgi işlem SQL veri tabanından tarama yapılarak elde edilmiştir. Çalışmanın uygulanabilmesi için hastalara ilişkin cinsiyet, yaş, hastaneye başvuru tarihi, ICD kodu verilerinin kullanım izni Bursa Uludağ Üniversitesi Etik Kurulu'ndan alınmıştır. (Etik Kurul No: 09/2018-05)

Veri tabanında ulaşılabilir olan 2005-2018 yılları arasında hipotiroidi hastalarına yönelik ICD kodları tanımlamasına göre E03: "Konjenital Hipotiroidizm, diğer", E03.9:" Hipotroidizm, tanımlanmamış" ve E06:" Akut Tiroidit" tanılarına sahip 173418 hasta çalışmaya dahil edilmiştir.

Tiroid fonksiyonlarının değerlendirilme aşamasında en iyi tanıyı TSH ölçümü koymakla birlikte; Serbest T3, Serbest T4 hormonlarının kan düzeylerinin ölçülmesiyle de karar verilmesi önerilmektedir. (Türkiye Endokrinoloji Metabolizma Derneği, Nisan 2019) Bu nedenle tez çalışmasında ilgilendiğimiz tanı kapsamında herhangi bir tedavi almamış ilk tanıdaki TSH, Serbest T3, Serbest T4 laboratuvar değerleri analize alınmıştır. Laboratuvar verilerine ek olarak, demografik verilerden yaş ve cinsiyet değişkenleri veri setine dahil edilmiştir. Laboratuvar ölçüm değerlerinden en az biri eksik olan hastalar çıkartılarak, verisi tam olan 21125 hasta çalışmaya alınmıştır.

Veriler "erkek" (n=5220), "kadın"(n=15905) ve "genel"(n=21125) olarak ve "18 yaş altı erkek ve kadın" (n_e=1511 , n_k=1918) , "18 yaş üstü erkek ve kadın" (n_e=3709 , n_k=13987) ve "18 yaş altı ve üstü genel" ($n_{<18}$=3429, $n_{\geq 18}$=17696) olarak büyük veri kapsamında değerlendirilmiştir.

k-ortalamalar kümeleme metodu yardımı ile yaş, Serbest T3, Serbest T4 ve TSH değerlerine bağlı olarak iki ayrı kümeye ayrılacak şekilde çözümleme yapılmıştır.

Oluşan kümelerdeki verilerin normal dağılıma uygunluğu Shapiro-Wilks testi ile test edilmiştir. Kümelerin oluşmasında etkili olan değişkenler etki büyüklüğü analizleri yapılarak araştırılmıştır. Verilerin normal dağılmamasından dolayı etki büyüklüğü olarak Cliff's Delta etki büyüklüğü katsayıları ve güven aralıkları hesaplandı (Cliff, 1996 ; Hogarty ve Kromrey, 1999).

$$\delta = \mathrm{P}(x_{i1} > x_{j2}) - \mathrm{P}(x_{i1} < x_{j2}) \qquad (3.1)$$

Burada,

x_{i1}: 1. Gruptaki i. skor değeri

x_{j2}: 2. Gruptaki j. skor değeri

Cliff Delta etki büyüklüğüne ilişkin güven aralıklarına ilişkin değerler dominant matris yardımıyla hesaplanmaktadır. (Cliff, 1996 ; Hogarty Kromrey, 1999).

Dominant matris satır ve sütun etiketleri sırasıyla birinci ve ikinci gruba ait elemanlarından oluşmaktadır. Satır elemanı etiketi sütun elemanı etiketinden büyük olduğu durumda -1, tersi durumda 1 ve eşit olduğu durumda ise 0 değerini alır. Marjinal değerler ise oluşan matrisin satır ya da sütun elemanlarının toplamlarının ilgili gruptaki gözlem sayısına bölünmesi ile elde edilir.

Etki büyüklüğü, δ'nın varyansının tutarlı tahmin edicisi,

$$S_{\delta \mathrm{t}} = \frac{(n_2 - 1){S_{\delta \mathrm{i.}}}^2 + (n_1 - 1){S_{\delta \mathrm{.j}}}^2 + {S_{\delta \mathrm{ij}}}^2}{n_1 n_2} \qquad (3.2)$$

ile hesaplanır.

Burada,

$${S_{\delta \mathrm{i.}}}^2 = \sum (\delta_{i.} - \delta)^2/(n_1 - 1) \qquad (3.3)$$

$$S_{\delta.j}{}^2 = \sum(\delta_{.j} - \delta)^2/(n_2 - 1) \qquad (3.4)$$

$$S_{\delta ij}{}^2 = \sum\sum(\delta_{ij} - \delta)^2/[(n_1 - 1)(n_2 - 1)] \qquad (3.5)$$

δ katsayısına ilişkin güven aralıkları formül (3.2) yardımıyla, formül (3.6) ile hesaplanmaktadır.

$$\frac{\delta - \delta^3 \pm Z_{\propto/2} S_{\delta t}[(1-\delta^2)^2 + Z_{\propto/2}{}^2 S_{\delta t}{}^2]^{1/2}}{1 - \delta^2 + Z_{\propto/2}{}^2 S_{\delta t}{}^2} \qquad (3.6)$$

burada $Z_{\propto/2}$, normal dağılımın $(1-\propto/2)$. yüzdeliğe karşılık gelen z değeridir.

Cliff Delta etki büyüklüğü değerlerinin yorumlanmasında Tablo 3.1'de belirtilen aralıklar dikkate alınmıştır (Romano ve ark. , 2006).

Tablo- 3.1 Cliff delta etki büyüklüğünün yorumlanması

Etki Büyüklüğü	**Yorumlama**
$\lvert\delta\rvert$ ***<0.147***	Önemsiz (Negligible)
$\lvert\delta\rvert$ ***<0.330***	Küçük (Small)
$\lvert\delta\rvert$ ***<0.474***	Orta (Medium)
$\lvert\delta\rvert$ ***≥0.474***	Büyük (Large)

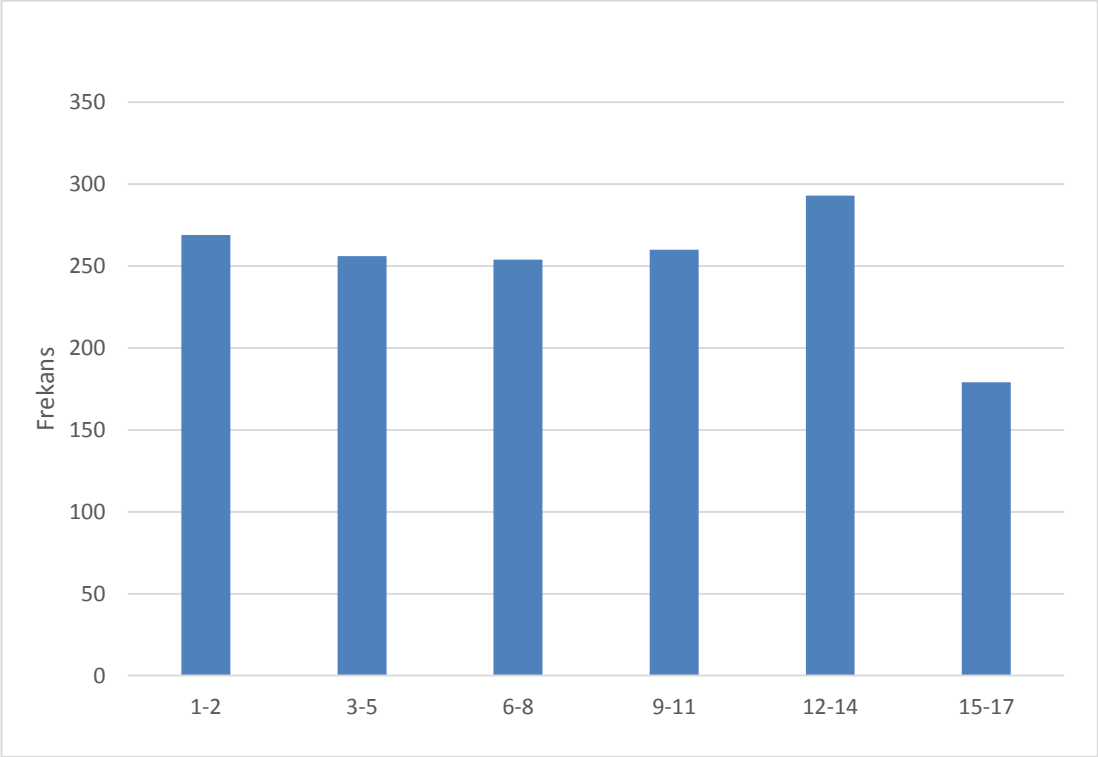
Verilerin, veri tabanından çekilmesi aşamasında SQL, düzenlenmesi aşamasında Excel, analizlerin gerçekleştirilmesi aşamasında ise SPSS 22, R paket programı 3.5.2 ile “effsize” kütüphanesi kullanılmıştır.

