NWSA-Engineering Sciences
ISSN: 1306-3111/1308-7231
NWSA ID: 2014.9.2.1A0352

Status : Original Study
Received: December 2013
Accepted: April 2014

E-Journal of New World Sciences Academy

Yunus Kökver
Kırıkkale University, yunuskokver@gmail.com, Kırıkkale-Turkey
Necaattin Barışçı
Kırıkkale University, nbarisci@kku.edu.tr, Kırıkkale-Turkey
Aydın Çiftçi
Kırıkkale University, dr.aydin.71@hotmail.com, Kırıkkale-Turkey
Yakup Ekmekçi
The Güven Hospital, yakupekmekci@hotmail.com, Kırıkkale-Turkey

http://dx.doi.org/10.12739/NWSA.2014.9.2.1A0352

HİPERTANSİYONA ETKİ EDEN FAKTÖRLERİN VERİ MADENCİLİĞİ YÖNTEMLERİYLE İNCELENMESİ

ÖZET

Bu çalışmada 150 kişiden alınan; Yaş, Cinsiyet, Vücut Kütle İndeksi, HDL, LDL, Trigliserid, Ürik Asit ve Sigara Kullanımı verileri, veri madenciliği sınıflandırma yöntemleriyle incelenmiştir. Veriler normal veya hasta olacak şekilde iki sınıfa ayrılmıştır. Böylelikle hipertansiyon hasta adaylarının, hipertansiyon olup olmadığını tahmin edecek bir teşhis sistemi geliştirilmiştir. Ayrıca elde edilen sonuçlara göre bir karar ağacı oluşturularak, hipertansiyona doğrudan ve dolaylı olarak etki eden faktörler belirlenmiştir. Sınıflandırma algoritmalarından C4.5, Naive Bayes ve Çok Katmanlı Algılayıcının kullanıldığı bu çalışmada, C4.5 algoritmasının daha başarılı sonuçlar verdiği görülmüştür.

Anahtar Kelimeler: Hipertansiyon, Veri Madenciliği, C4.5, Çok Katmanlı Algılayıcı, Naive Bayes

DETERMINING AFFECTING FACTORS OF HYPERTENSION WITH DATA MINING TECHNIQUES

ABSTRACT

In this paper, Age, Gender, Body Mass Index, HDL, LDL, Triglyceride, Uric Acid and The Use of Smoking data gathered from 150 patients are analyzed with data mining classification algorithms. The data is divided into two different classes which are normal and patient. Thus, a diagnostic system is developed which predicts whether a candidate patient has hypertension or not. Besides, a decision tree is created and factors affecting hypertension directly and indirectly are determined. In this study, C4.5, Naive Bayes and Multilayer perceptron classification algorithms are used, and shown that C4.5 algorithm gives better results.

Keywords: Hypertension, Data Mining, C4.5, Multilayer Perceptron, Naive Bayes

1. GİRİŞ (INTRODUCTION)

Hipertansiyon hastalığı, günümüzde önemli bir sağlık sorunu olarak karşımıza çıkmaktadır. Hipertansiyonu olan hastalar yıllarca bu hastalığı fark etmeyebilir. Hipertansiyon yavaş ve sinsice yıllarca herhangi bir belirti vermeden hastanın böbrek, kalp ve diğer damarlarına zarar verebilir. Toplumlarda hipertansiyona maruz kalma sıklığı yaşla beraber artar. İlerleyen yıllarda çoğu kimsede spot olarak yapılan ölçümlerde rastlanabilir. Bu bağlamda hipertansiyonun erken teşhis ve tedavisi büyük önem arz etmektedir. Bu önemin sebebi hipertansiyonun organlarda yaptığı hasar ve bunun sonucunda ortaya çıkan komplikasyonların tedavi maliyetleri ile iş gücü kaybının yüksek olmasıdır.

Ülkemizde hipertansiyon konusunda henüz yeterince farkındalık oluşmamıştır. Bunun sonucu olarak hipertansiyon tedavi ve komplikasyonları önlemede kritik önem arz eden erken teşhis istenilen düzeyde değildir. Bu açıdan bakılacak olursa erken teşhis çok önemlidir. Hipertansiyonun erken teşhisinde kamuoyu oluşturma yanında yeni metotlara da ihtiyaç vardır. Bu çalışmada, hipertansiyonun erken teşhisinde yeni bir yaklaşım olarak düşünülebilecek veri madenciliği yöntemleri incelenmiştir.

