Uludağ Üniversitesi
İktisadi ve İdari Bilimler Fakültesi Dergisi
Cilt XXV, Sayı 1, 2006 s. 79-88

VERİ MADENCİLİĞİNDE KULLANILACAK VERİLERİN STANDARTLAŞTIRILMASI VE İYİLEŞTİRİLMESİ

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

Özet

Bilişim sistemlerindeki hızlı gelişim ile idari, endüstriyel ve akademik ortamlardaki veri toplama ve saklama kapasiteleri hem artmakta hem de bu verilerden daha çok yararlanma istek ve ihtiyaçları ortaya çıkmaktadır.

Bu ihtiyacı karşılayabilecek tekniklerden birisi olan "Veri Madenciliğinin" uygulanabilmesi için verilerin düzenlenmesi ve ayıklanması sonuç almak açısından çok önemli bir faktördür. Ancak ülkemiz gibi gelişmekte olan ülkelerde bilişim sitemlerinin kurulması ve kullanılması köklü bir geçmişe sahip olmaması, saklanan verilerin içerik ve yapılarında önemli farklılıkların oluşmasına yol açmıştır.

Çalışmamızda Atatürk Üniversitesi öğrenci verilerine, veri madenciliği teknikleri uygulanabilmesi için bahsedilen nedenlerden kaynaklanan veri uyumsuzluklarının giderilebilmesi amaçlanarak, veri madenciliği tekniklerinin uygulanabileceği standart veriler elde edilmesi amaçlanmıştır.

Anahtar Kelimeler: *Veri Madenciliği, Veri Hazırlama, Veri Standardizasyonu.*

Abstract

Result of the quick revolution in information technologies, both the data acquiring and storage capacities in management, industrial and academic media, and the requirements and requests of the use of these data are arisen.

---

[*] *Bozok Üniversitesi, Meslek Yüksekokulu, İktisadi ve İdari Programlar Bölümü.*
[**] *Atatürk Üniversitesi Erzurum Meslek Yüksekokulu, Bilgisayar Teknolojisi ve Programlama.*
[***] *Atatürk Üniversitesi İletişim Fakültesi, Gazetecilik Bölümü.*

To apply Data mining techniques to these requires to managing these requirements, data must be sorting and arranged. So it is possible to get quite results. But in developing countries like our country, information systems had no quite long history, so stored data have different structures and different contents.

In our study, it is aimed to apply data mining techniques to Ataturk University Student database. To achieve this aim, firstly it is intended to correct data inconsistency occurred from mentioned reasons.

Key Words: *Data mining, Data Preparation, Data Standardization.*

1. GİRİŞ

Bilgisayarlarda sabit disk kapasitesi ve işlemci hızı açısından hızlı bir gelişme süreci yaşanırken, maliyetler de gelişmeyle ters orantılı olarak düşmektedir. Bu düşüş, verilerin depolanma maliyetlerinin de düşmesine yol açmaktadır. Tüm bunların sonucu oluşan veri miktarlarındaki üssel artış, kendisini günlük alanda her an hissettirmektedir. Bu artış, bilişim çalışanları kadar, kurumların pazarlama, satın alma, karar destek gibi birim çalışanları tarafından da bir sorun olarak görülmekte ve veri kirliğinden kurtulmak için çeşitli çözümler aranmaktadır. Aynı zamanda hızlı bir rekabet ortamının yaşandığı çağımızda, kuruluşların öncelikle kendi verileri içerisinde saklı olan bilgileri/örüntüleri ortaya çıkarmak ve kullanmak, bu rekabette bir adım öne geçmeyi sağlayacak, çok önemli bir etken olarak ta görülmektedir.

Günümüzden 5-10 yıl önce sadece veri istatistik sonuçlar kuruluşlar için yeterli bilgi sağlarken, zaman içerisinde veritabanlarında saklı bilgilerin de olabileceği ortaya konulmaya başlandı ve veritabanlarında bilgi keşfi çokça başvurulan bir yöntem oldu. Veritabanlarında bilgi keşfi uygulamaları ile veri madenciliği kavramı da birçok uygulama alanı buldu (Kamrani 2001:361).

