LİSE TÜRÜ VE LİSE MEZUNİYET BAŞARISININ, KAZANILAN FAKÜLTE İLE İLİŞKİSİNİN VERİ MADENCİLİĞİ TEKNİĞİ İLE ANALİZİ

Y.Ziya AYIK[*]
Abdulkadir ÖZDEMİR[**]
Uğur YAVUZ[***]

Özet: Kurumların veritabanı boyutlarının, her gün üretmekte oldukları bilgi miktarındaki büyük artışlar ile binlerce Terabaytı aştığını söyleyebiliriz. Buna paralel olarak bilgisayar teknolojisinin sürekli gelişiyor ve ucuzluyor olması miktarı gittikçe artan bu verinin saklanabilmesine imkân sağlamaktadır. Ancak söz konusu verilerin tek başlarına bir anlam oluşturmadıkları bilinmektedir. Veriler ancak belirli bir amaç doğrultusunda işlendikleri takdirde anlam kazanabilmektedirler. Günümüzde veritabanlarındaki anlamsız verilerden anlamlı ve kullanılabilir bilgiler elde etmede önemli bir tekniğin Veri Madenciliği olduğu genel kabul görmektedir.
Bu çalışmada, Atatürk Üniversitesi öğrencilerinin mezun oldukları lise türleri ve lise mezuniyet dereceleri ile kazandıkları fakülteler arasındaki ilişki, veri madenciliği teknikleri kullanılarak incelenmiştir. Elde edilen sonuçların, üniversitemizi sonraki yıllarda tercih edecek öğrenci profilinin belirlenmesine yardımcı olacağı düşüncesindeyiz.

Anahtar Kelimeler: Veri madenciliği, veritabanlarında bilgi keşfi, mineset

I. Giriş

Veri miktarında meydana gelen olağan üstü artış, bu verilerden nasıl yararlanılabileceği konusunu ön plana çıkarmıştır. Veri değerlendirmenin klasik yöntemleri veya geleneksel bilişim teknikleri ile bu kadar çok veriden anlamlı bilgilerin elde edilmesinin pek mümkün olamayacağı anlaşılmıştır. Bilişim teknolojilerinin gelişimi ve tahmin edilemeyecek oranda biriken ve derlenen bilgi dağının oluşmasının sonucu olarak, her alanda strateji geliştirme konusunda kurumları ve bireyleri desteklemek amacıyla Veri Madenciliği (VM) adlı bir teknik son yıllarda yaygın olarak uygulanmaya başlanmıştır. Veri Madenciliği Teknikleri, verinin yığın halde bulunduğu, akla gelebilecek bütün alanlarda gizli bilgilerin açığa çıkarılabilmesi ve gelecekteki eğilim ve davranış şekillerinin tahmin edilebilmesinde kullanılabilmektedir (Zhang Dongsong and Zhou Lina, 2004). Veri Madenciliği büyük miktardaki veri yığını içerisinden gelecekle ilgili tahmin yapmamızı sağlayacak, bağıntı ve kuralların bilgisayar programları kullanılarak aranmasıdır (Babadağ Kadir, 2006). Maliyetli ve zahmetli bir süreç olan veri toplama yatırımlarından en yüksek faydayı

[*] Y.Doç.Dr., Atatürk Üniversitesi, Erzurum MYO, Bilgisayar Teknolojileri ve Programları ABD
[**] Uzm.Dr., Atatürk Üniversitesi, Mühendislik Fakültesi, Elektrik-Elektronik Müh. ABD
[***] Doç.Dr., Atatürk Üniversitesi, İletişim Fakültesi, Gazetecilik ABD

sağlamak veri madenciliği ile mümkündür (Kecman Vojislav, 2001). Veri Madenciliği, insan davranışlarının önceden tahmin edilebilmesini sağlar. Örneğin önceden biliniyor olsaydı;

1.Hastanelere yapılan tedavi taleplerinin bölgelere, zamana ve ihtiyaca göre değerlendirilmesi, salgın hastalık riskinin ilk aşamada tespiti ve kaynak planlama açısından faydalı olmaz mıydı?

2.Kaçak enerji kullananların profillerini tespit eden bir model, olası kaçak enerji kullanıcılarını tahmin etmenizi sağlasa idi, düşük maliyet ile kaçaklarla etkin mücadele edilmez miydi?

