ÖĞRENCİ BAŞARILARININ SINIFLANDIRILMASINDA LOJİSTİK REGRESYON ANALİZİ ve SİNİR AĞLARI YAKLAŞIMI

Nuray GÜNERİ *
Ayşen APAYDIN **

ÖZET

Bu çalışma öğrenci başarısızlıklarının nedenlerini tanımlamak ve böylece gelecekte karşılaşılabilecek başarısızlıkları kestirmek için sinir ağları ile lojistik regresyon yöntemini karşılaştırmayı hedeflemektedir. Lojistik regresyon ve sinir ağları yöntemleri bireylerin doğru sınıflandırılma oranlarına göre karşılaştırılmıştır. Yöntemler, Gazi Üniversitesi Ticaret ve Turizm Eğitim Fakültesi'nden alınan veriler üzerine uygulanmış ve sinir ağları yönteminden elde edilen oranın, lojistik regresyon yönteminden elde edilen orana eşit olduğu görülmüştür.

Anahtar Kelimeler : Sinir ağları, geri besleme, lojistik regresyon

LOGISTIC REGRESSION ANALYSIS and NEURAL NETWORKS APPROACH in THE CLASSIFICATION of STUDENTS' ACHIEVEMENT

ABSTRACT

This study aimed to compare neural networks with logistic regression method to identify causes of student failures and therefore predicting future failures. Logistic regression and neural networks methods have been compared with respect to their correct classification probabilities of individuals. These methods have been applied to a data set taken from Gazi Üniversity, Faculty of Commerce and Tourism Education and it is observed that the correct classification probability obtained from neural networks is equals to the correct classification probability obtained from logistic regression.

Key Words: Neural network, backpropagation, logistic regression

---

* Gazi Üniversitesi Ticaret ve Turizm Eğitim Fakültesi (Arş. Gör.)
** Ankara Üniversitesi Fen Fakültesi (Prof. Dr.)

GİRİŞ

Öğrenci başarısızlıklarının nedenlerini tanımlamak ve böylece gelecekte karşılaşılabilecek başarısızlıkları kestirerek öğrencilerinin başarı olasılıklarını en büyüklemek herhangi bir akademik bölümün en önemli amaçlarından biridir. Bu amaçla bir tahmin modeli kurabilmek için çalışmalar yapılmaktadır. Lojistik regresyon analizi, bağımlı değişkenin tahminini olasılık olarak hesaplayarak çok değişkenli istatistiksel verilerin sınıflandırılmasında kullanılan bir yöntemdir. Regresyon modellerinin çözüm yöntemlerinden biri de son yıllarda kullanılmaya başlanan sinir ağları yaklaşımıdır. Sinir ağları yaklaşımının kullanılması ile, lojistik regresyon analizinde olduğu gibi verilerin sınıflandırılması mümkün olabilmektedir. Bu çalışmada, öğrencilerin başarılarına göre sınıflandırılmasında sinir ağları analizi ile lojistik regresyon analizi kullanılacak ve bulunan sonuçlar karşılaştırılacaktır.

Sinir ağları insan beyninin çalışmasından esinlenerek, biyolojik sinir sistemi gibi hareket edecek ağ modellerinin kurulabilmesi için geliştirilen bir yöntemdir. Yapay sinir ağları farklı uygulamalarda kesin hesaplamaları yerine getirmek için günümüzdeki mevcut en hızlı bilgisayarlardan daha hızlıdır. Karmaşık ya da belirsiz olan verilerden anlam çıkarabilir. Bu özelliklerinden dolayı bilgi sınıflama ve bilgi yorumlamanın da içinde bulunduğu çok değişik problemlerin çözümünde kullanılmasının yanı sıra, iş hayatı, finans, endüstri ve eğitimde var olan yöntemlerin yerine veya doğrusal olmayan sistemlerde başarıyla uygulanmaktadır. Geri beslemeli ağ yapıları kullanılarak, regresyon çözümlemesinde olduğu gibi bağımlı değişken için uygun bir tahmin modeli kurulabilmektedir.

Sinir Ağları'nın başlangıcı 1940'lara dayanmaktadır. İlk olarak 1943 yılında bir sinir hekimi olan Warren McCulloch ile bir matematikçi olan Walter Pitts beynin hesaplama gücünün kaynağını araştırıp sinir sistemini inceleyerek, sinir sisteminde çok sayıda basit sinirin oluşturduğu bir birlik olduğunu bulmuşlar ve elektrik devreleriyle ilk sinir ağı modelini oluşturmuşlardır. 1949 yılında ise Hebb "Organization of Behavior" isimli kitabında öğrenme ile ilgili temel teoriyi ele almıştır. Hebb, öğrenebilen ve uyum sağlayabilen sinirler ve sinirlerin aralarındaki bağlantılar için öğrenme kuralını geliştirmiştir. Bunlar Hebb Kuralı olarak bilinmekte ve günümüzde kullanılan sinir ağları modellerinin temelini oluşturmaktadır (Fausset 1994:22, Elmas 2003:27). Fakat bundan sonra yaklaşık yirmi yıllık bir periyodda sinir ağları üzerinde yapılan çalışmalar önemli ölçüde azalmıştır. 1970'li yıllarda yapılan çalışmaların en önemlisi çok tabakalı ağlar için eğitim modelinin bulunuşu olmuştur. Geri besleme ağı olarak adlandırılan

yeni bir öğrenme algoritması geliştirilmiştir. Geri besleme yöntemi ile doğrusal ayrılabilir olmayan problemler çözülebilmektedir (Warner ve Mısra 1996:285).

1980'li yılların sonunda sinir ağlarının istatistikte kullanımı ile ilgili yoğun çalışmalar yapılmış ve 1990'lı yıllarda çalışmalar yeniden hız kazanmıştır. White (1989), sinir modelleri ile istatistiksel yaklaşım arasındaki ilişkiyi tanımlamıştır. Dutta ve Shekkar (1988) sinir ağlarının çoklu regresyon üzerinde parametre tahmininde bulunabildiğini göstermiş; Denton, Hung ve Osyk (1990), sinir ağlarının diskriminant analizinden daha iyi sonuç verdiğini bulmuş; Tang, Almeida ve Fishwish (1991) ve Sharda ve Patil (1990), zaman serileri ile sinir ağlarının çalışmalarını karşılaştırmışlardır (Collins ve Clark 1993:504, Adıgüzel 1999:16-17). Warner ve Mısra (1996), yayınladıkları çalışmalarında sinir ağlarını regresyon modelleriyle karşılaştırmışlardır. Stergiou ve Siganos (1996), yapay sinir ağlarının farklı çeşitleri ve uygulama alanlarını araştırmışlar ve matematiksel modellemesi üzerinde çalışmışlardır.

