Kadinlarda demir eksikliğine bağli kansizlik tanisina ilişkin bir veri madenciliği çalişmasi

Yüksel Yurtay[1] yyurtay@sakarya.edu.tr
Ziynet Yılmaz[2] ziynet@sakarya.edu.tr
Özgür Çiftçi[1] ociftci@sakarya.edu.tr
Kayhan Ayar[1] kayar@sakarya.edu.tr

[1]Bilgisayar Müh. Bölümü
Sakarya Üniversitesi,
Türkiye

[2]Elektrik-Elektronik Müh. Bölümü
Sakarya Üniversitesi,
Türkiye

Özet : Tıp alanında kullanılan veri madenciliği uygulamaları; veri ambarları, karmaşık terimler, çok sayıda hasta ya da potansiyel hastadan elde edilen çok büyük sayıda verilerden oluşmaktadır. Karar destek sistemi kullanılarak bahsi geçen verileri iyi analiz ederek, doğru karar verebilmek oldukça önemlidir. Bu çalışmada, kadınlarda demir eksikliğine bağlı kansızlık tanısına ilişkin bir veri ambarı kullanılarak, Gini algoritması işletilmekte ve tıbbi karar ağacı elde edilmektedir. Karar ağacının başarısı ROC analizi ile irdelenmekte ve sonuçları açıklanmaktadır.

Anahtar kelimeler: Demir eksikliği anemisi, veri madenciliği, gini algoritması, anemi teşhisi

Diagnosis of iron deficiency anemia women regarding a study of data mining

Abstract: In the medicine used data mining applications which consisting data warehouses, complex terms, a large number of data obtain from large numbers of patients or potential patients. Using a decision support system, mention data correct analyzing is very important to give the right decision. In this study, diagnose iron deficiency anemia in women related by using data warehouse, operated Gini algorithm and obtained medical decision tree. The success of the decision tree is evaluated by ROC analysis and the results describes.

Keywords:. Iron deficiency anemia, data mining, gini algorithm, anemia diagnosis**.**

Giriş

Kansızlık, kandaki kırmızı kan hücrelerinin miktarının veya hemoglobin oranının normalden daha az olması şeklinde tanımlanmaktadır. Tedavi edilmediği durumda birçok değişik kalp hastalıklarının oluşmasına neden olmaktadır. Anemi hastaları yaşamlarını olması gerektiği şekilde devam ettirememekte, vücut fonksiyonlarındaki düzensizlik nedeniyle de anemi hastalarının ölüm oranları artmaktadır.

2005 yılı TUIK istatistiklerine göre demir eksikliğine bağlı kansızlık tanısı konan kadın hastalar, tüm hastalıklar arasında % 12,2 lik bir yere sahiptir (Tuik, 2012). Bu sadece yatarak tedavi olan hasta sayısıdır. Ayakta tedavi edilen hasta sayısının daha çok olduğu tahmin edilmektedir. Demir eksikliği de kansızlığın en önemli sebeplerinden birisi olduğu bilinmektedir (Linker, 2003). Clark (2008), Amerika'da demir eksikliğine bağlı olarak gerçekleşen kansızlık değerlerini 1999-2000 yılları için, 12-49 yaş arası kadınlarda %12, 50-69 yaş için %9, 70 ve yukarı için de %6 olarak ifade etmiştir. Çocuklarda ve erkeklerde ise bu oranların daha düşük olduğu görülmüştür.

Özellikle tıbbi veri tabanlarında veri analizi, karar destek sistemlerinin oluşturulması, yönetim biriminde bilgilere etkili ve hızlı bir şekilde ulaşılabilmesi bakımından bilgisayarlar uzmanlara büyük kolaylıklar sağlamaktadır. Bu hedef doğrultusunda önceden bilinmeyen, ilk bakışta fark edilemeyen, veri içinde gizli kalmış anlamlı ve değerli bilgiler elde edilebilmesinden dolayı veri madenciliği etkili bir çözüm olmuştur. Özellikle sınıflandırma işlemi, tıbbi karar destek sistemlerinde önemli bir yer tutar. Karar ağaçları da sınıflandırma yöntemlerinden biridir. Gini, Twoing, ID3, C4.5 gibi pek çok karar ağacı algoritması geliştirilmiştir(Hand ve diğerleri,2001).

Bu çalışmada demir eksikliğine bağlı kansızlık tanısı amacı ile Karar ağacı Gini algoritması ile geliştirilmiş ve %92,97 doğrulukla tanı sağlanmıştır.

Metot

Veri Kaynağı

Zonguldak Devlet Hastanesi'nin 2010 yılına ait 2640 adet kadın hastasının demir eksikliğine bağlı kansızlık tanısına yönelik laboratuvar kan analiz sonuçları kullanılmıştır. Bu veriler 567 kansızlık var, 2073 kansızlık yok tanısı içermektedir. (Yılmaz ve Bozkurt, 2011), RBC,HGB,HCT,MCV,MCH,MCHT değerlerini içeren laboratuvar verileri ile FFN, CFN, DDN,TDN,PNN VE LVQ yapay sinir ağlarını kullanmışlar ve sonuçlarını irdelemişlerdir. Bu çalışmada da kullanılan laboratuvar verilerinde yer alan hematolojik parametreler Tablo 1' de belirtilmiştir.

