İstatistik Araştırma Dergisi Cilt 03 No 01 Sayfa 73-83 Nisan 2004

HİPERTANSİYONUN TAHMİNİ İÇİN ÇOKLU TAHMİN MODELLERİNİN KARŞILAŞTIRILMASI

Mevlüt TÜRE*, İmran KURT**, Ebru YAVUZ**,
Turhan KÜRÜM***

ÖZET

Bu çalışmada, kontrol ve hipertansiyonlu hasta grubunun tahmini için lojistik regresyon analizi (LR), flexible diskriminant analizi (FDA) ve neural networks (NNs) karşılaştırıldı. Aile hikayesi, lipoprotein A, trigliserid, sigara kullanımı ve vücut kitle indeksi tahminleyici değişken olarak ele alındı. Veriler, 2001 yılında Trakya Üniversitesi Tıp Fakültesi Kardiyoloji Kliniğinden elde edildi. Bütün modellerin ROC eğrisi altındaki alanları 0.793-0.984 aralığında yer aldı. NNs'nin duyarlılık, özgüllük ve doğruluk oranları %90'dan daha yüksek bulundu. NNs ve LR ile NNs ve FDA'nın ROC eğrisi altında kalan alanları istatistiksel olarak farklı bulundu (sırasıyla $p<0.0005$ ve $p<0.0005$). FDA ve LR'nin ROC eğrisi altında kalan alanları istatistiksel olarak farklı bulunmadı ($p=0.394$). NNs'nin performansının LR ile FDA'dan istatistiksel olarak daha iyi olduğuna karar verildi.

Anahtar Kelimeler: *Lojistik Regresyon Analizi, Neural Networks, Flexible Diskriminant Analizi, ROC Eğrisi*

1. GİRİŞ

Hipertansiyon, normal düzeyin üzerine çıkan kan basıncı seviyesi olarak tanımlanır. Bu durum kalp, beyin, böbrekler ve gözlerde hasara neden olabilir. Çünkü kalp ve arterler, kendilerine oksijen ve besinleri göndermek için daha sıkı çalışmak zorundadırlar. Hipertansiyon için bir çok risk faktörü vardır. Bu çalışmada risk faktörü olarak aile hikayesi, lipoprotein A, trigliserid, sigara kullanımı ve vücut kitle indeksi değişkenleri yer almaktadır (Jo vd, 2001; Cheitlin vd,1993; Hall vd, 1994; Cruickshank vd, 1989; Williams, 1989; Williams vd, 1989; Spers vd, 1986; Elford vd, 1990; Folkow, 1993, Kaplan, 1994, Schieken, 1993).

* Yard. Doç. Dr., Trakya Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim Dalı, EDİRNE (Haberleşme Adresi)
** Araş. Gör., Trakya Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim Dalı, EDİRNE
*** Doç. Dr., Trakya Üniversitesi Tıp Fakültesi Kardiyoloji Anabilim Dalı, EDİRNE

Bu çalışmada, kontrol ve hipertansiyonlu bireylerden oluşan grupların tahminlenmesinde lojistik regresyon analizi (LR), flexible diskriminant analizi (FDA) ve neural networks (NNs) yöntemlerinin karşılaştırılması amaçlandı.

- Doğrusal diskriminant analizi (DDA), grupları ortalamalarına göre ortak ortalamadan farklı olmalarını sağlayacak bir ayırma kriteri geliştirmeyi amaçlayan bir yöntemdir. Bu nedenle veri setlerine DDA uygulanabilmesi için veri setlerinin aşağıdaki varsayımları taşıması gerekir (Özdamar, 1999; Tatlıdil, 1996).

  I. Veri matrisi çok değişkenli normal dağılım göstermelidir.
  II. Değişkenlerin varyans ve kovaryansları homojen olmalıdır.
  III. Değişkenlerin ortalamaları ve varyansları arasında ilişki bulunmamalıdır.
  IV. Değişkenler arasında çoklu bağımlılık bulunmamalıdır.
  V. Veri matrisi grupların birbirlerinden ayrılmasında rol oynamayacak gereksiz değişkenler içermemelidir.