Öğrenci başarısızlıklarının tahmininde sinir ağlarının kullanımı ile ilgili çalışmalar ilk kez 1994 yılında yapılmıştır. Gorr, Nagin ve Szczypula (1994), öğrencilerin ağırlıklı not ortalamasının tahmininde istatistik metotlarından çoklu lineer regresyon ve stepwise lineer regresyon analizleri ile sinir ağlarını karşılaştırmış; Hardgrave, Wilson ve Walstrom (1994), öğrencilerin akademik başarılarının tahmininde beş farklı modeli (en küçük kareler, stepwise, diskriminant analizi, lojistik regresyon analizi, sinir ağları) inceleyerek, karşılaştırmalar yapmıştır. Flitman (1997) da, öğrencilerin akademik ortalamaları üzerinde yaptığı çalışmada, sinir ağları ile diskriminant analizi ve lojistik regresyon analizini karşılaştırmış, sonuçta sinir ağları ile yapılan tahminlerin daha iyi olduklarını göstermiştir.

Bu çalışmanın Birinci Bölümünde, yapay sinir ağları, yapay sinir ağı modelleri ve regresyon çözümlemesinde kullanılan geri beslemeli ağ sistemleri konuları incelenecektir.

İkinci Bölümünde lojistik regresyon modeli tanımlanacaktır.

Üçüncü Bölümünde öğrencileri başarılarına göre sınıflandırmak için lojistik regresyon ve sinir ağları yönteminin gerçek veriler üzerinde uygulaması yapılacaktır.

Sonuç bölümünde ise iki yöntemden elde edilen sonuçlar karşılaştırılacak ve yorumlanacaktır.

1. YAPAY SİNİR AĞLARI

Yapay sinir ağları, insan beyninin özelliklerinden olan öğrenme yolu ile yeni bilgiler türetebilme, yeni bilgiler oluşturabilme ve keşfedebilme gibi yetenekleri herhangi bir yardım almadan otomatik olarak gerçekleştirmek amacı ile geliştirilen bilgisayar sistemleridir. İnsan beynine benzer şekilde öğrenme,

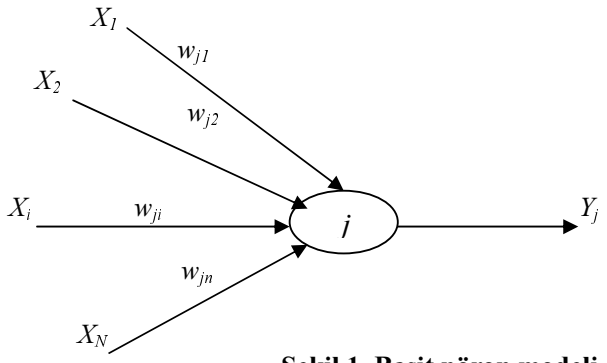
ilişkilendirme, sınıflandırma, genelleme, özellik belirleme ve optimizasyon gibi konularda uygulanmaktadırlar. Sinirler, örneklerden elde ettikleri bilgiler ile kendi deneyimlerini oluşturur ve daha sonra benzer kararları verirler (Öztemel 2003:29).

Yapay sinir ağları, birbirleriyle bağlantılı çok sayıda basit sinirin oluşturduğu, merkezi sinir sisteminin basitleştirilmiş modelleridir. Bu model bir bilgi süreci sistemi gibi düşünülebilir. Ağı oluşturan sinirler, girdi sinyallerine tepki verme ve çevreye adaptasyonu öğrenme yeteneğine sahiptir ve beynin yapısını taklit ederek matematiksel hesaplamalar yapabilmektedir. Bu hesaplamalar sistemin en temel elemanı olan nöronlar (sinirler) ve nöronların birbirleriyle olan bağlantıları yardımıyla gerçekleşmektedir. Sinir ağları birbirleriyle yüksek bağlantıya sahip bu nöronlardan oluşmaktadır. Nöronların sayısı ağın yapısını belirler. Sayı büyük olduğunda bağlantılar karışabilmektedir.

Basit bir nöron modeli Şekil 1.' de görülmektedir. Burada, X_i girdileri ($i=1,2,...,N$), w_{ji} ağırlık katsayılarını ($i=1,2,...,N$; $j=1,2,...,M$), Y_k çıktıları ($k=1,2,...,Z$) göstermektedir.

Şekil 1.' de oklar bağlantıların ve akış sinyalinin yönünü gösterirler. Sinyaller sadece oklarla gösterilen yön boyunca iletilirler. İki nöron arasındaki bağlantı, girdi nöronlarının çıktı nöronları üzerindeki etkisini gösteren sayısal değerlere sahiptir. Bu değerlere ağırlık denir (Hwang ve Ding 1997:748).

Her bir nörondan çıkan sinyaller, nöronların arasındaki bağlantıların ağırlık değerleriyle modüle edilirler. Ağırlıklar girdi sinyallerinin yoğunluğunu gösteren adaptasyon katsayıları olarak da tanımlanabilir (Warner ve Mısra 1996:287).



Şekil 1. Basit nöron modeli

Bir sinir ağı modeli için, t zamanında, ağ içerisindeki bir nörona gelen net girdi sinyallerinin toplamı,

$$net_j(t) = \sum_{i=1}^{N} w_{ji}(t)\, x_i(t) \tag{1.1}$$

eşitliği ile tanımlanır. Eşitlik (1.1)' de

$net_j(t)$: ağın j. nöronuna gelen net girdi sinyali

$w_{ji}(t)$: i. nöron ile j. nöron arasındaki bağlantının ağırlığı

$x_i(t)$: i. nörondan j. nörona giden çıktılar

N : j. nöron ile bağlantıya sahip diğer birimlerin sayısı

dır.