4. BULGULAR

Çalışmaya dahil olan hastaların %24,7'si erkek (n=5220) , %75.3'ü (n=15905) kadındır. Hipotiroid tanısına sahip kadınların oranı erkeklerin oranından daha fazladır.

Kadınların %23.25'i (n=3698), erkeklerin ise %16.07'si (n=839) 45-55 yaş grubundadır.

4.1 Erkek Hastaların 18 Yaş Altına İlişkin Verilerin Değerlendirilmesi



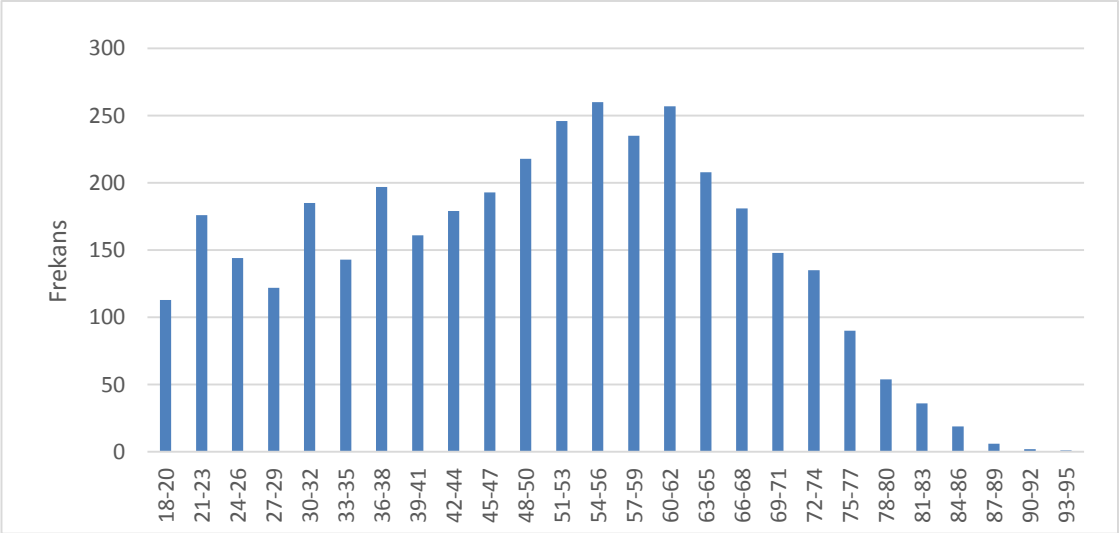
Şekil-4.1 18 yaş altı erkek hastaların yaşlarına göre frekansları

Tablo-4.1 18 yaş altı erkek hastaların kümelenmesi, yaş ve laboratuvar ölçüm değerlerinin kümelere göre karşılaştırılması

ERKEK (YAŞ<18)	K1(n=779)	K2(n=732)	Cliff's Delta	
	Medyan (Min-Maks)	Medyan (Min-Maks)	δ	%95 G. A. Alt:Üst
YAŞ	4 (1-8)	13 (9-17)	1.000	0.999:1.000
TSH	2.46 (0.003-44.87)	2.09 (0-26.23)	0.137	0.079:0.194
Serbest T3	3.65 (1-8.27)	3.5 (1-9.78)	0.151	0.093:0.208
Serbest T4	1.19 (0.54-2.44)	1.11 (0.43-3)	0.222	0.165:0.278

k-ortalamalar kümeleme yöntemi ile kümelenen gruplar yaş ve ölçüm değerleri değerlendirildiğinde farklılık göstermektedir. Yaş değişkeni bakımından iki küme arasında büyük etki büyüklüğü görülürken, yaş değerleri düşük olan küme-1 ile daha büyük olan küme-2 arasında Serbest T3 ve Serbest T4'ün değerleri bakımından küçük TSH değişkeni bakımından ise önemsiz etki büyüklüğü görülmüştür.

4.2 Erkek Hastaların 18 Yaş Üstüne İlişkin Verilerin Değerlendirilmesi



Şekil-4.2 18 yaş ve üzeri erkek hastaların yaşlara göre frekansları

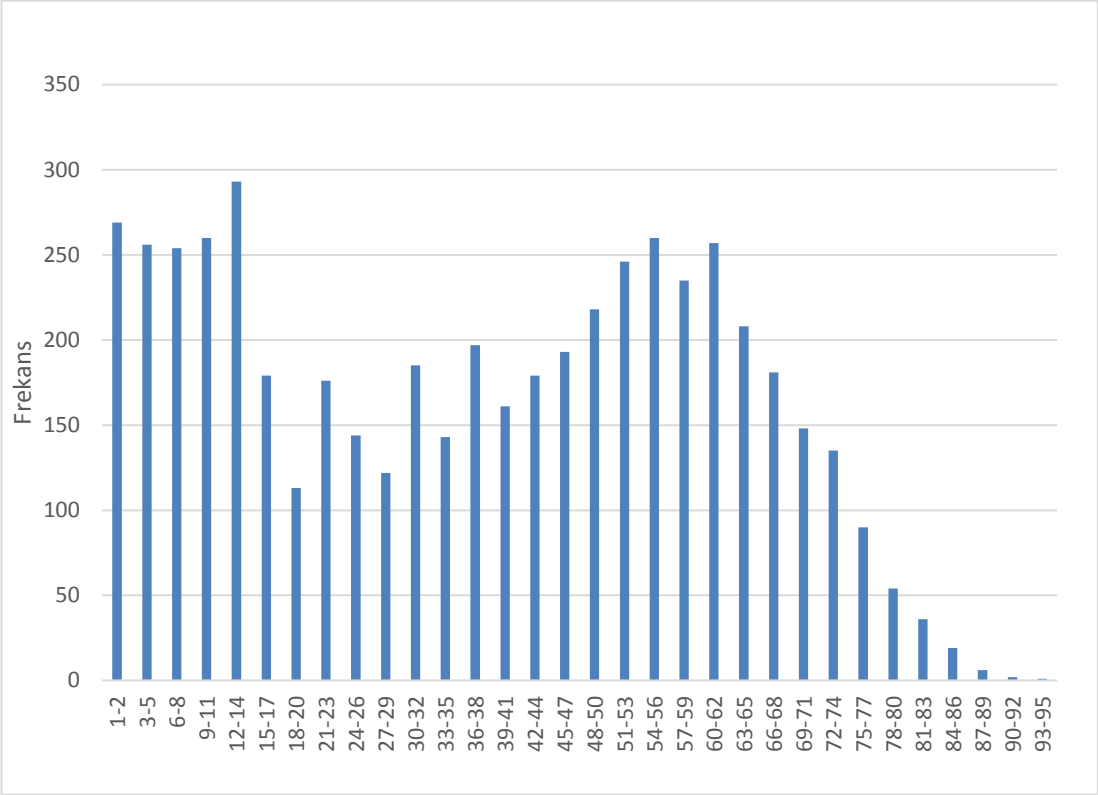
18 yaş ve üzeri erkek hastaların frekansları incelendiğinde, başlangıçta yaş arttıkça hipotiroidi görülmesinde artış gözlemlendiği 54-56 yaş ve 60-62 yaş arasında en yüksek frekansa ulaştığı, daha sonra ise yaş arttıkça düşüş olduğu görülmektedir.

