KPSS SONUÇLARININ VERİ MADENCİLİĞİ YÖNTEMLERİYLE TAHMİN EDİLMESİ

Pamukkale Üniversitesi
Fen Bilimleri Enstitüsü
Yüksek Lisans Tezi
Bilgisayar Mühendisliği Anabilim Dalı

Hüseyin ÖZÇINAR

Danışman: Yard. Doç. Dr. Sezai TOKAT

Mayıs 2006
DENİZLİ

YÜKSEK LİSANS TEZİ ONAY FORMU

Hüseyin ÖZÇINAR tarafından Yard. Doç. Dr. Sezai TOKAT yönetiminde hazırlanan **"KPSS Sonuçlarının Veri Madenciliği Yöntemleriyle Tahmin Edilmesi"** başlıklı tez tarafımızdan okunmuş, kapsamı ve niteliği açısından bir Yüksek Lisans Tezi olarak kabul edilmiştir.

Prof. Dr. Hüseyin KIRAN
Jüri Başkanı

Yard. Doç. Dr. Özcan MUTLU
Jüri Üyesi

Yard. Doç. Dr. Sezai TOKAT
Jüri Üyesi (Danışman)

Pamukkale Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun …./…./……. tarih ve …… sayılı kararıyla onaylanmıştır.

Prof. Dr. Mehmet Ali SARIGÖL
Müdür

Bu tezin tasarımı, hazırlanması, yürütülmesi, araştırılmalarının yapılması ve bulgularının analizlerinde bilimsel etiğe ve akademik kurallara özenle riayet edildiğini; bu çalışmaların doğrudan birincil ürünü olmayan bulguların, verilerin ve materyallerin bilimsel etiğe uygun olarak kaynak gösterildiğini ve alıntı yapılan çalışmalara atfedildiğini beyan ederim.

İmza :
Öğrencinin Adı Soyadı : Hüseyin ÖZÇINAR

TEŞEKKÜR

Bu araştırmadaki yardımları, önerileri ve araştırma boyunca gösterdiği ilgisi ve sabrı için değerli hocam Yard. Doç. Dr. Sezai Tokat'a sonsuz teşekkürlerimi sunarım.

Verdiği destek ve veri kaynaklarına ulaşmamdaki yardımlarından dolayı Prof. Dr. Hüseyin Kıran'a teşekkürlerimi sunarım.

Bilgisi ve deneyimleriyle araştırmaya katkıda bulunan Yard. Doç. Dr. Ramazan Baştürk, Yard. Doç. Dr. Bertan Badur, Yard. Doç. Dr. Erkan Korkmaz ve Yard. Doç. Dr. Özcan Mutlu'ya sonsuz teşekkürlerimi sunuyorum. Yardımlarını ve desteklerini esirgemeyen değerli arkadaşlarım Arş. Gör. Serap Samsa, Öğr. Gör. Ebru Mutlu'ya teşekkür ederim.

Her zaman yanımda olan çok değerli aileme, verdikleri destek, gösterdikleri sabır ve anlayış için şükranlarımı sunuyorum.

ÖZET

KPSS SONUÇLARININ VERİ MANDECİLİĞİ YÖNTEMLERİ İLE TAHMİN EDİLMESİ

Özçınar, Hüseyin
Yüksek Lisans Tezi, Bilgisayar Mühendisliği ABD
Tez Yöneticisi: Yard. Doç. Dr. Sezai TOKAT

Mayıs 2006, 53 Sayfa

Araştırmada materyal olarak sınıf öğretmenliği A.B.D. öğrencilerinin lisans eğitimleri süresince bazı derslerden aldıkları ders geçme notları, genel not ortalamaları, öğretim türleri ve KPSS puanları kullanılmıştır. Çalışmada ilk olarak toplanan veriler temizlenip birleştirilmiş ve veri madenciliği uygulamasında kullanılabilecek şekilde düzenlenmiştir. Daha sonra veriler veri madenciliği uygulamasında kullanılmak üzere öğrencilerin üniversiteye giriş yılına göre ayrılarak dört farklı veri kümesi oluşturulmuştur.

Toplanan verinin anlaşılabilmesi için frekans analizi ve regresyon analizi yöntemleri kullanılarak derslere ve yıllara göre verinin özellikleri incelenmiştir. Yapılan inceleme sonucunda bazı derslerde A1 ve A2 gibi yüksek notlarla geçen öğrencilerin oranı % 5-6 civarında iken C ile geçen öğrencilerin tüm veri kümesi için oranının % 38,6 olduğu görülmüştür. Bu aşamada bazı derslerin not dağılımlarının yıllara göre % 10 ile %20 arasında değişimler gösterebildiği saptanmıştır.

Modelleme aşamasında tahmin doğruluklarının karşılaştırılabilmesi için yapay sinir ağı ve regresyon modelleri oluşturulmuştur. Yapay sinir ağı modelini oluşturmak için öğrenme yöntemi olarak geriye yayılım algoritmasını kullanan çok katmanlı perseptron kullanılmıştır. Regresyon modelini oluşturmak için çoklu doğrusal regresyon yöntemi kullanılmıştır.

Frekans analizi yöntemiyle veri kümesinin özellikleri belirlenmiştir. Oluşturulan regresyon modeli ile KPSS sonuçlarının değişimi üzerinde anlamlı katkısı olan değişkenler incelenmiş ve oluşturulan modellerin tahmin doğrulukları, ortalama mutlak hata ve ortalama hata kareler kökü değerleri kullanılarak karşılaştırılmıştır.

Anahtar Kelimeler: Veri Madenciliği, Yapay Sinir Ağları, Regresyon Analizi, Öngörü, KPSS

Prof. Dr. Hüseyin KIRAN
Yard. Doç. Dr. Sezai TOKAT
Yard. Doç. Dr. Özcan MUTLU

ABSTRACT

PREDICTING KPSS RESULTS USING DATA MINING METHODS

Özçınar, Hüseyin
M. Sc. Thesis in Computer Engineering
Supervisor: Asst. Prof. Sezai TOKAT

May 2006, 53 Pages

In this study, general point average, grades of lessons and type of school were to predict KPSS results. Initially collected data were cleaned, merged and formated which could be used in data mining application. Data tables were splitted according to universtiy entry date of students and four data sets were created.

Data sets were examined with frequency and regression analysis techniques to get a better understanding of collected data. The analysis results showed that while the percentage of high grades like A1, A2 were about %5-6, the percentage of students who get C was % 38,6. At this stage it was noted that the grade distributions of some of lessons were changed % 10, % 20 with respect to years.

At the modeling stage, artificial neural networks model and regression model were created in order to compare predictive accuracy of these data mining techniques. Multilayer perceptron with backpropagation used for artificial neural network model and multiple linear regression technique used for regression model.

Frequency analysis was used to explore data set charecteristics and variables which have signicant effect on KPSS results found using regression model. The error term of models were compared using mean absolute error and root mean squared error.

Key Words: Data Mining, Artificial Neural Networks, Regression Analysis, Prediction, KPSS

Prof. Dr. Hüseyin KIRAN
Asst. Prof. Dr. Sezai TOKAT
Asst. Prof. Dr. Özcan MUTLU

İÇİNDEKİLER

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

TABLOLAR DİZİNİ

Sayfa

SİMGE VE KISALTMALAR DİZİNİ

KPSS	Kamu Personeli Seçme Sınavı
ÖSS	Öğrenci Seçme Sınavı
ÖSYM	Öğrenci Seçme ve Yerleştirme Merkezi
YSA	Yapay Sinir Ağları
R^2	Çoklu Açıklayıcılık Katsayısı
ε	Öğrenme Katsayısı
μ	Momentum Katsayısı
DPT	Devlet Planlama Teşkilatı
PAÜ	Pamukkale Üniversitesi

1. GİRİŞ

Tarihsel olarak elektronik veri yönetiminin başlangıcı 1950'lerin sonuna rastlar. Dönemin standartları bugüne nazaran oldukça ilkel, yazılım ve donanım maliyetleri açısından da oldukça pahalıydı. Sonraki yıllarda toplanan veri miktarındaki hızlı artış, daha gelişmiş elektronik veri yönetim tekniklerine olan gereksinimi de artırdı (Schumann 2005).

Elektronik veri saklama ve analiz araçlarının gelişimi büyük miktarlarda veriyi işleme yeteneğine sahip teknolojilerin üretilmesini sağladı. Bu teknolojilerin en yenileri veri ambarları ve veri madenciliğidir. Veri madenciliği 1980'lerin sonunda geliştirilen, 90'lı yılarda büyük bir gelişme gösteren ve uygulama alanları artan, yeni binyılda da bu gelişimini sürdürmesi beklenen, veri temelli karar alma süreçlerinde önemli katkıları olan bir teknolojidir (Beitel 2005, Han ve Kamber 2000).

Modern bilim ve mühendislik fiziksel, biyolojik ve sosyal sistemleri tanımlamada hipoteze dayalı modelleri kullanmaktadır. Böyle bir yaklaşım temel bilimsel modelin elde edilmesi ve bu model üzerine çeşitli uygulamaların oluşturulması esasına dayanır. Bu yaklaşımda toplanan veri daha önce oluşturulan hipotezi doğrulamak ve doğrudan ölçülmesi zor veya imkansız olan parametreleri tahmin etmek için kullanılır. Ancak birçok durumda oluşturulması gereken hipotezler bilinmemektedir ya da sistem matematiksel olarak modellemek için çok karmaşıktır. Bilgisayar kullanımının artmasıyla birlikte bu tür sistemlerden toplanan verilerin de artması, herhangi bir hipotez olmaksızın sistem parametreleri arasındaki ilişkileri tahmin etmeye yarayan tekniklere gereksinimi ortaya çıkardı. Bu nedenle günümüzde klasik modelleme ve hipoteze dayalı analizlerden, gelişen modellere ve veriden doğrudan analiz yapmaya yarayan tekniklere doğru bir geçiş yaşanmaktadır (Kantardzic 2003).

Bilgisayarlarda, bilgisayar ağlarında terabytelar büyüklüğünde verilerin saklandığı günümüzde kamu kurumları, bilim kuruluşları ve şirketler veri toplama ve saklama işlemleri için oldukça büyük miktarlarda finansal kaynak ayırmaktadırlar. Toplanan verilerin hacimlerinin yönetmek için çok büyük olması ve veri yapılarının etkin bir veri analizi yapmak için çok karmaşık olması pratikte bu verilerin çok küçük bir kısmının

kullanılabilmesine neden olmaktadır. Bu durumun temel nedeni veri kümesi oluşturulması esnasında verinin nasıl kullanılıp analiz edileceği ile ilgili planların yerine veri saklama alanının etkin kullanımına yönelik kaygıların göz önünde bulundurulmasıdır.

Geniş, karmaşık ve bilgi bakımından zengin verilerin anlaşılması hemen hemen bütün bilim, iş ve mühendislik çevreleri için ortak bir gereksinimdir. İş dünyasında şirket ve müşteri bilgileri stratejik bir değerdir. Veri tabanlarındaki verilerden faydalı bilgileri çıkartmak ve bu bilgiyi işlemek rekabetçi çağdaş dünya için büyük bir öneme sahiptir.

Veri madenciliği, verinin işlenip bilgi üretilmesi işlevini yerine getirmek için tanımlayıcı ve öngörüye yönelik modeller sunmaktadır. Öngörü yönteminin karar alma sürecinde başarılı kararları beraberinde getireceği ve bu şekilde fayda maksimizasyonu sağlanabileceği gerçeği, öngörü yöntemine olan ilgiyi artırmaktadır. Artan bu ilgiyle beraber öngörü modelleri hakkında yapılan çalışma ve kullanılan yöntem çeşitliliği de hızla artmaktadır. Yapay sinir ağları ve regresyon analizi teknikleri bunlardan en önemlileridir (Yurtoğlu 2005).

Basit bir şekilde insan beyninin çalışma şeklini taklit eden yapay sinir ağı modelleri birçok alanda yaygın olarak kullanılmaktadır. Evrensel fonksiyon yakınsayıcı yöntem (Universal Function Aproximators) olarak tanımlanan yapay sinir ağları, veriden öğrenebilme, genelleme yapabilme ve çok sayıda değişkenle çalışabilme gibi önemli özelliklere sahiptir. Bu özellikleri sayesinde önemli avantajlar sağlayan yapay sinir ağları yöntemi öngörü modellemesinde son yıllarda yaygın olarak kullanılmaktadır (Yurtoğlu 2005).

İstatistik biliminin en önemli konularından birisini regresyon analizi oluşturmaktadır. Regresyon analizi matematik, finans, ekonomi, tıp gibi bilim alanlarında yoğun olarak kullanılmaktadır. Regresyon analizinin temelinde; gözlenen bir olayın değerlendirilirken, hangi olayların etkisi içinde olduğunun araştırılması yatmaktadır. Regresyon analizi yapılırken, gözlem değerlerinin ve etkilenilen olayların bir matematiksel gösterimle yani bir fonksiyon yardımıyla ifadesi gerekmektedir. Kurulan bu modele regresyon modeli denir. Bağımsız değişkenin birden fazla olduğu regresyon modellerine ise çoklu regresyon modeli denir.

Veri analizinden değerli ve kazanç getiren bilgi sağlamada başarılı olan veri madenciliği tekniği iş çevrelerinde yıllardır kullanılmaktadır. Ancak eğitim alanında veri madenciliği kullanımı, teknik bilgiye ve doğru veri madenciliği tekniğini seçmek için istatistik bilgisine sahip insan kaynakları kıtlığı ve veri madenciliği teknikleri için ayrılması gereken finansal kaynak azlığı gibi nedenlerle sınırlı kalmıştır (Beitel 2005).

Bilişim sektöründeki büyük miktarlardaki üretim ve rekabet nedeniyle ucuzlayan yazılım ve donanım fiyatları ve kullanım için daha az teknik bilgi gerektiren kullanıcı dostu veri madenciliği araçlarının üretilmesi bu teknolojinin eğitim alanında da daha yaygın olarak kullanımına olanak sağlamaktadır. Ayrıca makinelerin insan kaynaklarından çok daha ucuz bir yöntem olduğu kurumlar tarafından tecrübe edilmiştir. İnsanlar zihinlerinde işleyebilecekleri veri miktarı bakımından sınırlıdır. İnsan beyni kurumlar tarafından toplanan çok büyük miktarlardaki veriyi işleyecek analitik kapasiteye sahip değildir. Bu analizlerin günümüz gelişmiş veri madenciliği teknikleri ile bile anlaşılması zordur (Schumann 2005).

Kamu Personeli Seçme Sınavı (KPSS), kamu sektöründe çalışmak isteyen bireyler arasında seçme yapabilmek için Öğrenci Seçme ve Yerleştirme Merkezi (ÖSYM) tarafından yapılan bir sınavdır. Öğretmen istihdamının çok önemli bir kısmının devlet tarafından sağlandığı ülkemizde eğitim fakültesi öğrencilerinin mezuniyet sonrası mesleklerini kamu sektöründe icra edebilmeleri için bu sınavda yüksek bir başarı elde etmeleri gerekmektedir.

Bu çalışmanın amacı son yıllarda iş dünyasında mühendislik, tıp, ekonomi gibi alanlarda gittikçe artan bir oranda kullanılan veri madenciliği yöntemini tanıtmak, göreceli olarak yeni bir öngörü tekniği olan yapay sinir ağları yöntemi ile regresyon analizi yöntemini karşılaştırmak ve eğitim fakültesi öğrencilerinin KPSS'den aldıkları puanları lisans eğitimleri süresince aldıkları ve KPSS'de soru çıkan çeşitli derslerden aldıkları ders geçme notu, genel not ortalamaları, öğretim türleri gibi parametreleri kullanarak tahmin eden bir model oluşturmaktır. Bu çalışma esnasında yapay sinir ağları tekniği kullanılarak oluşturulan öngörü modeli ile çoklu regresyon analizi yöntemi kullanılarak elde edilen modelin tahmin başarısı açısından karşılaştırılması yapılmış bu karşılaştırmada yapay sinir ağları yöntemiyle oluşturulan modelin performansının, eğitim ve test veri kümesi büyüklüğüne, kullanılan ağ yapısına, öğrenme yöntemine ve öğrenme katsayısı, momentum, eğitim için kullanılan yineleme sayısı gibi öğrenme parametrelerine göre değişimi incelenmiştir. Toplanan verilerin

görselleştirme ve özetleme gibi veri madenciliği teknikleri kullanılarak herkes tarafından kolay anlaşılabilir bilgiler üretmek bu çalışmanın bir diğer amacıdır.