3.Web sitenizi ziyaret eden kişiler, ilk birkaç klikten sonra ihtiyaçları doğrultusunda yönlendirilseler, kişilerin ihtiyaçları doğrultusunda içerik yönetimi yapabilseniz, e-devlet hedefleri doğru yönetilmez miydi?

Bütün bu soruların cevabı elbette "Evet"tir. Ancak bu, veriye uygulanacak doğru veri madenciliği modelleri sayesinde olabilecektir.

Veri madenciliği ile büyük veri yığınlarından oluşan veritabanı sistemleri içerisinde gizli kalmış bilgilerin çekilmesi sağlanır. Bu işlem, istatistik, matematik disiplinleri, modelleme teknikleri, veritabanı teknolojisi ve çeşitli bilgisayar programları kullanılarak yapılır.

II. Veri Tabanlarında Bilgi Keşfi

Veri madenciliğinin uygulanabilmesi için kullanılan algoritma bize Veritabanlarında Bilgi Keşfi sürecinin gerçekleşmesini sağlamaktadır. Bu süreç içerisinde modelin uygulanacağı verilerin özelliklerinin çok iyi bilinmesi gerekmektedir. Sürecin başarılı olabilmesi için şu adımların dikkatle uygulanması zorunludur (Chen Ming ve diğerleri, 1996).

A. Problemin Tanımlanması

Veri madenciliği çalışmalarında başarılı olmanın ilk şartı, uygulamanın hangi amaç için yapılacağının açık bir şekilde tanımlanmasıdır. İlgili işletme amacı, işletme problemi üzerine odaklanmış ve açık bir dille ifade edilmiş olmalıdır. Elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceği tanımlanmalıdır. Ayrıca, yanlış tahminlerde katlanılacak olan maliyetlere ve doğru tahminlerde kazanılacak faydalara ilişkin tahminlere de bu aşamada yer verilmelidir.

B. Verilerin Hazırlanması

Modelin kurulması aşamasında ortaya çıkacak sorunlar, bu aşamaya sık sık geri dönülmesine ve verilerin yeniden düzenlenmesine neden olacaktır. 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.

Verilerin hazırlanması aşaması kendi içerisinde toplama, değer biçme, birleştirme ve temizleme, seçme ve dönüştürme adımlarından meydana gelmektedir.

1. Toplama (Collection)

Tanımlanan problem için gerekli olduğu düşünülen verilerin ve bu verilerin toplanacağı veri kaynaklarının belirlenmesi adımıdır. Verilerin toplanmasında, kuruluşun kendi veri kaynaklarının dışında kalan kuruluşların veritabanlarından faydalanılabilir.

2. Değer Biçme (Assessment)

Veri madenciliğinde kullanılacak verilerin farklı kaynaklardan toplanması, doğal olarak veri uyumsuzluklarına neden olacaktır. Bu uyumsuzluklar verilerin farklı zamanlara ait olmaları, kodlama farklılıkları ve farklı ölçü birimlerinin kullanılması olabilir. Bu nedenlerle, iyi sonuç alınacak modeller ancak iyi verilerin üzerine kurulabileceği için, toplanan verilerin ne ölçüde uyumlu oldukları bu adımda incelenerek değerlendirilmelidir.

3. Birleştirme ve Temizleme (Consolidation and Cleaning)

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

4. Seçim (Selection)

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

5. Dönüştürme (Transformation)

Veritabanı veya veri ambarlarında özet veya birbirleriyle bağlantılı bulunan veriler, daha anlamlı bir yapıya dönüştürülürler. Örneğin bir uygulamada bir yapay sinir ağı 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.

C. 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 (Zaki Mohammed, 2003).

Model kuruluş süreci denetimli (Supervised) ve denetimsiz (Unsupervised) öğrenimin kullanıldığı modellere göre farklılık göstermektedir.

Örnekten öğrenme olarak da isimlendirilen denetimli öğrenimde, 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.

Denetimsiz öğrenmede, kümeleme analizinde olduğu gibi ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden hareket ederek sınıfların tanımlanması amaçlanmaktadır (www.sas.com/software)

D. Modelin Kullanılması ve Güncellenmesi

Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak kullanılabilir. Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında doğrudan kullanılabileceği gibi, promosyon planlaması simülasyonuna entegre edilebilir veya tahmin edilen envanter düzeyleri yeniden sipariş noktasının altına düştüğünde, otomatik olarak sipariş verilmesini sağlayacak bir uygulamanın içine gömülebilir.

Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirecektir.

III. Veri Madenciliği Modelleri

Veri Madenciliğinde kullanılan modeller, tahmin edici (Predictive) ve tanımlayıcı (Descriptive) olmak üzere iki ana başlık altında incelenmektedir (Bigus Joseph, 1996).

Tahmin edici modellerde, sonuçlar bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümelerinin sonuç değerlerinin tahmin edilmesi amaçlanmaktadır.

Tanımlayıcı modellerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır (Akpınar Haldun, 2000).

Veri Madenciliği modelleri fonksiyonlarına göre şu şekilde sınıflandırılırlar. Sınıflama ve Regresyon Modelleri tahmin edici, Kümeleme, Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler ise tanımlayıcı modellerdir.

A. Sınıflama (Classification) ve Regresyon (Regression) Modelleri

Tahmin etmede faydalanılan ve veri madenciliği teknikleri içerisinde en yaygın kullanıma sahip olan sınıflama ve regresyon modelleridir. Sınıflamada

tahmin edilen bağımlı değişken kategorik, Regresyonda ise süreklilik gösteren bir değere sahiptir. Sınıflama ve regresyon modellerinde karar ağaçları, yapay sinir ağları, genetik algoritmalar, K-en yakın komşu ve naïve-bayes gibi teknikler kullanılmaktadır.

1. Karar Ağaçları (Decision Trees)

Veri madenciliğinde karar ağaçları, kurulmasının ucuz olması, yorumlanmalarının kolay olması, veritabanı sistemleri ile kolayca entegre edilebilmeleri ve güvenilirliklerinin iyi olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahip tekniktir. Karar ağacı, adından da anlaşılacağı gibi bir ağaç görünümünde, tahmin edici bir tekniktir (Berry Michael and Linoff Gordon, 1999). Ağaç yapısı ile, kolay anlaşılabilen kurallar yaratabilen, bilgi teknolojileri işlemleri ile kolay entegre olabilen en popüler sınıflama tekniğidir (Curtarolo Stefano and Morgan Dane, 2003).

Karar ağacı karar düğümleri, dallar ve yapraklardan oluşur (Han Jiawei and Kamber Micheline, 2000). Karar düğümü, gerçekleştirilecek testi belirtir. Bu testin sonucu ağacın veri kaybetmeden dallara ayrılmasına neden olur. Her düğümde test ve dallara ayrılma işlemleri ardışık olarak gerçekleşir ve bu ayrılma işlemi üst seviyedeki ayrımlara bağımlıdır. Ağacın her bir dalı sınıflama işlemini tamamlamaya adaydır. Eğer bir dalın ucunda sınıflama işlemi gerçekleşemiyorsa, o dalın sonucunda bir karar düğümü oluşur. Ancak dalın sonunda belirli bir sınıf oluşuyorsa, o dalın sonunda yaprak vardır. Bu yaprak, veri üzerinde belirlenmek istenen sınıflardan biridir. Karar ağacı işlemi kök düğümünden başlar ve yukarıdan aşağıya doğru yaprağa ulaşana dek ardışık düğümleri takip ederek gerçekleşir.

2. Yapay Sinir Ağları (Artifical Neural Networks)

Yapay sinir ağları (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 İnsan beynine oranla henüz çok geride olduğunu söylemek yanlış olmaz (Edelstein Herbert, 1999).

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.

3. Genetik Algoritmalar (Genetic Algorithms)

Genetik algoritmalar, doğada gözlemlenen evrimsel sürece benzer bir şekilde çalışan arama ve eniyileme yöntemidir. Karmaşık çok boyutlu arama uzayında en iyinin hayatta kalması ilkesine göre bütünsel en iyi çözümü arar. Genetik algoritmalar problemlere tek bir çözüm üretmek yerine farklı çözümlerden oluşan bir çözüm kümesi üretir. Böylelikle, arama uzayında aynı anda birçok nokta değerlendirilmekte ve sonuçta bütünsel çözüme ulaşma olasılığı yükselmektedir. Çözüm kümesindeki çözümler birbirinden tamamen bağımsızdır. Her biri çok boyutlu uzay üzerinde bir vektördür.

