Sağlık Alanında Yapılan Araştırmalarda Kümeleme Algoritmalarının Kullanımı: Bir Uygulama

Usage of Cluster Algorithms in Health Studies: An Application

Özge PASİN,[a]
Handan ANKARALI[a]

[a]Biyoistatistik ve Tıbbi Bilişim AD, Düzce Üniversitesi Tıp Fakültesi, Düzce

Geliş Tarihi/*Received:* 09.12.2015
Kabul Tarihi/*Accepted:* 16.02.2016

Bu çalışma, 17. Ulusal Biyoistatistik Kongresi (5-9 Kasım 2015, Girne, KKTC)'nde sözlü bildiri olarak sunulmuştur.

Yazışma Adresi/*Correspondence:*
Özge PASİN
Düzce Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim AD, Düzce, TÜRKİYE/TURKEY
ozgepasin@duzce.edu.tr

ÖZET Amaç: Farklı kümeleme algoritmalarını, algoritmaların nasıl ve hangi durumlarda doğru bir şekilde kullanılabileceğini tanıtmaktır. Aynı zamanda sağlık araştırmalarından elde edilmiş gerçek bir veri seti üzerinde uygulanabilir olan farklı kümeleme algoritmalarının sonuçlarını karşılaştırmaktır. **Gereç ve Yöntemler:** Kardiyovasküler rahatsızlığına sebep olabilecek risk faktörleri incelenerek bireyler düşük, orta ve yüksek riskli olarak gruplandırabilmek için farklı kümeleme algoritmaları kullanıldı. Kümeleme algoritması olarak EM (expectation maximization), En uzak ilk, Yoğunluk kümeleme, K-Medoid ve Cascade K-ortalama, K-ortalama yöntemleri kullanılmıştır. **Bulgular:** Yapılan değerlendirmeler sonucunda kullanılan iki farklı veri seti için hesaplanan uyum katsayıları istatistik olarak anlamlı bulundu ancak bu katsayılar orta derecede bir uyumu gösterdi. Gerçekleştirilen uygulama sonucunda her iki veri seti içinde kappa katsayısı bakımından en uygun ve en hızlı sonuçlar üreten algoritmanın en uzak ilk kümeleme yöntemi olduğu sonucuna varıldı. Framingham risk grupları oluşturulan kümeler arasında çapraz tablolar oluşturularak grupların dağılımı incelendiğinde ise, en isabetli kararların Make Density Based ve EM algoritmalarına ait kümelerden elde edildiği görüldü. **Sonuç:** Sonuç olarak kümeleme yöntemlerinin hastalıklara ait risk faktörlerinin incelenmesi ve klinik bilgiler de dikkate alınarak hastalık gruplarının oluşturulmasında, doğru hastalık teşhislerinin konulmasında önemli kullanım alanlarına sahip olacağı düşünülmektedir. Ayrıca kümeleme algoritmaları veri dağılımları ve özellikleri dikkate alınarak kullanıldığında sağlık alanında her türlü planlama ve teşhis için kullanılabilir, sonuçta iyi bir araç olacağı düşüncesindeyiz.

Anahtar Kelimeler: Küme analizi; veri madenciliği; algoritma; kalp ve damar hastalıkları; risk faktörleri

ABSTRACT Objective: The purpose of this study is to introduce different clustering algorithms and show how and which cases should be correctly used. At the same time, different clustering algorithms results which can be applied on a real data set were compared**. Material and Methods:** Individuals grouped to low, medium and high risk clusters with using different clustering algorithms by examining risk factors of cardiovascular disease. EM, Denstiy based clustering, K-Medoid, Cascade K-Means and K-Means used for evaluating risk groups. **Results:** According to the evaluations, for two different data sets the kappa coefficients were statistically significant and its degree are intermediate. In terms of both data sets the most convenient and fastest algorithm is farthest clustering algorithm. The results obtained by Make Density Based and EM algorithms gave the most accurate desicions in terms of the distribution of the groups among Framingham risk groups cross tables. **Conclusion:** As a result, with taking into account the criterion of clinical information it is thought that the examination of clustering of risk factors of the disease, will be played an important role for introduction of accurate disease diagnosis. In addition we believe that when considering data distribution and characteristics of data sets clustering algorithms can be used as a diagnostic tool for the plannings and diagnosis of diseases in the field of health.

Key Words: Cluster analysis; data mining; algorithms; cardiovascular diseases; risk factors

doi: 10.5336/medsci.2015-48928

Turkiye Klinikleri J Med Sci 2016;36(1):40-52

Bilgi ve teknolojinin takip edilemez hızda arttığı ve ilerlediği herkes tarafından kabul edilmektedir. Yoğun bilgi kümesinden faydalı ve yararlı sonuçlar elde edebilmek için daha kapsamlı ve daha ileri istatistik yöntemlerin kullanımı nerdeyse zorunlu hale gelmektedir. Teknolojinin özellikle internet sonrası hızla gelişmesi teorisi ortaya atılmış istatistik yöntemlerin uygulama alanına hızla girmesine neden olmuştur. Bu yöntemlerin kullanılması ise karmaşık bilgiyi daha iyi anlamamıza ve gerçek dünyayı daha iyi yorumlamamıza neden olmaktadır.

Çok sayıda kişinin çok sayıda özelliğine ait bilgilerin yer aldığı durumlarda verileri daha iyi değerlendirmek amacıyla geliştirilen yöntemler, "Veri Madenciliği" genel başlığı altında toplanmıştır. Veri Madenciliği başlığı altında yer alan ve özellikle son 10 yılda yoğun kullanım alanı bulan tanı koyma, sınıflama, gruplama (kümeleme) ve tahmin etme amaçlarıyla kullanılan çok sayıda algoritma mevcuttur. Sağlık alanı araştırmalarında da birçok hipotezin temel amacı bunlardan biri veya birkaçıdır. Araştırmacıların kümeleme yöntemleri yardımıyla elde edecekleri bilgiyi nerede ve nasıl kullanılacakları ve sonuçlarını nasıl yorumlayacakları konularında bilgi eksiği olduğunu düşündürmüştür. Ayrıca yaygın kullanılan istatistik paket programlarında var olan birkaç kümeleme algoritmasının dışında özellikle son yıllarda geliştirilmiş yeni kümeleme algoritmalarının literatürde kullanımı sınırlı sayıdadır. Söz konusu eksikliklerden yola çıkarak bu araştırmada teorisi literatüre geçmiş farklı kümeleme algoritmalarının karşılaştırmalı olarak tanıtılması amaçlanmış ayrıca uygulamada yaygın kullanılmamış bazı algoritmaların sağlık alanından elde edilen bir veri seti üzerinde uygulaması yapılarak elde edilen sonuçların nasıl yorumlanacağı gösterilmiştir. Araştırma sonucunda ülkemizde yapılacak bilimsel çalışmalarda kümeleme algoritmalarından ne zaman ve nasıl yararlanılacağı ortaya konmuş olacak ve yeni değerlendirme yöntemleri yardımıyla bilimsel çalışmaların ürettiği bilgilerin kalitesi daha da yükselmiş olacaktır.

GEREÇ VE YÖNTEMLER

VERİ MADENCİLİĞİ VE KÜMELEME ANALİZİ

Veri madenciliği, büyük miktardaki verilerin ayıklanması veya maden araması olarak ifade edilmektedir. Veri madenciliğinin amacı, terabyte boyutunda çok büyük miktardaki verileri işe yarayabilir şekile dönüştürmektir. Veri madenciliği birbirini tekrarlayan aşamalar sonucunda gerçekleşmektedir.[1,2]

Kümeleme benzer nesneleri gruplandırma işlemidir. Bu analiz denetimsiz (unsupervised) bir öğrenme şeklidir. Bu yöntemler sayesinde sadece nesneler değil özellikler de kümelenebilmektedir.[3-5]

Etkili ve doğru bir kümeleme algoritmasının taşıması gereken bazı temel özellikler mevcuttur. Uygun bir kümeleme algoritması veritabanını tek bir seferde tarayarak farklı şekillere sahip ve farklı genişliklerdeki küme yapılarını keşfetmeli ve aynı zamanda niteliksel ve niceliksel olmak üzere tüm veri türlerine uygulanabilir olmalıdır. Etkili bir kümeleme yöntemi, veritabanı büyüklüğü ayırt etmeden büyük ve küçük veritabanlarının her ikisi içinde elverişli olmalıdır. Bu aynı zamanda kümeleme algoritmasının ölçeklenebilirlik özelliğine sahip olup olmadığını göstermektedir. İyi bir kümeleme algoritması etkili ve sapan verilere karşı ne yapması gerektiğini bilmeli ve etkilenmemelidir. Bahsedilen kriterlerin yanında iyi bir kümeleme algoritması uygulanması kolay, yorumlanabilir, fonksiyonel ve anlaşılır olmalıdır.[3-5]

KÜMELEME ALGORİTMALARI

Veri madenciliğinde kullanılan kümeleme algoritmaları, altı temel grup altında incelenir.