Kore Tıbbi Sigorta Kurumu tarafından hazırlanan bir veri tabanı üzerinde yüksek tansiyon ile ilgili bir çalışma yapılmıştır. Bu çalışma 1998 yılına ait 127886 kayıt üzerinde yapılmıştır. İlk aşamada yüksek tansiyona sahip 9103 kayıt üzerinde, daha sonra aynı sayıda yüksek tansiyonu olmayan kayıtlar üzerinde çalışılmıştır. Bu örnek 13689 kayıttan oluşan öğrenme ve 4588 kayıttan oluşan test setine bölünerek modelin eğitimi yapılmıştır. Öğrenim algoritmasında karar ağaçları algoritmalarından CHIAD, C4.5, C5.0 kullanılmıştır. Bu çalışmalar sonucunda yüksek tansiyon tahmininde etkili değerler BMI, idrar proteini, kan glikozu, kolesterol değerleridir. Yaşam koşullarının (diyet, alınan tuz miktarı, alkol, tütün gibi) hiçbirinin tahminde etkili olmadığı ayrıca grafiksel değerlerde de yalnızca yaşın etkili olduğu saptanmıştır [1].

Almazyad ve arkadaşları [2], hipertansiyonla ilgili 15-64 yaş arası erkeklerde yaptıkları çalışmada, belirledikleri 5 tedavi türünün, hangi yaş grubunda daha etkili olduğu sonuçlarının analizinde veri madenciliği algoritmalarından yararlanmışlardır.

Cheng-Ding Chang ve arkadaşları [3], ortak risk faktörlerine göre Hipertansiyon ve Hiperlipidemi çoklu hastalık tahmini modellemesi için Veri Madenciliği teknikleri kullanmışlardır. Bu çalışmada aynı anda Hipertansiyon ve Hiperlipidemi hastalık tahmini için iki aşamalı bir analiz yöntemi önerilmektedir. Öncelikle, bu iki hastalığın bireysel risk faktörlerini belirlemek için altı veri madenciliği yaklaşımı kullanılmıştır ve sonra oy ilkesi ile ortak risk faktörleri belirlenmiştir. Önerilen analiz yöntemi, bu iki hastalığın ortak risk faktörlerinin Sistolik Kan Basıncı, Trigliserid, Ürik Asit, Glutamat Piruvat Transaminaz(GPT) ve cinsiyet olduğunu göstermektedir.

Mevlut Türe ve arkadaşları [4], yaptıkları çalışmada, hipertansiyon hastalığı riskini tahmin etmek amacıyla, sınıflandırma tekniklerinin performansını karşılaştırmışlardır. 694 veri üzerinde retrospektif bir analiz yapılmıştır. 3 Karar Ağacı, 4 İstatistiksel Algoritma ve 2 Yapay Sinir Ağının performansları karşılaştırılmıştır. Belirleyici değişkenler; yaş, cinsiyet, hipertansiyon aile öyküsü, sigara alışkanlığı, lipoprotein(a), trigliserid, ürik asit, total kolesterol ve vücut-kütle indeksidir. Yapay Sinir Ağı algoritması olan Çok Katmanlı Algılayıcı (ÇKA), hipertansiyon tahmininde diğer yöntemlere göre daha iyi bir performans sergilemiştir.

2. ÇALIŞMANIN ÖNEMİ (RESEARCH SIGNIFICANCE)

Bu çalışmanın temel hedefi, hipertansiyon hastalığına etki eden kan parametrelerini ve demografik değerleri ve bu değerlerin hangi oranlarda hipertansiyona etki ettiklerini tespit etmek ve bu yönlere dikkat çekmektir. Bu amaçla çalışmada Naive Bayes, C4.5 ve Çok katmanlı algılayıcı ağı sonuçları karşılaştırılmıştır.

Bu çalışmada Kırıkkale Yüksek İhtisas Hastanesinde 150 kişiden, hipertansiyon hastalığına en çok etki edebilecek yaş, cinsiyet, vücut kütle indeksi, hdl, ldl, trigliserid, ürik asit ve sigara kullanımı verileri alınmıştır. Alınan bu veriler sonuçta 'normal' ve 'hasta' olacak şekilde 2 sınıfa ayrılmıştır. WEKA veri madenciliği aracı kullanılarak bu veriler üzerinde sınıflandırma yapılmıştır. Verilerin %76'sı eğitim için, geri kalan %24'ü de test verisi olarak kullanılmıştır. Bu çalışmada Naive Bayes, C4.5 ve Çok katmanlı algılayıcı ağı sonuçları karşılaştırılmıştır.

3. MATERYAL VE YÖNTEMLER (MATERIAL AND METHODS)

3.1. Hipertansiyon (Hypertension)

Hipertansiyon iki şekilde tanımlanmaktadır:

- Sistemik arteriyel kan basıncının normal sınırların üstünde olması (sistolik kan basıncı (SKB)≥140 mmHg, diyastolik kan basıncı (DKB) ≥ 90 mmHg) ya da kişinin antihipertansif ilaç kullanıyor olması,
- Bir sağlık profesyoneli tarafından en az iki kere yüksek kan basıncı olduğunun söylenmesi olarak tanımlanmaktadır [5].

