T.C.
Fırat Üniversitesi
Eğitim Bilimleri Enstitüsü
Bilgisayar ve Öğretim Teknolojileri Eğitimi

ÖĞRENCİLERİN AKADEMİK BAŞARILARININ VERİ MADENCİLİĞİ METOTLARI ile TAHMİNİ

Yüksek Lisans Tezi

Dönüş ŞENGÜR

Danışman: Yrd. Doç. Dr. Ahmet TEKİN

Elazığ-2013

T.C.
Fırat Üniversitesi
Eğitim Bilimleri Enstitüsü
Bilgisayar ve Öğretim Teknolojileri Eğitimi

ÖĞRENCİLERİN AKADEMİK BAŞARILARININ VERİ MADENCİLİĞİ METOTLARI ile TAHMİNİ

Yüksek Lisans Tezi

Danışman **Hazırlayan**
Yrd. Doç. Dr. Ahmet TEKİN **Dönüş ŞENGÜR**

Dönüş ŞENGÜR'ün hazırlamış olduğu **"**Öğrencilerin Akademik Başarılarının Veri Madenciliği Metotları ile Tahmini" başlıklı tez, Eğitim Bilimleri Enstitüsü Yönetim Kurulunun…….tarih ve ……sayılı kararı ile oluşturulan jüri tarafından…..…… tarihinde yapılan tez savunma sınavı sonunda yüksek lisans/doktora tezini oy birliği/oy çokluğu ile başarılı saymıştır.

Jüri Üyeleri:

1. Yrd. Doç. Dr. Murat KARABATAK (Jüri Başkanı)
2. Yrd. Doç. Dr. Ferhat BAHÇECİ
3. Yrd. Doç. Dr. Ahmet TEKİN (Danışman)

Fırat Üniversitesi Eğitim Bilimleri Enstitüsü Yönetim Kurulunun ….. tarih ve …….sayılı kararıyla bu tezin kabulü onaylanmıştır.

Doç. Dr. Mukadder BOYDAK ÖZAN
Eğitim Bilimleri Enstitüsü Müdürü

BEYANNAME

Fırat Üniversitesi Eğitim Bilimleri Enstitüsü tez yazım kılavuzuna göre, Yrd. Doç. Dr. Ahmet TEKİN danışmanlığında hazırlamış olduğum "Öğrencilerin Akademik Başarılarının Veri Madenciliği Metotları ile Tahmini" adlı yüksek lisans tezimin bilimsel etik değerlere ve kurallara uygun, özgün bir çalışma olduğunu, aksinin tespit edilmesi halinde her türlü yasal yaptırımı kabul edeceğimi beyan ederim.

Dönüş ŞENGÜR

TEŞEKKÜR

Bu tez çalışmam boyunca, ilgi ve yardımlarını esirgemeyen danışmanım Sayın Yrd. Doç. Dr. Ahmet TEKİN'e, çalışmam boyunca beni sabır ve özveri ile destekleyen eşime, yardımlarından ötürü Sayın Yrd. Doç Dr. Murat KARABATAK'a teşekkürlerimi ve şükranlarımı sunarım.

ÖZET

Yüksek Lisans Tezi

Öğrencilerin Akademik Başarılarının Veri Madenciliği Metotları ile Tahmini

Dönüş ŞENGÜR

Fırat Üniversitesi
Eğitim Bilimleri Enstitüsü
Bilgisayar ve Öğretim Teknolojileri Eğitimi
Elazığ-2013; Sayfa: XI+47

Kaliteli bir eğitim için Yükseköğretim kurumları yönetimsel ve eğitimsel anlamda doğru kararlar verebilmelidir. Yanlış veya eksik yapılan akademik planlama, başarısız olabilecek öğrenciler, mezun öğrencilerin yol haritaları, okuldan ayrılabilecek öğrenciler gibi konular Yükseköğretim kurumlarının problemlerindendir. Bu problemlerin çözülmesi ve tedbirlerin alınması eğitimin kalitesi için son derece önemlidir. Yükseköğretim kurumlarında eğitime ait giderek artan veriler bulunmaktadır. Giderek artan bu verilerin yönetime, eğitimcilere veya eğitime hiçbir yararı yoktur. Bahsedilen problemler hakkında yüksek oranlardaki doğruluklarla tahminler yapılabilmekte ve anlamlı sonuçlar, veri madenciliği yöntemleri ile ortaya çıkarılabilmektedir. Veri madenciliği yöntemleri akademik müdahaleler için güçlü bir araçtır. Bu tez de, veri madenciliği yöntemlerinden olan Yapay Sinir Ağları (YSA) ve Karar Ağaçları (KA) kullanılarak Fırat Üniversitesi, Eğitim Fakültesi, Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü (BÖTE) öğrencilerinin mezuniyet notlarının tahmin edilmesi gerçekleştirilmiştir. Bu bağlamda 2011 yılında mezun olmuş 127 BÖTE öğrencisinin 4 yıl süresince almış olduğu toplam 49 kültür ve mesleki dersin yılsonu notları kullanılmıştır. Mezuniyet notunun tahmini için iki farklı senaryo denenmiştir. İlk senaryoda, öğrencilerin sadece birinci ve ikinci sınıfa ait derslerinin yılsonu notları kullanılarak mezuniyet notu tahmin edilmiştir. İkinci senaryo da ise ilk üç sınıf notları ile mezuniyet notlarının tahmini gerçekleştirilmiştir. Gerçekleştirilen

benzetim çalışmalarında YSA'nın, KA'ya oranla daha iyi tahmin başarımı sağladığı ve ikinci senaryonun, birinci senaryoya oranla daha iyi tahminler yaptığı görülmüştür.

Anahtar Kelimeler: Eğitsel Veri Madenciliği, Öğrencilerin Başarılarının Tahmini, Yapay Sinir Ağları, Karar Ağaçları

ABSTRACT

Master Thesis

Prediction of Student's Academic Achievements by Using the Data Mining Methods

Dönüş ŞENGÜR

Firat University,
Institute of Education Sciences
Computer Education and Instructional Technology
Elazig-2013; Page: XI+47

For an eligible education, higher education institutions are to conclude the right decision by means of administrative and educational aspect. For academic planning being missing or incorrect, students being unsuccessful, determining the roadmap for graduates, student's dropping out are the subject of higher education institution's problem. Solving these problems and taking measures are very important for eligible education. There is a bulk of increasing data belonging to education of higher education institutions. These increasing data has nothing to do with administration, academician and education. These data can be meaningful by processing with the methods of data mining, estimations can be done with higher accuracy rate and measures can be taken. Data mining methods are powerful tools for academic interventions. In this thesis, several prediction techniques in data mining such as artificial neural networks and Decision tree methods are used to help the educational institutions to predict the students' graduation scores. In this context, 127 unique student records of Computer Education and Instructional Technology department are used, where the students were in the university between 2006-2010 and 2007-2011. Two different scenarios are investigated. Firstly, we used the students' first two years scores. Secondly, students' first three years scores are used for prediction. According to the experimental results, it

is observed that the ANN yielded better performance than the Decision tree and the second scenario yielded better results than the first scenario.

Key Words: Educational Data Mining, Prediction of Student's Academic Achievements, Artificial Neural Networks, Decision Trees.

İÇİNDEKİLER

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

KISALTMALAR LİSTESİ

VM: Veri Madenciliği

VTBK: Veri Tabanında Bilgi Keşfi

BÖTE: Bilgisayar ve Öğretim Teknolojileri

YSA: Yapay Sinir Ağları

KA: Karar Ağaçları

ÇAİ: Çevrimiçi Analitik İşleme

SQL: Structured Query Language (Yapısal Sorgulama Dili)

GA: Genetik Algoritmalar

K-NN: K-En Yakın Komşu

BTN: Bellek Temelli Nedenleme

LR: Lojistik Regresyon

BİRİNCİ BÖLÜM

I. GİRİŞ

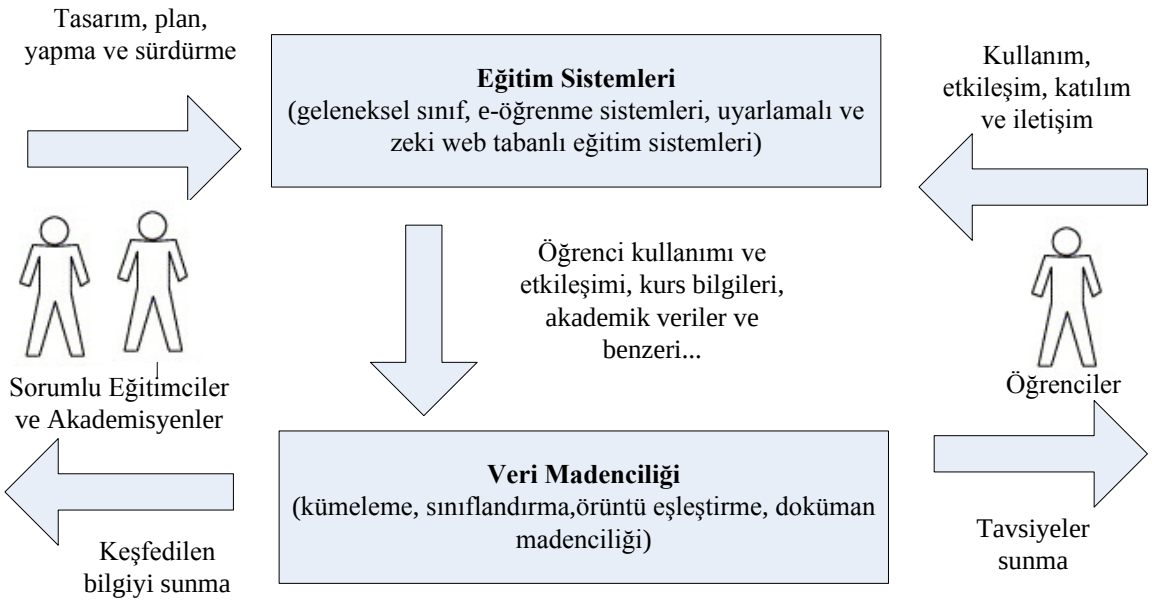
Veri Madenciliği (VM) ve Veri Tabanında Bilgi Keşfi (VTBK) büyük veri kümelerinden kesin ve ilginç verilerin otomatik olarak çıkarımıdır (Agrawal, Imielinski ve Swami, 1993; Agrawal ve Srikant, 1994). VTBK yalnızca öğrenci modelleme veya öğrenme sürecinde model öğrenme için değil aynı zamanda e-öğrenme sistemlerini geliştirmek ve değerlendirmek için de kullanılır (Ding, 2001; Han ve Kamber, 2006). Diğer bir ifade ile VM, eldeki verilerden üstü kapalı, çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilginin çıkarılmasıdır. Bu da; kümeleme, veri özetleme, değişikliklerin analizi, sapmaların tespiti gibi belirli sayıda teknik yaklaşımları içermektedir. Kısaca veri madenciliği; geniş veri tabanlarındaki veriler arasından bilgi çıkarma işlemidir.

Öğrenmeyi geliştirmek için eğitim sistemlerine bilgi çıkarımı tekniklerinin uygulanması biçimlendirici bir değerlendirme tekniği olarak görülebilir. Biçimlendirici değerlendirme hâlen gelişmekte olan eğitim programının değerlendirmesidir ve sürekli olarak uygulanan programı iyileştirme amaçlıdır. Öğrencilerin sistemi nasıl kullandığının sınanması biçimlendirici bir şekilde öğretim tasarımını değerlendirmek için kullanılan bir yoldur ve bu yol, öğretim materyalleri geliştirmek için eğitimcilere yardımcı olabilir. Veri madenciliği teknikleri, eğiticinin, öğretme ortamını tasarlarken veya geliştirirken oluşturacağı kararlar için pedagojik bir temel oluşturmasına yardımcı, faydalı bilgileri sunabilir.

Yükseköğretim kurumlarındaki öğrenciler, dersler, akademik ve idari personel, yönetim sistemleri vb. veriler stratejik verilerdir. Stratejik verilerin çözümlenerek anlamlı bilgilerin ortaya çıkarılması, yükseköğretim kurumlarının birtakım tedbirler alarak, eğitimdeki kaliteyi artırmasını sağlayacaktır. Veriyi çözümlemek ve anlamlı bilgileri ortaya çıkarmada istatistiki yöntemler her zaman kullanışlı olmayabilmektedir. Bu durumlarda verileri işlemek ve çözümlemek için veri madenciliği yöntemleri kullanılmaktadır. Yükseköğretim kurumları, öğrenci ve mezunların yol haritalarını tahmin etmeye odaklanmıştır (Luan, 2002). Kurumlar hangi öğrencilerin özel ders programlarına kayıt yaptıracağı, hangi öğrencilerin mezun olabilmesi için akademik yönlendirmeye ihtiyacı olacağı veya hangi öğrencilerin okuldan ayrılabileceği gibi

soruların cevaplarını bilmek isterler. Bu gibi soruların cevapları kurumlar için önemlidir ve bu soruların cevapları veri madenciliği yöntemleri ile bulunabilmektedir.

Eğitim sistemlerinde veri madenciliği uygulaması, hipotez oluşumu, test ve arıtma döngüsü tekrarıdır. Elde edilen bilgi, sistem döngüsüne ve rehbere kaydedilmelidir, bir bütün olarak öğrenmeyi kolaylaştırmalı ve artırmalıdır. Veri sadece bilgiye dönüştürülmez, aynı zamanda karar mekanizması için filtre (temizlenir ayıklanır) edilir.



Şekil 1. Eğitimde veri madenciliği uygulama döngüsü (Romero ve Ventura, 2007)

Şekil 1'de görüldüğü gibi, eğitimciler ve akademisyenler eğitim sistemlerini korumak, kurmak, planlamak ve tasarlamakla sorumludur. Öğrenciler onlarla etkileşim halindedirler. Kurslar, öğrenciler, kullanım ve etkileşim hakkındaki mevcut tüm bilgilerden başlayarak, farklı VM teknikleri, e-öğrenme sürecini geliştirmeye yardım eden kullanışlı bilgiyi keşfetmek için uygulanabilir. Keşfedilen bilgi yalnızca eğitimciler tarafından değil aynı zamanda öğrenciler tarafından da kullanılabilir. Bu yüzden eğitsel VM uygulamaları, farklı bakış açısıyla farklı katılımcılara uyarlanabilir (Romero ve Ventura, 2007).

1.1. Problem Durumu

Öğrencilerin akademik başarılarını geliştirmeye yardımcı, çeşitli gizli bilgileri barındıran eğitsel veri tabanlarında büyük miktarlarda veriler birikmektedir. Veri tabanlarında depolanan bu verilerden, gelecekte kullanılabilecek gizli ve faydalı bilgiler ortaya çıkarılabilir. Bir kursa veya derse kayıt yaptıracak öğrencilerin profillerinin tahmininde, geleneksel sınıfta öğretme modelinin aksaklıklarının belirlenmesinde, çevrimiçi sınavlardaki olumsuzlukların tespitinde, öğrenci sınav kâğıtlarındaki anormal değerlerin tespitinde, öğrenci performanslarının tahminde, vb. gizli bilgiler ortaya çıkarılarak eğitsel amaca yönelik kullanılabilir.

Öğrencilerin okullardaki akademik başarısı gerek öğrenci gerekse eğitmen ve aileler tarafından önemsenmektedir. Her aile çocuğunun başarılı olmasını, her öğrenci kendisinin başarılı olmasını ister. Veri madenciliği yöntemleri ile öğrencinin akademik başarı durumu tahmin edilebilir. Başarısız öğrencilere, başarıyı artırmak için yönlendirmeler yapılabilir, ders çalışma programları oluşturulabilir, destek ders kaynakları tavsiye edilebilir veya seçmeli dersler seçtirilebilir. Bu doğrultuda "Öğrencilerin Akademik Başarılarının Veri Madenciliği Metotları ile Tahmini" araştırma konusu olarak seçilmiştir.

1.2. Araştırmanın Amacı

Bu tezin temel amacı, veri madenciliği yöntemlerinden olan YSA ve karar ağaçları kullanılarak Fırat Üniversitesi, Eğitim Fakültesi, Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü (BÖTE) öğrencilerinin mezuniyet notlarının tahmin edilmesi gerçekleştirilmiştir. Bu bağlamda 2011 yılında mezun olmuş 127 BÖTE öğrencisinin 4 yıl süresince almış olduğu toplam 49 kültür ve mesleki dersin yılsonu notları kullanılmıştır. Bu notlar Fırat Üniversitesi, Öğrenci İşleri veri tabanından gerekli izinler alındıktan sonra elde edilmiştir. Mezuniyet notunun tahmini için iki farklı senaryo denenmiştir. İlk senaryoda, öğrencilerin sadece birinci ve ikinci sınıfa ait derslerinin yılsonu notları ile mezuniyet notu tahmin edilmiştir. İkinci senaryo da ise ilk üç sınıf notları ile mezuniyet notlarının tahmini gerçekleştirilmiştir.