Her bir nörondaki hesaplama bir dönüşüm sürecidir. Bu süreç aktivasyon fonksiyonu adı verilen bir transfer fonksiyonu yardımıyla gerçekleşmektedir (Veaux vd. 1998:274). Ağlarda yer alan sinir gövdeleri yardımıyla karşılaştırma yapılmaktadır. Herhangi iki nöronun arasından geçen sinyal, hem aktivasyon fonksiyonuna hem de bağlantının ağırlığına bağlıdır (Flitman 1997:368). Modelde yer alan bir nöronun girdilerinin ağırlıklandırılmış toplamı yani net girdi değeri bir aktivasyon fonksiyonundan geçer. Bu fonksiyondan elde edilen sonuç değer, bu nöronun çıktı değeridir. Sonuçta elde edilen çıktılar matematiksel olarak,

$$Y_j(t) = g_j(net_j(t)) \tag{1.2}$$

eşitliği ile hesaplanır. Eşitlik (1.2)' de

$g_j(.)$: aktivasyon fonksiyonunu,

$Y_j(t)$: j. nöronun çıkışını

göstermektedir.

Bir modelde kullanılan aktivasyon fonksiyonunun belirlenmesi bazı faktörlere bağlıdır. En çok kullanılan aktivasyon fonksiyonu tipleri katı sınırlayıcı, doğrusal, rampa, adım, eşik, sigmoid, hiperbolik tanjant fonksiyonlarıdır (Flitman 1997:368).

1.1. Yapay Sinir Ağı Modelleri

Bir yapay sinir ağı modeli birbirleriyle bağlantılı sinirlerin yer aldığı tabakalardan oluşmaktadır. Girdi tabakası, çıktı tabakası ve gizli tabaka olmak üzere temelde üç tabaka bulunmaktadır.

Girdi tabakası ilk tabakadır ve istatistikte bağımsız değişkenlere karşılık gelen girdi değişkenlerinden meydana gelir. Son tabaka çıktı tabakası olarak adlandırılır ve istatistikte bağımlı değişkenlere karşılık gelen çıktı değişkenlerinden meydana gelir. Modeldeki diğer tabakalar ise girdi tabakası ile çıktı tabakası arasında yer alır ve gizli tabaka olarak adlandırılır. Gizli tabakada bulunan sinirlerin dış ortamla bağlantıları yoktur. Yalnızca girdi tabakasından gelen sinyalleri alırlar ve çıktı tabakasına sinyal gönderirler (Warner ve Mısra 1996:287).

Yapay sinir ağları işleyiş şekillerine göre ileri beslemeli ve geri beslemeli yapay sinir ağları olmak üzere iki şekilde incelenir. İleri beslemeli yapay sinir ağlarında sinyaller sadece tek bir yönde, girdi tabakasından çıktı tabakasına doğru yönelir. Bir tabakadan elde edilen çıktı değeri, aynı tabakadaki sinirleri etkilemez. İleri beslemeli ağlarda, sinirler yalnızca bir sonraki tabakada bulunan sinirlerle bağlantıya sahiptir. Geri beslemeli yapay sinir ağlarında, sinyalin yönü girdi tabakasından çıktı tabakasına doğrudur. Fakat aynı zamanda, bir tabaka üzerinde yer alan sinirler, kendisinden, tabakadaki diğer sinirlerden veya diğer tabakalardaki sinirlerden sinyal alabilmektedir. Bu nedenle geri beslemeli ağlarda bir sinirin çıkışı, sinirin o andaki girdileri ve ağırlık değerleriyle belirlenmesinin yanında bazı sinirlerin bir önceki süredeki çıkış değerlerinden de etkilenir. Bu tür ağlar çok güçlü ve karmaşıktır. Geri beslemeli yapay sinir ağları genellikle en iyileme problemlerinin çözümünde kullanılır (Stergio ve Siganos 1996:16).

Yapay sinir ağlarında bilgi, ağdaki bağlantıların ağırlıklarında depolanır. İstenen bir işlevi yerine getirecek şekilde ağırlıkların ayarlanması süreci yapay sinir ağlarında "Öğrenme ya da Eğitim Süreci" olarak adlandırılır. Bu süreç, ağ iyi bir performans elde edinceye kadar devam eder. Sinir ağları modelinde önce açıklayıcı değişkenler belirlenir. Daha sonra meydana gelen çıktı ile arzu edilen çıktı değerleri karşılaştırılır. Karşılaştırmanın sonucuna göre ağ, tabakaların arasındaki ağırlıkların değerlerini değiştirerek ayarlama yapar. Böylece öğrenme gerçekleşir.

1950'li yıllardan bu yana birçok araştırmacı Hebb'in kurallarını temel alarak öğrenmenin nasıl daha iyi olacağı konusunda araştırmalarını sürdürmüşler ve yeni öğrenme yöntemleri geliştirmeye çalışmışlardır. Temelde bu öğrenme yöntemleri, gözetimli ve gözetimsiz öğrenme olarak iki gruba ayrılır. Gözetimli öğrenme, eğitmenli öğrenme olarak da adlandırılır. Sistemde yer alan her bir girdi değişkeni ile ilişkide olan hedef çıktı değerleri bilindiği zaman, gözetimli öğrenmeye ihtiyaç duyulur. Başka bir deyişle sistemdeki girdilere karşılık üretilmesi arzu edilen çıktılar belirtilir. Bu girdi değişkenlerini ve bunlara karşılık üretilmesi istenen çıktı değerlerini içeren veri seti, eğitim seti olarak adlandırılır. Gözetimli öğrenme sürecinde, ağın oluşturduğu çıktılar ile arzu edilen çıktı değerleri karşılaştırılır ve aralarındaki fark hesaplanır. Bu fark ağın

eğitiminde kullanılır. Fark en küçük olacak şekilde bağlantı ağırlıkları düzenlenir. Gözetimli öğrenmeyi kullanan sinir ağları için bir çok farklı algoritma vardır. Geri besleme algoritması, bunlar içinde en yaygın olanıdır. Gözetimsiz öğrenme, eğitmensiz öğrenme olarak da adlandırılır. Burada girdi değişkenlerine karşılık arzu edilen çıktılar belirtilmez. Ağ yalnızca girdi modelini öğrenir. Öğrenme süreci üzerindeki ileri dönüşün kaynağı belli değildir. Tabakalar arasındaki ağırlıkların ayarlanması ağ tarafından kendiliğinden gerçekleştirilir (Warner ve Mısra 1996:288, Elmas 2003:36).