Veri madenciliği uygulamaları için temel özelliklerden birisi çok miktarda veri barındıran ortamların bulunması, ikincisi ise bu veriler içerisinden kullanılabilir ve anlamlı bilgileri çıkarılma ihtiyacının olmasıdır, şeklinde özetlenebilir. Veri madenciliği uygulamalarında sonuç almada çok önemli etkenlerden bekli de en önemlisi verilerin amaca göre birleştirilmesi, ayıklanması ve kirlilikten ayıklanmasıdır (Adriaans 1998).

Veri madenciliği için kullanılan verilerin farklı veritabanlarından, tablolardan ve tarihsel olarak farklı verilerden alınması dolayısı ile kullanılan verilerin farklı standartlarda olması, özellikle bilişim altyapısı hızla değişen Türkiye gibi ülkelerde, sıklıkla karşılaşılan bir durumdur. Bunun yanı sıra, eldeki verilerin amaca uygun olarak yeniden yapılandırılması da bir zorunluluktur. Dolayısıyla, veri madenciliği sürecinin yaklaşık %60'ını oluşturarak en önemli ve uzun evresi olan (Mlynarski 2006:273) veri

hazırlama evresinde standart olmayan veriler nedeniyle istenilen sonuçlar elde edilememekte ve veri madenciliğinin istenen hedefe ulaşması sekteye uğrayabilmektedir. Bu gibi olumsuzluklarla karşılaşmamak için standart olmayan verilerin veri hazırlamadan önce standardize edilmesi veri madenciliği sürecinin başarısı için önemlidir.

Atatürk Üniversitesi Öğrenci veritabanı kullanılarak bir veri ambarı oluşturulması amaçlanmıştır. Ancak bu veri ambarının oluşturulmasında veri yapılarının standart olmaması dolayısıyla çeşitli problemler ortaya çıkmıştır. Örneğin lise mezuniyet notları standart değildir. Mezuniyet tarihine bağlı olarak bazı yıllarda lise mezuniyet notu 5 üzerinde hesaplanırken, bazı yıllarda 10 üzerinden hesaplanmıştır. Günümüzde ise 100 üzerinden hesaplanmaktadır. Ayrıca veriler girilirken de standartlara uyulmamış, bazı veriler (örneğin mezuniyet tarihleri) yalnızca yıl olarak girilmişken, bazı veriler gün/ay/yıl olarak girilmiştir. Bu ve benzeri problemler giderilmeden veri ambarının oluşturulması doğru analizler yapılmasını engelleyecektir.

Bu çalışmada, Atatürk üniversitesi veri tabanında, veri girişinden veya tarihsel süreçteki mevzuat değişikliklerinden kaynaklanan veri tür ve içerik farklılıklarının giderilmesi amacıyla yapılan çalışmalar ve yöntemler açıklanmıştır.

2. VERİ TOPLAMA

Henüz gelişim aşamasını tamamlamamış olan ülkemizin, yönetim bilişim altyapılarında yapısal ve içerik değişiklikleri ortaya çıkabilmektedir. Devlet organlarının ve kuruluşların yönetim kademelerindeki bir kısım değişiklikler, veri toplama işiyle uğraşan birimlerin zaman zaman veritabanlarında köklü değişiklikler yapmasına yol açabilmektedir. Bu değişiklikler kimi zaman tüm veritabanına hemen uygulanabilirken, kimi zaman ise verilerin alındığı kaynakta değişiklik yapıldığından, başka veritabanlarındaki verilerin düzeltilmesine olanak bulunamamaktadır. Bu aslında verilerin yönetimsel olarak doğru planlanmadığından ve veriler arası ilişkilerin göz ardı edildiğinden kaynaklanmaktadır (Rajagopalan 2001:460).

Doğru yapılanmış bir yönetim bilişim sisteminde veritabanına girilen verilerin, internet veya diğer sayısal ortamlardan doğrudan alınması ve veritabanına buradan alına bilgilerin kaydedilmesi gerekirken (Haag 1998:221), nerdeyse tüm kamu ve özel kuruluşlarda verilerin veritabanına girilmesi insanlar tarafından gerçekleştirilmektedir. Bu tür bir veri girişinde ise ülkemiz gibi gelişmekte olan ülkelere özgü bazı veri tutarsızlıkları ve yanlışlıkları yanında standart dışı verilerin oluşması da söz konusu olabilmektedir.

Veritabanlarında bilgi keşfinin bir gereği olarak veri madenciliği uygulanacak verilerin bir veri ambarında olması, verinin durağan hale gelmesi açısından önemlidir.