Başka bir tanıma göre hipertansiyon "insan sağlığını, yaşam kalitesini ve yaşam süresini kötü yönde etkileyebilecek kadar yüksek olan kan basıncı değerleri" olarak ifade edilmektedir. Hipertansiyon tanısı almış hastalarda SKB, DKB veya her ikisi birden yüksek olabilir; bu değerlerdeki heyecan, korku veya egzersiz gibi durumlar nedeniyle oluşabilecek geçici yükselmeler hipertansiyon olarak kabul edilmez [6].

3.2. Hipertansiyonun Sınıflandırılması (Classification of Hypertension)

Hipertansiyon sınıflamasının amacı her hastanın durumuna uygun bir profil elde etmede güvenilir ve kolay bir yöntem sunmaktır. Sınıflama ile hastalığın ciddiyeti hakkında değerlendirme yapılabilir ve risk tanımlanarak tedavi sağlanabilir [7].

Erişkinlerde kan basıncı derecesinin sınıflandırılması niteldir. Pratikte tedaviye yaklaşım kolaylığı sağlamak için kan basıncı değerleri dikkate alınmaktadır. Amerika Birleşik Devletleri Ulusal Komitesi(JNC)-7 raporunda 18 yaş ve üstündeki erişkinlerin kan basınçları optimal, normal, yüksek-normal ve hipertansiyon olarak dört dereceye ayrılmıştır. Tablo 1'de hipertansiyonun sınıflandırılması gösterilmiştir [8].

Tablo 1. Hipertansiyonun sınıflandırılması
(Table 1. Classification of hypertension)

Kan Basıncı Derecesi	Sistolik	Diastolik
Optimal	<120	<80
Normal	<130	<85
Yüksek normal	130-139	85-89
Evre1	140-159	90-99
Evre2	>160	>100

3.3. Naive Bayes Algoritması (Naive Bayes Algorithm)

Eldeki verilerin belirlenmiş olan sınıflara ait olma olasılıklarını öngören bir algoritmadır. Temeli, istatistikteki Bayes teoremine dayanır. Bu teorem; belirsizlik taşıyan herhangi bir durumun modelinin oluşturularak, bu durumla ilgili evrensel doğrular ve gerçekçi gözlemler doğrultusunda belli sonuçlar elde edilmesine olanak sağlar. Belirsizlik taşıyan durumlarda karar verme konusunda oldukça başarılıdır.

Genellikle belirsizlik durumlarında sınıflandırma ve tahmin yapmak için kullanılır. En önemli dezavantajı değişkenler arası ilişkinin modellenmemesi ve değişkenlerin birbirinden tamamen bağımsız olduğu varsayımıdır [9].

Bayes Kuralı;

A ve B rastgele sayılar olsun;

$P(A \mid B) = P(B \mid A)P(A) / P(B)$ (1)

P(A) : A olayının bağımsız olasılığı öncül olasılık

P(B) : B olayının bağımsız olasılığı

P(B| A) : A olayının olduğu bilindiğinde B olayının olasılığı (şartlı olasılık)

P(A | B) : B olayının olduğu bilindiğinde A olayının olasılığı (artçıl) olasılık

Bayes kuralına dayanarak P(A | B)'yi maksimum yapan durumlar hesaplanabilir.

"E" A olayının bütün durumlarının kümesi;

$$A_{MAX} = \operatorname{argmax} A \in E\, P(A|B)$$

$$= \operatorname{argmax} A \in E \frac{P(B|A)P(A)}{P(B)}$$

$$= \operatorname{argmax} A \in E\, P(B|A)P(A) \quad (2)$$

3.4. C4.5 Algoritması (C4.5 Algorithm)

C4.5 algoritması ile sayısal değerler içeren veri tabanları üzerinde karar ağaçlarının oluşturulma olanağı sağlanmıştır.

C4.5 karar ağacının oluşturulma algoritması:

- Çıkış değerlerini en fazla farklılaştıran öznitelik seçilir.
- Seçilen özniteliğin her değeri için farklı bir dal oluşturulur.
- Seçilen düğümdeki öznitelik değerlerini yansıtacak şekilde örnekler alt gruplara ayrılır.
- Her alt grup için öznitelik seçimi durdurulur; Eğer
  - Alt gruptaki tüm üyeler aynı çıkış değerini üretiyorsa, ağacın ilerlemesi durdurulur ve çıkış değeri olarak son belirlenen değeri atanır.
  - Alt grupta tek düğüm kaldıysa veya ayırt edici öznitelikler belirlenemiyorsa ağacın ilerlemesi durdurulur.
- 3.aşamada belirlenen her alt grup için yukarıdaki işlem tekrarlanır.

Sayısal nitelikleri belirli aralıklara bölme konusunda bazı zorluklar görülebilir. Ancak en uygun t eşik değerini hesaplamak için çeşitli yöntemler bulunmaktadır. Nitelik değerleri sıralanır ve $\{V_1, V_2, \ldots, V_n\}$ şeklini alır. Nitelik değerler kümesi iki parçaya ayrılır ve Eşik değeri olarak $[V_i, V_{i+1}]$ aralığının orta noktası alınabilir: [10].