Bu çalışmanın önemi yapay sinir ağları ve regresyon analizi yöntemlerinin veri madenciliği öngörü modeli olarak performanslarının karşılaştırılması, yapay sinir ağları yöntemi kullanılarak oluşturulan modellerin, eğitim alanında geçmişten beri kullanılan çoklu regresyon analizi yöntemine bir alternatif oluşturup oluşturamayacağının belirlenmesi, veri madenciliği tekniklerinin eğitim alanındaki kullanım alanlarıyla ilgili yapılan çalışmalara katkıda bulunulması ve eğitim fakültesi yöneticilerine, öğretim elemanlarına ve öğrencilerine fayda sağlayabilecek bilgiler çıkartılması olarak özetlenebilir.

2. VERİ MADENCİLİĞİ

Bu bölümde veri madenciliği kavramının nasıl oluştuğu anlatılacak veri madenciliğinin çeşitli tanımları verilecektir. Standart veri madenciliği süreci olarak CRISP-DM sürecine yer verilerek veri madenciliği metodolojisi aşamalarıyla incelenecektir. Veri madenciliği uygulamaları, bu uygulamalarda kullanılan teknikler incelenecek ve veri madenciliği uygulama alanları tanıtılacaktır.

2.1. Veri Madenciliği Kavramı

Veritabanı ve bilgi teknolojileri 1960'lardan beri ilkel dosya işlem sistemlerinden büyük güçlü veritabanı sistemlerine doğru sistematik olarak gelişiyor. Bu gelişme 1970'lerden itibaren ilişkisel veritabanı sistemlerinin oluşmasına, 1980'lerin ikinci yarısından itibaren multimedya uzay verileri gibi hacimli verileri tutmaya olanak sağlayan nesne-tabanlı, geliştirilmiş-ilişkisel veri tabanları gibi gelişmiş veritabanı sistemlerinin oluşmasına neden oldu. Veri tabanı sistemlerindeki bu gelişmeler 1980'lerin sonunda veri ambarları ve veri madenciliği gibi kavramların oluşmasını sağladı (Han ve Kamber 2000).

Veri madenciliği ismi ve tanımı birbiriyle çelişkili olmasa da farklı çevreler arasında değişim gösterdi.

Veri Madenciliği kavramı ile ilgili tanımlardan ikisi şu şekildedir:

- Veri Madenciliği, istatistik, matematiksel yöntemler, örüntü tanıma tekniklerini kullanarak büyük miktardaki verilerin içinden anlamlı ve yeni örüntüleri bulma sürecidir (Web_1 2006),
- Veri Madenciliği; geniş veritabanlarından bilgi çıkartabilmek amacıyla makine öğrenmesi, örüntü tanıma, istatistik, görselleştirme gibi alanların tekniklerini bir araya getiren disiplinler arası bir alandır (Cabena vd 1998),

Veri madenciliği ve veri tabanlarında bilgi keşfi süreci kavramları birçok kaynakta birbirinin yerine kullanılmaktadır. Veri madenciliği, veri tabanlarında bilgi keşfi

sürecinde bir adım olmasına rağmen birçok çalışmada tüm süreci anlatmak için kullanılmaktadır. Bu çalışmada veri madenciliği kavramı sürecin tamamını ifade etmek için kullanılacaktır.

Veri madenciliği tanımlarından da anlaşılacağı gibi istatistik, makine öğrenmesi, veritabanı yönetimi, görselleştirme gibi alanlardan faydalanan disiplinler arası bir alandır. İşlenmemiş veriden, son kullanıcının kolayca anlayıp karar alma sürecine dahil edebileceği bilgiyi oluşturana kadar geçen tüm süreci kapsayan bir yöntem olmasından, hipotez doğrulamaya yönelik değil yeni, gizli örüntüler bulmaya yönelik bir alan olmasından ve çok çeşitli teknikleri aynı uygulama içinde kullanabilmeye olanak sağlamasından dolayı veri madenciliği kullanıcılarına kendisini oluşturan makine öğrenmesi, istatistik matematik gibi yöntemlerden daha farklı bir perspektif sunar (Feelders vd 2000).

Son yıllarda gelişmiş arayüzleri ile son kullanıcı için kullanım kolaylığı sağlayan veri madenciliği programları üretilmektedir. Bu tür programların varlığına rağmen veri madenciliği sürecinde veri ve alan uzmanlarına gereksinim duyulmaktadır.

2.1.1. Veri madenciliği sürecinde insan faktörü

Birçok yazılım üreticisi veri madenciliği yazılımlarını pazarlarken ürünlerinin tak-kullan olduğu yönünde sloganlar kullanmaktadırlar. Bazı kitaplar veri madenciliğinin tanımında "otomatik" kelimesine yer vermektedirler ancak zamanla veri madenciliği sürecinde iyi yetişmiş alan ve analiz bilgilerine sahip uzmanların projenin başarısı açısından mutlak bir gereklilik olduğu ortaya çıkmıştır(Larose 2005). Örneğin Berry ve Linoff (1997) veri madenciliğini otomatik ya da yarı otomatik süreçlerle büyük miktarlardaki verilerin örüntü ve kurallar bulmak için işlenmesi süreci olarak tanımlarken daha sonraki çalışmalarında Berry ve Linoff (2000) veri madenciliği tanımı için kullandıkları otomatik ve yarı otomatik tanımlamasının veri madenciliğinin bir disiplin değil satın alınan bir ürün olarak anlaşılmasına yol açtığını ve bu kanının çok yanlış olduğunu söylemişlerdir.

Veri madenciliği birçok aşamasına kullanıcı tarafından karar verilen yinelemeli ve etkileşimli bir süreçtir (Fayyad vd 1996). Proje için belirlenen amaca ulaşabilmesi için alan ve veri uzmanlarına gereksinim vardır.

Veri madenciliği projelerinden anlamlı bir sonuç elde edebilmek için veriyi anlamak oldukça önemlidir. Ayrık değerler, doğum tarihi ve yaş gibi birbiriyle beraber değişen

özelliklerin tespit edilmesi, projenin amacının, proje boyunca cevap aranılacak soruların, bu soruları cevaplamak için kullanılabilecek veri kümelerin belirlenmesi her aşamada çıkan sonuçların değerlendirilmesi alan uzmanının sorumluluğundaki görevlerdir (Web_3 1999).

Veri tabanlarındaki verilerin amaca yönelik seçilerek örnek veri kümelerinin oluşturulması bu verilerin değerlendirilmesi için istatistiksel yargı yeteneği olan bir veri uzmanı gerekir (Hand 1998). Veri uzmanı kullanılacak algoritmaları seçer, ki bu veri madenciliği sürecinde sonucu en çok etkileyen adımlardan biridir, verileri bu algoritmalarla kullanılabilecek yapıya koyar, süreci takip eder ve sonucu alan uzmanının anlayabileceği bir dile çevirir (Feelders vd 2000).

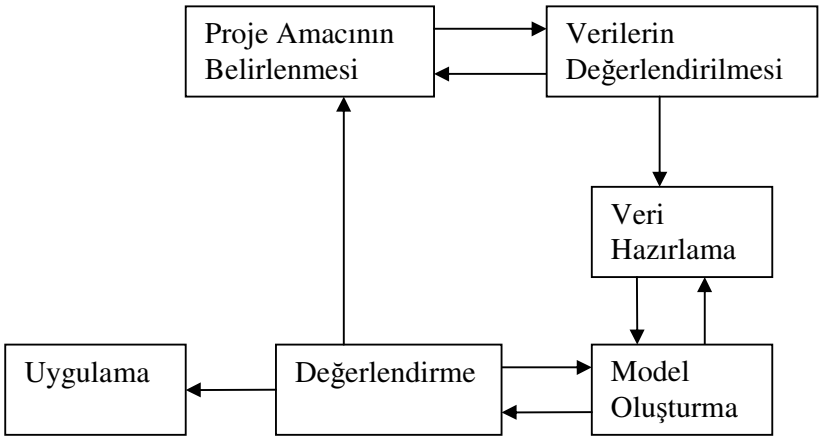
2.1.2. Veri madenciliği süreci

Birçok kurum kendi problemlerine, verilerine ve sahip oldukları diğer kaynaklara göre kendi veri madenciliği sürecini oluşturmaktadır ancak veri madenciliği sürecinin oluşturulmasında yapılan yanlışlıklar, sürecin etkinliğine zarar vermektedir (Larose 2005).

Veri madenciliği sürecinin standartlaştırılması konusunda farklı grup, kurum ve şirketler çeşitli standartlar oluşturmuşlardır bunlardan en çok takip edileni DaimlerChrysler ve SPSS tarafından 1996 yılında oluşturulan Veri Madenciliği için Sektörler Arası Standart Sürecidir (Web_2 2003). Bu çalışmada bu süreç CRISP-DM adıyla anılacaktır.

CRISP-DM sürecine göre veri madenciliği süreci altı aşamadan oluşan etkileşimli ve yinelemeli bir süreçtir. Şekil 1.1'de gösterilen akış şemasının herhangi bir aşamasında elde edilen sonuçlara göre sonraki aşamaya ya da önceki bir aşamaya geçilip yeni belirlenen problemlere, ilgi alanlarına göre iyileştirmeler ya da farklı işlemler yapılabilir (Larose 2005). Veri madenciliği süreci aşağıdaki adımlardan oluşmaktadır.

- Proje amacının belirlenmesi
- Verilerin değerlendirilmesi
- Verilerin hazırlanması
- Modelleme
- Değerlendirme
- Uygulama



Şekil 1.1 CRISP-DM süreci

2.1.2.1. Proje amacının belirlenmesi

Bu aşamada projenin hangi sektörde ne amaçla kullanılacağının, projenin sonunda neyin hedeflendiğinin, nelere ihtiyaç duyulduğunun, proje sonunda elde edilecek bilginin nasıl değerlendirileceğinin açıkça ortaya koyulması gerekir. Ortaya koyulan amaçlar, gereksinimler ve kısıtlamalar veri madenciliği problem tanımı formuna dönüştürülür ve bu amaçlara ulaşmak için bir strateji oluşturulur.

2.1.2.2. Verilerin değerlendirilmesi

Bu aşama verinin toplanmasıyla başlar. Veri analizcisi ve alan uzmanları açıklayıcı veri analizi gibi yöntemlerle veriyi tanımaya, kalitesi hakkında fikir sahibi olmaya çalışırlar. Bu aşamada proje hakkında ilk izlenimlere sahip olmak için veriden küçük ilginç örnekler seçilerek hipotezlerde oluşturulabilir (Chapman vd 2000).

2.1.2.3. Verilerin hazırlanması

Bu aşamada işlenmemiş verinin projede kullanılabilecek duruma getirilmesi amaçlanır. Hatalı veya analizin yanlış yönlenmesine neden olabilecek veriler temizlenir. Veri farklı kaynaklardan toplanmışsa ve aralarında farklılıklar varsa gerekli dönüşümler yapılarak bu farklılıklar ortadan kaldırılır. Eksik verilerin bulunduğu kayıtlar proje için fazla enformasyon taşımıyor ise silinir ya da eksik veriler çeşitli yöntemler kullanılarak tahmin edilmeye çalışılır. Bu aşama en çok iş gücü gerektiren ve toplam süreç içinde en fazla zaman alan aşamadır (Larose 2005).

2.1.2.4. Modelleme

Bir veri madenciliği problemi için birden fazla teknik kullanılabilir, problem için uygun olan teknik veya tekniklerin bulunabilmesi için birçok teknik oluşturulup bunların içinden en uygun olanlar seçilir. Genetik algoritmalar en iyi sonuç veren tekniğin seçimi için kullanılabilir. Model oluşturulduktan sonra kullanılan tekniğin gereksinimlerine uygun olarak veri hazırlanması aşamasına tekrar dönülüp gerekli değişiklikler yapılabilir (Chapman vd 2000).

2.1.2.5. Değerlendirme

Bu aşamada, daha önce oluşturulmuş olan model, uygulamaya koyulmadan önce son kez tüm yönleriyle değerlendirilir, kalitesi ve etkinliği ölçülür. Modelin ilk aşamada oluşturulan proje amacına ulaşmada etkin olup olmadığı ve problemin tüm yönleri için bir çözüm sağlayıp sağlamadığı karara bağlanır. Modelin anlaşılabilirliği ve doğruluk oranı gibi konularda da model amaç için yeterli kaliteyi sağlıyorsa uygulama aşamasına geçilir(Chapman vd 2000).

2.1.2.6. Uygulama

Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi başka bir uygulamanın alt parçası olarak da kullanılabilir**.** İşlenen veri kullanıcının anlayabileceği, karar alma sürecinde kullanılabilecek bir şekilde son kullanıcıya verilir.

2.1.3. Veri madenciliği uygulama alanları

Son yıllarda iş ve bilim çevreleri veri madenciliği yöntemlerini sıklıkla kullanmaya başlamıştır. Veri madenciliği uygulamalarının kullanıldığı sektörler ve uygulama alanları Güvenç (2001) tarafından aşağıdaki gibi sıralanmıştır:

Pazarlama

- Müşteri guruplandırmasında,
- Müşterilerin demografik özellikleri arasındaki bağlantıların kurulmasında,
- Çeşitli pazarlama kampanyalarında,
- Mevcut müşterilerin elde tutulması için geliştirilecek pazarlama stratejilerinin oluşturulmasında,
- Pazar sepeti analizinde,
- Çapraz satış analizleri,
- Müşteri değerleme,

- Müşteri ilişkileri yönetiminde,
- Çeşitli müşteri analizlerinde,
- Satış tahminlerinde,

Bankacılık

- Farklı finansal göstergeler arasındaki gizli korelasyonların bulunmasında,
- Kredi kartı dolandırıcılıklarının tespitinde,
- Müşteri segmentasyonunda,
- Kredi taleplerinin değerlendirilmesinde,
- Usulsüzlük tespiti,
- Risk analizleri,
- Risk yönetimi,

Sigortacılık

- Yeni poliçe talep edecek müşterilerin tahmin edilmesinde,
- Sigorta dolandırıcılıklarının tespitinde,
- Riskli müşteri tipinin belirlenmesinde.

Perakendecilik

- Satış noktası veri analizleri,
- Alış-veriş sepeti analizleri,
- Tedarik ve mağaza yerleşim optimizasyonu,

Borsa

- Hisse senedi fiyat tahmini,
- Genel piyasa analizleri,
- Hisse tespitlerinde,
- Alım-satım stratejilerinin optimizasyonu.

Telekomünikasyon

- Kalite ve iyileştirme analizlerinde,
- Hatların yoğunluk tahminlerinde,

Sağlık ve İlaç

- Test sonuçlarının tahmini,
- Ürün geliştirme,
- Tıbbi teşhis
- Tedavi sürecinin belirlenmesinde

Endüstri

- Kalite kontrol analizlerinde

- Lojistik,
- Üretim süreçlerinin optimizisyonunda,

Bilim ve Mühendislik

- Deneysel veriler üzerinde modeller kurarak bilimsel ve teknik problemlerin çözümlenmesi.

Eğitim

- Öğrenci davranışlarının öngörülmesi.
- Öğrencilerin ders seçme eğilimlerinin belirlenmesi.

2.1.4. Veri madenciliği modelleri

Veri madenciliğinde kullanılan modeller, temel olarak tahmin edici ve tanımlayıcı olmak üzere iki ana başlık altında toplanabilir. Tahmin edici modeller ile tanımlayıcı modeller arasındaki fark kesin sınırlarla ayrılmamıştır. Tahmin edici modeller anlaşılabilir olduğu ölçüde tanımlayıcı model olarak, tanımlayıcı modeller de tahmin edici model olarak kullanılabilirler (Velickov ve Solomatine 2000).