Durağan hale gelecek olan bu veriler üzerinde öncelikle veri tanımlama veya belirleme işlemi yapılarak veri kümesi oluşturulur. Veri kümesi üzerinde sağlıklı bir veri madenciliği yapılabilmesi için gereksiz ve tekrarlı olan veriler ayıklanmalı ve veri kirliliği oluşturan anlamsız veya gereksiz veriler temizlenmelidir. Bu işlemlerden sonra veri madenciliği uygulamak, anlamlı sonuçlar alma açısından önemlidir [4].

3. TOPLANAN VERİLERİN ANALİZİ

Veritabanlarında bilgi keşfinin sonuç almada en önemli etkenlerinden birisi verilerin temizlenmesi ve ayıklanması olarak ifade edilmektedir (Adriaans 1998). Toplanan verilerin analizi yapılırken öncelikle mevcut verilerin istenilen yapıda olup olmadığına bakılmalı, eğer toplanan verilerde veri madenciliği için önemli olabilecek detay veriler göz ardı edilmişse buradan elde edilecek analizler de yüzeysel olacağından veri madencili açısından kayda değer sonuçlar elde edilemeyecektir. Buna karşın verilerde gereksiz detay bilgilerde varsa bunlarda sonuç almayı engelleyici etkenlerdir (Riccardi 2001),(Witten 1999).

3.1. Verilerin Yapısal Analizi

Veri madenciliğinde kullanılan veritabanlarının çok hacimli veriler barındırdığı bir gerçektir. Bu açıdan verilerin yapısal analizi ve iyileştirilmesi en az veriler kadar önem taşımaktadır. Çok büyük veritabanları söz konusu olduğundan bu verilerin yapısal sorunları, bilgi elde etmenin önünde büyük bir engel olabilmektedir.

Yapısal olarak iyi tasarlanmamış veritabanlarında, sayısal olarak saklanması gereken veri alanları bazen metin veya çift duyarlıklı sayı şeklindeki alanlarda saklanabilmektedir. Bu ise veritabanının gereksiz olarak büyümesine ve yapılacak analizlerin çok uzun zaman almasına ve hatta sonuç alınamamasına kadar çeşitli sorunlar ortaya çıkarabilmektedir (Riccardi 2001). Bu sebeple bu tür veriler eğer mümkünse en az yer tutacak şekle çevrilmeli ve bu şekilde saklanmalıdır.

3.2. Verilerin İçerik Analizi

Verilerin içerik yönünden analiz edilerek, içeriğinde uygun veri bulunmayan verilerin düzeltilmesi veya ayıklanması yoluna gidilmelidir. Bu ayıklamada veriler içerisinde bulunan tekrarlı verilerin veya gereksiz verilerin veritabanından çıkarılması sonuç alma açısından önemli bir

adımdır. Bu adımın bir parçası olarak ve veri madenciliğinde doğru sonuçlar almayı sağlayacak bir işlem olarak, verilerde bulunan metinsel ifadelerin tümünün büyük veya tümünün küçük harfe çevrilmesi de gereklidir (Adriaans 1998), (Riccardi 2001).

3.3. Verilerin Standartlaştırılması

Ülkemiz gibi gelişmekte olan ülkelerde kuruluşların ve devlet organlarının gelişmesi sürekli devam ettiğinden, bu ortamlarda oluşan verilerde yapısal değişiklikler yanında, aynı yapı içerisindeki verilerin özellik olarak ta değiştiği bir gerçektir.

Örneğin okulların not sistemi, üniversite giriş sınavı puanı hesaplama yöntemleri, enflasyon nedeniyle parasal veriler, önceden öngörülmemiş olup ta sonradan ortaya çıkan ilave veriler gibi, aynı veri alanında farklı özelliklerle kayıtlı veriler bulunabilmektedir.

Bu farklılıklar ortadan kaldırılmadan yapılacak veri madenciliği analizleri ile elde edilecek bilgiler yanlış sonuçlar elde edilmesine yol açacaktır. Bunu önüne geçmek için verilerin belli standart verilere dönüştürülmesi gerekmektedir. Bu dönüşüm kimi uygulamalarda çok kolaylıkla uygulanabilirken, kimi uygulamalarda daha zor olacağı da açıktır (Kamrani 2001:361).

Verilerin standartlaştırılması işleminde verilerin oluşmasındaki aşamaların bilinmesi standartlaştırma açısından bir kolaylık sağlayabilir. Ancak verilerin standartlaştırılması gerekliliği verilerin derinlemesine analizi ve yorumu ile de bulunabilir.

