

Üniversite Öğrenci Başarısı Üzerine Etki Eden Faktörlerin
Veri Madenciliği Yöntemleri İle İncelenmesi

Çağrı Kılınç

YÜKSEK LİSANS TEZİ

Elektrik-Elektronik Mühendisliği Anabilim Dalı

Ocak 2015

Examining The Effects On University Student
Success By Data Mining Techniques

Çağrı Kılınç

MASTER OF SCIENCE THESIS

Department of Electrical Electronics Engineering

January 2015

Üniversite Öğrenci Başarısı Üzerine Etki Eden Faktörlerin
Veri Madenciliği Yöntemleri İle İncelenmesi

Çağrı Kılınç

Eskişehir Osmangazi Üniversitesi
Fen Bilimleri Enstitüsü
Lisansüstü Yönetmeliği Uyarınca
Elektrik-Elektronik Mühendisliği Anabilim Dalı
Telekomünikasyon ve Sinyal İşleme Bilim Dalında
YÜKSEK LİSANS TEZİ
Olarak Hazırlanmıştır

Danışman: Yrd. Doç. Dr. Celal Murat Kandemir

Ocak 2015

ONAY

Elektrik Elektronik Mühendisliği Anabilim Dalı Yüksek Lisans öğrencisi Çağrı Kılınç'ın YÜKSEK LİSANS tezi olarak hazırladığı “Üniversite Öğrenci Başarısı Üzerine Etki Eden Faktörlerin Veri Madenciliği Yöntemleri İle İncelenmesi” başlıklı bu çalışma, jürimizce lisansüstü yönetmeliğin ilgili maddeleri uyarınca değerlendirilerek oybirliği ile kabul edilmiştir.

Danışman : Yrd. Doç. Dr. Celal Murat Kandemir

İkinci Danışman : -

Yüksek Lisans Tez Savunma Jürisi:

Üye : Yrd. Doç. Dr. Celal Murat Kandemir

Üye : Yrd. Doç. Dr. Nihat Adar

Üye : Yrd. Doç. Dr. Kemal Özkan

Üye : Yrd. Doç. Dr. Efnan Şora Günel

Üye : Yrd. Doç. Dr. Uğur Gürel

Fen Bilimleri Enstitüsü Yönetim Kurulu'nun tarih ve
..... sayılı kararıyla onaylanmıştır.

Prof. Dr. Hürriyet ERŞAHAN
Enstitü Müdürü

ETİK BEYAN

Eskişehir Osmangazi Üniversitesi Fen Bilimleri Enstitüsü tez yazım kılavuzuna göre, Yrd. Doç. Dr. Celal Murat KANDEMİR danışmanlığında hazırlamış olduğum “Üniversite Öğrenci Başarısı Üzerine Etki Eden Faktörlerin Veri Madenciliği Yöntemleri İle İncelenmesi” başlıklı YÜKSEK LİSANS tezimin özgün bir çalışma olduğunu; tez çalışmamın tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı; tezimde verdiğim bilgileri, verileri akademik ve bilimsel etik ilke ve kurallara uygun olarak elde ettiğimi; tez çalışmamda yararlandığım eserlerin tümüne atıf yaptığımı ve kaynak gösterdiğimi ve bilgi, belge ve sonuçları bilimsel etik ilke ve kurallara göre sunduğumu beyan ederim. 07/01/2015

Çağrı Kılınç

İmza

ÖZET

Veri boyutları giderek artmakta ve bu verilerden anlamlı bilgiler elde etmek zorlaşmaktadır. Bu sebeple veri madenciliği giderek önem kazanmaktadır. Birçok alanda veri madenciliği yöntemleri yaygın bir şekilde kullanılmaktadır. Bu alanlardan birisi de eğitsel veri madenciliğidir. Eğitsel veri madenciliğinde geleneksel ve uzaktan eğitim türleri üzerine çalışılmaktadır. Geleneksel eğitim veren birçok üniversitede öğrenci başarısını etkileyen unsurlar araştırılmaktadır. Değişen eğitim politikaları da öğrenci başarısını etkileyebilecek unsurlardan birisidir. Bu tezin amacı, 2011 yılında değişen atılma politikasının, öğrencilerin parasal durumlarının ve demografik özelliklerinin, etkilerinin veri madenciliği yöntemleri ile incelenmesidir. Bu tez kapsamında, Eskişehir Osmangazi Üniversitesi Bilgisayar Mühendisliği Bölümü'nde 2008'den 2011'e kadar olan dört yıl içinde birinci sınıf derslerine başlayan öğrencilerin verileri kullanılmıştır. Yararlı olabilecek bilgiler çıkarılabilmesi için deneysel bir yaklaşımda bulunulmuş ve farklı veri setleri oluşturulmuştur. Bu veri kümeleri üzerinde veri madenciliği yöntemlerinden sınıflama ve birliktelik kuralları algoritmaları uygulanmıştır. Sınıflama algoritmalarından kNN ve J48, birliktelik kuralları algoritmalarından Apriori ve Predictive Apriori kullanılmıştır. Uygulanan yöntemler sonucunda, atılma politikasındaki değişiklik ile öğrencilerin not durumları arasındaki ilişki bulunmuş, öğrencilerin eğitim sürelerinin, burs veya kredi alınmasıyla değişiklik gösterdiği ve öğrencilerin parasal durumlarıyla annelerinin meslekleri arasında bağlantı olduğu görülmüştür.

Anahtar Kelimeler: Eğitsel veri madenciliği, eğitim politikası, üniversiteden atılma.

SUMMARY

Process of information retrieval from raw data is getting difficult due to increasing size of data. Therefore, data mining has a great importance to deal with this problem. Data mining techniques can be applied on many different areas. Educational data mining is one of those fields. Educational data mining is applied on two separate types of education, traditional and distance education. In many universities that give traditional education, researches are done on factors effecting students' success. Alteration on education policy is one of the factors that may affect student achievement. The aim of this thesis is examining the effects of financial status and demographic characteristics of students and expulsion policy that changed in 2011. Data mining techniques are performed to data that belongs to students from Computer Engineering Department of Eskişehir Osmangazi University. An experimental approach is used to obtain useful information from the data. For achieving this goal, student data is divided into different data sets. Classification and rule mining algorithms are used on these data sets. J48 and kNN are chosen for classification algorithm. Apriori and Predictive Apriori are chosen for rule mining algorithm. As a result of the methods applied, there are relations between following couples, expulsion policy and grades of students, scholarship and training times of students, financial status of students and profession of their mothers.

Keywords : Educational data mining, policy of education, expulsion from universities.

TEŞEKKÜR

Danışmanım Yrd. Doç. Dr. Celal Murat KANDEMİR'e gerek lisansüstü eğitimimde gerekse çalışmamda bana yol gösterdiği ve yönlendirdiği için teşekkür ederim.

Yrd. Doç. Dr. Nihat ADAR'a takıldığım sorularda her türlü yardımda bulunduğu için teşekkür ederim.

Bu süreçte beni her daim destekleyen ve yanımda bulunan aileme ve arkadaşlarıma teşekkür ederim.

İÇİNDEKİLER

Sayfa

ÖZET	vi
SUMMARY	vii
TEŞEKKÜR.....	viii
İÇİNDEKİLER.....	ix
ŞEKİLLER DİZİNİ.....	xi
ÇİZELGELER DİZİNİ.....	xiii
1. GİRİŞ.....	1
2. LİTERATÜR ARAŞTIRMASI.....	3
3. EĞİTSEL VERİ MADENCİLİĞİ	7
3.1. Geleneksel Eğitim.....	7
3.2. Uzaktan Eğitim	8
3.3. Veri Kaynakları.....	9
3.3.1. Anket	9
3.3.2. Web	10
3.3.3. Öğrenci bilgi sistemi	10
3.4. Veri Tabanlarında Bilgi Keşfi.....	10
3.4.1. Veri ön işleme	12
<u>3.4.1.1. Veri temizleme</u>	12
<u>3.4.1.2. Veri bütünleştirme</u>	13
<u>3.4.1.3. Veri azaltma</u>	15
<u>3.4.1.4. Veri dönüştürme</u>	15
3.4.2. Veri madenciliği	16
3.4.3. Örüntü değerlendirme ve sunum	16
3.5. Veri Madenciliği Yöntemleri	17
3.5.1. Tahmin edici.....	18
<u>3.5.1.1. Sınıflama</u>	18
<u>3.5.1.2. Regresyon</u>	19
<u>3.5.1.3. Kestirim</u>	19
<u>3.5.1.4. Zaman serileri analizi</u>	20

İÇİNDEKİLER (devam ediyor)

Sayfa

3.5.2. Tanımlayıcı.....	20
<u>3.5.2.1. Birliktelik kuralları</u>	20
<u>3.5.2.2. Özetleme</u>	21
<u>3.5.2.3. Kümeleme</u>	21
<u>3.5.2.4. Sıralı dizileri bulma</u>	22
4. VERİ MADENCİLİĞİ UYGULAMASI	23
4.1. Kullanılan Programlar	23
4.1.1. Weka.....	23
4.1.2. Microsoft Excel	25
4.2. Kullanılan Algoritmalar	25
4.2.1. Knn.....	26
4.2.2. J48	27
4.2.3. Apriori ve predictive apriori.....	28
4.3. Veri Seti ve Ön İşleme	28
4.4. Veri Madenciliği ve Çıktıları	37
4.4.1. Öğrenci işleri bilgi sistemi	37
4.4.2. Demografik özelliklerin öğrenci başarısı üzerine etkisi.....	48
4.4.3. Atılma politikasının öğrenci başarısı üzerine etkisi	56
5. SONUÇLAR	61
KAYNAKLAR DİZİNİ	63