1.2. Regresyon Problemlerine Sinir Ağları Yaklaşımı

Bilindiği gibi istatistikte en çok kullanılan tekniklerden biri de regresyon analizi tekniğidir. Çoğu bilim adamı sinir ağlarını daha iyi anlatabilmek için, sinir ağları ile regresyon modelleri arasındaki ilişkiyi açıklamaya çalışmıştır. Yapılan çalışmalarda sinir ağları doğrusal, doğrusal olmayan, basit, çoklu, parametrik, parametrik olmayan, lojistik, vb. gibi çok sayıda regresyon modeli ile karşılaştırılmıştır.

Tek tabakalı bir ağ yapısında, girdi değişkenleri ile çıktı değişkenleri verilir. Ağın eğitim modelinde gözetimli öğrenme kullanılır. Verilen girdiler ve çıktılar kullanılarak aradaki ilişkinin fonksiyonu tahmin edilmeye çalışılır. Sinir ağlarında kullanılan hata fonksiyonu hesaplanırken,

$$E = \frac{1}{2} \sum_{p=1}^{n} \sum_{k=1}^{Z} (y_{pk} - \hat{y}_{pk})^2 \qquad (1.3)$$

eşitliği ile tanımlanan hata kareler ortalaması ölçüt olarak kullanılır (Bishop 1995:194, Warner ve Mısra 1996:288). Burada ,

p : gözlem sayısını ($p= 1, ...,n$),

k : çıktı sayısını ($k=1, ..., Z$),

y : gözlenen çıktıları,

$\hat{y}$: tahmini çıktıları,

göstermektedir.

Yapay sinir ağında, ileri doğru ilk bilgi geçişi gerçekleştikten sonra geri besleme süreci başlamış olur. Sürecin başlaması için önce girdi tabakasında yer alacak girdi değişkenleri tanımlanmalıdır. Girdi nöronlarının bilgiler üzerinde işlem yapabilme yetenekleri yoktur. Yalnızca bilgilerin bir sonraki tabaka olan gizli tabakaya geçişlerini sağlar. Bilgi geçişi gerçekleştikten sonra girdi tabakasından j. gizli nörona gelen girdi sinyallerinin toplamı,

$$h_{pj} = \sum_{i=1}^{N} w_{ji} x_{pi} \qquad (1.4)$$

eşitliği ile hesaplanır. Burada,

N : girdi nöronlarının sayısı,

w_{ji} : i. girdi nöronu ile j. gizli tabaka nöronu arasındaki ağırlık,

x_{pi} : p modeli için i. girdinin değeri,

olarak tanımlanır. Eşitlik (1.4) ile tanımlanan girdi sinyallerinin toplamı hesaplandıktan sonra j. gizli nöron, net girdilere ve çıktılara aktivasyon fonksiyonunu uygular. Tahmin problemlerinde aktivasyon fonksiyonu olarak, genellikle hatayı minimum yapan sigmoid fonksiyonu seçilir (Hwang ve Dıng 1997:748). j. gizli nörona gelen net girdilere aktivasyon fonksiyonu uygulanarak

$$v_{pj} = g(h_{pj}) = \frac{1}{1 + e^{-h_{pj}}} \tag{1.5}$$

eşitliği elde edilir. Benzer şekilde k. çıktı nöronu için net girdi toplamı

$$f_{pk} = \sum_{j=1}^{M} W_{kj}\, v_{pj} \tag{1.6}$$

eşitliği ile hesaplanır. Burada,

M : gizli nöronların sayısı,

W_{kj} : j. gizli nöron ile k. çıktı nöronu arasındaki ağırlık,

dır.

Sonuç olarak,

$$\hat{y}_{pk} = g(f_{pk}) = \frac{1}{1 + e^{-f_{pk}}} \tag{1.7}$$

eşitliği ile genel çıktı hesaplanır. (1.4) - (1.7) arasındaki eşitlikler, (1.3) de yerine yazılırsa

$$E = \frac{1}{2} \sum_{p=1}^{n} \sum_{k=1}^{Z} \left(y_{pk} - g\left(\sum_{j=1}^{M} W_{kj}\, g\left(\sum_{i=1}^{N} w_{ji}\, x_{pi} \right) \right) \right)^2 \tag{1.8}$$

eşitliği ile hata fonksiyonu elde edilir (Warner ve Mısra 1996:288).

Geri besleme süreci başlamadan önce çıktı tabakasında yer alacak nöronların sayısı verilmelidir. Çünkü gizli tabakadaki her bir nöron, çıktı tabakasındaki bütün nöronlarla bağlantıya sahiptir (Warner ve Mısra 1996:289).

2. LOJİSTİK REGRESYON ANALİZİ

Çok değişkenli istatistiksel verilerin sınıflandırılması, bu verilere uygulanabilecek çeşitli istatistiksel yöntemler için gerekli bir ön analiz olmasının yanı sıra pratikte başlı başına bir analiz olarak da sıkça kullanılmaktadır. Lojistik regresyon analizi de sınıflandırma ve atama işlemlerini yapmak için kullanılan yöntemlerden biridir. Bu yöntem normal dağılım varsayımı veya süreklilik varsayımı gibi ön koşul gerektiren yöntemlere bir alternatif olarak geliştirilmiştir. Lojistik model, bağımlı değişkenin 0, 1 gibi ikili ya da ikiden çok düzey içeren kesikli değişken olması durumunda normallik varsayımı kısıtı olmaması nedeniyle kullanım rahatlığı sağlamaktadır (Tatlıdil 1996:289, Özdamar 1997:461). Lojistik regresyon, bağımlı değişkenin tahminini olasılık olarak hesaplayarak olasılık kurallarına uygun sınıflama işlemi yapma imkanı vermektedir. Değişik gösterim biçimleri olan genel doğrusal regresyon modeli, koşullu beklenen değer biçiminde,

$$E(y_i/x_{i1},...,x_{im}) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} \quad ; \; i = 1, ..., n \text{ için} \tag{2.1}$$

olarak tanımlanabilir. Bu modelde bağımsız değişkenler üzerinde herhangi bir koşul yoktur. Ancak, y bağımlı değişkeninin sürekli değişken olması gerekir. *i*. gözlem için,

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \varepsilon_i \tag{2.2}$$

biçiminde ifade edilen modelde, bağımsız değişkenler için herhangi bir koşul olmadığından dolayı y_i bağımlı değişkeninin sonuç değeri -∞ ile +∞ arasında tüm değerleri alabilmektedir. Bağımlı değişkenin 0, 1 gibi değerler aldığı durumlarda bu kural bozulmaktadır. Bu durumda çeşitli deformasyon dönüşümleri yardımıyla y_i değerleri belli bir aralıkta sürekli hale getirilebilir.