2.1.4.1. Tanımlayıcı modeller

Tanımlayıcı modeller analiste daha önceden bir hipoteze sahip olmaksızın, veri kümesinin içinde ne tür ilişkiler olduğunu anlama imkanı sunar. Analizcinin çok geniş veri tabanlarındaki bilgileri incelemek, örüntüleri keşfetmek için doğru soruları sorup hipotezler geliştirmesi pratikte zor olduğundan, ilginç örüntüleri keşfetme insiyatifi veri madenciliği programına bırakılır. Keşfedilen bilginin kalitesi ve zenginliği, uygulamanın kullanışlılığını ve gücünü oluşturur (Güvenç 2001). Kümeleme, birliktelik kuralları, çok kullanılan tanımlayıcı modellerdir.

Kümeleme yöntemi, danışmansız sınıflama modeli olarak da bilinir (Pryke 1998). Kümeleme heterojen veri kümelerini veri karakteristikleri bakımından homojen sayılabilecek gruplara bölme bir başka değişle diğerlerinden çok farklı ancak üyeleri çok benzer olan grupları bulma işidir (Web_3 1999, Güvenç 2001). Kümeleme modelinde; veri tabanındaki kayıtların hangi kümelere ayrılacağı veya kümelemenin hangi değişken özelliklerine göre yapılacağı, konunun uzmanı olan bir kişi tarafından belirlenebilir (Akpınar 2005).

Tahmin edici modeller kümeleme modelini, homojen veri grupları oluşturması için veri ön işleme aşaması olarak ta kullanmaktadırlar.

Birliktelik kuralları, bir arada olan olayların ya da özelliklerin keşfedilmesi sürecidir, ilişki analizi ya da pazar sepet analizi olarak da adlandırılır. Birliktelik kuralları genellikle "*eğer şu olursa daha sonra bu olur*" şeklindedir. Birliktelik kuralları oluşturmada en çok kullanılan algoritmalar Apriori ve GRI'dir.

Özetleme tanımlayıcı istatistikleri kullanarak verinin betimlenmesidir, genellikle açıklayıcı veri analizi için uygulanır (Fayyad vd 1996). Görselleştirme, verinin grafik öğeleri yardımıyla betimlenmesidir, genellikle ayrık değerleri tespit etmede, veri ön işlemede, trend ve ilişkilerin bulunmasında kullanılır (Güvenç 2001).

2.1.4.2. Tahmin edici modeller

Tahmin, geçmiş tecrübelerden elde edilen bilgiler ve mantık kullanılarak, gelecekte olması muhtemel durumlar hakkında öngörüde bulunmaktır. Tahmin edici modeller karar alma süreçlerinde önemli bir rol oynar. Tahmin edici modellerde sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerinin tahmin edilmesi amaçlanır (Akpınar 2005). Tahmin edici modellerin temel iki türü sınıflandırma ve regresyondur.

Sınıflandırma, veri nesnesini daha önceden belirlenen sınıflardan biriyle eşleştirme sürecidir (Wang 1992). Verileri ve karşı gelen sınıfları içeren eğitim kümesi ile eğitilen sistem, sonraki aşamalarda sınıf bilgisine sahip olunmayan verilerin ait olduğu sınıfların bulunması için kullanılır (Pryke 1998).

Müşteri segmentasyonu, kredi analizi, iş modellemesi ve benzeri birçok alanda kullanılan sınıflandırma yöntemi günümüzde en çok kullanılan veri madenciliği yöntemidir (Pryke 1998).

Regresyon, sürekli sayısal bir değişkenin, aralarında doğrusal ya da doğrusal olmayan bir ilişki bulunduğu varsayılan diğer değişkenler yardımıyla tahmin edilmesi yöntemidir (Bidgoli 2004).

Regresyon modeli, sayısal değerleri tahmin etmeye yönelik olması dışında sınıflandırma yöntemine benzetilebilir. Çok terimli lojistik regresyon gibi kategorik değerlerin de tahmin edilmesine olanak sağlayan tekniklerin geliştirilmesi ile sınıflandırma ve regresyon modelleri giderek birbirine yaklaşmakta ve dolayısıyla aynı tekniklerden yararlanılması mümkün olmaktadır. Sınıflandırma ve regresyon modellerinde kullanılan başlıca teknikler;

- Yapay Sinir Ağları
- Genetik Algoritmalar
- K-En Yakın Komşu
- Naïve-Bayes
- Çoklu Regresyon, Lojistik Regresyon
- Faktör ve Ayırma analizleri
- Karar Ağaçları

şeklinde sıralanabilir (Akpınar 2005).

Karar ağacı, çoklu regresyondaki sınırlılıkları aşmak amacıyla geliştirilmiştir. Bu yöntemde karar ağaçları kullanılarak veri kümesi sonlu sayıda sınıfa ayrılır. Karar ağacındaki düğümler nitelik isimleriyle, dallar nitelik değerleriyle, yapraklar da farklı sınıf isimleriyle etiketlenir (Sörensen ve Jassens 2003). Kök düğüm olarak da adlandırılan ilk eleman en yüksek karar düğümüdür, kullanılan algoritmaya bağlı olarak her düğüm iki veya daha fazla dala sahip olur. İki dala sahip olan karar ağaçları ikili ağaç, daha fazla dala sahip olanlar ise çok yollu ağaç olarak adlandırılır. Her dal bir başka karar düğümüyle, ya da ağacın sonuyla yani yaprak düğümle sonlanır. Karar düğümlerinde gerçekleştirilen her bölünmede oluşturulan gruplar arasındaki mesafenin maksimum olması bir başka değişle elde edilen grupların mümkün olduğu kadar saf olması istenir. Kategorik değerleri sınıflandırmak için oluşturulan karar ağaçlarına sınıflandırma ağacı, sürekli sayısal değişkenleri tahmin etmek için kullanılan karar ağaçlarına ise regresyon ağacı denilmektedir. Karar Ağacı oluşturmak için CHAID, CART, Quest ve C5.0 gibi algoritmalar kullanılır.

En çok bilinen ve kullanılan evrim algortiması olan genetik algoritmalar kavramı, 1975 yılında Michigan Üniversitesi'nde John Holland ve arkadaşları tarafından oluşturulmuştur. Genetik algoritmaların adı ve işleyiş mekanizması doğal seleksiyon modelinden esinlenerek oluşturulmuştur (Collard ve Francisci 2001). Genetik algoritmalar optimum çıktıları elde etmek için gerekli olan girdileri üretmeye ve test etmeye olanak sağlayan bilgisayar tabanlı bir arama metodudur (Alkan 2001). Bu tekniğin veri madenciliğindeki ilk uygulamaları yapay sinir ağları gibi öğrenme araçlarının optimizasyonuydu ancak genetik algoritmalarda bireylerden oluşan bir

popülasyon kullanıldığından günümüzde popülasyondaki bireyler, örüntüleri sembolize etmek için kullanılabilmektedir (Collard ve Francisci 2001).

Kural çıkarsama yöntemi farklı olayları sınıflandırmak için "eğer ise" kuralları oluşturma tekniğidir. Kural oluşturmaya yönelik diğer bir yöntem olan karar ağacı yönteminden farklı olarak, kural çıkarsama yönteminde bağımsız kurallar oluşturulabilir, yani kuralların bir ağaç oluşturması gerekmez. Kural çıkarsama yöntemiyle oluşturulan kurallar tüm olasılıkları kapsamayabilir. Bu yöntemin karar ağacı yönteminden farklı olduğu bir diğer nokta da kuralların çelişme ihtimali olmasıdır.

İnsanlar yeni problemleri çözmeye çalışırken genellikle daha önce çözdükleri benzer problemlerin çözümlerine bakarlar. K en yakın komşuluk algoritması problem çözümü için benzer bir tekniği kullanan sınıflandırma tekniğidir. Bu teknikte yeni bir durum daha önce sınıflandırılmış benzer, en yakın komşuluktaki k tane olaya bakılarak sınıflandırılır. K en yakın komşuluğundaki olayların ait olduğu sınıflar sayılır ve yeni durum sayısı fazla olan sınıfa dahil edilir (Web_3 1999). Bu yöntemde ilk olarak nitelikler arasındaki mesafeyi ölçmek için bir ölçme yöntemi oluşturulur. Olaylar arasındaki uzaklıklar hesaplandıktan sonra, yeni olayların sınıflandırılması için halihazırda sınıflandırılmış olan durumlar temel olarak alınır. Uzaklık karşılaştırmasına kaç adet olayın dahil edileceği (k'nın belirlenmesi) ve komşuluk hesaplamalarının nasıl yapılacağına karar verilir. Komşuluk hesaplamaları yapılırken, daha yakın komşulara daha büyük ağırlık değerleri atanabilir (Güvenç 2001).

K en yakın komşuluk yönteminde sınıflandırılmak istenen olay sayısı arttıkça hesaplamalar için gereken sürede hızlı bir şekilde artar, k en yakın komşuluk modelinin işlem hızını artırmak için genellikle bütün veri hafızada tutulur. Bellek tabanlı nedenselleştirme (reasoning) bellekte tutulan k en yakın komşu sınıflandırmasını ifade eder (Web_3 1999).

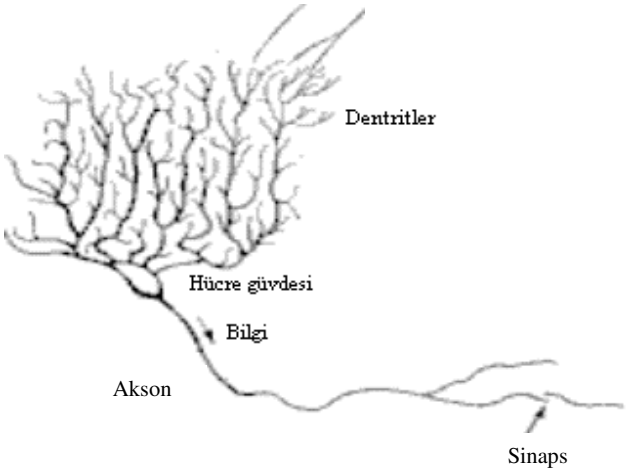
3. YAPAY SİNİR AĞLARI

Yapay sinir ağları biyolojik sinir sisteminin taklit edilmesi, matematiksel olarak modellenmesi çabalarının bir sonucu olarak ortaya çıkmıştır (Yurtoğlu, 2005). Biyolojik sinir sistemindeki bilinen yapılar ve işlevleri yapay sinir ağlarında matematiksel modellerle, biyolojik sinir sistemindeki eşleniklerinin görevlerini yerine getirecek şekilde modellenmiştir. İnsan beyninin yaklaşık olarak 10^{11} tane nöron olarak adlandırılan hesap elemanından oluştuğu ve bu nöronlar arasında 10^{15} bağlantı bulunduğu düşünülmektedir (Kantardzic 2003). Biyolojik sinir ağını oluşturan nöronlar

- Soma
- Akson
- Dentrit

olmak üzere üç bölgeye ayrılır. Bu bölgelerin her biri bilgilerin girişinde ve iletiminde belirli bir rol oynamaktadır (Nabiyev 2003).

Biyolojik sistemlerde öğrenme, nöronlar arasındaki sinaptik bağlantıların ayarlanması ile oluşturulur. İnsan yaşamı süresince tecrübeler edinir, bu tecrübelerin sinaptik bağlantıları etkilediği ve öğrenmenin bu şekilde geliştiği düşünülmektedir. Yapay sinir ağlarında bu ayarlamayı yapmak ve öğrenmeyi sağlamak için ağırlık fonksiyonları kullanılmaktadır, insanın deneme yanılma yoluyla öğrenmesi yapay sinir ağlarında yinelemeli eğitim sayesinde gerçekleştirilmektedir. (Yurtoğlu 2005). Şekil 3.1'de bir biyolojik sinir hücresinin yapısı gösterilmiştir.



Şekil 3.1 Sinir hücresi (Web_5 2006)

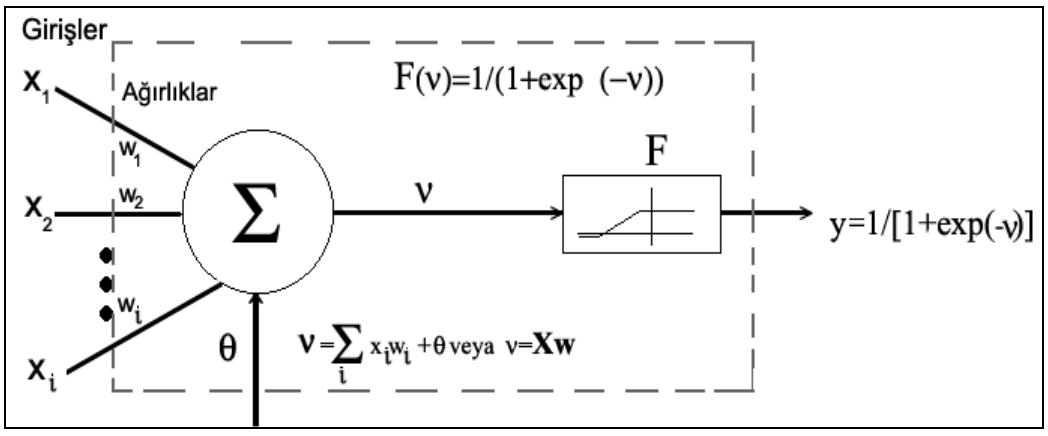
3.1. Yapay Nöron Modeli

Şekil 3.2'de temel bir yapay sinir ağı nöronunu ($X_1, X_2 \ldots\ldots X_n$) girdileri ve bu girdilere karşı düşen ($W_1, W_2 \ldots\ldots\ldots\ldots W_n$) ağırlıklarını göstermektedir. Girdiler işlenmeden önce ağırlıklarla çarpılarak toplanır.

$$\nu = X_1 . W_1 + X_2 . W_2 + \ldots\ldots\ldots + X_n . W_n \qquad (3.1)$$

Üretilen işaret (ν) aktivasyon fonksiyonuna gönderilir ve çıktı (y) elde edilir.

Ağırlık değerleri, kullanılan yapay sinir ağı yapısına göre eğitim esnasında yenilenir (Ramaswamy 1997).



Şekil 3.2 Yapay nöron modeli (Aydın 2005)

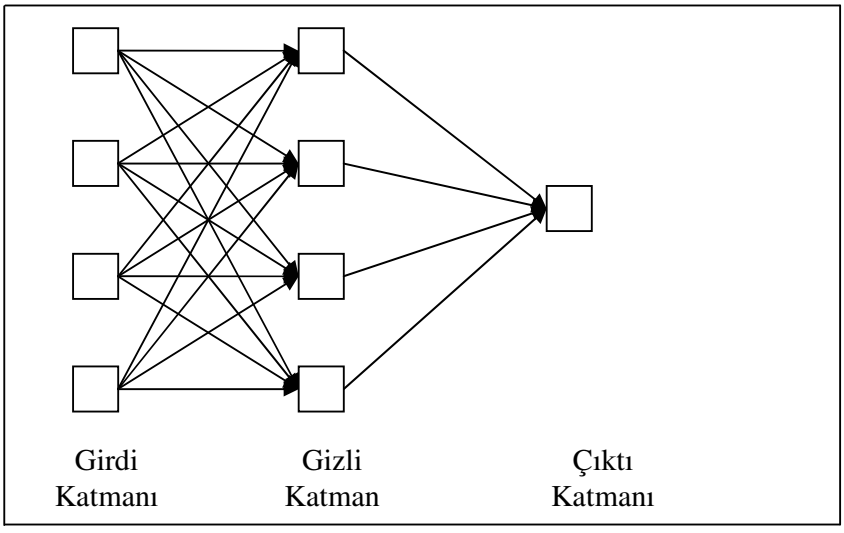
3.2. Yapay Sinir Ağlarının Sınıflandırılması

Yapay sinir ağlarını için standart bir sınıflandırma yöntemi yoktur. Nöronların bağlanma şekillerine, kullanılan öğrenme algoritmasına, zaman gecikmesine ve işlenilen verinin türüne göre çeşitli sınıflandırmalar yapmak mümkündür (Kim 2003).

Nöronların bağlanma biçimlerine göre yapay sinir ağları ileri beslemeli ve geri beslemeli yapay sinir ağları olarak ikiye ayrılır (Slaughter 2003).