1.3. Araştırmanın Sınırlılıkları

Bu tez çalışması;

1- 2007-2011 öğretim yılı ile sınırlıdır.

2- Çalışma sadece BÖTE öğrencilerini kapsamaktadır.

3- Tezde VM yöntemlerinden olan YSA ve KA yöntemleri kullanılmıştır.

4- İlgili bütün yazılımlar MATLAB ortamı ile sınırlıdır.

1.4. Tezin Organizasyonu

Tezin birinci bölümünde, teze genel bir bakış açısı kazandırmaya yönelik olarak temel bilgiler, problem durumu, tezin amacı ve tezin sınırlılıkları verilmiştir. Diğer bölümlerin organizasyonu aşağıda sunulmuştur:

Bölüm 2 de, veri madenciliği kavramı tanımlanarak, sahip olduğu her bir bileşen açıklanmıştır. Tez ile ilgili literatür incelenmiştir.

Bölüm 3 de, YSA ve KA kullanılarak, öğrencilerin mezuniyet notları tahmin edilmiştir.

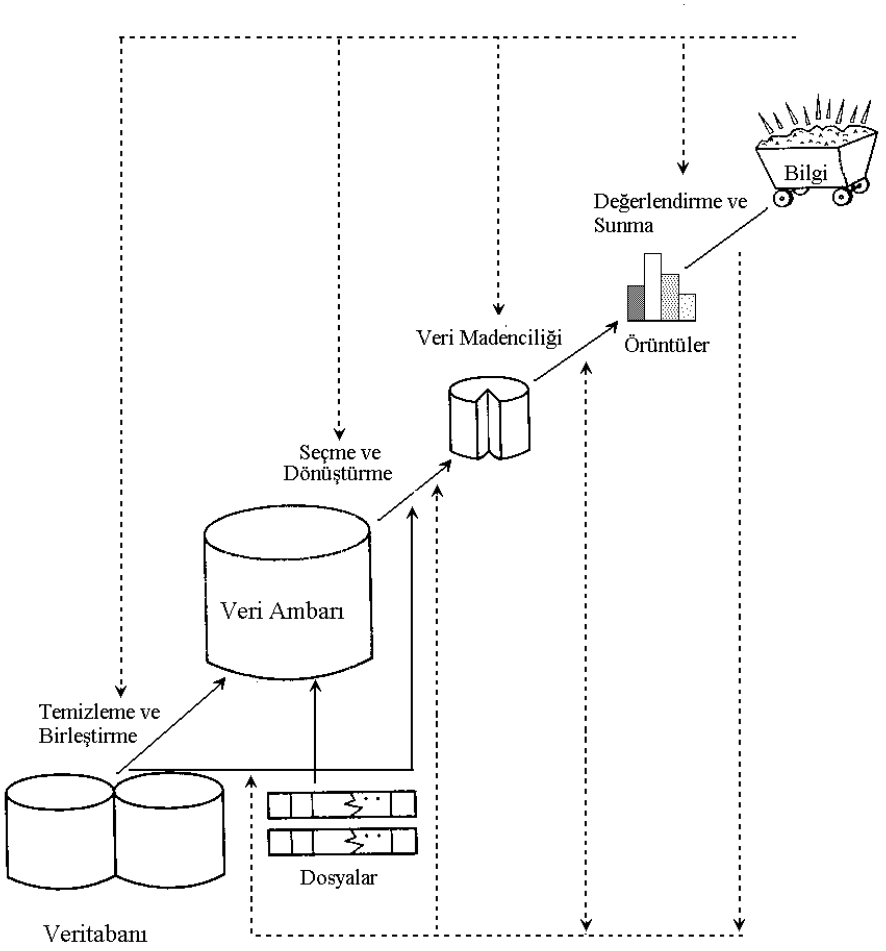
Bölüm 4 de, tezin sonuçları irdelenmiştir. Ayrıca ileriye dönük uygulama alanları ve öneriler tartışılmıştır.

İKİNCİ BÖLÜM

II. VERİ MADENCİLİĞİ

Veri madenciliği (VM), bir veri kümesi içerisinden ileriye yönelik tahmin yapmayı sağlayan bağıntı ve kuralların çıkarılmasıdır. VM, öbekleme, veri özetleme, değişikliklerin analizi, sapmaların tespiti gibi belirli sayıda teknik yaklaşımları içermektedir (Agrawal ve diğerleri, 1993; Agrawal ve Srikant., 1994).

Diğer bir ifade ile VM, önemli olan bilginin geniş veri tabanlarındaki veriler arasından işlenerek çıkarılmasıdır. Yer altındaki madenlerin araştırılması, çıkarılması ve işletilmesine madencilik deniyorsa, verilerden de bilgi keşfetme işine veri madenciliği denilmektedir. Veri madenciliği ile aynı anlamı taşıyan, veri tabanlarından bilgi keşfi, bilgi çıkarma, veri analizi, veri tarama gibi terimler de yaygın olarak kullanılmaktadır.



Şekil 2. Veri tabanında bilginin keşfi, (Karabatak, 2008)

Literatürde en çok VTBK terimi, VM ile eş anlamlı olarak kullanmaktadır. VTBK sürecinin nasıl gerçekleştiği Şekil 2'de adım adım gösterilmiştir.

- **Veri Temizleme:** Keşfedilmesi istenen bilginin kalitesini arttırmak için tutarsız ve gürültülü verilerin veri tabanından çıkarılma aşamasıdır.
- **Veri Birleştirme:** Veri tabanında bulunan aynı veya benzer verileri birleştirme aşamasıdır.
- **Veri Seçimi:** Veri tabanından, konuyla ilgili istenen verilerin tespit edilip onları seçme aşamasıdır. Bu adım, birkaç veri kümesini birleştirerek sorguya uygun örneklem kümesini elde etmeyi gerektirir.
- **Veri Madenciliği:** Veriden örüntüler veya benzerlikler çıkarmak için kullanılan ve çeşitli yöntemleri içeren en önemli aşamadır.
- **Örüntü/Model Değerlendirme:** Veri tabanından gerçekte ilginç ve doğru olan verilerin tanımlanmasıdır.
- **Bilgi Sunumu:** Keşfedilen ve elde edilen bilgilerin geçerlilik, yenilik, yararlılık ve basitlik kıstaslarına göre değerlendirilmesi ve sunulmasıdır.

Yıllardır yapılan çalışmalar sonucunda geniş veri tabanlarında gizli olan bilgileri elde etmek için birçok yöntem geliştirilmiştir. Uzun zamandan beri özellikle batı ülkelerinin üzerinde çalıştığı bir konu olan VM, ileri teknoloji ürünü yazılımlar sayesinde daha çok kullanılmaya başlanmıştır. Ayrıca veri madenciliği, makine öğrenmesi, istatistik, veri tabanı yönetim sistemleri, veri ambarlama ve paralel programlama gibi farklı disiplinlerde kullanılan yaklaşımları birleştirmektedir. Özellikle veri madenciliği algoritmalarında paralel programlama kullanılması önemli bir ihtiyaç haline gelmiştir. (Karabatak, 2008). VM'yi en etkili şekilde uygulayabilmek için önemli olan noktalar aşağıdaki gibi özetlenebilir (Karabatak, 2008);

Farklı tipteki verileri ele alma: Sembolik veya kategorik, tamsayı, kesirli sayılar, çoklu ortam verisi, coğrafi bilgi içeren veriler gibi farklı tipteki veriler üzerinde işlem yapılmasına olanak tanıyan veri tipine özgü adanmış veri madenciliği algoritmaları geliştirilmektedir.

Veri madenciliği algoritmasının etkinliği ve ölçeklenebilirliği: Veri madenciliği algoritmasının bazı veri kümeleri içinden bilgi elde etmek etkin ve ölçeklenebilir olması gerekir. Bunun için veri madenciliği algoritmasının çalışma

zamanı, öngörü yapılabilir ve kabul edilebilir bir zaman diliminde olmalıdır. Zaman karmaşıklığı, üstel veya çok terimli bir karmaşıklığa sahip ise veri madenciliği algoritmasında uygulanması kullanışlı değildir.

Sonuçların yararlılık, kesinlik ve anlamlılık kıstaslarını sağlaması: Edilen kuralların kalitesini belirlemede önemli bir rol oynayan bu süreçte analizi yapılan veri tabanını doğru biçimde yansıtan sonuçlar elde edilmesi gerekir. Ayrıca gürültülü ve aykırı veriler ele alınmalıdır.

Keşfedilen kuralların çeşitli biçimlerde gösterimi: Yüksek düzeyli bir dil kullanarak hazırlanmış grafik ara yüzü kullanarak keşfedilen bilginin gösterim biçiminin seçilebilmesini sağlar.

Farklı birkaç soyutlama düzeyi ve etkileşimli veri madenciliği: Veri tabanlarından elde edilen bilgiler ile ilgili öngörü yapılması zordur. Bu yüzden veriler üzerine etkileşimli olarak sorgulama, sorguyu değiştirme, farklı açılardan ve farklı soyutlama düzeylerinden keşfedilen bilgiyi inceleyebilme esnekliği olmalıdır.

Farklı ortamlarda yer alan veri üzerinde işlem yapabilme: Yerel ağlar üzerinden dağıtık ve heterojen veri tabanı üzerinde işlem yapan kurumların bu veriler üzerinde analiz yapabilmesini gerektirir. Farklı veri tabanlarındaki verinin büyüklüğü ve dağıtık olması, paralel ve dağıtık veri madenciliği algoritmalarının kullanılmasını gerektirir.

Gizlilik ve veri güvenliğinin sağlanması: Veri tabanlarından istenilen bilgi keşfedildikten sonra verinin güvenliği ve gizliliği veri madenciliği sistemini kullanan kullanıcının haklarına ve erişim yetkilerine göre sağlanmalıdır.

Veri madenciliği teknikleri günümüzde çeşitli alanlarda başarı ile kullanılmaktadır. Bu uygulamalardan bazıları ilgi alanlarına göre aşağıda verilmiştir.

Pazarlama

- Müşterilerin satın alma örüntülerinin belirlenmesi,
- Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması,
- Posta kampanyalarında cevap verme oranının artırılması,

Bankacılık

- Farklı finansal göstergeler arasında gizli korelasyonların bulunması,
- Kredi kartı dolandırıcılıklarının tespiti,

Sigortacılık

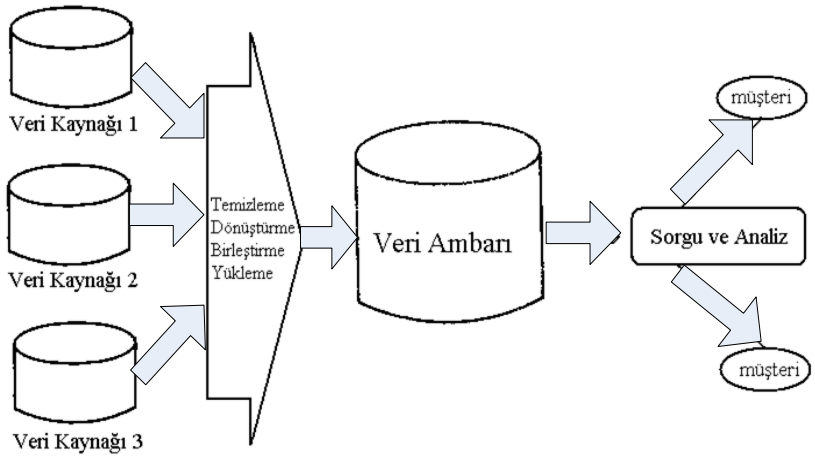
- Yeni poliçe talep edecek müşteriler ile ilgili öngörü yapılması,
- Sigorta dolandırıcılıklarının tespiti,

Eğitim

- Öğrencilerin bir dersteki akademik başarımlarının tespiti,
- Mezun öğrenciler ile ilgili tahminler ve yönlendirmeler,
- Yükseköğretim kurumlarında akademik planlama.

2.1. Veri Ambarları

Veri ambarları, veri tabanlarının birleştirerek veriyi işlemeye daha uygun bir özetini saklar. Günlük kullanımda amaç doğrultusunda gerekli veri ambardan alınarak veri madenciliği çalışması için standart bir forma çevrilmektedir (Karabatak, 2008)



Şekil 3. Tipik bir veri ambarı

Tipik bir veri ambarı yapısı Şekil 3'de verilmiştir. Veri ambarında veri oluşturulduktan sonra bu verinin manüel analizi Çevrimiçi Analitik İşleme (ÇAİ) programları kullanılarak yapılabilmektedir. Bu programlar, her boyutu veride bir alana karşılık gelen çok boyutlu bir küp olarak veriye bakmayı veya incelemeyi sağlamaktadırlar. Böylece boyut bazında gruplama, boyutlar arasındaki korelasyonları inceleme ve sonuçları grafik veya rapor olarak sunma olanağı sağlanmaktadır (Karabatak, 2008).

2.1.1. Çevrimiçi Analitik İşleme ve Veri Madenciliği

Veri ambarı üzerinde yürütülen ve iki önemli fonksiyon olan çevrimiçi analitik işleme ve veri madenciliği kavramları birbirini tamamlayan öğeler olmasına rağmen, birbirinden farklıdır. Her iki fonksiyonun da amacı, ham veri içinde gizli duran işle ilgili, yararlı bilgileri ortaya çıkarmaktır.

Veri madenciliğinde; istatistik, matematik, makine öğrenmesi ve yapay zekâ disiplinlerinden oldukça fazla yararlanılmaktadır. Veri madenciliği sürecini çevrimiçi analitik işleme sürecine üstün kılan özelliğini şu şekilde açıklamaktadır. Veri madenciliği tekniği, "ne?" sorusunun arkasında yatan "niçin?" sorusundan başka "başka ne?" ve "neden o?" gibi soruları yanıtlayarak çevrimiçi analitik işleme tekniğini geçmeye çalışır (Agrawal ve Srikant, 1994).

Standart sorgulama dilinin (*Structured Query Language-SQL*) saatle veya gün ile ölçülen cevaplama süresi, çevrimiçi analitik işlemede dakikalarla ölçülmektedir. Bunun dışında elle analiz edebilme imkânı da sağlamaktadır. Çevrimiçi analitik işlemeye verilen diğer bir isim çok boyutlu veri tabanıdır. İyi tasarlanmış bir çevrimiçi analitik işleme küpünde her kayıt sadece bir alt kümeye girmelidir. Bu, çok boyutlu veri tabanı olmanın en önemli kuralıdır (Ding, 2001).

2.1.2. Veri Ambarının Yapısı

Veri ambarında, verinin akışı ve son kullanıcıya kadar ulaşma süreci yakından incelenecek olursa, veri ambarının çok katmanlı yapısı ile karşılaşılmaktadır. Bu katmanlar kısaca özetlenirse (Han ve Kamber, 2006);

- **Kaynak Sistemler:** Verinin ilk toplandığı ve soyutlama seviyesinin en düşük olduğu kısımdır. Elde edilen bu ilk verinin karar destek için kullanılması söz konusu değildir.
- **Veri Nakli ve Temizlenmesi:** Verinin kaynak sistemlerinden çıkararak veri ambarı ve analiz ortamına gönderen yazılımların kullanılmasıdır.
- **Merkezi Depo:** Verilerin içinde bulunduğu çok büyük bir veri tabanıdır. Veri ambarının teknik olarak en gelişmiş kısmıdır.
- **Meta Veri:** Meta veri, verinin fiziksel alt yapısını hazırlamaktadır. Analiz için gerekli olan kısımların ön plana çıkarılmasına ve indeks, tablo, alan sayılarının belirlenmesine çalışılır. Veriye kendisi hakkında bilgi sağlar.

- **Datamartlar:** Datamartlar bir işletmede gerekli olan bilgiyi merkezileştirme özelliğine sahip bir sistemdir.
- **Operasyonel Geri Besleme:** Bu noktaya kadar olan veri işleme sonuçlarının geri besleme olarak operasyonel sisteme verilmesi sürecidir.
- **Son Kullanıcı:** Veri ambarının yapısı içindeki en önemli kısımdır. Son kullanıcıdan amaç analizciler, uygulama geliştiriciler ve işletmecilerdir.