1) Hiyerarşik Kümeleme Algoritmaları

2) Yoğunluğa Dayalı Kümeleme Algoritmaları

3) Bölünmeye Dayalı Kümeleme Algoritmaları

4) Izgaraya Dayalı Kümeleme Algoritmaları

5) Kategorik Kümeleme Algoritmaları

6) Olasılık Modellerine Dayalı Kümeleme Algoritmaları

Yukarıda verilen kümeleme algoritmalarının arasında yer alan ve uygulamada kullanılan K-ortalama, Cascade K-ortalama, En uzak ilk, Expectation Maximization (EM), Make Density Based, K-medoid kümeleme algoritmaları sırasıyla aşağıdaki başlıklar altında incelenmektedir.

K-Ortalama Kümeleme Algoritması (K-Means Clustering Algorithm)

En eski kümeleme yöntemlerinden biri olan *K*-ortalamalar yöntemi 1957 yılında ilk kez Hugo Steinhaus'un öne sürdüğü bir fikir olmasına rağmen 1967 yılında J.B. MacQueen tarafından geliştirilmiştir.[6]

K-ortalama yönteminin performansını etkileyen en önemli faktörler başlangıç merkezlerinin seçimidir. *K*-ortalama kümeleme yöntemi sayısal veriler için kullanılmaktadır. Gürültülü ve uç değerlerden aşırı derecede etkilenmektedir. Hesaplama karmaşıklığı *O(tkn)*'dir. Burada *t* iterasyon sayısını, *K* küme sayısını, *n* ise nesne sayısını göstermektedir. Karmaşıklığı diğer yöntemlere göre azdır ve uygulaması kolaydır. *K*-ortalama kümeleme yöntemi büyük veri setleri için başarılı sonuçlar üretebilmektedir. *K*-ortalama algoritması genel olarak küresel kümeleri bulmada daha başarılı bir yöntemdir. Ancak yöntemin dezavantajı küme sayısını belirlemekteki zorluktur. Bunun yanı sıra kümeleme sonucunda bazı kümelerde eleman olmayabilir yani kümeleme sonucunda boş küme oluşabilmektedir ki bu durumda uygun optimizasyon yöntemleri kullanılmalıdır. Kümeleri oluşturmada rol oynayan özelliklerin hangisinin daha çok kümenin oluşmasında etkisi olduğu ise bilinmemektedir.[7,8]

Cascade K-Ortalama Kümeleme Algoritması (Cascade K-Means Clustering Algorithm)

Bilinen K-ortalamalar yönteminin performansını etkileyen en önemli faktörler başlangıç merkezlerinin seçimi ve başlangıçta küme sayısının belirlenmesidir. Başlangıç küme seçimi için Hartigan ve Gong, K-ortalama ++ gibi birçok farklı yöntem söz konusudur. Ancak bu yöntemlerin en büyük dezavantajı K değerine karar verebilmektir. Bu dezavantajı giderebilmek için Cascade K-ortalamalar yöntemi önerilmiştir. Bu yöntem 1974 yılında grup içi kareler toplamı gruplar arası kareler toplamına oranlaması ile elde edilen Calinski Harabas kriterini kullanarak en iyi K değerini belirlemek için kullanılmaktadır.

En Uzak İlk Kümeleme Algoritması (Farthest First Clustering Algorithm)

En uzak ilk kümeleme algoritması Hochbaum ve Scmoys tarafından 1985 yılında geliştirilmiştir. Bu algoritma *K*-ortalama yöntemine benzerdir. Yöntem merkez noktaları seçerek, nesneleri kümelere atamaktadır. Ancak bu işlemi uzaklıkları maksimum yaparak gerçekleştirmektedir. Başlangıç çekirdek noktaları, ortalama değerlerden en büyük uzaklığa sahip değerdir. Başlangıç çekirdekler ve *K* küme sayısı belirlenmesinin ardından kümeleme işlemleri yapılmaktadır.[9,10]

En uzak ilk kümeleme yöntemi *K*-merkezin belirlenmesi problemini çözmektedir ve büyük veri setleri için de oldukça etkili bir kümeleme algoritmasıdır. Algoritma merkezleri hesaplamak için ortalamaları bulmamaktadır. Merkezleri keyfi olarak ele alır ve merkezler arasındaki uzaklıkların maksimum olması sağlanmaktadır. Geliştirilen bu kümeleme algoritması oldukça hızlıdır ve veri madenciliği uygulamalarında büyük ölçekli veriler için de uygundur.[9,10]

K-Medoids Kümeleme Algoritması (K-Medoids Clustering Algorithm)

Yöntem ilk olarak 1987 yılında Kaufman ve Rousseeuw tarafından *K*-ortalama algoritmasının dezavantajlarını gidermek için geliştirilmiştir.[11-13] *K*-ortalama yönteminde veri setinde çok büyük değeri alan nesne, kümenin ortalamasını ve merkezini aşırı derecede etkileyebilmektedir. Ancak *K*-medoids algoritmasında ortalama yerine temsili noktalar kullanıldığı için bu dezavantaj ortadan kaldırılmıştır.

Medoid kelimesi küme içinde en merkeze yerleştirilmiş olan nesne anlamına gelmektedir. *K*-medoid yöntemi veri setindeki nesneler ile temsili nokta arasındaki benzersizliklerin toplamını minimum yapmayı amaçlamaktadır.[14,15] *K*-medoids algoritmasında birinci aşama yapılandırma aşamasıdır. Bu aşama kümelemenin başlangıç aşamasıdır,

k adet temsilci nesne seçilene kadar devam eder. Başlangıç merkezleri rastgele atanabileceği gibi çeşitli işlemler sonucu da belirlenebilir. Algoritmanın ikinci aşaması değiştirme (Swap) aşamasıdır. Bu aşama temsilci nesneleri geliştirerek kümeleme işleminin verimini arttırmak için uygulanır. Her bir (*i,h*) çifti için hesaplama yapılır. *i* seçilmiş, h ise seçilmemiş bir nesnedir. Değişim ihtimallerinin kümelemeye nasıl bir etkisi olduğu incelenerek her bir kombinasyon için kümeleme kalitesi hesaplanır.[16]

EM Kümeleme Algoritması (Expectation Maximization Clustering Algorithm)

EM kümeleme algoritması denetimsiz bir öğrenme yöntemidir. Veri noktalarının yoğunluklarını tahmin etmek için kullanılmaktadır. Bu yöntem birçok alanda farklı şekillerde kullanılmaktadır. EM kümeleme algoritması olasılığa dayalı kümeleme algoritmaları arasında yer almaktadır. Algoritma E ve M olmak üzere iki farklı adımdan oluşmaktadır. E aşamasında, veri setindeki her bir nesne için beklenen küme olasılıkları hesaplanmaktadır. M aşamasında ise, E aşamasında elde edilen parametre tahminleri yeniden tahmin edilmektedir. M aşamasında elde edilen parametre değerleri, takip eden E aşamasında girdi parametresi olarak kullanılmaktadır. E ve M aşamaları durağan bir sonuç elde edilene kadar devam etmektedir. EM kümeleme algoritması yoğun istatistik temellerine sahip bir yöntemdir. Gürültülü verilere karşı dayanıklı ve yüksek boyutlu veriler için de kullanılabilir bir algoritmadır. EM kümeleme yönteminin adımları oldukça basit ve anlaşılması kolaydır. Diğer kümeleme algoritmalarına göre maliyeti daha azdır. Ayrıca kayıp veriler olduğunda bu verileri tahmin edebilme yeteneğine de sahiptir. Yöntem uygulanmadan önce toplam küme sayısı, maksimum tolere edilebilir hata, maksimum iterasyon sayısı gibi parametrelerin belirlenmesi gerekmektedir.