ŞEKİLLER DİZİNİ

<u>Sekil</u>	<u>Sayfa</u>
3.1 Veri tabanlarında bilgi keşfi süreci (Fayyad, Shapiro ve Smyth, 1996).....	11
3.2 Veri bütünleştirme örneği a) birinci veri seti, b) ikinci veri seti, c) bütünleşmiş veri seti	14
3.3 Veri madenciliği yöntemleri (Aydın'dan, 2007)	17
4.1 “.csv” dosya uzantılı veri seti örneği	24
4.2 “.arff” dosya uzantılı veri seti örneği.....	24
4.3 En yakın komşuluk algoritmasının örnek çalışması	26
4.4 Hava durumu veri setinden elde edilebilecek ağaç yapısı	27
4.5 Genel not ortalamalarının dokuz kategoriye çevirim grafiği.....	35
4.6 Hazırlık sınıfını geçme notlarının altı kategoriye çevirim grafiği	36
4.7 2008 yılı, J48 algoritmasının oluşturduğu BursKredi sınıfının ağaç yapısı	39
4.8 2008 yılı, J48 algoritmasının oluşturduğu GNO sınıfının ağaç yapısı a) V1 veri seti için b) V2 veri seti için.....	40
4.9 2008 yılı, J48 algoritmasının oluşturduğu TopEgtYılı sınıfının ağaç yapısı.....	41
4.10 2009 yılı, J48 algoritmasının oluşturduğu BursKredi sınıfının ağaç yapısı	42
4.11 2009 yılı, J48 algoritmasının oluşturduğu GNO sınıfının ağaç yapısı	42
4.12 2009 yılı, J48 algoritmasının oluşturduğu TopEgtYılı sınıfının ağaç yapısı a) V1 veri seti için b) V2 veri seti için	43
4.13 2010 yılı, J48 algoritmasının oluşturduğu GNO sınıfının ağaç yapısı	44
4.14 2008-2011 yılları, J48 algoritmasının V1 veri seti için oluşturduğu BursKredi sınıfının ağaç yapısı.....	45
4.15 2008-2011 yılları, J48 algoritmasının V2 veri seti için oluşturduğu BursKredi sınıfının ağaç yapısı.....	46
4.16 2008-2011 yılları J48 algoritmasının oluşturduğu TopEgtYılı sınıfının ağaç yapısı	46
4.17 2008-2011 yılları için Apriori algoritmasının çıkardığı kurallar	47
4.18 2008-2011 yılları için Predictive Apriori algoritmasının çıkardığı kurallar.....	47

ŞEKİLLER DİZİNİ (devam)

<u>Sekil</u>	<u>Sayfa</u>
4.19 2008 yılı, demografik verilere uygulanan J48 algoritmasının oluşturduğu BursKredi sınıfının ağaç yapısı.....	49
4.20 2008 yılı, demografik verilere uygulanan J48 algoritmasının oluşturduğu TopEgtYılı sınıfının ağaç yapısı.....	49
4.21 2008 yılı, demografik verilere uygulanan J48 algoritmasının oluşturduğu GNO sınıfının ağaç yapısı.....	50
4.22 2009 yılı, demografik verilere uygulanan J48 algoritmasının oluşturduğu ağaç yapıları a) AnneEgt sınıfı için b) BabaEgt sınıfı için.....	51
4.23 2009 yılı, demografik verilere uygulanan J48 algoritmasının oluşturduğu BursKredi sınıfının ağaç yapısı.....	52
4.24 2010 yılı, demografik verilere uygulanan J48 algoritmasının oluşturduğu AnneEgt sınıfının ağaç yapısı.....	53
4.25 2008-2011 yılları, demografik verilere uygulanan J48 algoritmasının oluşturduğu BursKredi sınıfının ağaç yapısı.....	55
4.26 2008-2011 yılları demografik verilerde elde edilen birliktelik kuralları a) Apriori ile b) Predictive Apriori ile	55
4.27 J48 algoritmasının 2011 öncesi ve sonrası için çıkardığı ağaç yapısı	59
4.28 2011 yılı öncesi ve sonrası için elde edilen birliktelik kuralları a) Apriori ile b) Predictive Apriori ile.....	59
4.29 2011 öncesi, 2011 sonrası sınıflarının girdi sayısına göre dağılımı	60

ÇİZELGELER DİZİNİ

<u>Cizelge</u>	<u>Sayfa</u>
4.1 Öğrencilerin sağlık bilgilerinin bulunduğu veri setinin içeriği	29
4.2 Öğrencilerin demografik bilgilerinin bulunduğu veri setinin içeriği	30
4.3 Öğrencilerin aldıkları derslerle ilgili olan veri setinin içeriği	31
4.4 Öğrencilerin bilgi sistemindeki bilgilerinin bulunduğu veri setinin içeriği	32
4.5 Temel olarak oluşturulan Excel dosyaları ve kısa açıklamaları	33
4.6 Dörtlük not sistemi ile harf kodu karşılıkları.....	35
4.7 2008 yılı, V1 ve V2 veri setlerine uygulanan sınıflama algoritmalarının çıktıları	39
4.8 2009 yılı, V1 ve V2 veri setlerine uygulanan sınıflama algoritmalarının çıktıları	41
4.9 2010 yılı V1 ve V2 veri setlerine uygulanan sınıflama algoritmalarının çıktıları.....	44
4.10 2008-2011 yılları V1 ve V2 veri setlerine uygulanan sınıflama algoritmalarının çıktıları	45
4.11 2008 yılı demografik verilere uygulanan sınıflama algoritmalarının çıktıları	48
4.12 2009 yılı Demografik verilere uygulanan sınıflama algoritmalarının çıktıları.....	51
4.13 2010 yılı Demografik verilere uygulanan sınıflama algoritmalarının çıktıları.....	53
4.14 2008-2011 yılları Demografik verilere uygulanan sınıflama algoritmalarının çıktıları	54
4.15 2011 yılı öncesi ve sonrası sınıflaması için oluşturulmuş veri seti örneği	56
4.16 Algoritmaların zaman sınıfı için ürettiği sonuçlar.....	57
4.17 Algoritmaların üç öznitelikle zaman sınıfı için ürettiği sonuçlar	58

1. GİRİŞ

Bilgisayarların depolama özellikleri arttıkça saklanan verilerin de boyutları çok hızlı bir şekilde artmaktadır. Sürekli çoğalan veri yığınlarının arasından anlamlı bilgiler çıkarabilmek için veri madenciliğine ihtiyaç duyulmaktadır. Bilgisayar teknolojilerinin getirdiği veri işleme zorluğunu, bilgisayarlar üzerinde çalıştırılan veri madenciliği algoritmalarıyla aşmak mümkündür.

Veri madenciliği, verilerin sayısal olarak saklandığı her alanda kullanılabilir. Eğitim alanında yapılan çalışmalar ve uygulamalar, gün geçtikçe çoğalmaktadır. Öğrenci profili çıkartma, öğrenci destek sistemi, ders yönetim sistemi gibi benzeri birçok şekilde eğitim alanına uygulanabilir. Ülkemizde de eğitim alanında yapılan bu türden araştırmalar artış göstermektedir. Bu artışın sebebi, verilen eğitimin kalitesinde olumlu gelişmelerin sağlanması amacıyla.

Ülke genelindeki eğitim seviyesi, devlet politikalarına bağlı olduğu için, eğitim üzerine yapılan planlarda eğitsel veri madenciliğinin kullanılması, doğru kararların alınmasında etkili olacaktır. 2011 yılında yapılan değişiklikte üniversitelerden atılma kalkmıştır. Bu değişikliğin, üniversite öğrencileri üzerindeki etkisinin eğitsel veri madenciliği yöntemleriyle keşfedilmesi bu tezin ana amacıdır.

Eskişehir Osmangazi Üniversitesi öğrenci bilgilerinin tutulduğu sistemdeki 2008-2011 yıllarında birinci sınıf derslerine başlayan Bilgisayar Mühendisliği Bölümü öğrencilerinin verileri kullanılmaktadır. Üniversitelerden atılmanın kalkmasının etkilerinin yanı sıra bilgi sisteminde bulunan diğer verilerle ilgili de veri madenciliği çalışması yapılmaktadır.

Birinci bölümde, çalışmada gerçekleştirilmek istenen hedefler ortaya konmakta ve çalışmanın ne şekilde yürütüleceği anlatılmaktadır.

İkinci bölümde, veri madenciliği alanında yapılan çalışmaların ana uygulama alanlarındaki örnekleri verilmektedir. Eğitsel veri madenciliği alanında ise literatürde, bu çalışmada gerçekleştirilmek istenen hedeflere benzer amaçlar güden çalışmalardan bahsedilmektedir.

Üçüncü bölümde, eğitsel veri madenciliği alanından ve bu alanda kullanılan yöntemlerden bahsedilmektedir. Veri tabanlarında bilgi keşfi sürecinin işleyişi hakkında bilgiler verilmektedir.

Dördüncü bölümde, bu çalışmada kullanılan programlardan ve algoritmalarından, veri setlerinin oluşturulma sürecinden, bu süreçte yapılan ön işlemlerden ve veriler üzerine uygulanan veri madenciliği aşamasından elde edilen çıktılarından bahsedilmektedir. Veri madenciliği aşamasında özniteliklerin birbirleriyle olan ilişkileri, demografik özelliklerin başarıyı nasıl etkilediği ve atılmanın kalkması ile öğrenci başarısında nasıl değişiklikler olduğu incelenmektedir.

Beşinci bölümde, elde edilen sonuçlar değerlendirilmektedir ve gelecek çalışmalarda nelere dikkat edilmesi gerektiği hususunda fikirler sunulmaktadır.

2. LİTERATÜR ARAŞTIRMASI

Veri madenciliği uygulanabilecek alanlar açısından çok geniş bir yelpazeye sahiptir. Bunun altında yatan sebep, verilerin olduğu her alanda veri madenciliğinin kullanılabilmesidir. Veri madenciliğinin başlıca kullanıldığı alanlar aşağıda verilmiştir.