Koşullu beklenen değer biçiminde yazılan doğrusal regresyon modelinde y_i değişken değerlerinin sadece 0 ve 1 gibi değerler aldığı durumda, $P(y_i = 1)$, *i*. gözlemin 1 değerini alma olasılığı olmak üzere beklenen değer,

$$E(y_i) = 1 \times P(y_i = 1) + 0 \times P(y_i = 0) = P(y_i = 1) \tag{2.3}$$

olmaktadır. Bu sonuç regresyon denklemi olarak yazılırsa,

$$E(y_i) = P(y_i = 1) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} \qquad (2.4)$$

eşitliği elde edilir. Sol tarafı 0-1 arasında olasılık değerleri alan (2.4) eşitliğine "doğrusal olasılık modeli" adı verilmektedir (Tatlıdil 1996:290, Gürcan 1998).

Eşitlik (2.4) ile verilen doğrusal olasılık modelinin sol tarafı [0,1] aralığında değerler alırken açıklayıcı değişkenler üzerinde bir kısıtlama olmadığından sağ taraf bütün reel sayıları alabilmektedir. Bu nedenle eşitlik her zaman sağlanamamaktadır. Böylesi bir durumla karşılaşıldığında eşitliğin sağlanabilmesi için sürekli deformasyon dönüşümleri yardımıyla ya $P(y_i=1)$ olasılık değerlerinin IR' ye genişletilmesi ya da reel sayıların [0,1] kapalı aralığında bir değer alması sağlanır. Bu nedenle geliştirilen dönüşümlerden yaygın olarak kullanılanlar, lojit (logit) ve probit (probability unit) dönüşümlerdir. Bu dönüşümler birbirine yakın sonuçlar vermektedir.

Lojit $p = \log(p/(1-p))$ dönüşümünün bazı özellikleri şöyledir:

i. p arttıkça lojit(p) de artar,

ii. $p \in [0,1]$ iken lojit(p) tüm reel değerleri alır,

iii. Eğer $p < 0.5$ ise lojit(p) < 0 ve eğer $p > 0.5$ ise lojit(p) > 0'dır. Bu özellik gözlemlerin sınıflara atanmasında kullanıldığı için çok önemlidir (Aldrich ve Nelson 1984).

Lojistik modelde yer alan parametre tahmin değerleri, en çok olabilirlik, yeniden ağırlıklandırılmış en küçük kareler ve minimum lojit khi-kare yöntemleri ile hesaplanabilir (Aldrich ve Nelson 1984:68, Gürcan 1998).

3. UYGULAMA

Herhangi bir eğitim kurumu öğrenci seçerken, seçeceği öğrencinin başarı olasılığı en yüksek öğrenci olmasını ister. Bu durum yükseköğretim programlarına daha başarılı ve uygun öğrencilerin alınması açısından da önem taşır. Öğrenci başarısızlıklarının nedenlerini tanımlamak ve gelecekte karşılaşılabilecek başarısızlıkları önceden tahmin etmek amacıyla yapılan çalışmalarda, öğrenci başarısızlıkları analiz edilerek, öğrenciler başarılarına göre sınıflandırılmaya çalışılmıştır. Bu nedenle kullanılan çok değişkenli istatistiksel yöntemlerden biri sınıflandırma ve atama işlemlerini yapmak için kullanılan lojistik regresyon analizidir. Lojistik regresyon analizi ile öğrencilerin başarısını etkilediği düşünülen değişkenler bilindiğinde, gelecekteki başarı durumlarının tahmini, olasılık kurallarına uygun olarak hesaplanarak, öğrenciler "başarılı" veya "başarısız" olarak sınıflandırılır.

Lojistik regresyon analizinde olduğu gibi öğrencileri başarı durumlarına göre sınıflandırmak amacı ile kullanılabilecek bir diğer yöntem ise sinir ağları

yaklaşımıdır. Bu çalışmada sinir ağları yaklaşımı, lojistik regresyon analizi ile karşılaştırılacaktır.

Uygulamada, Gazi Üniversitesi Ticaret ve Turizm Eğitim Fakültesi'nden alınan gerçek veriler kullanılmıştır. Çözüm sürecinde, lojistik regresyon analizi için SPSS, sinir ağları analizi için ise Neural Connection paket programlarından yararlanılmıştır.

Araştırmanın kitlesi Gazi Üniversitesi Ticaret ve Turizm Eğitim Fakültesi'nde kayıtlı bütün öğrencilerdir. Fakültede eğitim-öğretim yapan Muhasebe, Büro, Turizm-Seyahat ve Turizm-Konaklama adı altında 4 program bulunmaktadır. 2003-2004 öğretim yılında 3.sınıfa kayıtlı 352 öğrenci örnek olarak seçilmiştir. Araştırma içerisindeki verileri, 2003-2004 öğretim yılı güz yarıyılı sonu itibariyle Gazi Üniversitesi Ticaret ve Turizm Eğitim Fakültesi'nde 3.sınıfa kayıtlı öğrencilerin dosyalarındaki bazı bilgiler ve her bir öğrencinin 3 sene süresince almış olduğu notların ortalaması oluşturmaktadır.

Bağımlı değişken, akademik başarıdır. Akademik başarıyı etkilediği düşünülen pek çok etken vardır. Bunlardan; üniversitede kayıtlı olduğu program, cinsiyet, lise türü, lise ortalaması, Öğrenci Seçme Sınavı (ÖSS) puanı, ailenin yaşadığı şehir ve yaş etkenleri üzerinde durulmuş, bu etkenler bağımsız değişkenler olarak kabul edilmiştir.

Tanımlanan bu bağımsız değişkenler,

x_1: programlar (Muhasebe Programı ise 1, Büro Programı ise 2, Turizm-Seyahat Programı ise 3, Turizm-Konaklama Programı ise 4 değerli),
x_2: cinsiyet (erkek ise 1, kız ise 2 değerli),
x_3: lise ortalaması,
x_4: ÖSS puanı,
x_5: ailenin yaşadığı şehir (Ankara ise 1, Ankara dışı ise 2 değerli),
x_6: mezun olduğu lise türü (Ticaret Meslek Lisesi ise 1, Anadolu Ticaret Meslek Lisesi ise 2, Otelcilik ve Turizm Meslek Lisesi ise 3, Anadolu Otelcilik ve Turizm Meslek Lisesi ise 4 değerli) ve
x_7: yaş

olarak belirlenmiştir.