3.2.1. İleri beslemeli yapay sinir ağları

Şekil 3.3'te fiziksel yapısı gösterilen ileri beslemeli yapay sinir ağlarında nöronlar katmanlar şeklinde düzenlenir ve bir katmandaki nöronların çıkışları bir sonraki katmana ağırlıklar üzerinden giriş olarak verilir. Aynı katmandaki nöronlar arasında veya bir önceki katmana bağlantı yani geri besleme çevirimi yoktur. Giriş katmanı, dış ortamlardan aldığı bilgileri hiçbir değişikliğe uğratmadan gizli katmandaki nöronlara iletir. Bilgi orta katmanlarda ve çıkış katmanında işlenerek ağ çıkışı belirlenir. İleri beslemeli yapay sinir ağlarına örnek olarak çok katmanlı perseptron ileriki bölümlerde ayrıntılı olarak incelenecektir (Web_4 2006).

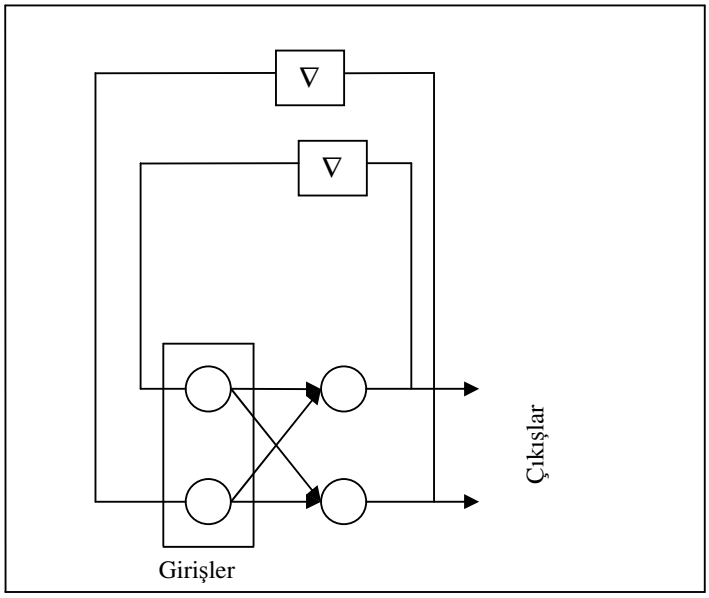


Şekil 3.3 İleri beslemeli ağ modeli

3.2.2. Geri beslemeli ağlar

Geri beslemeli ağlarda en az bir tane geri besleme çevirimi bulunur. Geri beslemenin yapılış biçimi Şekil 3.4'de gösterilmiştir. Geri besleme, aynı katmandaki hücreler

arasında olabileceği gibi farklı katmanlardaki nöronlar arasında da olabilir. Geri beslemenin yapılış şekline göre farklı yapı ve davranışta geri beslemeli yapay sinir ağı yapıları elde edilebilir (Web_4 2006).



Şekil 3.4 Geri beslemeli iki katmanlı ağ modeli

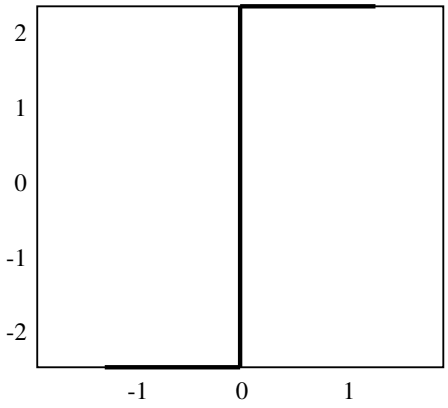
3.3. Aktivasyon Fonksiyonları

Aktivasyon fonksiyonları nörona gelen net girdiyi işleyerek nöronun bu girdiye karşılık vereceği tepkiyi belirler. Aktivasyon fonksiyonları temel olarak üç grupta incelenebilir.

3.3.1. Eşik aktivasyon fonksiyonu

McCulloch-Pitts modeli olarak bilinen eşik aktivasyon fonksiyonlu hücreler, mantıksal çıkış verir ve sınıflandırıcı ağlarda tercih edilir. Eşik fonksiyonlu hücrelerin matematiksel modeli (3.2) numaralı denklemde verilmiştir (Web_4 2006). Eşik aktivasyon fonksiyonun grafiği Şekil 3.5'te çizilmiştir.

$$y = \begin{cases} 1 & \nu > x \\ 0 & \nu <= x \end{cases} \tag{3.2}$$

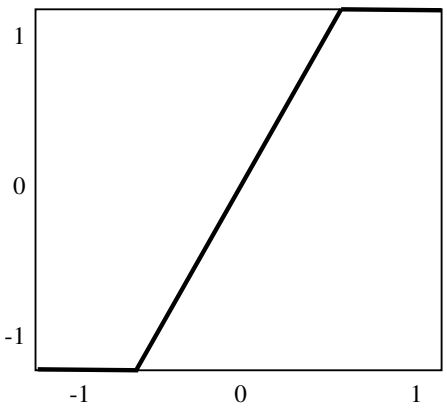


Şekil 3.5 Eşik aktivasyon fonksiyonu

3.3.2. Doğrusal ve doyumlu-doğrusal aktivasyon fonksiyonu

Doğrusal bir problemi çözmek amacıyla kullanılan doğrusal aktivasyon fonksiyonu, hücrenin net girdisini doğrudan hücre çıkışı olarak verir. Doğrusal aktivasyon fonksiyonu matematiksel olarak y=v şeklinde tanımlanabilir. Doyumlu doğrusal aktivasyon fonksiyonu ise aktif çalışma bölgesinde doğrusaldır ve hücrenin net girdisinin belirli bir değerinden sonra hücre çıkışını doyuma götürür. Doyumlu doğrusal aktivasyon fonksiyonun matematiksel modeli (3.3) nolu denklemde grafiği Şekil 3.6'da gösterilmiştir(Web_4 2006).

$$y = \begin{cases} 1 & v > 1 \\ v & -1 < v < 1 \text{ ise} \\ -1 & v < -1 \end{cases} \tag{3.3}$$

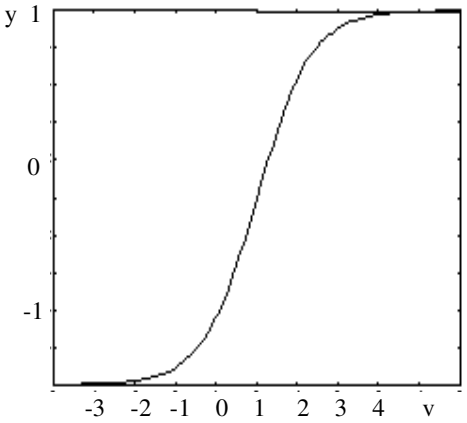


Şekil 3.6 Doyumlu doğrusal aktivasyon fonksiyonu

3.3.3. Sigmoid aktivasyon fonksiyonu

S şekilli aktivasyon fonksiyonu olarak da isimlendirilen bu fonksiyon sinir ağlarında en çok kullanılan aktivasyon fonksiyonudur (Haykin 1998). Çok katmanlı perseptron gibi bazı yapay sinir ağı modelleri aktivasyon fonksiyonunun türevlenebilir olmasını gerektirmektedir, bu şart sigmoid aktivasyon fonksiyonunda sağlanır. Bu fonksiyonun -5 ile +5 arasındaki giriş değerleri için ürettiği çıkış değerleri Şekil 3.7'de gösterilmiştir.

$$Y=\frac{1}{1+e^{-v}} \tag{3.4}$$



Şekil 3.7 Sigmoid aktivasyon fonksiyonu

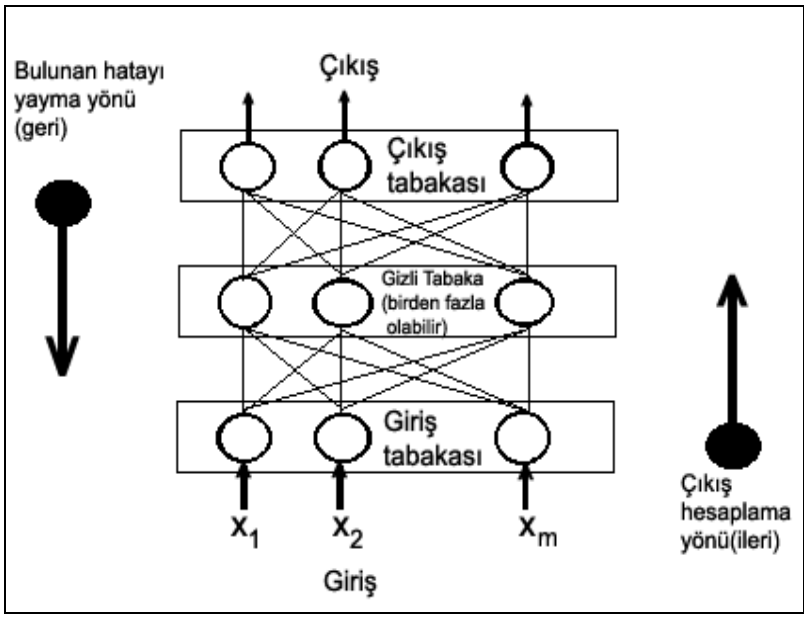
Bu eğrinin eğimi v 'nin alacağı değerlere göre değişir, ve eğim sonsuza gittiğinde sigmoid fonksiyonu, eşik değer fonksiyonuna benzer.

3.4. Çok Katmanlı Perseptronlar

Çok katmanlı ileri beslemeli ağlar, bir girdi katmanı, bir veya daha fazla gizli katman ve bir çıktı katmanından oluşur. Şekil 3.8'de yapısı gösterilen çok katmanlı perseptron modelinde girdi katmanı, gizli katmanlar ve çıktı katmanındaki nöronlar tamamen ya da bölgesel olarak ileri yönde bağlıdırlar. Çok katmanlı perseptronlarda aynı katman nöronları arasında veya önceki katman nöronlarına geribesleme bağlantıları bulunmaz. Çok katmanlı perseptron modelinde doğrusal olmayan aktivasyon fonksiyonu kullanan en az bir gizli katman bulunur. Birçok algoritma için eğitim süresi gizli katman ve nöron sayısına bağlı olarak hızla artmaktadır. Pratikte karşılaşılan problemlerin çözümünde çoğunlukla iki katman yeterlidir. (Web_4 2006).

Çok katmanlı perseptronlarda girdi katmanı nöron sayısı genellikle problemdeki girdi sayısına, çıktı katmanı nöron sayısı da istenilen çıktı sayısına eşit tutulur. Gizli katman sayısı ve gizli katmandaki nöronların sayısı deneylerle tespit edilir. Gizli katman ve gizli katmanlarda yer alan nöron sayısının fazla olması eğitim kümesinde hata oranını düşürür ancak test kümesinde hata oranının artmasına neden olur. Nöron sayısının artması bağlantı sayısının dolayısıyla eğitilmesi gereken ağırlık sayısın artmasına neden olur (Auclair 2004).

Girdi katmanındaki nöronlar, girdiler üzerinde bir değişiklik yapmadan sonraki katmana gönderir. Gizli katmanda ve çıkış katmanında bağlantıların ağırlık değerleriyle çarpılarak gelen veriler toplanır ve aktivasyon fonksiyonundan geçirilir. Aktivasyon fonksiyonunun türevlenebilir olması istendiğinden genellikle sigmoid fonksiyonu kullanılır (Mohamed 2004).



Şekil 3. 8 Çok katmanlı perseptron (Aydın 2005)

3.5. Yapay Sinir Ağlarında Öğrenme

Eğitim, yapay sinir ağları uygulamalarındaki en önemli süreçlerden biridir, girdi bilgisi ve çıktı arasında bir eşleştirme yapılması amacını taşır. Ağ yeterince eğitildiğinde daha önceden görmediği girdilere karşı uygun çıktılar üretir. Bu özellik yapay sinir ağlarının genelleme kapasitesi olarak adlandırılır ve çoğunlukla ağın yapısına ve eğitim için uygulanan yineleme sayısına bağlıdır (Alkan 2001).

Yapay sinir ağlarının eğitiminde kullanılacak olan ağın modeli ve problemin yapısına bağlı olarak danışmanlı, danışmansız ve destekleyici öğrenme stratejileri uygulanmaktadır (Web_4 2006).

3.5.1. Danışmanlı öğrenme

Bu tip öğrenmede hedef çıktı, eğitim kümesinin bir elemanı olarak ağa sunulur. Ağın görevi girdileri, sunulan çıktı değerleriyle eşleştirmek, dolayısıyla girdiler ile çıktılar arasındaki ilişkiyi belirlemektir. Öğrenme, ağın ürettiği değerlerle hedef değerler arasındaki farkın, hata kriteri tarafından istenilen düzeye gelene kadar ya da belirli bir yineleme sayısına kadar devam eder (Alkan 2001, Web_4 2006). Bu kriter sağlandığında yapay sinir ağının genelleme performansı daha önce ağa verilmeyen test verileri ile değerlendirilir(Alkan 2001).

3.5.2. Danışmansız öğrenme

Bu tip öğrenmede eğitim kümesi hedef çıktı değerlerini içermez, verilerdeki parametreler arasındaki ilişkilerin ağ tarafından bulunması beklenir. Eğitim, ağ tutarlı değerler üretinceye kadar yani üretilen hata oranı belirli bir aralığa düşünceye kadar devam eder. Danışmansız öğrenme daha çok sınıflandırma problemleri için kullanılan bir eğitim yöntemidir.

3.5.3. Destekleyici öğrenme

Destekleyici öğrenme yönteminde eğitim kümesinde hedef çıktılar bulunmaz sadece ağa ürettiği çıktıların doğru ya da yanlış olduğu söylenir, ağ bunu dikkate alarak öğrenme sürecini devam ettirir. LVQ ağları bu öğrenme tipini kullanan ağlara örnek olarak gösterilebilir (Web_4 2006).

3.6. Geri Yayılımlı Öğrenme

Geri yayılım algoritması adını hatayı yayma biçiminden alır. Bu algoritmada elde edilen çıktı ve olması gereken çıktı arasındaki fark yani hata, tüm ağırlıklara yansıtılır (Ethridge ve Zhu 1996). Geri yayılım algoritmasında eğitime rastgele bir ağırlık kümesi ile başlanır, birçok uygulamada ağın başarısı ağa atanan ilk ağırlık değerlerinin uygun seçilmesine bağlıdır (Nabiyev 2003).

Örneklere eğitim algoritmasını uygulamanın iki yolu vardır.

- Eğitim kümesindeki her örnek için tüm ağın hatası en aza indirilir.

- Eğitim kümesindeki tüm örnekler uygulanarak bir nöronun hatası en aza indirilir daha sonra zincirdeki diğer nörona geçilir.

Yapay sinir ağlarının eğitiminde hatayı en aza indirmek için genellikle, hata fonksiyonunun yönünü bulmaya ve hata fonksiyonunu azaltmaya çalışan dereceli azaltma (gradient descent) tabanlı algoritmalar kullanılır (Auclair 2004).

Geri yayılım algoritmasında hata tespti için ölçüt olarak ortalama hatanın karesi kullanılır. Ortalama hatanın karesi (3.5) numaralı denklem yardımıyla bulunur.

$$E=\frac{1}{2}\sum_{k}(t_k - y_k)^2 \tag{3.5}$$

t_k =Çıktı katmanındaki k. nöronun hedef çıktısı

y_k = Çıktı katmanındaki k. nöronun gerçek çıktısı

Geri yayılım algoritmasında her bir ağırlığın değişim miktarı ise (3.6) numaralı denklem yardımıyla elde edilir.

$$\Delta W_{jk} = -\varepsilon.\frac{\partial E}{\partial W_{jk}} \tag{3.6}$$

ε : Öğrenme katsayısı

Geri yayılım algoritmasının öğrenme hızını artırmak için denkleme momentum terimi(μ) eklenebilir(Alkan 2001).

$$\Delta W_{jk} = -\varepsilon.\frac{\partial E}{\partial W_{jk}}\ (t+1)+\ \mu\ \Delta W_{jk}\ (t) \tag{3.7}$$

W_{jk} = j. Birimden k. birime bağlantının ağırlık değeri

$$\Delta W_{jk}\ (t+1) = -(1-\mu)\varepsilon.\frac{\partial E}{\partial W_{jk}}\ (t+1)+\ \mu\ \Delta W_{jk}\ (t) \tag{3.8}$$

(3.7) numaralı denkleme $(1-\mu)$ teriminin dahil edilmesi momentum katsayısı (μ) arttıkça, öğrenme katsayısı ε'nin azaltılması gereksinimini ortadan kaldırır (Alkan 2001).