- Pazarlama
- Bankacılık
- Sağlık
- İnternet
- Telekomünikasyon
- Astronomi
- Eğitim

Pazarlama alanında yapılan çalışmalarda, müşteri profilinin belirlenmesi, malların müşteri özelliklerine göre nasıl pazarlanacağı ve müşterilerin elde nasıl tutulacağı gibi amaçlar güdülmektedir. Ling ve Li (1998), tarafından yapılmış çalışmada veri madenciliğinin pazarlama dalında karşılaştığı sorunlar ve çözümleri üzerinde durulmuştur. Veri madenciliğinin pazarlama alanında yapılan uygulamalarında net kârın ve yatırımın geri dönüşünün unutulmaması yönünde öneride bulunulmuştur.

Bankacılık sektöründe, kredi vermeye uygun müşterilerin belirlenmesi, kredi kartı dolandırıcılıklarının tespit edilmesi, müşterilere uygun yatırım seçenekleri sunulması gibi çalışmalar yapılmaktadır. Fabris'e (1998) göre, değişen faiz oranlarına müşterilerin nasıl tepki vereceğinin tahmin edilmesinde, yeni sunulacak tekliflerden hangi müşterilerin yararlanmak isteyeceğinin belirlenmesinde ve benzeri durumlarda bankanın en çok kâr elde etmesinde veri madenciliği yöntemleri kullanılacaktır (Hormozi ve Giles, 2004).

Sağlık alanında, veri madenciliği hastalıklı dokuların tespit edilmesinde, gen dizilimleri ile ilgili veriler üzerindeki çalışmalarda, tanı konulmasında destek olacak sistemlerin tasarlanmasında kullanılabilir. Brossette vd. (1998) tarafından yapılan çalışmada, hastanelerde çıkabilecek salgınların tespit edilmesi için veri madenciliği yöntemlerini kullanan gözetim sistemi önerilmiştir. Hastanenin eski kayıtlarında bulunan

P. Aeruginosa hastalığının kontrol verileri üzerinde birliktelik kuralları ile yapılan çalışmadan elde edilen yararlı bilgilerle sistemin faydalı olabileceği yönler ortaya konulmuştur.

İnternet alanındaki çalışmalar çok geniş bir alanı kapsamaktadır. Arama motorlarının, kullanıcıların arama alışkanlıklarına göre örüntü çıkarıp, kullanıcının tercihlerine uygun sitelere yönlendirmesi, bu çalışmalardan birine örnek gösterilebilir. Cooley vd. (1998) tarafından yapılan çalışmada, internet sitesinin kullanılma analizi yapılarak, sitenin tasarımında yararlanabilecek bilgiler elde edilmek istenmiştir. Web sitesinden elde edilen verilerin kullanılabilmesi için gereken ön işlemler üzerinde özellikle durulmuştur.

Telekomünikasyon alanında, firmaların arasındaki rekabetin de etkisiyle veri madenciliği yöntemleri yaygın bir şekilde kullanılmaktadır. Özellikle müşteri davranışlarını incelemek ve müşterilerin isteklerine göre kampanyalar sunmak, müşteri memnuniyetini arttırmaktadır. Aldığı hizmetten müşterinin hoşnut olması da telekomünikasyon şirketini rekabette öne geçmesini sağlayabilmektedir. Guifung ve Youshi (2010), tarafından yapılmış çalışmada telekomünikasyon şirketlerinin, müşteri ilişkilerini yönetme konusundaki uygulamaları özetlenmiştir.

Astronomi alanında veri madenciliği yöntemlerinin kullanılması zorunluluk haline gelmiştir. Gelişen teleskop teknolojileriyle veri boyutları da yükselmiştir. Borne (2009), tarafından yapılmış çalışmada veri madenciliğinin astronomi alanında kullanılması anlatılmıştır. Kümeleme ve sınıflama algoritmalarının yaygın olarak, astronomi alanında yıldız gruplarının kümelenmesi ya da galaksilerin sınıflanması gibi problemlerde kullanılmakta olduğunu belirtmiştir.

Veri madenciliğinin eğitim alanına uygulanması ile birlikte “Eğitsel Veri Madenciliği” olarak adlandırılan bir sektör oluşturmuştur ve bu sektörde yapılan çalışmalar katlanarak artmaktadır. Romero and Ventura (2007), 1995-2005 arası EVM’de yapılmış araştırmaları inceledikleri bir çalışma yapmışlardır ve 2005 yılından bu zamana on yıl geçmiştir. Bu süre zarfında yapılan çalışmaların sayısı önceki yıllara göre çok daha fazladır. Bu süreçte yapılan araştırmaların hepsinin incelenmesi, tez amacına uygun olmadığı için üniversite öğrencilerinin başarılarını etkileyen faktörler ve üniversiteden

atılma durumlarını inceleyen çalışmalarla sınırlandırıldı. Bu kısıtlar göz önünde bulundurularak, yapılan çalışmalar metnin devamında belirtilmiştir.

İnan'ın (2003) çalışmasında, Konya Selçuk Üniversitesi öğrenci işleri otomasyonundan elde edilen veriler veri madenciliği yöntemleriyle incelenmiştir. Üniversiteyi yeni kazanan bir öğrencinin başarısına etki eden etmenler ve öğrencinin üniversite eğitimi boyunca izleyeceği evreler çıkarılmıştır.

Erdoğan'ın (2004) çalışmasında, Maltepe Üniversitesi öğrenci işleri veri tabanından elde edilmiş öğrenci kayıtları kullanılmıştır. 2013 yılında Maltepe Üniversitesi'ni kazanan 3876 öğrencilik girdi, ön işleme aşamasından sonra 722' ye düşürülmüştür. Öğrencilerin yerleşme puan yüzdeleri, başarı notları, lise türü ve fakülte kodu gibi öznitelikler seçilmiş olup K-means algoritması uygulanmıştır.

Altınışık'ın (2006) çalışmasında, Kocaeli Üniversitesi öğrenci bilgi sistemi üzerinde Apriori algoritması kullanılarak veri madenciliği uygulaması gerçekleştirilmiştir. Öğrencilerin başarısız olduğu dersler arasındaki birliktelikler bulunup, öğrencilerin ders seçimi aşamalarında yardımcı olunması amaçlanmıştır.

Saygılı'nın (2013) çalışmasında, mühendislik fakültesi öğrencilerinin demografik bilgileri, ÖSYS puanları ve üniversiteye yerleştirilme sıraları ile kazandıkları bölümlerdeki ağırlıklı başarı not ortalamaları giriş verisi olarak kullanılmıştır. Öğrencilerin gelmiş oldukları bölgeler ve okul türleri üzerinde kümeleme analizi yapılmıştır. Öğrencilerin belli bir kısmına, aile bilgileri için anket yapılmıştır ve bu verilerin öğrencilerin başarılarını nasıl etkilediği incelenmiştir.

Taşdemir'in (2012) çalışmasında, 2010 yılında Dicle Üniversitesi'ne yerleşen -tıp fakültesi öğrencileri ve hazırlık sınıfı okumuş öğrenciler hariç- öğrencilerin not ortalaması, kaldıkları ders sayısı, fakülteleri, giriş yaşı, ÖSS puanları, cinsiyetleri, ebeveynlerinin eğitim durumları gibi bilgilerinden yararlanılmıştır. Örneğin "Cinsiyetin not ortalamasına etkisi vardır." şeklinde hipotezler oluşturulup çıkan sonuçlarla karşılaştırılmıştır.

Ekim'in (2011) çalışmasında, Selçuk Üniversitesi öğrenci işleri otomasyonundan elde edilen veriler kullanılmıştır. Veriler Apriori ve karar ağacı algoritmaları ile incelenmiştir. Öğrencilerin ailelerinin eğitim seviyelerinin ve gelir düzeylerinin, başarılarına en çok etki eden faktörler olduğu sonucuna ulaşılmıştır.

Gülçe'nin (2010) çalışmasında, Pamukkale Üniversitesi'ne 2009 yılında kayıt yaptıran 3588 öğrencinin bilgilerinden yararlanılmıştır. Öğrencilerin fakülte, not ortalaması, cinsiyet, giriş tipi, medeni hal, yaş, lisedeki derecesi, ÖSS puanı, tercih sırası ve yüzdelik dilim bilgilerinden yararlanılmıştır. Başarısız sayılabilecek öğrencilerin -risk grubundaki öğrencilerin- önceden belirlenmesi amaçlanmıştır.

Güvenç'in (2001) çalışmasında, son iki sene içinde mezun olmuş öğrencilerin profilleri çıkarılıp performansları ölçülmüştür. Aynı öğrencilerin üniversite eğitimleri süresince aldıkları dersler ve bu derslere ait notlar göz önüne alınarak, öğrencilerin ders seçimi davranışları ve başarıları incelenmiştir. Son aşama olarak, öğrencilere ait kişisel bilgilerle de birleştirilerek sınıflama problemi şeklinde bir yaklaşımda bulunulmuştur. Birlikte kuralları, karar ağaçları ve istatistiksel analize dayalı veri madenciliği teknikleri kullanılmıştır.

Özarslan ve Barışçı'nın (2014) çalışmasında, Kırıkkale Üniversitesi öğrenci bilgi sisteminden alınan ön lisans ve lisans öğrencilerine ait 672 adet veri kullanılmıştır. Öğrencilerin Temel Bilgi Teknolojileri Kullanımı dersine ait başarı notları, üniversiteye geliş puan türleri, fakülte-yüksekokul türleri, başarı durumları, yaş aralıkları, örgün ya da uzaktan eğitim aldıkları ve cinsiyet bilgileri öznitelik olarak kullanılmıştır. Başarılarıyla ilişkilendirilebilecek faktörler belirlenmeye çalışılmıştır.

3. EĞİTSEL VERİ MADENCİLİĞİ

Eğitsel veri madenciliği, eğitim alanından gelen birçok türdeki verileri, veri madenciliği algoritmalarıyla irdeleyen bir alandır. Veri madenciliğinin eğitim alanında kullanılması giderek yaygınlaşmaktadır ve bu alanda yapılan çalışmaların sayısı da artmaktadır (Romero ve Ventura, 2007).