- FDA, DDA'nın varsayımlarının getirmiş olduğu kısıtlamalar nedeniyle ortaya çıkan problemleri ortadan kaldıran ve DDA'ya göre daha esnek parametrik olmayan bir yöntemdir (Hastie vd, 1995; Hastie vd, 1994; Hastie ve Tibshirani, 1996; Kim ve Loh, 2002).

- LR, bağımsız değişkenlerin dağılımı ile ilgili her hangi bir varsayımı olmayan bir yöntemdir. Bu yüzden LR, doğrusal diskriminant analizinden daha geniş bir araştırma alanında kullanılır (Özdamar, 1999; Tatlıdil, 1996).

- NNs ise değişken yapıları konusunda her hangi bir varsayım gerektirmeyen, sınıflandırma için uygun çok esnek bir yöntemdir (Lee, 2000). NNs, ağ hatasının optimal değerini bulmak için geriye doğru yayılma algoritmasını kullanarak aşama aşama en iyi tahmine ulaşmaya çalışır (Abdul-Kareem vd, 2001; Fine, c1999; Haykin, 1999; Hassoun, c1995).

2. LOJİSTİK REGRESYON ANALİZİ

LR, sınıflama ve atama işlemi yapmaya yardımcı olan bir regresyon yöntemidir. Normal dağılım varsayımı, süreklilik varsayımı ön koşulu yoktur (Özdamar, 1999).

P bağımsız değişken için LR modeli aşağıdaki gibi yazılabilir;

$$P(Y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}}$$

Burada $\beta_0, \beta_{1,\ldots,}\beta_p$, regresyon katsayılarıdır. Bu katsayılar,

$$\ln\left(odds\left(\frac{P(Y)}{1-P(Y)} \right) \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

$$odds = \frac{P(Y)}{1-P(Y)} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+....+\beta_p X_p} = e^{\beta_0} e^{\beta_1 x_1} e^{\beta_2 X_2} .....e^{\beta_p X_p}$$

şeklinde hesaplanır (Özdamar, 1999).

3. FLEXIBLE DİSKRİMİNANT ANALİZİ

FDA, doğrusal diskriminant analizine göre esnek ve parametrik olmayan alternatif bir yöntemdir.

θ: {1,...,j} $\rightarrow R^1$, doğrusal regresyon kullanılarak X bağımsız değişkenleriyle optimal olarak tahmin edilen dönüştürülmüş sınıfların atanan skorlarının bir fonksiyonu olsun. Bu, sınıflar arası tek boyutlu bir ayrımı meydana getirir. Daha genel olarak, θ_1, θ_2,..., θ_k sınıfları için K bağımsız skorlar setini ve R^p uzayında çoklu regresyon için optimal olarak seçilen $\eta_k(X) = X^T\beta_k$ (k=1,...,K) K doğrusal grafiği bulacağız. Eğer örneklem (g_i, x_i), i=1,2,...,N, biçiminde ise, $\theta_k(g)$ skorları ve β_k gösterimleri ortalama karesel artık (ASR) değerini minimuma indirmek için seçilir,

$$ASR = \frac{1}{N}\sum_{k=1}^{K}\left[\sum_{i=1}^{N} \left(\theta_k(g_i) - x_i^T\beta_k \right)^2 \right]$$

Skorlar setinin, karşılıklı olarak ortagonal olduğu varsayılır ve önemsiz çözümleri önlemek için uygun bir iç çarpımla ilgili olarak normalleştirilir. Bu, u_k kanonik vektörlerin sırasının bir sabite kadar β_k sırası ile aynı olduğunu gösterir.