Genetik algoritmalar problemlerin çözümü için evrimsel süreci bilgisayar ortamında taklit ederler. Diğer eniyileme yöntemlerinde olduğu gibi çözüm için tek bir yapının geliştirilmesi yerine, böyle yapılardan meydana gelen bir küme oluştururlar. Problem için olası pek çok çözümü temsil eden bu küme genetik algoritma terminolojisinde nüfus adını alır. Nüfuslar vektör veya birey adı verilen sayı dizilerinden oluşur. Birey içindeki her bir elemana gen adı verilir. Nüfustaki bireyler evrimsel süreç içinde genetik algoritma işlemcileri tarafından belirlenir.

4. K-En Yakın Komşu (K-Nearest Neighbor)

Veri madenciliğinde sınıflama amacıyla kullanılan bir diğer teknik ise örnekseme yoluyla öğrenmeye dayanan k-en yakın komşu algoritmasıdır. Bu teknikte tüm örneklemler bir örüntü uzayında saklanır. Algoritma, bilinmeyen bir örneklemin hangi sınıfa dahil olduğunu belirlemek için örüntü uzayını araştırarak bilinmeyen örnekleme en yakın olan k örneklemi bulur. Yakınlık Öklid uzaklığı ile tanımlanır. Daha sonra, bilinmeyen örneklem, k en yakın komşu içinden en çok benzediği sınıfa atanır. K-en yakın komşu algoritması, aynı zamanda, bilinmeyen örneklem için bir gerçek değerin tahmininde de kullanılabilir (www.spss.com)

5. Naïve-Bayes

Naive bayes algoritmasında her kriterin sonuca olan etkilerinin olasılık olarak hesaplanması temeline dayanmaktadır. 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 rüzgar durumu bilgileri de alınmış olsun. Biz bu bilgileri değerlendirdiğimizde varsayılan tahmin yöntemleri ile hava bugün rüzgarlı 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 rüzgar yok şeklinde bir bilgiyi verdiğimizde sistem eğitildiği daha önce gerçekleşmiş istatistiklerden faydalanarak tenis maçının oynanma ve oynanmama ihtimalini hesaplar ve bize tahminini bildirir.

B. Kümeleme Modelleri (Clustering)

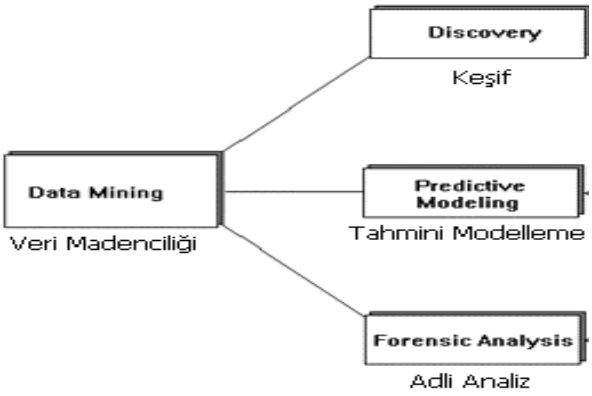
Kümeleme modellerinde amaç üyelerinin birbirlerine çok benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veritabanındaki kayıtların bu farklı kümelere bölünmesidir. Kümeleme analizinde; veritabanındaki kayıtların hangi kümelere ayrılacağı veya kümelemenin hangi değişken özelliklerine göre yapılacağı konunun uzmanı olan bir kişi tarafından belirtilebileceği gibi veritabanındaki kayıtların hangi kümelere ayrılacağını geliştirilen bilgisayar programları da yapabilmektedir.

C. Birliktelik Kuralı (Association Rule) ve Ardışık Zamanlı Örüntü (Sequential Pattern)

Bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal 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 sepeti analizi (Market Basket Analysis) adı altında veri madenciliğinde yaygın olarak kullanılmaktadır. Bununla birlikte bu teknikler, tıp, finans ve farklı olayların birbirleri ile ilişkili olduğunun belirlenmesi sonucunda değerli bilgi kazanımının söz konusu olduğu ortamlarda da önem taşımaktadır (Alpaydın Ethem, 2000).

IV. Veri Madenciliği İşlevi

Veri madenciliğine işlevleri açısından bakılacak olursa, veri madenciliği aktiviteleri 3 sınıf altında toplanmıştır. Keşif (discovery), tahmini modelleme (predictive modeling) ve adli analiz (forensic analysis) (Vahaplar Alper ve İnceoğlu Mustafa, 2001).