Standartlaştırmada kullanılabilecek yöntemler verinin türüne göre değişiklik gösterebilir. Eğer veri yıllara bağlı olarak değişmişse bu yılların bilinmesi ile veri belli bir yıl temel alınarak standardize edilebileceği gibi verinin değişim aralıkları göz önüne alınarak ta bu dönüşüm yapılabilir.

4. UYGULAMA

Atatürk Üniversitesi Öğrenci İşleri veritabanı üzerinde bir veri madenciliği uygulaması yapılması planlandı. Çalışmaya esas olan veritabanında 1976 yılında liseden mezun olmuş öğrencilere ait verilerin bulunması, veri madenciliği açısından bir avantaj olarak görüldü, ancak daha sonra yapılan çalışmada bu verilerde bulunan "öğrenci lise mezuniyet notu", "öğrenci ÖSYM puanı" gibi verilerin bazı yıllarda değiştiği görüldü.

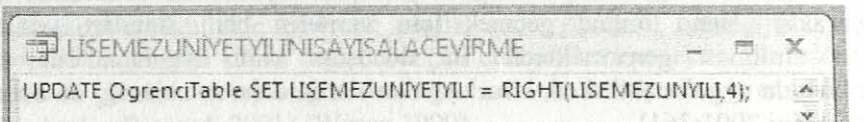
Bu değişimler yapılacak analizleri amacından saptıracak boyut ve önemdeydi. Bu sebeple, veri madenciliğinde verilerin ayıklanması ve temizlenmesi kadar önemli bir diğer konunun da verilerin standart hale getirilmesi olduğu ortaya çıktı.

Atatürk Üniversitesi Öğrenci Veritabanında karşılaşılan en temel problem eksik veri girişidir. Örneğin üç bin civarında öğrencinin lise mezuniyet notu girilmemiştir. Uygulamada çalışılan veri alanı ile ilgili eksik kayıtlar çıkarılarak, veriler, veri madenciliği uygulamasına hazırlandı.

4.1. Veri Yapılarının Düzeltilmesi

Öğrenci İşleri veritabanından seçilen veri setinde, verilerin bir kısmının sayısal veri olmalarına karşın metinsel veri şeklinde kaydedildiği görüldü. Bu yapının veri madenciliği tekniklerinden bazılarının uygulanmasında sorunlar çıkaracağı açıktır.

Bu sebeple oluşturulan veri ambarında bu verilerin yapısal olarak düzeltilmesi işlemi yapıldı. Düzeltme işleminde sayısal veri içermesine karşın metinsel alanlara kaydedilmiş veriler sayısal veriye dönüştürülme işlemi gerçekleştirildi. Bu dönüşüm işleminde sayısal yapıdaki yeni veri alanları kullanıldı. Bu amaçla kullanılan SQL sorgularından birisi şöyledir:



Şekil 1. Mezuniyet yılının sayısal veriye çeviren sorgu

Böylece hem gün/ay/yıl olarak girilen veriler sadece yıl formatına dönüştürülmüş, hem de metin türü olan bu veriler sayısal formata çevrilmiştir.

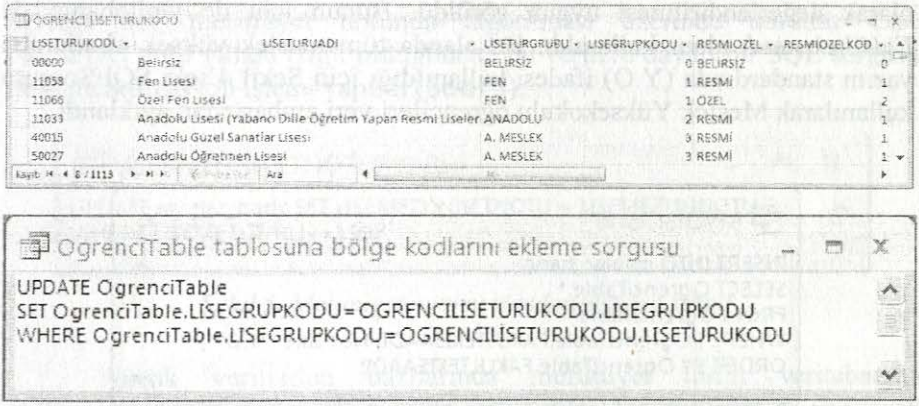
4.2. Verilerin İçerik Yönünden İyileştirilmesi