$t_i = (V_i + V_{i+1}) / 2$ (3)

3.5. Çok Katmanlı Algılayıcı (Multilayer Perceptron)

Bu sınıflandırıcı, örnekleri sınıflandırmak için arka planda yayılma yöntemini kullanır. Bu ağ elle oluşturulabileceği gibi bir algoritmayla ya da her ikisiyle de oluşturulabilir. Eğitim zamanı sürecinde bu ağ görüntülenebilir ve değiştirilebilir. Çok Katmanlı Algılayıcı (ÇKA), yapay sinir ağları yapısının bir sınıfıdır. ÇKA, paralel bağlı ağ modeli kurarak insan beyninin öğrenme mekanizmasını taklit eden akıllı kodlar oluşturmak üzerine yoğunlaşır. Bir ÇKA modelinde önce sistem eğitilir ve ağ, en son güncellenen ağ parametreleri kullanılarak fonksiyonel haritalayıcı olarak çıkışları hesaplayabilir [11].

Çok Katmanlı Algılayıcı, standart geri yayılım algoritması ile sürekli geri beslenerek eğitilir. ÇKA denetimli ağlardır ki, bu yüzden ÇKA istenilen cevabın eğitilmesi ihtiyacı duyar. ÇKA giriş verisinin istenilen cevap verisine nasıl dönüştürüleceğini öğrenir ve bu yüzden ÇKA örüntü sınıflamasında yaygınca kullanılır. Bir veya iki gizli katmanla ÇKA giriş ve çıkış haritasını sanal olarak birbirine yaklaştırır. Birçok yapay sinir ağları uygulaması ÇKA içerir.

ÇKA, algılayıcılar olarak adlandırılan basit sinir ağıdır. Algılayıcı, giriş ağırlıklarına göre lineer kombinasyon formuna dönüştürülerek çoklu gerçek değerli girişlerden ve sonra da bazı lineer olmayan aktivasyon fonksiyonları içinde yer almış mümkün çıkışlardan tek bir çıkış hesaplar [12].

Genellikle dış dünya bilgileri analog bilgi olduğundan bu bilgilerin, sayısal hale dönüştürülmesi ve bu sayısal bilginin 0-1 arası skalaya indirgenmesi gerekmektedir. Bu durum Eşitlik 4'de gösterilmiştir.

$$F_k = G_k \tag{4}$$

Burada;

F_k: Giriş katmanındaki k. nöronun çıkışını ifade eder.

G_k: Giriş katmanına dış dünyadan gelen bilgiyi ifade eder.

Ara katmanda ve çıkış katmanındaki nöronların çıkısının hesaplanabilmesi için ilgili nörona gelen net girdinin hesaplanması gerekir. Bunun için gelen bilgi ve ağırlık çarpımı kullanılır. Bu durum Eşitlik 5'de gösterilmiştir.

$$Net_j = \sum_{i=1}^{n} W_{ij} F_i \tag{5}$$

Burada;

Net_j : j. prosesin net girdisini ifade eder.

F_i : j. nörona bilgi gönderen nöronların çıkış bilgisidir.

W_{ij}: j. nörona bilgi gönderen i. nöron j. nöron arası ağırlığı ifade eder.

Ara katman ve çıkış katmanı nöronları için net bilginin sigmoid fonksiyonu kullanılarak nöron çıkışına dönüştürülmesi işlemi yapılır. Bu işlem Eşitlik 6'da gösterilmiştir.

$$F_j = \frac{1}{1+e^{(-Net_j+\beta)}} \tag{6}$$

Çıkış katmanında elde edilen çıkış bilgisi ile olması gereken çıkış bilgisi arası fark hatayı oluşturur. Hata kavramı Eşitlik 7'de gösterilmiştir.

$$E_m = B_m - C_m \quad (7)$$

Ağın test başarısı Eşitlik 2.8' de görüldüğü gibi hesaplanmaktadır.

$$P = \frac{D}{T} x100 \quad (8)$$

3.6. Sınıflandırma Algoritmalarının Doğruluğunu Test Etme Yöntemi (Testing Method of Classification Algorithm's Accuracy)

Bütün veri madenciliği modellerinin performansını hesaplamak için standart bir ölçütün kullanılması önemlidir. Veri madenciliğinde sınıflama modellerinin karşılaştırılması için en sık kullanılan yöntem hata oranını hesaplamaktır. İki sınıflı model için sınıflama matrisi Tablo 2'de verilmiştir.