Öğrencilerin üç sene sonundaki not ortalamaları olan y bağımlı değişkeni, not ortalaması 0-1.99 arasında olan öğrenciler için 0 (başarısız) olarak, 2.00-4.00 arasında olan öğrenciler için ise 1 (başarılı) olarak alınmıştır.

3.1. Lojistik Regresyon Analizi

Lojistik regresyon yöntemi ile yapılan analiz sonucunda, bulunan lojistik model yardımı ile bireylerin başarı olasılıkları hesaplanmaktadır. Hesaplanan olasılık değeri 0.5'ten küçük olan bireyler "başarısız"; hesaplanan olasılık değeri 0.5'ten büyük olan bireyler ise "başarılı" sınıfına atanmaktadır. Gözlemler sınıflandırılarak atandıktan sonra gerçek durumlarıyla karşılaştırılarak doğru sınıflandırılma oranları hesaplanmaktadır. Bulunan sonuçlar Tablo 1' de verilmiştir.

Tablo 1. Lojistik regresyon uygulamasına göre sınıflandırma tablosu

		Tahmin			
		0	1	Toplam	Doğruluk%
Gerçek	0	1	17	18	5.55
	1	0	334	334	100
	Toplam	1	351	352	95.17

Tablo 1 incelendiğinde, gerçekte başarısız (0) olan öğrencilerden, lojistik regresyon uygulamasına göre başarısız sınıfına atanmış olanların sayısının 1, başarılı sınıfına atanmış olanların sayısının ise 17 olduğu görülmektedir. Gerçekte başarılı (1) olan öğrencilerden ise, lojistik regresyon uygulamasına göre başarısız sınıfına atanmış olanların sayısı 0, başarılı sınıfına atanmış olanların sayısı ise 334' dir.Doğru sınıflandırılma oranları,

$$\text{başarısızlık durumu için} \quad : \frac{1}{18} = \%5.55$$

$$\text{başarı durumu için} \quad : \frac{334}{334} = \%100$$

biçiminde hesaplanmıştır.Lojistik regresyon analizi sonucu elde edilen genel doğruluk yüzdesi ise,

$$\frac{1+334}{352} = \%95.17$$

olarak bulunmuştur. Modele eklenecek yeni bir öğrencinin gelecekteki başarısı tahmin edilmek istendiğinde, kurulan lojistik model kullanıldığı zaman tahminin doğru olma olasılığı %95.17 olacaktır.

3.2. Sinir Ağları Yaklaşımı

Sinir ağları uygulamasında çoklu ağ tabakalarından yararlanılmıştır. Sinir ağı modeli girdi tabakası, gizli tabaka ve çıktı tabakasından olmak üzere toplam üç tabakadan oluşmaktadır. Uygulamada 7 bağımsız değişken olduğu için girdi tabakasında 7 sinir bulunmaktadır. Çıktı değişkeni ise 3 yılın sonundaki başarı ortalaması olduğu için çıktı tabakasında da 1 sinir bulunmaktadır.

Ağ modelinde, sinirler arasındaki bağlantıların ağırlık değerleri, uygulamanın başında SPSS' de rasgele olarak üretilir. Ağ, bu ağırlık değerleri kullanılarak test edilmektedir.

Veri setinde yer alan veriler rasgele olarak; eğitim, geçerlilik ve test seti olmak üzere üç bölüme ayrılmaktadır. Eğitim seti, verilerin ağırlıklarına uygun olan öğrenme için kullanılmaktadır. Geçerlilik seti, bir sınıflandırıcının ağırlıklarına uygun olarak kullanılır. Örneğin, sinir ağındaki gizli ünite sayısını seçmek için geçerlilik seti kullanılır. Test seti ise tamamen belirli bir sınıflandırıcının performansını değerlendirmekte kullanılır. Eğitim seti, ağın eğitimine yönelik olarak, test seti ise eğitimin uygulanmasının performansını ölçmede kullanılır. Veri setinin %80'ini eğitim seti, %10'unu geçerlilik seti, %10'unu da test seti oluşturmaktadır.

Veriler karar ya da önerilere eşit katkıda bulunduğundan, ölçü birimi etkisinden arındırılmak için standartlaştırılır. Kullanılan paket program ilk aşamada verileri standartlaştırır. Daha sonra aktivasyon fonksiyonu seçilir. Bu çalışmada aktivasyon fonksiyonu sigmoid fonksiyon olarak seçilmiştir.

Eğitim işlemleri ağ üzerindeki ağırlıkların rasgele olarak hesaplanmasıyla başlar. Eğitim modeli, girdi değişkenlerinin uygulanması ve 1. gizli tabakadaki aktivasyonların hesaplanmasıyla oluşur. Aktivasyon fonksiyonu aracılığıyla bu sinirler tarafından üretilen çıktılar takip eden tabakadaki sinirlere uygulanır. İleri doğru işleyen bu süreç çıktı tabakasından bir çıktı sinyali gelinceye kadar devam eder. Gerçek çıktı değerleri ile arzu edilen çıktı değerleri arasındaki farklılık ölçülür ve sonuca göre ağ modelinin bağlantı ağırlıkları değişir. Bağlantı ağırlıkları sonucu oluşan geri dönüş geçişi, çıktı tabakalarının bağlantıları ile başlayan ve girdi tabakalarının bağlantıları ile sona eren ağın üretilmesiyle gerçekleşir.

Öğrenme kuralı basittir. Ağ tarafından üretilen çıktı, arzu edilen çıktı değeriyle karşılaştırılır ve çıktı sinirlerinden bütün girdi sinirlerine doğru olan bağlantılar değişmektedir. Eğer ağdan elde edilen çıktı değeri, arzu edilen çıktı değerinden büyükse o zaman çıktı sinirleri ile tüm girdi sinirleri arasındaki bağlantıların ağırlık değerleri azalır. Eğer çıktılar arzu edilenden küçükse o zaman da bağlantıların ağırlık değerleri artar. Öğrenme süreci hakkında belirtilmesi gereken iki önemli nokta bulunmaktadır. İlki; algoritma, ilk bağlantı ağırlık değerlerinin herhangi bir düzenlenmesinden doğan hata kareler

ortalamasındaki en yakın yerel minimum dereceyi bulan eğim iniş sürecini kullanır. Bir çok minimum nokta vardır. Bağlantıların başka düzenlemelerini karşılaştıran daha iyi bir minimum nokta bulunabilir. İyi bir optimum nokta bulabilmek için birçok farklı başlangıç değerinden algoritmaya doğru gitmek gerekli olabilir. İkinci nokta; öğrenme oranı ve momentum katsayısı kullanıcı tarafından sürecin başında seçilir.

