Kansızlık Tanısına İlişkin Bir Veri Madenciliği Uygulaması

Yüksel Yurtay[1], Yavuz Salman[1], Mehmet Emin Salman[1], Fatih Gençali[1]
[1]Bilgisayar Mühendisliği Bölümü, Sakarya Üniversitesi, Türkiye.

Özet

Tıbbi verilerin elektronik ortamlarda birikimiyle başlayan süreçte, farklı disiplinlerde farklı çözüm arayışları başlamıştır. Sürekli genişleyen bu tıbbi veri tabanları üzerinde, geleneksel sorgulama ve raporlama araçlarını kullanmak yetersiz kalmaktadır. Büyük tıbbi veri tabanlarından, üstü kapalı örüntülerin, birlikteliklerin, anormalliklerin ortaya çıkarılması önem kazanmıştır. Bu sebeplerle veri madenciliği, tıbbi alanlardaki veriler üzerinde ilgi çekici bir hale gelmiştir. Bu çalışmada, bir veri madenciliği çalışması yapılmış, kan değerlerinden oluşan laboratuvar verilerinin kümeleme algoritması ile sonuçları irdelenmiştir.

Anahtar Kelimeler: Anemi, Veri madenciliği, Kümeleme algoritması.

Abstract

Process starting with the accumulation of medical data in electronic environments, different solutions have been developed in different disciplines. The ever-expanding medical data bases this on the use of traditional query and reporting tools are insufficient. Large medical data bases, covered patterns, associations, the discovery of abnormalities has gained importance. For these reasons, data mining, data on medical fields have become attractive. In this study, a data mining work was done, the blood results were analyzed with a clustering algorithm consisting of the values of laboratory data.

Keywords: Anemia, data mining, clustering algorithm.

1.Giriş

Demir yetersizliği, aneminin dünya çapında başlıca nedenlerinden biridir. Bütün vücudun demir oranı 2 gram ile 4 gram arasındadır. Erkeklerde 50 mg/kg, kadınlarda 35 mg/kg dolaylarındadır. Demirin çoğu hemoglobinin kırmızı kan hücrelerinin dolaşımında bulunur. Erkeklerde kırmızı kan hücresi 30 ml/kg'a yakın olup 70 kg bir adamın 2100 ml kırmızı kan hücresi vardır.[1] Kadınlarda regl ve doğum gibi sebeplere bağlı olarak meydana gelen kan kayıpları, mide, bağırsak kanamaları gibi nedenlerle kan kaybına bağlı demir yetersizliği oluşabilir bu durum anemiye yol açabilir. Kadınlarda ise kırmızı kan hacmi 27 ml/kg ye yakın örneğin 50 kilo bir kadının 1350 ml kırmızı kan hücresi vardır.

Amerika'da demir eksikliğine bağlı olarak gerçekleşen kansızlık değerlerinin 1999-2000 yılları için, 12-49 yaş arası kadınlarda %12, 50-69 yaş için %9, 70 ve yukarı için de %6 olduğu ifade edilmiştir

Çocuklarda ve erkeklerde ise bu oranların daha düşük olduğu görülmüştür [2]. Demir eksikliği anemisi ile ilgili bir çalışmada, sistemin işleyişi veri madenciliği tekniklerinden olan karar ağaçları yapısı ile sağlanmaktadır. Sisteme giriş olarak, biyokimya parametrelerinden demir eksikliği anemisi hastalığı için temel belirleyiciler olan Serum demiri, Serum demir bağlama kapasitesi (SDBK) ve Ferritin enzimleri kullanılarak, çıkış olarak ta Anemi(+) ve Anemi(-) değerlendirmelerinde bulunulmuştur. Tasarlanan sistemde 96 hasta verisi değerlendirilmiştir. Karar destek sisteminin sonuçları, doktorun verdiği kararlarla tamamen örtüşmüştür.[3]

Bu çalışmada, kansızlık tanısı amacı ile yapılmış laboratuvar kan testleri kullanılarak kümeleme çalışması yapılmıştır. Amaç hangi hastaların benzerlik gösterdiklerini, hangi gruptaki hastaların kansızlık düzeylerinin ne kadar ilerlediğini incelemektir. Sonuçlar, RapidMiner 5.3 programı yardımıyla elde edilmiştir. Çalışma sonucunda hastalar yaş, HGB ve HCT değerlerine göre gruplandırılmış ve hastalık seviyeleri belirlenmiştir.