2.1.3. Veri Madenciliği ve Veri Ambarı

Veri ambarı, son yıllarda hemen hemen her alanda kullanılmaya başlanmıştır. Günümüzde hipermarket satışlarından bankacılığa, astronomiden fiziğe birçok alanda büyük veri tabanları kullanılmaktadır. Veri ambarları, veri madenciliğinin kaynak olarak değerlendirildiği alanlardır. Verinin bir arada, temizlenmiş olarak bulunması ve veri madenciliğinin hayati döngüsünü tamamlayıcı özelliği, veri ambarının veri madenciliği için ne kadar önemli olduğunu kısaca özetlemektedir. Veri madenciliği, bilginin bu alanlardan çekilip çıkarıldığı araçlar kümesi sağlamaktadır (Karabatak, 2008).

Veri ambarı, bilgi çıkarımı konusunda çok etkin olmamasına rağmen, raporlama ve faturalama gibi faaliyetleri kolaylaştırarak veri madenciliğine yardımcı olmaktadır. Bunun dışında, veri ambarı veri madenciliğinde kullanılacak veriyi soyutlayarak temizlemektedir. Veri ambarından veriler çok hızlı elde edildiği için veri madenciliği sürecinin zaman karmaşıklığı azalmakta ve işini kolaylaştırmaktadır. Veri madenciliğini standart istatistiksel yöntemlere üstün kılan özelliği, büyük veriler ile çalışılabilir olmasıdır. Standart istatistikte, ana kütleden seçilen bir örneklem üzerinde çalışarak genelleştirme yapılır. Fakat bu durumun, ileriye yönelik tahmin veya öngörü yapamama, gelişmelere ve değişimlere cevap verememe gibi olumsuz yönleri vardır.

2.2. Veri Madenciliğinde Karşılaşılan Problemler

Veri madenciliği uygulamalarında, veri tabanından kaynaklanabilecek problemlerle karşılaşmak mümkündür. Veri madenciliği uygulanan ve doğru çalışan bir sistem veritabanı büyüdükçe tamamen farklı davranabilir. Veriye gürültü de eklendiğinde, sistemin başarımı olumsuz yönde etkilenebilir. Veri madenciliği uygulamalarında genel olarak aşağıdaki problemlerle karşılaşılır:

- Veri tabanının boyutu
- Gürültülü veri
- Boş değerler
- Eksik veri
- Artık veri
- Dinamik veri

2.2.1. Veri Tabanının Boyutu

Veri tabanı boyutunun çok büyük olması veri madenciliği sistemlerinde karşılaşılan en önemli sorunlardan biridir. Öğrenme algoritmalarının çoğu küçük örneklemleri ele alabilecek biçimde geliştirilmiştir. Örüntülerin gerçekten var olduğunu göstermek açısından örneklemin büyük olması avantaj olmasına rağmen örneklemden elde edilebilecek olası örüntü sayısı çok büyüktür. Aynı algoritmaların çok büyük örneklemlerde kullanılabilmesi için veri madenciliği yöntemleri ya sezgisel/buluşsal bir yaklaşımla arama uzayını taramalıdır ya da örneklemi yatay/dikey olarak indirgemelidir.

Yatay indirgeme, nitelik değerlerinin önceden belirlenmiş genelleme sıradüzenine göre, bir üst nitelik değeri ile değiştirilme işlemi yapıldıktan sonra aynı olan çokluların çıkarılması işlemidir. Dikey indirgeme, artık niteliklerin indirgenmesi işlemidir. Özellik seçimi yöntemleri ya da nitelik bağımlılık çizelgesi uygulanarak yapılır.

2.2.2. Gürültülü Veri

Verilerin ilk elden toplanması veya veri girişi sırasında oluşan sistem dışı hatalara gürültü adı verilir. Veri tabanlarındaki birçok özellik veya nitelik kullanıcı veya sistemden kaynaklanan problemlerden dolayı yanlış değer olabilir. Veri girişi sırasında oluşan hataları otomatik olarak gidermek veya eksik verileri tahmin etmede az da olsa destek veren veri tabanları bulunmaktadır. Hatalı veri, veri tabanlarında ciddi problemler oluşturduğundan veri madenciliği yönteminin, kullanılan veri kümesinde bulunan gürültülü veriyi tanımalı ve ihmal etmelidir. Quinlan (1986), gürültünün sınıflama üzerindeki etkisini araştırmak için bir dizi analiz yapmıştır. Deneysel sonuçlar, etiketli öğrenmede etiket üzerindeki gürültü öğrenme algoritmasının

başarımını doğrudan etkileyerek düşmesine sebep olmuştur. Buna karşın eğitim kümesindeki nesnelerin özellikleri/nitelikleri üzerindeki en çok %10'luk gürültü miktarı ayıklanabilmektedir. Chan ve Wong (1991), gürültünün etkisini analiz etmek için istatistiksel yöntemler kullanmışlardır.

2.2.3. Boş Değerler

Veri tabanlarında boş değer birincil anahtarda yer almayan herhangi bir niteliğin değeri olabilir. Eğer bir nitelik değeri boş ise o nitelik bilinmeyen ve uygulanamaz bir değere sahiptir. Bu durum ilişkisel veri tabanlarında sıkça karşımıza çıkmaktadır. Bir ilişkide yer alan tüm çoklular aynı sayıda niteliğe, niteliğin değeri boş olsa bile, sahip olmalıdır. Örneğin kişisel bilgisayarların özelliklerini tutan bir ilişkide bazı model bilgisayarlar için ses kartı modeli niteliğinin değeri boş olabilir.

Lee (1992), boş değeri, bilinmeyen, uygulanamaz ve bilinmeyen veya uygulanamaz olacak biçimde üçe ayıran bir yaklaşımı ilişkisel veri tabanlarını genişletmek için öne sürmüştür. Mevcut boş değer taşıyan veri için herhangi bir çözüm sunmayan bu yaklaşımın dışında bu konuda sadece bilinmeyen değer üzerinde çalışmalar yapılmıştır (Luba ve Lasocki, 1994; Grzymala-Busse, 1991). Boş değerli nitelikler veri kümesinde bulunuyorsa, ya bu çoklular tamamıyla ihmal edilmeli ya da bu çoklularda niteliğe olası en yakın değer atanmalıdır (Quinlan, 1986).

2.2.4. Eksik Veri

Veri tabanında tutulan veriler, veri madenciliği uygulanacak sistemi ne kadar çok yansıtırsa başarım o kadar yüksek olur. Ama veriler kurum ihtiyaçları göz önünde bulundurularak düzenlenip toplandığından, mevcut veri sistemi yeterince yansıtmayabilir. Örneğin sadece çok yaşlı insanların belirtilerinin bulunduğu bir veri kümesi kullanılarak hastalığın tanısını koymak için kurallar üretilseydi, bu kurallara dayanarak bir çocuğa tanı koymak pek doğru olmazdı. Bu gibi koşullarda bilgi keşif modeli, belirli bir güvenlik derecesinde öngörüsel kararlar alabilmelidir (Karabatak, 2008).

2.2.5. Artık Veri

Verilen veri kümesi, gerçek sisteme uygun olmayan veya artık nitelikler içerebilmektedir. Bu duruma, veri kümesi içerisinde yapılacak pek çok işlem sırasında karşılaşmak mümkündür. İstenen veriyi elde etmek için iki ilişki birleştirildiğinde, elde edilen ilişkide kullanıcının farkında olmadığı artık niteliklerin ortaya çıkma ihtimali vardır. Bu artık nitelikleri elemek için özellik seçimi olarak adlandırılan geliştirilmiş algoritmalar kullanılmaktadır.

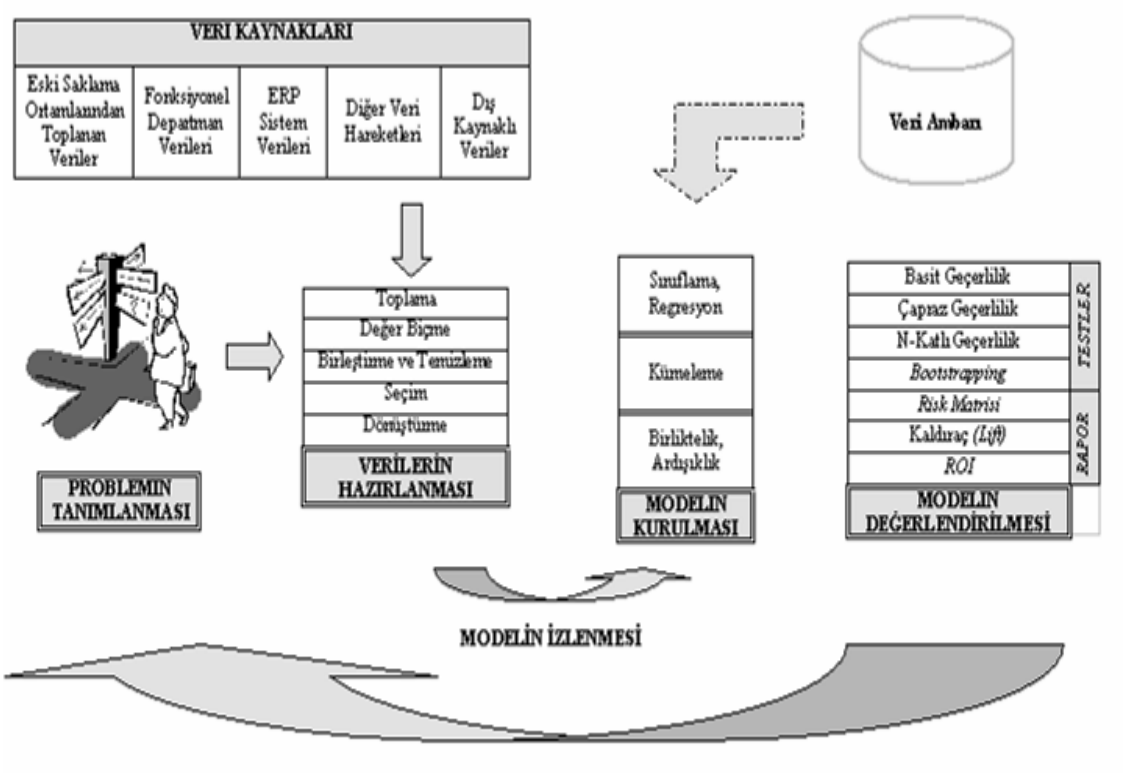
Özellik seçimi, tümevarıma dayalı öğrenmede budama öncesi yapılan işlem, hedef bağlamı tanımlamak için yeterli ve gerekli olan niteliklerin küçük bir alt kümesinin seçimi problemidir. Özellik seçimi yalnız arama uzayını küçültmekle kalmayıp, sınıflama işleminin kalitesini de arttırmaktadır (Kira ve Rendell, 1992; Almuallim ve Dietterich, 1991).

2.2.6. Dinamik Veri

Kurumsal çevrimiçi veri tabanlarının içeriği sürekli olarak değiştiğinden dinamik bir yapıdadır. Veri tabanının dinamik olması, bilgi keşfi yöntemleri için önemli sakıncalar meydana getirmektedir. Bu sakıncalardan ilki, veri tabanı uygulaması olarak sadece okuma yapan ve uzun süre çalışan bilgi keşfi metodu, mevcut veri tabanı ile birlikte çalıştırıldığında mevcut uygulamanın da başarımı önemli ölçüde azalmaktadır. Diğer bir sakınca ise, veri tabanında bulunan verilerin değişmediği düşünülüp, bilgi keşif metodu çevrimdışı veri üzerinde çalıştırıldığında, değişen verinin elde edilen örüntülere yansıması gerekmektedir. Bu işlem, bilgi keşfi metodunun ürettiği örüntüleri zaman içinde değişen veriye göre sadece ilgili örüntüleri yığmalı olarak günleme yeteneğine sahip olmasını gerektirir. Aktif veri tabanları tetikleme mekanizmalarına sahiptir ve bu özellik bilgi keşif yöntemleri ile birlikte kullanılabilir (Karabatak, 2008).

2.3. Veri Tabanlarında Bilgi Keşfi Süreci (VTBKS)

Veri madenciliği algoritmalarının üzerinde araştırma yapılan işin ve verilerin özelliklerinin bilinmemesi durumunda başarıyı sağlaması mümkün değildir. Başarının ilk şartı olarak öncelikle iş ve veri özelliklerinin öğrenilmesi gerekmektedir. Daha sonra Şekil 4'te verilen veri tabanında bilgi keşfi süreci aşamaları uygulanmalıdır (Karabatak, 2008).



Şekil 4. Veri tabanında bilgi keşfi süreci (Karabatak, 2008)

2.3.1. Problemin Tanımlanması

Uygulama amacının açık bir şekilde tanımlanması veri madenciliği çalışmalarının en önemli aşamasıdır. Bunun için öncelikle problem üzerine odaklanarak amacı açık bir dille ifade etmek, daha sonra elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceği tanımlanmak gerekir. Ayrıca yanlış öngörülerde katlanılacak olan maliyetlere ve doğru öngörülerde kazanılacak faydalara ilişkin öngörülere de bu aşamada yer verilmelidir.

2.3.2. Verilerin Hazırlanması

Verilerin hazırlanması aşaması, toplama, değer biçme, birleştirme ve temizleme, seçme ve dönüştürme adımlarından meydana gelmektedir. Verilerin hazırlanması sırasında ortaya çıkacak sorunlar modelin yanlış kurulmasına neden olur. Bu durum verilerin hazırlanması ve modelin kurulması aşamaları için, bir analistin veri keşfi sürecinin toplamı içerisinde enerji ve zamanının % 50 - % 85'ini harcamasına neden olmaktadır (Karabatak, 2008).

2.3.2.1. Toplama

Tanımı yapılan problem için gerekli verilerin toplanacağı veri kaynaklarının belirlenmesi adımıdır. Verilerin toplanmasında kuruluşun kendi veri kaynaklarının dışında, farklı kuruluşlara ait veri tabanlarından faydalanılabilir (Karabatak, 2008).

2.3.2.2. Değer Biçme

Farklı veri kaynaklarından toplanan verilen aynı nitelikler için veri uyumsuzluklarına neden olabilir. Farklı veri tabanlarındaki kodlamaların ve kullanılan ölçü birimlerinin aynı olmaması ve güncel olamayan verilen bulunması, bu uyumsuzlukların başlıca sebepleri olarak sıralanabilir. Ayrıca verilerin nasıl, nerede ve hangi koşullar altında toplandığı da önem taşımaktadır (Karabatak, 2008).

Bu nedenlerle, iyi sonuç alınacak modeller ancak iyi verilerin üzerine kurulabileceği için, toplanan verilerin ne ölçüde uyumlu oldukları bu adımda incelenerek değerlendirilmelidir.

2.3.2.3. Birleştirme ve Temizleme

Bu adımda, farklı kaynaklardan toplanan verilerde bulunan ve bir önceki adımda belirlenen sorunlar mümkün olduğu ölçüde giderilerek veriler tek bir veri tabanında toplanır.

2.3.2.4. Seçim

Bu adımda, kurulacak modele bağlı olarak veri seçimi yapılır. Örneğin öngörü yapan bir model için, bu adım bağımlı ve bağımsız değişkenlerin ve modelin eğitiminde kullanılacak veri kümesinin seçilmesi anlamını taşımaktadır.

Sıra numarası, kimlik numarası gibi anlamlı olmayan ve diğer değişkenlerin modeldeki ağırlığının azalmasına da neden olabilecek değişkenlerin modele girmemesi gerekmektedir. Bazı veri madenciliği algoritmaları konu ile ilgisi olmayan bu tip değişkenleri otomatik olarak elese de, pratikte bu işlemin kullanılan yazılıma bırakılmaması daha akılcı olacaktır.

2.3.2.5. Dönüştürme

Kredi riskinin öngörüsü için geliştirilen bir modelde, borç/gelir gibi önceden hesaplanmış bir oran yerine, ayrı ayrı borç ve gelir verilerinin kullanılması tercih edilebilir. Ayrıca modelde kullanılan algoritma, verilerin gösteriminde önemli rol oynayacaktır. Örneğin bir uygulamada bir YSA algoritmasının kullanılması durumunda, kategorik değişken değerlerinin evet/hayır olması; bir karar ağacı algoritmasının kullanılması durumunda ise örneğin gelir değişken değerlerinin yüksek/orta/düşük olarak gruplanmış olması modelin etkinliğini artıracaktır.

2.3.3. Modelin Kurulması ve Değerlendirilmesi

Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir.