Şekil 1. Veri Madenciliği Aktiviteleri

Keşif, ne olabileceği konusunda önceden belirlenmiş bir fikir ya da hipotez olmadan, veritabanı içerisinde gizli desenleri arama işlemidir. Geniş veritabanlarında kullanıcının pratik olarak aklına gelmeyecek ve bulmak için gerekli doğru soruları bile düşünemeyeceği birçok gizli desen olabilir. Buradaki asıl amaç, bulunacak desenlerin zenginliği ve bunlardan çıkarılacak bilginin kalitesidir. Tahmini modellemede, veritabanından çıkarılan desenler, geleceği tahmin için kullanılır. Bu model, kullanıcının bazı alan bilgilerini bilmese bile kayıt etmesine izin verir. Sistem, bu boşlukları, önceki kayıtlara bakarak tahmin yoluyla doldurur. Keşif, verideki desenleri bulmaya yönelikken, tahmini modelleme, bu desenleri yeni veri nesnelerini bulmak için uygular. Adli analiz, normal olmayan ya da sıra dışı veri elemanlarını bulmak için, çıkarılmış desenleri uygulama işlemidir. Sıra dışı olanı bulmak için ilk önce sıradan kısmı tespit etmek gerekir (Glymour Clark and Madigan David, 1997).

V. Veri Madenciliği Uygulaması

Veritabanlarında Bilgi Keşfi ve Veri Madenciliği Tekniklerinin uygulama alanı yığın halde verilerin bulunduğu aklımıza gelebilecek bütün ortamlar olarak düşünülebilir. Elektronik Ticaret, Bankacılık, Sigortacılık, Endüstri, Ticaret, Sağlık, Eğitim ve Bilim başta olmak üzere daha birçok kamu ve özel sektör faaliyet alanı bu kapsamda kabul edilir (http://www.datamining.com).

Bu çalışmada Atatürk Üniversitesinden 1976 yılından itibaren mezun olan ve halen okumakta olan öğrenci bilgilerinin bulunduğu veritabanı üzerinde Veri Madenciliği teknikleri uygulanmıştır. Verilerin tanımlanması, anlaşılması ve hazırlanması çalışmanın başında oldukça önemli bir zaman almıştır. Veri hazırlama işlemi, toplama, birleştirme temizleme ve dönüştürme aşamalarından oluşmuştur. Daha sonra modelin kurulması aşamasına geçilmiştir. Bu sırada Veri madenciliği tekniğinin uygulanabilmesi için, bu amaca uygun Purple Insight MineSet™ 3.2 programı lisanlı olarak satın alınmıştır. Purple Insight MineSet 3.2 programı Windows tabanlıdır ve veri madenciliği ve görsel araçların bütünleşmesi uygundur. Ayrıca veri madenciliği süreci ve kapsamının hızlı ve doğru gerçekleşmesini sağlamaktadır. Programın kurulumu yapıldıktan sonra, problemin çözümü için en uygun model araştırılmaya başlanmıştır. En uygun modelin bulunabilmesi için veri madenciliği tekniğine uygun bütün modeller denenmiştir. Yapılan denemelerden veritabanımızda bulunan verilerin anlaşılabilir en uygun analizinin Sınıflama (Classification) modeli ile yapılabildiği gözlenmiştir. Atatürk Üniversitesi mezunlarına ait veritabanı verilerinin hazırlandıktan sonraki yapısı şu şekildedir.

Tablo 1. Veritabanı Alanları ve Kodları

Alan Adı	**Alan Kısa Adı**	**Kod Anlamı**
Lise Türü	LTUR	**1.** Fen **2.** Anadolu **3.** Anadolu Öğretmen **4.** Genel **5**. Meslek
Fakülte Adı	FA	İkt. ve İd. Bil. Tıp Mühendislik Eczacılık vb.
Lise Mez. Bşr.	LMZN	0-2, 2-2.5, 2.5-3, 3-3.5, 3.5-4, 4-4.5, 4.5-5
Rsmi-Özel	RO	**1** Resmi **2** Özel
Lise Bölge Kodu	LBLG	**1.** Marmara **2.** Ege **3.** Akdeniz **4.** İç Anadolu **5.** Karadeniz **6.** Doğu Anadolu **7.** Güney Doğu Anadolu
Cinsiyet	CNS	**1.** Erkek **2.** Kız

Model uygulanmaya hazır hale gelmiş bu veriler yaklaşık 50000 kayıttan oluşmaktadır. Bu çalışmada bir öğrencinin lisede okuduğu bölümün, yani lise türünün ve lisedeki başarı seviyesinin Üniversiteye girerken kazandığı fakülte ile bir ilişkisinin var olup olmadığının anlaşılabilmesi için veri madenciliği tekniği uygulanmıştır. Durum MineSet 3.2 programı yardımı ile değerlendirilmiş ve elde edilen sonuçlar aşağıda verilen birkaç örnek tablo yardımı ile anlatılmıştır.