Eğim iniş metodu her bir iterasyondan sonra hata yüzeyinin eğimini ölçer ve iniş eğiminin yönü üzerindeki ağırlıkları değiştirir. Minimuma ulaşıldığında yeni bir eğim ölçülür ve ağırlıklar yeni yön üzerinde değişir. Momentum katsayısı ve öğrenme oranı olarak bilinen iki parametre değişebilmektedir. Öğrenme oranının değişimi bağlantı ağırlıklarındaki değişimdir. Eğer öğrenme oranı çok yüksekse öğrenme algoritması minimumu aşacak, eğer çok düşükse algoritma minimuma çok uzun bir yolda ulaşacaktır. Momentum katsayısı, ortalama eğim yönündeki bağlantılarda meydana gelen değişimleri düzenler. Eğim iniş metodu her geçişten sonraki hata yüzeyinin eğimini ölçmekte eğimin yönü ile bir önceki değişimin yönü arasında bir karşılaştırma yaparak sinir girişlerinin ağırlıklarını düzenlemektedir (Neural Connection Version2.0 1997).

Bir tabakadaki sinir sayısı ağlar tarafından otomatik olarak seçilebilir ya da bağlantılı olarak düzenlenebilir. Bir çok durumda sinirlerin sayısını artırmak eğitim verileri üzerindeki çoklu tabaka ağlarının performansını geliştirir. Ancak bu geçerlilik verilerinde gerekli değildir.

Bir problemdeki gizli tabakaların sayısının etkisini değerlendirmek için geçerlilik verilerinin performansına bakılır. Ağ yapısının performansını ölçmek için mutlak hata ortalaması (M.H.O.) ve hata kareler ortalaması (H.K.O.) kullanılmaktadır.

Problemin modeli 7 girdi tabakasından ve 1 çıktı tabakasından oluşmaktadır. Gizli tabaka sayısını belirlemek için uygulamada, önce gizli tabaka sayısı 1 alınarak 7-1-1 modeli için hatalar hesaplanmıştır. Daha sonra gizli tabakaların sayısı artırılmış ve geçerlilik verilerine ilişkin hatalar hesaplanmıştır. Bu sonuçlar Tablo 2' de verilmiştir.

Tablo 2. Gizli tabaka sayısının belirlenmesinde oluşturulan modellerin sonuçları

Model	M.H.O.	H.K.O.
7-1-1	0.088193	0.206884
7-2-1	0.099415	0.217241
7-3-1	0.098259	0.217449
7-4-1	0.098065	0.218129
7-5-1	0.098759	0.217929

Tablo 2 incelendiğinde, 1 gizli tabakalı modelin (7-1-1 modeli) 0.088193 mutlak hata ortalaması ve 0.206884 hata kareler ortalaması ile en küçük hataya sahip olduğu görülmektedir. Gizli tabaka sayısı artırıldığında hata değerlerinin artmaya başladığı görülmüştür. Gizli tabakaların sayısı artırıldığında, her bir yeni gizli tabaka, veri setindeki özelliklerden birini daha göstermeye başlayacağından geçerlilik setindeki ağ performansı da artmaktadır. Çok sayıda tabaka eklendiğinde performansta bir azalma görülebilir. Bunun nedeni genel güçteki kayıptır ve bu durumda ağ, verilerden gürültü öğrenmeye başlar. Geçerlilik seti üzerinde hata ölçümleri yapılarak aşırı öğrenmenin tehlikesi azaltılmış olur (Neural Connection Version2.0 1997). 1 gizli tabakalı model (7-1-1 modeli) en küçük hata değerlerine sahip olduğundan uygulamada model olarak seçilmiştir.

Ağın eğitimi için 10000 iterasyon gerçekleştirilmiştir. Sinir ağları analizi sonucunda, her bir set için ayrı sınıflandırma tabloları elde edilmiştir. Her setten elde edilen doğruluk yüzdeleri farklıdır. Genel doğruluk yüzdesi hesaplanırken üç setten elde edilen sonuçlar birleştirilmektedir.

Eğitim seti için doğru sınıflandırma yüzdesi %95.01 olarak bulunmuştur. Mutlak hata ortalaması 0.089281, hata kareler ortalaması 0.210847' dir. Sınıflandırma tablosu Tablo 3' de görülmektedir.

Tablo 3. Eğitim seti için sınıflandırma

		Tahmin			
		0	1	Toplam	Doğruluk%
Gerçek	0	0	0	0	0
	1	14	267	281	95.01
	Toplam	14	267	281	95.01

Geçerlilik seti için doğru sınıflandırma yüzdesi %94.44 olarak bulunmuştur. Mutlak hata ortalaması 0.088193, hata kareler ortalaması 0.0206884' dir. Sınıflandırma tablosu Tablo 4' de görülmektedir.

Tablo 4. Geçerlilik seti için sınıflandırma

		Tahmin			
		0	1	Toplam	Doğruluk%
Gerçek	0	0	0	0	0
	1	2	34	36	94.44
	Toplam	2	34	36	94.44

Test seti için doğru sınıflandırma yüzdesi %97.14 olarak bulunmuştur. Mutlak hata ortalaması 0.085973, hata kareler ortalaması 0.179409' dir. Sınıflandırma tablosu Tablo 5' de görülmektedir.

Tablo 5. Test seti için sınıflandırma

		Tahmin			
		0	1	Toplam	Doğruluk%
Gerçek	0	0	0	0	0
	1	1	34	35	97.14
	Toplam	1	34	35	97.14

Sinir ağları uygulamasına göre sınıflandırma tablosunu elde etmek için eğitim, geçerlilik ve test setleri birleştirilmiştir. Birleştirme işlemi yapılırken, aynı hücrelerde yer alan atama değerleri toplanmaktadır. Elde edilen tablo Tablo 6' da verilmiştir.