Eğitsel veri madenciliğinin giderek yaygınlaşmasının altında yatan ana sebeplerden birisi öğrenci sayılarının artması ve buna karşın öğretim üyelerinin sayısının yetersiz kalmasıdır. Bu durum, öğrencilerle bire bir ilgilenilmesini zorlaştırmakta ve eğitim kalitesinin belirli bir seviyede tutulmasına engel olmaktadır. Bu sebeple öğrencilerle daha çok ilgilenilmesi ve aldıkları eğitimin kalitesinin artırılabilmesi için veri madenciliği eğitim alanında kullanılmaktadır.

Eğitsel veri madenciliği uygulamaları, verileri edindikleri ortamlara göre sınıflandırılacak olursa, geleneksel ve uzaktan eğitim olmak üzere ikiye ayrılmaktadır. Geleneksel eğitim, öğrenci ve öğretmenin aynı çatı altında bulundukları bir ortamken, uzaktan eğitimde öğrenci ve öğretmen farklı iletişim yollarıyla eğitimi sürdürmektedirler. Eğitim türleri arasındaki bu fark toplanan verilerin ve toplanma şekillerinin de farklı olmasına sebep olmaktadır.

Eğitim alanında veri madenciliği çalışmaları yapılabilmesi için öncelikle iki eğitim türü arasındaki farklılıkların, net bir biçimde ortaya konulması gerekmektedir. Eğitim türlerinin özelliklerinin belirlenmesi, yapılacak çalışmalarda uygulanabilecek yöntemlere karar verilmesinde kolaylık sağlamaktadır.

3.1. Geleneksel Eğitim

Eğitsel veri madenciliği, toplanan verilere göre farklı amaçlarla uygulanabilir. Geleneksel eğitim türünde öğretmen, öğrencilerin sözlü ve yazılı sınavları, projeleri ya da ödevleri gibi kaynaklardan gelen verileri değerlendirmekle kalmayıp, aynı zamanda

öğrencilerin sınıf içindeki tutumları, derse katılım şekilleri ve ders içindeki sorulara verdikleri cevapları da hesaba katmaktadır. Öğretmenin gerçekleştirdiği olgu aslında veri madenciliğidir.

Öğrencilere ait tuttuğu verileri, öğrenciler hakkında model oluşturmak için kullanır. Bu modele göre eğitim tarzında geliştirmeler yapabilir, başarısız olarak nitelendirdiği öğrencilerine farklı yaklaşımlarda bulunup, başarılı olmaları için yönlendirebilir. Az mevcutlu sınıflarda öğretmenler, bu şekilde çalışmalar yapabilirken giderek artan öğrenci sayısı, verimliliği düşürmüştür.

3.2. Uzaktan Eğitim

Uzaktan eğitim alanında, öğrenciler ile öğretmenler yüz yüze iletişim halinde olamadıkları için, öğretmenler öğrenciler üzerinde kişisel gözlemlerde bulunamamaktadır. Öğrencilerle ilgili veriler notlarla sınırlı kalmaktadır. Öğrenciler hakkında daha fazla veri elde etmek için anket uygulamak, düşünülen ilk çözümlerden birisidir. Anket uygulanmasındaki amaç öğrenciler hakkında mümkün olduğunca bilgi alınması ve bu bilgiler ışığında eğitim kalitesini artırılmasıdır.

Örneğin, Talavera ve Gaudioso (2004) tarafından yapılan çalışmada, öğrencinin öğrenme yönetim sistemiyle olan etkileşim verileri kullanılmıştır. Ziyaret edilen ders sayfaları, dersle ilgili foruma öğrenci tarafından atılan mesaj sayısı gibi bilgiler etkileşim verileri olarak kullanılmıştır. Bu bilgilerin daha da zenginleştirilebilmesi için de, demografik ve benzeri bilgiler (öğrenci ile ilgili daha farklı bilgiler) anketler aracılığıyla temin edilmiştir.

İnternetin gelişmesiyle, öğrencilerle ilgili edinilebilecek veriler de artmıştır. Çevrimiçi yapılan dersler öğrenci-öğretmen arasındaki bağlantıyı, sanal da olsa kurduğu gibi öğrencinin dersle ilgili, site içi hareketlerinin log dosyaları şeklinde tutulması sayesinde öğrenci modellemesi yapılabilir. Örneğin, öğrencilerin dersle ilgili site içi hareketleri ve başarıları arasındaki ilişkiler bulunabilir. Ders yürütücüleri, bulunan örüntüleri dikkate alarak planlarını oluşturabilir.

Singh ve Lal (2013), öğrenim yönetim sistemi olan Moodle ortamındaki derslerin iyileştirilmesi için birliktelik kurallarını kullanarak bir çalışma yapmışlardır. Çalışmada öğrencilerin Moodle ortamında hareketlerinin dokümanları kullanılarak, sayısal ortamdaki derslerin iyileştirilmesi hedeflenmiştir.

3.3. Veri Kaynakları

Eğitsel veri madenciliği, farklı kaynaklardan gelen verileri kullanabilmektedir. Geleneksel eğitimdeki kişisel gözlemlerle edinilen veriler dahil edilmezse, yöntemler aşağıdaki biçimde ana başlıklar altında toplanabilir.

- Anket
- WEB
- Öğrenci Bilgi Sistemi

3.3.1. Anket

Öğrencilerle ilgili veriler birçok farklı kaynaktan elde edilebilir. Bu kaynaklardan birisi de hedeflere yönelik oluşturulan anketlerdir. Gerçekleştirilmek istenen çalışmada, öğrencilerle ilgili gerekli görülen bilgiler, oluşturulacak bir anketle elde edilebilmektedir. Anket çalışmasında, Likert tipi anketlerde olduğu gibi sorulara yönelik hazır bulunan cevaplardan seçim yapılmasını beklemek şart değildir. Oluşturulan sorularda, anket hazırlayıcısı tarafından önceden belirlenmemiş cevaplar istenmesi yoluna gidilmesi de kullanışlı olabilmektedir. Böylece öğrenciler, seçeneklerle sınırlandırılmamış olmaktadır. “Eğitiminizi engelleyen herhangi bir rahatsızlığınız var mı? Varsa yazınız.” şeklinde bir soru iyi bir örnek olabilir. Yalnız bu tür sorularda dikkat edilmesi gereken husus, verilen cevapların nasıl değerlendirileceğinin iyi düşünülmesi gerektiğidir. Cevaplar rahatsızlığın olup olmadığı şeklinde değerlendirilebileceği gibi rahatsızlığın türüne yönelik bir değerlendirme de yapılabilmektedir.

3.3.2. Web

İnternetin gelişmesi ve kullanımının yaygınlaşması sonucunda, eğitim aynı zamanda sayısal ortamda da verilmektedir. Geleneksel eğitim destek olarak interneti kullanırken, uzaktan eğitimin büyük bir kısmı internet ortamında yapılmaktadır. Öğrencilerin internete erişebildikleri her ortamdan eğitimlerini sürdürebilmeleri uzaktan eğitimin başka boyutlara gelmesini sağlamaktadır.

İnternet üzerinde gerçekleştirilen eğitimde öğrencilerle ilgili bilgiler değişik şekillerde olabilmektedir. Eğitim için oluşturulmuş bir site ele alınırsa, öğrencilerin o sitede bulundukları süre, site içerisinde etkileşime geçtikleri bağlantılar, görüntüledikleri ders dokümanları, site içi yazışma imkanı varsa yazışma konuları, sınavlar site içinde gerçekleştirilmekteyse sınavı tamamlama süreleri, sınav sonuçları gibi birçok bilgi veri madenciliği uygulamalarında kullanılabilmektedir.

3.3.3. Öğrenci bilgi sistemi

Eğitim kurumları öğrenciler hakkında geniş kapsamlı bilgiler saklamakta ve bu verileri öğrenci bilgi sistemleri için oluşturulmuş veri tabanlarında tutmaktadırlar. Bu veri tabanları öğrencilerin bulundukları kurumda, aldıkları notlar, yaşları, memleketleri, önceki kurumlardan gelmiş olabilecek bilgiler ve benzeri birçok veri açısından zengindir. Dolayısıyla bu keşfedilmeyi bekleyen zenginlik, veri madenciliği uygulamaları için çok elverişli bir ortam sağlamaktadır.

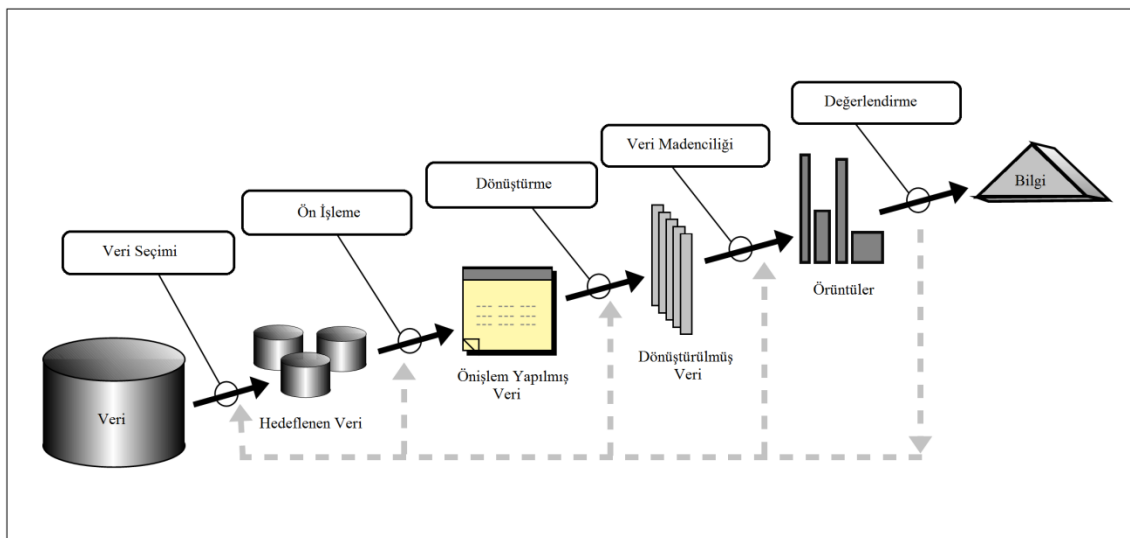
3.4. Veri Tabanlarında Bilgi Keşfi

Gelişen teknolojiyle büyük boyutlarda verilerin saklanması kolaylaşmıştır. Bu nedenle, özel kuruluşların, devlet kurumlarının ve hatta bireysel bilgisayar kullanıcılarının, daha sonra faydalı olabilir düşüncesiyle sakladıkları veriler kargaşaya sebep olmuştur.