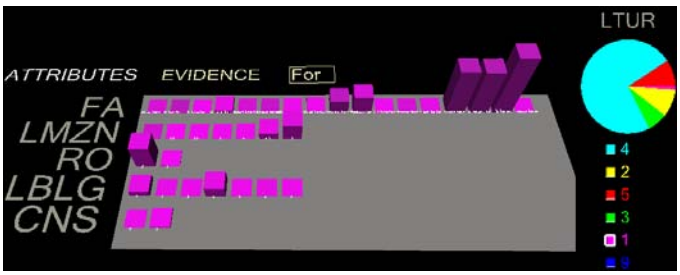
A. Lise Türünün Kazanılan Fakülte İle İlişkisi

Atatürk Üniversitesini kazanan öğrencilerin genel dağılımları incelendiğinde, öğrencilerin büyük bir kısmının genel liselerden geldiği, diğer lise türlerinden gelen öğrenci sayılarının ise birbirine yakın oldukları görülmektedir. Ayrıca öğrencilerin büyük bir çoğunluğu resmi liselerden mezun olduğu, cinsiyet olarak aralarında çok önemli bir farkın olmadığı, bölgeler itibariyle de en çok Doğu Anadolu Bölgesi olmak üzere, sırasıyla Karadeniz, Akdeniz ve İç Anadolu bölgesi ve diğer bölge liselerinden mezun öğrencilerin bulunduğu anlaşılmaktadır.

Şekil 2. Lise Türlerine Göre Kazanılan Fakültelerin Dağılımı

1. Fen Lisesi Mezunu Olmanın Kazanılan Fakülte İle İlişkisi

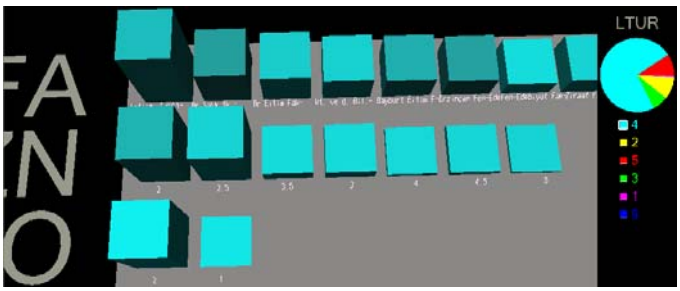
Tıp, Diş ve Eczacılık fakültelerine giren öğrencilerde fen Lisesi'nin ağırlığının diğer fakülteler göre daha fazla olduğu görülmektedir. Öğrencilerin çoğu özel fen liselerinden mezun olup, bölgeler itibariyle de en kalabalık grubu Doğu Anadolu Bölgesi mezunları oluşturmaktadır.



Şekil 3. Fen Lisesi Mezunlarına Göre Kazanılan Fakültelerin Dağılımı

2.Genel Lise Mezunu Olmanın Kazanılan Fakülte İle İlişkisi

Genel lise mezunlarının Fen-Edebiyat Fakültesi, Eğitim Fakültesi ve İktisadi ve İdari Bilimler Fakültelerinde oranlarının daha yüksek olduğu gözlenmektedir.



Şekil 4. Genel Lise Mezunlarına Göre Kazanılan Fakültelerin Dağılımı

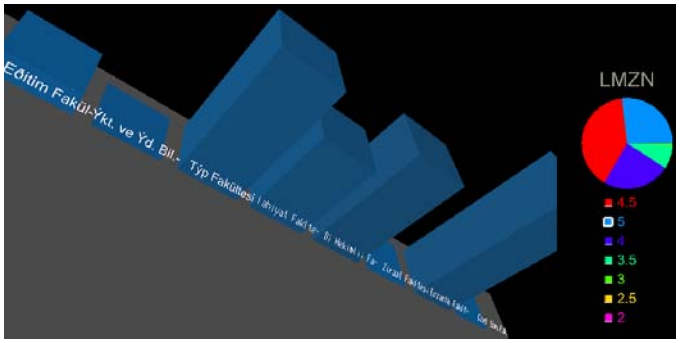
B. Lise Başarısının Kazanılan Fakülte İle İlişkisi

Atatürk Üniversitesine ait fakülteleri kazanan öğrencilerin lise genel başarı ortalamasının 5 üzerinden 3.5 olduğu, daha sonra 4, 4.5 ve 5 ortalamalı öğrencilerin de sırasıyla önemli yer tutukları tespit edilmiştir.