Ayrıca, birinci K kanonik vektörüyle tanımlanan alt uzayla sınırlanan j. sınıf merkezi $\hat{\mu}_k$ için bir x test noktasının Mahalanobis uzaklığı,

$$\delta_K(x, \hat{\mu}_j) = \sum_{k=1}^{K} w_k \left(\eta_k(k) - \overline{\eta}_k^j \right)^2$$

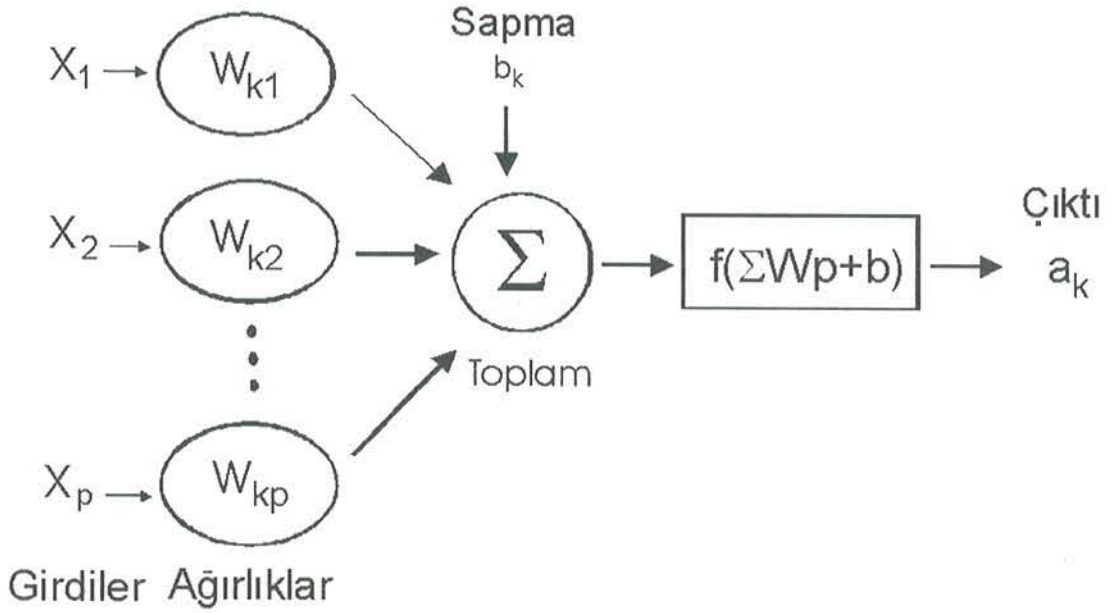
biçiminde tanımlanır. Burada $\overline{\eta}_k^j$, $g_i = j$ için $\eta_k(x_i)$'nın ortalamasıdır. w_k, k. optimal olarak skorlanmış uyumun ortalama karesel artığı r_k^2'ye dayanarak tanımlanan koordinat ağırlıklardır (Hastie vd, 1995; Hastie vd, 1994; Hastie ve Tibshirani, 1996; Kim ve Loh, 2002):

$$w_k = \frac{1}{r_k^2\left(1-r_k^2\right)}$$

4. NEURAL NETWORKS

NNs, beynin çalışma ilkelerinin bilgisayarlar üzerinde taklit edilmesi fikri ile ortaya çıkmış ve ilk çalışmalar beyni oluşturan biyolojik hücrelerin yada nöronların matematiksel olarak modellenmesi üzerinde yoğunlaşmıştır. Günümüzde NNs, bir çok nöronun belirli biçimlerde bir araya getirilip bir işlevin gerçeklenmesi üzerindeki yapısal olduğu kadar matematiksel ve felsefi sorunlara yanıt arayan bir alan olmuştur. Bu yöntem, nöronlar olarak bahsedilen basit hesap hücrelerinin birbirleriyle bağlantılarını kullanarak insan beyninde olduğu gibi bilgiyi kaybetmeden hedefe doğru minimum hata ile ulaşılmasını sağlar (Abdul-Kareem vd, 2001; Demuth vd, 2000; Francis, 2001; Fine, c1999; Haykin, 1999).