Karşılaşılan problemin sebebi bu şekilde depolanan verilerden yararlanılmak istenildiğinde, büyük boyutlu ve aynı düzene sahip olmayan verilerin içinden işe yarar bilgilerin çekilmesinin zor olmasıdır.

Veri analizleri istatistiksel yöntemlerle uzun bir süredir yapılmaktadır. Fakat artan veri boyutları karşısında istatistiksel analizler yetersiz kalmaktadır. Ham verilerin artış hızıyla karşılaştırıldığında, bu ham verilerden elde edilen yararlı bilgilerin artış hızının çok yavaş olduğu gözlenmektedir. Sürekli artan bu farkın kapatılıp, verilerden daha çok faydalı bilgiler çıkarılabilmesi için veri tabanlarında bilgi keşfi olarak bilinen süreç doğmuştur.

Bu süreç, makine öğrenimi, örüntü tanıma, istatistik, yapay zeka gibi birçok farklı alandan faydalanmaktadır. Sürecin genel olarak işleyişi Şekil 3.1’de verilmiş olup, sürecin birbirini takip eden aşamalardan oluştuğu görülmektedir. Süreç aynı zamanda kendini tekrar eden bir yapıya sahiptir. Değerlendirme aşamasından sonra elde edilen bilginin anlamsız olması ya da istenilen kalitede olmaması gibi durumlarda sürecin önceki aşamalarından birine dönmek gerekebilmektedir. Dönülen aşamadan itibaren yapılan işlemler kontrol edilerek süreç tekrar edilebilmektedir. Veri tabanlarında bilgi keşfi bu özelliğiyle, deneysel bir yaklaşıma da sahip olmaktadır.



Şekil 3.1: Veri tabanlarında bilgi keşfi süreci (Fayyad, Shapiro ve Smyth, 1996)

3.4.1. Veri ön işleme

Veri tabanlarında bilgi keşfi sürecinin en önemli ve üzerinde en çok zaman harcanması gereken adımı olduğu söylenebilir. Verilerin kullanıma hazır hale getirilmesine, ön işleme süreci denir. Veriler ham haliyle, yanlış girilen, eksik olan veya büyük boyutlarda değerlere sahip olabilir. Verilerde, insan kaynaklı bir sorundan, bir yerden bir yere iletimi sırasındaki kayıptan, sensörlerden alınan değerlerde oluşabilecek donanımsal sorunlardan vb. durumlardan oluşabilen hatalar olabilir. Ham veriler üzerinde çalışılmaya başlandığında, ilk yapılması gereken, verileri kullanılabilecek bir biçime çevrilmesidir.

Ham verilerde karşılaşılan sorunlar şunlardır; zamanla değişebilen, alakasız, eksik ve gürültülü verilerdir. Veriler, veri seti oluşturulmadan önce bu tür sorunlar giderilip kullanıma uygun hale getirilmelidir. Eksik olan değerler doldurulmalı, yanlış olan değerler düzeltilmeli ve büyük boyutlardaki değerler kabul edilebilir bir ölçeğe çekilmelidir. Eğer ön işleme kısmı gerektiği gibi yapılmazsa, verilerden çıkarılacak sonuçlar hatalı olacaktır.

Ön işleme süreci birden fazla alt aşamalardan oluşmaktadır ve aşamalar şunlardır:

- Veri temizleme
- Veri bütünleştirme
- Veri azaltma
- Veri dönüştürme

3.4.1.1. Veri temizleme

Veriler ham haliyle birçok hataya ve tutarsızlıklara sahip olabilmektedir. İlk olarak bu hataların ve tutarsızlıkların giderilmesi gerekmektedir. Hataların giderilmesi daha kaliteli verilerle çalışılması anlamına gelmekte ve düzeltilmiş verilerden çıkarılacak sonuçların da kaliteli olmasını sağlamaktadır.

Verilerin kaydı sırasında, yazım yanlışlıklarından, girilmeyen değerlerden, aynı girdinin birden fazla kez yer almasından vb. sebeplerden kaynaklanan hatalar ve tutarsızlıklar olmaktadır. Örneğin, bir hastalığa yakalanma riskinin yaşla ilgisi olup olmadığının araştırıldığı bir durumda, yaş bilgisinin hatalı yazıldığı veya hiç yazılmadığı bireylerin girdileri çalışmanın doğruluğunu zedeleyecektir. Eğer yaş bilgileri, doğruluğuna güvenilen kaynaklardan düzeltilemiyorsa, eksik bilgisi olan bireyin kaydı tamamen çıkarılmalıdır. Birebir aynı olan girdilerden, yalnızca bir tanesi bırakılmalıdır. Çünkü bu durum girdinin, veri setindeki etkisini arttırmaktadır ve yanlış sonuçlara yol açmaktadır.

Marquez-Vera vd. (2013) tarafından yapılan çalışmada veri temizleme kısmı üzerinde özellikle durulmuştur. Sonuçları etkilememesi için eksiksiz bilgileri olan girdiler ve öznitelikler kullanılmıştır. Veri temizleme kısmında veri ön işlemede uygulanan (sürekli verileri, kesikli verilere dönüştürme gibi) diğer aşamalar da tatbik edilmiştir.

Sürecin özünde, veri temizlemekten kasıt, verilerin düzeltilmesidir. Çoğu durumda, elle yapılmasını gerektiren sorunlarla karşılaşmakta ve bu durum çok zaman harcanmasına neden olmaktadır.

3.4.1.2. Veri bütünleştirme

Kullanılacak veri kümesi oluşturulurken birçok farklı kaynaktan yararlanmak gerekebilir. Bu kaynaklardan gelen verilerin, birbirleriyle uyumlu bir biçimde, tek bir kaynakta toplanması önem arz etmektedir.

Örnek verilecek olursa, bir kaynakta öğrencilerin her biri bir girdi olsun ve öğrencilerle ilgili özelliklerde girdilerin özniteliklerini oluştursun. İkinci kaynakta ise girdi olarak, öğrencilerin eğitim hayatları boyunca aldıkları dersler ve girdilerin öznitelikleri olarak da bu derslerle ilgili (dersin kredisi, alınan başarı notu, harf kodu gibi) bilgiler bulunsun. İki kaynağın birleştirilmesindeki sorun, hangi kaynaktaki girdinin temel olarak kabul edileceğidir. Burada göz önünde bulundurulması gereken nokta, çözümün yapılması istenilen çalışmaya göre değişecektir.

Eğer öğrencilerin başarılarıyla ilgili bir çalışma düşünülüyorsa, öğrenciler girdi olarak kullanılır ve derslerle ilgili bilgiler özniteliğe dönüştürülür. Aksi durumda, dersler girdi olarak ele alınıp, dersleri alan öğrencilerin verileri öznitelik haline dönüştürülmelidir. Şekil 3.2'deki temsili veri setleri ile veri bütünleştirme örneği verilmiştir.

<i>Öğrenci No</i>	<i>Not Ortalaması</i>	<i>Yaş</i>
1	3,43	21
2	3,55	19
3	2,3	20
4	3	18

(a)

<i>Öğrenci No</i>	<i>Dersler</i>	<i>Kredi</i>	<i>Not</i>
1	Matematik	4	3
1	Türkçe	3	4
2	Matematik	4	4
2	Türkçe	3	3
2	Fizik	3	3,5
3	Matematik	4	2
3	Türkçe	3	2,5
3	Fizik	3	2,5
4	Matematik	4	3
4	Türkçe	3	4
4	Fizik	3	2

(b)

<i>Öğrenci No</i>	<i>Not Ortalaması</i>	<i>Yaş</i>	<i>Matematik</i>	<i>Türkçe</i>	<i>Fizik</i>
1	3,43	21	3	4	-
2	3,55	19	4	3	3,5
3	2,3	20	2	2,5	2,5
4	3	18	3	4	2

(c)

Şekil 3.2: Veri bütünleştirme örneği a) birinci veri seti, b) ikinci veri seti, c) bütünleşmiş veri seti

3.4.1.3. Veri azaltma

Veri kümesinin büyüklüğü, üzerinde çalıştırılacak algoritmaların verimini doğrudan etkileyecektir. Aynı sonuçları verebilecek daha küçük bir veri kümesi oluşturulması işlem süresini düşürmeye yarayacaktır.

Veri boyutlarının düşürülebilmesi için örnekleme yapılabilir. Veri kümesinin girdi dağılımına, eşdeğer özellikte olan daha az girdili örnekler oluşturulabilir. Başka bir yaklaşım da, veri kümesinin istenilen bir özelliği benzer olan, eşit sayıda girdilerden oluşturulması olabilir. Bu yaklaşım, seçilen özelliğe göre sınıflandırma yapılması istenildiğinde, yanlış sınıflandırma yapılmasını önleyebilir.

Gereksiz özniteliklerin elenmesi de veri boyutlarının azaltılmasında etkili bir eylemdir. Sonuca etki etmeyen özniteliklerin veya gürültü olarak tanımlanabilecek (algoritmanın başarılı sonuç almasını engelleyebilecek) özniteliklerin çıkarılması, sürece olumlu yönde etki edecektir.