2. Metot

2.1. Veri Kaynağı

Özel bir hastanenin 10769 adet hasta verisi üzerinde çalışılmıştır. Bu veriler içinden demir eksikliği anemisi tanısına yönelik laboratuvar kan analiz sonuçları seçilmiş ve kullanılmıştır. Veriler 1457 kansızlık var, 9312 kansızlık yok tanısı içermektedir. Nitelik olarak RBC, HGB, HCT, MCV, MCH, MCHC değerleri yer almaktadır (Tablo 1).

Tablo 1. Demir eksikliğine bağlı kansızlık için hematolojik parametreler [4],[5].

Parametre	Açıklama	Değerler
RBC	Kırmızı kan Hücreleri (Red Blood Cells)	4,5-6
HGB	Hemoglobin	12-16
HCT	Hematocrit	36-48
MCV	Ortalama Eritrosit Hacmi (Mean Corpuscular Volume)	80-100
MCH	Ortalama Hemoglobin Hacmi (Mean Corpuscular Hemoglobin)	27-34
MCHC	Ortalama Hemoglobin Hacmi Yoğunluğu (Mean Corpuscular Hemoglobin Concentration)	31-37

Kayıtlı hastaların bu laboratuvar verileri istatistik olarak incelendiğinde, aneminin kadınlarda erkeklere göre 4 kat daha sık görüldüğü, aneminin erkeklerde 66-75'li yaşlarda daha sık olduğu, kadınlarda da 36-45'li yaşlar arasında yaygın olduğu ortaya çıkmıştır.

2.2. Kümeleme

Kümeleme analizi, verilerin farklı gruplaşmalarını arar ve varsa bunları ortaya koyar [6]. Kümeleme işlemi birbirine benzeyen nesnelerin aynı grupta toplanmasıdır. Bu analizde en büyük etken hangi kriterlere göre kümeleme yapılacağıdır. Bu kriterler konu ile ilgili uzman

tarafından tahmin edilir. İstatistik tekniklerden farklı olarak verilerin normal dağılım göstermesi kümeleme analizi için önemli değildir.

Bu çalışmada Hiyerarşik kümeleme algoritmalarından olan ve Kaufman ve Rousseeuw tarafından 1990 yılında sunulan AGNES (AGglomerative NEsting) algoritması kullanılmıştır [7]. Algoritma, her nesneyi ayrı bir küme kabul ederek çalışmaya başlar, benzer özellik gösteren kümeleri birleştirir ve bu birleştirme işlemi, herhangi bir koşulla sonlandırılmadıkça, veriler tek kümede toplandığında sona erer [8]. Nesneler arasındaki benzerlikler belirlenirken

Öklid

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

ya da Manhattan

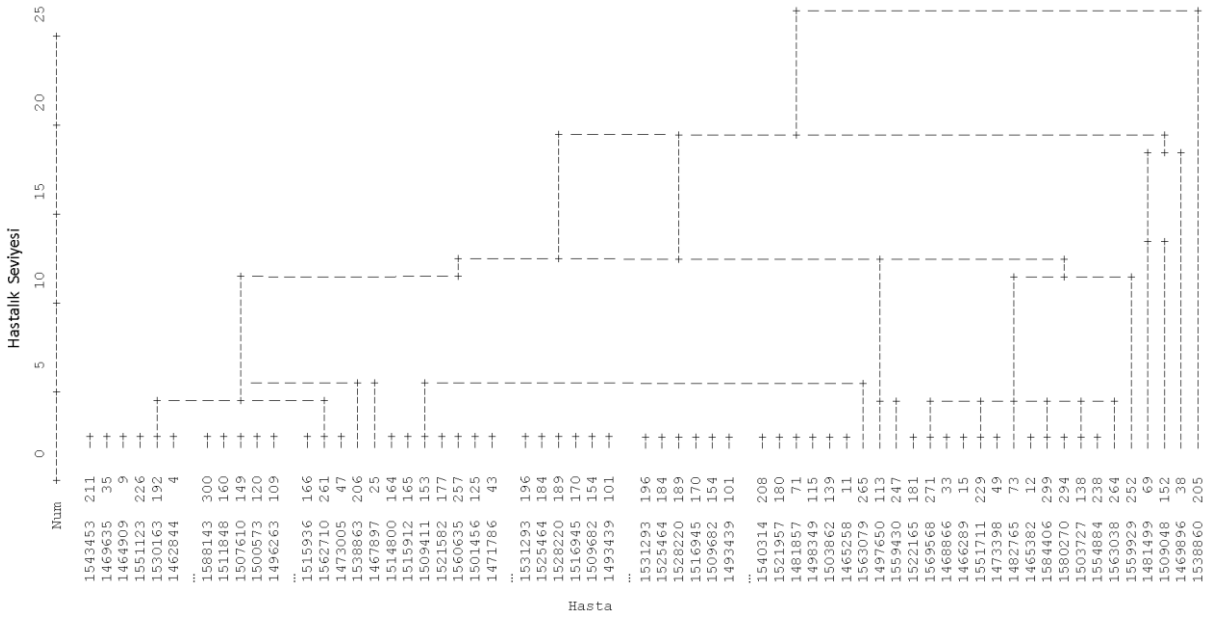
$$d(i,j) = |x_{i1} - x_{j1}| + \cdots + |x_{ip} - x_{jp}|$$

uzaklık formülleri kullanılabilir.