Make Density Based Kümeleme Algoritması

Bu kümeleme algoritmasında kümeler oluşturulurken veri tabanındaki yoğunluk özellikleri dikkate alınmaktadır. Yöntemde kümeler ve dolayısıyla da sınıflar kolayca tanımlanabilmektedir. Çünkü sahip olunan nokta sayısına göre yoğunluk artışı meydana gelmektedir. Veri tabanındaki elementler çekirdek ve sınır nokta olmak üzere iki farklı şekilde sınıflandırılmaktadır.[17]

P(x) olasılık yoğunluk fonksiyonu altında, $x_1, x_2,\ldots,x_n$'ler noktaların bir örnekleridir. *P(x)* kümülatif olasılık fonksiyonu ise aşağıdaki gibi tanımlanmaktadır:[17]

$$P(x)=\sum x_i \le xp\ (x_i)$$

P(x), x'in tahmin fonksiyonudur. *P(x)* değeri, *x* merkezli, *h* genişliğinde bir pencere dikkate alınarak tahmin edilmektedir. *h* parametresinin genişliği yoğunluk tahmininin yayılımını ve düzgünlüğünü göstermektedir. Eğer yayılım çok büyük ise, daha büyük ortalama bir değer elde edilir. Küçük olduğunda ise, pencerede yeteri kadar noktanın olmadığı sonucuna varılır. Pencere içerinde kalan noktalar için ($|z|\le 1/2$),*p(x)* olasılık fonksiyonuna $1/hn$'in katkısı olacaktır. Pencere dışında kalan noktaların ($|z|>1/2$) ise olasılık fonksiyonuna katkısı sıfırdır.[17]

$$K(z)=1/\sqrt{2}exp\ \{-z^2/2\}$$

Yukarıdaki formüldeki $z=x'-x_i/2$'dir.

K(z) eşitliğine göre *x*, dağılımın ortalaması olarak *h* ise dağılımın standart sapması olarak işlem görmektedir.[17]

VERİLER

Çalışmamızda kullanılan veriler, 2012-2014 yılları arasında 3 yıllık periyot içerisinde kilo vermek ve periyodik muayene için Düzce Üniversitesi Tıp Fakültesi Aile Hekimliği Polikliniğine başvuran kişiler üzerinden toplanmıştır. Toplamda 4788 kişide ölçümler yapılmıştır. Kümeleme analizi yapmadan önce verilerde var olan eksik gözlemler için normal dağılmayan değişkenler için medyan değerleri, normal dağılan değişkenler için ortalama, kategorik değişkenler için ise mod değerleri hesaplanarak kayıp veri sorunu ortadan kaldırılmıştır. Ayrıca çalışmaya alınan kişilerin Framingham risk skoru yazılmış bir makro yardımıyla hesaplanmış ve bireyler risk skorlarına göre düşük riskli grup, orta riskli grup ve yüksek riskli grup olarak gruplandırılmıştır.

Framingham risk cetveline göre bireyler, %10'un altında riske sahip ise "düşük riskli grup", %10 ile %20 arasında "orta düzeyde riskli grup" ve %20' nin üzerinde "yüksek riskli grup" olarak sınıflandırılmaktadır. Bu durumda 3 farklı grup ortaya çıkmaktadır. Bu bilgiden hareketle gerçekleştirilen bu tez çalışmasının uygulama bölümünde kullanılan kümeleme yöntemlerine göre 3 kümenin elde edilmesi hedeflenmiştir. Uygulamada iki amaç ortaya konmuştur. Bunlardan birincisi, Framingham risk skoru hesaplanırken kullanılan değişkenler yardımıyla kümeleme analizi yapıldığında kişilerin hangi gruba düştüğünün belirlenmesi, ikincisi ise literatürde önerilen ve klinik pratikte önemsenen ilave risk faktörlerini de dikkate alarak yeniden kümeleme analizi yapıldığında bireylerin hangi risk grubuna düştüğünün belirlenmesidir. Bu iki amacı gerçekleştirmek için farklı kümeleme algoritmaları kullanılmış ve ortaya çıkan kümelerin Framingham risk skoruna göre belirlenen gruplarla uyumu incelenmiştir. Ayrıca Framingham risk skoruna göre risk grupları oluşturulurken kadın ve erkekler için ayrı ayrı ve aile öyküsü olan ve olmayanlar için ayrı ayrı hesaplamalar yapılmaktadır. Bu çalışmada ise kümeleme algoritmalarında cinsiyet ve aile öyküsü de bir risk faktörü olarak modelde yer almıştır.

Uygulamada öne sürülen birinci ve ikinci amacı gerçekleştirmek için kümeleme analizlerinde kullanılan risk faktörleri Tablo 1'de topluca verilmiştir (Tablo 1).

UYGULAMADA KULLANILAN KÜMELEME ALGORİTMALARI VE PAKET PROGRAMLAR

Değerlendirmelerin ilk aşamasında veri setinde yer alan sayısal değişkenlere ait tanımlayıcı değerler ortalama ve standart sapma olarak, kategorik yapıdaki değişkenlere ait tanımlayıcı değerler ise sayı ve yüzde olarak hesaplandı ve tablolar halinde verildi. Framingham risk skoru hesaplanırken, yaş, cinsiyet, sigara kullanma durumu, sistolik kan basıncı, total kolesterol, HDL ve tansiyon ilacı kullanma durumuna ait veriler kullanıldı. Uygulamadaki amacı gerçekleştirmek için bu değişkenler

TABLO 1: Kümeleme analizlerinde kullanılan risk faktörleri.

Risk faktörü numarası	Birinci amacı gerçekleştirmede kullanılan risk faktörleri	İkinci amacı gerçekleştirmede kullanılan risk faktörleri
1	Yaş (sürekli değişken)	Yaş (sürekli değişken)
2	Cinsiyet (kadın-erkek)	Cinsiyet (kadın-erkek)
3	Sigara içme durumu (içmiyor-içiyor)	Sigara içme durumu (içmiyor-içiyor)
4	Aile öyküsü (yok-var)	Aile öyküsü (yok-var)
5	Sistolik kan basıncı (sürekli değişken)	Sistolik kan basıncı (sürekli değişken)
6	Tansiyon ilacı kullanma durumu (evet-hayır)	Tansiyon ilacı kullanma durumu (evet-hayır)
7	HDL (sürekli)	HDL (sürekli)
8	Total kolesterol (sürekli)	Total kolesterol (sürekli)
9		Diyastolik kan basıncı (sürekli değişken)
10		Kalça çevresi (sürekli)
11		Vücut kütle İndeksi (sürekli)
12		İç yağlanma değeri (sürekli)
13		Trigliserid (sürekli)
14		Ürik asit (sürekli)
15		Bel çevresi (sürekli)
16		Açlık kan şekeri (sürekli)
17		HOMA (IR) (sürekli)
18		Viseral şişmanlık indeksi (sürekli)
19		Vücut Şİşmanlık İndeksi (sürekli)
20		ATPIII (negatif, pozitif)