Tablo 2. Sınıflama matrisi
(Table 2. Classification matrix)

Modelin Sınıf Tahmini	Gerçek Sınıf		
		Pozitif	Negatif
	Pozitif	Doğru Pozitif Sayısı (DP)	Yanlış Pozitif Sayısı (YP)
	Negatif	Yanlış Negatif Sayısı (YN)	Doğru Negatif Sayısı (DN)

Modeli oluşturan toplam örnek sayısı = N = DP + YP + YN + DN (9)
Modelin doğru sınıflama oranı (Doğruluk) = (DP + DN) / N (10)
Doğru Pozitif Oranı(Duyarlılık) = DP / (DP + YN) (11)
Doğru Negatif Oranı (Belirlilik) = DN / (DN + YP) (12)
Hassaslık (Precision) = DP / (DP + DN) (13)

4. BULGULAR VE TARTIŞMA (FINDINGS AND DISCUSSION)

Hipertansiyona etki etmesi muhtemel faktörler, uzman hekimlerle yapılan ortak bir çalışmayla belirlenmiş ve bu faktörlerin; Yaş, Cinsiyet, Boy, Kilo, Lipid Profili (HDL ve LDL), Trigliserid, Ürik Asit ve Sigara Kullanımı (günde içtiği paket sayısı ve kaç yıldır içtiği bilgisi) olması gerektiği vurgulanmıştır. Hastalardan bu değerler alınırken uyulması gereken kriterler ise, hastanın 30 yaş ve üzeri olması, hamile olmaması ve herhangi bir ilaç tedavisine başlamamış olmasıdır.

Bu kriterlere uygun olarak, Kırıkkale Yüksek İhtisas Hastanesi Dahiliye Bölümü'nde 150 kişiden değerler alınmış, sonuçta bu kişiler hipertansiyon hastası ve sağlıklı olarak 2 sınıfa ayrılmıştır.

Çalışmanın temel amacı olan hipertansiyona etki eden faktörler, Weka veri madenciliği aracı kullanılarak incelenmiştir.
Veri madenciliği çalışması için kullanılacak veri seti üzerindeki verilerin dağılımı Tablo 3'de gösterilmiştir.

Tablo 3. Veri madenciliği çalışması için kullanılacak verilerin dağılımı
(Table 3. The distribution of data to be used for the data mining application)

Sınıf	Sayı
Normal	65
Hasta	85
Toplam	150

150 adet kayıttan 85 tanesi hasta olarak belirlenmiş, kalan 65 kayıt ise normal olarak sınıflandırmaya katılmıştır.

Çalışmada kullanılan Weka 3.6.6 sürümünde yer alan sınıflandırıcılar sırayla seçilmiştir. Test seçeneği olarak tüm sınıflandırıcılar için yüzde ayırma yöntemi kullanılmıştır.

Verilerin %76'sı eğitim için, geri kalan %24'ü de test verisi olarak kullanılmıştır. Yani bir nevi seçilen algoritma, verilerin %76'sı arasında ilişki ve kuralları belirleyerek çeşitli örüntüler oluşturmuştur. Oluşan örüntü desenlerine göre de algoritmanın doğruluğu, verilerin kalan %24'ü üzerinde test edilmiştir [13].

4.1. Naive Bayes Algoritması İçin Veri Modellemesi (Data Modelling for Naive Bayes Algorithm)

Tablo 4'de Naive bayes algoritmasının hipertansiyon tahmini için Düzensizlik Matrisi verilmiştir.

Tablo 4. Naive bayes algoritması için düzensizlik matrisi
(Table 4. Confusion matrix for naive bayes algorithm)

a	b	
20	3	a=hasta
0	13	b=normal

Tablo 4'de verilen Düzensizlik Matrisine göre gerçekte 23 "hasta" değerli test verisinin 20 tanesi hasta, 3 tanesi "normal" ve gerçekte 13 "normal" değerli test verisinin 13 tanesi de "normal" olarak tahmin edilmiştir.

Tablo 4'de verilen Düzensizlik Matrisine göre 36 verinin detaylandırılmış doğruluk tablosu Tablo 5.'de verilmiştir.

Tablo 5. Naive bayes algoritması için detaylandırılmış doğruluk tablosu
(Table 5. Detailed confusion matrix for naive bayes algorithm)

Doğruluk Yüzdesi	Duyarlılık (DP Oranı)	Yanlış Hassaslık Oranı (YP Oranı)	Hassaslık (Precision)	Sınıf (Class)
%91,67	0,87	0	1	a = hasta
	1	0,13	0,813	b = normal

Tablo 5'e göre Naive Bayes Algoritması için doğruluk yüzdesi, %91,67 olarak hesaplanmıştır. Duyarlılık, Yanlış Hassaslık Oranı ve Hassaslık değerleri, hasta ve normal sınıfları için ayrı ayrı hesaplanmıştır [13].

4.2. C4.5 Algoritması İçin Veri Modellemesi (Data Modelling for C4.5 Algorithm)

Tablo 6'da C4.5 algoritmasının hipertansiyon tahmini için Düzensizlik Matrisi verilmiştir.

Tablo 6. C4.5 Algoritması için düzensizlik matrisi
(Table 6. Confusion matrix for C4.5 algorithm)

a	**b**	
22	1	a=hasta
2	11	b=normal

Tablo 6'da verilen Düzensizlik Matrisine göre gerçekte 23 "hasta" değerli test verisinin 22 tanesi "hasta", 1 tanesi "normal" ve

gerçekte 13 "normal" değerli test verisinin 11 tanesi "normal", 2 tanesi ise "hasta" olarak tahmin edilmiştir.