Eşlenik gradyan veya quasi newton gibi yöntemler basit dereceli azaltma (gradient descent) yönteminden daha çabuk yakınsar ancak bu yöntemlerde genellikle hata

yüzeyinin karesel bir fonksiyonla modellenebileceği varsayılır, bu varsayımın tutmadığı durumlarda bu yöntemler çok başarılı sonuçlar üretemez (Auclair 2004) .

Yavaş olması ve basit bir problemin çözümünde bile yapay sinir ağının eğitiminin binlerce yineleme gerektirmesi geriye yayılım algoritmasının dezavantajlarıdır. Bu algoritmanın başarısı ağırlık katsayılarının ilk değerlerine, momentum ve öğrenme katsayısı değerlerinin seçimine bağlıdır.

3.7. Yapay Sinir Ağı Parametreleri

Yapay sinir ağlarının performansları farklı parametreler için önemli ölçüde değişir. Dolayısıyla oluşturulan modelin başarısı seçilen ağ yağısı kadar bu ağ için seçilecek öğrenme katsayısı, gizli katman ve nöron sayısı gibi parametrelere de bağlıdır.

3.7.1. Gizli katman ve nöron sayısının belirlenmesi

Gizli katmanlar hatayı geri yayma algoritmasının merkezini oluşturur ve yapay sinir ağlarının gücünü oluşturan katmanlardır. Gizli katmanlar üst seviye özelliklerin tespit edilmesinde ve yapay sinir ağının genelleme özelliğini kazanmasında önemli bir yere sahiptir (Iskandar 2005). Yapay sinir ağı modelinin tahmin kapasitesinin maksimum olmasında optimum gizli katman ve nöron sayısının bulunması önemlidir.

Kim (2003)'in aktardığına göre Hornick Stinhcombe ve White çalışmalarında sigmoid çıkış fonksiyonlu üç katmanlı bir perseptronun yeterli eğitim ile evrensel yakınsayıcı (Universal Approximator) olarak kullanılabileceğinin ancak kaynak kullanımı ve veri kümesinin özelliği gibi etkenlerde gözönüne alındığında katman sayısının her uygulama için farklı olabileceğini göstermişlerdir.

Yapay sinir ağının sahip olması gereken gizli nöron sayısının tespiti için çeşitli formüller üretilmiştir. Bir yapay sinir ağının yapısını InHmOp şeklinde tanımlarsak,

I: Girdi Katmanı
H: Gizli Katman
O: Çıktı Katmanı
n,m,p: Nöron Sayıları

bu ağdaki toplam parametre sayısı (mxn)+(mxp) olarak bulunabilir (Kim 2003).

Baum ve Hausler (1989) bir yapay sinir ağı için optimum parametre sayısının eğitim kümesindeki kayıt sayısının %10'unu aşmaması gerektiğini söylemişlerdir. Örneğin

eğitim kümesi 600 kayıttan oluşan bir problem için oluşturulan yapay sinir ağı modelinin 60'dan fazla parametre içerecek şekilde yapılandırılmaması gerekir. Ancak birçok çalışma bu kuralın çalışmadığını göstermiştir (Kuligowski ve Barros 1998, Swingler 1996).

Yapay sinir ağlarında gizli katman ve nöron sayısının artması verinin içindeki gizli örüntüleri yakalama konusunda ağın şansını artırırken, ağın veri kümesindeki özellikleri ezberlemesine neden olabilir. Bu durumda ağ veri kümesi için küçük hata oranları üretirken test kümesi için ürettiği hata oranı artmaya başlar bir başka değişle ağ genelleme özelliğini kaybeder.

3.7.2. Sonlandırma kriteri

Yapay sinir ağları yinelemelerle öğrenir ancak yineleme sayısının büyük seçilmesi, ağın eğitim kümesini ezberlemesine neden olur. Yineleme sayısı arttıkça ağın eğitim kümesi için ürettiği hata azalırken bir noktadan sonra test kümesi için üretilen hata artmaya başlar, bu durum ağın verinin özelliklerini modellemek yerine ezberlemeye başladığını gösterir.

Yapay sinir ağları uygulamalarında genel olarak üç farklı yöntem kullanılır.

1. Tüm bağlantı ağırlık değerleri değişimleri belirli bir eşik değerin altında kaldığında eğitim sonlandırılır.
2. Ağın ürettiği çıktı değerleri ile istenen çıktı değerler arasındaki hata belirlenen bir oranın altına düştüğünde eğitim sonlandırılır.
3. Belirlenen yineleme sayısına ulaşıldığında eğitim sonlandırılır (Han ve Kamber 2000).

3.7.3. Öğrenme katsayısı

Öğrenme katsayısı (ε), hemen hemen bütün yapay sinir ağlarında kullanılan, 0 ile 1 arasında değer alan bir sabittir. Öğrenme katsayısı öğrenme hızını kontrol eder. Danışmanlı öğrenmede ağın bağlantı ağırlıklarının üretilen çıktıdan istenen çıktıya daha yakın değerler elde edecek şekilde uyarlanmasında kullanılır (Pigus 1996). Uygun öğrenme katsayılarının seçilmesi yapay sinir ağının karar uzayındaki yerel minimumlara sıkışmasını önler (Han ve Kamber 2000).

Öğrenme katsayısının büyük seçilmesi ağın istenen değere yakınsaması için geçen süreyi azaltır ancak ağın yetersiz iki çözüm arasında salınıma girmesine neden olabilir. Öğrenme katsayısının küçük seçilmesi ise ağın yakınsama zamanını uzatır (Larose 2005). Genellikle öğrenme katsayısının eğitimin başlangıcında yüksek olması zamanla azalması istenir. Bunu gerçekleştirmek için bazı uygulamalarda öğrenme katsayısı yineleme sayısına bölünmektedir. Bu yöntemle öğrenim katsayısının küçük seçildiği durumlarda oluşan geç yakınsama sorunu azaltılır bu yapılırken öğrenme katsayısı büyük seçildiğinde oluşan yetersiz çözümler arasında salınıma girme riski ortadan kaldırılır.

4. REGRESYON ANALİZİ

Regresyon analizi bir bağımlı değişken ile bir veya daha fazla sayıda bağımsız değişken arasındaki ilişkiyi sayısal hale dönüştürmek için kullanılan istatistiksel analiz yöntemidir. Regresyon analizi esas olarak değişkenler arasındaki ilişkinin niteliğini saptamayı amaçlar. Bağımsız değişken olarak bir değişken kullanılırsa basit regresyon, iki veya daha fazla değişken kullanılırsa çoklu regresyon analizi olarak adlandırılır. Regresyon analizinde amaç her bağımsız değişkenin bağımlı değişkendeki değişmeye katkısının hesaplanması dolayısıyla tahmin değişkenlerinin değerinden hareketle bağımlı değişkenin değerinin tahmin edilmesidir (Xu 2003).

4.1. Basit Doğrusal Regresyon

Aralarında doğrusal ilişki bulunan bir bağımlı ve bir bağımsız değişken arasındaki ilişkiyi çözümleyen regresyon analizi türüdür. Basit doğrusal regresyon modeli (4.1) numaralı denklemde verilmiştir.

$$Y_i = b_0 + b_1 X_i + e_i \tag{4.1}$$

b_0 : Doğrunun y eksenini kestiği nokta

b_1 :Regresyon katsayısı

e_i :Hata değeri

4.2. Çoklu Regresyon

Çoklu regresyon analizi bağımsız değişkenlerden yola çıkarak bağımlı değişkenin tahmin edilmesi ve hangi bağımsız değişkenlerin bağımlı değişken üzerinde anlamlı bir etkiye sahip olduğunun bulunması için kullanılır. Regresyon analizinde örnek veri noktalarını en iyi temsil eden çizgi ya da düzlem bulunmaya çalışılır (Xu 2003).

Çoklu doğrusal regresyon modeli (4.2) numaralı denklemde verilmiştir.

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + ... + b_p X_{pi} + e_i \tag{4.2}$$

b_0 : Doğrunun y eksenini kestiği nokta

$b_1, b_2, \ldots, b_p$:Regresyon katsayıları, e_i :Hata değeri

Çoklu regresyon modeli $X_{1i}, X_{2i}, \ldots, X_{pi}$ girdi değerlerine bağlı olarak Y_i değerinin tahmin edilmesini sağlar. Modelde yer alan regresyon katsayıları en küçük kareler yöntemi kullanılarak hesaplanır.

Oluşturulan regresyon modelinin veriyi ne kadar iyi açıkladığının ölçülmesi için çeşitli ölçüler kullanılmaktadır. Bu ölçülerden en çok kullanılanı açıklayıcılık katsayısı R^2 ve düzeltilmiş R^2'dir.Bu ölçüler 0-1 aralığında değerler alır ve değerin büyüklüğü modelin uygunluğunu gösterir. Çoklu açıklayıcılık katsayısı (R^2) ve düzeltilmiş R^2 bağımsız değişkenlerin bağımlı değişkeni ne kadar açıkladığını anlamak içinde kullanılır. Regresyon varsayımlarının yerine getirilmediği durumlarda, uygun olmayan modeller için de R^2'nin yüksek değerler alması mümkündür dolayısıyla R^2 model uygunluğu için güvenilir bir ölçü değildir Düzeltilmiş R^2 model uygunluğu ölçümünde göreceli olarak daha iyi bir performansa sahiptir ve birçok istatistik yazılımı bu ölçütü kullanmaktadır (Xu 2003).

4.2.1. Çoklu regresyon analizinde kullanılan yöntemler

Çoklu regresyon analizinde kullanılan pek çok yöntem vardır. Standart çoklu regresyon, hiyerarşik çoklu regresyon ve istatistiksel çoklu regresyon bunlardan en çok kullanılanlarıdır.

4.2.1.1. Standart çoklu regresyon

Bu yöntemde, bütün bağımsız değişkenler aynı anda denkleme girer. Bağımsız değişkenlerin her biri, diğer bağımsız değişkenlerin hepsi denkleme girdikten sonra denkleme alınmış gibi değerlendirilir. Her bir bağımsız değişkenin bağımlı değişkeni tahmin etmede ne kadar katkıda bulunduğu ortaya konulur (Tabachnick ve Fidell 2001).

4.2.1.2. Hiyerarşik çoklu regresyon

Bu çoklu regresyon yönteminde bağımsız değişkenlerin modele dahil edilme sırasına, çalıştığı konuyu göz önüne alarak araştırmacı karar verir. Araştırmacı değişkenleri modele dahil etme sırasını modele en çok katkısı bulunan değişkenden, en az katkısı bulunana doğru seçebileceği gibi bunun tam tersini de seçebilir (Tabachnick ve Fidell 2001).

4.2.1.3. İstatistiksel çoklu regresyon

İstatistiksel çoklu regresyon analizi üç farklı yöntemle yapılabilir. İleriye doğru seçme yönteminde her bir bağımsız değişkenle bağımlı değişken arasındaki korelasyon hesaplanır ve öncelikle bağımlı değişkenle en yüksek korelasyonu veren bağımlı değişken analize dahil edilir. Bu değişkenin katkısı R^2 terimi incelenerek değerlendirilir. Daha sonra, ikinci en yüksek korelasyon katsayısına sahip bağımsız değişken analize alınır ve açıklayıcılık katsayısındaki artışa göre söz konusu değişkenin modele katkısı incelenir. Bu işlem bağımsız değişkenlerin bağımlı değişkeni açıklamada anlamlı bir katkılarının olmadığı görülene kadar devam eder. Anlamlılık ölçütü olarak daha önceden belirlenen α değeri kullanılır (Erdem vd 2006).

Adım adım regresyon yöntemi ileriye doğru seçme yönteminin daha gelişmişi olarak da düşünülebilir. Bu yöntemde, her adımda o an modelde bulunan tüm bağımsız değişkenler sanki modele en son girmiş gibi değerlendirilir. Bu şekilde her bir değişkenin modele girmesiyle yeniden tüm modelin değerlendirilmesi sayesinde başta iyi bir tahmin edici olarak görülen bir değişkenin daha sonra tüm model içinde etkili bir katkısının olmadığı görülebilir (Erdem vd 2006).

Geriye doğru çıkarma yöntemine bütün bağımsız değişkenlerin analize dahil edildiği bir modelle başlanır. Daha sonra, her bir bağımsız değişkenin *p* değeri daha önceden belirlenmiş α değeriyle kıyaslanır ve *p* değeri α'dan büyük olan değişkenler model dışında bırakılır (Erdem vd 2006).

5. YÖNTEM VE MODEL OLUŞTURMA

Bu çalışmada veri madenciliği uygulaması Bölüm 2.1.2'de açıklanan CRISP-DM referans modeli takip edilerek gerçekleştirilmiştir. Bu bölümde CRISP-DM referans modelinin problemin değerlendirilmesi ve amacın belirlenmesi, verinin incelenmesi, verinin hazırlanması ve model oluşturma aşamalarında gerçekleştirilen işlemler anlatılacaktır.

5.1. Problemin Değerlendirilmesi ve Amacın Belirlenmesi

Bu çalışmada KPSS sonuçlarının veri madenciliği yöntemi kullanılarak tahmin edilmesinde çoklu regresyon analizi ve yapay sinir ağları yöntemlerinin başarılarının karşılaştırılması amaçlanmıştır.

Bu karşılaştırmayı gerçekleştirmek için, Pamukkale Üniversitesi, Eğitim Fakültesi, İlköğretim Bölümü, Sınıf Öğretmenliği A.B.D öğrencilerinin KPSS'den aldıkları puanları, öğrencilerin lisans eğitimleri süresince bazı derslerden aldıkları geçme notları, genel not ortalamaları ve öğretim türleri tahmin edici değişkenler olarak kullanılarak öngörülmeye çalışılmıştır.

Bu çalışmada aşağıdaki süreç izlenilmiştir;

- KPSS'de soru çıkan dersleri belirlemek,
- Ulaşılması mümkün veri kümesi büyüklüğünü bulmak,
- Bu derslere ait not ortalamaları, öğrencilerin genel not ortalamaları ve KPSS puanlarına ilişkin verileri temin etmek,
- Bu çalışma için kullanılabilecek uygun program ve teknikleri belirlemek ve uygulamak.

Veri madenciliği öngörü modeli ile ilgili daha önce yapılan çalışmalar incelendiğinde, birden çok tahmin edici değişkene sahip ve tahmin edilmesi istenen değişkenin veri türünün sürekli sayısal değer olduğu durumlarda öğrenme modeli olarak

hatayı geri yayma metodunu kullanan ileri beslemeli yapay sinir ağları, genetik algoritmalar ve çoklu regresyon tekniklerinin kullanıldığı görülmüştür.

5.2. Veri Değerlendirme

Bu çalışmada PAÜ Eğitim Fakültesi, İlköğretim Bölümü, Sınıf Öğretmenliği A.B.D'na 1999, 2000 ve 2001 yıllarında kayıt olan öğrencilere ait veriler kullanılmıştır. Kullanılan veri kümesi Pamukkale Üniversitesi Öğrenci İşleri Bölümü'nden ve ÖSYM internet sitesinden edinilen verilerin birleştirilmesi suretiyle oluşturulmuştur. Bu çalışma içerisinde öğrencilerin ders geçme notlarını barındıran ve PAÜ Öğrenci İşleri Bölümü'nden edinilen verileri içeren tablo, not veri kümesi, genel not ortalamalarını içeren tablo, ortalama veri kümesi, KPSS puanları için oluşturulan tabloda puan veri kümesi olarak isimlendirilecektir. Aynı ders için normal öğretim ve ikinci öğretimde farklı optik kodlar kullanılmasına karşın, veri kümesine her öğrencinin öğretim türünü gösteren bir alan eklenmiş ve bu dersler içi normal öğretim optik kodları kullanılmıştır. Tablo 5.1 ve 5.2'de veri kümelerinde tutulan verilerin türleri gösterilmiştir.