Tablo-4.2 18 yaş ve üzeri erkek hastaların kümelenmesi, laboratuvar ölçüm değerlerinin kümelere göre karşılaştırılması

ERKEK ***(YAŞ>=18)***	K1(n=1610)	K2(n=2099)	Cliff's Delta	
	Medyan (Min-Maks)	Medyan (Min-Maks)	δ	%95 G. A. Alt:Üst
YAŞ	34 (18-47)	61 (47-93)	1.000	0.999:1.000
TSH	1.49 (0-97.95)	1.12 (0-97.23)	0.108	0.070:0.144
Serbest T3	3.03 (1-30)	2.72 (1-11.3)	0.309	0.273:0.344
Serbest T4	1.09 (0.40-6)	1.09 (0.40-4.72)	0.004	-0.033:0.041

18 yaş ve üzeri erkek hastalarda oluşan iki küme etki büyüklükleri değerlendirildiğinde, yaş bakımından iki küme arasında büyük etki büyüklüğü görülmektedir. Yaş değişkeni büyük olan küme ile daha düşük olan küme arasında TSH ve Serbest T4 bakımından önemsiz, Serbest T3 değişkeni için küçük etki büyüklüğü olarak farklılık göstermektedir.

4.3 Erkek Hastaların Tamamına İlişkin Verilerin Değerlendirilmesi



Şekil-4.3 Erkek hastaların yaşlara göre frekansları

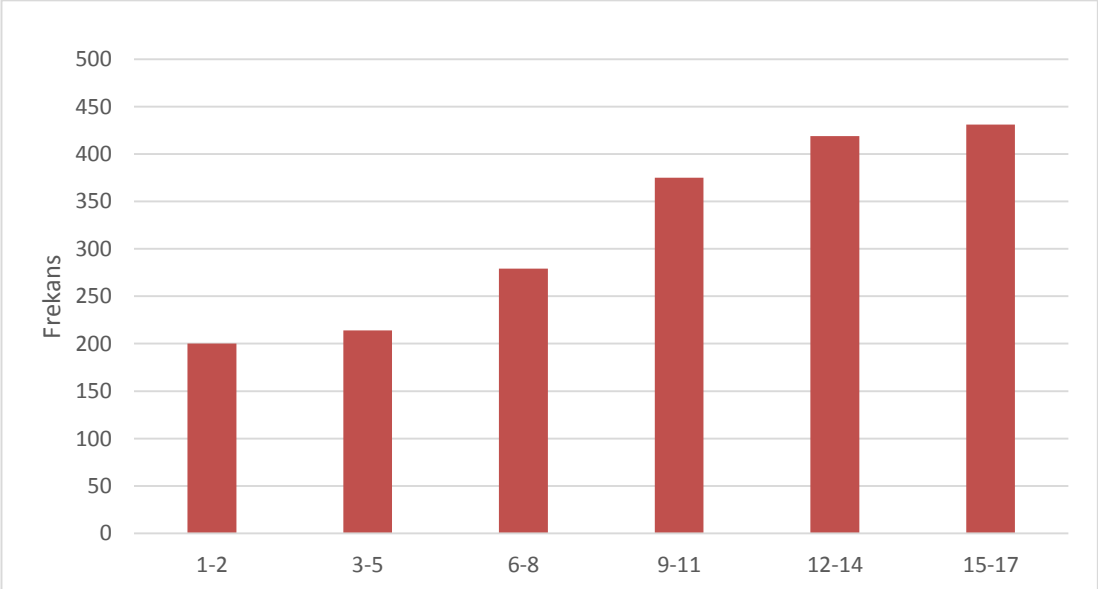
Şekil- 4.3 incelendiğinde**,** 1-14 ve 48-65 yaş arası hipotiroidi hastalıkları tanısına sahip erkek hastalarda diğer yaş gruplarına göre sayıca daha fazladır. Genç yaşlarda hipotiroidi erkekler için daha düşük frekansa sahiptir.

Tablo-4.3 Erkek hastaların kümelenmesi, laboratuvar ölçüm değerlerinin kümelere göre karşılaştırılması

ERKEK *(GENEL)*	K1(n=2393)	K2(n=2827)	Cliff's Delta	
	Medyan (Min-Maks)	Medyan (Min-Maks)	δ	%95 G. A. Alt:Üst
YAŞ	13 (1-35)	56 (35-93)	1.000	0.999:1.000
TSH	2.01 (0-83.36)	1.18 (0-97.95)	0.271	0.240:0.301
Serbest T3	3.39 (1-30)	2.78 (1-30)	0.545	0.519:0.570
Serbest T4	1.13 (0.40-5.86)	1.09 (0.40-6)	0.127	0.096:0.158

Erkek hastaların tamamı değerlendirildiğinde iki küme arasında yaş değişkeni için büyük etki büyüklüğü görülmektedir. Yaşı daha küçük olan küme ile daha büyük olan küme arasında Serbest T3 için büyük etki büyüklüğü, TSH değişkeni için küçük, Serbest T4 bakımından ise önemsiz etki büyüklüğü olduğu Tablo-4.3 üzerinde görülmektedir.

4.4 Kadın Hastaların 18 Yaş Altına İlişkin Verilerin Değerlendirilmesi



Şekil-4.4 18 yaş altı kadın hastaların yaşlara göre frekansları

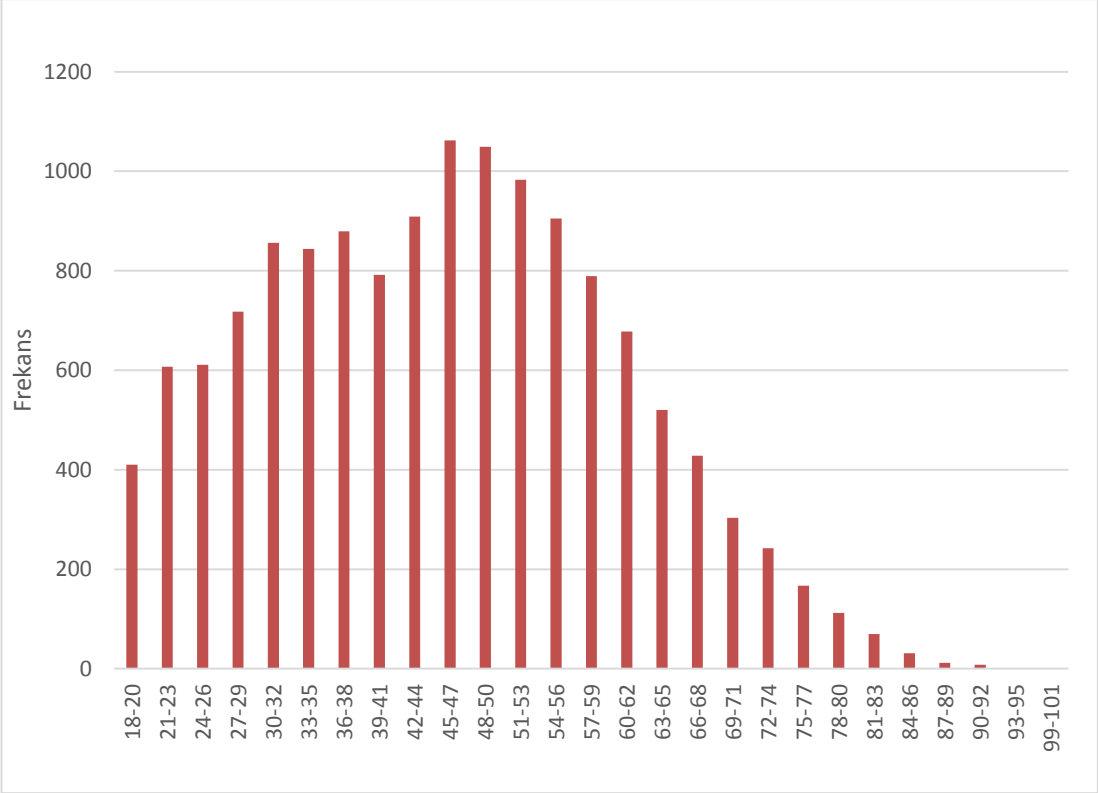
18 yaş altı kadın hastalarda, yaş arttıkça hipotiroidi tanısına sahip hastaların frekansının arttığı görülmektedir.