Talavera ve Gaudioso (2004) tarafından, öğrenme yönetim sistemleri üzerinde öğrenci karakterlerini oluşturmaya yönelik yapılan çalışmada veri azaltma yoluna gidilmiştir. Öğrencilerin hiçbir bilgi sağlamadıkları öznitelikleri ya da sistemin otomatik olarak aynı bilgilerle doldurduğu öznitelikler veri setinden çıkarılmıştır. Buna ek olarak çok az öğrencinin verileri bulunan öznitelikler de kaldırılmıştır.

3.4.1.4. Veri dönüştürme

Her algoritmanın farklı bir özelliği vardır. Çalışabildikleri veri türleri de bu özelliklerden birisidir. Verileri kullanıma hazır hale getirilirken uygulanması gereken işlemlerden bir diğeri de dönüştürmedir. Kullanılacak algoritmalar sayısal, kategorik veya mantıksal bir ve sıfır değerlerinden hangisiyle çalışıyorsa o yönde bir dönüştürme uygulanmalıdır.

Değerlerin büyüklüğü de ayrı bir sorun olabilmektedir. Çok büyük sayılara sahip veri kümesinde, değerler $[0,1]$ ya da $[-1,1]$ aralıklarına ölçeklendirilebilir. Farklı kaynaklardan gelen aynı özneliliklerin birimleri de farklı olabilmektedir. Bu durumda birimler arası dönüştürme yapılması gerekmektedir. Kumar ve Chadha (2012) tarafından yapılan çalışmada öğrencilerin yüzlük sistemdeki notları üniversitenin harf kodlarına dönüştürülmüştür.

3.4.2. Veri madenciliği

Bilgi keşfi sürecinin akıllı yöntemler uygulandığı aşamasıdır. Uygulamada ulaşılmak istenilen amaçlara göre uygun olan algoritmalar bu aşamada çalıştırılır ve bunun sonucunda elde edilen çıktılar bir sonraki aşamada yorumlanmak üzere değerlendirmeye alınır.

Veri madenciliği aşaması için üretilmiş birçok algoritma vardır. Tatbik edilecek algoritmanın belirlenmesi, verilerin özelliklerine ve kullanıcının tercihi kalmaktadır. Veri madenciliği aynı zamanda deneysel bir çalışma olduğu için birden fazla algoritmanın ayrı ayrı sınanmasında fayda vardır. Bu şekilde üzerinde çalışılan verilerde en başarılı sonuçları veren algoritma ya da algoritmalar belirlenebilir. Bilgi keşfi sürecinin gidişatı konusunda karar verilmesinde de yardımcı olacaktır.

3.4.3. Örüntü değerlendirme ve sunum

Algoritmaların çıktıları, ilk halleri ile kolay anlaşılır durumda olmayabilmektedir. Çıktılar görsel olarak düzgün bir şekilde sunulursa, elde edilen bilgilerin doğru yorumlanmasında sunumun yardımı olacaktır.

Elde edilen sonuçlar aşağıdaki yöntemlerle sunulabilmektedir.

- Çizelgelerle
- Doğru grafikleriyle

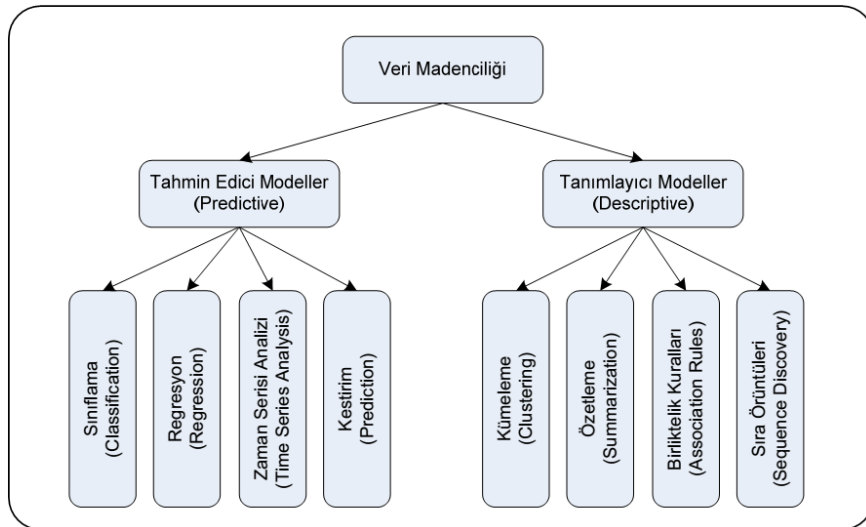
- Ağaç yapısında
- Kurallar aracılığıyla
- Girdi bazında
- Kümeleme yoluyla

3.5. Veri Madenciliği Yöntemleri

Veri madenciliği alanında tasarlanan her bir algoritmanın farklı özellikleri bulunmaktadır. Bu farklılıkların oluşmasındaki ana unsur, çözüm olarak üretildikleri problemlere özgü olmalarındandır.

Bu algoritmaların özellikleri ayrıntılı bir biçimde incelendiğinde, birbirleriyle olan farklılıkları net bir şekilde görülmektedir. Fakat farklı problemlerde hangi algoritmanın kullanılabileceğinin belirlenmesinde, ilk olarak çözülmesi gereken sorun ana hatlarıyla ortaya konulmalıdır.

Bu düşünceyle veri madenciliği yöntemleri ana hatlarıyla sınıflandırılmıştır. Yaygın olarak kullanılan yaklaşım Şekil 3.3'teki gibidir.



Şekil 3.3: Veri madenciliği yöntemleri (Aydın'dan, 2007)

3.5.1. Tahmin edici

Elde bulunan veriler işlenerek bir örüntü oluşturulup, bu örüntü yardımıyla farklı veriler üzerinde tahminde bulunulmaya çalışılıyorsa, kullanılabilecek yöntemler tahmin edici başlığı altında gruplandırılmıştır.

3.5.1.1. Sınıflama

Eldeki verilerin önceden tanımlanmış olan kategorilerden hangisine ait olduğunu tespit eden yöntemler bütünüdür. Sınıflama problemi olarak ele alınabilecek her problemde kullanılabilmektedir. Sınıflama algoritmaları iki ana adımdan oluşmaktadır. Birinci adımda hangi sınıfta olduğu belirli olan veri seti yardımıyla algoritma bir model oluşturmaktadır. Bu sürece algoritmanın öğrenme süreci de denilmektedir. Algoritmanın ikinci adımında ise girdileri hangi sınıfa ait olduğu bilinmeyen veri seti önceki adımda oluşturulan model yardımı ile sınıflandırılmaktadır. Bu adıma da deney aşaması denilmektedir. Deney aşamasında öğrenme sürecinde oluşturulan model ile girdilerin hangi sınıflara ait olduğu belirlenmektedir.

Sınıflama algoritmaları kullanılacakları probleme, veri setine ve bu gibi özelliklere göre değişiklikler göstermektedirler. Yaygın bir şekilde kullanılan algoritmalar şunlardır (Kesavaraj ve Sukumaran, 2013).

- Karar ağaçları
- Kural tabanlı
- Hafıza tabanlı öğrenme
- Yapay sinir ağları
- Bayes ağları
- Destek vektörü makineleri

Shah ve Jivani (2013) tarafından yapılan çalışmada göğüs kanserinin tespitinde *karar ağaçları*, *bayes ağları* ve *k-en yakın komşuluk* sınıflama algoritmalarının başarıları karşılaştırılmıştır.

3.5.1.2. Regresyon

Kudyaba ve Hoptroff'a (2001) göre, regresyon bağımsız değişkenler ile bağımlı değişken arasındaki ilişkiyi en iyi tanımlayan fonksiyonu elde etmek için uygulanan istatistiksel tekniklerdir (Aydın,2007). Veri setinde sınıf olarak tanımlanmış öznitelik bağımlı değişken, geri kalan öznitelikler ise bağımsız değişkenler şeklinde tanımlanmaktadır. Regresyon sürekli verilerle çalışmakla beraber kesikli verilerle de çalışabilmektedir. Kesikli verilerle elde edilen regresyon modeli sürekli veriler için de uygulanabilmektedir. Model olarak fonksiyon oluşturduğu için gelecekteki sonuçların da tahmin edilmesinde kullanılabilir. Doğrusal, sayısal ve lojistik regresyon gibi, farklı verilere uygulanabilen türleri mevcuttur.

Taşdemir'in (2012), çalışmasında öğrenci başarısına etki eden faktörler regresyon analizi ile tespit edilmiştir. Çalışmada *fakülte, giriş yaşı, öss puanı, tercih sırası, cinsiyet, anne öğrenim durumu, baba öğrenim durumu, bekleme süresi* öznitelikleri bağımsız değişken olarak kullanılmıştır. Bağımlı değişken olarak ise *not ortalaması* ve *kaldığı ders sayısı* kullanılmıştır.

3.5.1.3. Kestirim

Sınıflamanın aksine, kestirimde önceden belirlenmiş sınıflar yer almamaktadır (Han ve Kamber, 2006). Örneğin sınıflamada kategoriler oluşturulmaktadır ve sınıflayıcı sonuç olarak bu kategorilerden birisini seçmek zorundadır. Yaptığı seçimin doğruluğuna göre doğru sınıflama oranı değişmektedir.

Kestirim ise belirli herhangi bir sınıf bulmaz, bunun yerine gelecekte elde edilebilecek bir sonucu tahmin eder. Bu özelliğiyle regresyon modelleriyle benzerlik göstermektedir. Kestirim algoritmaları tahminde bulunabilmek için regresyon yöntemlerini kullanabilmektedirler. Dikkat edilmesi gereken husus zaman serileri gibi farklı yöntemler de kullanılabildiği için regresyon yöntemleriyle karıştırılmaması gerektiğidir (Tüzüntürk, 2010).