dikkate alınarak 6 farklı kümeleme algoritması yardımıyla Framingham risk skoruna göre 3 farklı risk grubu elde edildiği için, 3 farklı küme elde edildi. Ardından bu değişkenlere aile öyküsü de eklenerek yeniden kümeler oluşturuldu. Kümelemede, *K*-ortalama kümeleme yöntemi, başlangıç küme seçiminde daha iyi sonuçlar verdiği, daha kısa sürede sonuca ulaşılabildiği ve modelin küme içi hata kareler toplamını düşürdüğü için tercih edildi. Uzaklık ölçüsü olarak Öklid uzaklığı kullanıldı ve küme sayısı başlangıçta 3 olarak belirlendi. Öklid uzaklığının kullanılmasının nedeni aşırı değerlerden kaynaklanabilecek olan olumsuzluklardan etkilenmemesidir. Çekirdek değeri olarak ise farklı değerlerin uygulanması sonucunda küme içi hata kareler toplamını en düşük yapan değer olan 35 seçildi. Verilerdeki ölçüm farklılıklarının etkilerini ortadan kaldırmak için gözlem değerlerinden aritmetik ortalamaları çıkartılıp bu farka ait standart sapmaya bölünerek standartlaştırıldıktan sonra *K*-ortalama kümeleme yöntemi uygulandı. Ardından veri setine EM kümeleme, *En uzak ilk* (Farthest First) kümeleme, yoğunluk (density) kümeleme, *K*-medoid ve cascade *K*-ortalama kümeleme yöntemleri uygulandı. K-ortalama yöntemi sadece sayısal veriler için uygulanabilen bir yöntem olmasına rağmen çok değişkenli olarak uzaklık ölçülerinin normal dağılması sebebiyle yine bu algoritmadan yararlanılabilmektedir. Ayrıca veri setinde oldukça az sayıda kategorik veriler mevcuttur. Uygulamanın ikinci amacında, risk gruplarının daha iyi tahmin edilmesini sağlamak hedeflendiği için modele yeni değişkenler eklendi. Her iki amaç sonrasında oluşturulan kümeler arasında anlamlı düzeyde ayrım yapan değişkenleri belirlemek için, sayısal değişkenler için tek yönlü ANOVA kullanıldı. ANOVA sonrasında anlamlı düzeyde farklı kümeleri belirlemek için Tukey çoklu karşılaştırma testinden yararlanıldı. Benzer amaçla kategorik yapıdaki cinsiyet, tansiyon ilacı kullanma, ailede kalp rahatsızlığı olma ve sigara içme durumları ile oluşturulan kümeler arasındaki ilişkiler Pearson Ki-kare testi ile incelendi. Son olarak, Framingham risk skoruna göre belirlenen risk grupları ile kümeleme algoritmalarından elde edilen kümeler arasındaki uyum Kappa istatistiği ile değerlendirildi. İstatistiksel anlamlılık düzeyi 0,05 olarak alındı ve analizlerde WEKA (ver. 3,4,11), Rapid Miner (ver. 6,4) ve SPSS (ver. 18) paket programlarından yararlanıldı.

BULGULAR

Çalışmaya katılan kişilerin %83,5'ü (n=3998) kadın, %31,3'ü (n=1501) sigara tüketmekte, %12,1'i (n=579 kişi) tansiyon ilacı kullanmakta olup, %42,6'sının ise (n=2039 kişi) ATPIII değeri pozitifti. Yaş, cinsiyet, sigara içme durumu, sistolik kan basıncı, total kolesterol, HDL, tansiyon ilacı kullanma durumlarının değerlendirilmesi sonucunda elde edilen Framingham risk skoru sonucunda kişilerin %4,2' sinde (n=203) yüksek risk tespit edildi. Aile öyküsü de dikkate alındığında bu oran %4,4' e (n=212) yükseldi.

Veri setindeki yaş, cinsiyet, sigara kullanma durumu, sistolik kan basıncı, total kolesterol, HDL ve tansiyon ilacı kullanma durumu değişkenleri dikkate alınarak *K*-ortalama, cascade *K*-ortalama, *en uzak ilk*, EM, density ve *K*-medoid kümeleme yöntemlerinin uygulanması sonucunda iterasyonlar 5 adımda sonlandırılmış yani 5 adımda uygun kümeler tespit edilmiştir. Altı farklı kümeleme algoritması sonucunda elde edilen küme özellikleri değerlendirildiğinde küme 1; kardiyovasküler riski düşük olan bireyleri, küme 2; orta riskli olanları ve küme 3; yüksek riskli olanları ifade ettiği görüldü. Küme sonuçları incelendiğinde ise EM kümeleme algoritmasının riskli bireyleri bulmaya karşı daha eğilimli olduğu, en uzak ilk yönteminin ise kardiyovasküler riski düşük olan bireyleri bulma eğiliminde olduğu belirlendi. Ancak genel olarak *K*-ortalama, cascade *K*-ortalama, *en uzak ilk* ve density kümeleme algoritmalarının sonucunda kümelere yerleşen bireylerin benzer oranda olduğu gözlendi.

Kümeleme işlemlerine yaş, cinsiyet, sigara kullanma durumu, sistolik kan basıncı, total kolesterol, HDL, tansiyon ilacı kullanma durumuna ilaveten aile öyküsü değişkeni de ilave edilerek *K*-ortalama, cascade *K*-ortalama, *en uzak ilk*, EM, density ve *K*-medoid algoritmalarının uygulan-

ması sonucunda elde edilen kümeleme sonuçları aşağıdaki gibi elde edildi.

Aile öyküsünün olup olmama durumu da dikkate alındığında, density haricinde kalan diğer kümeleme algoritmalarının sonuçları aile öyküsü alınmadan elde edilen kümeleme sonuçları ile benzer çıktığı ancak aile öyküsü modele alındığında density kümeleme algoritmasında kardiyovasküler riski düşük olan bireylerin birçoğunun orta riskli veya çok riskli gruplara dâhil edildiği görüldü.

Tablo 1' de yer alan tüm risk faktörleri dikkate alınarak *K*-ortalama, cascade *K*-ortalama, *en uzak ilk*, EM, density ve *K*-medoid algoritmalarının uygulanması sonucunda elde edilen kümeleme sonuçları değerlendirildiğinde, Cascade *K*-ortalama yöntemine göre yüksek riskli kişilerin oranı (%17), diğer kümeleme yöntemlerinin yüksek riskli olarak sınıflandırdığı bireylerden daha fazla bulundu. *En uzak ilk* kümeleme algoritmasının sonuçları değerlendirildiğinde ise kardiyovasküler riski düşük olan bireylerin oranının, diğer kümeleme algoritmalarının sonuçlarına göre daha fazla olduğu görüldü.

Framingham risk skoru hesaplanırken kullanılan sayısal yapıdaki değişkenler dikkate alınarak (aile öyküsü dâhil) farklı kümeleme algoritmaları sonucunda elde edilen kardiyovasküler riski düşük, orta, yüksek grupları arasında, söz konusu sayısal değişkenlerin ortalamalarının karşılaştırılması sonucunda elde edilen tanımlayıcı istatistikler ve *p* değerleri Tablo 2' de verildi. Uygulamada kullanılan 6 farklı kümeleme algoritması yardımıyla elde edilen kümeler arasında yaş, total kolesterol ve HDL ortalamaları bakımından anlamlı fark bulundu (her bir karşılaştırma için $p<0,001$). Ancak *K*-ortalama kümeleme yöntemi sonucunda elde edilen kümeler arasında sistolik kan basıncı ortalaması bakımından anlamlı fark yok iken ($p=0,441$), cascade *K*-ortalama, *en uzak ilk*, EM, Density ve *K*-medoid algoritmaları sonucunda anlamlı farklılık bulundu (her biri için $p<0,001$). Anlamlı düzeyde farklı olan kümeler,

TABLO 2: Aile öyküsü değişkeni modele alınarak elde edilen kümelerde sayısal değişkenlerin tanımlayıcı değerleri ve kümelerin karşılaştırma sonuçları.