Tablo-4.4 18 yaş altı kadın hastaların kümelenmesi, yaş ve laboratuvar ölçüm değerlerinin kümelere göre karşılaştırılması

KADIN ***(YAŞ<18)***	K1(n=774)	K2(n=1144)	Cliff's Delta	
	Medyan (Min-Maks)	Medyan (Min-Maks)	δ	%95 G. A. Alt:Üst
YAŞ	5 (1-15)	13 (9-17)	0.981	0.924:0.995
TSH	2.82 (0.003-87.33)	1.80 (0-24.71)	0.342	0.292:0.390
Serbest T3	3.69 (1.04-8.86)	3.31 (1.13-19.06)	0.333	0.283:0.381
Serbest T4	1.19 (0.40-2.96)	1.10 (0.55-3.58)	0.220	0.168:0.270

18 yaş altı kadın hastalar kümeler arası karşılaştırıldığında, yaş bakımından büyük etki büyüklüğü görülmektedir. TSH ve Serbest T3 için yaşı daha büyük olan küme ile daha küçük olan iki küme karşılaştırıldığında orta etki büyüklüğü, Serbest T4 için ise küçük etki büyüklüğü görülmektedir.

4.5 Kadın Hastaların 18 Yaş Üstüne İlişkin Verilerin Değerlendirilmesi



Şekil-4.5 18 yaş ve üzeri kadın hastaların yaşlara göre frekansları

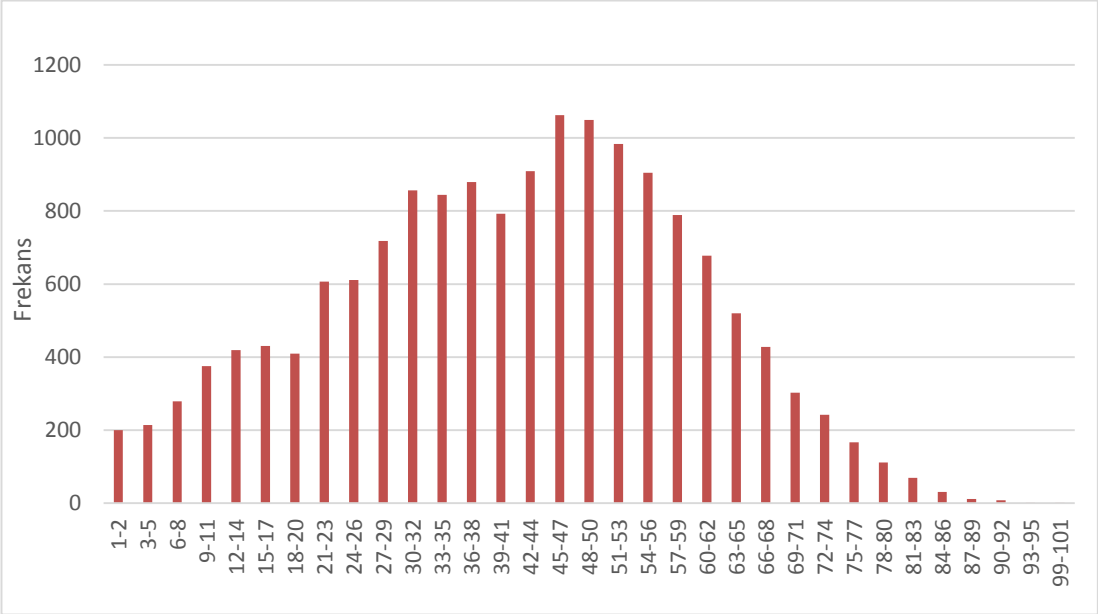
Şekil-4.5 incelendiğinde, kadın hastalarda 45-50 yaş aralığı, diğer yaş gruplarına göre en yüksek frekansa sahiptir. Hipotiroidi tanısı orta yaşlarda en yüksek frekansa, genç ve yaşlılarda daha düşük frekansa sahiptir.

Tablo-4.5 18 yaş ve üzeri kadın hastaların kümelenmesi, yaş ve laboratuvar ölçüm değerlerinin kümelere göre karşılaştırılması

KADIN ***(YAŞ>=18)***	K1(n=6618)	K2(n=7369)	Cliff's Delta	
	Medyan (Min-Maks)	Medyan (Min-Maks)	δ	%95 G. A. Alt:Üst
YAŞ	33 (18-44)	55 (43-101)	1.000	0.999:1.000
TSH	1.70 (0-98.14)	1.47 (0-99.12)	0.057	0.038:0.076
Serbest T3	2.85 (1-30)	2.71 (1-21.11)	0.179	0.160:0.197
Serbest T4	1.09 (0.40-5.54)	1.13 (0.40-5.41)	0.089	0.070:0.108

18 yaş ve üzeri kadın hastalara ilişkin kümeler arası karşılaştırma yapıldığında, yaş değişkeni için büyük etki büyüklüğü vardır. İki küme arasında TSH ve Serbest T4 için önemsiz, Serbest T3 değişkeni için küçük etki büyüklüğü vardır.

