Üniversite Öğrencilerinin İnterneti Eğitimsel Amaçlar İçin Kullanmalarını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti

Ahmet Selman BOZKIR[1] , Bilge GÖK[2] ve Ebru SEZER[3]

[1]Hacettepe Üniversitesi Bilgisayar Mühendisliği Bölümü, 06532, Beytepe-Ankara
selman@cs.hacettepe.edu.tr
[2]Hacettepe Üniversitesi Sınıf Öğretmenliği Bölümü, 06532, Beytepe-Ankara
bilgeb@hacettepe.edu.tr
[3]Hacettepe Üniversitesi Bilgisayar Mühendisliği Bölümü, 06532, Beytepe-Ankara
esezer@cs.hacettepe.edu.tr

ÖZET

Bu çalışmanın amacı, üniversitede okumakta olan lisans öğrencilerinin eğitim amacıyla internet kullanımlarının içerdiği gizli veya açık örüntüleri, veri madenciliği yöntemleriyle tespit etmektir. Hacettepe Üniversitesi'nde 2007-2008 öğretim yılı bahar döneminde gelişigüzel seçilen 380 öğrenciye uygulanan konu ile ilgili ölçekten elde edilen verilere üç adet veri madenciliği algoritması (karar ağaçları, kümeleme ve birliktelik kuralları) uygulanarak öğrencilerin eğitimsel amaçlar için internet kullanmalarına yönelik davranış ve düşüncelerinde sık görülen örüntüler bulunmuştur. Yapılan analizler sonucunda öğrencilerin konu ile ilgili teknik yeterliliklerini etkileyen faktörler belirlenmiş, çeşitli öğrenci profilleri çıkarılmış ve internetin kullanımına ilişkin görüşlerindeki benzerlikler tespit edilmiştir.

Anahtar Kelimeler :*Veri Madenciliği; Karar Ağaçları; Kümeleme; Birliktelik Kuralları.*

ABSTRACT

The aim of this study is to detect hidden or visible patterns of university students' internet usage for educational purposes via data mining techniques. Frequently observed patterns on behaviors and thoughts of and the students are found out through processing three data mining algorithms (decision trees, clustering, association rules) to the data which are collected from a poll applied to randomly chosen 380 students at Hacettepe University in 2007- 2008 Spring Semester. As a result of analyses, the factors that affect the technical sufficiency of the students are defined, the profiles of some students are extracted, and similar views among students about internet usage are discovered.

Key Words: *Data Mining; Decision Trees; Clustering; Association Rules.*

1. GİRİŞ

İnternet şüphesiz yeryüzündeki en büyük bilgi kaynağıdır ve bu kaynak her gün milyonlarca kez farklı amaçlarla kullanılmaktadır. Bu çalışmada internetin ülkemizde lisans öğrencileri tarafından eğitimsel amaçlar için kullanımı ele alınmış olup bu konudaki etkin faktörlerin ve görüşlerin araştırılması adına Hacettepe Üniversitesin'de 2007-2008 öğretim yılında bir ölçek geliştirilmiş ve bu ölçek üniversite öğrencilerine uygulanmıştır. Öğrencilerin ölçeğe verdikleri cevaplar araştırmamızın temel veri kaynağı olmuştur. Elde edilen veriler çeşitli işlemlerden geçirilirek veri madenciliği yöntemleriyle anlamlı bilgiler keşfedilmeye çalışılmıştır.

Bu çalışmada, öncelikle veri madenciliğinin ne olduğu ve yöntemleri kısaca özetlenecek daha sonra ise çalışmanın yapılış aşamaları anlatılarak ayrıntılı sonuçlar verilecektir. Son bölümde elde edilmiş anlamlı sonuçlardan yola çıkılarak ülkemizde lisans düzeyinde eğitim – öğrenim gören üniversite öğrencilerinin mevcut durumları ile gelecekte neler yapılması gerektiği üzerine bir tartışma yapılacaktır.