Tablo 1- Demir eksikliğine bağlı kansızlık için hematolojik parametreler(Özaslan and Delibaşı,2008,Yılmaz and Bozkurt, 2011).

Parametre	Açıklama	Değerler
RBC	Kırmızı Kan Hücreleri (Red Blood Cells)	4,5-6
HGB	Hemoglobin	12-16
HCT	Hematocrit	36-48
MCV	Ana Korpuskuler Hacmi (Mean Corpuscular Volume)	80-100
MCH	Ana Korpuskuler Hemoglabin (Mean Corpuscular Hemoglobin)	27-34
MCHC	Ana Korpuskuler Hem. Hacmi (Mean Corpuscular Hemoglobin Concentration)	31-37

Yöntem

Gini Algoritması, ikili bölünmeler şeklinde gerçekleşen bir sınıflandırma yöntemi olup, ikili yinelemeli bölümleme için en iyi bilinen kurallardandır. Her bir ağaç farklı bir stil ile gelişir. Algoritma nitelik değerlerinin sol ve sağda olmak üzere ikili bölünmeler şeklinde ayrılması temeline dayanır(Özkan,2008).

L_i : *Sol daldaki i grubundaki örnek(lerin) sayısı*

R_i : *Sağ daldaki i grubundaki örnek(lerin) sayısı*

k :*Sınıfların sayısı*

T :*Düğümdeki örnekler*

$|T_{sol}|$: *Sol daldaki örnek(lerin) sayısı*

$|T_{sağ}|$:*Sağ daldaki örnek(lerin) sayısı*

Tanımlamaları ile aşağıdaki bağıntılar hesaplanabilecektir.

$$Ginisol = 1 - \sum_{i=1}^{k} \left(\frac{L_i}{|T_{sol}|} \right)^2 \quad Ginisağ = 1 - \sum_{i=1}^{k} \left(\frac{R_i}{|T_{sağ}|} \right)^2$$

Her bir j niteliği için n öğrenme setindeki eleman sayısı olmak üzere aşağıdaki bağıntı hesaplanır.

$$Gini_j = \frac{1}{n} (|T_{sol}| Gini_{sol} + |T_{sağ}| Gini_{sağ})$$

Performans Değerlendirme-ROC Analizi

Bu çalışmanın performansı için ROC (Receiver Operating Characteric) analizinden faydalanılmıştır. ROC analizi II. Dünya Savaşı sırasında Britanya'da radarda tespit edilen sinyallerin doğru tanımlanması, dost ve düşman ayrımının sağlanması için geliştirilmiştir ve 1967 yılında tıpta karar vermede kullanımı önerilerek, 1969 yılında medikal görüntüleme cihazlarında kullanımını sağlamıştır. Sonraki yıllarda tıpta tanı testlerinin performansının değerlendirilmesinde kullanımı giderek yaygınlaşmıştır (Tomak ve Bek, 2010). Tanı testlerinde olumlu ya da olumsuz kararın doğruluk derecesi önemlidir. Pozitif ya da negatif kararların her biri için doğruluk düzeyini gösteren ölçütler vardır. ROC analizi de bu ölçütleri bulmayı sağlar.

Tablo 2. ROC analizi için kullanılan parametreler

	Gerçek Durum		
Test Sonucu	Pozitif	Negatif	Toplam
Pozitif	Doğru pozitif (DP)	Yanlış Pozitif (YP)	(DP+YP)
Negatif	Yanlış Negatif (YN)	Doğru Negatif (DN)	(YN+DN)
Toplam	(DP+YN)	(YP+DN)	(DP+YN+YP+DN)

Tablo 2'yer alan bu parametreler aşağıdaki gibi açıklanabilir:

DP: Gerçek durum pozitifken test sonucu da pozitif çıkan durumlar

YN: Gerçek durum pozitifken test sonucu negatif çıkan durumlar

YP: Gerçek durum negatifken test sonucu pozitif çıkan durumlar

DN: Gerçek durum negatifken test sonucu da negatif çıkan durumlar

Bu ROC parametreleri kullanılarak doğruluk değeri aşağıdaki formülle hesaplanır:

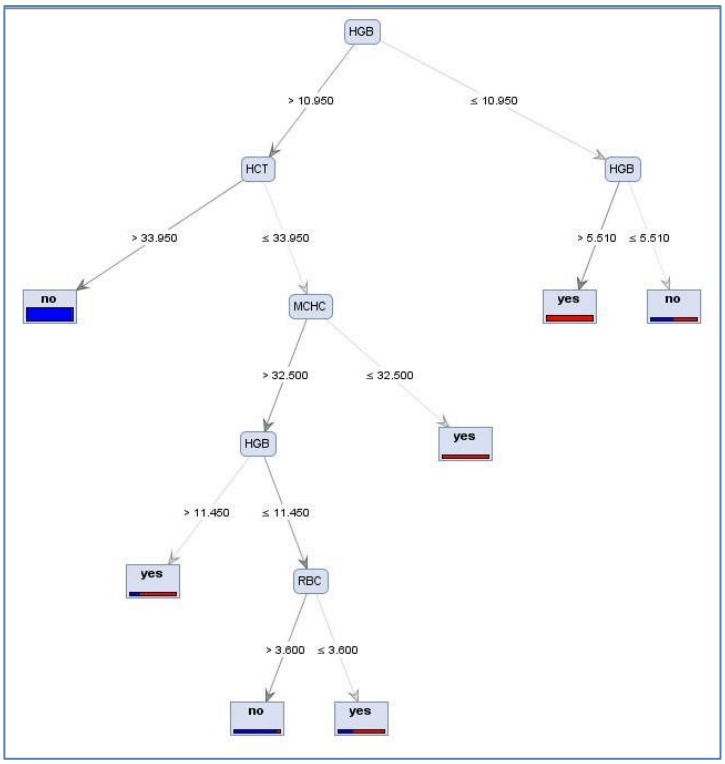
Doğruluk(Accuracy): (DP+DN) / (DP+YP+YN+DN)

Bu çalışmada 2640 adet veri temizlendikten sonra kalan 2599 adet veri 3 kutuya random olarak ayrılmış ve 2 kutu eğitim ve 1 kutu test olasılıklarının tümü değerlendirilerek ortalama doğruluk değerleri elde edilmiştir. Tablo 3'de 3 kutu için olası durumlar ve doğruluk değerleri ve ortalama doğruluk değeri gösterilmektedir.

Tablo 3- Demir eksikliğine bağlı kansızlık verileri için Doğruluk değerleri

Kutu seçimi(kutu1:867 kayıt, kutu2:866, kutu 3: 866			Ortalama doğruluk değeri
Eğitim	Test	Doğruluk(%)	
1- Kutu1+ Kutu 2	1- Kutu 3	89,5	92,97
2- Kutu 1+ Kutu 3	2- Kutu 2	96,2	
3- Kutu 2+ Kutu 3	3- Kutu 1	93,2	

Şekil 1'de 2 nolu eğitim ve test seti için Rapidminer programı ile elde edilmiş karar ağacı gösterilmektedir (Rapidminer,2012). Bu karar ağacı 2 numaralı fold ile test edildiğinde %96,2 lik bir doğrulukla tanı yapabilecek kuralları içermektedir



Şekil 1. Demir eksikliğine bağlı kansızlık verileri için elde edilen karar ağacı

Sonuç

Bu çalışmada, kadınlarda demir eksikliğine bağlı kansızlık hastalığı araştırılırken, bu hastalıkla ilgili benzer semptomları gösteren kişilerin önceden kayıtlı bilgileri doğrultusunda bir veri madenciliği çalışması yapılmış, Gini algoritması uygulanarak karar ağacı çıkarılmıştır. Bu ağaç ve kuralları sayesinde yeni hastaların hangi gruba girdiği görülebilmektedir. Karar ağaçları belirlemede kullanılan diğer veri madenciliği algoritmaları da bu veriler üzerinde denenebilir ve olası başarıları yüksek görülmektedir.

Veri madenciliği kullanılıp oluşturulacak karar ağacı çalışmasında örnek sayısı çoğaldıkça elle çalışmak oldukça uzun ve zahmetli olabileceğinden bu tip çalışmalarda yazılımlar kullanmak ve büyük sayıda nitelik içeren öğrenme setine cevap verebilecek programlardan yararlanmak, sonuçların daha doğru ve kısa sürede alınması açısından daha faydalı olacaktır.

Kaynaklar

TOMAK. L., BEK.Y. (2010). İşlem Karakteristik Eğrisi Analizi Ve Eğri Altında Kalan Alanların Karşılaştırılması, Journal of Experimental and Clinical Medicine, Vol:27, no:2, s:58-65.

Hand, D., Mannila, H., Smyth, P. (2001). Principles of data mining, MIT Press.

Linker, C.A. (2003). Current Medical Diagnosis & Treatment , Chapter 13 ,page 470.

Özaslan, E., Delibaşı, T. (2008). Tusem Book, Tusem Publisher, 46-48.

Özkan, Y. (2008). Veri Madenciliği Yöntemleri, Papatya Yayıncılık.

Rapidminer, http://rapid-i.com/content/view/181/190/ Erişim Tarihi: 22-11-2012.

Susan F. Clark (2008), 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.

Tuik, http://www.tuik.gov.tr/PreHaberBultenleri.do?id=8620, Erişim Tarihi: 22-11-2012.

Yılmaz, Z., Bozkurt, M.R. (2011). Determination of Women Iron Deficiency Anemia Using Neural Networks, Journal of Medical System.