Tablo 5.1 Not veri kümesi veri türleri

Alan Adı	Veri Türü
S.No	SAYI
Öğr. No	SAYI
Ders Kodu	SAYI
Ders Adı	METİN
Ders Geçme Notu	METİN
Dönem	SAYI

Tablo 5.2 Ortalama veri kümesi veri türleri

Alan Adı	Veri Türü
S.No	SAYI
Öğr. No	SAYI
TC Kim. No	SAYI
ÖSS Puanı	SAYI
Akademik Ortalama	SAYI

ÖSYM internet sitesinden alınan KPSS puanları ortalama veri kümesine dahil edilerek puan veri kümesi oluşturulmuştur.

5.3. Verinin Hazırlanması

Not veri kümesi, Microsoft Office XP Access programı üzerinde yürütülen SQL sorgularıyla, her ders için bir tablo oluşturacak şekilde bölünmüştür. Veri temizleme aşamasında, aynı dersi bir kereden fazla alan öğrencilerin dersi ilk aldıkları dönemki ders geçme notları yine aynı şekilde bir kereden fazla KPSS sınavına giren öğrencilerin ilk girişlerinde aldıkları puanlar kullanılmış diğer değerler çıkartılarak ayrı bir tablo oluşturulmuştur. Lisans eğitimi süresince başka okullardan yatay geçiş ile PAÜ'ye gelen öğrencilere ait kayıtlar veri kümesinden çıkartılmıştır. Veri toplama aşamasında edinilen her iki veri kümesinde de eksik nitelik barındıran kayıtlar ilk aşamada veri kümesinden çıkarılmıştır. Veri temizleme aşamasından sonra elde edilen kayıtlar SQL sorgularıyla tek bir tablo haline getirilmiştir.

PAÜ Eğitim Fakültesi'nde harfe dayalı not sistemi kullanılmaktadır, kullanılan harfler, 4'lük sistemindeki karşılıkları ve veri madenciliği uygulaması için oluşturulan modellerde kullanılabilmesi için 0-1 arasındaki sayılara dönüştürülmüş karşılıkları Tablo 5.3'te verilmiştir.

Tablo 5.3 Not sistemleri

HARF SİS.	DÖRTLÜK SİS.	KULLANILAN DEĞER
A1	4,0	1
A2	3,5	0,8
B1	3,0	0,6
B2	2,5	0,4
C	2,0	0,2
F3	0,0	0

Öğrencilerin genel not ortalamaları incelendiğinde, notların 2.20-3.67 arasında dağıldığı görülmüş ve bu puanlar yapay sinir ağı modelinde kullanılabilmesi için (5.1) numaralı denklemi kullanılarak normalize edilmiştir.

$$G' = \frac{G - G\min}{G\max - G\min} \tag{5.1}$$

G': Normalize edilmiş not ortalaması

G : Gerçek not ortalaması

$G\min$: Veri kümesindeki en küçük not ortalaması

$G\max$: Veri kümesindeki en büyük not ortalaması

Aynı teknik KPSS puanları içinde uygulanmış ve 50-90 aralığında olduğu görülen puanlar (5.2) numaralı denklemi kullanılarak normalize edilmiştir.

$$P' = \frac{P - P\min}{P\max - P\min} \tag{5.2}$$

P : Gerçek puan

P' : İzdüşürülmüş puan

$P\min$: Veri kümesindeki en düşük KPSS puanı

$P\max$: Veri kümesindeki en yüksek KPSS puanı

Veriler Microsoft Office Access ve Excel programları kullanılarak temizlenip yapay sinir ağları ve regresyon analizi için kullanılabilir biçime getirildikten sonra, herhangi bir eksik veri içermeyen 1031 kayıttan oluşan bir veri kümesi elde edildi. Elde edilen veri kümesindeki değişkenler, veri türleri ve değişkenlerin açıklamaları Tablo 5.4'de verilmiştir.

Tablo 5.4 Veri özellikleri

ALAN ETİKETİ	VERİ TÜRÜ	AÇIKLAMA
OGR	2 SEÇENEKLİ	ÖĞRETİM TÜRÜ
113101	KATEGORİK	TEMEL MATEMATİK I
113102	KATEGORİK	TEMEL MATEMATİK II
113107	KATEGORİK	COĞRAFYAYA GİRİŞ
113108	KATEGORİK	TÜRKİYE COĞRAFYASI VE JEOPOLİTİĞİ
113203	KATEGORİK	TÜRK DİLİ I SES VE ŞEKİL BİLGİSİ
113204	KATEGORİK	TÜRK DİLİ II CÜM.VE METİN BİLGİSİ
113205	KATEGORİK	ÜLKELER COĞRAFYASI
113207	KATEGORİK	CUM.DÖNEMİ TÜRK EDEBİYATI
113473	KATEGORİK	VATANDAŞLIK BİLGİSİ
127101	KATEGORİK	ÖĞRETMENLİK MESLEĞİNE GİRİŞ
127201	KATEGORİK	GELİŞİM VE ÖĞRENME
127202	KATEGORİK	ÖĞRETİMDE PLAN VE DEĞERLEN.
127302	KATEGORİK	SINIF YÖNETİMİ
127402	KATEGORİK	REHBERLİK
ORT	SAYI	NOT ORTALAMASI
KPSS	SAYI	KPSS PUANI

Oluşturulan veri kümesi 1999, 2000 ve 2001 yıllarında üniversiteye kayıt olan öğrencilerden oluştuğundan, bu öğrencilerin farklı yıllarda mezun olduğu göz önüne

alınarak veri kümesi yıllara göre bölünmüş üç veri kümesi daha elde edilmiş ve veri madenciliği uygulamasında kullanılmak üzere toplam dört veri kümesi oluşturulmuştur. Elde edilen veri kümelerinin içerdikleri kayıt sayıları Tablo 5.5'de verilmiştir.

Tablo 5.5 Veri kümeleri

Adı	İçerdiği Yıllar	Kayıt Sayısı
Veri Kümesi_1	Genel	1031
Veri Kümesi_2	1999	367
Veri Kümesi_3	2000	361
Veri Kümesi_4	2001	302

Bu işlemler yapıldıktan sonra veri madenciliği çalışması için kullanılacak programlara uygun dosya formatlarının oluşturulması aşamasına geçilmiştir.

JavaNNS programı bu çalışmada yapay sinir ağları tekniğinin ile oluşturulacak modellerden birini oluşturmak için kullanılmıştır. Bu program veri girdilerini örüntü dosyalarından almaktadır. Bu dosya formatını oluşturmak için Microsoft Office Access programından Microsoft Office Excel programına aktarılan veriler aktarılan veriler Excell programıyla açılmış ve sekmeyle ayrılmış metin formatında kaydedilmiştir. Excel sayıların ondalık kısımlarını ayırmak için virgül kullanırken JavaNNS nokta ile ayrılmış sayıları kabul ettiğinden bu değişiklik yapılmış ve bu dosyayı örüntü dosyasına çevirmek için gerekli başlık bilgisi dosyaya eklenerek örüntü dosyası elde edilmiştir. Örüntü dosyası için örnek başlık biçimi ve açıklaması aşağıda verilmiştir.

SNNS pattern definition file V4.2

generated at Thu Apr 13 02:15:03 2006 (Dosyanın yaratıldığı anı ve kullanılan JavaNNS versiyonunu belirtir)

No. of patterns : 901 (Toplam kayıt sayısı)

No. of input units : 16 (Tahmin edici değişken (bağımsız değişken) olarak kullanılacak değişken sayısı)

No. of output units : 1 (Tahmin edilecek değişken sayısı)

JavaNNS örüntü dosyası alan isimleri içermediğinden, tahmin edilmek istenen değerlerin yer aldığı alanlar, diğer alanlardan sonra girilir.

WEKA uygulamalarında veri girişi için ARFF dosya biçimi kullanılmaktadır. Bu dosya biçimini oluşturmak için excell biçimindeki dosya virgülle ayrılmış biçimde kaydedilmiştir. Bölgesel ayar farklılığından dolayı veri kümesindeki alanlar noktalı

virgülle, sayıların ondalık kısmı virgülle ayrıldığından, virgül noktayla, noktalı virgülde virgülle değiştirilmiştir. Elde edilen dosya MS DOS biçimli metin dosyası olarak kaydedilmiş ve başlık bilgisi eklenmiştir. Arff dosyası için gerekli olan başlık biçimi ve açıklamaları aşağıdaki gibidir.

@relation <dosya_adi>: Dosyanın başında yer alır, Veri kümesini içeren dosyanın adını belirtir.

@attribute <değişken> <tür>: Sütunda yer alan değişkenin ismini ve türünü içerir.

@data: başlık alanının bittiğini ve veri alanının başladığını gösterir.

5.4. Model Oluşturma

Bu çalışmada oluşturulan çok katmanlı perseptron modelleri JavaNNS ve WEKA 3.4.7 programları yardımıyla hazırlanmış, veri inceleme amaçlı regresyon analizi için SPSS 12.0 modellerin tahmin netliğini karşılaştırmak için oluşturulan regresyon analizi için WEKA 3.4.7 programı kullanılmıştır.

5.4.1. JavaNNS

Java sinir ağları simulasyonu (JavaNNS), Almanya Tübingen Üniversitesi Wilhelm-Shickard Bilgisayar Bilimleri Enstitüsü tarafından geliştirilmiş bir yapay sinir ağı simülatörüdür. JavaNNS, Stutgard sinir ağları simulatörü (SNNA) 4.2 kabuğu üzerine Java programlama dili ile yeni bir kullanıcı arayüzü yazılarak oluşturulmuştur. SNNS'in JavaNNS olarak yeniden yorumlanması esnasında üç boyutlu yapay sinir ağı gösterimi gibi pek kullanılmayan bazı özellikleri dışarıda bırakılırken log paneli gibi kullanışlı olacağı düşünülen yeni özellikler eklenmiştir.

Kullanıcı arayüzündeki değişiklerin yanı sıra SNNS, Unix tabanlı sistemler için tasarlanmışken, JavaNNS Java Runtime Environment kurulu olması koşuluyla, Windows, Linux, Mac OS ve Solaris işletim sistemleri üzerinde de çalışabilecek bir şekilde, platform-bağımsız olarak tasarlanmıştır.

JavaNNS kullanıcı arayüzü, kullanıcıya doğrudan yaratma, konfigüre etme ve görselleştirme imkanları sunar. JavaNNS kullanıcıya yapay sinir ağlarının çok çeşitli parametrelerini kurma şansı verir, dolayısıyla oldukça esnek ve kullanışlı bir simulatördür. JavaNNS tek başına veri madenciliği programında kullanılabilecek bir program değildir. CRISP-DM standart modeli kullanılan bu çalışmanın modelleme

aşamasında kullanılmıştır. Yapay sinir ağı yaratılması, öğrenme yöntemi seçilmesi vb. birçok yapay sinir ağı parametresi seçimi ve üretilmesi kullanıcıya bırakıldığından, analizcinin yapay sinir ağları konusunda bilgi sahibi olması gerekmektedir(Fischer 1996).

5.4.2. WEKA

WEKA Yeni Zelenda'daki University of Waikato tarafından java programlama dili kullanılarak yaratılmış, Linux, Mac OS ve Windows işletim sistemleri altında denenmiş içerisinde birçok makine öğrenmesi algoritması barındıran bir veri madenciliği programıdır. WEKA'nın asıl alanı birçok makine öğrenmesi algoritmasının uygulandığı sınıflandırma problemi olmasına rağmen bünyesinde çeşitli kümele, birliktelik kuralları ve regresyon analizi algoritmaları da barındırmaktadır.

5.4.3. SPSS

SPSS sosyal bilimlerde yapılan araştırma verilerinin istatistiksel yöntemlerle incelenmesi için oluşturulmuştur. SPSS sahip olduğu kullanıcı dostu grafiksel arayüzü sayesinde teknik bilgi gerektirmeksizin kullanıcıya faktör analizi, regresyon hesabı, varyans analizi, kümeleme analizi gibi istatistiksel teknikleri uygulama ve veriyle ilgili grafikler oluşturma imkanı verir.

5.5. Çok Katmanlı Perseptron Modelinin Oluşturulması

Çok katmanlı perseptron modeli oluşturulurken ilk olarak en uygun ağ yapısının belirlenebilmesi için rastgele seçilen 20 ağ yapısı, öğrenme katsayısı (ε) 0.3 momentum katsayısı (μ) 0.2 için denenmiş, başarılı bulunan ağ yapıları farklı öğrenme ve momentum katsayıları ve yineleme sayıları için denenerek en başarılı model seçilmiştir.

Gizli katman ve nöron sayısının belirlenmesi için geçen süreyi azaltmak için en iyi ağ modelinde yer alacak nöron sayısı denemeleri (5.3) numaralı denklem ile elde edilen 8 sayısından başlanarak yapılmıştır.

$$\text{Gizli Katman Nöron Sayısı}=(N_g + N_ç)/2 \tag{5.3}$$

N_g : Girdi sayısı

$N_ç$: Çıktı sayısı

Tablo 5.6 Çok katmanlı perseptron modelleri

Ağ Adı	1. Gizli Kat. Nöron Sayısı	2. Gizli Kat. Nöron Sayısı	3. Gizli Kat. Nöron Sayısı
YSA1	2	2	0
YSA2	4	0	0
YSA3	2	0	0
YSA4	4	2	0
YSA5	6	4	0
YSA6	6	0	0
YSA7	8	6	0
YSA8	10	4	2
YSA9	16	8	4
YSA10	8	0	0
YSA11	8	4	2
YSA12	40	20	10
YSA13	8	0	0
YSA14	12	6	3
YSA15	30	10	5
YSA16	20	8	0
YSA17	40	10	0
YSA18	10	0	0
YSA19	30	10	0
YSA20	10	0	0

Bu modellerin değerlendirilmesi sonucu en başarılı ağ yapısı, gizli katman sayısı bir, gizli katman nöron sayısı sekiz olan YSA 10 modeli seçilmiştir.

Oluşturulan ağ yapısı JavaNNS yardımı ile farklı geri yayılım öğrenme algoritmaları ile denenmiş momentumlu geri yayılım algoritmasının tüm veri kümelerinde daha başarılı olduğu tespit edilmiştir.

Öğrenme katsayısı seçiminde farklı denemeler yapılmış ve 0.1-0.7 aralığında seçilen tüm katsayılar için ağın başarısı yetersiz bulunmuş, WEKA 3.4.7 programında, her yineleme için öğrenme katsayısını yineleme sayısına bölerek oluşturulan yeni ve daha küçük bir öğrenme katsayısı kullanılmıştır. Öğrenme katsayısı bu şekilde seçildiğinde, momentum teriminin ağın ürettiği hata değerleri üzerindeki etkisi azalmıştır. 0.1-0.3

aralığındaki momentum terimleri için de denemeler yapılmış ve her veri kümesi için en başarılı öğrenme parametreleri kombinasyonu bulunmaya çalışılmıştır.

Oluşturulan modelin test edilebilmesi için veri kümesini beş parçaya bölen ve her seferinde farklı bir kümeyi test kümesi geriye kalan kümeleri eğitim kümesi olarak kullanan 5 kümeli çapraz doğrulama tekniği kullanılmıştır. Eğitim için yineleme sayısına farklı yineleme sayıları için hata oranları karşılaştırılarak karar verilmiştir.

6. BULGULAR VE YORUM

Bu bölümde ilk olarak elde edilen veri kümesinin özelliklerinin anlaşılabilmesi için SPSS 12.0 programı yardımıyla elde edilen frekans dağılımı tablosu ve regresyon analizi sonuçları incelenecek, daha sonra WEKA 3.4.7 ve JavaNNS yardımı ile oluşturulan regresyon modeli ve çeşitli yapay sinir ağı modellerinin öngörü netlikleri kıyaslanacaktır.

6.1. Veri Özelliklerinin İncelenmesi

Veri madenciliğinde verinin özelliklerinin anlaşılması modelin başarısı açısından çok önemlidir. Verinin iyi anlaşılması uygulanacak modelin ve model parametrelerinin doğru seçilmesini, dolayısıyla proje başarısını beraberinde getirir.