1.1 Veri Madenciliği

Basit bir tanım yapmak gerekirse **veri madenciliği**, büyük ölçekli veriler arasından bilgiye ulaşma, bilgiyi madenleme işidir [1]. Diğer bir deyişle, veri madenciliği tek başına birşey ifade etmeyen veriler içindeki gizli örüntüleri ve ilişkileri ortaya çıkarmak için istatistik, yapay zekâ ve makine öğrenmesi gibi yöntemlerin ileri veri çözümleme araçlarıyla kullanılmasını kapsayan süreçler topluluğudur. Geleneksel sorgu (Query) ve raporlama araçlarının veri yığınları karşısında yetersiz kalması, saklı ve işlenmemiş bilgiye olan büyük ihtiyaç Veritabanlarında Bilgi Keşfi (VTBK) ve Veri Madenciliği (VM) gibi alanların keşfiyle anlaşılabilir ve yorumlanabilir hale gelmiştir [2]. Veri madenciliği uygulamaları başta pazarlama, bankacılık, tıp, mühendislik, endüstri, borsa analizleri ve ulusal güvenlik alanlarında kullanılmaktadır. Örneğin, müşteri ilişkileri yönetiminde, kredi kartı dolandırıcılıklarının tespitinde, üretim süreçlerinin iyileştirilmesinde, hatların yoğunluk tahmininde, kalite kontrol analizlerinde, hisse senedi fiyat tahmininde veri madenciliğinden etkin şekilde faydalanılmaktadır. Veri madenciliği çalışmaları yapmak üzere birçok ticari yazılım üretilmiştir. Oracle DM, Microsoft SQL Server 2005 Analysis Services, SPSS Clementine, SAS Enterprise Miner bu ürünlerden sadece birkaçıdır. Çalışmamızda daha

sonrada bahsedilecek olan yararlarından dolayı Microsoft SQL Server 2005 Analysis Services uygulaması kullanılmıştır.

Veri madenciliği modelleri iki kesimde toplanmaktadır:

1) Kestirimsel Modeller (sınıflandırma, eğri uydurma, zaman serileri vs.)
2) Tanımlayıcı Modeller (kümeleme, özetleme, birliktelik kuralları, sıralı diziler vs.)

Kestirimsel modellerde amaç mevcut verileri kullanarak geleceğe yönelik kestirimler yapabilmek iken, tanımlayıcı modellerde amaç yine mevcut veri içindeki gizli ilişkileri, kümeleri ve veriyi niteleyebilecek olan özellikleri ortaya çıkarmaktır. Bu kapsamda çalışmada her iki modelde yer alan tekniklerden yararlanılmıştır. Aşağıda çalışmamızda kullanılan teknikler kısaca özetlenmiştir:

Karar Ağaçları:

Karar ağaçları kestirimsel yöntemler içerisinde başarım oranının yüksekliği ve kolay anlaşılır bir grafiksel yorumunun bulunması sebebiyle en çok tercih edilen sınıflandırma tekniğidir. Karar ağaçları kullanılarak verinin sınıflandırılması, iki basamakta gerçekleşmektedir. İlk basamakta mevcut veriler temiz ve tutarlı bir hale dönüştürülüp karar ağacı algoritması öğrenme (training) aşamasından geçirilir. Bu aşamanın tamamlanması ile daha sonra algoritmanın tahminsel amaçlar için kullanabileceği bir model oluşur. Bu model sınıflama kuralları olarak da adlandırabilecek olan karar ağaçlarıdır. İkinci aşamada, eğitimi tamamlanmış olan model ile yeni verilerde tahmin edilmesi beklenen nitelikler tahmin edilebilir. Karar ağacı algoritmalarından ID3, C4.5, C5.0, CART (The Classification and Regression Trees), CHAID, Microsoft Decision Trees algoritmaları tanınmış algoritmalardır. Çalışmamızda karar ağacı üretimi sonunda bağımlılık ağı gösterim desteği bulunduğu için Microsoft Decision Trees algoritması kullanılmıştır.