4.6 Kadın Hastaların Tamamına İlişkin Verilerin Değerlendirilmesi



Şekil-4.6 Kadın hastaların yaşlara göre frekansları

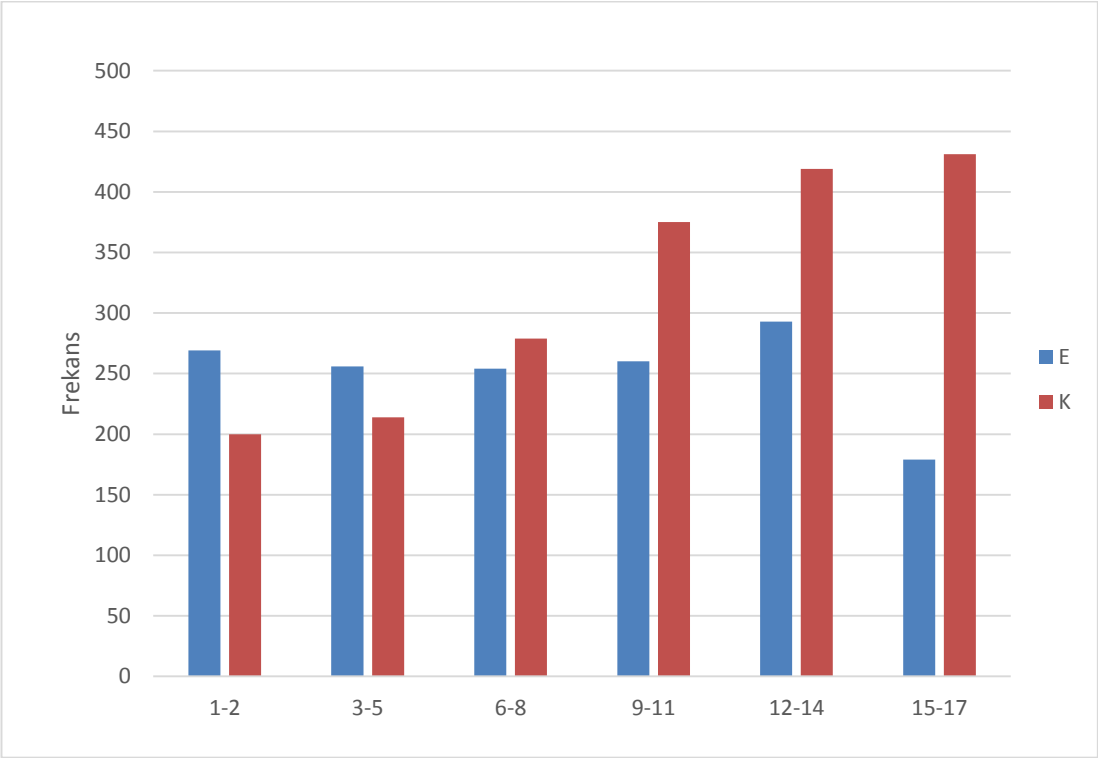
Şekil- 4.6'da kadın hastalara ilişkin yaşlara göre frekanslar incelendiğinde, yaş arttıkça tiroid rahatsızlığına sahip hastaların sayısının arttığı, 45-50 yaş arasında pik noktasına ulaştığı ve 50 yaşından sonra frekansların azalarak devam ettiği görülmektedir.

Tablo-4.6 Kadın hastaların kümelenmesi, yaş ve laboratuvar ölçüm değerlerinin kümelere göre karşılaştırılması

KADIN ***(GENEL)***	K1(n=7092)	K2(n=8813)	Cliff's Delta	
	Medyan (Min-Maks)	Medyan (Min-Maks)	δ	%95 G. A. Alt:Üst
YAŞ	26.5 (1-39)	53 (38-101)	1.000	0.999:1.000
TSH	1.84 (0-98.14)	1.49 (0-99.12)	0.091	0.073:0.109
Serbest T3	2.98 (1-30)	2.72 (1-21.11)	0.316	0.298:0.332
Serbest T4	1.10 (0.40-5.54)	1.12 (0.40-5.41)	0.040	0.023:0.058

Kadın hastaların tamamı incelendiğinde ise yaş değişkeni için iki küme arasında fark etki büyüklüğü bakımından büyük olduğu gözlemlenmiştir. Serbest T3 bakımından ise küçük etki büyüklüğü farkı gözlemlenmiştir. TSH ve Serbest T4 değerleri için kümeler arası etki büyüklüğü önemsizdir.

4.7 Hastaların 18 Yaş Altına İlişkin Verilerin Değerlendirilmesi



Şekil-4.7 18 yaş altı hastaların cinsiyete ve yaşlara göre frekansları

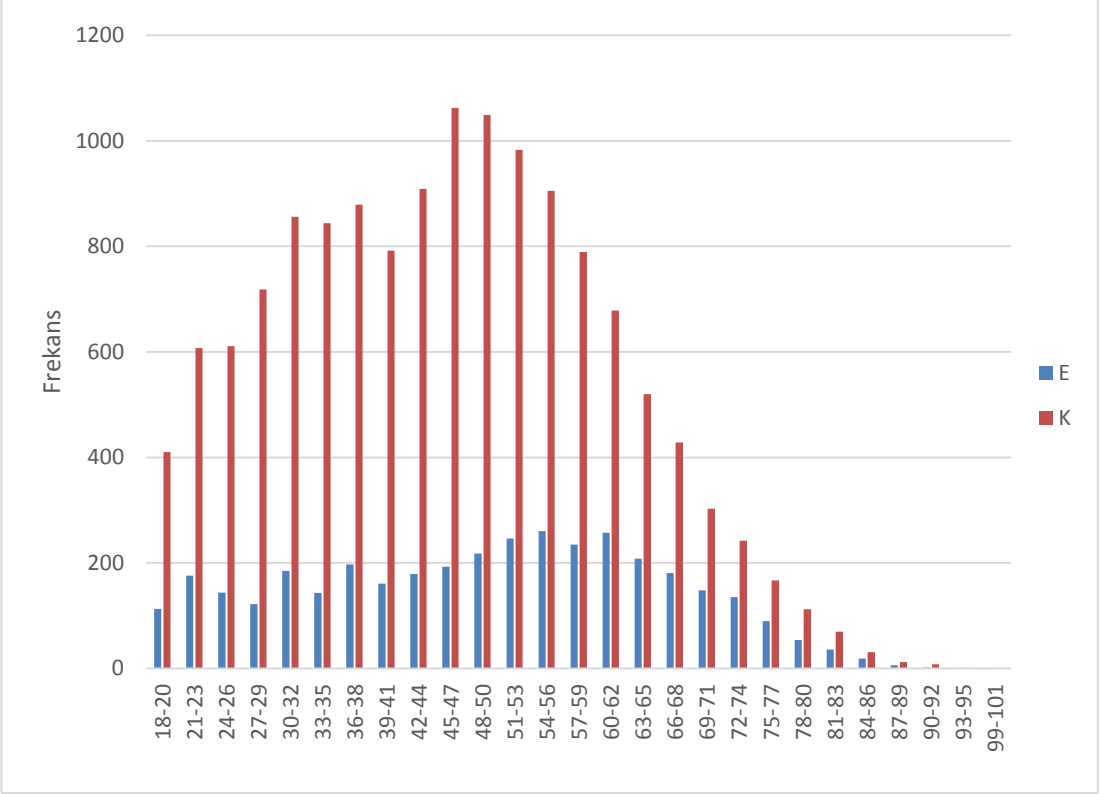
Tablo-4.7 18 yaş altı hastaların kümelenmesi, yaş ve laboratuvar ölçüm değerlerinin kümelere göre karşılaştırılması

GENEL ***(YAŞ<18)***	K1(n=1531)	K2(n=1898)	Cliff's Delta	
	Medyan (Min-Maks)	Medyan (Min-Maks)	δ	%95 G. A. Alt:Üst
YAŞ	4 (1-15)	13 (9-17)	0.991	0.994:0.986
TSH	2.61 (0.003-87.33)	1.93 (0-26.23)	0.249	0.211:0.286
Serbest T3	3.66 (1-8.86)	3.39 (1-19.06)	0.253	0.215:0.290
Serbest T4	1.19 (0.40-2.96)	1.11 (0.43-3.58)	0.225	0.188:0.263

18 yaş altı hastalar bakımından iki küme karşılaştırıldığında iki küme arasında büyük etki büyüklüğü görülmüştür. Yaşı büyük olan küme için ölçüm değerleri yaşı

daha küçük olan kümeye göre laboratuvar ölçüm değerleri ilişkin medyan değeri daha küçük olmakla birlikte iki küme arasında küçük etki büyüklüğü gözlemlenmiştir.