Model kuruluş süreci, denetimli ve denetimsiz öğrenmenin kullanıldığı modellere göre farklılık göstermektedir. Örnekten öğrenme olarak da isimlendirilen denetimli öğrenmede, bir denetçi tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilir. Sistemin amacı verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunması ve bu özelliklerin kural cümleleri ile ifade edilmesidir. Öğrenme süreci tamamlandığında, tanımlanan kural cümleleri verilen yeni örneklere uygulanır ve yeni örneklerin hangi sınıfa ait olduğu kurulan model tarafından belirlenir. Eğiticisiz öğrenmede, kümeleme analizinde olduğu gibi ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden hareket ederek sınıfların tanımlanması amaçlanmaktadır. Eğiticili öğrenmede seçilen algoritmaya uygun olarak ilgili veriler hazırlandıktan sonra, ilk aşamada verinin bir kısmı modelin öğrenimi, diğer kısmı ise modelin geçerliliğinin test edilmesi için ayrılır.

Modelin öğrenimi, öğrenim kümesi kullanılarak gerçekleştirildikten sonra, test kümesi ile modelin doğruluk derecesi belirlenir. Kurulan modelin değerinin belirlenmesinde kullanılan diğer bir ölçü, model tarafından önerilen uygulamadan elde edilecek kazancın bu uygulamanın gerçekleştirilmesi için katlanılacak maliyete bölünmesi ile edilecek olan yatırımın geri dönüş oranıdır. Kurulan modelin doğruluk derecesi ne kadar yüksek olursa olsun, gerçek dünyayı tam anlamı ile modellediğini garanti edebilmek mümkün değildir. Yapılan testler sonucunda geçerli bir modelin doğru olmamasındaki başlıca nedenler, model kuruluşunda kabul edilen varsayımların ve modelde kullanılan verilerin doğru olmamasıdır. Örneğin modelin kurulması sırasında varsayılan enflasyon oranının zaman içerisinde değişmesi, bireyin satın alma davranışını belirgin olarak etkileyecektir.

2.3.4. Modelin Kullanılması

Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak da kullanılabilmektedir. Kurulan modeller farklı amaçlar için doğrudan kullanılabileceği gibi, daha karmaşık uygulamaların içine de gömülebilir.

2.3.5. Modelin İzlenmesi

Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirecektir. Öngörüsü yapılan ve gözlenen değişkenler arasındaki farklılığı gösteren grafikler model sonuçlarının izlenmesinde kullanılan yararlı bir yöntemdir.

2.4. Veri Madenciliği Modelleri

Veri madenciliği modelleri, temelde iki ana başlıkta incelenmektedir. Birincisi, elde edilen örüntülerden sonuçları bilinmeyen verilerin öngörüsü için kullanılan öngörü yapan model, diğeri ise eldeki verinin tanımlanmasını sağlayan tanımlayıcı modeldir (Agrawal ve diğerleri, 1993).

Öngörü yapan modellerde, sonuçları bilinen veriler kullanılarak bir model geliştirilir. Oluşturulan bu model kullanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerleri ile ilgili öngörü yapılması amaçlanmaktadır. Örneğin bir banka, daha

önceki dönemlerde müşterilerine verdiği tüm kredilerle ilgili bilgilere sahiptir. Bu bilgileri kullanarak daha sonraki dönemlerde müşterilere vereceği kredinin geri dönüp dönmeyeceğini müşteri bilgilerini kullanarak öngörü yapılabilir.

Tanımlayıcı modeller ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanmasını sağlamaktadır. Belirli özelliklere sahip insanların bazı davranışlarının birbirine benzerlik göstermesi tanımlayıcı modele bir örnek olabilir. Veri madenciliği modellerini gördükleri işlevlere göre ise üç ana başlık altında incelemek mümkündür. Bunlar;

- Sınıflama ve Regresyon,
- Kümeleme,
- Birliktelik kuralları ve ardışık zamanlı örüntülerdir.

2.4.1. Sınıflama ve Regresyon Modelleri

Sınıflama ve regresyon, veri madenciliği tekniklerinde en çok kullanılan yöntemlerden biridir. Mevcut verilerden hareket ederek gelecekteki durumlar ile ilgili öngörü yapılması durumunda faydalanılır ve yeni bir veri elemanını daha önceden belirlenmiş sınıflara atamayı amaçlar. Sınıflama ve regresyon arasındaki temel fark, öngörü yapılan bağımlı değişkenin kategorik veya süreklilik gösteren bir değere sahip olmasıdır. Ancak her iki model de birbirine giderek yaklaşmakta ve bunun sonucu olarak aynı tekniklerden yararlanılması mümkün olmaktadır. Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler şunlardır;

- Karar Ağaçları
- Yapay Sinir Ağları
- Genetik Algoritmalar
- K-En Yakın Komşu
- Bellek Temelli Nedenleme
- Lojistik Regresyon

KA tekniğini kullanarak verinin sınıflanması iki basamaklı bir işlemdir. İlk basamak öğrenme basamağıdır. Öğrenme basamağında önceden bilinen bir eğitim verisi, model oluşturmak amacıyla sınıflama algoritması tarafından analiz edilir. Öğrenilen model, sınıflama kuralları veya KA olarak gösterilir. İkinci basamak ise sınıflama basamağıdır. Sınıflama basamağında test verisi, sınıflama kurallarının veya

KA doğruluğunu belirlemek amacıyla kullanılır. Eğer doğruluk kabul edilebilir oranda ise, kurallar yeni verilerin sınıflanması amacıyla kullanılır. Test verisine uygulanan bir modelin doğruluğu, yaptığı doğru sınıflamanın test verisindeki tüm sınıflara oranıdır. Her test örneğinde bilinen sınıf, model tarafından tahmin edilen sınıf ile karşılaştırılır. Eğer modelin doğruluğu kabul edilebilir bir değer ise model, sınıfı bilinmeyen yeni verileri sınıflama amacıyla kullanılabilir.

YSA temelde tamamen insan beyni örneklenerek geliştirilmiş bir teknolojidir. Bilindiği gibi; öğrenme, hatırlama, düşünme gibi tüm insan davranışlarının temelinde sinir hücreleri bulunmaktadır. İnsan beyninde tahminen 10^{11} adet sinir hücresi olduğu düşünülmektedir ve bu sinir hücreleri arasında sonsuz diyebileceğimiz sayıda sinaptik birleşme denilen sinirler arası bağ vardır. Bu sayıdaki bir birleşimi gerçekleştirebilecek bir bilgisayar sisteminin dünya büyüklüğünde olması gerektiği söylenmektedir; ancak 50 yıl sonra bunun büyük bir yanılgı olmayacağını bu günden kimse söyleyemez. İnsan beyninin bu karmaşıklığı göz önüne alındığında, günümüz teknolojisinin 1.5 kg'lık insan beynine oranla henüz çok geride olduğunu söylemek yanlış olmaz (Türkoğlu, 1996). YSA'nın hesaplama ve bilgi işleme gücünü, paralel dağılmış yapısından, öğrenebilme ve genelleme yeteneğinden aldığı söylenebilir. Genelleme, eğitim ya da öğrenme sürecinde karşılaşılmayan girişler için de YSA'nın uygun tepkileri üretmesi olarak tanımlanır. Bu üstün özellikleri, YSA'nın karmaşık problemleri çözebilme yeteneğini gösterir.

GA; en iyinin korunumu ve doğal seçilim ilkesinin benzetim yoluyla bilgisayarlara uygulanması ile elde edilebilir bir arama yöntemidir (Goldberg, 1989). Standart bir GA'da, aday sonuçlar eşit boyutlu vektörler olarak ifade edilir. Başlangıçta, bu vektörlerden bir grup, rastlantısal olarak seçilerek belirli bir büyüklükte bir popülasyon (toplum) oluşturulur. Kromozom adı verilen bu vektörler, yeni nesiller (nesil) oluşturarak değişikliklere uğrar. Bir kromozomun üzerindeki genler, *n* boyutlu vektörlerin bir boyutuna karşılık gelmektedir. Her yeni nesilde kromozomların iyiliği ölçülür, yani her vektör (kromozom), amaç fonksiyonuna yerleştirilerek vermiş olduğu sonuç hesaplanır. Bir sonraki nesil oluşturulurken, bazı kromozomlar yeniden üretilir, çaprazlanır ve mutasyona uğratılır.

K-NN veri madenciliğinde sınıflama amacıyla kullanılan bir diğer teknik ise örnekseme yoluyla öğrenmeye dayanan en yakın komşu algoritmasıdır (Biçer, 2002).

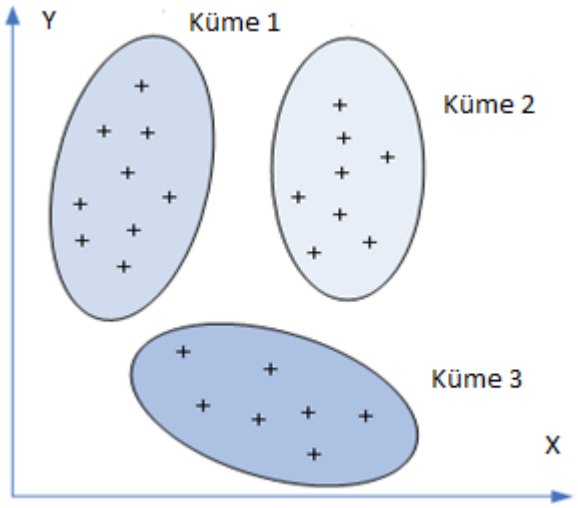
Bu teknikte tüm örneklemler bir örüntü uzayında saklanır. Algoritma, bilinmeyen bir örneklemin hangi sınıfa dâhil olduğunu belirlemek için örüntü uzayını araştırarak bilinmeyen örnekleme en yakın olan k örneklemi bulur. Yakınlık Öklid uzaklığı ile tanımlanır. Daha sonra, bilinmeyen örneklem, en yakın komşu içinden en çok benzediği sınıfa atanır.

BTY denetimli öğrenmenin kullanıldığı veri madenciliği tekniklerindendir. Bu tekniğin temel özelliği, daha önceki deneyimlerimizden faydalanarak elimizdeki problemlere benzer durumları tanımlayıp geçmiş benzer problemlere getirdiğimiz uygun çözümleri mevcut problemlerimize uygulamaya çalışmaktır.

Naive bayes; bu algoritma da her kriterin sonuca olan etkilerinin olasılık olarak hesaplanması temeline dayanmaktadır (Duda ve Hart, 1989). Veri Madenciliği işlemini en çok verilen örneklerden biri ile açıklayacak olursak elimizde tenis maçının oynanıp oynanmamasına dair bir bilgi olduğunu düşünelim. Ancak bu bilgiye göre tenis maçının oynanması veya oynanmaması durumu kaydedilirken o anki hava durumu, sıcaklık, nem ve rüzgâr durumu bilgileri de alınmış olsun. Biz bu bilgileri değerlendirdiğimizde varsayılan tahmin yöntemleri ile hava bugün rüzgârlı tenis maçı bugün oynanmaz şeklinde kararları farkında olmasak da veririz. Ancak Veri Madenciliği bu kararların tüm kriterlerin etkisi ile verildiği bir yaklaşımdır. Dolayısıyla biz ileride öğrettiğimiz sisteme bugün hava güneşli, sıcak, nemli ve rüzgâr yok şeklinde bir bilgiyi verdiğimizde sistem eğitildiği daha önce gerçekleşmiş istatistiklerden faydalanarak tenis maçının oynanma ve oynanmama ihtimalini hesaplar ve bize tahminini bildirir.

2.4.2. Kümeleme

Kümeleme, veri tabanındaki verileri alt kümelere ayıran bir yöntemdir (Everitt, 1993). Her bir kümede yer alan elemanlar birbirlerine çok benzemekte, özellikleri farklı olan elemanlar ise farklı kümelerde bulunmaktadır. Başlangıç aşamasında veri tabanındaki kayıtların hangi kümelere ayrılacağı veya kümelerin hangi özelliklere sahip olacağı bilinmemekte ve konunun uzmanı olan bir kişi tarafından bu özellikler belirlenmektedir. Şekil 5'de üç kümeye ayrılmış örnek bir veri seti gösterilmiştir.



Şekil 5. Kümeleme

Kümeleme analizinde amaç birbirine en çok benzeyen nesneleri aynı grupta toplamaktır. Benzemekten kasıt, geometrik anlamda uzaklık olarak birbirine en yakın nesnelerin seçilmesidir. Bu nedenle nesnelerin sayısal değer olması gerekir. Bu noktada değişkenler dörde ayrılmaktadır. Bunlar (Everitt, 1993);

- Kategorik Değişkenler: Bu değişkenler arasında sadece birbirine benzeme söz konusudur. Sıralama mümkün değildir. (örneğin siyah>beyaz durumu söz konusu değildir.
- Sıralama Değişkenleri: $X>Y$ şeklinde bir sıralama yapmak mümkündür. Ama büyüklüğün ne kadar olduğu belli değildir. (X-Y bulunamaz)
- Aralık Ölçekli Değişkenler: İki nokta arasındaki uzaklık hesaplanır. Fakat bu tür değişkenlerde gerçek "0" değeri yoktur.
- Oran Ölçekli Değişkenler: Anlamlı "0" noktasının bulunduğu, her türlü dört işleme açık değişkenler topluluğudur. (örneğin 20 yaşındaki bir kişi 10 yaşındaki bir kişiden iki katı yaştadır denilebilir.)

Kategorik ve sıralama değişkenlerinin matematiksel hesaplamaların yapılabileceği sayısal değerlere dönüştürülmesi şarttır.

2.4.3. Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler

Birliktelik kuralları, işlemlerden oluşan ve her bir işlemin de ürünlerin birlikteliğinden oluştuğu düşünülen bir veri tabanında, bütün ürün birliktelilerini

tarayarak, sık tekrarlanan ürün birlikteliklerini veri tabanından ortaya çıkarmaktır (Karabatak, 2008; Karabatak ve İnce, 2004). Bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi ürün veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi, müşteriye daha fazla ürünün satılmasını sağlama yollarından biridir. Satın alma eğilimlerinin tanımlanmasını sağlayan birliktelik kuralları ve ardışık zamanlı örüntüler, pazarlama amaçlı olarak pazar sepet analizi adı altında veri madenciliğinde yaygın olarak kullanılmaktadır.

Birliktelik kuralları eş zamanlı olarak gerçekleşen ilişkilerin tanımlanmasında kullanılır. Örneğin, bira satın alan müşterilerin %75 ihtimalle patates cipsi de almaları veya ekmek ve yağ alanların %90'ının süt de satın almaları birliktelik kuralları kapsamında tespit edilebilir.

Ardışık zamanlı örüntüler ise birbiri ile ilişkisi olan ancak birbirini izleyen dönemlerde gerçekleşen ilişkilerin tanımlanmasında kullanılır. *X* ameliyatı yapıldığında, 15 gün içerisinde %45 ihtimalle *Y* enfeksiyonu oluşması, çekiç alan bir müşterinin ilk üç ay içerisinde %15, bu dönemi izleyen süre içerisinde ise %10 ihtimalle çivi alması ardışık zamanlı örüntüler olarak tanımlanmaktadır.

2.5. Literatür ve İlgili Araştırmalar

Birçok araştırmacı ve yazar özellikle yükseköğretimde veri madenciliğinin çeşitli uygulamalarını araştırmıştır ve tartışmışlardır. Yazarlar, yükseköğretimde veri madenciliği uygulamalarının önemini anlamak ve ilgili alandaki kaliteyi artırmak için kapsamlı literatür taraması yapmaktadırlar. Bu bağlamda, Ranjan ve Khalil (2008), özellikle danışma süreci ve akademik anlamda kabul görmüş ve genelde eğitim yönetimi alanına yönelik bir veri madenciliği süreci önermiştir. Çalışmada veri madenciliği teknikleri kullanılarak hazırlanmış örnekler, geleceğe yönelik yöntemler, öğrenilen dersler ve çalışmanın sınırlılıkları ayrıntılı olarak incelenmiştir. Beikzadeh ve Delavari (2005), veri madenciliği süreçlerini gösteren bir analiz modeli önermiştir. Önerilen analiz modeli, yükseköğretim sistemi için veri madenciliği sistemini geliştirmede bir karar destek aracı olarak kullanılabilir. Ayrıca veri madenciliği teknolojileri sayesinde gelişebilen eğitim süreçlerinin parçalarını belirlemek için bir rehber ya da yol haritası olarak da görev alınabilir. Sembiring, Zarlis, Hartama, Ramliana ve Wani (2011), öğrenci davranışlarını ve başarılarını analiz etmek ve öğrenci performans tahmin modelini geliştirmek için veri madenciliği tekniklerini

kullanmışlardır. Bresfelean, Bresfelean ve Ghisoiu (2008), "FarthestFirst" algoritması ve "Weka J48" de bulunan öğrenme yoluyla sınıflandırma ve veri kümeleme yöntemlerini öğrencilerin akademik başarı ya da başarısızlığını saptama amaçlı kullanmışlardır. Mamcenko, Sileikiene, Lieponiene ve Kulvietiene (2011), veri madenciliğindeki birliktelik kuralları ve kümeleme yöntemlerini kullanarak elektronik sınav verilerini analiz etmeyi önermişlerdir. Önerilen çalışmanın amacı elde edilen sonuçları değerlendirmek, yorumlamak ve tanımlayıcı bir model kullanarak elektronik sınav sistemini geliştirmektir. Zhang, Oussena, Clark ve Kim (2010), veri madenciliğinin risk altındaki öğrencilere nasıl yardım edebileceğini, dersin veya modülün uygunluğunun nasıl değerlendirebileceğini ve elde edilen sonuçların öğrencilere nasıl uyarlanacağını inceleyen bir çalışma sunmuşturlar. Ramaswami ve Bhaskaran (2010), deneysel bir yöntem ile birinci ve ikinci kaynaklardan elde edilen verilerin oluşturacağı bir veri tabanı geliştirilmiştir. Normal öğrenciler birinci kaynağı oluştururken, ikinci kaynağı da özel öğretim öğrencileri oluşturmaktadır. 2006 yılında farklı bölgelerdeki farklı okullardan toplanan 1000 öğrenci verisi, veri kümesini oluşturmaktadır. İşlenmemiş veri ilkönce bir dizi ön işlemden geçirilmiş ve daha sonra ilgili özellikler oluşturulmuştur. Sonuç olarak 772 öğrenci verisi ilgili tahmin modelinin oluşturulması için kullanılmıştır.