Kümeleme:

Kümeleme, anlamından da anlaşılabileceği üzere, belli bir veri kümesindeki her varlığı niteliklerine göre bir kümeye atama işlemidir. Bu teknik ile müşteri profilleri gibi anlamlı bilgiler çıkarılabilmektedir. K-Means, Expectation Maximization (EM), Hierarchical Clustering algoritmaları bilinen ve sıklıkla kullanılan algoritmalardır. Çalışmamızda Microsoft SQL Server 2005 ürünü ile birlikte gelen EM algoritması kullanılmıştır.

Birliktelik Kuralları:

Birliktelik kuralları, veri kümesi içindeki hareketlerin (transactions) analiz edilerek bu hareketler ya da kayıtlar arasında sıklıkla bir arada görülenlerin tespit edilmesi işlemidir. Birliktelik kuralları ticaret, mühendislik, fen ve sağlık sektörlerininde içinde bulunduğu diğer birçok alanda uygulanmaktadır [3]. Birliktelik kurallarının kullanımında bazı destek ve güven parametreleri önemli rol oynamaktadır. Bununla birlikte çalışmamızda çıkarılan kurallara ait öğelerin birbirleriyle korelasyonlarını ortaya koyan **önem** (importance, lift) değeri de baz alınmıştır.

1.2 Eğitimde Veri Madenciliği Çalışmaları

Ülkemizde eğitim üzerinde veri madenciliği yöntemleriyle yapılan çalışmalara örnek vermek gerekirse "Kütüphane Kullanıcılarının Erişim Örüntülerinin Keşfi" başlıklı çalışmada [4], kütüphane sitesi web günlüklerine dayalı olarak kütüphane kullanıcılarının erişim örüntüleri bulunmaya çalışılmıştır. Diğer bir çalışmada [2], ÖSS sınavına giren öğrencilerin profillerinin ve tercihlerinin veri madenciliği yöntemleriyle belirlenmesi amaçlanmıştır. Yine başka bir çalışmada [5] veri madenciliği teknikleri yardımıyla Fırat Üniversitesi Teknik Eğitim Fakültesi Bilgisayar Eğitimi Bölümü öğrencilerinin notları kullanılarak öğrenci başarılarının analizi yapılmıştır. Bu analizi yapmak için veri madenciliğinde, birliktelik kuralı çıkarım algoritmalarından biri olan Apriori algoritması kullanılmıştır. Bu çalışma da veri madenciliği yöntemlerini kullanarak intenetin eğitim amaçlı kullanımını araştırmayı hedeflemiştir.

2. UYGULAMA

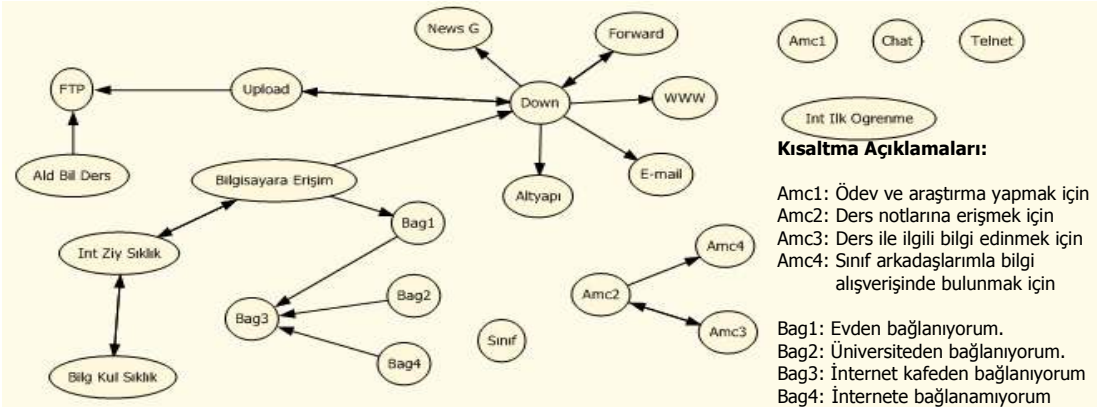
Çalışmamızda internetin eğitimsel amaçlar için kullanımını tespit edebilmek için daha önce yapılan bir çalışmada [6] kullanılmış olan tutum ölçeğinden yararlanılmıştır. Bu ölçekte 2 kısım yer almaktadır. Birinci kısımda öğrencilerin mevcut eğitimsel ve teknik durumlarını öğrenmeye yönelik 12, ikinci kısımda ise konuya ait görüşlerini tespit etmeye yönelik 50 adet soru yer almıştır. Anket Hacettepe Üniversitesi'nde 2007-2008 bahar döneminde 380 öğrenciye gelişigüzel örnekleme yöntemiyle uygulanmıştır. Veriler SPSS 15. uygulamasında ham veri olarak toplanmıştır. Toplanan ilk grup verilere ait ayrıntılı dağılım Tablo 1'de sunulmuştur.