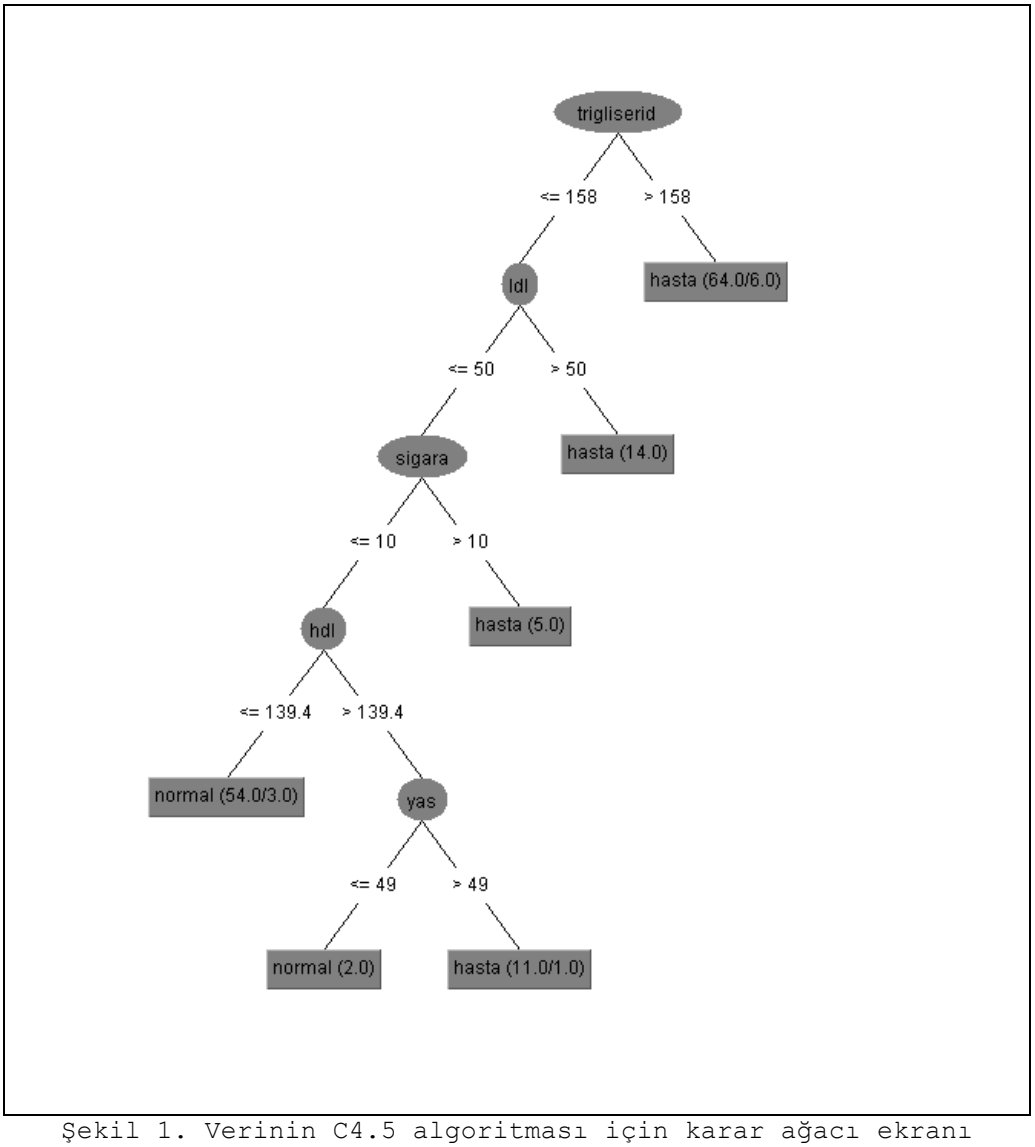
Tablo 6'da verilen Düzensizlik Matrisine göre 36 verinin detaylandırılmış doğruluk tablosu Tablo 7'de verilmiştir.

Tablo 7. C4.5 algoritması için detaylandırılmış doğruluk tablosu
(Table 7. Detailed confusion matrix for C4.5 algorithm)

Doğruluk Yüzdesi	Duyarlılık (DP Oranı)	Yanlış Hassaslık Oranı (YP Oranı)	Hassaslık (Precision)	Sınıf (Class)
%91,67	0,957	0,154	0,917	a = hasta
	0,846	0,043	0,917	b = normal

Tablo 7'ye göre C4.5 Algoritması için doğruluk yüzdesi, %91,67 olarak hesaplanmıştır. Duyarlılık, Yanlış Hassaslık Oranı ve Hassaslık değerleri, hasta ve normal sınıfları için ayrı ayrı hesaplanmıştır.