4.8 Hastaların 18 Yaş Üstüne İlişkin Verilerin Değerlendirilmesi



Şekil-4.8 18 yaş üstü hastaların cinsiyete ve yaşlara göre frekansları

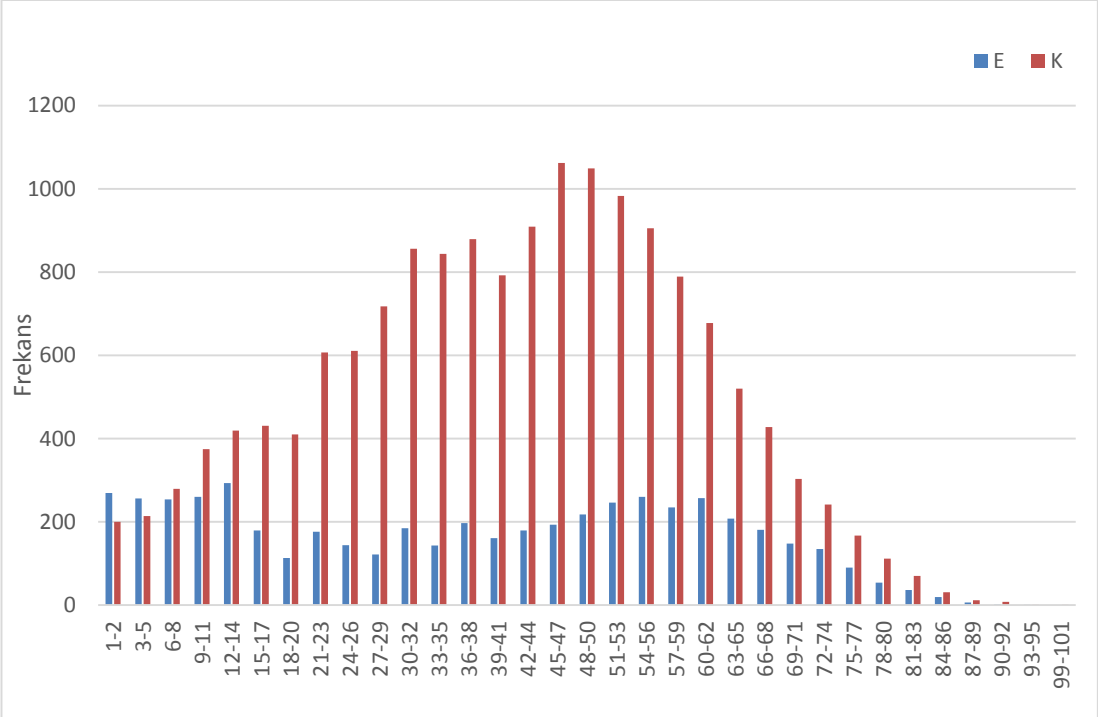
Şekil-4.8 incelendiğinde, 18 yaş üstü hastalar için frekanslar, kadınlarda 48-50 yaş aralığında erkeklerde ise 54-56 yaş aralığında en yüksek değere ulaşmıştır.

Tablo-4.8 18 yaş ve üzeri hastaların kümelenmesi, yaş ve laboratuvar ölçüm değerlerinin kümelere göre karşılaştırılması

GENEL ***(YAŞ>=18)***	K1(n=8856)	K2(n=8840)	Cliff's Delta	
	Medyan (Min-Maks)	Medyan (Min-Maks)	δ	%95 G. A. Alt:Üst
YAŞ	34 (18-46)	57 (45-101)	1.000	0.999:1.000
TSH	1.65 (0-98.26)	1.36 (0-99.12)	0.074	0.057:0.091
Serbest T3	2.87 (1-30)	2.71 (1-21.11)	0.198	0.182:0.215
Serbest T4	1.09 (0.40-6)	1.12 (0.40-5.41)	0.067	0.050:0.084

18 yaş ve üzeri yaş ve laboratuvar ölçüm değerleri karşılaştırıldığında iki küme arasında benzer şekilde yaş bakımından büyük etki büyüklüğü gözlemlenmiştir. Serbest T3 için iki küme arasında etki büyüklüğü küçük, TSH ve Serbest T4 bakımından önemsizdir.

4.9 Hastaların Tamamına İlişkin Verilerin Değerlendirilmesi



Şekil-4.9 Hastalara ilişkin cinsiyet ve yaşlara göre frekansları

Şekil- 4.9 incelendiğinde ise hipotiroid tanısına sahip kadın hastaların yaş gruplarına göre frekansları, erkeklere göre çok daha yüksek olduğu açıkça görülmektedir.

Tablo-4.9 Hastaların tamamının kümelenmesi, yaş ve laboratuvar ölçüm değerlerinin kümelere göre karşılaştırılması

GENEL ***(GENEL)***	K1(n=9420)	K2(n=11705)	Cliff's Delta	
	Med (Min-Maks)	Med (Min-Maks)	δ	%95 G. A. Alt:Üst
YAŞ	23 (1-38)	54 (37-101)	1.000	0.999:1.000
TSH	1.88 (0-98.14)	1.41 (0-99.12)	0.130	0.115:0.146
Serbest T3	3.08 (1-30)	2.73 (1-30)	0.382	0.368:0.397
Serbest T4	1.11 (0.40-5.86)	1.13 (0.40-6)	<0.001	-0.016:0.015

Hastaların tamamı kümeler arası karşılaştırıldığında, yaş değişkeni için büyük etki büyüklüğü gözlemlenmiştir. Serbest T3 değişkeni bakımından ise orta etki büyüklüğü gözlenmiştir. Yaşı büyük olan kümenin medyanı Serbest T3 değişkeni iki yaşı büyük olan kümeye göre daha düşüktür. TSH ve Serbest T4 bakımından iki küme arasında etki büyüklüğü önemsiz olarak görülmektedir.

5. TARTIŞMA VE SONUÇ

Büyük veri kullanımı süresince yalnızca veri toplamak ve eklenecek veriler için depolama alanlarının oluşturmak yeterli değildir. Verinin aynı zamanda analize uygun hale getirilmesi ve çalışmanın amacına uygun metodun seçilip analiz edilmesi aynı zamanda da yorumlanması gerekmektedir (Altunışık, 2015). Oluşan bilgi yoğunluğundan çalışmaya yönelik verilerin elde edilmesi, düzenlenmesi, karar verme aşamalarında süreci hızlandırması sonucu zaman ve maliyet bakımından olumlu yönde etkilemektedir. Büyük veri kullanımının olumlu yönlerinin yanında ayrıca olumsuz olabilecek yönleri de bulunmaktadır. Veri düzenleme işlemi konu uzmanı yardımı ile kontrollü şekilde yapılmalı ve veri manipüle edilmeden titizlikle çalışılmalıdır. Büyük sayıda veri ile çalışılması nedeniyle yapılacak her yanlış işlem çok ciddi sorunlar ortaya çıkaracaktır. Çalışılacak veri setinin güvenilir kaynaklardan elde edilmemesi ve kişisel bilgi mahremiyetine önem verilmemesi hukuki sorunlara neden olacak ve telafisi mümkün olmayan sonuçlar ile karşı karşıya kalınacaktır.

Büyük veri kullanımında öngörülen sorunlara karşı önlemler alınsa da önlenemeyen ve kontrol edilmesi mümkün olmayan bazı durumlar ortaya çıkmaktadır. Hipotiroidi tanısı, TSH ölçüm değerinin yüksekliği ile ilişkilidir. Önlenemeyen durumlar söz konusu olabilir. TSH ölçüm değerinin belirtilen referans aralıklarından düşük çıkmasının farklı nedenlerinden birisi, başka bir hastanede tedavi almış olan hastanın Bursa Uludağ Üniversitesi Sağlık Uygulama ve Araştırma Merkezi'ne ilk kez başvurmuş olması olabilir. Başka bir sorun ise hastanın hipofizer ya da bir başka ifadeyle sekonder hipotiroidi olma durumu olabilir. Hipofizer hipotiroidi, TSH hormonu salgısının yetersizliği sonucu ortaya çıkan bir rahatsızlıktır (Sağlam ve Çakır, 2012). Uzman tarafından tanının yanlış konulması durumları da göz önünde bulundurulduğunda büyük veri kullanımında ortaya çıkabilecek ve önlenemeyen bazı sorunlar olarak gözlemlenmiştir.