Şekil 5. Lise Başarı Notlarına Göre Kazanılan Fakültelerin Dağılımı

1. Lise Mezuniyet Notlarının Yüksek Olması İle Kazanılan Fakülte İlişkisi

Lise mezuniyet notları 4.5 ve 5 olan öğrencilerin daha çok Tıp, Diş, Eczacılık ve İlahiyat Fakültelerini tercih ettikleri görülmektedir.



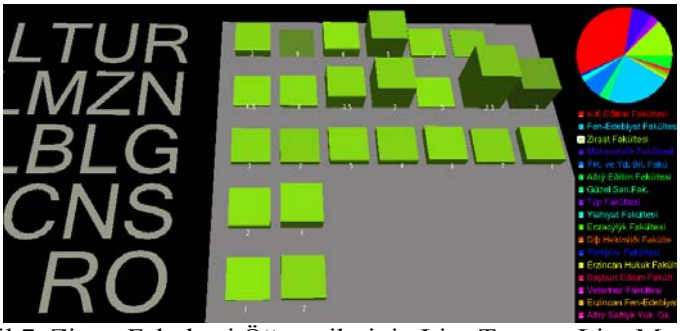
Şekil 6. Lise Mezuniyet Notlunun Yüksek Olması Durumuna Göre Kazandıkları Fakülte Dağılımı

C. Fakültelerde, Lise Türü ve Başarı Notuna Göre Dağılımlar

Fakülte Adı baz alınarak yapılan veri madenciliği uygulamasında, Lise Türü ve Lise Mezuniyet Notuna Göre dağılımlarda her fakülte için farklı sonuçların elde edildiği görülmüştür. Örnek olarak Ziraat Fakültesi ile Tıp fakültesinin dağılımları gösterilmiştir.

1. Ziraat Fakültesi Öğrencilerinin Lise Türü ve Lise Başarı Durumları

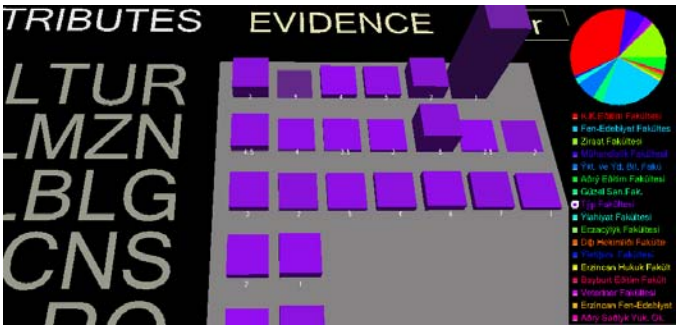
Ziraat Fakültesini kazanan öğrencilerin büyük oranda Meslek Liselerinden geldikleri ve en çok lise mezuniyet notu 2.5 olan öğrencilerden oluştukları gözlenmektedir.



Şekil 7. Ziraat Fakültesi Öğrencilerinin Lise Türü ve Lise Mezuniyet Notuna Göre Dağılımları

2. Tıp Fakültesi Öğrencilerinin Lise Türü ve Lise Başarı Durumları

Tıp Fakültesini kazanan öğrencilerin büyük oranda Fen Liselerinden geldikleri ve en çok lise mezuniyet notu 5 olan öğrencilerden oluştukları tespit edilmiştir.



Şekil 8. Tıp Fakültesi Öğrencilerinin Lise Türü ve Lise Mezuniyet Notuna Göre Dağılımları

VI. Sonuç

Veri miktarındaki baş döndürücü artışla, insan ömrünün bu veri ile mücadelede daha verimli harcanması gerektiği düşüncesi oluşmuştur. Veriler ön elemeden geçirilmeli ve gizli bilgiler mutlaka çıkarılmalıdır. Bu amaçla kullanılan Veri Madenciliği Teknikleri ve Veritabanlarında Bilgi Keşfi süreci son yıllarda oldukça olumlu sonuçlar vermiştir. Veri elde edildikten sonra tanımlanmakta, anlaşılmakta ve faydalı bir hale getirilmesi için hazırlanmaktadır. Daha sonra Veri madenciliği tekniği uygulanabilmesi için tekrar tekrar denemeler yapılarak veri yapısına uygun bir model belirlenmektedir (Jing Luan, 2002). Model değerlendirilerek kullanıma sunulmakta, böylece yığın haldeki verilerden daha anlamlı bilgiler daha kısa zamanda elde edilebilmektedir (http://www.dwinfocenter.org.)