Şekil 1'de C4.5 algoritmasının hipertansiyon tahmini için karar ağacı sonuç ekranı görülmektedir.



Şekil 1. Verinin C4.5 algoritması için karar ağacı ekranı
(Figure 1. Decision tree screen for C4.5 algorithm)

Şekil 1'e göre trigliserid düzeyinin 158'den büyük olması durumu, doğrudan hipertansiyonun varlığına işaret etmektedir. Trigliserid düzeyine bağımlı olarak LDL değerinin 50'den büyük olması durumu, yine hipertansiyonun varlığına işaret etmektedir. Sigara kullanımının 10'dan büyük olması da hipertansiyona işaret eden farklı bir durum olarak karşımıza çıkmaktadır. Hipertansiyona etki eden bir diğer etken ise yaş değişkenidir. Şekil 1'e göre yaş değişkeni, doğrudan değil, dolaylı olarak hipertansiyona etki etmektedir [13].

4.3. Çok Katmanlı Algılayıcı (ÇKA) İçin Veri Modellemesi (Data Modelling For Multilayer Perceptron)

Tablo 8'de ÇKA'nın hipertansiyon tahmini için Düzensizlik Matrisi verilmiştir.

Tablo 8. ÇKA için düzensizlik matrisi
(Table 8. Confusion matrix for multilayer perceptron)

a	b	
19	4	a=hasta
1	12	b=normal

Tablo 8'de verilen Düzensizlik Matrisine göre gerçekte 23 "hasta" değerli test verisinin 19 tanesi hasta, 4 tanesi "normal" ve gerçekte 13 "normal" değerli test verisinin 12 tanesi "normal", 1 tanesi "hasta" olarak tahmin edilmiştir.

Tablo 8'de verilen Düzensizlik Matrisine göre 36 verinin detaylandırılmış doğruluk tablosu Tablo 9'da verilmiştir.

Tablo 9. ÇKA için detaylandırılmış doğruluk tablosu
(Table 9. Detailed confusion matrix for multilayer perceptron)

Doğruluk Yüzdesi	Duyarlılık (DP Oranı)	Yanlış Hassaslık Oranı (YP Oranı)	Hassaslık (Precision)	Sınıf (Class)
%86,11	0,826	0,077	0,95	a = hasta
	0,923	0,174	0,75	b = normal

Tablo 9'a göre Çok Katmanlı Algılayıcı için doğruluk yüzdesi, %86,11 olarak hesaplanmıştır. Duyarlılık, Yanlış Hassaslık Oranı ve Hassaslık değerleri, hasta ve normal sınıfları için ayrı ayrı hesaplanmıştır.

Weka'daki verilere uygun algoritmalar uygulanarak doğru sınıflandırma yüzdeleri ayrı ayrı bulunmuş ve bu sonuçlar Tablo 10.'da karşılaştırılmıştır.

Tablo 10. Sınıflandırma algoritmaları ve doğruluk yüzdeleri
(Table 10. Classification algorithms and accuracy percentages)

Algoritma Adı	Doğruluk Yüzdesi
Naive Bayes Algoritması	91,667
Çok Katmanlı Algılayıcı	86,111
C4.5 Algoritması	91,667

Çeşitli insan grupları üzerinde yapılan bu çalışmada yaş, cinsiyet, vücut kütle indeksi, Ürik Asit, HDL, LDL, Trigliserid düzeylerinin ve sigara kullanımının kan basıncını nasıl etkilediği incelenmiştir. Kişilerden elde edilen verilere, veri madenciliği sınıflandırma algoritmaları uygulanarak başarılı sonuçlar elde edilmiştir.

Veri madenciliği sınıflandırma algoritmalarından Naive Bayes ve C4.5, %91,667 ile en yüksek başarı oranını veren algoritmalar

olmuştur. Bu üç algoritmanın başarılarını kıyaslamak için ise Düzensizlik Matrisi'ne bakılması gerekir. Düzensizlik Matrisine göre C4.5 algoritması, 1 kişiyi gerçekte "hasta" iken "normal" olarak sınıflandırmış, Naive Bayes algoritması ise 3 kişiyi gerçekte "hasta" iken "normal" olarak sınıflandırmıştır. Gerçekte hasta olan bir kişiye sağlıklı teşhisi koymak, tıbben daha sakıncalı olduğu için, C4.5 algoritması, Naive Bayes algoritmasına göre daha başarılıdır.

C4.5 algoritmasının karar ağacı ekranına göre hipertansiyona etki eden faktörler incelendiğinde, Trigliserid düzeyinin 158'den büyük olması durumu, doğrudan hipertansiyon riskini ortaya koymaktadır. LDL'nin 50'den büyük olması durumu da hipertansiyon riskini ortaya koymaktadır. Hipertansiyona etki eden diğer bir faktör ise sigara kullanımıdır. Trigliserid ve LDL'ye bağlı olarak sigara kullanımının 10'dan büyük olması durumunda hipertansiyondan bahsedilebilir. HDL seviyesinin 139,4'ten büyük olması durumunda ise bakılması gereken faktör yaştır. Böyle bir durumda yaşın 49'dan büyük olması durumunda hipertansiyondan bahsedilebilir.