Basit bir ağda, **p** girdi vektörü ile **W** ağırlık matrisi çarpılır ve bu çarpıma **b** sapmasının eklenmesiyle f dönüşüm fonksiyonunun girdisi elde edilir. Ağ, elde edilen bu yeni girdiyi f dönüşüm fonksiyonu yardımıyla ağ içinden ileterek **a** çıktısını oluşturur (Şekil-1) (Abdul-Kareem vd, 2001; Demuth vd, 2000; Fine, c1999; Haykin, 1999).



Şekil1. f dönüşüm fonksiyonlu en basit neural networks

NNs'de sıkça kullanılan dönüşüm fonksiyonları: lojistik dönüşüm fonksiyonu,

$$f(\sum wp+b)=\frac{1}{\left(1+e^{-(\sum wp+b)}\right)}$$

ve hiperbolik tanjant sigmoid fonksiyondur (Lee, 2000; Abdul-Kareem vd, 2001).

$$f(\sum wp + b) = \frac{2}{\left(1 + e^{-2(\sum wp+b)}\right)} - 1$$

Genellikle ağ yapıları üç temel sınıfta tanımlanır: tek tabakalı ağlar, çok tabakalı ağlar ve yinelenen ağlar. Tek tabakalı ağ, tabakalar biçiminde düzenlenmiş nöronlardan oluşan ağdır. Çok tabakalı ağ, gizli nöronlar olarak isimlendirilen düğümlerden oluşan bir yada daha çok gizli tabakaya sahip olan ağlardır. Yinelenen ağ, en az bir geri bildirim döngüsüne sahip olan ağdır (Fine, c1999; Haykin, 1999; Hassoun, c1995).

Bu çalışmada çok tabakalı ağ kullanılmıştır. Bu ağ zor ve farklı problemleri başarılı olarak çözen geriye doğru yayılma algoritması olarak bilinen bir algoritmayla denetlenerek çalıştırılır. Geriye doğru yayılma algoritması, ağ hatasının minimize edilmesi için ağın ağırlıkları ve sapmalarının hesaplanması sürecidir (Abdul-Kareem vd, 2001; Fine, c1999; Haykin, 1999; Hassoun, c1995).

5. UYGULAMA

Çalışmamız, 2001 yılında Trakya Üniversitesi Tıp Fakültesi Kardiyoloji Polikliniğine gelen 236 hasta ve 123 bireylik kontrol gruplarından oluşmaktadır. Aile hikayesi, lipoprotein A, trigliserid, sigara kullanımı ve vücut kitle indeksi bağımsız değişkenler olarak ele alındı. Değişkenlere ilişkin tanımlayıcı istatistikler Tablo-1'de verilmiştir.

Tablo 1. Bağımsız değişkenlerin gruplara göre tanımlayıcı istatistikleri

Bağımsız Değişkenler	**Hasta Grubu n=236**	**Kontrol Grubu n=123**
Aile Hikayesinde Hipertansiyon		
- Olan	%77.7	%33.6
- Olmayan	%22.3	%66.4
Lipoprotein A (mg/dl)	33.42±24.14	26.60±16.88
Trigliserid (mg/dl)	164.13±82.22	130.07±59.18
Sigara Kullanımı		
-İçen	%23.8	%32.7
-İçmeyen	%76.2	%67.3
Vücut Kitle İndeksi (kg/m^2)	28.80±3.62	27.28±3.64