3.5.1.4. Zaman serileri analizi

Verilerin belirli zaman aralıklarına göre elde edildikleri ve birbirlerini takip etme sıralarının da önemli olduğu, veri setleri üzerine çalışılan veri madenciliği alanıdır. Borsada, sismik hareketlerin incelenmesinde, hava durumunun gözlenmesinde ve bu gibi birçok problemde yaygın bir şekilde kullanılmaktadır. Veriler günlük, haftalık, yıllık ve benzeri şekillerde toplanabilmektedir. Brown ve Oxford (2001), Virginia'nın Richmond bölgesindeki haneye tecavüz ve hırsızlık suçlarına yönelik zaman serileri analizi yapmışlardır. Suçların belirli zaman aralıklarında yapılıp yapılmadığı, içerisinde regresyon analizi de olan farklı yöntemlerle incelenmiştir. İlgili veri setinde en başarılı sonuçlar, haftalık zaman dilimleri şeklinde yaklaşıldığında alınmıştır. Elde edilen bilgilerin yerel polis kaynaklarının verimli kullanılmasında yardımcı olması hedeflenmiştir.

3.5.2. Tanımlayıcı

Elde bulunan veriler hakkında bilgi çıkarılmaya ve bu verilerin özellikleri farklı şekillerde ifade edilmeye çalışılıyorsa, tanımlayıcı başlığı altında toplanmıştır.

3.5.2.1. Birliktelik kuralları

Birliktelik kuralları, veri kümelerinin üyeleri arasındaki ilişkileri ortaya çıkaran veri madenciliği yöntemlerinden birisidir. Pazarlama alanında, (Pazar sepeti analizi adıyla) sıkça kullanılmaktadır (Oğuzlar, 2004).

Algoritması temel olarak iki kısımdan oluşmaktadır. Birinci kısımda veri kümesinin öğelerinden alt kümeler oluşturulur. İkinci kısımda ise belirli bir ölçüte göre oluşturulan alt kümeler arasından kural çıkarmak için seçim yapılır. Elde edilen kurallar şekil olarak matematikteki mantık alanındaki önermelerle benzerlik göstermektedir. “ $A \Rightarrow B$ ”, “A” ifadesi doğru olduğu takdirde “B” ifadesi de doğrudur anlamına gelen

gösterim şekli, veri madenciliği için “A” ögesinin bulunduğu altkümelerde “B” ögesi de bulunmaktadır şeklinde yorumlanmaktadır.

Alt kümelerin elde edilmesinde kullanılan ölçütler destek (support) ve güven (confidence) ölçütleridir. Bu ölçütlerin hesaplanmasıyla kuralların geçerliliği de hesaplanmış olur. Kumar ve Chadha (2012), mezun olmuş öğrencilerin lisans öğrenimlerinden sonraki başarıları ile lisans öğrenimlerindeki başarıları arasında birliktelik kuralları yardımıyla yararlı olabilecek ilişkiler bulmuşlardır.

3.5.2.2. Özetleme

Veri kümelerinin daha iyi anlaşılabilmesi için özetleme yöntemlerinin kullanılması çoğu durumda veriler üzerinde çalışılmasını kolaylaştırmaktadır (Han ve Kamber, 2006). Özetlemede kullanılan birçok yöntem istatistik alanından gelmektedir ve bu yöntemler, veri kümelerini ifade edebilecek alt kümeler ya da tanımlayıcılar oluşturmada da kullanılabilirler. Veri kümelerinin varyansları, ortalama değerleri ve benzeri bulgular örnek olarak düşünülürse, bu bulgular verilerin dağılımıyla ilgili yararlı bilgiler içerebilmektedir. Bu bilgiler ışığında uygulanabilecek yöntemlere karar verilebilmektedir.

3.5.2.3. Kümeleme

Veri madenciliğinin tanımlayıcı başlığı altında yer alan kümeleme, sınıfları belli olmayan veri kümesinde, birbirine belirli bir ölçüde benzer olan öğeleri gruplayan yöntemlerden oluşmaktadır (Han ve Kamber, 2006). Örneğin araba ve motosikletle ilgili verilerin bulunduğu bir veri kümesi ele alınacak olursa. Girdilerin motosiklet için mi yoksa araba için mi olduğu belirtilmezse bu verilerin gruplanması kümeleme problemi olacaktır. Bu türdeki problemlerde veri setindeki öznitelikler oluşacak küme sayısını doğrudan etkileyecektir. Tek bir öznitelik üzerinden kümelenmeye çalışılırsa ve bu öznitelik aracın sahip olduğu tekerlek sayısı ise iki aracın arasındaki farklılık net bir şekilde ortaya konulacaktır. Fakat araçların kullandıkları yakıt türünün öznitelik olarak

kullanıldığı düşünülürse, bu durumda oluşabilecek öbekler tahmin edilememektedir. Yakıt türü öznel olarak seçildiği için sonuç olarak araçlar motor türlerine göre gruplandırılabilir.

3.5.2.4. Sıralı dizileri bulma

Birliktelik kurallarına benzerlik göstermektedir farkları ise birliktelik kurallarında öğelerin sıraları gözetilmezken, sıralı dizileri bulma algoritmalarında öğelerin birbirlerini takip etme sıralarının gözetilmesidir. Müşterilerin alışveriş düzenlerinin bulunması, zamana bağlı bilimsel verilerin işlenmesi, DNA dizilimlerinin incelenmesi gibi alanlarda kullanılmaktadır ve geniş bir kullanım yelpazesi vardır.

Pei vd. (2004) tarafından yapılan çalışmada, sıralı dizilerin farklı yöntemleri irdelenmiştir. Yöntemlerle ilgili yaptıkları deneysel çalışmanın sonucunda, geliştirilmiş *PrefixSpan* algoritmasının çoğu denemede daha iyi sonuçlar verdiği gözlenmiştir.

4. VERİ MADENCİLİĞİ UYGULAMASI

Bu bölümde, öğrencilere ait veriler üzerinde uygulamalar gerçekleştirilmektedir. Uygulamalarda kullanılacak yöntemler, programlar ve veri setlerine ait açıklamalar da bu bölümde yer almaktadır

4.1. Kullanılan Programlar

Veri setlerinin derlenmesi ve düzenlenmesinde Microsoft Excel programı kullanılmıştır. Algoritmaların ve ön işlemlerin bir kısmının gerçekleştirilmesinde ise Weka programı kullanılmıştır.

4.1.1. Weka

Weka, Yeni Zelanda'da bulunan Waikato Üniversitesi'nde geliştirilmiş Java tabanlı bir uygulamadır. Literatürde olan çoğu veri madenciliği algoritmasını içinde barındıran Weka, aynı zamanda ön işleme sürecinde kullanılabilecek yararlı araçlara da sahiptir. Genel Kamu Lisansı'na sahip Weka, ücretsiz bir uygulama olduğu gibi, Java'da yazıldığı için de çoğu sistemde sorunsuz bir şekilde çalışır (Witten vd., 2011).

Weka uygulaması kendine özgü *.arff* dosya uzantısına sahiptir. Microsoft Excel programında *.csv* uzantılı olarak kaydedilmiş direk çalıştıramamaktadır. Yapılması gereken ilk iş *.csv* uzantılı dosyanın metin editörü yardımıyla açılıp tüm noktalı virgül işaretlerinin, virgül işaretiyle değiştirilip kaydedilmesidir. Sonrasında Weka ortamında veri seti ayarlanıp *arff* dosyasına çevrilebilmektedir.

Weka ortamı için geliştirilen *Arff* dosyalarının *csv* dosyalarından farkı, özniteliklerin türlerinin belirtilmesidir. Özniteliklerin türlerinden kastedilmek istenen, özniteliğin alacağı değerlerin kategori şeklinde ya da sürekli sayısal değerler şeklinde olmasıdır. Şekil 4.1' de *csv* dosya örneği verilmiştir. Görüldüğü üzere ilk satırdan itibaren

öznitelik adları aralarına virgül konularak yazılmaktadır. İkinci satıra geçildiğinde o özniteliklerin sırasıyla sahip olduğu değerler yazılmaktadır. Şekil 4.2’de ise *arff* dosya örneği verilmiştir. csv dosyasından farkı değerlerin sınırlarının belli olması ve verilerin alacağı değerlerin “@data” başlığıyla belirtilmesidir.

```
gokyuzu, sicaklik, ruzgar, oyun?
gunesli, 30, yok, hayır
gunesli, 27, var, hayır
kapali, 28, yok, evet
yagmurlu, 20, yok, evet
yagmurlu, 19, yok, evet
yagmurlu, 15, var, hayır
kapali, 17, var, evet
gunesli, 25, yok, evet
```

Şekil 4.1: “.csv” dosya uzantılı veri seti örneği

```
@relation hava
@attribute gokyuzu { gunesli, kapali, yagmurlu }
@attribute sicaklik numeric
@attribute ruzgar { var, yok }
@attribute oyun? { evet, hayır }
@data
gunesli, 30, yok, hayır
gunesli, 27, var, hayır
kapali, 28, yok, evet
yagmurlu, 20, yok, evet
yagmurlu, 19, yok, evet
yagmurlu, 15, var, hayır
kapali, 17, var, evet
gunesli, 25, yok, evet
```

Şekil 4.2: “.arff” dosya uzantılı veri seti örneği

4.1.2. Microsoft Excel

Microsoft Excel, Microsoft Windows, Mac OSX ve IOS gibi işletim sistemleri üzerinde çalışabilen bir elektronik tablo uygulamasıdır. Veri işlemeyi kolaylaştıran, alanındaki en iyi programlardan birisidir. Günlük ofis işlerinde kullanıldığı gibi VBA(Visual Basic for Applications) ortamında makro geliştirebilme özelliği sayesinde daha gelişmiş çalışmalara da imkân sağlamaktadır.