Bozkır, Mazman ve Sezer (2010), "Facebook" kullanma süresini ve sıklığını tahmin eden bir çalışma önermişlerdir. Çalışma, 570 "Facebook" kullanıcısına uygulanan anket verilerinden oluşturulan veri tabanında uygulanmıştır. Ayrıca çalışmada "Facebook'un" eğitime katkısı öğrenci görüşlerine bağlı olarak irdelenmiştir. Mardikyan ve Badur (2011), karar ağaçları ve regresyon yöntemlerini kullanarak öğretim üyelerinin öğretme performanslarını değerlendirmişlerdir. İlgili veri tabanı Boğaziçi üniversitesi, Yönetim Bilişim sistemleri bölümü öğrencilerinin görüşleri doğrultusunda hazırlanmıştır. Ayrıca veri tabanında dersler ve diğer öğretim üyeleri hakkında da bilgiler bulunmaktadır. Gaafar ve Khamis (2009) tarafından makine mühendisliği bölümü öğrencilerinden mezun olabilecek öğrencilerin profillerini belirleyebilecek bir yöntem önerilmiştir. Çalışma, Kahire Amerikan üniversitesinde gerçekleştirilmiştir. Farklı veri madenciliği yöntemleri kullanılarak, farklı birimlerden alınan verilerle bir veri tabanı oluşturulmuştur. Böylece, iki farklı öğrenci profili

modellenmiştir. Birinci öğrenci profili başarılı yani mezun olabilecek öğrenciler, ikinci öğrenci profili ise başarısız ya da okulu bırakabilecek öğrenciler oluşturmaktadır.

Delavari, Phon-amnuaisuk ve Beikzadeh (2008), veri madenciliği yöntemlerini kullanarak yükseköğretim kurumları için bir karar destek sistemi geliştirmişlerdir. Chao, Huang ve Chang (2010), "Bilgi Tabanlı Destek Servis Merkezi" olarak adlandırılan bir sistem önermiştir. Sistem öğrencilerin özel sorunlarını çevrimiçi iletebilecekleri bir form sunmaktadır. Ayrıca sistem, iletilen problemin cevaplarını veri tabanında aramakta ve ilgili çözümü öğrenciye geri iletmektedir. Erdoğan ve Timor (2005) tarafından yapılan çalışmada Maltepe üniversitesi öğrencilerinin bazı karakteristikleri k-ortalamalar kümeleme algoritması ile kümelenmiştir. Çalışmada 722 öğrenci verisi kullanılmıştır. Sistemin girişini üniversite giriş sınavı puanı oluşturmaktadır. 2002 yılında öğrencilerin yükseköğretim ile ilgili memnuniyetlerini inceleyen bir veri madenciliği yaklaşımı Luan (2002) tarafından önerilmiştir. Vranic, Pintar ve Skocir (2007), veri madenciliği teknikleri ile öğretim kalitesinin nasıl artırılacağı ile ilgili bir çalışma gerçekleştirmiştir. Çalışma belirli bir dersteki öğrencilere uygulanmıştır.

Minaei-Bidgoli, Kashy, Kortmeyer ve Punch (2003), web tabanlı eğitsel bir veri tabanında bulunan verilere bağlı olarak öğrencilerin final sınavında alacakları notu tahmin etmiştir. Çalışmada genetik algoritma kullanılmıştır. Kotsiantis, Patriarcheas ve Nikxenos (2010) artımsal Bayes, en yakın komşu ve WINNOW algoritmalarını birleşimsel olarak kullanan bir yapı önermiştir. Çalışma, uzaktan eğitim alan bir gurup öğrenci üzerinde test edilmiştir. Bharadwaj ve Pal (2011), veri madenciliği yöntemleri ile bilgisayar uygulama dersini alan 5 farklı üniversiteden toplam 300 lisans öğrencisinin başarım değerlendirmesini gerçekleştirmiştir. 17 adet öznitelik, Bayes sınıflandırıcısı kullanılarak sınıflandırılmıştır. Bu özniteliklerden bazıları sırası ile lise ortalaması, öğrencinin ikamet ettiği bölge, öğrenci velisinin yeterliliği ve ailenin ekonomik durumudur. Daha sonra Bharadwaj ve Pal (2011), öğrencilerin yarıyıl sonundaki başarımlarını modellemek için öğrencinin devam durumu, sınıf testleri ve seminer notlarını gibi öznitelikleri kullanmıştır. Ramaswami ve Bhaskaran (2009), farklı öznitelik seçme yöntemlerinin karşılaştırarak öğrenci performanslarını değerlendirmiştir. Çalışmada öğrencilere ait öznitelikler, sırası ile öğrencinin vizyonu, yemek alışkanlıkları ve ailesi ile ilgili birçok ilginç parametreyi kapsamaktadır. Sen ve diğerleri (2012), orta öğretim yerleştirme sınavı sonuçlarını tahmin etmek için çeşitli

veri madenciliği modellerini kullanmışlardır. Çalışmada yerleştirme sınavına etki eden en önemli parametrenin belirlenmesi için duyarlılık analizleri gerçekleştirilmiştir. Çalışmada C.5 karar ağacı yönteminin bu uygulama için destek vektör makineleri, lojistik regresyon ve yapay sinir ağları modellerinden daha iyi sonuçlar ürettiği bildirilmiştir. Yine benzer bir çalışma Kovacic (2010) tarafından sunulmuştur. Kovacic (2010), öğrenci kayıt bilgilerini veri madenciliği yöntemleri ile değerlendirerek öğrenci başarısını modellemeyi amaçlamıştır. Bu amaçla CHAID ve CART algoritmaları kullanılmıştır. Ben-Zadok, Hershkovitz, Mintz ve Nachmias (2007), veri madenciliği teknikleri ile öğrencilerin öğrenme davranışlarını analiz etmiş ve final sınavları öncesi risk altındaki öğrencilerin uyarılmasını sağlayan bir örnek çalışma önermişlerdir.

Al-Radaideh, Al-Shawakfa ve Al-Najjar (2006), yüksek öğretim kalitesini artırmak için veri madenciliği yöntemlerini öğrencilerin akademik verilerini analiz etmekte ve değerlendirmekte kullanmışlardır. Böylece yükseköğretim yöneticileri bu sınıflandırma modelini kullanarak ders çıktılarını geliştirebileceklerdir. Karabatak ve İnce (2004), yaptığı çalışmada öğrencilerin tüm derslerini dikkate almış ve 4 yıllık öğretim sürecindeki tüm ders notları arasındaki ilişkileri birliktelik kuralı ile ortaya çıkarmıştır. Bu sayede öğrencilerin gelecekte derslerden alacağı notların tahmini yapılmıştır. Ayesha, Tasleem, Sattar ve Khan (2010), tarafından geliştirilen yazılım sayesinde, okula kayıt yaptıracak öğrencilerin sayıları karar ağacı yöntemi ile tahmin edilmektedir. Böylelikle idareciler, yeni kayıt yaptıracak öğrenciler için gerekli kaynakları hazırlayıp yönetebilecektir. Jadric, Garaca ve Cukusic (2010), öğrencilerin okul türlerine ve giriş puanlarına göre, veri madenciliği yöntemlerini kullanarak, öğrencilerin okuldan ayrılma durumlarını incelemiştir. Ayrıca yazar veri madenciliği uygulamalarının bu alanda henüz yeterli olmadığını vurgulamıştır. Romero, Ventura ve Salcines (2008), eğitim ve öğrenme yönetim sistemlerinden biri olan Moodle'ın verilerine birliktelik kuralı, kümeleme, sınıflandırma gibi veri madenciliği teknikleri uygulamıştır. E-öğrenme yöneticileri ve çevrimiçi eğitimcileri hem teorik hem de pratik olarak tanıştırmak amaçlanmıştır. Zaiane ve Luo (2001), eğitimciler ve öğrenme süreçlerinin daha iyi değerlendirmesi amacıyla web-tabanlı öğrenme ortamının tasarımında veri madenciliği ve makine öğrenmesi tekniklerinin kullanılabilirliğini tartışmıştır.

ÜÇÜNCÜ BÖLÜM

III. ÖĞRENCİLERİN AKADEMİK BAŞARILARININ VERİ MADENCİLİĞİ METOTLARI ile TAHMİNİ

Öğrencilerin başarımlarını geliştirmeye yardımcı, çeşitli gizli bilgileri barındıran eğitsel veri tabanlarında büyük miktarlarda veriler birikmektedir. Veri tabanlarında depolanan bu verilerden, gelecekte kullanılabilecek gizli ve faydalı bilgiler ortaya çıkarılabilir. Bir kursa veya derse kayıt yaptıracak öğrencilerin profillerinin tahmininde, geleneksel sınıfta öğretme modelinin aksaklıklarının belirlenmesinde, çevrimiçi sınavlardaki olumsuzlukların tespitinde, öğrenci sınav kâğıtlarındaki anormal değerlerin tespitinde, öğrenci performanslarının tahminde, vb. gizli bilgiler ortaya çıkarılarak eğitsel amaca yönelik kullanılabilir.

3.1. Kullanılan Veri

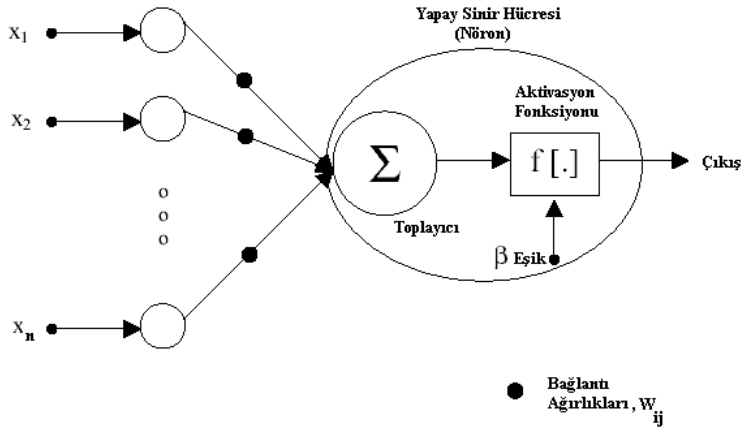
Bu tez çalışmasında kullanılan veriler, öğrencilerin özel bilgileri (adı soyadı, TC kimlik no, vb.) gizli tutularak, Fırat Üniversitesi Öğrenci İşleri veri tabanından temin edilmiştir. Veriler, 2007 ve 2011 yılları arasında Fırat Üniversitesi, BÖTE bölümünden mezun olan öğrencilerin ders ve yılsonu notlarından oluşmaktadır. Öğrencilerin mezun olması için gerekli olan 49 adet mesleki ve kültürel dersleri veri kümesinin özniteliklerini oluşturmaktadır. Veri kümesi oluşturulurken 127 öğrencinin yılsonu notları göz önüne alınmıştır. Böylece 127x49'luk bir veri matrisi elde edilmiştir. Öğrencilerin aldıkları dersler için başarı puanları bağıl değerlendirme sistemi ile hesaplanmaktadır. Bir dersin sınavı 100 üzerinden değerlendirilir ve dersin yılsonu notu ara sınav notunun 0,4 katı ile genel sınav notunun 0,6 katının toplamı olarak hesaplanır. Daha sonra bağıl değerlendirme sistemine göre öğrencinin, o dersteki ders notu harfli sisteme göre bulunur. Harfli notlar Tablo 1'e göre dönüştürülerek öğrencinin ders notu hesaplanır.

Tablo 1. Not dönüştürme tablosu

Ders Notu	Başarı Katsayısı
AA	4
BA	3.5
BB	3
CB	2.5
CC	2
DC	1.5
DD	1
FF	0

3.2. Yapay Sinir Hücresi

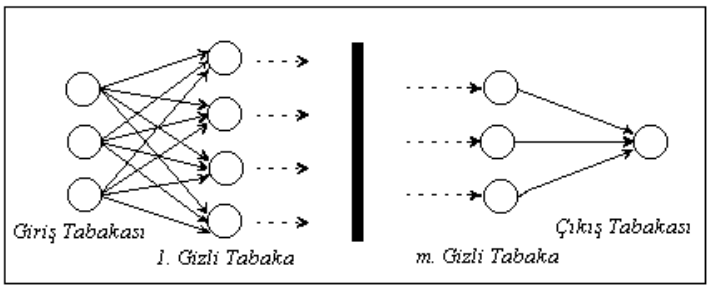
Sinir ağları paralel olarak çalışan basit elemanlardır. Gerçek bir sinir hücresine eşdeğer bileşenlere sahip yapay sinir hücresi Şekil 6'da gösterilmiştir. Gövdenin giriş birimi olan bağlantıların her birinin kendine ait bir ağırlık çarpanı vardır. Bu ağırlık değeri pozitif veya negatif olabilir. Uygulanan girişlerin ağırlık değeriyle çarpımları, iki kısımdan oluşan gövdenin ilk kısmında toplanır. Bu toplam, ikinci kısmı tanımlayan aktivasyon fonksiyonunun girdisidir.



Şekil 6. Yapay Sinir Hücresi

3.3. Yapay Sinir Ağı (YSA)

YSA, öngörülen bir sayıda yapay sinir hücresinin, bazı amaçlarla belirli bir mimaride yapılandırılmasıyla ortaya çıkar. Bu mimari yapı, genellikle, birkaç katmandan oluşur. İlk katman, giriş katmanıdır. Giriş katmanındaki elemanların ağırlık çarpanları ve aktivasyon fonksiyonlarının olmaması sebebiyle veri girişinden başka bir işlem yapmamalarıdır. Çıkış katmanı da son katmandır. Tercihe bağlı olarak farklı sayıda olabilen diğer ara katmanların ortak adı gizli katmandır (Türkoğlu, 1996). Gizli katmanların amacı giriş ve çıkış katmanları arasında gerekli bir takım işlemler yapmaktır. Giriş katmanı geniş olduğu zaman gizli katmanlar sayesinde yüksek dereceli istatistiksel veri elde edilebilir. Çok katmanlı yapılarda (n). katmanın çıkış sinyalleri (n+1). katmanın giriş sinyalleri olarak kullanılır. m adet giriş düğümü, ilk gizli katmanında h_1 adet nöron (sinir hücresi), ikinci gizli katmanında h_2 adet nöron ve çıkış katmanında q adet nöron bulunan bir çok katmanlı ileri besleme ağı m-h_1-h_2-q ağı olarak adlandırılır. Eğer her katmanda bulunan nöronlar bir sonraki katmanın tüm nöronlarına bağlı ise bu tip ağa tam bağlantılı ağ denir. Eğer bu sinaptik bağlantılardan bazıları eksikse ağ, kısmi bağlantılı ağ adını alır (Türkoğlu, 1996). Şekil 7'de çok katmanlı bir YSA ağı gösterilmiştir.