Çalışmada, Atatürk Üniversitesi öğrencilerinin yıllar boyunca birikmiş karmaşık verileri üzerinde uzun bir hazırlık çalışması yapılarak, Lise Türü ve Lise Mezuniyet Başarısının, Kazanılan Fakülte üzerindeki öneminin tespit edilebilmesi amaçlanmıştır. Çalışma sonucunda, lise türünün arzu edilen bir fakültenin kazanılmasında çok büyük öneminin olduğu, yine lise başarısının da aynı derecede önemli olduğu tespit edilmiştir.

Abstract: It can be suggested that the enormous increase in the capacity of data of institutions has exceeded thousands of terabytes due to increase in the storage of data. In parallel to this increase, the advancement in the computer technology and the accessibility of these facilities to the larger public make it possible for the storage of ever increasing data. However, it is obvious that these data do not have any significance in themselves alone. They gain meaning if they are only mined in accordance with specific purpose. Today, data mining is generally accepted to be a very important technique in the process of obtaining sensible and meaningful results from arbitrarily accumulated data.

This study is intended to make an analysis of the correlation between the type of high school and high school graduation score and University Placement Examination using data mining technique. We believe that the results obtained from this investigation can be used in determining the profile of the students who would like to attend our institution in the future.

Key Words: Data mining, knowledge discovery in databases, mineset

Kaynakça

Akpınar Haldun, (2000), Veritabanlarında Bilgi Keşfi ve Veri Madenciliği, İ.Ü. İşletme Fakültesi Dergisi, c:29, 1-22

Alpaydın Ethem, (2000), Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri, Bilişim2000, 1-5

Babadağ Kadir, (2006), Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri, Industrial Application Software, 85-87

Berry Michael and Linoff Gordon, (1999), Mastering Data Mining: The Art and Science of Customer Relationship Management, John Wiley & Sons, 1st Ed.

Bigus Joseph, (1996), Data Mining With Neural Networks: Solving Business Problems from Application Development to Decision Support, McGrawHill Text.

Chen Ming, Hun Jiawei and Yu Philip, (1996), Data Mining: An Overwiev From a Database Perspective and Knowledge Discovery, IEEE Transaction on Knowledge and Data Engineering, Vol:8, No:6 866

Curtarolo Stefano and Morgan Dane, (2003), Predicting Crystal Structures with Data Mining of Quantum Calculations , Phys. Rev. Lett. 91, in pres.

Edelstein Herbert, (1999), Introduction to Data Mining and Knowledge Discovery, Two Crows Corporation.

Glymour Clark and Madigan David, (1997), Statistical Themes and Lessons for Data ining; Data Mining and Knowledge Discovery1,11-28

Han Jiawei and Kamber Micheline, (2000), Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, 1st Ed., San Francisco, USA

Jing Luan, (2002), Data Mining and Its Applications in Higher Education, New Dırectıons For Instıtutıonal Research, no. 113.

Kecman Vojislav, (2001), Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models, The MIT Press, Cambridge.

Vahaplar Alper ve İnceoğlu Mustafa, (2001), Veri Madenciliği ve Elektronik Ticaret, VII. Türkiye'de İnternet Konferansı, 1–3 Kasım 2001

Zaki Mohammed, (2003), Introduction to Data Mining, Springer Verlag.

Zhang Dongsong and Zhou Lina, (2004), Discovering Golden Nuggets: Data Mining in Financial Application, IEEE Transactions on Systems, Applications and Reviews, Vol: 34, No:4, 513-515

İnternet Kaynakları

Data Warehousing Information Center - http://www.dwinfocenter.org

Information Discovery Inc. – http://www.datamining.com

SPSS Inc. Extend Your Data Mining Capabilities with Advanced Analysis. www.spss.com

SAS Institute Inc. The Data Mining Challenge: Turning Raw Data Into Business Gold. www.sas.com/software