Kümeleme		Yaş		Sistolik Kan Basıncı		Total Kolesterol		HDL	
Algoritmaları	Kümeler	Ort±SD*	*p*	Ort±SD	*p*	Ort±SD	*p*	Ort±SD	*p*
K-Ortalama	Düşük Risk	37,06^{a}±12,556		124,80±18,225		190,48^{b}±39,433		52,94^{a}±12,093	
	Orta Risk	39,27^{b}±11,457	<0,001	124,56±17,135	0,441	197,35^{a}±40,860	<0,001	51,10^{b}±12,633	<0,001
	Yüksek Risk	34,59^{c}±10,449		125,66±14,965		193,43^{ab}±39,462		42,15^{c}±9,538	
Cascade	Düşük Risk	36,61^{a}±2,474		124,65^{a}±18,028		190,06^{a}±39,181		52,79^{a}±11,963	
K-Ortalama	Orta Risk	37,58^{b}±10,730	<0,001	123,34^{a}±16,482	<0,001	195,51^{b}±41,245	<0,001	51,38^{b}±12,903	<0,001
	Yüksek Risk	40,54^{c}±11,988		128,42^{b}±16,550		200,72^{c}±39,879		42,23^{c}±10,034	
En Uzak İlk	Düşük Risk	35,36^{a}±10,849		23,88^{b}±16,676		190,56^{b}±39,551		50,76^{b}±12,266	
	Orta Risk	42,33^{b}±13,696	<0,001	127,23^{a}±19,314	<0,001	197,77^{a}±40,452	<0,001	52,01^{a}±12,883	<0,001
	Yüksek Risk	38,43^{c}±10,158		126,26^{ab}±19,568		198,12^{ab}±39,316		50,08^{ab}±12,065	
EM	Düşük Risk	27,49^{a}±5,904		114,61^{a}±11,170		170,45^{a}±29,665		53,65^{a}±12,249	
	Orta Risk	39,00^{b} 8,865	<0,001	127,13^{b}±10,671	<0,001	193,87^{b}±30,924	<0,001	42,44^{b}±7,043	<0,001
	Yüksek Risk	49,54^{c}±9,401		136,90^{c}±21,878		222,97^{c}±40,918		56,79^{c}±12,374	
Density	Düşük Risk	27,49^{a}±5,904		114,61^{a}±11,170		170,45^{a}±29,665		53,65^{a}±12,249	
	Orta Risk	39,00^{b}±8,865	<0,001	127,13^{b}±10,671	<0,001	193,87^{b}±30,924	<0,001	42,44^{b}±7,043	<0,001
	Yüksek Risk	49,54^{c}±9,401		136,90^{c}±21,878		222,97^{c}±40,918		56,79^{c}±12,374	
K-Medoid	Düşük Risk	29,57^{a}±8,925		115,50^{a}±12,970		143,70^{a}±17,421		49,78^{a}±11,666	
	Orta Risk	37,73^{b}±11,585	<0,001	126,97^{b}±17,668	<0,001	191,79^{b}±18,442	<0,001	49,24^{a}±11,195	<0,001
	Yüksek Risk	44,85^{c}±11,581		128,87^{c}±18,065		250,21^{c}±28,868		58,19^{b}±14,256	

*Tabloda ortalamaların üzerinde ifade edilen harfler çoklu karşılaştırma sonuçlarına yöneliktir. Farklı harfler gruplar arasında anlamlı farklılık olduğunu ifade ederken,aynı harfler gruplar arasında anlamlı farklılık olmadığını göstermektedir.

Tablo 2'de ortalamaların yanına yerleştirilen tamamen farklı harflerle gösterildi (Tablo 2).

Framingham risk skoru hesaplanırken kullanılan kategorik yapıdaki değişkenlerin kategorilerinin, oluşturulan kümelerdeki dağılımı ve kümelerin bu değişkenler bakımından karşılaştırıldığında tüm kategorik değişkenlerin dağılımları (k-medoid yönteminde kullanılan *aile öyküsü* değişkeni haricindeki) bakımından kümeler arasında anlamlı farklılıklar gözlendi (her bir karşılaştırma için $p<0,05$).

Uygulamadaki ikinci amacı değerlendirmek için literatürde kardiyovasküler hastalıkların risk faktörü olarak tanımlanan ve klinik pratikte önemsenen ilave risk faktörleri de dikkate alındı ve altı farklı kümeleme algoritması yardımıyla risk grupları yeniden oluşturuldu. Bu amaç için dikkate alınan risk faktörleri Tablo 1'in üçüncü sütununda topluca verildi. Kümeleme işlemleri sonrasında ortaya çıkan kümelerde dikkate alınan sayısal yapıdaki risk faktörlerinin tanımlayıcı değerleri ve kümelerin bu risk faktörleri bakımından karşılaştırma sonuçları değerlendirildiğinde tüm sayısal değişkenlerin ortalamaları bakımından (HOMA hariç) kümeleme sonucunda oluşturulan risk grupları arasında anlamlı farklılıklar tespit edildi (her biri için $p<0,001$). Uygulanan kümeleme algoritmaları sonucunda kardiyovasküler risk gruplarında HOMA ortalamaları benzer bulunmasına rağmen literarürde HOMA değişkeninin kardiyovasküler hastalıklar için bir risk faktörü olduğu bilgisi ile sıklıkla karşılaşılmaktadır.[15] Kategorik değişkenler ile (ikinci modeldeki) kümeleme sonucunda elde edilen risk grupları arasındaki ilişkiler ise birinci model ile benzer bulunmuştur.

Framingham skoru hesaplanırken kullanılan risk faktörleri ile birlikte aile öyküsü de modele alınarak elde edilen kümeleme sonuçlarının birbirleriyle uyumları Tablo 3'de topluca verilmiştir. Tablo 3 değerlendirildiğinde, her ne kadar uygulamada kullanılan 6 algoritmanın kümeleri arasında anlamlı uyum gözlense de çalışmadaki toplam birey sayısının büyük olması, kappa değeri küçük olan bazı karşılaştırma sonuçlarının uyumlarının istatistik olarak anlamlı çıkmasına neden olmuştur. Kappa değeri büyük olan uyumlar incelendiğinde, *K*-ortalama kümeleri, Cascade *K*-ortalama kümeleri ve *En Uzak İlk* kümeleri arasında oldukça yüksek uyum olduğu, Density kümeleri ile EM kümeleri ve *K*-Medoid kümeleri arasında %100 uyum sağlandığı ve EM ile *K*-Medoid arasında ise orta düzeyde uyumun var olduğu gözlenmiştir (Tablo 3).

Risk faktörü olarak düşünülen değişkenlerin tamamı dikkate alınarak uygulanan kümeleme algoritmaları yardımıyla elde edilen kümelerin uyumları değerlendirilmiş ve elde edilen sonuçlar Tablo 4' de topluca verilmiştir. Tablo 4 incelendiğinde, tüm kümeleme algoritmaları arasındaki uyumu ölçen Kappa katsayısının istatistiki olarak anlamlı olduğu sonucuna varılmıştır. Uyumlar daha detaylı olarak incelendiğinde, en yüksek uyumun EM ve Density kümeleme algoritmaları ara-

TABLO 3: Framingham skoru hesaplanırken kullanılan risk faktörlerine ilaveten aile öyküsü de dikkate alınarak elde edilen kümeleme sonuçlarının birbirleriyle uyumları.

		Cascade K-ortalama	En Uzak İlk	EM	Density	K-Medoid
K-ortalama	KAPPA	0,860	0,564	0,063	0,063	0,015
	p	**<0,001**	**<0,001**	**<0,001**	**<0,001**	**0,05**
Cascade K-ortalama	KAPPA		0,608	0,053	0,053	0,024
	p		**<0,001**	**<0,001**	**<0,001**	**<0,001**
En Uzak İlk	KAPPA			0,031	0,031	0,023
	p			**0,001**	**<0,001**	**0,003**
EM	KAPPA				1.000	0,334
	p				<0,001	**<0,001**
Density	KAPPA					1.000
	p					**<0,001**

TABLO 4: Değişkenlerin tamamı nıiçeren kümelerin birbirleriyle uyumları

		Cascade K-ortalama	En Uzak İlk	EM	Density	K-Medoid
K-ortalama	KAPPA	0,495	0,252	0,252	0,248	0,175
	p	**<0,001**	**<0,001**	**<0,001**	**<0,001**	**<0,001**
Cascade K-ortalama	KAPPA		0,285	0,195	0,169	0,225
	p		**<0,001**	**<0,001**	**<0,001**	**<0,001**
En Uzak İlk	KAPPA			0,374	0,333	0,147
	p			**<0,001**	**<0,001**	**<0,001**
EM	KAPPA				0,894	0,043
	p				**<0,001**	**<0,001**
Density	KAPPA					0,026
	p					**0,015**

TABLO 5: Framingham skoruna göre oluşturulan risk grupları ile Framingham değişkenleri ile birlikte aile öyküsü de dikkate alınarak yapılan kümeleme analizleri sonucunda elde edilen kümeler arasındaki uyumlar.