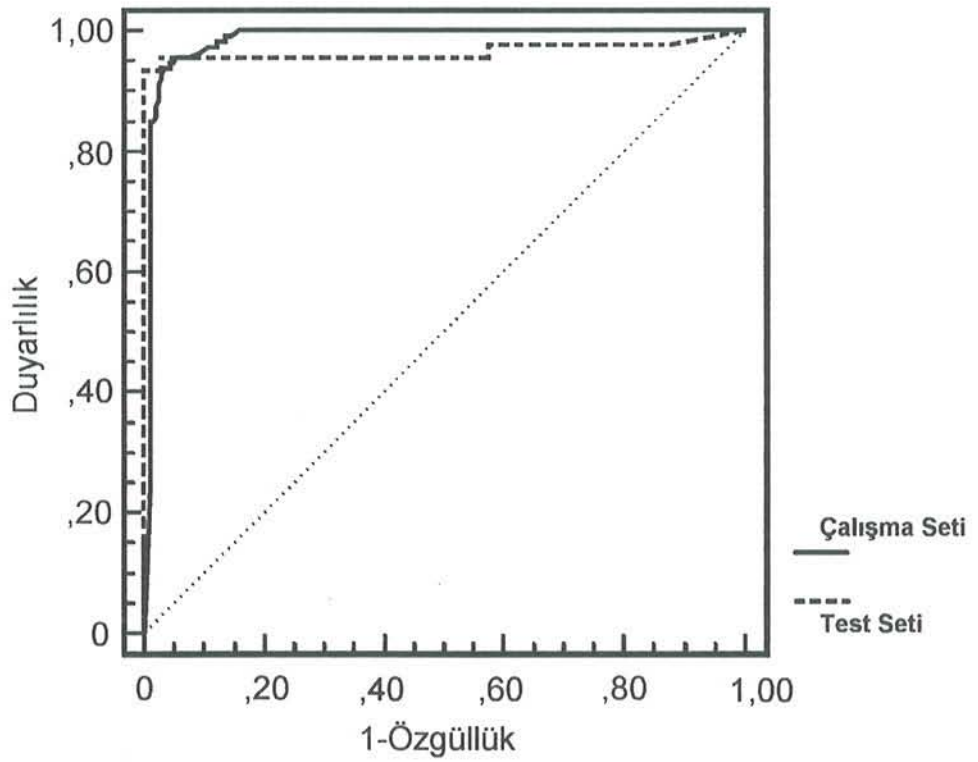
Bu çalışma, 2 aşamadan oluşmaktadır. 1. aşama NNs'nin çalışma ve test sonuçlarının incelenmesi, 2. aşama ise NNs ile LR ve FDA'nın karşılaştırılmasıdır.

NNs'nin çalışma ve test sonuçlarının karşılaştırılması

NNs modeli oluşturulmadan önce veri seti, 276 birimlik çalışma (359 birimin %77'si) ve 83 birimlik (359 birimin %23'ü) test seti olmak üzere iki gruba ayrıldı.

NNs, 15 gizli nöronlu (hiperbolik tanjant sigmoid dönüşüm fonksiyonlu) ve 1 ikili (hasta-kontrol) çıktılı (hiperbolik tanjant sigmoid dönüşüm fonksiyonlu) ağdan oluşmaktadır. NNs'de hatayı minimize etmek amacıyla geriye doğru yayılma algoritması kullanıldı.

Çalışma setinden elde edilen NN modeli test setine uygulandı. Çalışma ve test setinin performansları karşılaştırıldığında sonuçlarının benzer oldukları görüldü (Tablo-2, Şekil-2).



Şekil 2. Çalışma ve test setlerinin ROC eğrileri

Tablo 2. Çalışma ve test setlerinin duyarlılık ve özgüllük oranları, ROC eğrisi altında kalan alanları, standart hataları ve ROC eğrisi altında kalan alanın güven aralığı.

	Çalışma Seti	Test Seti
Duyarlılık (%)	97.1	96.6
Özgüllük (%)	92.7	93.8
ROC Eğrisi Altında Kalan Alan	0.984	0.967
Standart Hata	0.006	0.019
ROC Eğrisi Altında Kalan Alanın %95 Güven Aralığı	0.964 - 0.995	0.901 - 0.994

NNs'nin LR ve FDA ile karşılaştırılması

Tablo-3: Lojistik regresyon ve flexible diskriminant analizi sonucunda elde edilen bağımsız değişkenlere ilişkin parametre tahminleri