Veri madenciliği uygulamasında verileri değişim aralığının çok geniş olması sonuç alma ve sonuçları görselleştirme açısından sorunlar çıkardığından, bazı verilerin veritabanında olmayan, ancak daha anlamlı olan verilere dönüştürülmesi veya kodlanması gerektiği bu çalışmanın sonraki adımında ortaya çıktı. Bu amaçla yapılan işlemlerin bir kısmı şunlardır:

Liselerin Gruplandırılması:

Bu aşamada yapılan çalışmada Üniversitemizi kazanmış olan öğrencilerin mezun oldukları okul türlerinin çok çeşitli olduğu ve veri madenciliği açısından yanlış sonuçlar ortaya çıkarabileceği anlaşıldı. Bu durumu ortadan kaldırmak için okul türlerinin sınıflandırılması ve sınıfları içeren bir verinin veritabanına eklenmesi yolu seçildi. Bunun için lise

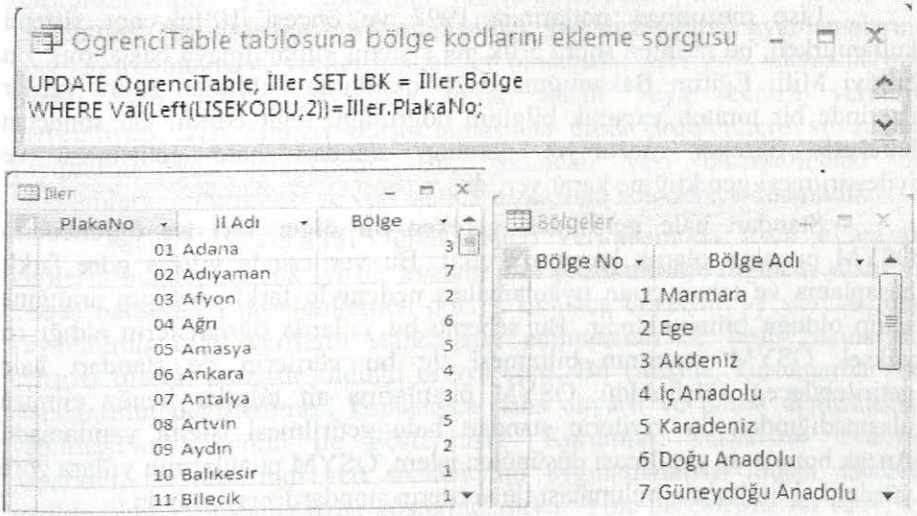
türlerinin bulunduğu tablodaki LİSETURGRUBU alanı kullanılarak Öğrenci Tablosu güncellendi (Şekil 2)



Şekil 2. *Öğrenci_LiseTurukodu Tablosu ve Öğrenci Tablosunu Güncelleyen SQL Sorgusu*

Coğrafi Bölgelerin Eklenmesi:

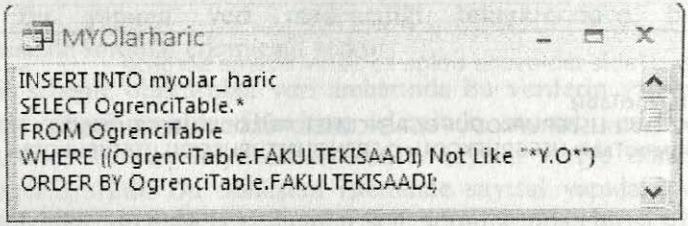
Diğer bir gereklilik ise, üniversite öğrencilerinin mezun oldukları liselerin bulunduğu illerin bölgelere göre gruplandırılmasıdır. OgrenciTable tablosundaki LISEKODU alanının ilk iki karakterinin il plaka kodunu temsil ettiği göz önüne alınarak, Şekil 3'teki Bölgeler referans tablosu esas alınarak illerin coğrafi bölgeleri aşağıdaki SQL sorgusu ile elde edildi:



Şekil 3. *İl ve Bölge Tabloları*

Meslek Yüksekokullarının Çıkarılması:

Meslek Yüksekokulları 2 yıllık oldukları ve OSYM giriş puanları (çoğunlukla) olmadığı için veri ambarından çıkarılması ve ayrı bir kategori olarak değerlendirilmesi uygun görüldü. Bunun için de veritabanındaki *Fakültekısaadı* alanı kullanıldı. Bu alanda tüm meslek yüksek okullarının yazım standardında (Y.O) ifadesi kullanıldığı için Şekil 4'teki SQl Sorgusu kullanılarak Meslek Yüksekokulu öğrencileri veri ambarından ayıklandı.