	K-ortalama	Cascade K-ortalama	En Uzak İlk	EM	Density	K-Medoid
Kappa	0,128	0,177	0,206	0,146	0,146	0,054
p	**<0,001**	**<0,001**	**<0,001**	**<0,001**	**<0,001**	**<0,001**

sında, en düşük uyumun ise Density ve *K*-medoid kümeleme algoritmaları arasında olduğu görülmüştür. Ayrıca *K*-ortalama, Cascade *K*-Ortalama, *En Uzak ilk* ve Density yöntemlerinin sonuçları birbirleriyle anlamlı derecede uyumlu olup uyumun derecesi ise orta düzeydedir (Tablo 4).

Verileri en iyi kümeleyen algoritmaları bulabilmek amacıyla ilk olarak Kappa katsayıları incelendi. Ardından Framingham skoruna göre oluşturulan risk grupları içinde düşük riskli olarak belirlenen bireylerin, kümeleme analizleri sonucunda düşük veya orta riskli bulunması önemsendi ve Framingham skoruna göre riskli olan bireylerin ise kümeleme analizlerinde riskli çıkma oranının yüksek olması tercih edildi. Ayrıca Framingham skoruna göre orta riskli bulunan bireylerin, kümeleme analizleri sonucunda orta riskli veya yüksek riskli olması durumu tercih edildi.

Yapılan değerlendirmeler sonucunda, aile öyküsü dikkate alınarak hesaplanan Framingham risk grupları ile aile öyküsü dikkate alınarak hesaplanan kümeleme grupları arasındaki en yüksek uyuma sahip kümeleme algoritmasının 0,206 ile *en uzak ilk* yönteminin olduğu belirlenmiştir. Risk şüphesi taşıyan değişkenlerin tamamı dikkate alınarak oluşturulan kümeler ile Framingham skoruna gören elde edilen risk grupları arasındaki uyum ise en yüksek 0,325 ile yine *en uzak ilk* yönteminin olduğu belirlenmiştir. Hesaplanan uyum katsayıları istatistik olarak anlamlı ancak orta derecede bir uyumu göstermektedir. Framingham skoruna göre oluşturulan risk grupları ile Framingham değişkenleri ile birlikte aile öyküsü de dikkate alınarak yapılan kümeleme analizleri sonucunda elde edilen kümeler arasındaki uyumlar incelendiğinde Tablo 5'deki sonuçlara ulaşıldı. Tablo 5 değerlendirildiğinde, kümelerle Framingham risk grupları arasında düşük veya orta düzeye yakın ancak istatistik olarak anlamlı uyumun elde edildiği görüldü (Tablo 5).

Framingham skoruna göre oluşturulan risk grupları ile tüm değişkenler dikkate alınarak yapılan kümeleme analizleri sonucunda elde edilen kümeler arasındaki uyumlar incelendiğinde Tablo 6'daki sonuçlara ulaşıldı. Kümelerle Framingham risk grupları arasında orta düzeye yakın ancak istatistik olarak anlamlı uyumun elde edildiği görüldü (Tablo 6).

TABLO 6: Framingham skoruna göre oluşturulan risk grupları ile tüm değişkenler dikkate alınarak yapılan kümeleme analizleri sonucunda elde edilen kümeler arasındaki uyumlar.

	K-ortalama	Cascade K-ortalama	En Uzak İlk	EM	Density	K-Medoid
Kappa	0,249	0,198	0,325	0,177	0,152	0,165
p	**<0,001**	**<0,001**	**<0,001**	**<0,001**	**<0,001**	**<0,001**

Tablo 5 ve Tablo 6 'daki Kappa değerleri incelendiğinde Framingham risk grupları ile en yüksek uyumun *en uzak ilk* yönteminden elde edildiği görüldü. Framingham risk grupları ve Framingham risk faktörleri (aile öyküsü dâhil) kullanıldığında farklı algoritmalardan elde edilen kümeler arasında çapraz tablolar oluşturularak grupların dağılımı incelendi. Sonuç olarak Framingham risk faktörleri (aile öyküsü dâhil) kullanıldığında en isabetli kararların EM algoritmasına ait kümelerden elde edilebileceği kanaatine varıldı. Framingham skoruna göre düşük riskli olan bireylerin %20,9'u EM algoritması sonucunda yüksek riskli olarak bulundu ve Framingham skoruna göre yüksek riskli olarak adlandırılan bireylerin ise %71,7' si EM algoritması sonucunda da yüksek riskli olarak tespit edildi. Framingham skoruna göre yüksek riskli olarak adlandırılan bireylerin tamamı EM algoritması yardımıyla ya orta veya yüksek riskli olarak belirlendi. Bunlara ilaveten, Framingham skoruna göre orta düzeyde riskli bulunan bireylerin, EM algoritması yardımıyla %33,6'sı orta riskli bulundu.

Framingham risk grupları ve tüm değişkenler dikkate alınarak oluşturulan kümeler arasında çapraz tablolar oluşturularak grupların dağılımı incelendiğinde ise, en isabetli kararların Density algoritmasına ait kümelerden elde edildiği görüldü. Framingham skoruna göre düşük riskli olan bireylerin %12,9'u Density algoritması sonucunda yüksek riskli olarak bulundu ve Framingham skoruna göre yüksek riskli olarak adlandırılan bireylerin ise %54,7' si Density algoritması sonucunda da yüksek riskli, %17' si düşük ve %28,3' ü ise orta riskli olarak tespit edildi. Ayrıca Framingham skoruna göre orta düzeyde riskli bulunan bireylerin, Density algoritması yardımıyla %33,6'sı orta riskli bulundu.

Her iki veri setinde de Framingham skoruna göre düşük riskli bulunan bireylerin en fazla oranda *en uzak ilk* kümelemede de düşük riskli olduğu görüldü.

TARTIŞMA VE SONUÇ

Gerçekleştirilen çalışma ile bağlantılı olarak, Zheng ve ark. 2005 yılında EM, *En uzak ilk* ve *K*-ortalama kümeleme algoritmalarını bir veri setinde uygulamalı olarak karşılaştırmışlardır. EM algoritmasının hangi kritere bakılırsa bakılsın, *K*-ortalama ve *En uzak ilk* algoritmalarından daha üstün olduğu sonucuna varmışlardır. EM'nin tüm veri setleri için *K*-ortalama ve *En uzak ilk* yönteminden daha küçük standart sapmaya sahip olduklarını tespit etmişler ve bu durumun EM algoritmasının diğer iki yöntemden daha durağan sonuçlar verdiğini ifade ettiğini belirtmişlerdir.[18]

Abbas 2008 yılında farklı kümeleme algoritmalarını karşılaştırmış, *K*-ortalama ve EM algoritmalarının performansının hiyerarşik kümeleme yöntemlerinden daha iyi olduğu sonucuna varmıştır. Ayrıca büyük veri tabanları için EM ve *K*-ortalama algoritmalarının oldukça iyi sonuçlar üretebileceklerini vurgulamıştır. Ancak *K*-ortalama ve EM algoritmalarının gürültülü veriler için oldukça hassas olduğu belirlenmiştir.[19]

Madhulatha 2011 yılında yaptığı çalışmada *K*-ortalama ve *K*-medoid kümeleme algoritmalarını karşılaştırmış ve küçük veri setleri için *K*-ortalama yönteminin kaliteli ve hızlı sonuçlar ürettiğini gözlemlemiştir. *K*-medoid yönteminin ise daha büyük veri setleri için kullanabilir olduğunu ifade etmişlerdir.[20]