Tablo 1. Veri kümesinde yer alan niteliklere ait ayrıntılı dağılım

Nitelik Adı.	Nitelik Türü	Dağılım
Cinsiyet	Metinsel	Kız: **246**, Erkek: **134**
Sınıf	Sayısal (Ayrık)	1.Sınıf:**155**, 2.Sınıf: **55**, 3.Sınıf:**112**, 4.Sınıf:**98**
Not Ortalaması	Metinsel	0 – 1.50: **11**, 1.50 - 2.00: **24**, 2.00 - 2.50: **81**, 2.50 – 3.00: **115**, 3.00 – 3.50: **127**, 3.50 - 4.00:**22**
Bilg. Eriş. Koşul.	Metinsel	Kendimin Var: **213**, Yakın çevremde var:**144**, Ulaşmam çok zor:**17**, Çevremde yok: **6**
Bilg. Kul. Sıklık.	Metinsel	Hergün, sürekli:**101**, Hergün, birkaç saat: **106**, Haftada birkaç gün: **100**, Haftada birkaç saat: **59**, Ayda birkaç saat: **11**, Hiç: **3**
Alınan. Bilg. Ders	Sayısal (Ayrık)	Hiç: **66**, 1 Ders: **188**, 2 Ders: **67**, 3 Ders: **13**, 3den fazla ders:**46**
İnt. Bağlan. Durum:	Metinsel	(Bag 1) Evden bağlanıyorum:**177** Evet, **209** Hayır (Bag 2) Üniversiteden bağlanıyorum: **148** Evet, **232** Hayır (Bag 3) İnternet kafeden bağlanıyorum: **144** Evet, **236** Hayır (Bag 4) İnternete bağlanamıyorum: **18** Evet, **362** Hayır
İnt. Eğitim. Amaç:	Metinsel	(Amaç 1) Ödev ve araştırma yapmak için: **361** Evet, **19** Hayır (Amaç 2) Ders notlarına erişmek için: **234** Evet, **146** Hayır (Amaç 3) Sınıf arkadaşlarımla bilgi alış verişi için: **230** Evet, **150** Hayır (Amaç 4) Diğer: **230** Evet, **160** Hayır
Bilg. Altyapısı:	Metinsel	Hiç yok: **13**, Çok sınırlı: **46**, Biraz var:**185**, Epeyce var: **110**, Çok iyi: **26**
Kul. Özellikler:	Metinsel	WWW: **293** Evet, **87** Hayır, E-Posta: **301** Evet, **79** Hayır, FTP: **47** Evet, **333** Hayır, Chat Odaları: **36** Evet, **344** Hayır, Newsgroup (Haber grupları): **101** Evet, **271** Hayır, Forwarding: **144** Evet, **236** Hayır, Downloading: **196** Evet, **184** Hayır Uploading: **75** Evet, **305** Hayır
İnt. Kul. Sıklık:	Metinsel	Hiçbir zaman: **3**, Ayda bir kez: **10**, 15 Günde bir kez: **33**, Haftada bir kez: **122**, Hergün: **222**
İnt. İlk Öğrenme:	Metinsel	Derste: **48**, Kitap/dergi aracılığıyla: **7**, Sunum yapmak için:**12**, Arkadaşlarımın yardımıyla: **70**, Kütüphane aracılığıyla: **10**, Kendi kendime: **219**, Diğer: **14**