2.3. Yöntemlerin Sonuçları

Bu çalışmada kümeleme analizi için, RapidMiner 5.3 AGglomerative NEsting algoritması kullanılmıştır. Kümeleme analizi kadınlar üzerinde yaş, HGB, HCT verileri üzerinden gerçekleştirilmiştir.

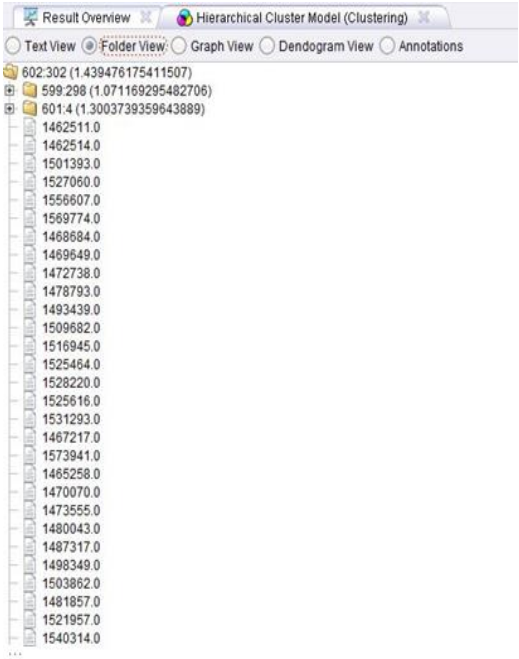
Hiyerarşik kümeleme analizini daha iyi irdelemek için, 36-45 yaşlar arasındaki hasta verileri ele alınmıştır. Bilindiği üzere kadınlarda anemi görülme sıklığı erkeklere göre daha fazladır. Özellikle 36-45'li yaşlarda maksimum düzeydedir. Hiyerarşik kümeleme için kullanılan 36-45 yaş aralığındaki kadın hasta sayısı 302 adettir.



Şekil 1. Kümeleme analizi dendogram grafiği

Şekil 1 deki kümeleme analizi sonucu elde edilen dendogram grafiğinin çarpıcı kısımları aktarılmıştır. Bu grafik ile hastaların kümeleme analizi sonucunda hiyerarşik dağılımı görülmektedir.

	A	B	C	D
1	Değerlendirme Verileri			
2	HASTA NO	YAS	HGB	HCT
3	1462511	38	11	34
4	1462514	38	11	34
5	1462780	40	11	32
6	1462844	37	10	33
7	1463461	42	11	34
8	1463916	41	11	33
9	1464085	42	9	29
10	1464487	44	10	34
11	1464909	37	10	31
12	1464994	36	10	30
13	1465258	37	11	34
14	1465382	43	8	27
15	1465697	44	11	32
16	1465806	42	11	33
17	1466289	41	8	26
18	...	...	...	...



Şekil 2. İşlenmiş ve sıralanmış hasta verileri

Şekil 2 de, hastalara ait verilerin temizlenmiş ve sıralanmış hali listelenmiştir. Şekil 1'de görülen dendogram grafiğinde hastaların hiyerarşik kümelenmesi görülmektedir. Dendogram grafiği incelendiğinde soldan sağa doğru yığılmanın olduğu görülmektedir. HCT, HGB değerlerinin eksende soldan sağa ilerledikçe azaldığı, hastalık seviyesinin ise arttığı görülmüştür. Ancak bu durumun yaş parametresinden bağımsız olduğu gözlemlenmiştir.

Aynı küme içerisinde yer alan hastaların yaşlarının birbirine çok yakın, HCT ve HGB değerlerinin de aynı olduğu görülmüştür.

Elde edilen dendogram grafiğinde hastalar arasındaki farklılıkların ya da benzerliklerin diğer etkenlerden(doğum, diğer rahatsızlıklar, beslenme vb.) olduğu tahmin edilmektedir.

Ayrıca rahatsızlığın ilk evrelerinde hastalara ait hemogram testi değerleri(MCV, MCH, MCHC vs.) incelediğinde normal değere çok yakın seyrettiği görülmüştür. Bu durumun elde ettiğimiz dendogram grafiğinin ilk kümelerinden alınan sonuçlarla aynı olduğu tespit edilmiştir.