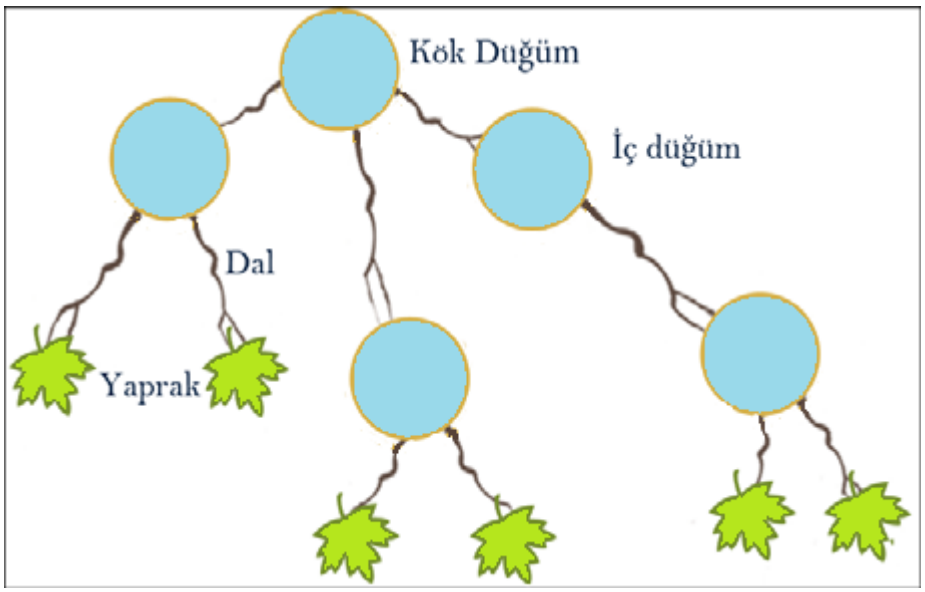


Şekil 7. Çok katmanlı yapay sinir ağı

3.4. Karar Ağaçları

KA, verilen bir problemin yapısına bağlı olarak bir ağaç yapısı şeklinde sınıflandırma ve regresyon modeli oluşturmaktadır. Ağaç yapılarının oluşturulmasında kullanılan kuralların anlaşılabilir olması yöntemin kullanımını kolay ve uygulanabilir bir hale getirmiştir. KA sınıflandırma ve regresyon bir probleminin çözümünde çok aşamalı ve ardışık bir yaklaşım ile basit bir karar verme işlemini gerçekleştirmektedir (Safavian ve Landgrebe, 1991). Tahmin edilecek hedef öznitelikler ayrık verilerden veya belirli kategorilerden oluşuyorsa kullanılan model sınıflandırma ağacı, öznitelik

verileri sürekli değişkenlerden oluşuyorsa model regresyon ağacı olarak adlandırılmaktadır (Nefeslioglu vd., 2010). Basit bir regresyon ağacı yapısı Şekil 8'de gösterilmiştir. Bu yapıda her bir öznitelik bir düğüm tarafından temsil edilirken, ağaç yapısının en üst kısmı kök ve en alt kısmı yapraklardan oluşmaktadır. Kök ve yapraklar arasında kalan ve üst düğümler ile alt düğümler arasındaki ilişkiyi sağlayan kısımlar ise dal olarak ifade edilmektedir (Quinlan, 1993). KA yapısı oluşturulmasında temel prensip verilere ilişkin bir dizi sorular sorularak karar kurallarının oluşturulmasıdır. Bu işlem için ağaç yapısının temel elemanı olan kök düğümünde sorular sorulmaya başlanır ve ağaç yapısının son elemanı olan yapraklara ulaşılıncaya kadar ağacın büyümesi veya dallanması devam eder (Pal ve Mather, 2003).



Şekil 8. Düğüm, dal ve yapraklardan oluşan basit bir karar ağacı yapısı (Kavzoğlu, T., Şahin, E.K. ve Çölkesen, İ, 2012).

Ağaç yapısının oluşturulmasındaki en önemli aşama ağaçtaki dallanmanın hangi kritere ya da öznitelik değerine göre olacağının belirlenmesidir. Literatürde bu problemin çözümü için geliştirilmiş çeşitli yaklaşımlar vardır. Bunlardan en önemlileri bilgi kazancı ve bilgi kazanç oranı (Quinlan, 1993), Gini indeksi (Breiman vd., 1984), Twoing kuralı (Breiman vd., 1984) ve Ki–Kare olasılık tablo istatistiği (Mingers, 1989) yaklaşımlarıdır. Tek değişkenli karar ağaçlarından ID3 algoritması bilgi kazancı yaklaşımını kullanırken, C4.5 algoritması bölünme bilgisi kavramı ile bilgi kazancından yararlanmaktadır. Sınıflandırma ve regresyon ağacı olarak bilinen CART algoritması ise

Twoing kuralını kullanmaktadır (Breiman vd., 1984). CART algoritmasının en önemli özelliği regresyon ağaçları oluşturma yeteneğidir. Regresyon ağaçlarının yapraklarında tahmin edilecek öznitelik değeri kategorik bir sınıf değeri değil süreklilik gösteren bir gerçek sayı değeridir. Bu regresyon probleminin çözümü için CART algoritması tahmin edilecek değerlerin karesel ortalama hatasını minimum yapacak bölünmeleri hesaplayarak ağacın büyümesini ve dallanmasını gerçekleştirir. Her bir yaprakta ulaşılan tahminler düğüm için hesaplanan ağırlıklı ortalamalara bağlı olarak hesaplanır (Rokach ve Maiman, 2008). CART algoritması ile oluşturulan ağaç yapısında ikili dallanmalar söz konusu olup, her bir karar düğümünden itibaren ağacın iki alt dala ayrılması prensibi esas alınmaktadır (Breiman vd.,1984; Lawrence vd., 2001). Diğer bir değişle bir düğümde seçme işlemi yapılmasının ardından düğümlerden sadece iki dal ayrılabilir. CART algoritmasında, bir düğümde belirli bir kriter (Twoing kuralı) uygulanarak bölünme işlemi gerçekleştirilir. Bunun için tüm özniteliklere ait değerler göz önüne alınır ve tüm eşleşmelerden sonra iki bölünme elde edilerek seçme işlemi gerçekleştirilir (Özkan, 2008). Twoing algoritmasında özniteliklerin içerdiği değerler göz önüne alınarak eğitim kümesi aday bölünme olarak adlandırılan iki ayrı dala ayrılır. Bir t düğümünde sağ ($t_{sağ}$) ve sol (t_{sol}) şeklinde kümelerden oluşan iki dal bulunur. Regresyon ağacı oluşturulmasında kullanılacak her bir veri sağ ve sol dala bölünmeye adaydır. Twoing kuralında öncelikle her bir aday için sağ ve sol taraftaki dalda olma olasılıkları hesaplanır. Her bir aday verinin sol taraftaki dala bölünme olasılığı p_{sol} ve $P(j/t_{sol})$, sağ taraftaki dala bölünmesi olasılığı ise $p_{sağ}$ ve $P(j/t_{sağ})$ şeklinde ifade edilir. Olasılıkların hesaplanmasının ardından *t* düğümündeki *s* aday bölünmelerinin uygunluk ölçüsü,

şeklinde hesaplanır. Bu eşitlikte *j* özniteliklere ait sınıf değerini göstermektedir. Hesaplama sonucu elde edilen değerler içerisinden en büyük olanı seçilir ve bu değere karşılık ilgili aday bölünme satırı dallanmayı oluşturacak satır olarak belirlenir. Dallanma bu şekilde yapılarak regresyon ağacının ilk ikili dallanması gerçekleştirilir. Ağacın aşağıya doğru ikili dallanmalarını gerçekleştirmek için alt kümelere söz konusu işlemler tekrar uygulanır (Larose, 2005).

3.5. Geçerlilik Analizi

Bir modelin doğruluğunun test edilmesinde kullanılan en basit yöntem basit geçerlilik analizidir. Bu analiz tipik olarak verilerin % 5 ile % 33 arasındaki bir kısmı test verileri olarak ayrılır ve kalan kısım üzerinde modelin öğrenimi gerçekleştirildikten sonra, bu veriler üzerinde test işlemi yapılır. Sınırlı miktarda veriye sahip olunması durumunda, kullanılabilecek diğer bir analiz yöntem, çapraz geçerlilik testidir. Bu yöntemde veri kümesi rastgele iki eşit parçaya ayrılır. İlk aşamada bir parça üzerinde model eğitimi ve diğer parça üzerinde test işlemi; ikinci aşamada ise ikinci parça üzerinde model eğitimi ve birinci parça üzerinde test işlemi yapılarak elde edilen hata oranlarının ortalaması kullanılır. Bir kaç bin veya daha az satırdan meydana gelen küçük veri tabanlarında, verilerin *k* gruba ayrıldığı *k* katlı çapraz geçerlilik testi tercih edilebilir. Verilerin örneğin 5 gruba ayrıldığı bu yöntemde, ilk aşamada birinci grup test, diğer gruplar öğrenim için kullanılır. Bu süreç her defasında bir grubun test, diğer grupların öğrenim amaçlı kullanılması ile sürdürülür. Sonuçta elde edilen on hata oranının ortalaması, kurulan modelin tahmini hata oranı olacaktır (Esen ve diğerleri, 2008; Esen ve diğerleri, 2009).

3.6. Başarım Değerlendirmesi

Başarım değerlendirmesi için, global istatistiki yöntemlerden korelasyon katsayısı (R) ve ortalama karesel hata fonksiyonları (OKH) kullanılmıştır. Ancak bu yöntemler hatanın dağılımı hakkında herhangi bir bilgi vermemektedirler. Bu nedenle bu çalışma için diğer global metotlara ilaveten modelin performansını daha etkili değerlendirmek için ortalama mutlak hata (OMH) yöntemi de kullanılmıştır. Korelasyon katsayısının karesine determinasyon katsayısı denmektedir. Determinasyon katsayısının 1'e yakın olması durumu X ve Y değişkenlerinin arasında doğrusal bağımlılığın kuvvetlendiğini göstermektedir. OKH, gerçek veri ile tahmin edilen veri değerlerinin farkının toplanıp, toplam veri sayısına bölünmesiyle elde edilen değerdir. Bu değerin sıfıra yakın olması, tahmin edilen değerin kuvvetli biçimde doğruya yakınsadığını göstermektedir. OMH ise gerçek veri ve tahmin edilen veri değerlerinin farkının, gözlenen değere bölündükten sonra her bir sonuç için yüzde olarak toplanmasıyla elde edilen değerdir (Esen ve diğerleri, 2008).

3.7. Uygulama

Bu tezin temel amacı, öğrencilerin mezuniyet notlarını erken tahmin edebilecek bir veri madenciliği uygulamasının gerçekleştirilmesidir. Böylece, mezun olamayacak öğrenciler uyarılabilecek veya ortalaması belirli bir değerin altında kalan öğrencilerin daha yoğun çalışmaları önerilebilecektir. Bu bağlamda Bölüm 3.1'de oluşturulan veri kullanılarak iki farklı senaryo MATLAB ortamında gerçekleştirilmiştir. Bunların ilkinde, öğrencilerin sadece ilk iki yılda aldıkları yılsonu notları göz önüne alınmıştır. Böylece toplam 24 adet dersin yılsonu notlarından öğrencilerin mezuniyet notları tahmin edilmiştir. İkinci senaryoda ise öğrencilerin ilk üç yılsonunda almış oldukları derslerin yılsonu notları kullanılarak, öğrencilerin mezuniyet notları tahmin edilmiştir. İlk üç sene sonunda öğrenciler toplam 38 adet ders almışlardır.

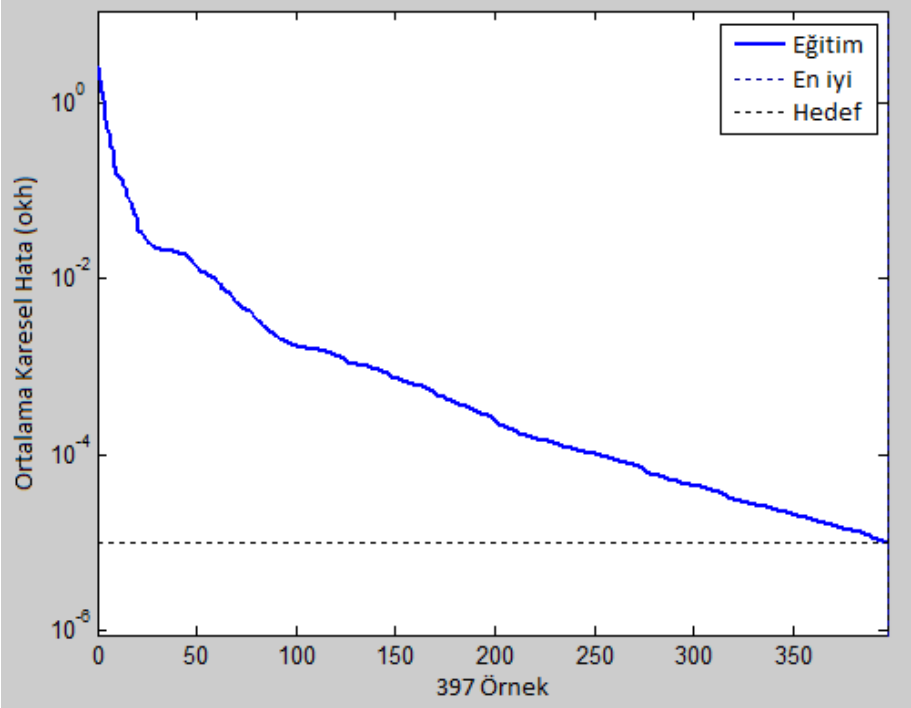
Daha önceki bölümlerde de bahsedildiği üzere, bu tez çalışmasında veri madenciliği yöntemlerinden olan YSA ve KA kullanılarak öğrencilerin mezuniyet notları tahmin edilmiştir. Bu bağlamda kullanılan bu yöntemlerin parametrelerinin ayarlanması gerçekleştirilen bilgisayar benzetimleri ile sağlanmıştır. YSA modeli, tanjant sigmoid aktivasyon fonksiyonu kullanan tek gizli katman içermektedir. Birinci senaryo için YSA'nın giriş katmanında 24 hücre ve benzer şekilde ikinci senaryo için ise giriş katmanında 38 hücre bulunmaktadır. Her iki senaryo için çıkış katmanında tek bir hücre bulunmaktadır. Çıkış katmanında lineer aktivasyon fonksiyonu ve giriş katmanında ise tanjant sigmoid aktivasyon fonksiyonu kullanılmıştır. Gizli katmanda ise birinci ve ikinci senaryo için sırası ile 25 ve 39 hücre kullanılmıştır.

Bilgisayar benzetimlerinde 5 katlı çapraz geçerlilik kullanılmıştır. Böylece, YSA'nın eğitimi için yaklaşık 101 örnek ve eğitilen YSA'nın testi için de 26 örnek kullanılmıştır. Tablo 2'de 5 katlı çapraz geçerlilik kullanılarak elde edilen ortalama sonuçlar verilmiştir.

Tablo 2. YSA kullanılarak birinci senaryo için elde edilen başarım değerleri

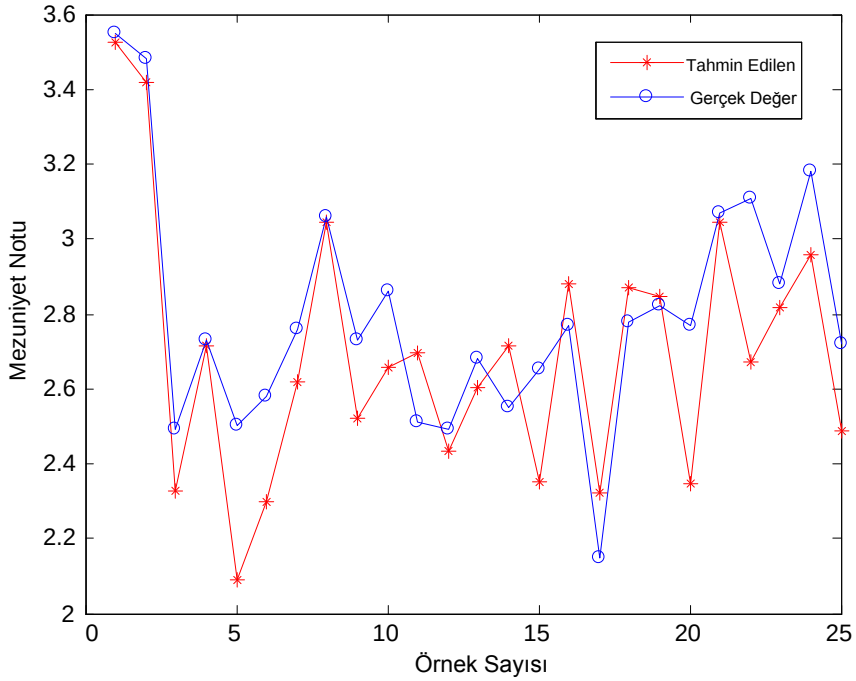
Kullanılan Yöntem	OKH	Korelasyon Katsayısı	OMH
YSA	0.2068	0.8494	7.4005

Tablo 2'de de gösterildiği gibi birinci senaryo için ortalama 0.2068 OKH değeri, 0.8494 korelasyon katsayısı ve 7.4005 OMH değerleri elde edilmiştir.