Bu durumda hipertansiyona en çok etki eden faktörlerin Trigliserid düzeyi, LDL değeri ve sigara kullanımı olduğu görülmüş, HDL değeri ve yaşın ise dolaylı olarak hipertansiyona etki ettiği saptanmıştır. Çalışmada kullanılan Vücut Kütle İndeksi (BMI) ve Ürik Asit değerlerinin ise herhangi bir etkisi gözlemlenememiştir. Bu durum ise, çalışmada kullanılan veri sayısının ve çeşitliliğinin azlığına yorulmuştur [13].

İleriki dönemlerde, bu çalışmada kullanılan veri miktarı artırılarak algoritmaların daha kapsamlı örüntüler oluşturup veri madenciliğinden daha iyi sonuçlar alınması sağlanabilir. Gelecekte bu uygulama, daha geniş bir veri seti kullanılarak, daha yüksek bir doğruluk oranıyla gerçekleştirilebilir.

5. SONUÇ (CONCLUSION)

Bu çalışmada kullanılan yöntem, hastaların demografik değerlerini ve kan değerlerini inceleyerek, hastaya cerrahi bir tetkik yapmadan hipertansiyon teşhisi yapmayı sağlamaktadır. Çalışma sayesinde hastaneye başka bir şikâyetinden dolayı giden bir kişi bile hipertansiyon hastası olduğunu öğrenebilir. Böylece birçok ciddi hastalıkların önlenmesi ve ömür boyu ilaç kullanılmasına gerek kalmadan sadece yaşam şeklini değiştirerek hipertansiyon ile başa çıkılabilir.

TEŞEKKÜR (ACKNOWLEDGEMENT)

Bu çalışma 2011/71 numaralı Kırıkkale Üniversitesi Bilimsel Araştırma Projesi olarak desteklenmiştir.

KAYNAKLAR (REFERENCES)

1. Chae, Y.M., Ho, S.E.,Cho, K.W.,Lee, D.H., and Ji, S.E., (2001). Data mining approach to policy analysis in a health insurance domain. International Journal of Medical Informatics, Volume: 62, Issues: 2-3, ss: 103-111.
2. Almazyad, A.S., Ahamad, M.G., and Siddiqui,M.K., (2010). Effective Hypertensive Treatment Using Data Mining In Saudi Arabia. Journal of Clinical Monitoring and Computing, Volume: 24, ss: 391-401.
3. Chang, C.D., Wang, C.C., and Jiang, B.C., (2011). Using Data Mining Techniques For Multi-Diseases Prediction Modeling of Hypertension and Hyperlipidemia by Common Risk Factors. Expert Systems with Applications, Volume: 38, ss :5507-5513.
4. Ture, M.,Kurt,I., Kurum, A.T., and Ozdamar, K., (2005). Comparing Classification Techniques for Predicting Essential

Hypertension. Expert Systems with Applications, Volume: 29, ss: 583-588.
5. Heart Disease and Stroke Statistics-2007 (2007). Update: A Report From the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Circulation. Volume: 115, ss: 69-171.
6. Summary of 1993 World Health Organisation-International Society of Hypertension Guidelines for The Management of Mild Hypertension Subcommittee of WHO/ISH Mild Hypertension Liaison Committee. (1993). BMJ.Volume: 307, ss: 1541-1546.
7. Şarlı, Ş., (2011). Hipertansiyon Hastalığı Olanlarda Tedaviye Uyum, Etkileyen Faktörler ve Yaşam Kalitesinin Değerlendirilmesi. Tıpta Uzmanlık Tezi, Kayseri: Erciyes Üniversitesi Tıp Fakültesi Halk Sağlığı Anabilim Dalı.
8. Lenfant C., Aram V., Daniel W., and Roccella, E.J., (2003). The seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure (JNC -7 Report). Hypertension, Volume: 41, ss: 1178.
9. Wang, J., (2006). Encyclopedia of Data Warehousing and Mining. Information Science Reference, Volume: 49, ss: 140.
10. Özkan, Y., (2008). Veri Madenciliği Yöntemleri. İstanbul: Papatya Yayıncılık.
11. Göktepe, A.B., Agar, E. and Lav, A.H., (2004). Comparison of Multilayer Perceptron and Adaptive Neuro-Fuzzy System on Backcalculating the Mechanical Properties of Flexible Pavements. ARI The Bulletin of the Istanbul Technical University, Volume: 54(3), ss: 65-77.
12. Haykin, S., (2010). Neural Networks and Learning Machines, PHI Learning Private Limited,.
13. Kökver, Y., (2012). Veri Madenciliğinin Nefroloji Alanına Uygulanması. Yüksek Lisans Tezi. Kırıkkale: Kırıkkale Üniversitesi Fen Bilimleri Enstitüsü.