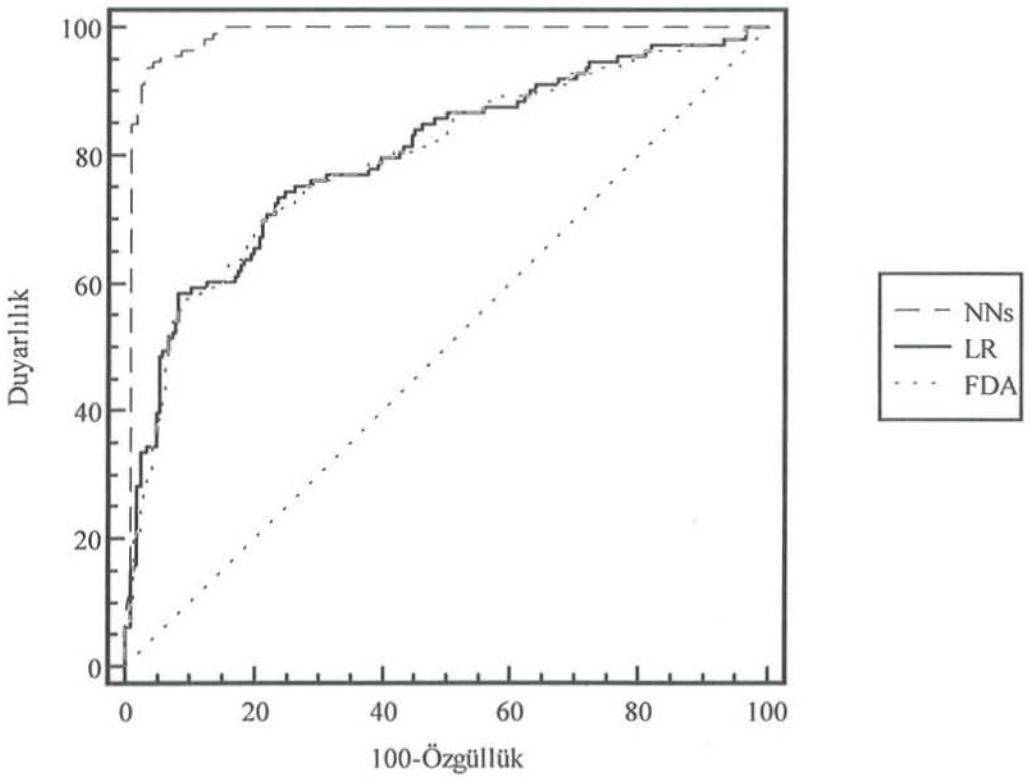
Bağımsız Değişkenler	Lojistik Regresyon Analizi (β_i)	Flexible Diskriminant Analizi (β_i)	p*
Sabit	-5.028	-5.545	
Aile Hikayesinde Hipertansiyon	1.919	1.853	<0.0005
Lipoprotein A (mg/dl)	0.014	0.011	0.042
Trigliserid (mg/dl)	0.008	0.005	<0.0005
Sigara Kullanımı	0.761	0.640	0.015
Vücut Kitle İndeksi (kg/m^2)	0.088	0.075	0.021

*: LR sonucu elde edilen p değerleri

LR ve FDA sonucunda elde edilen parametre tahminleri Tablo-3'de verilmiştir. LR, FDA ve NNs'nin ROC eğrisi altındaki alanları 0.793-0.979 aralığında elde edildi (Tablo-4). Hanley ve McNeil tarafından önerilen yöntem kullanılarak iki ROC eğrisi altındaki alanlar arasındaki farkın istatistiksel anlamlılığı test edildi. LR, FDA ve NNs'nin ROC eğrisi altında kalan alanları Şekil-3'de gösterilmiştir. NNs ile LR ve NNs ile FDA'nın ROC eğrisi altında kalan alanları istatistiksel olarak farklı bulundu (sırasıyla $p<0.0005$ ve $p<0.0005$). FDA ile LR'nin ROC eğrisi altında kalan alanları istatistiksel olarak faklı bulunmadı ($p=0.394$). Ayrıca eğri altında kalan alanların standart hataları incelendiğinde, NNs'nin standart hatasının diğerlerine göre daha küçük olduğu görüldü.