Toplanan ham veriler Excel 2007 üzerine kurulu "SQL Server 2005 Data Mining Add-Ins for Office 2007" paketi yardımıyla temizlenmiş ve niteleliklere ait uygun etiketlemeler yapılmıştır. Daha sonrasında SQL Server 2005 ürünü içindeki "Analysis Services" paketinde bulunan veri madenciliği modülleri, Visual Studio 2005 üzerinde çalıştırılarak analizleri yapılmıştır.

"Microsoft Decision Trees" algoritmasıyla yapılan analizlerde varsayılan ayarlar kullanılmış, "Cinsiyet" ve "Sınıf" nitelikleri sadece girdi amaçlı kullanılırken diğer nitelikler tür olarak "Predict" şeklinde hem girdi hem de tahminsel amaçlı kullanılmıştır. Gerçekleştirilen madencilik modeli içinde niteliklerin birbirlerine etkilerini gösteren bağımlılık ağları mevcuttur. Ağdaki bağlara %50 seyreltme uygulanarak kuvvetli olanları aşağıda sunulmuştur (Şekil 1).



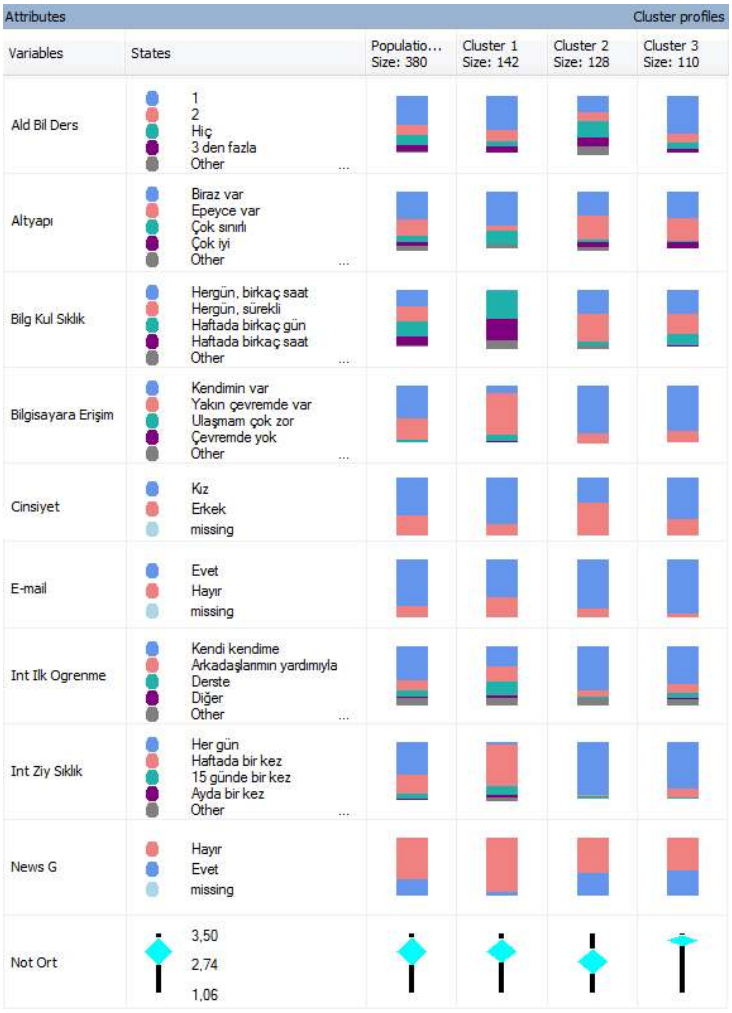
Şekil 1. Karar ağacı algoritması sonuçlarından çıkarılan bağımlılık ağı.