TEMD'in yayınladığı Tiroid Hastalıkları Tanı ve Tedavi Kılavuzuna göre genç ve orta yaşlılar için TSH değerleri 0.5 ile 2.5 mU/mL olması gerektiğini belirtmektedir. Yaşlılar için TSH değerleri ise 3-6 mU/mL arasında olabileceğini belirtilmiştir (Türkiye Endokrinoloji Metabolizma Derneği, 2019). Çalışmamızda Bursa Uludağ Üniversitesi Sağlık Uygulama ve Araştırma Merkezi Hastanesinde tiroid rahatsızlıklarına sahip ve ilk tanı değerleri dikkate alınmıştır. Genç bireyler için medyan değeri 1.88 olup 0 ile 98.14 arasında değişmekte, orta ve yaşlı bireyler için medyan değeri 1.41 olmak üzere 0 ile 99.12 arasında olması nedeniyle medyan değerleri belirtilen referans aralıklarının altında olduğu gözlemlenmiştir.

Dünyada ve ülkemizde de yaygın olarak görülen tiroid hastalıkları cinsiyet bazında incelendiğinde kadınlarda daha yaygın olarak görülmektedir. (Hueston ve ark. , 2008) Birçok çalışmada da tiroid hastalıklarında kadın erkek arasındaki oranın 3.5 ile 12 arasında olduğu belirtilmektedir. Kadın predominansı, hipotiroid ve hipertiroid, aşikâr ve subkliniktiroid hastalıklarında, düşük ve yüksek iyot tüketiminin bulunduğu bölgelerde gözlemlenmektedir. (Iglesias ve Diez, 2008) Çalışmamızda kadın/erkek oranı 3.05 olarak hesaplanmış olup hastaların önemli kısmı kadınların oluşturduğunu göstermektedir.

Tiroid hastalığı benzer şekilde ilerleyen yaşlarda da yaygın olarak görülebilmektedir. (Levy, 1991) Kadınlarda, 48-50 yaş arasında frekans en yüksek seviyeye ulaşmıştır. Kadınların %23.25'ini 45-55 yaş aralığı oluşturmakta olup benzer şekilde erkeklerde de %16.07'sini 45-55 yaş aralığı oluşturmaktadır.

Kapelari ve arkadaşlarının (2008) hastaların doğumundan yetişkinlik durumuna kadar dönemi incelediklerinde, Serbest T3, Serbest T4 ve TSH laboratuvar ölçüm değerleri ile yaş arasında istatistiksel olarak negatif ilişki olduğu sonucuna varmışlardır. Yaş arttıkça laboratuvar ölçüm değerlerinde düşüş gözlemlemişlerdir. Çalışmamızda benzer şekilde 18 yaşından küçük hastalar incelendiğinde, yaşı küçük olan hastalar ile yaşı daha büyük hastalar arasında TSH, Serbest T3 ve Serbest T4 bakımından kümeler arasında farklılık gözlemlenmiştir. Laboratuvar ölçüm değerlerinin iki küme arasında medyan değerleri incelendiğinde yaşı daha küçük olan hastaların bulunduğu kümenin medyanı, yaşı daha büyük olan hastalarının oluşturduğu kümeye göre daha yüksek çıkmıştır.

K. Takeda ve arkadaşlarının (2009), 1007 Japon hasta üzerinden yaptığı çalışmada erkeklerde, Serbest T4 ölçüm değişkenine ilişkin değerler ile yaş arasında istatistiksel olarak anlamlı ve negatif ilişkiye sahip olduğu sonucuna varılmıştır. Yaş arttıkça Serbest T4 ölçüm değerinin azaldığını gözlemlemişlerdir. Çalışmamızda erkekler yaşı büyük olan küme ile yaşı küçük olan kümenin Serbest T4 bakımından iki küme arasında farklılığın olmadığı gözlemlenmiştir.

Jammah ve arkadaşları (2015) verilerinin tamamını incelediklerinde, yaş ile TSH arasında negatif ilişkiye sahip olduğu, yaş ile TSH arasında da istatistiksel olarak ilişkiye rastlanılmadığı belirtilmiştir. Çalışmamızda benzer şekilde yaşı küçük olan küme ile yaşı büyük olan küme arasında TSH bakımından kümeler arasında farklılığı olmadığı gözlemlenmektedir.

Çalışmamızda da bazı laboratuvar değerlerinde hastalık tanısında kabul gören laboratuvar değerlerine göre farklılıklar görülmüştür. Tanı koymada referans alınan Serbest T3 ve Serbest T4 laboratuvar değerleri çalışmamızdaki büyük veri analiz sonuçlarıyla uyumlu çıkarken TSH laboratuvar değerleri uyumsuz çıkmıştır. Görülen farklılıkların nedenleri olarak, veri tabanına hatalı girişlerden yapılmasından, hatalı ICD kodlarının verilmesinden, hastaların daha önceden belirtmedikleri ama almış oldukları veya kullanmış oldukları ilaçlardan kaynaklanabilir. Bununla birlikte bu farklılıklar bir hata olmaksızın da ortaya çıkmış olabilir. Büyük veri analizinin sonrasında ortaya çıkan bu farklılıklar, kontrollü çalışmalar ile planlanarak farklılıkların değerlendirilmesi ve araştırılması gerektiğini düşündürmektedir.

6. KAYNAKLAR

Aksoy B, Bayrakçı H, Bayrakçı E ve ark. (2017) Büyük Verinin Kurumlarda Kullanımı. Süleyman Demirel Üniversitesi, İktisadi ve İdari Bilimler Fakültesi Dergisi, 22, 1853-1878.

Aldenderfer M, Blashfield R (1989) Cluster Analysis, 6th Edition. London, Sage Publications Inc.

Alpar R (2011) Uygulamalı Çok Değişkenli İstatistiksel Yöntemler. Ankara, Detay Yayıncılık.

Altunışık R (2015) Büyük Veri: Fırsatlar Kaynağı mı Yoksa Yeni Sorunlar Yumağı mı? Yıldız Social Science Review, 1(1), 45-76.

Baro E, Degoul S, Beuscart R et al (2015) Toward a Literature-Driven Definition of Big Data in Healthcare. BioMed Research International, 1-9.

Berkhin P (2006) A Survey of Clustering Data Mining Techniques. In Kogan J., Nicholas C., Teboulle M. (eds) Grouping Multidimensional Data. Springer, Berlin, Heidelberg.

Berry M, Linoff D (2004) Data Mining Techniques. Wiley Publishing Inc.

Chaorasiya V, Shrivastava A (2015) A survey on Big Data : Techniques and Technologies. International Journal of Research and Development in Applied Science and Engineering, 8(1), 1-4.

Cliff N (1996) Ordinal Methods for Behavioral Data Analysis. Hillsdale, NJ, Erlbaum.

Dumbill E (2013). Making Sense of Big Data. Big Data, Mary Ann Liebert, Inc., 1(1), 1-2.

Fisher D, DeLine R, Czerwinski M et al (2012). Interactions with Big Data Analytics. interactions, 19(3), 50-59.

Grasso M, Comer A, DiRenzo D et al (2015) Using Big Data to Evaluate the Association between Periodontal Disease and Rheumatoid Arthritis. AMIA Annu Symp Proc, 589-593.

Han J, Kamber M, Pei, J (2012) Data Mining Concepts and Techniques , Third Edition. Morgan Kaufmann.

Hasan M, Mohibullah M, Hossain Z (2015) Comparison of Euclidean Distance Function and Manhattan Distance Function Using K-Mediods. International Journal of Computer Science and Information Security, 13, 61-71.

Hogarty K, Kromrey J (1999) Using SAS to Calculate Tests of Cliff's Delta. Proceedings of the Twenty-Foursth Annual SAS User Group International Conference, Miami Beach, Florida, 238.

Hosmer D, Lemeshow S (2000) Applied Logistic Regression Second Edition. John Wiley & Sons, Inc.

Hueston W, Carek P, Allweiss, P (2008) Endocrine Disorders in Chapter 35. Current Diagnosis & Treatment Family Medicine 2nd Edition, McGraw-Hill Companies, 392.