6.1.1. Frekans analizi

Yüzlük not sistemi ile yapılan sınavların dörtlük not sistemine dönüştürülmesi veri madenciliği açısından önemli oranda bilgi kaybına neden olmaktadır. Dörtlük not sisteminde de belirli notlarda yığılmalar olması veri kümesinin tahmin edici özelliğini zayıflatmaktadır.

Tablo 6.1 incelendiğinde birçok dersi A1 ile geçen öğrencilerin oranının %1 veya daha düşük olduğu görülmektedir. Veri kümesinin geneli için bu oran % 2.2'dir. A2 için ise bu oran %5.5'dir. Veri kümesinde yer alan derslerden C ile geçilme oranı ortalama olarak % 38,6'dır. Not dağılımdaki bu yığılma bazı derslerde daha yoğundur. Örneğin 113101, 113102 ve 127402 optik kodlu derslerden veri kümesinde yer alan öğrencilerin % 50'den fazlası aynı notla geçmiştir.

Tablo 6.1 Derslerde alınan notların frekans dağılımı

Ders Kodu	Sayı/%	F3	C	B2	B1	A2	A1
113101	Sayı	173	541	134	160	17	6
	%	16,8	52,5	13,0	15,5	1,6	0,6
113102	Sayı	325	556	66	67	11	6
	%	31,5	53,9	6,4	6,5	1,1	0,6
113107	Sayı	33	287	206	398	85	22
	%	3,2	27,8	20,0	38,6	8,2	2,1
113108	Sayı	168	440	168	211	31	13
	%	16,3	42,7	16,3	20,5	3,0	1,3
113203	Sayı	104	486	187	224	24	6
	%	10,1	47,1	18,1	21,7	2,3	0,6
113205	Sayı	100	333	206	307	65	20
	%	9,7	32,3	20,0	29,8	6,3	1,9
113207	Sayı	196	483	162	164	22	4
	%	19,0	46,8	15,7	15,9	2,1	0,4
113473	Sayı	39	361	204	358	57	12
	%	3,8	35,0	19,8	34,7	5,5	1,2
127101	Sayı	133	509	198	168	14	9
	%	12,9	49,4	19,2	16,3	1,4	0,9
127201	Sayı	312	286	130	222	60	21
	%	30,3	27,7	12,6	21,5	5,8	2,0
127202	Sayı	338	408	138	132	10	5
	%	32,8	39,6	13,4	12,8	1,0	0,5
127302	Sayı	110	277	122	284	105	133
	%	10,7	26,9	11,8	27,5	10,2	12,9
127402	Sayı		112	146	516	199	58
	%		10,9	14,2	50,0	19,3	5,6
113204	Sayı	99	488	188	206	41	9
	%	9,6	47,3	18,2	20,0	4,0	0,9
ORTALAMA	%	14,8	38,6	15,6	23,7	5,1	2,2

Bu çalışmada farklı yıllarda eğitim gören öğrencileri kapsayan veriler üzerinde çalışılmıştır, dersi veren öğretim görevlisinin notlandırma sisteminin değişmesi veya dersi veren öğretim görevlisinin değişmesi KPSS puanın hesaplanmasında tahmin edici değişken olarak kullanılan ders geçme notlarının dağılımını etkileyebileceği için yıllara göre not dağılım yüzdeleri veri incelemesine dahil edilmiştir.

Tablo 6.2 incelendiğinde bazı dersler için geçme notu yüzdelerinin yıllara göre önemli miktarda değiştiği görülmektedir. Örneğin üniversiteye 1999 yılında giren öğrencilerin 113101 kodlu dersten F3 notu alma yüzdeleri 10,8 iken 2001 girişli öğrenciler için bu oran 26,8'dir.1999 girişli öğrencilerin 113108 kodlu dersten F3 notu alma yüzdeleri 29,4 iken 2000 yılı girişli öğrenciler için bu oran 6,4'dür. Bu çalışmada öğrencilerin her dersi ilk kez alışlarındaki geçme notları kullanıldığından üniversiteye giriş yılları farklı olan öğrencilerin çok büyük bir kısmının belirli bir dersi alma yılları da farklıdır.

Pamukkale Üniversitesi ilköğretim Bölümü Sınıf Öğretmenliği A.B.D'nın her yıl ÖSYM tarafından yapılan ÖSS sınavının hemen hemen aynı yüzdelik dilimine giren adaylar arasından öğrenci aldığı düşünülürse, not dağılım oranlarındaki bu farklılığın öğretim elemanlarının uyguladıkları ölçme yöntemi farklılıklarından kaynaklandığı öne sürülebilir.

Aynı dersin not dağılımının yıllara göre önemli oranlarda değişmesi o dersin tahmin edici değişken olarak tutarlılığını azaltmakta ve oluşturulan veri madenciliği öngörü modellerinin tahmin netliğini olumsuz yönde etkilemektedir. Ders geçme notları için oluşturulmuş standart bir yaklaşım olmaması, bir dersi aynı yıl içinde birden fazla öğretim elemanın vermesi gibi durumlarda da öngörü modelinin başarısı düşecektir. Ders not dağılımlarındaki bu farklılıklar KPSS puanlarının ders notlarından yola çıkarak öngörülmesi ile ilgili çalışmalarda dersi veren öğretim elemanının da tahmin edici bir parametre olarak modele dahil edilmesinin modelin tahmin netliğini artıracağı söylenebilir.

Tablo 6.2 Yıllara göre notların frekans dağılımı

DERS KODU	YIL	F3(%)	C(%)	B2(%)	B1(%)	A2(%)	A1(%)
113101	1999	10,1	50,7	15,8	20,7	1,9	0,8
113101	2000	15,0	51,5	13,0	17,2	2,8	0,6
113101	2001	26,8	56,0	9,6	7,3	0,0	0,3
113102	1999	31,9	55,9	6,0	4,4	1,1	0,8
113102	2000	23,3	55,7	8,3	10,2	1,9	0,6
113102	2001	40,7	49,7	4,6	4,6	0,0	0,3
113107	1999	5,4	30,5	22,1	35,7	5,2	1,1
113107	2000	1,7	31,6	18,0	35,7	9,7	3,3
113107	2001	2,3	20,2	19,5	45,7	10,3	2,0
113108	1999	29,4	43,6	13,6	12,3	1,1	0,0
113108	2000	6,4	37,1	17,5	28,8	7,5	2,8
113108	2001	12,3	48,0	18,2	20,5	0,0	1,0
113203	1999	3,0	30,0	22,6	39,0	4,9	0,5
113203	2000	13,6	51,2	18,0	15,2	0,8	1,1
113203	2001	14,6	62,9	12,9	8,6	1,0	0,0
113205	1999	24,8	45,8	13,9	12,5	1,9	1,1
113205	2000	1,1	26,6	19,9	35,2	13,0	4,2
113205	2001	1,3	22,8	27,5	44,4	3,6	0,3
113207	1999	10,1	55,0	19,1	14,7	0,5	0,5
113207	2000	21,1	46,3	17,5	13,6	1,1	0,6
113207	2001	27,5	37,4	9,6	20,2	5,3	0,0
113473	1999	3,0	33,5	21,0	36,2	4,9	1,4
113473	2000	3,9	40,2	17,7	30,5	6,6	1,1
113473	2001	4,6	30,8	20,9	37,7	5,0	1,0
127101	1999	15,3	51,8	14,4	13,9	2,7	1,9
127101	2000	10,2	43,5	24,7	21,1	0,3	0,3
127101	2001	13,2	53,3	18,5	13,6	1,0	0,3
127201	1999	40,3	26,7	9,3	17,2	4,4	2,2
127201	2000	34,3	31,3	10,2	15,2	6,1	2,8
127201	2001	13,2	24,5	19,5	34,4	7,3	1,0
127202	1999	24,0	32,4	16,9	22,9	2,5	1,4
127202	2000	39,9	40,7	11,6	7,5	0,3	0,0
127202	2001	34,8	47,0	11,3	7,0	0,0	0,0
127302	1999	19,3	41,7	8,7	10,4	4,4	15,5
127302	2000	10,2	17,5	7,8	25,2	20,2	19,1
127302	2001	0,7	19,9	20,5	51,3	5,3	2,3
127402	1999	0,0	4,6	11,4	52,0	22,6	9,3
127402	2000	0,0	13,3	17,7	51,5	15,2	2,2
127402	2001	0,0	15,6	12,9	46,0	20,2	5,3
113204	1999	15,5	56,1	16,9	9,8	1,6	0,0
113204	2000	5,3	39,9	17,7	29,6	6,6	0,8
113204	2001	7,6	45,7	20,2	20,9	3,6	2,0

6.1.2. Regresyon analizi

Bu bölümde PAÜ Eğitim Fakültesi Sınıf Öğretmenliği A.B.D'ye 1999 yılında giren öğrencilerden oluşan veri kümesi 2, 2000 yılında giren öğrencilerden oluşan veri kümesi 3, 2001 yılında giren öğrencilerden oluşan veri kümesi 4 ve tüm öğrencileri kapsayan veri kümesi 1 için regresyon analizi yapılacak. Bu veri kümeleri için modellerin genel olarak KPSS puanlarını açıklama oranları tartışılacak tüm derslerin modeldeki katkıları veri kümesi 1 için yapılan regresyon analizi sonucunda tartışılacaktır.

Tablo 6.3 Regresyon modelleri

Uygulanan Veri Kümesi	R	R^2	Düzeltilmiş R^2	Std. Öngörü Hatası
1999	0,571	0,326	0,295	0,152613
2000	0,457	0,209	0,172	0,172484
2001	0,516	0,266	0,225	0,189237
Genel	0,469	0,220	0,208	0,174466

SPSS kullanılarak yapılan regresyon analizi sonucu 1999 yılında sınıf öğretmenliği ABD'ye kayıt olan öğrencilerinin aldıkları KPSS puanlarının % 32,6 sının, 2000 yılında kayıt olan öğrencilerden oluşan veri kümesinde %20,9'unun, 2001 yılında kayıt olan öğrencilerden oluşan veri kümesinde de %22'sinin lisans eğitimleri boyunca aldıkları ve KPSS'de soru çıkan 14 ders, genel not ortalamaları ve öğretim türleri tarafından açıklanabilmiştir.

Öğrencilerin ders başarı düzeylerinin KPSS de gösterdikleri başarının ancak bu kadar küçük bir yüzdesini açıklayabilmesi, eğitimciler için üzerinde düşünülmesi gereken bir durumdur.

Bu durumun nedenlerinin anlaşılabilmesi için,

- Lisans derslerinde uygulanan ölçme ve değerlendirme yöntemlerinin,
- KPSS için uygulanan ölçme ve değerlendirme yöntemlerinin,
- Lisans dersleri müfredatı ile KPSS'nin içeriğinin uyumluluğunun,

- Öğrencilerin lisans derslerine çalışma motivasyonlarının,

sorgulanması gerekir.

Veri kümesinde yer alan tahmin edici değişkenlerin modele katkılarını belirlemek için regresyon analizinden elde edilen katsayılar tablosu kullanılmıştır. Herhangi bir değişkenin modele anlamlı bir katkısının olup olmadığı F-testinin p değerine bakılarak tespit edilmektedir.

Tablo 6.4 incelendiğinde bütün değişkenler girilerek oluşturulan regresyon modeline anlamlı katkısı olan değişkenler OGR (Öğretim türü), 113101 (Temel Matematik 1),113473 (Vatandaşlık Bilgisi), 127202 (Öğretimde Planlama ve Değerlendirme), 127402 (Rehberlik) ve ORT (Genel not ortalaması), 113107 (Coğrafyaya Giriş) değişkenleridir. Regresyon analizinde değişkenlerin modele giriş sırası, o değişkenin modeldeki katsayısını dolayısıyla modele yaptığı katkıyı değiştirebilmektedir. Dolayısıyla p (sig) değeri 0.05'in biraz üzerinde olan değişkenlerde model incelemesi esnasında dikkate alınabilir.

Tablo 6.4 Veri kümesi_1 için regresyon analizi katsayılar tablosu

	Katsayılar		Standart Katsayılar		
Değişkenler	Std. Hata	Beta	Std. Hata	T	Sig.
(sbt)	0,257	0,024		10,783	0,000
OGR	0,112	0,014	0,273	8,149	0,000
113101	0,066	0,031	0,069	2,168	0,030
113102	-0,014	0,035	-0,013	-0,406	0,685
113107	0,046	0,027	0,052	1,697	0,090
113108	0,018	0,028	0,021	0,652	0,515
113203	-0,031	0,030	-0,032	-1,034	0,301
113205	0,035	0,028	0,043	1,262	0,207
113207	0,018	0,031	0,020	0,586	0,558
113473	0,199	0,030	0,220	6,707	0,000
127101	-0,030	0,030	-0,031	-0,992	0,322
127201	0,006	0,024	0,008	0,243	0,808
127202	-0,095	0,031	-0,103	-3,030	0,003
127302	0,002	0,019	0,003	0,092	0,927
127402	0,094	0,031	0,096	3,024	0,003
113204	-0,004	0,029	-0,005	-0,146	0,884
ORT	0,304	0,063	0,181	4,799	0,000

Bu değişkenler ile adım adım regresyon yöntemi kullanılarak oluşturulan model KPSS puanlarındaki değişimin % 21,6'sını açıklamaktadır. Bu değişkenler ile oluşturulan model için katsayılar Tablo 6.5'de verilmiştir.

Tablo 6.5 Modele anlamlı katkısı olan değişkenler için katsayılar tablosu

Değişkenler	Katsayılar		Standart Katsayılar	T	Sig. Std. Hata
	B	Std. Hata	B		
(Sbt)	0,261	0,022		11,918	0,000
113473	0,208	0,029	0,230	7,244	0,000
OGR	0,109	0,012	0,264	9,115	0,000
ORT	0,320	0,058	0,190	5,558	0,000
127202	-0,104	0,030	-0,113	-3,510	0,000
127402	0,086	0,030	0,087	2,811	0,005
113101	0,059	0,027	0,061	2,151	0,032
113107	0,054	0,026	0,061	2,070	0,039

Sadece Tablo 6.5'deki 7 değişkenle, 16 değişken kullanılarak oluşturulan modeldekine çok yakın bir performans elde edilmesi diğer değişkenlerin modele anlamlı bir katkısının olmaması, değişkenler arasındaki ilişkilere ve bu değişkenleri oluşturan dersler için kullanılan ölçme stratejisine bağlanabilir.

6.2. Veri Madenciliği Modellerinin Öngörü Netliğinin Karşılaştırılması

Bu bölümde yapıları daha önce anlatılan çok katmanlı yapay sinir ağı ve çoklu regresyon yöntemleri kullanılarak oluşturulan modellerin öngörü performansları, 5 kümeli çapraz doğrulama tekniği kullanılarak modellerin ürettikleri ortalama mutlak hata ve ortalama hata kareler kökü değerleri karşılaştırılarak bulunacaktır.