Bağımlılık ağından çıkarılabilecek sonuçlar şu şekilde yorumlanabilir;

- Öğrencilerin interneti ziyaret sıklığı, bilgisayar kullanım sıklığı ve bilgisayara erişim ile doğrudan ilişkilidir.
- FTP kullanımına etkiyen etmenler, alınan bilgisayar dersi sayısı ile yükleme (upload) yapabilmedir.
- Dosya indirebilme bilgisi, diğer birçok işlemi (E-mail, WWW, Posta İletimi (Forwarding), Haber grubu (Newsgroup) vs..) etkilemektedir.
- Sınıfın, chat yapma durumunun, interneti kullanmayı ilk öğrenme yönteminin ve internetin ödev ve araştırma yapma için kullanımının diğer niteliklerle kuvvetli bir etkileşimi görülmemiştir.

Yapılan kümeleme çalışmasında küme sayısının otomatik tespit edilmesi amaçlanmıştır. Bu sebeple uygulama içerisinde "CLUSTER_COUNT" parametresi 0 olarak verilerek EM algoritması kullanılmıştır. Çalışmamızda EM algoritması K-Means algoritmasından daha iyi sonuçlar vermiştir. Analiz sonucunda üç adet öğrenci profili çıkarılmıştır (Şekil 2).

Şekil 2'de yer alan nitelikler, kümeleme sonucunda kümeler açısından belirgin farklılıklar taşıyan niteliklerdir. Analiz sonucunda 2. ve 3. kümeler arasında benzerlik tespit edilmiş olmasına karşın 2. küme (Toplam 128 katılımcı) %56 oranında erkek katılımcı içermekte ve genel not ortalamasında (2.33 +/- 0.52) en düşük sıralamada yer almaktadır. 3. küme (Toplam 110 kişi) ise %71 oranında bayan katılımcı içeriyor olup en yüksek not ortalamasına (3.20 +/- 0.23) sahip gruptur. Bu iki küme internete bağlanma durumu, bilgisayarı ve interneti ziyaret sıklığı, bilgisayara erişim durumu, bilgisayar konusundaki altyapı, interneti kullanmayı ilk öğrenme yöntemi ve e-mail ile haber grubu kullanımı gibi alanlarda yüksek benzerliğe sahip olmakla birlikte, derslerde gösterilen başarı ve bilgisayara yönelik alınan ders sayısı bakımından farklılıklar taşımaktadır. Birinci kümede ise (Toplam 142 katılımcı, not ortalaması: 2.76 +/- 0.48) %80 oranında bayan katılımcı tespit edilmiştir. Birinci kümedeki katılımcılar gerek bilgisayara/internete erişim ve kullanım açısından gerekse de internete bağlı teknolojilerin kullanımı konusunda daha geride kalmaktadırlar oysaki bu küme konu hakkında yüksek oranda (%99) ders görmüş bir grup olmasına karşın bilgisayara ve internete erişimde yaşanan sıkıntılar bu öğrenci profilinin, internet teknolojilerini kullanma konusunda geride kalmasına sebep olmaktadır.



Şekil 2. Kümeleme algoritmasının sonucunda çıkarılan kümelerin farklılık gösteren nitelikleri

İnternetin kullanımına yönelik görüşlerde sıklıkla görülen benzerliklerin araştırıldığı birliktelik kuralları analizinde, destek değeri %5, güven değeri ise %50 olarak belirlenerek seyreltme uygulanmış ve çıkarılabilecek binlerce kuraldan önemli olanların tespit edilmiştir.