Tablo 6. Sinir ağları uygulamasına göre sınıflandırma

		Tahmin			
		0	1	Toplam	Doğruluk%
	0	0	0	0	0
Gerçek	1	17	335	352	95.17
	Toplam	17	335	352	95.17

Sinir ağları analizi sonucu elde edilen genel doğruluk yüzdesi,

$$\frac{0+335}{352} = \%95.17$$

olarak bulunmuştur. Modele eklenecek yeni bir öğrencinin gelecekteki başarısı tahmin edilmek istendiğinde, sinir ağları kullanıldığı zaman tahminin doğru olma olasılığı %95.17 olacaktır.

SONUÇ VE ÖNERİLER

Öğrencilerin akademik başarılarını tahmin etmek için kullanılan bir çok istatistiksel yöntem vardır. Son yıllarda pek çok alanda kullanılan sinir ağları yaklaşımı bu çalışmada lojistik regresyona alternatif bir yöntem olarak incelenmiştir. İki yöntem kullanılarak öğrenci başarısızlıkları analiz edilmiş, sinir ağları yaklaşımının sınıflandırma işlemleri için geçerliliği sorgulanmıştır. Uygulamada Gazi Üniversitesi Ticaret ve Turizm Eğitim Fakültesi'nden alınan gerçek veriler önce lojistik regresyon analizi daha sonra da sinir ağları yaklaşımı kullanılarak sınıflandırılmıştır. Uygulama sonucunda elde edilen değerler Tablo 7' de verilmiştir.

Tablo 7. Başarı sınıflandırmasında lojistik regresyon ve sinir ağları yaklaşımı sonucunda elde edilen doğru sınıflandırma oranları

Model	Başarı durumu		
	Başarısız (0)	Başarılı (1)	
	Doğruluk%	Doğruluk%	Genel%
Lojistik regresyon	5.55	100	95.17
Sinir Ağları	0	95.17	95.17

Tablo 7' den de görüleceği gibi verilerin doğru sınıflandırma olasılıkları lojistik regresyon uygulaması ve sinir ağları yaklaşımı için %95.17 olarak bulunmuştur. İki yöntemin aynı sonucu vermiş olması sinir ağlarının atama problemlerinde kullanılabilirliğini göstermektedir.

Örnek olarak; muhasebe bölümünde kayıtlı, erkek, lise ortalaması 3.00, ÖSS puanı 185, ailesi Ankara dışında yaşayan, Ticaret Meslek Lisesi Mezunu, 18 yaşındaki bir öğrencinin gelecekte başarılı olup olamayacağını tahmin etmek için sinir ağları yaklaşımı kullanıldığında; değişken değerleri veri setine eklenerek yapılan analiz sonucunda bu öğrencinin "başarısız" sınıfına atandığı görülmüştür. Bu tahminin doğru olma olasılığı %95.17'dir. Başarısız olacağı tahmin edilen bu öğrencinin başarısını arttırılabilmek için farklı öğretim teknikleri kullanılabilir. Böylece herhangi bir akademik bölüm, öğrencilerinin başarılarını yükselterek en önemli amaçlarından birini gerçekleştirmiş olur.

Sonuç olarak, başarıyı etkileyen değişkenler bilindiğinde, modele yeni katılan öğrencilerin gelecekteki başarı durumlarını tahmin etmek için sinir ağları yaklaşımı, lojistik regresyon analizine alternatif bir yöntem olarak kullanılabilir. Böylece öğrencilerin gelecekte karşılaşabilecekleri başarısızlıklar tahmin edilerek bu öğrencilerin başarısını artırabilmek için gerekli önlemler alınabilir.

KAYNAKÇA

Adıgüzel, F. (1999). **Neural Networks as a Statistical Tool**. Master thesis (unpublished), Middle East Technical University, 100 p. , Ankara.

Aldrich, J.H. and Nelson, F.D. (1984), **Linear Probability, Logit and Probit Models,** Sage Publications, Inc., 95 p. , London.

Bishop, M.C. (1995), **Neural Networks for Pattern Recognition**, Oxford University Press, 482 p. ,New York.

Collins, J.M. and Clark, M.R. (1993), *An Aplication of The Theory of Neural Computation to The Prediction of Workplace Behavior: An Illustration and Assesment of Network Analysis,* **Personnel Psychology**, 46; 503-524.

Elmas, Ç. (2003), **Yapay Sinir Ağları,** Seçkin Yayıncılık. 192s. , Ankara.

Fausset, L. (1994), **Fundementals of neural networks**, Prentice Hall. Englewood Cliffs, 461 p. , New Jersey.

Flitman, A.M. (1997), *Towards Analysing Student Failures: Neural Networks Compared with Regression Analysis and Multiple Discriminant Analysis*, **Computers Ops. Res.**, 24(4); 367-377

Gorr, W.L. , Nagin, D. and Szcypula, J. (1994), *Comparative Study of Artificial Neural Network and Statistical Models for Predicting Student Grade Point Avarages*, **International Journal of Forecasting**, 10; 17-34.

Gürcan, M. (1998), **Lojistik Regresyon Analizi ve Bir Uygulama**, Yüksek Lisans Tezi (basılmamış), Ondokuz Mayıs Üniversitesi, 63 s. , Samsun.

Hardgrave, B.C. , Wilson, R.L. and Walstrom, K.A. (1994), *Predicting Graduate Student success: A Comparison of Neural Networks and Traditional Techniques*, **Computers Ops. Res**., 21(3); 249-263.

Hwang, G.J.T. and Ding, A.A. (1997), *Prediction Intervals for Artificial Neural Networks*, **Journal of the American Statistical Association** 92(438); 748-757.

Neural Connection Version2.0 Copyright 1995-97. Recognition Systems Ltd.

Özdamar, K. (1997). **Paket Programlar ile İstatistiksel Veri Analizi**, Anadolu Üniversitesi Yayınları, 512 s. , Eskişehir.

Öztemel, E. (2003), **Yapay Sinir Ağları**, Papatya Yayıncılık, 232s. , İstanbul.

Stergiou, C. and Siganos, D. (1996), **Neural Networks**, http://www-dse.doc.ic.ak.uk./~nd/surprise .

Tatlıdil, H. (1996), **Uygulamalı Çok Değişkenli İstatistiksel Analiz**, Cem Web Ofset Ltd. Şti. , 424 s. , Ankara.

Veaux, D.R. , Schumi, J. , Schweinsberg, J. and Ungar, H.L. (1998), *Prediction Intervals for Neural Networks via Nonlinear Regression*, **Technometrics**, 40(4); 273-282.

Warner, B. and Mısra, M. (1996), *Understanding Neural Networks as Statistical Tools*, **The American Statistician**, 50(4); 284-293.