Rahatsızlık seviyesi ilerlemiş olan hastaların hemogram testine göre MCH, MCHC ve MCV değerlerinin normalin çok altında olduğu görülmüştür. Dendogram grafiğindeki aynı kümede yer alan hastaların HGB, HCT değerlerinin de normalin altında olduğu görülmüştür.

Yeni hasta verisine ait HCT, HGB ve yaş kriterleri, dendogram grafiği üzerinde değerlendirildiğinde hastaya ait durum bilgisi elde edilebilmektedir.

3. Sonuç

Bu çalışma farklı bir açıdan ve veri madenciliği yöntemlerini kullanarak tıbbi alanlardaki teşhis ve tedavi süreçlerine destek olmayı amaçlamaktadır. Tıp alanında biyolojik/medikal, klinik ve yönetsel anlamda veri madenciliğinden faydalanmak olasıdır. Elde edilen sonuçların teşhis ve tedavi süreçlerinin herhangi bir aşamasında güvenilir bir şekilde kullanılması oldukça önemlidir. Bu sonuçlar bir veri madenciliği uygulaması üzerinden gerek karar ağaçlarının kullanılması gerekse kümeleme algoritmalarından çıkarılan analizler ile gerçekleştirilebilir.

Veri madenciliği algoritmaları uygulanırken sıklıkla kullanılan yöntemlerden biri de kümelemedir. Gerçek hasta verilerinden yola çıkarak yapılan bu çalışmada ham verilere ait temizleme, indirgeme, normalleştirme vb. işlemleri gerçekleştirilmiştir. Sonrasında veri madenciliği algoritmaları ile işlenebilir veri haline getirilmiştir. Üzerinde çalışılabilir 302 adet hasta verisinden hareketle hiyerarşik kümeleme algoritmalarından agglomerative nesting algoritması kullanılarak kümeleme sonuçları elde edilmiştir. Bu sonuçlar dendogram grafiği ile okunabilir biçimde ifade edildikten sonra hemogram testinden elde edilen sonuçlarla karşılaştırılarak, uzman doktor yardımıyla ile yorumlanmıştır. Elde edilen sonuçlar 36-45 yaşları arasındaki kadın hastalarda HCT, HGB azalan değerlerine bağlı olarak hastalık seviyelerinin arttığı gözlemlenmiştir, Aynı küme içerisinde farklı yaş gruplarındaki hastaların bulunması ise diğer etkenlerden(hamilelik, geçirilen rahatsızlıklar, beslenme vs.) kaynaklandığı tespit edilmiştir.

Bundan sonraki benzer çalışmalarda farklı veri gruplarıyla ve algoritmalar yardımıyla sağlık alanındaki teşhis ve tedavi süreçlerine destek olunabilir. Dolayısıyla bu çalışmalar tıp alanında hizmet üreten tüm uzmanlara, işin ve sürecin kalitesi açısından son derece faydalı olacağı düşünülmektedir.

Kaynaklar:

[1] Linker, C.A., Current Medical Diagnosis & Treatment , Chapter 13 ,page 470, 2003
[2] Susan F. Clark, Iron Deficiency Anemia,Nutrition in Clinical Practice, American Society for Parenteral and Enteral Nutrition,23(2):128-141, http://ncp.sagepub.com/content/23/2/128.full.pdf+html.2008.
[3] Doğan Ş., Türkoğlu İ., Iron-Deficiency Anemia Detection From Hematology Parameters By Using Decision Trees, International Journal of Science & Technology Volume 3, No 1, 85-92, 2008.
[4] Özaslan, E. ,Delibaşı, T., Tusem Book, Tusem Publisher, 46-48, 2008.
[5] Yılmaz, Z., Bozkurt, M.R., Determination of Women Iron Deficiency Anemia Using Neural Networks, Journal of Medical System, July 2011.
[6] Çelik H .Coskun, Satıcı Ömer, Çelik M. Yusuf, Kronik Sigara İçme Alışkanlığı Olan Üniversite Öğrencilerinin Tutumlarına İlişkin Değişkenlerin Aşamalı Kümeleme Yöntemi Çözümlemesi İnönü Üniversitesi Tıp Fakültesi Dergisi 11(4) 217-222 ,2004.
[7] Kaufman, L., Rousseeuw, P.,J., Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley and Sons Publishers, 1990.
[8] Demiralay M., Çamurcu A.,Y., CURE, AGNES VE K-MEANS Algoritmalarındaki Kümeleme Yeteneklerinin Karşılaştırılması, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, Yıl: 4, Sayı: 8, Güz 2005/2, s.1-18.