Tablo 2. Birliktelik Kuralları Analizinden Çıkarılan Kurallar

Bulunan Kural.	**Destek % (Support)**	**Önem (İmportance)**
İnternet üzerinden tarama yapmaktan hoşlanmıyorum.=> İnternette araştırma yapmak bana sıkıcı gelir.	0,549	**0,784**
Araştırma yaparken internetten yararlanmam, Dersler için interneti kullandığımda bunaldığımı hissederim. => Bilgi alışverişini internet aracılığı ile yapmam.	0,714	**0,761**
İnternette araştırma yapmak bana sıkıcı gelir. => İnternet üzerinden tarama yapmaktan hoşlanmıyorum.	0,634	**0,712**
İnterneti bilgiye erişmek için kullanmak oldukça zordur.=> Ödev yapmada internetin yararlı olduğunu düşünmüyorum.	0,730	**0,674**

Uygulanan ölçekteki elli soruluk görüşler kısmındaki sorulara "Tam katılıyorum" ya da "Katılıyorum" şeklinde cevap alınanlar hedef veri kümesine veri çekme ve dönüştüme

işlemleri yapılarak dahil edilmiştir. Güven (confidence) değerinin kimi zaman yanlış değerlendirmelere sebep olabileceği ve çıkarsanan kuraldaki öğelerin korelasyonunu ifade ettiği için kuralın önem (importance, lift) değeri temel alınarak sıralama yapılmıştır. Tablo 2'de analiz sonucunda çıkarılan bir dizi ilginç kural sunulmuştur.

3. SONUÇLAR VE ÖNERİLER

Çalışmamızda internetin üniversite öğrencileri tarafından eğitimsel amaçlar için kullanımına yönelik imkânları ve görüşleri veri madenciliği yöntemleri ile analiz edilmiştir. Öğrencilerin bilgisayara ve internete erişim koşullarının bu teknolojileri kullanımlarıdaki sıklığa doğrudan etkisi olduğu görülmüştür. Bilgisayara yönelik alınan derslerin öğrencilerin altyapısında çok fazla değişiklik getirmediği gözlemlenmiştir. Bu sebeple öğrencilerin teknik imkânlarının iyileştirilmesinin bilgisayar derslerinden daha önemli olduğunu düşünmekteyiz. Ayrıca öğrencilerin internette araştırma yaparken genellikle sıkıldıkları bunun sonucunda da internete ait bakış açılarının olumsuz yönde geliştiği görülmektedir. Bu nedenle internette bilgi erişimi için yapılan aramalarda öğrenciyi doğru başlangıç noktalarına yönlendirme, internette arama yapma ipuçları ile destekleme ve etkin sorgu üretebilme gibi bilgi desteklerinin öğrenciye verilmesi gerektiğini düşünmekteyiz. Sonuç olarak; ülkemizde internetin öğrenciler tarafından daha yararlı ve etkin kullanılabilmesi için üniversitelerdeki teknik imkânların geliştirilmesi ve öğrencilere etkin çevrimiçi araştırma yöntemleri üzerine eğitim verilmesi gerekliliğinin ortaya çıktığı düşünülmektedir.

4. KAYNAKLAR

1. Veri Madenciliği", http://tr.wikipedia.org/wiki/Veri_madencili%C4%9Fi
2. Dolgun, M. Ö., Özdemir, T.G., Deliloğlu, S., "Öğrenci Seçme Sınavında (ÖSS) Öğrencilerin Tercih Profillerinin Veri Madenciliği Yöntemleriyle Tespiti", Bilişim" 07 Kongresi, Ankara, 2007
3. Dolgun, M. Ö., "Büyük Alışveriş Merkezinden Yapılan Satışlar İçin Sepet Analizi", Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Ankara, 2006.
4. Takçı, H. ve Soğukpınar, İ., " Kütüphane Kullanıcılarının Erişim Örüntülerinin Keşfi ", Bilgi Dünyası Dergisi, 3 (1): 12-26, 2002.
5. Karabatak, M ve İnce M. C., "Apriori Algoritması ile Öğrenci Başarısı Analizi", Eleco' 2004 Elektrik-Elektronik ve Bilgisayar Mühendisleri Sempozyumu", Bursa, 2004
6. Gök, B ve Erdoğan T., "Sınıf Öğretmeni Adaylarının İnternet'in Eğitimsel Amaçlar İçin Kullanımına Yönelik Tutumlarının Belirlenmesi", Uluslararası Eğitim Teknolojileri Sempozyumu, Eskişehir, 2008