Tablo 4. Duyarlılık, özgüllük, pozitif tanımlama, negatif tanımlama ve doğruluk oranları, ROC eğrisi altında kalan alanlar, kesim noktaları ve standart hatalar.

	Lojistik Regresyon Analizi	Flexible Diskriminant Analizi	Neural Networks
Duyarlılık (%)	85.4	82.0	97.1
Özgüllük (%)	60.2	63.7	92.7
Pozitif Tanımlama Oranı (%)	79.6	80.5	98.5
Negatif Tanımlama Oranı (%)	69.4	66.1	88.0
Doğruluk (%)	76.5	75.5	94.4
ROC Eğrisi Altında Kalan Alan	0.800	0.793	0.984
Kesim Noktası	0.499	0.498	0.433
Standart Hata	0.028	0.027	0.006



Şekil-3: LR, FDA ve NNs için ROC eğrileri

6. SONUÇ

Bu çalışmada LR, FDA ve NNs'nin sınıflandırma performansları karşılaştırıldı. Bütün modellerin ROC eğrisi altındaki alanları 0.793-0.979 aralığında elde edildi. NNs, LR ve FDA'dan istatistiksel olarak daha farklı bulundu. NNs'de duyarlılık, özgüllük ve doğruluk oranları %90'dan daha büyük bulundu.

Sonuç olarak, aile hikayesi, lipoprotein A, trigliserid, sigara kullanımı ve vücut kitle indeksi değişkenlerinin, kontrol ve hipertansiyonlu hasta gruplarını tahmin etmede bir kriter olarak kullanılabileceğini, NNs'nin grupları sınıflandırma gücünün klasik çok değişkenli tekniklere göre daha iyi olduğunu söyleyebiliriz.

KAYNAKLAR

ABDUL-KAREEM, S., BABA, S. and ZUBAIRI Y.Z. (2001). *Back Propagation Neural Network for Medical Prognosis: A Comparison of Different Training Algorithms*, Erişim: [http://www.sat.ait.ac.th/ejat/ articles/3.1/main.html]. Erişim Tarihi: 20.03.2002

CHEITLIN, M.D., SOKOLOW, M. and MCLLROY, M.B. (1993). *Systemic Hypertension. Clinical Cardiology*, Prentice-Hall Int. Inc., A Lange Medical Book.

CRUICKSHANK, J.M., NEIL-DWYER, G., et al. (1989). *Acute Effects of Smoking on Blood Pressure and Cerebral Blood Flow,* J. Hum. Hypertens, 3: 443.

DEMUTH, H., and BEALE, M. (2001). *Neural Network Toolbox User's Guide*, The Mathworks, Inc.: USA.

ELFORD, J., PHILLIPS, A., et al. (1990). *Migration and Geographic Variations in Blood Pressure in Britain,* Br. Med. J., 300:291.

FINE, T.L. (c1999). *Feedforward Neural Network Methodology,* Springer: New York.

FOLKOW, R. (1993). *The Pathophysiology of Hypertension. Differences Between Young and Erderly Patients,* Drugs, 46 (suppl) 2:3-7.

FRANCIS, L. (2001). *The Basics of Neural Networks Demystified*, Erişim: [http://www.contingencies.org/novdec01/workshop.pdf]. Erişim Tarihi: 9.01.2002

HALL, W.D., WOLLAN, G.L. and TUTTLE, E.P. (1994). *Diagnostic Evaluation of The Patient with Systemic Arterial Hypertension*: An Overview, Hurst et al. (der.), The Heart. New York: McGraw-Hill Inc., 10.

HASSOUN, M.H. (c1995). *Fundamentals of Artificial Neural Networks.* Cambridge, Mass:MIT Press.