Şekil 9. Birinci senaryo için YSA eğitim başarımı

Şekil 9'da YSA modelinin eğitim başarımı gösterilmiştir. YSA modeli 10^{-5} hata değerine 397 iterasyon sonucunda varmıştır. Şekil 10'da gösterilen sürekli eğri, YSA'nın öğrenme başarımını, kesikli gösterilen seviye ise hedef hata değerini göstermektedir.



Şekil 10. Birinci senaryo için YSA tahmin sonuçları

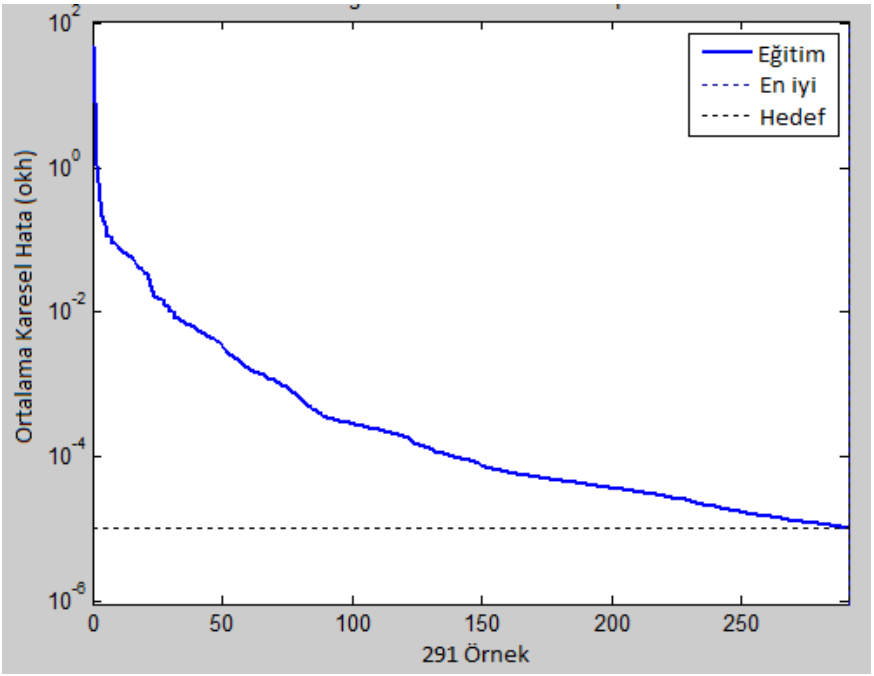
Şekil 10'da YSA modelinin 25 örnek için ürettiği tahmin sonuçları ve gerçek değerler gösterilmiştir. Burada -* tahmin edilen değerleri, -o ise gerçek değerleri göstermektedir. Şekil 10 dikkatle incelendiğinde 5, 6, 20 ve 22. örnekler dışında diğer örnekler için tahmin edilen değer gerçek değere oldukça yakındır.

İkinci senaryo için elde edilen başarım değerleri Tablo 3'de verilmiştir. 0.1329 OKH, 0.9376 korelasyon katsayısı ve 4.7547 OMH değerleri ikinci senaryo için elde edilmiştir. Bu değerler gerçekleştirilen modellemenin birinci senaryoya göre daha gerçekçi olduğunu göstermektedir.

Tablo 3. YSA kullanılarak ikinci senaryo için elde edilen başarım değerleri

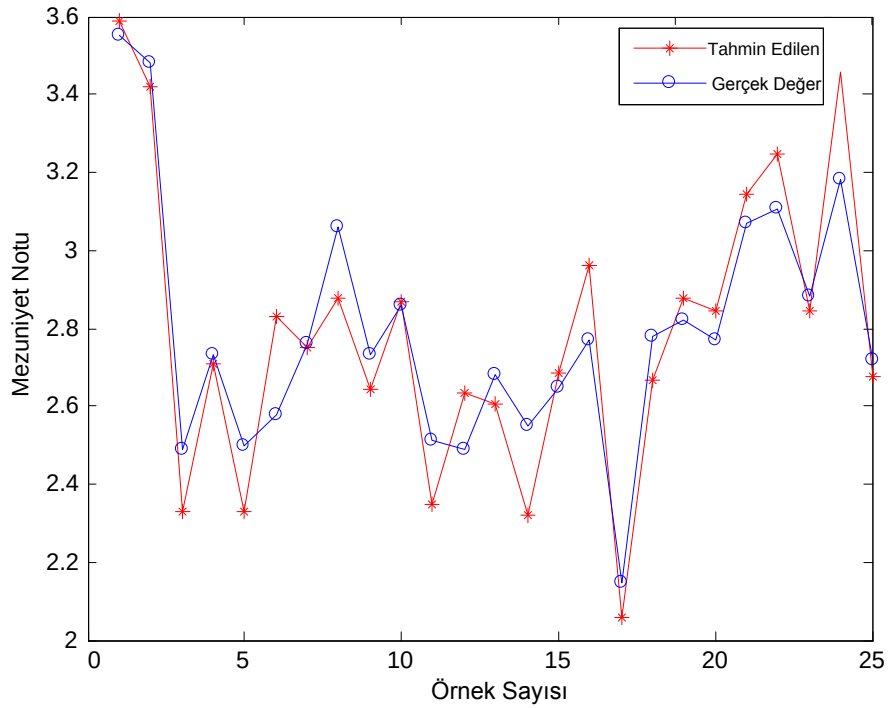
Kullanılan Yöntem	OKH	Korelasyon katsayısı	OMH
YSA	0.1329	0.9376	4.7547

Diğer taraftan, ikinci senaryo için elde edilen YSA eğitim başarımı Şekil 11'de gösterilmiştir. Bu grafikten, ikinci senaryo için YSA'nın hedeflenen hata değerine daha kısa bir zaman da (291 iterasyon) ulaştığı görülmektedir.



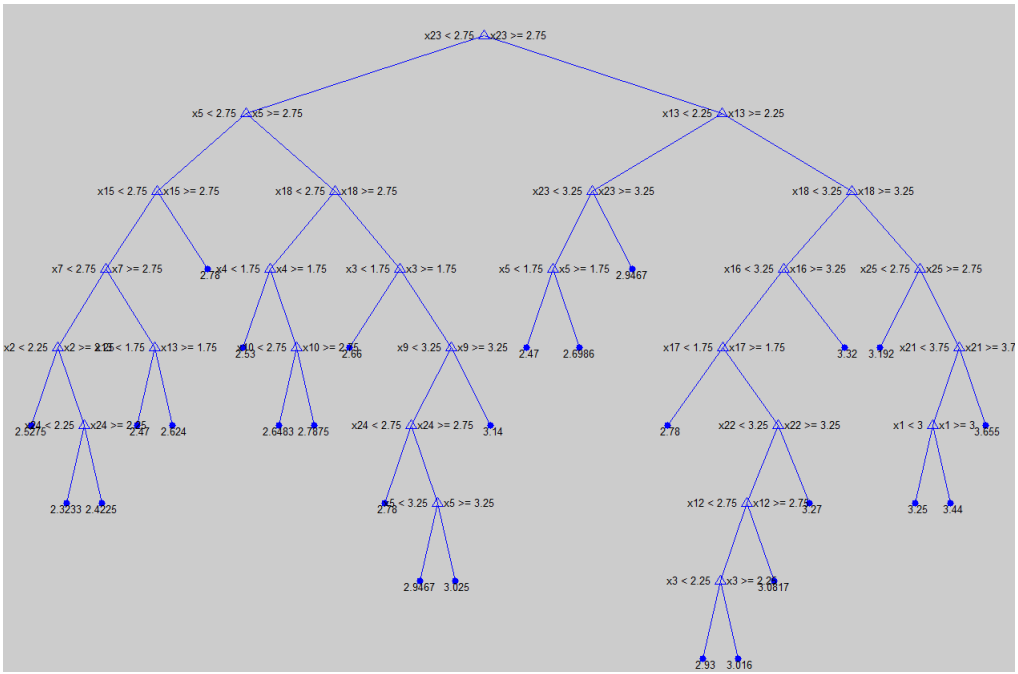
Şekil 11. İkinci senaryo için YSA eğitim başarımı

Şekil 11'de ikinci senaryo için YSA'nın eğitim başarımı ve Şekil 12'de ise gerçek ve tahmin edilen değerler verilmiştir. Tablo 3 ve Şekil 12 incelendiğinde ikinci senaryo için gerçekleştirilen tahminlerin daha iyi olduğu görülmektedir.



Şekil 12. İkinci senaryo için YSA tahmin sonuçları

Karar ağaçları ile regresyon işlemi için tek değişkenli karar ağacı algoritmalarından CART algoritması kullanılmıştır. Her iki senaryo içinde aynı ağaç yapısı kullanılmıştır. Regresyon ağaç modeli oluşturulmasında dallanmaya esas olacak özniteliklerin seçiminde Twoing algoritması, oluşturulan karar ağacının sadeleştirilmesi amacıyla ön budama yöntemi kullanılmıştır.



Şekil 13. Birinci senaryo için kullanılan regrasyon ağacı modeli

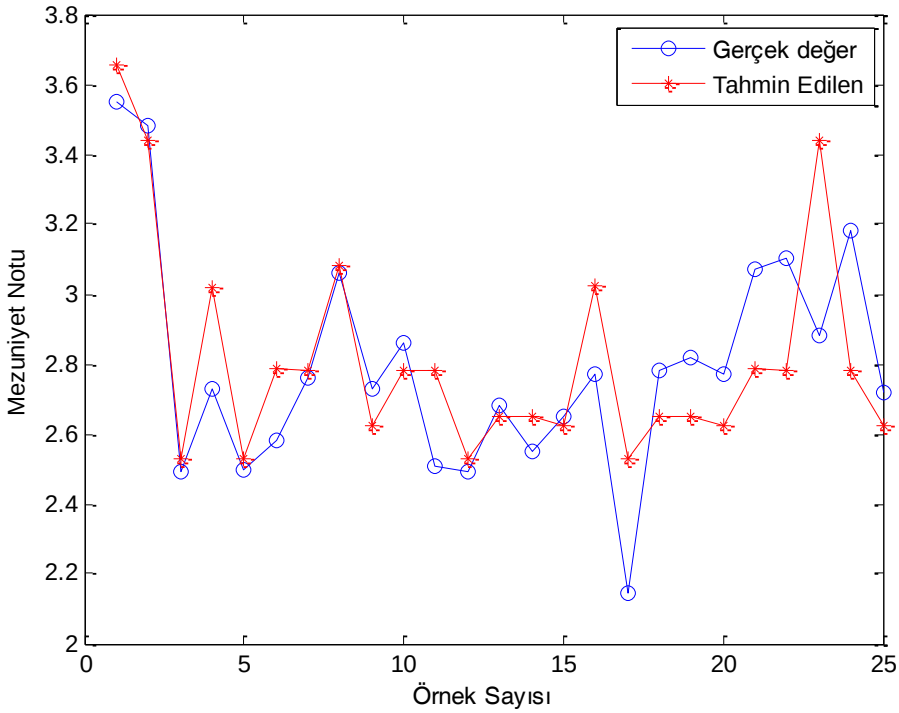
Söz konusu regresyon ağacı modeli Şekil 13'de gösterilmiştir. Şekilde 25 adet ders için elde edilen ağaçta 26 düğüm ve 27 yapraktan oluştuğu görülmektedir. Şekilde dallar üzerinde bulunan değerler regresyon ağaç yapısının dallanmasında kullanılan eşik değerlerini ifade etmektedir. İlk senaryo için 23 nolu ders kök düğüm olarak belirlenmiştir. Daha sonraki dallanmalar 5 ve 13 numaraları dersler üzerinden devam etmiştir. Modelin yapraklarını ise 24, 10, 5, 3 ve 1 numaralı dersler oluşturmaktadır.

Tablo 4. KA kullanılarak birinci senaryo için elde edilen başarım değerleri

Kullanılan Yöntem	**OKH**	**Korelasyon katsayısı**	**OMH**
Karar Ağaçları	0.2180	0.7444	7.7480

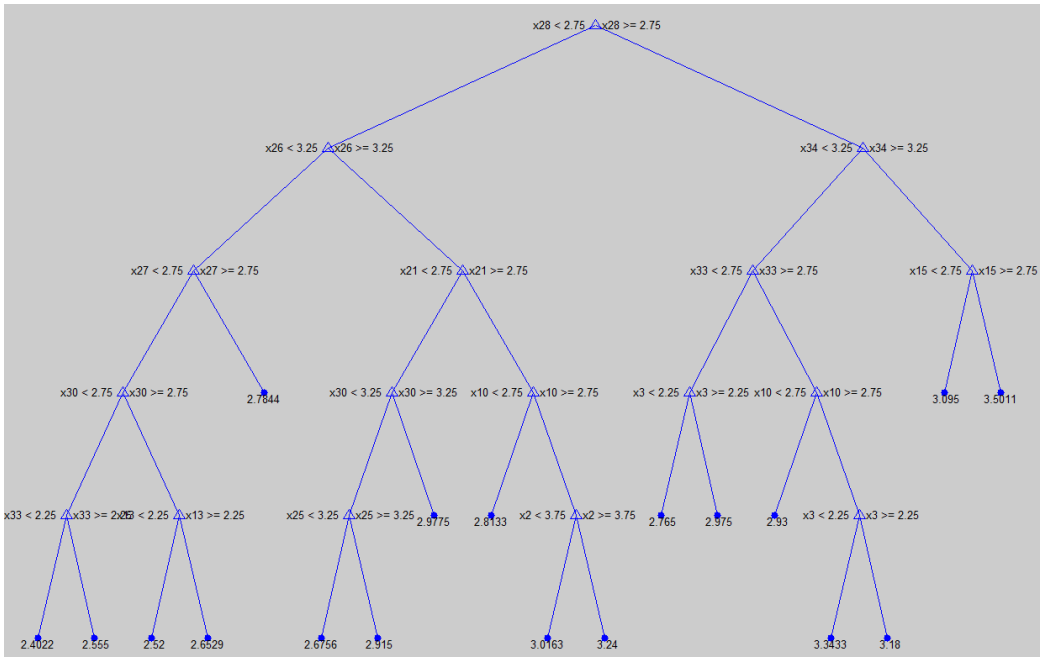
Tablo 4'de de gösterildiği gibi karar ağaçları kullanılarak gerçekleştirilen benzetim çalışmalarında birinci senaryo için ortalama 0.2180 OKH değeri, 0.7444

korelasyon katsayısı ve 7.7480 OMH değerleri elde edilmiştir. Diğer taraftan Şekil 14'de Karar Ağaçları modelinin 25 örnek için ürettiği tahmin sonuçları ve gerçek değerler gösterilmiştir. Burada yine, -* tahmin edilen değerleri, -o ise gerçek değerleri göstermektedir. Şekil 14 dikkatle incelendiğinde 17, 22, 23 ve 24. örnekler dışında diğer örnekler için tahmin edilen değer gerçek değere yakın oldukları görülmektedir.



Şekil 14. Birinci senaryo için KA tahmin sonuçları

Benzer şekilde ikinci senaryo için elde edilen Karar ağacı yapısı Şekil 15'de gösterilmiştir. Şekil 15'de 38 adet ders için elde edilen ağaçta 17 adet düğüm ve 18 adet yapraktan oluştuğu görülmektedir.