Şekil 4. Meslek Yüksekokullarının Asıl Tablodan Ayıklayan SQL Sorgusu

4.3. Standartlaştırılacak Verilerin Belirlenmesi

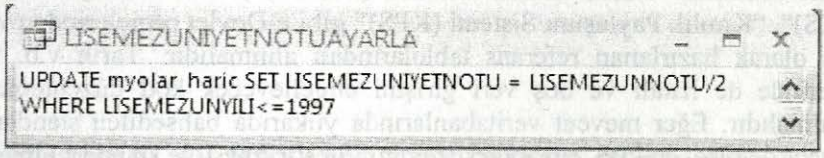
Atatürk Üniversitesini kazanan öğrencilerin lise mezuniyet notları ve ÖSYM puanlarını da içeren bir veri madenciliği uygulaması yapılmadan önce bu verilerin standart hale getirilmesi gerektiği görüldü. Bu verilerden "Öğrenci Lise Mezuniyet Notu"nda 1 ila 10 arası değerler içerdiği ve bu sınırların şimdiki lise mezuniyet notları ile uyuşmadığı belirlendi. Bunun üzerine lise not sisteminin değiştirildiği tarihin belirlenmesi yoluna gidildi.

Lise mezuniyet notlarında 1997 ve öncesi 10'luk not sistemi kullanılırken, bu tarihten sonra 5'lik not sistemi kullanılmaya başlanmış. Bu bilgiyi Milli Eğitim Bakanlığına bağlı okullardan aldıktan sonra veriler üzerinde bir tarama yaparak bilginin doğruluğu teyit edildi. Bu sonuçtan hareketle öğrenci mezuniyet notunun standart hale getirilmesi ve iyileştirilmesi gerektiğine karar verildi.

Standart hale getirilmesi gereken bir diğer veri ise öğrencilerin ÖSYM puanları olarak karşımıza çıktı. Bu verilerinde yıllara göre farklı hesaplama ve taban puan uygulamaları nedeniyle farklı değişim aralığına sahip olduğu bilinmektedir. Bu sebeple bu yıllarda öğrencilerin aldığı en yüksek ÖSYM puanının bilinmesi ile bu verilerin de standart hale getirilebileceği düşünüldü. ÖSYM puanlarına ait bilgiler henüz elimize ulaşmadığından bu verilerin standart hale getirilmesi işlemi yapılamadı. Ancak burada da yapılması düşünülen işlem, ÖSYM puanlarının yıllara göre yüzdelik dilimlerinin bulunması ile verilerin standardizasyonuydu.

4.4. Veri Standartlaştırma Algoritması

Verilerin standartlaştırılmasına ilişkin algoritmalar belirlenirken her bir verinin ayrı ele alınması gerektiği ortadadır. Bu bakımdan ilk olarak öğrenci lise mezuniyet notunun algoritması üzerinde duruldu. Lise mezuniyet notu yıllara bağlı olduğundan bu verilere dayalı bir SQL sorgusu ile standardizasyon işlemi yapıldı (Şekil 5).



Şekil 5. *Mezuniyet notunu standartlaştıran sorgu*

Ancak verilerden bazılarında mezuniyet tarihi veritabanına girilmemiş olduğu da bu işlemler ile ortaya çıktı. Bu durumda da yapılacak iki işlem vardı bunlardan biri bu verilerin güncellenmesi, diğeri ise bu verilerin atılması idi. Bu tür sorunlu veri sayısı az olduğundan bu verilerin atılması yolu seçildi.

5. SONUÇ

Veri ambarının iyi bir şekilde oluşturulması, veri madenciliği uygulamalarının başarılı olmasının öncelikli şartıdır. Bu sebeple veri madenciliği uygulamalarının en uzun ve karmaşık aşaması veri ambarının oluşturulmasıdır. Veri ambarının oluşturulmasında pek çok problemler mevcuttur. Öncelikle mevcut veritabanının, veri madenciliği uygulamalarına göre tasarlanmaması, verilerin standart olmaması, veri giriş elemanlarının ihmali veya başka sebeplerle eksik, hatalı veya tekrarlı verilerin mevcudiyeti, veri ambarı oluşturulmasında ciddi problemlere ve zaman kayıplarına yol açmaktadır. Bu bakımdan veri tabanlarındaki bu problemlerin giderilmesi ve veri standardizasyonu son derece önemlidir.