HASTIE, T., TIBSHIRANI, R. and BUJA, A.(1995). *Flexible Discriminant and Mixture Models*, In Kay, J. & Titterington, D., (eds.) Neural Networks and Statistics. Oxford University Press.

HASTIE, T., TIBSHIRANI, R. and BUJA, A.(1994). *Flexible Discriminant Analysis by Optimal Scoring,* Journal of The American Statistical Association 89: 1255-1270.

HASTIE, T. and TIBSHIRANI, R. (1996). *Discriminant Analysis by Gaussian Mixtures*, Journal of The Royal Statistical Society (B), 58, 155-176.

HAYKIN, S. (1999). *Neural Network: A Comprehensive Foundation*, Upper Saddle River, NJ:Prentice Hall.

JO, I., AHN, Y., LEE, J., SHIN, K.R., et al. (2001). *Prevalence, Awareness, Treatment, Control and Risk Factors of Hypertension in Korea: The Ansan Study*, Journal of Hypertension, 19 (9): 1523-1532.

JOHNSON, R,A., WICHERN, D.W. (1982). *Applied Multivariate Statistical Analysis,* Prentice-Hall: New Jersey.

KAPLAN, N.M. (1994). *Clinical Hypertension,* Baltimore, Williams and Wilkins.

KIM, H. and LOH, W.-Y. (2002*). Classification trees with bivariate linear discriminant node models, Journal of Computational and Graphical Statistics*, in press. [This paper extends CRUISE to fit linear discriminant models in terminal nodes.]

LEE, H.K.H. (2000). *Model Selection for Neural Network Classification*, Erişim: [http://ftp.isds.duke.edu/WorkingPapers/00-18.pdf]. Erişim Tarihi: 02.04.2002

ÖZDAMAR, K. (1999). *Paket Programlar ile İstatistiksel Veri Analiz 1,* Eskişehir, Kaan Kitabevi.

ÖZDAMAR, K. (1999). *Paket Programlar ile İstatistiksel Veri Analiz 2,* Eskişehir, Kaan Kitabevi.

SCHIEKEN, R.M. (1993). *Genetic Factors That Predispoze The Child to Develop Hypertension,* Pediatr Clin North Am., 40:1.

SHARMA, S. (1996). *Applied Multivariate Techniques*, John Wiley & Sons: New York.

SPERS, M.A., KASL, S.V., et al. (1986). *Blood Pressure Concordance Between Spouses,* Am. J. Epidemiol., 123:818.

TATLIDİL, H. (1996). Uygulamalı Çok Değişkenli İstatistiksel Analiz, Ankara, Akademi Matbaası.

WILLIAMS, R.R., HUNT, S.C., et al. (1989). *Current Knowledge Regarding The Genetics of Human Hypertension,* J. Hypertens., 7(Suppl 6):8.

WILLIAMS, R.R. (1989). *Will Gene Markers Predict Hypertension?,* Hypertension, 14: 610.

COMPARISON OF MULTIPLE PREDICTION MODELS FOR HYPERTENSION

ABSTRACT

In this study, we compared logistic regression analysis (LR), flexible discriminant analysis (FDA) and neural networks (NNs) for predict of control and hypertension groups. Predictor variables were family history, lipoprotein A, triglyceride, smoking and body mass index. The data were collected from Cardiology Clinic of Trakya University Medical Faculty in Turkey, 2001. All models had areas under the receiver operating characteristic curve (ROC) in the 0.793-0.984 range. NNs had sensitivity, specificity, and accuracy greater than 90% at ideal threshold. ROC curve areas of NNs and LR, and NNs and FDA were statistically different ($p<0.0005$ and $p<0.0005$ respectively). ROC curve areas of FDA and LR were not statistically different ($p=0.394$). We concluded that performance of NNs was statistically better than LR and FDA.

Key Words: *Neural Networks, Flexible Discriminant Analysis, Logistic Regression, ROC Curve*