Singh ve Chauhan 2011 yılında en çok kullanılan kümeleme algoritmaları arasında yer alan *K*-ortalama ve *K*-medoid yöntemlerini uygulamalı olarak karşılaştırmışlardır. Araştırma sonucunda, her iki yöntemin de küresel şekile sahip küçük veya orta hacimdeki veri setleri için iyi sonuçlar üretti-

ğini gözlemlemişler; her iki kümeleme algoritması için de başlangıçta *K* değerinin doğru bir şekilde belirlenmesi gerektiğini belirtmişlerdir. *K*-ortalama yönteminin hesaplama maliyetini daha düşük bulmuşlar ancak gürültülü ve aşırı uç değerlere karşı oldukça hassas olduğunu, *K*-medoid yönteminin ise hassas olmadığını görmüşlerdir.[21]

Shrivastava ve Arya tarafından 2012 yılında *K*-ortalama, yoğunluk yöntemi ve *En uzak ilk* kümeleme algoritmaları uygun bir veri seti üzerinde uygulanarak kümeleme yöntemlerinin performansları yanlış sınıflandırılan örneklerin yüzdeleri ile ölçülmüştür. *En uzak ilk* kümeleme algoritması, *K*-ortalama ve yoğunluk yöntemlerine göre daha iyi sonuçlar vermiştir. *En uzak ilk* yöntemin küme sayısından bağımsız olduğunu ancak *K*-ortalama yönteminin küme sayısına oldukça bağımlı sonuçlar ürettiğini gözlemlemişlerdir. *En uzak ilk* kümeleme algoritması hızlı sonuç üretmektedir.[22]

Sharma ve ark. 2012 yılında WEKA'da kullanılan algoritmaları karşılaştırmışlar ve EM kümeleme algoritmasının gerçek veri setleri için oldukça kullanışlı olduğu sonucuna varmışlardır. *En uzak ilk* kümeleme algoritmasının ise büyük ölçekli veritabanları için kullanışlı olduğunu öne sürmüşlerdir. *K*-ortalama yöntemi bakımından değerlendirildiklerinde ise yöntemin fazla miktarda değişkenler içeren veri setleri için hiyerarşik kümelemeye göre daha hızlı sonuçlar verdiğini belirtmişlerdir. Ayrıca *K*-ortalama yöntemi daha sıkı yapıda kümeleri oluşturmaktadır. Yöntemin dezavantajının ise, kümeleme sonucunda oluşturulan kümelerin kalitelerinin karşılaştırılmasının zor olduğunu öne sürmüşlerdir.[23]

Godara ve Yadav 2013 yılında gerçekleştirdikleri çalışmada *K*-ortalama, hiyerarşik ve density kümeleme yöntemlerini incelemişlerdir. Araştırma sonucunda zaman ve doğruluk kriterlerinin her ikisi de birlikte değerlendirildiğinde, bahsedilen üç algoritma içerisinde en iyi yöntemin *K*-ortalama kümeleme algoritması olduğu sonucuna varmışlardır.[24]

Singh ve Dubey 2013 yılında *K*-ortalama ve *En uzak ilk* algoritmalarını karşılaştırmışlardır. Yaptıkları çalışma sonucunda doğruluk oranlarının *K*-ortalama yöntemine göre daha yüksek olduğunu gözlemlemişler ve daha hızlı sonuçlar almışlardır.[25]

Revathi ve Nalini 2013 yılında, çalışmamızda kullandığımız *K*-ortalama, *En uzak ilk* ve make density based algoritmalarını karşılaştırmışlardır. Tüm algoritmalarda küme sayısı arttıkça kümeleri oluşturmada da harcanan zamanın arttığını gözlemlemişler, bunun yanı sıra *En uzak ilk* yöntemindeki uygulamaların çok kısa bir zaman sürdüğü, *K*-ortalamanın ise en uzun zamanda kümeleme sonuçlarını verdiği sonucuna varmışlardır. Böylece *K*-ortalama yönteminin çok büyük veri setleri için kullanımının uygun olmayacağını önermektedirler.[26]

Balabantaray ve ark. 2015 yılında *K*-ortalama ve *K*-medoid yöntemlerini karşılaştırmışlardır. Karşılaştırmalar sonucunda *K*-ortalama yönteminin *K*-medoid kümeleme algoritmasından daha iyi sonuçlar verdiğini görmüşlerdir. Benzerlik ölçüsünü geliştirdiği için büyük veri setlerinde *K*-ortalama yönteminin daha iyi sonuç ürettiğini belirtmişlerdir.[27]

Kakkar ve Parashar 2014 yılında WEKA'da kullanılan *K*-ortalama, hiyerarşik yöntem, EM ve yoğunluğa dayalı kümeleme algoritmalarını karşılaştırarak en iyi yöntemin *K*-ortalama olduğu sonucuna varmışlardır.[28]

Jung ve ark. 2014 yılında *K*-ortalama ve EM kümeleme algoritmalarını karşılaştırmışlardır. Yaptıkları çalışma sonucunda *K*-ortalama yönteminin EM algoritmasına göre doğruluğunun daha fazla olduğunu tespit etmişlerdir. Ancak *K*-ortalama yönteminin EM yöntemine göre daha fazla zaman aldığını belirlemişlerdir.[29]

Goyal 2014 yılında WEKA'da kullanılan COBWEB, DBSCAN, EM, *En uzak ilk* ve *K*-ortalama kümeleme algoritmaları veri setleri üzerinde uygulayarak en iyi yöntemlerin EM ve *K*-ortalama olduğu sonucuna varmıştır.[30]

Çalışmamızda kullandığımız Framingham risk skoru birçok uluslararası ve ulusal çalışmada kardiyovasküler hastalık geçirme riskinin tahmininde kullanılmaktadır. Bu skor 1200 üzerindeki

makalede kardiyovasküler hastalığın dönüm noktasının bulunmasında kullanılmıştır. Ancak bu risk skoru hesaplanırken modelde sınırlı sayıda risk faktörü yer almaktadır. Bu faktörler; yaş, cinsiyet, total kolesterol seviyesi, sistolik kan basıncı, sigara kullanma ve diyabet olma durumlarını içerir.

Biz de bu doğrultuda en çok ölüme sebebiyet veren hastalıklar arasında yer alan kardiyovasküler hastalıkların risk faktörlerinden hareketle kişileri düşük riskli, orta riskli ve yüksek riskli olmak üzere üç kümeye ayırarak, Framingham skorundan elde edilen skor grupları ile farklı kümeleme algoritmaları sonucunda elde edilen kümeleri karşılaştırdık. Ancak çalışmamızda bazı kısıtlılıklar mevcuttur. Literatürde birçok kümeleme algoritması olmasına rağmen yalnızca altı yöntem gerçek veri seti üzerinde tartışılmıştır. Kullanılan yöntemler WEKA ve Rapid Miner programları içerisinde yer alan yöntemlerdir.