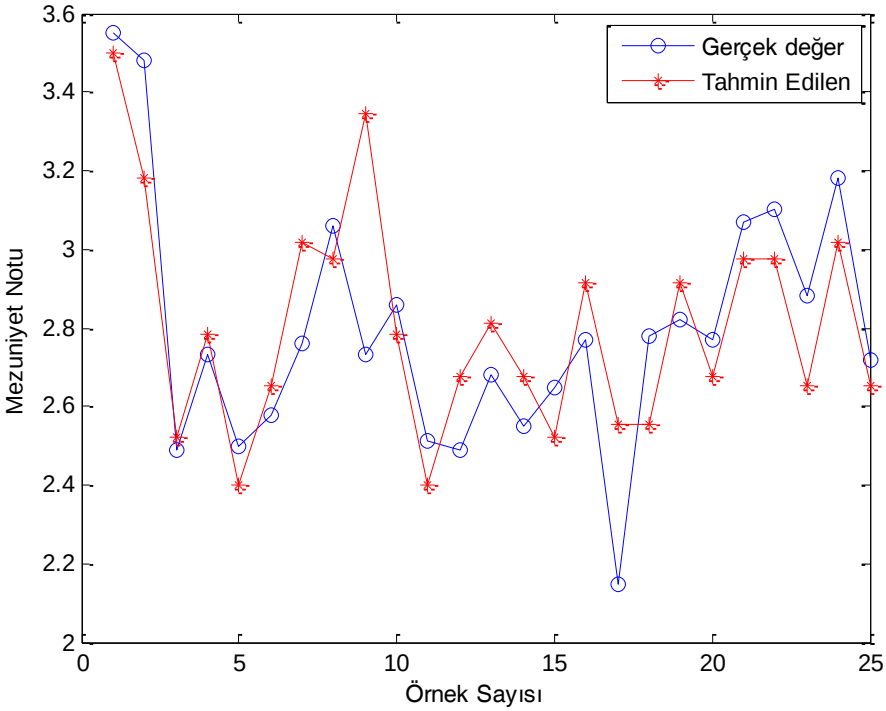
Şekil 15. İkinci senaryo kullanılan regresyon ağacı modeli

Ayrıca 28 numaralı dersin kök düğüm olarak belirlendiği ve ilgili dallanmaların da 26 ve 34 numaralı dersler üzerinden devam ettiği görülmektedir. Ağacın yapraklarında ise sırası ile 33, 13, 25, 2 ve 3 numaralı dersler bulunmaktadır.

Tablo 5. KA kullanılarak ikinci senaryo için elde edilen başarım değerleri

Kullanılan Yöntem	**OKH**	**Korelasyon katsayısı**	**OMH**
Karar Ağaçları	0.2026	0.7634	7.2228

Tablo 5'de de gösterildiği gibi karar ağaçları kullanılarak gerçekleştirilen benzetim çalışmalarında birinci senaryo için ortalama 0.2026 OKH değeri, 0.7634 korelasyon katsayısı ve 7.2228 OMH değerleri elde edilmiştir. Diğer taraftan Şekil 16'da Karar Ağaçları modelinin 25 örnek için ürettiği tahmin sonuçları ve gerçek değerler gösterilmiştir.



Şekil 16. İkinci senaryo için KA tahmin sonuçları

Benzer şekilde, Şekil 16 incelendiğinde 9 ve 17. örnekler dışında diğer örnekler için tahmin edilen değer gerçek değere yakın oldukları görülmektedir.

DÖRDÜNCÜ BÖLÜM

IV. TARTIŞMA ve ÖNERİLER

Bu tezin temel amacı, YSA ve KA gibi veri madenciliği yöntemleri kullanılarak Fırat Üniversitesi, Eğitim Fakültesi, BÖTE bölümü öğrencilerinin mezuniyet notlarının erken tahmin edilmesini gerçekleştirmektir. Böylece mezuniyet notları belirli bir değerin altında kalacak öğrenciler uyarılabilecek ve öğrencinin başarısı artırılabilecektir. Mezuniyet notunun tahmini için iki farklı senaryo denenmiştir. İlk senaryoda, öğrencilerin sadece birinci ve ikinci sınıfa ait derslerinin yılsonu notları kullanılarak mezuniyet notu tahmin edilmiştir. İkinci senaryo da ise ilk üç sınıf notları kullanılarak mezuniyet notlarının tahmini gerçekleştirilmiştir. Elde edilen çıkarımlar aşağıda sunulmuştur;

1-) Her iki senaryo ve her iki veri madenciliği yöntemi ile de belirli bir tahmin başarımı elde edilmiştir. Elde edilen başarım hem rakamsal hem de görsel sonuçlarla desteklenmiştir.

2-) Gerçekleştirilen benzetim çalışmalarında YSA'nın her iki senaryo için de karar ağaçları yönteminden daha iyi tahmin başarımı elde ettiği görülmüştür.

3-) Her iki tahmin yöntemi içinde, ikinci senaryonun birinci senaryoya oranla daha iyi tahmin gücüne sahip olduğu görülmüştür. Burada her iki tahmin yöntemine de giriş olarak verilen ders sayısının artmasının etkili olduğu anlaşılmaktadır.

4-) Her iki tahmin yönteminin ilgili parametrelerin ayarlanması için birçok deneme yapılması gerekmiştir. Özellikle YSA modeli için, gizli katman hücre sayısı, öğrenme oranı ve yöntemi gibi önemli parametrelerin, iyi bir başarım için ayarlanması gerekmektedir. Benzer şekilde karar ağaçları yöntemi için de yine bazı parametrelerin uygun seçilmesi gerekmektedir.

4.1. Öneriler ve gelecek çalışmalar

1-) Bu tez çalışmasında elde edilen başarımın daha da artırılması için öncelikle daha farklı veri madenciliği yöntemleri kullanılabilir. Özellikle son zamanlarda popüler olan Destek Vektör Makineleri, Bulanık mantık tabanlı regresyon yöntemleri ve istatistiksel bazı modeller kullanılabilir.

2-) Önerilen senaryolar farklı bölümler için de kullanılarak, böyle bir tahminin yapılabileceği yani bir genellemenin olabileceği gösterilebilir.

3-) Giriş öznitelik vektörünün normalizasyonu sağlanarak başarım değerlendirilmesi gerçekleştirilebilir.

KAYNAKLAR

Agrawal, R. and Srikant, R. (1994). Fast Algorithms for Mining Association Rules, *Proc. of the 20th Int'l Conference on Very Large Databases,* Santiago, Chile.

Agrawal, R., Imielinski, T., Swami, A. (1993). Mining association rules between sets of items in large databases, *ACM SIGMOD Conf. Management of Data.*

Almuallim, H. and Dietterich, T. (1991). Learning with many irrelevant features, *Proceeding of AAAI 91*, (Menlo Park, CA), 547-552, AAAI Press.

Al-Radaideh Q., Al-Shawakfa e. M. and Al-Najjar M. I. (2006). Mining Student Data Using Decision Trees, *International Arab Conference on Information Technology (ACIT'2006)*

Ayesha S.,Tasleem M.,Sattar A.R., Khan M.I. (2010). Data mining model for higher education system, *Europen Journal of Scientific Research*, 43(1), 24-29.

Beikzadeh, M. and Delavari, N. (2005). A New Analysis Model for Data Mining Processes in Higher Educational Systems, *On the proceedings of the 6th Information Technology Based Higher Education and Training*, 7-9.

Ben-Zadok G., Hershkovitz, A., Mintz, R. and Nachmias, R. (2007). Examining online learning processes based on log files analysis: a case study. *Research, Reflection and Innovations in Integrating ICT in Education*.

Bharadwaj, B.K. and Pal, S. (2011). Mining Educational Data to Analyze Students' Performance, *International Journal of Advance Computer Science and Applications (IJACSA)*, 2(6), 63-69.

Bharadwaj, B.K. and Pal, S. (2011). Data Mining: A prediction for performance improvement using classification. *International Journal of Computer Science and Information Security (IJCSIS)*, 9(4), 136-140.

Biçer, P. (2002). *Veri madenciliği: Sınıflandırma ve tahmin yöntemlerini kullanarak bir uygulama*, Yıldız Tek. Ünv. Sosyal Bil. Ens. Yüksek Lisans Tezi, İstanbul.

Bozkir, A.S., Mazman, S.G., Sezer, E.A. (2010). Identification of User Patterns in Social Networks by Data Mining Techniques: Facebook Case, *IMCW2010. Number 96 in CCIS (Communications in computer and information science)*, 145-153

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). Classification and Regression Trees, Monterey, CA: Wadsworth.

Bresfelean P.,Bresfelean M.,Ghisoiu N. (2008). Determining Students' Academic Failure Profile Founded on Data Mining Methods, *Proceedings of the ITI 2008 30th International Conference on Information Technology Interfaces*, 23-26.

Chan K.C.C., Wong, A.K.C. (1991). A statistical technique for extracting classificatory knowledge from databases, *Knowledge Discovery in Databases* (G. Piatetsky-Shapiro and W. J. Frawley, eds.), 107-123, Cambridge, MA: AAAI/MIT.

Chao R.M.,Huang S.Y., Chang J. (2010). Applying data mining and fuzzy technology on learning material recommendation mechanism, *International Journal of Business, Management and Social Sciences*, 1(1), 1-8.

Delavari N., Phon-amnuaisuk S. and Beikzadeh M. (2008). Data mining application in higher learning institutions, *Int. Educ. J.*, 7(4), 31-53.

Ding, Q. (2001). *Association rule mining survey*. Springer-Verlag Berlin Heidelberg.

Duda R.O., Hart P.E. (1989). *Pattern Classification and Scene Analysis*, John Wiley and sons.

Erdoğan Ş., Timor M. (2005). A Data Mining Application in a Student Database, *Havacılık ve Uzay Dergisi*, 2(2), 57-64.

Esen, H., Inalli, M., Sengur, A., Esen, M. (2008). Forecasting of a ground-coupled heat pump performance using neural networks with statistical data weighting pre-processing, *Int. J. Thermal Sciences*, 47(4), 431-41.

Esen, H., Ozgen, F., Esen, M. and Sengur, A., (2009). Modelling of a new solar air heater through least-squares support vector machines, *Expert Systems with Applications*, 36(7), 10673-10682.

Everitt, B. (1993). *Cluster Analysis for Applications*, Academic Press, New York.

Gaafar L. and Khamis M, (2009), Applications of Data Mining for Educational Decision Support, *Proceedings of the 2009 Industrial Engineering Research Conference*, 228-233

Goldberg, D.E., (1989), *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley Publishing Company Inc.ISBN:0-201-15767-5.

Grzymala-Busse, J.W., (1991), On the unknown attribute values in learning from examples, *Proceeding of Methodologies for Intelligent Systems*, Z. W. Ras and M. Zemankowa, (eds.), Lecture Notes in AI, 542, 368-377, New York: Springer-Verlag.

Han, J., Kamber , M. (2006), *Data Mining: Concepts and Techniques*, The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers.

Jadrić M., Garača Z., Ćukušić M. (2010). Student dropout analysis with application of data mining methods, *Management*, 15(1), 31-46

Karabatak M. (2008). *Özellik Seçimi, Sınıflama ve Öngörü Uygulamalarına Yönelik Birliktelik Kuralı Çıkarımı ve Yazılım Geliştirilmesi*, F.Ü. Fen Bilimleri Enstitüsü, Doktora Tezi, Elazığ, 116s.

Karabatak, M., İnce, M.C. (2004). Apriori Algoritması ile Öğrenci Başarısı Analizi, *ELECO 2004*, 348-352.

Kavzoğlu, T., Şahin, E.K. ve Çölkesen, İ. (2012). Heyelan Duyarlılığının İncelenmesinde Regresyon Ağaçlarının Kullanımı: Trabzon Örneği, *Harita Dergisi*, 147, 21-33.

Kira, K. and Rendell, L. (1992). The feature selection problem: Traditional methods and a new algorithm, *Proc. of AAAI 92*, 129-134, AAAI Press.

Kotsiantis S.B., Patriarcheas K., NikXenos M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowl.-Based Syst*. 23(6), 529-535.

Kovacic, Z. J. (2010). Early prediction of student success: Mining student enrollment data, *Proceedings of Informing Science & IT Education Conference*.

Larose, D.T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*, John Wiley & Sons, Inc., New York.

Lawrence, R.L., Wright, A. (2001). Rule-Based Classification Systems Using Classification and Regression Tree (CART) Analysis, *Photogrammetric Engineering and Remote Sensing*, 67, 1137-1142.

Lee, S. K. (1992). An extended relational database model for uncertain and imprecise information, *Proceeding of the ISth VLDB conference*, 211-218.

Luan, J. (2002). Data Mining, Knowledge Management in Higher Education, Potential Applications, *42nd Associate of Institutional Research International Conference* Toronto, Canada.

Luan, J. (2002). *Data mining applications in higher education*, John Wiley and Sons, New York.

Luba, T., Lasocki, R. (1994). On unknown attribute values in functional dependencies, *Proc. of the International Workshop on Rough Sets and Soft Computing*, (San Jose, CA), 490-497.

Mamcenko, J., Sileikiene, I., Lieponiene, J., Kulvietiene, R. (2011). Analysis of E-Exam Data Using Data Mining Techniques. *In: Proc of 17th International Conference on Information and Software Technologies* (IT 2011), Kaunas, Lithuania, 215–219.

Mardikyan, S.,Badur B. (2011). Analyzing Teaching Performance of Instructors Using Data Mining Techniques, *Informatics in Education*, 10(2), 245–257.

Minaei-Bidgoli, B., Kashy, D. A., Kortmeyer, G., and Punch, W. F. (2003). Predicting student performance: An application of data mining methods with an educational web-based system. *In The proceedings of the 33rd ASEE/IEEE frontiers in education conference*. Boulder, CO.

Mingers, J. (1989). An Empirical Comparison of Pruning Methods for Decision Tree Induction, *Machine Learning*, 4, 227–243.

Nefeslioglu, H.A., Sezer, E., Gokceoglu, C., Bozkir, A.S., Duman, T.Y. (2010). Assessment of Landslide Susceptibility by Decision Trees in the Metropolitan Area of Istanbul, Turkey, *Mathematical Problems in Engineering*, doi:10.1155/2010/901095.

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

Pal, M., Mather, P.M. (2003). An Assessment of the Effectiveness of Decision Tree Methods for Land Cover Classification, *Remote Sensing of Environment*, 86, 554-565.

Quinlan, J. R. (1986). *The effect of noise on concept learning. In Machine Learning: In An Artificial Intelligence Aproach*, R. Michalski, J. Carbonell, and T. Mitchell, (eds.), 2, 149-166, SanMateo, CA: Morgan Kauffmann Inc.

Quinlan, J. R. (1986). Introduction of decision trees, *Machine Learning*, l, 81-106

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.

Ramaswami M. and Bhaskaran R. (2010). A CHAID based performance prediction model in educational data mining, *IJCSI International Journal of Computer Science Issues*, 7(1), 156-168.

Ramaswami, M., &Bhaskaran, R. (2009). A study on feature selection techniques in educational data mining, *Journal of Computing*, 1(1), 7–11.

Ranjan, J., Khalil, S. (2008). Conceptual Framework of Data Mining Process in Management Education in India: An Institutional Perspective, *Information Technology Journal. Asian Network for Scientific Computing*, 1(7), 16-23.

Rokach, L., Maimon, O. (2008). *Data Mining with Decision Trees: Theory and Applications*, Series in Machine Perception and Artificial Intelligence, World Scientific Publishing, Singapore.

Romero, C., Ventura, S., Salcines, E. (2008). Data mining in course management systems: Moodle case study and tutorial, *Computer & Education*, 51(1), 368-384.

Romero, C., Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005, *Expert Systems with Applications*, 33(1), 135–146.

Safavian, S.R., Landgrebe, D. (1991). A Survey of Decision Tree Classifier Methodology, *IEEE Transactions on Systems Man and Cybernetics*, 21, 660-674.

Sembiring, S.,Zarlis, M., Hartama, D.,Ramliana S, Wani, E. (2011). Prediction Of Student Academic Performance By An Application Of Data Mining Techniques, *International Conference on Management and Artificial Intelligence, IPEDR* vol.6, IACSIT Press, Bali, Indonesia.

Sen, B.,Ucar, E. and Delen, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach, *Expert Systems with Applications*, 39, 9468–9476.

Türkoğlu İ. (1996). *Yapay sinir ağları ile nesne tanıma*, F.Ü. Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Elazığ, 112s.

Vranić M.,Pintar D.,Skoćır Z. (2007). The Use of Data Mining in Education Environment, *ConTEL 2007* Zagrep, 243-251.

Zaïane O. R., Luo J. (2001). Web usage mining for a better web-based learning environment, *Conference on Advanced Technology for Education*, 60-64

Zhang Y.,Oussena S., Clark T. and Kim H. (2010). Use data mining to improve student retention in higher education - a case study. *In ICEIS 2010: Proceedings of the 12th International Conference on Enterprise Information Systems*, Volume 1: Databases and Information Systems Integration, pages 190-197. INSTICC, Funchal, Portugal.

ÖZGEÇMİŞ

08.06.1984 İskenderun doğumluyum. Orta öğretimimi Kırşehir-Kaman'da, lise eğitimimi ise Kırşehir Anadolu lisesinde tamamladım. 2006 yılında Fırat Üniversitesi, Teknik Eğitim Fakültesi, Elektronik ve Bilgisayar Eğitimi bölümünü tamamladım. 2007 yılından beri Milli Eğitim Bakanlığı bünyesinde Bilgisayar Öğretmeni olarak çalışmaktayım.