Atatürk Üniversitesi Öğrenci İşleri Veritabanında, uzun yıllara ait verilerin bulunması çok iyi bir avantaj gibi gözükmesine rağmen, yukarıda kısaca bahsedilen problemlerden dolayı, oldukça düzensiz ve standart dışı kaydedilmiş olan verilerin standardize edilmesinin de başlı başına bir çalışma olacak düzeyde olduğu ortaya çıktı. Bu çalışma, kurumlarda veri standardının değiştirilmesi konusunda daha duyarlı ve planlı değişiklikler yapılmasının önemini de göstermiştir. Kurumlar, veritabanı tasarımı yaparken, bu tasarımın veri madenciliği uygulamalarına imkân verecek şekilde oluşturulmasına özen göstermelidirler. Yine bu çalışma ile verilerin sayısal ortamlarda ve sürekli güncellenecek bir veri alış-verişi yapısı

içerisinde bulunması gerektiği zorunluluğu ortaya çıkmıştır (Haags 1998). Ancak burada da kurumlara ait veritabanlarının güvenliği gibi sorunlarla karşılaşılması mümkündür. Bu gibi durumlarda ise verilerin sürekli incelenmesi ve hatalı veri giriş noktalarının düzeltilmesi yoluna gidilebilir. Bunun yanı sıra mümkün olduğu kadar standart olmayan veri girişi de engellenmelidir. Örneğin şehir, okul, il, ilçe v.s. bilgiler el ile yazılmak yerine "Merkezi Nüfus İdaresi Sistemi (Mernis)", "Adres Kayıt Sistemi (AKS)", "Kimlik Paylaşımı Sistemi (KPS)" gibi e-Devlet projelerinden veya özel olarak hazırlanan referans tablolarından alınmalıdır. Tarih v.b. veri türlerinde de hatalı ve boş veri girişini engelleyecek kod düzenlemeleri yapılmalıdır. Eğer mevcut veritabanlarında yukarıda bahsedilen standartlar mevcut değilse, en kısa zamanda sistemin iyileştirilmesi ve standart olmayan verilerin düzenlenmesi konusunda çalışmalar yapılmalıdır.

Veritabanlarının kullanılması ile verilerin oluşmasında kullanılan yazılımlar ve veri giriş yöntemlerinin de iyileştirilmesi, daha sonra yapılacak olan veri madenciliği ve bilgi keşfi uygulamaları için daha uygun zeminler hazırlayacaktır. Bu açıdan yazılım ve veritabanlarının da yeniden düzenlenmesi gerekmektedir.

KAYNAKÇA

Adriaans P., Zantinge D. (1998), *Data Mining*, Addison Wesley Longman, Boston-ABD.

Haag S., Cummings M., Dawkins J. (1998), *Managment Information Systems for the Information Age*, McGraw-Hill, New York-ABD.

Kamrani A., Rong W., Gonzalez R. (2001), *A Genetic Algorithm Methodology for Data Mining and Intelligent Knowledge Acquisition*, Computer & Industrial Engineering 40, 361-377.

Özdemir A., (2004), *Veritabanlarında Bilgi Keşfi ve Veri Madenciliği*, Atatürk Üniversitesi Sosyal Bilimler Enstitüsü, Doktora Tezi.

Rajagopalan B, Isken M. W. (2001), Exploiting Data Preparation to Enhance Mining and Knowledge Discovery, IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews 31(4), 460-467.

Riccardi G. (2001), *Principles of Database Systems with Internet and Java Aplications*, Addison Wesley.

Witten I., Frank E. (1999), *Data Mining- Practical Machine Learning Toosl and Techniques with Java Implementationns-*, Morgan Kaufmann Publishers.

Zhu X., Wu X. (2005), *Cost-Constrained Data Acqusition for Intelligent Data Preparation*, IEEE Transactions on Knowlegde and Data Engineering 17(11), 1542-1556.

Mlynarski, R., Ilczuk, G., Wakulicz-Deja, A., and Kargul, W., (2006), *A New Method of Data Preparation for Cardiological Decision Support,* IEEE Computers in Cardiology 33, 273–276.