Iglesias P, Diez J (2008) Hypothyrodism in Male Patients: A Descriptive, Observational and Cross-Sectional Study in a Series of 260 Men. The American Journal of the Medical Sciences, 336(4), 315-320.

Işık M, Çamurcu A (2007) k-Means, k-Medoids ve Bulanık c-Means Algoritmalarının Uygulamalı Olarak Performanslarının Tespiti. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi(11), 31-45.

Jammah A, Alshehri A, Alrakhis A et al (2015) Characterization of thyroid function and antithyroid antibody tests among Saudis. Saudi Medical Journal, 36(6), 692-7.

Jeong K, Lee J, Kang D et al (2018) A population-based epidemiological study of anaphylaxis using national big data in Korea: trends in age-specifc prevalence and epinephrine use in 2010–2014. Allergy, Asthma & Clinical Immunology, 14(31).

Johnson A, Wichern W (1988) Applied Multivariate Statistical Analysis. New Jersey, Prentice Hall.

Kapelari K, Kirchlechner C, Hogler W et al (2008) Pediatric reference intervals for thyroid hormone levels from birth to adulthood: a retrospective study. BMC Endocrine Disorders, 8(1), 15.

Kaufman L, Rousseeuw P (1990) Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons.

Kılınç D, Bozyiğit F, Borandağ E ve ark. (2016). Sınıflandırma Tabanlı Zombi Bilgisayar Tespit Sistemi. Akademik Bilişim.

Kumari S (2018) Breast Cancer Classification Using Big Data Approach. PARIPEX - Indian Journal of Research, 7(1), 401-403.

Levy E (1991) Thyroid Disease in the Elderly. Medical Clinics of North America, 75(1), 151-167.

MacQueen J (1967) Some Methods for Classification and Analysis. Proc. Symp. Math. Statist. and Probability (5th), 281-297.

Manyika J, Chui M, Brown B et al (2011) Big Data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, Washington.

Musa A (2014) Logistic Regression Classification for Uncertain Data. Research Journal of Mathematical and Statistical Sciences, 2(2), 1-6.

Olgun M, Özdemir G (2012) İstatistiksel Özellik Temelli Bayes Sınıflandırıcı Kullanarak Kontrol Grafiklerinde Örüntü Tanıma. Journal of the Faculty of Engineering and Architecture of Gazi University, 27(2), 303-311.

Pena J, Lozano J, Larranaga P (1999) An empirical comparison of four initialization methods for the K-Means algorithm. Pattern Recognition Letters, 20(10), 1027-1040.

Raghupathi W, Raghupathi V (2014) Big data analytics in healthcare: promise and potential. Health Information Science and Systems, 2(1), 1-10.

Ray S, Turi R (1999) Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation. Proc. of the 4th Int. Conf. on Advances in Pattern Recognition and Digital Techniques, Calcutta, India, Narosa Publishing House, New Delhi, 137-143.

Raymond T, Han, J (1994) Efficent and Effective Clustering Methods for Spatial Data Mining. VLDB, 144-155.

Romano J, Kromrey J, Coraggio J et al (2006) Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and other surveys? Annual Meeting of the Florida Association of Institutional Research.

Sağlam F, Çakır B (2012) Birinci Basamakta Tiroid Hastalıklarına Klinik Yaklaşım. Ankara Medical Journal, 12(3), 136-139.

Sangeetha S, Sreeja A (2015) No Science No Humans, No New Technologies No changes "Big Data a Great Revolution". International Journal of Computer Science and Information Technologies, 6(4), 3269-3274.

Sarıman G (2011) Veri Madenciliğinde Kümeleme Teknikleri Üzerine Bir Çalışma: k-Means ve k-Medoids Kümeleme Algoritmalarının Karşılaştırılması. Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü Dergisi, 15(3), 192-202.

Serin S, İlhan M, Ahcı S ve ark. (2006) Tiroid Hastalıklarında Bilinç Düzeyi. Şişli Etfal Hastanesi Tıp Bülteni, 50(3), 181-185.

Takeda K, Mishiba M, Sugiura H et al (2009) Evaluated Reference Intervals for Serum Free Thyroxine and Thyrotropin Using the Conventional Outliner Rejection Test without Regard to Presence of Thyroid Antibodies and Prevalence of Thyroid Dysfunction in Japanese Subjects. Endocrine Journal, 56(9), 1059-1066.

Tatlıdil H (1992) Uygulamalı Çok Değişkenli İstatistiksel Analiz. Ankara: Engin Yayınları.

Türkiye Endokrinoloji Metabolizma Derneği (2019) Tiroid Hastalıkları Tanı ve Tedavi Kılavuzu 2019, 37.

Ward J, Barker A (2013) Undefined By Data: A Survey of Big Data Definitions, https://arxiv.org/pdf/1309.5821.pdf.

Xu R, Wunsch D (2005) Survey of clustering algorithms. Neural Networks, IEEE(16(3), 645-678)

7. KISALTMALAR

Min. : Minimum değer
Maks. : Maksimum değer
TEMD: Türkiye Endokrinoloji ve Metabolizma Derneği
δ: Cliff Delta Katsayısı

EK I

8. EKLER

BURSA ULUDAĞ ÜNİVERSİTESİ
ARAŞTIRMA VE YAYIN ETİK KURULLARI
(Sağlık Bilimleri Araştırma ve Yayın Etik Kurulu)
TOPLANTI TUTANAĞI

OTURUM TARİHİ	OTURUM SAYISI
14 Eylül 2018	2018-05

KARAR NO 4: Üniversitemiz Tıp Fakültesi öğretim üyesi Prof. Dr. İlker ERCAN'ın "Otoimmün Tiroid Hastalıklardan Basedow Graves, Tiroid Orbitopati veya Hashimoto Tiroiditi Tanısı Konmuş Olguların Büyük Veri Analizi ile İncelenmesi" başlıklı araştırmasına ilişkin konunun değerlendirilmesine geçildi.

Yapılan görüşmeler sonunda; Üniversitemiz Tıp Fakültesi öğretim üyesi Prof. Dr. İlker ERCAN'ın "Otoimmün Tiroid Hastalıklardan Basedow Graves, Tiroid Orbitopati veya Hashimoto Tiroiditi Tanısı Konmuş Olguların Büyük Veri Analizi ile İncelenmesi" başlıklı araştırmanın, fikri, hukuki ve telif hakları bakımından metot ve ölçeğine ilişkin sorumluluğu başvurucuya ait olmak üzere uygun olduğuna oybirliği ile karar verildi.

Prof.Dr.Mehmet YÜCE
Kurul Başkanı

9. TEŞEKKÜR

Yüksek lisans eğitim boyunca ve tezimi gerçekleştirmem sırasında benden desteğini esirgemeyen değerli danışmanım Prof. Dr. İlker ERCAN'a güveni ve sabrı için çok teşekkür ederim. Tezimde klinik olarak yardımlarını esirgemeyen Prof. Dr. Canan ERSOY'a teşekkürlerimi sunarım.

Yüksek lisans öğrenim boyunca eğitimime katkıda bulunan anabilim dalımızdaki öğretim üyelerine teşekkürlerimi sunarım

Ayrıca yüksek lisans eğitimim boyunca ve tezimi gerçekleştirmem sırasında, maddi ve manevi göstermiş olduğu desteklerinden dolayı aileme teşekkürlerimi sunarım.

10. ÖZGEÇMİŞ

1992 yılında Bursa'da doğdum. Anadolu Üniversitesi İstatistik bölümünden 2017 yılında mezun oldum. 2015-2016 güz döneminde Varşova Teknoloji Üniversitesinde Erasmus eğitimimi tamamladım. 2017 yılı güz döneminde Uludağ Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik Anabilim Dalında yüksek lisans eğitimime başladım.