Veri setimizin büyük olması ve kategorik değişkenlerin az miktarda olması sebebiyle *K*-medoid yönteminin, *K*-ortalama yönteminden daha kullanabilir olduğu belirlendi. *K*-medoid yönteminin kappa uyumunun düşük olmasına rağmen, *K*-ortalama yönteminin büyük veri setlerinde kullanılamaması ve sayısal değişkenler için uygun olması sebebiyle, *K*-ortalama yöntemi veri setimiz için güvenilir sonuçlar üretmemesi sebebiyle *K*-medoid yönteminin kullanımının daha uygun olacağı belirlendi. Dolayısıyla *K*-ortalama algoritmasının kısıtlılıkları ve uygulama alanları dikkate alındığında çalışmamızda kullanılan veri seti için nispeten uygun olmayan bir kümeleme yöntemi olacağı kanaatine varılmıştır. Başlangıçta *K* değeri hakkında bilgi sahibi olmamız sebebiyle cascade *K*-ortalama yönteminin avantajlarından yararlanılamamıştır. Benzer olarak bu yöntemde *K*-ortalama algoritması gibi uygulanan veri seti için nispeten diğer yöntemlere göre daha az uygundur. Framingham risk skoru ile en uyumlu olan algoritma ise *En uzak ilk* algoritmasıdır. Algoritma büyük veri setleri için oldukça hızlı ve etkili sonuçlar üretmektedir. Aynı zamanda kullanılan veri seti için de en hızlı sonucu veren algoritmanın *En uzak ilk* kümeleme algoritması olduğu belirlendi. Olasılıklar değerlendirildiğinde ise en iyi yöntemlerin EM ve Make Density based yöntemleri olduğu görüldü. Bu yöntemler benzer olarak büyük veri setleri için uygundur. EM algoritması uygulama ve anlaşılma bakımından oldukça basittir. Maliyeti ve zaman kaybı oldukça azdır. Aynı zamanda veride kayıp nesneler olması durumunda bile bu verileri tahmin edebilme özelliğine sahiptir. Bu bilgilere ek olarak Make Density based kümeleme algoritmasının yüksek riskli bireyleri bulma eğilimine sahip olduğu görüldü. Bir başka ifadeyle gerçekte Framingham skoru sonucunda riskli olan bireylerin kümeleme algoritmalarının uygulanması sonucunda da gerçekten riskli olma oranı en yüksek Make density based kümeleme algoritması olduğu belirlendi. Her iki veri setinde de düşük riskli bireyleri bulmaya en elverişli yöntemin ise en uzak komşu algoritması olduğu tespit edilmiştir. Çünkü gerçekte Framingham skoru sonucunda düşük riskli olanların, en uzak komşu yöntemi sonucunda da yüksek bir oranın düşük riskli olduğu görülmüştür.

Ancak en iyi yöntem olarak belirlenen bu yöntemler klinik olarak değerlendirilmelidir. Tüm kümeleme algoritmaların sonuçları teker teker irdelenerek bu sonuçlar yorumlanmalıdır. Uygulanan bu çalışmada klinik olarak kardiyovasküler hastalığı geçirme riski yüksek olan bireyleri bulmak daha değerli olduğu için çapraz tablolarda bu oranın yüksek olması istenmektedir. Dolayısıyla kümeleme algoritmalarındaki kriterler, varsayımlar, algoritmaların kullanım koşulları, dezavantaj ve avantajları bir bütün olarak değerlendirilerek sağlık alanında klinik bilgiler ışığında kümeleme yöntemleri yorumlanmalıdır. İstatistiksel yöntemler bilindiği üzere klinik bulguları destekler nitelikte olmalıdır ki uygulamada rahatlıkla ve doğru bir şekilde kullanılma imkânı olsun.

Sonuç olarak sağlık alanında kümeleme algoritmalarının uygulamalarının artırılmasını önermekteyiz. Doğru yöntem kullanıldığı takdirde sağlık politikalarının geliştirilmesinde hastalık riski olan bireyler belirlenerek önlem alınır. Böylece halkımızın yaşam kalitesinin artması sağlanacak ve ortalama yaşam süresi uzayacaktır. Dolayısıyla basit bir kümeleme algoritması bile belki de sağlık alanında çok fazla değişiklerin olmasına ve tıpta gelişmelere yol açacaktır.

KAYNAKLAR

1. Han J, Kamber M. What is Data Mining?, What Kinds of Data Can be Mined? Data Mining: Concepts and Techniques. Morgan Kaufmann series in data management systems. 2nd ed. Amsterdam [u.a.]: Elsevier/Morgan Kaufmann; 2006. p.5-10.
2. Kob HC, Tan G. Data mining applications in healthcare. J Healthc Inf Manag 2005;19(9): 64-72.
3. Köktürk F, Ankaralı H, Sümbüloğlu V. [Overview to data mining methods]. Turkiye Klinikleri J Biostat 2009;1(1):20-5.
4. Fraley C, Raftery AE. How many clusters? Which clustering method? Answers via model-based cluster analysis. The Computer Journal 1998; 41(8):578-88.
5. Ferligoj A, Batagelj V. Some types of clustering with relational constraint. Psychometrika 1983;48(1):541-52.
6. MacQueen J. Some Methods for Classification and Analysis of Multivariate Obervations. Berkeley Symp. on Math. Statist. and Prob. University of California Press; 1967. p.281-97.
7. Zhanoxia T. K-Means clustering algorithm based on improved genetic algorithm. Journal of Chengdu University 2011;30(2):162-4.
8. Zhang C, Fang Z. An improved K-Means clustering algorithm. Journal of Information and Computational Science 2013;10(1):193-9.
9. Vadeyar D Yogish HK. Farthest first clustering in links reorganization. International Journal of Web & Semantic Technology 2014; 5(3):17-24.
10. Sharma N, Bajpai A, Litoriya R. Comparison the various clustering algorithms of Weka tools. International Journal of Emerging Technology and Advanced Engineering 2012; 2(5):73-80.
11. Park HS, Jun CH. A simple and fast algorithm for K-Medoids clustering. Expert Systems with Applications 2009;36(2):3336-41.
12. Bhat Aruna. K-Medoids clustering using partitioning around medoids for performing face recognition. International Journal of Soft Computing, Mathematics and Control 2014;3(3):1-12.
13. Salem SA, Nandi AK. New assessment criteria for clustering algorithms. Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing Mystic 2005; 1(1):285-90.
14. Han J, Kamber M. Partitioning methods. Data Mining: Concepts and Techniques. USA: Morgan Kaufmann Publishers Inc.; 2006. p.454-5.
15. Velmurugan T, Santhanam T. Computational complexity between K-Means and K-Medoids clustering algorithms for normal and uniform distribution of data points. Journal of Computer Science 2010;6(3):363-8.
16. Drineas P, Frieze A, Kannan R, Vempala S, Vinay V. Clustering large graphs via the singular value decomposition. Machine Learning 2004;56(1):9-33.
17. Priyadarishini A, Karthik S, Anuradha J, Tripathy BK. Diagnosis of psychopathology using clustering and rule extraction using rough set. Advances in Applied Science Research 2011;2(3):346-62.
18. Subhashree K, Prakash PS. Comparison of K-Means and K-Medoids clustering algorithms for big data using mapreduce techniques. International Journal of Innovative Science, Engineering & Technology 2015;2(4):627-32.
19. Abbas OA. Comparisons between data clustering algorithms. The International Arab Journal of Information Technology 2008;5(3): 320-5.
20. Madhulatha TS. Comparison between K-Means and K-Medoids clustering algorithms. Advances in Computing and Information Technology Communications in Computer and Information Science 2011;198(1):472-81.
21. Tiwari M, Singh R. Comparative investigation of K-Means and K-Medoid algorithm on iris data. International Journal of Engineering Research and Development 2012;4(8):69-72.
22. Shrisvastava V, Arya PN. A study of various clustering algorithms on retail sales data. International Journal of Computing, Communications and Networking 2012;1(2):98-74.
23. Sharma N, Bajpai A, Litoriya R. Comparison the various clustering algorithms of weka tools. International Journal of Emerging Techonology and Advanced Engineering 2012;2(5):73-80.
24. Godara S, Yadav R. Performance analysis of clustering algorithms for character recognition using Weka tool. International Journal of Advanced Computer and Mathematical Sciences 2013;4(1):119-23.
25 Singh B, Dubey G. A comparative analysis of different data mining using WEKA. International Journal of Innovative Research & Studies 2013;2(5):380-91.
26. Revathi S, Nalini T. Performance comparison of various clustering algoritm. International Journal of Advanced Research in Computer Science and Software Engineering 2013; 3(2):66-72.
27. Balabantaray RC, Sarma C, Jha M. Document clustering using K-means and K-medoids. International Journal of Knowledge Based Computer Systems 2015;1(1):7-13.
28. Kakkar P, Parashar A. Comparison of different clustering algrithms using WEKA tool. International Journal of Advanced Research in Technology, Engineering and Science 2014;1(2):20-2.
29. Jung YG, Kang MS, Heo J. Clustering performance comparison using K-means and expectation maximization algorithm. Biotechnol Biotechnol Equip 2014;28(Suppl 1):44-8.
30. Goyal VK. An experimental analysis of clustering algorithms in data mining using Weka tool. International Journal of Innovative Research in Science & Engineering 2014;2(4): 171-6.