6.2.1. Veri kümesi I

6.2.1.1. Regresyon modeli

KPSS = 0.1076 * OGR +0.0572 * 113101 +0.0472 * 113107 +
0.0411 * 113205 +0.2008 * 113473 -0.1017 * 127202 +
0.087 * 00127402 +0.2932 * GPA +0.2577

Ortalama mutlak hata 0.1418

Ortalama hata kareler kökü 0.1763

6.2.1.2. YSA modeli

Sekiz nörondan oluşan tek gizli katmana sahip yapay sinir ağı modelinin

Öğrenme Katsayısı: 0.1

Momentum Katsayısı: 0.1

Yineleme Sayısı :100

parametreleri ile ürettiği hatalar:

Ortalama mutlak hata 0.1413

Ortalama hata kareler kökü 0.1759

6.2.2. Veri kümesi II

6.2.2.1. Regresyon modeli

KPSS = 0.0995 * OGR + 0.1377 * 113101 -0.1053 * 113102 + 0.063 * 113107 + 0.0844 * 113203 + 0.0941 * 113207 + 0.2527 * 113473 + 0.0558 * 127302 +0.0935 * 127402 + 0.1684

Ortalama Mutlak Hata 0.1257

Ortalama hata kareler kökü 0.1568

6.2.2.2. YSA modeli

Sekiz nörondan oluşan tek gizli katmana sahip yapay sinir ağı modelinin

Öğrenme Katsayısı: 0.5

Momentum Katsayısı: 0.2

Yineleme Sayısı :500

parametreleri ile ürettiği hatalar:

Ortalama mutlak hata 0.1231

Ortalama hata kareler kökü 0.1555

6.2.3. Veri kümesi III

6.2.3.1. Regresyon modeli

KPSS = 0.0893 * OGR + 0.0964 * 113102 -0.065 * 113108 + 0.2142 * 113473 - 0.1086 * 127101 -0.0932 * 00127302 + 0.1232 * 127402 + 0.3042 * GPA + 0.392

Ortalama mutlak hata 0.1431

Ortalama hata kareler kökü 0.1777

6.2.3.2. YSA modeli

Sekiz nörondan oluşan tek gizli katmana sahip yapay sinir ağı modelinin

Öğrenme Katsayısı: 0.2

Momentum Katsayısı: 0.2

Yineleme Sayısı :200

parametreleri ile ürettiği hatalar:

Ortalama mutlak hata 0.1421

Ortalama hata kareler kökü 0.1772

6.2.4. Veri kümesi IV

6.2.4.1. Regresyon modeli

KPSS = 0.1567 * OGR + 0.098 * 113107 + 0.2418 * 113473 + 0.0903 * 127201 +0.1347 * 127402 + 0.2118

Ortalama mutlak hata 0.1555

Ortalama hata kareler kökü 0.196

6.2.4.2. YSA modeli

Sekiz nörondan oluşan tek gizli katmana sahip yapay sinir ağı modelinin

Öğrenme Katsayısı: 0.2

Momentum Katsayısı: 0.2

Yineleme Sayısı : 30

parametreleri ile ürettiği hatalar:

Ortalama mutlak hata 0.1531

Ortalama hata kareler kökü 0.1932

Tablo 6.6 Hata terimleri

Model	Hata türü	Veri Kümesi_1	Veri Kümesi_2	Veri Kümesi_3	Veri Kümesi_4
Regresyon Analizi	Ort. Mut. Hata	0.1418	0.1257	0.1438	0.1555
	Ort. Hat. Kar. Kökü	0.1763	0.1568	0.1777	0.196
YSA	Ort. Mut. Hata	0.1413	0.1231	0.1421	0.1531
	Ort. Hat. Kar. Kökü	0.1759	0.1555	0.1772	0.1932

Bu çalışmada kullanılan veri kümeleri aynı tahmin edici ve bağımlı değişkenleri kullanmasına rağmen WEKA 3.4.7 tarafından oluşturulan çoklu regresyon yöntemi her veri kümesi için oluşturduğu regresyon denkleminde farklı katsayılar oluşturmuştur. Bu veri kümeleri arasındaki farklılıklardan ve tahmin edici değişkenlerin birbirleriyle ilişkili olmasından kaynaklanmaktadır.

Tablo 6.6 incelendiğinde kullanılan yapay sinir ağı modelinin farklı ağ parametreleri ile her veri kümesi için çoklu regresyon modelinden tahmin doğruluğu açısından daha başarılı sonuçlar ürettiği görülmektedir.

7. SONUÇ VE ÖNERİLER

Günümüzde öğrencilerin üniversite ve bölüm tercihlerini etkileyen en önemli etken, eğitim aldıkları bölümün mezuniyet sonrası iş bulmalarında sağlayacağı kolaylıktır. Özel üniversitelerin hızla yaygınlaşması, üniversite ve üniversite eğitimi almak isteyen öğrenci sayısının artması yüksek öğretim kurumları için rekabetçi bir ortam yaratmaktadır. Bu ortamda iddialı olmak isteyen üniversiteler karar alma süreçlerinde bilime hizmet etmenin yanısıra öğrencilerin gereksinimlerini karşılayacak, onları diğer üniversitelerden mezun olan meslektaşları arasında avantajlı duruma getirecek yenilikleri gerçekleştirmek durumundadır.

Mezunlarının önemli bir kısmı kamu sektöründe istihdam edilen sınıf öğretmenliği A.B.D. öğrencileri için KPSS oldukça önemlidir. Dolayısıyla mezun olan öğrencilerin KPSS'de başarılı olma oranları üniversiteler arasında ayırt edici bir ölçüt haline gelmiştir.

Bu çalışmada PAÜ Eğitim Fakültesi Sınıf Öğretmenliği A.B.D.'den mezun olan öğrencilerin KPSS'de aldıkları puanlar, lisans eğitimleri süresince aldıkları ve KPSS'de soru çıkan derslerden geçme notları, öğretim türleri ve genel not ortalamaları tahmin edici değişkenler olarak kullanılarak tahmin edilmeye çalışılmış, bu amaçla oluşturulan yapay sinir ağı ve regresyon analizi modellerinin tahmin doğrulukları karşılaştırılmıştır.

Yapılan araştırma sonucunda KPSS puanlarındaki değişimin küçük bir kısmının (%22) veri kümesinde yer alan değişkenler tarafından açıklanabildiği sonucuna ulaşılmıştır. Yapılan frekans analizinde bazı derslerden A1, A2 gibi notlarla gecen öğrenci oranları % 1,2 civarında iken C, B gibi notlarla geçen öğrencilerin oranları % 50 civarında olduğu görülmüştür. Not dağılımının bu yapısı ölçmede bilgi kaybına neden olmakta ve modelin başarısını olumsuz yönde etkilemektedir.

Öğrencilerin önemli bir kısmının C, B2 gibi ortalama notlar ile dersten geçerken çok az bir kısmının A1, A2 gibi yüksek notlarla derslerinden geçmesi kullanılan ölçme sistemine ve öğrencilerin derslerden yüksek not almalarının kendilerine herhangi bir fayda sağlamamasına dolayısıyla daha iyi bir not ile dersten geçmek için

çabalamamalarına bağlanabilir. Bu durum modelin KPSS puanını açıklamaktaki başarısını da düşürmektedir.

Ders notlarının, öğretim türünün ve genel not ortalamalarının KPSS puanlarındaki değişimin bu kadar küçük bir kısmını açıklamasının nedenlerinin ortaya koyulabilmesi için lisans dersleri ve KPSS içerikleri arasındaki uyumun ve derslerde kullanılan ölçme yöntemlerinin sorgulanması gerekir.

Bu çalışmada kullanılan KPSS puanlarının tahmin edilmesi için yapay sinir ağı ve regresyon analizi modelleri dört farklı veri kümesi üzerinde uygulanmış ürettikleri hatalar karşılaştırılmıştır. Kullanılan yapay sinir ağı modeli tarafından yapılan tahminler ve gerçek sonuç arasındaki fark tüm veri kümeleri için regresyon analizi yönteminin bulduklarına göre daha küçük çıkmıştır.

Modellerin ürettikleri hata değerleri farkının çok yüksek olamasa da yapay sinir ağı modelinin çalışmada kullanılan bütün veri kümelerinde regresyon analizi modelinden daha başarılı olması, yapay sinir ağı tekniğinin öngörüye dayalı eğitim araştırmaları için klasik istatistik yöntemlere bir alternatif oluşturabileceğini göstermiştir.

Karar alma süreçlerine hipoteze dayalı klasik istatistik yöntemlerinden daha etkin katkı sağlama iddiasıyla gelişen ve tıp, sanayi, ticaret gibi alanlarda kullanımı hızla yaygınlaşan veri madenciliği yaklaşımı etkin olarak kullanıldığında eğitim yöneticilerine de alanlarında avantajlı duruma getirecek bilgi sağlamaya adaydır.

Bu alanda yapılacak çalışmalarda KPSS sonuçlarının tahmin edilmesi için uygulanacak veri madenciliği sürecine genetik algoritmalar ve doğrusal olmayan regresyon yöntemleri dahil edilerek tahmin doğrulukları artırılabilir. Ayrıca yapılacak benzer araştırmalarda öğrencilerin lise genel başarı not ortalamalarının ve dersi ilk alışlarında kalan öğrencilerin daha sonra aldıklarındaki geçme notlarının tahmin edici değişken olarak kullanılması, farklı üniversitelerden toplanacak verilerin çalışmaya dahil edilmesi benzer araştırmalar için önerilebilir.

KAYNAKLAR

Akpınar, H. (2006) Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=538 (24.03.2006)

Alkan, A. (2001) Predictive Data Mining With Neural Networks and Genetic Algorithms, Ph. D. Thesis, Institute of Science and Technology, Computer Engineering, ITU, İstanbul 51s

Auclair, A. (2004) Feed-Forward Neural Networks Applied to the Estimation of Magnetic Distributions, M.S. Thesis, Department of Electrical and Computer Engineering, McGill University, Montreal, Canada,128s

Aydın, Ö. (2005) Yapay Sinir Ağlarını Kullanarak Bir Ses Tanıma Sistemi Geliştirilmesi, Yüksek Lisans Tezi, Bilgisayar Mühendisliği ABD, Trakya Üniversitesi, Edirne, 74s

Baum, E. B. and Hausler, D. (1989) What Size Net Gives Valid Generalization, Neural Computation, 1: 151-160

Beitel, S. E. (2005) Applying Artificial İntelligence Data Mining Tools to the Challenges of Program Evaluation, Ph. D. Thesis, University of Connecticut, Connecticut, 156s

Berry, M. and Linoff G. (2000) Mastering Data Mining, Wiley, Hoboken, NJ, 340s

Berry, M. and Linoff G. (1997) Data Mining Techniques for Marketing, Sales and Customer Support, Wiley, Hoboken, NJ, 322s.

Bidgoli, B. M. (2004) Data Mining For a Web Based Educational System, Ph. D. Thesis, Department of Computer Science and Engineering, Michigan State University, 208s

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. and Zanasi, A. (1998) Discovering Data Mining: From Concept to Implementation, Prentice Hall, Upper Saddle River, NJ, 517s

Chapman, P. Clinton, J. Kerber, R. Khabaza and T. Reinartz, T. (2000) Step by Step Data Mining Guide, CRISP-DM, www.crispdm.org, s1-78

Collard, M. and Francisci, D. (2001) Evolutionary Data Mining: An Overview of Genetic-Based Algorithms, IEEE, 0-7803-7241-7/01: 3-8

Erdem, D., Kaan, M. ve Tanhan F. (2006) Regresyon Analizi, http://www.istatistik.gen.tr/erdem.htm (14.02.2006)

Ethridge, D. and Zhu, R. (1996) Prediction of Rotor Spun Cotton Yarn Quality: A Comparison of Neural Network and Regression Algorithms, Beltwide Cotton Conference, National Cotton Council, Memphis TN, s. 1314-1320

Fayyad, U. Patiesky, G. and Smyth, P.(1996) From Data Mining to Knowledge Discovery in Databases, www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf (02.05.2006)

Feelders, A., Daniels, H. and Holsheimer, M., (2000) Methodological and Practical Aspects of Data Mining, Information & Management 37: 271-281

Fischer, I., Hennecke, F., Bannes, C., Zell, A., (1996) JavaNNS User Manual, University of Tubingen, Tubingen, 33s

Güvenç, E. (2001) Yüksek Öğrenimde Öğrenci Performansının Veri Madenciliği Teknikleri ile Belirlenmesi, Yüksek Lisans Tezi, Endüstri Mühendisliği ABD, Fen Bilimleri Enstitüsü, Boğaziçi Üniversitesi, İstanbul, 120s

Han, J. and Kamber, M. (2000) Data Mining: Concepts and Techniques, CA: Morgan Kaufmann, San Francisco, 312s

Hand, D.J. (1998) Data Mining Statistics and More, SIGKDD Explorations 1,1: 16-19.

Haykin, S. (1998) Neural Networks for Pattern Recognition, Prentice Hall, Upper Saddle River, NJ, 842s

Iskandar, N. F. (2005) An Artificial Neural Network Approach For Short Term Modeling of Stock Price Index, Ph. D. Thesis, Industrial Systems Engineering, University of Regina, Saskatchewan, 85s

Kantradzic, M. (2003) Data Mining-Concepts, Models, Methods and Algorithms, John Wiley & Sons, Inc., Hoboken, 360s

Kim, J. Y. (2003) ANN Wawe Prediction Model For Winter Storms and Hurricanes, Ph. D. Thesis, The School of Marine Science, The College of William and Marry, Virginia, 246s

Kuligowski, R. J. and Barros, A. P.(1998) Experiments in Short Term Precipitatiun Forecasting Using Artificial Neural Networks, Monthly Weather Rev, 176: 470-482

Larose, D.T . (2005) Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley & Sons, Inc., Hoboken, 215s

Mohamed, N. M. S. (2004) High-Speed Network Traffic Prediction and its Applications Using Neural Networks and Self-Similar Models, Ph. D. Thesis, Graduate Faculty of the School of Engineering, Southern Methodist University, 160s.

Nabiyev, V.V. (2003) Yapay Zeka, Seçkin Yayıncılık San. Ve Tic. A.Ş., Ankara, 724s.

Pryke, A. N. (1998) Data Mining Using Genetic Algorithms and Interactive Visualization, Ph. D. Thesis, Faculty of Science, University of Birmingham, Birmingham, 187s

Ramaswamy, S. (1997) Decision Support System Using Neural Networks, M.S. Thesis, Computer Science, University of Nevada, Nevada, 79s

Schumann, J. A. (2005) Data Mining Methodologies in Educational Organizations, Ph. D. Thesis, University of Connecticut, Connecticut, 196s

Slaughter, G. E. F. (2003) Artificial Neural Netwok For Temporal Impedance Recognition of Neurotoxins, M.S. Thesis, The School of Engineering, Virginia Common Wealth University, Virginia, 115s

Sörensen, K. and Janssens G. K. (2003) Data Mining With Genetic Algorithms on Binary Trees, European Journal of Operational Research, 151: 253-264

Swingler, K. (1996) Applying Neural Networks: A Practical Guide, CA: Academic California, 303s

Tabachnick, B.G. and Fıdell, L. S.(2001) Using Multivariate Statistics. (4th ed.). Needham Heights, MA: Allyn & Bacon, 240s

Velickov, S. and Solomatine, D. (2000) Predictive Data Mining: Practical Examples, Artificial Intelligence in Civil Engineering, Germany, s. 1-17

Wang, W. (1999) Classification and Pattern Matching Methods, M.S. Thesis,Com Beijing Polytechnic University, 109s

WEB_1.(2006) The Gartner Grup, www.gartner.com (25.12.2005)

WEB_2.(2003) http://www.kdnuggets.com/websites/standards.html (03.03.2006)

WEB_3. (1999) Two Crows Corporation, Introduction to Data Mining And Knowledge Discovery, www.twocrows.com (10.11.2005)

WEB_4. (2006) Yesevi Net, Ders Notları, www.yesevi.net/muh/ysa.doc (12.02.2006)

WEB_5. (2006) Bilgi Üniversitesi, knuth.cs.bilgi.edu.tr/~robotik/dokumanlar/misc/ yapay_sinir_aglari.doc (15.03.2006)

Xu, Y. (2003) Using Data Mining In Educational Research: A Comparison of Bayesian Network With Multiple Regression in Prediction, Department of Educational Psychology, The University of Arizona, Arizona, 242s

Yurtoğlu, H. (2005) Yapay Sinir Ağları Modellemesi ile Öngörü Modellemesi: Bazı Makroekonomik Değişkenler için Türkiye Örneği, Uzmanlık Tezi, DPT, Yayın no:2683, Ankara, 192s

ÖZGEÇMİŞ

Hüseyin Özçınar 1980 yılında Manisa ilinin Sarıgöl ilçesinde doğdu. İlk ve orta okulu Dadağlı Köyü İlköğretim Okulu'nda tamamladı. Lise eğitimini 1997 yılında Aydın Lisesi'nde tamamladı. 1998 yılında girdiği İstanbul Teknik Üniversite Elektronik ve Haberleşme Mühendisliği Bölümü'nü 2003 yılında bitirdi. 2003 yılında Pamukkale Üniversitesi Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü'nde başladığı Araştırma Görevliliğine halen devam etmektedir**.**