Excel’de makro oluşturmak, normalde kullanıcının elle yaptığı işlemleri otomatik olarak bilgisayara yaptırabilmesini sağlamaktadır. Böylece zaman alıcı işlemlerin çoğu, çok kısa sürelerde tamamlanabilmektedir. Excel’de makrolar iki şekilde oluşturulabilmektedir. Birinci yöntem, makro kaydetme özelliği ile yapılması istenilen işin tamamı bir seferlik uygulanarak kaydedilir. Makro kaydedicisi kullanıcının yaptığı işlemler VBA diline dönüştürmektedir sonrasında oluşan makro çalıştırılarak kullanılabilir. İkinci yöntemde, VBA programlama dilinde yapılması gereken iş kodlanarak makro oluşturulmaktadır. Basic programlama dilinin Excel nesnelerini de içeren bir türevi olan VBA aracılığıyla makro oluşturma kolaylaştırılmıştır.

Makroların, Excel dosyası halinde bulunan veriler üzerinde, çalışıldığında ön işleme sürecini hızlandırmaya yönelik olumlu etkileri bulunmaktadır. Özellikle birden fazla veri setine, benzer işlemler uygulanması durumunda VBA’da makrolar oluşturmak, hali hazırda uzun olan ön işleme sürecini kısaltmaktadır.

4.2. Kullanılan Algoritmalar

Bu tez kapsamında gözetilen ana unsur kullanılan algoritmaların sayısı ve çeşitliliği olmadığı için veri madenciliği alanında tasarlanmış olan algoritmalarından aşağıdaki üçü kullanılmıştır.

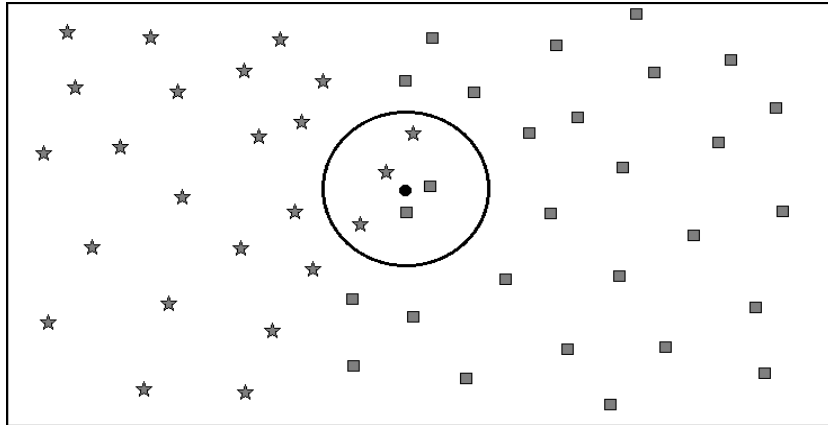
- KNN
- J48
- Apriori ve Predictive Apriori

Veri madenciliğinin tahmin edici alanından kNN ve J48 algoritmaları, tanımlayıcı alanından ise Apriori algoritması seçilmiştir. Bu algoritmaların seçilmesindeki sebep veri setlerine yönelik yapılan çalışmaya, sınıflama problemi gibi yaklaşmasının ve özniteliklerin arasındaki ilişkinin kurallar şeklinde ortaya çıkarılmasının istenmesidir.

4.2.1. Knn

K-Yakın Komşuluk (k-nearest neighbor, kNN) algoritması 1950’li yıllarda bulunmuş bir sınıflandırma algoritmasıdır. Bilgisayar teknolojileri gelişip, bilgisayarların işlem yapma gücü artınca kullanımı yaygınlaşmıştır (Han ve Kamber, 2006).

Karmaşık olmayan yapısına rağmen, çoğu zaman kNN etkili sonuçlar veren bir sınıflama algoritmasıdır. Test kümesindeki girdilerin hangi sınıfa ait olduğunu, eğitim setindeki girdilerle karşılaştırarak bulmaktadır ve bu işleyişi ile basit bir öğrenim algoritması da olmaktadır. Girdinin ait olduğu sınıfı bulmak için, eğitim setindeki her bir girdiyle benzerlik ya da benzeşmezlik değeri hesaplamaktadır. Girdiler koordinat sistemindeki birer noktalmış gibi ele alınır ve aralarındaki mesafe hesaplanır. Test girdisiyle, en düşük uzaklığa sahip olan eğitim girdilerinden, başlangıçta kullanıcı tarafından belirlenecek k sayısı kadar girdi seçilip bu k tane girdide çoğunluk hangi sınıfa aitse test girdisi o sınıfa atanacaktır. Şekil 4.3’te k sayısı 5 olarak belirlenmiştir ve yıldız sembolüyle gösterilmiş sınıftan 3 tane olduğu için test girdisi yıldızla temsil edilen sınıfa dahil olmuştur.

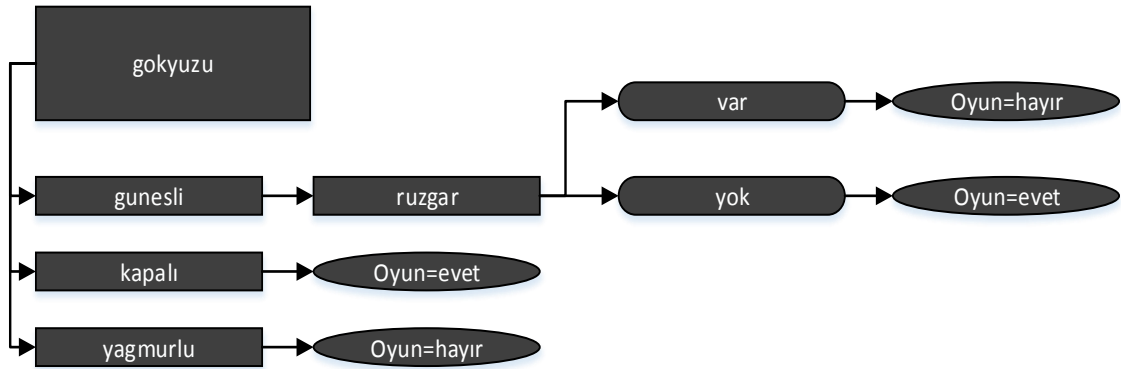


Şekil 4.3: En yakın komşuluk algoritmasının örnek çalışması

Uzaklık hesabında kullanılan yöntem, algoritmanın işleyişini doğrudan etkilemektedir. Bu yüzden probleme yönelik bir uzaklık ölçüm türü seçilmelidir. Buna ek olarak, eğer sayısal değerlerle çalışılıyorsa değerlerin tümü $[0, 1]$ aralığına çekilirse, özniteliklerden bazılarının baskın gelip sonucu kötü yönde etkilemesinin önüne geçilebilir.

4.2.2. J48

Karar ağaçları algoritmalarından C4.5'in Weka tarafından uyarlanmış halidir (Quinlan, 1993; Khan vd.'den, 2008) . Ağaç yapılarına benzemektedir, yaprak kısımları girdilerin sınıflarını temsil etmektedir. Hava durumuna göre dışarıda oyun oynanıp oynanamayacağı verilerinin bulunduğu veri seti ele alınacak olursa (Bkz. Şekil 4.1), karar ağaçları sınıflama yapılmasında kullanılabilir. Şekil 4.4'teki gibi bir ağaç yapısı oluşabilir.



Şekil 4.4: Hava durumu veri setinden elde edilebilecek ağaç yapısı

Ağaç yapısında gövdeden itibaren yapraklara doğru gidilerek kolay bir şekilde kurallar çıkarılabilmektedir. Örneğin gökyüzü güneşli ve rüzgar yoksa dışarı çıkılıp oyun oynanabilir kuralı elde edilmektedir.

4.2.3. Apriori ve predictive apriori

Apriori algoritması, veri kümelerinde sıkça geçen öğeleri bulmaktadır ve bu öğelerle birliktelik kuralları elde etmektedir (Agrawal ve Srikant, 1994). N tane elemanlı bir küme için önce bir elemanlı altkümeler oluşturulmaktadır. Algoritmanın başlangıcında belirlenen en küçük destek ölçütünden daha az sıklıkla geçen altkümeler elenmektedir. Sonraki aşamada elenmeyen öğelerden iki öğeli altkümeler oluşturulup yukardaki işlem tekrar edilerek destek ölçütünü sağlamayan altkümeler elenmektedirler. Böylece tüm altkümelerin taranması yerine destek ölçütünü sağlayan öğelerden türeyen altkümeler taranmış olmaktadır.

Kurallar elde edilmesi kısmında, kullanıcı tarafından belirlenmiş güven ölçütü temel alınmaktadır. Belirlenen güven ölçütünü sağlayamayan kurallar oluşturulmamaktadırlar. Liu (2010) tarafından yapılan çalışmada Apriori algoritmasının örnek bir veri kümesine uygulanması gösterilmiştir. Destek ve güven ölçütlerinin nasıl hesaplandığı ve algoritmanın işleyişi anlatılmaktadır.

Predictive (tahmin edici) Apriori algoritması Scheffer (2005) tarafından yapılan çalışmada Apriori algoritması üzerinde geliştirilmiştir. En uygun olacak güven ve destek ölçütleri tahmin edilerek yüksek doğrulukta kurallar elde edilmesi amacı güdülmektedir. Gereksiz olan kurallar üretilmediği için algoritmanın çalışma hızı da arttırılmış olmaktadır.

4.3. Veri Seti ve Ön İşleme

Eskişehir Osmangazi Üniversitesi Bilgisayar Mühendisliği Bölümü'nde eğitim gören ve 2008, 2009, 2010, 2011 yıllarında birinci sınıf derslerine başlayan öğrencilerin verileri kullanılmıştır. Bu yılların seçilmesindeki ana etken, 2011 yılında gerçekleşen üniversitelerden atılmanın kalkmasının gözlemlenebilmesi için 2011 öncesi yılların da verilerine ihtiyaç duyulmasıdır. 2008 yılından itibaren, öğrencilerle ilgili daha geniş kapsamlı bilgiler (öğrencilerin üniversiteye yeni kayıt olduklarında istenilen formlardan edinilen bilgiler) toplandığı için başlangıç yılı 2008 seçilmiştir. 2011 yılından sonra birinci

sınıfa başlayan öğrencilerin çoğu eğitimlerine devam ettikleri için, 2011 yılından sonra birinci sınıfa başlayan öğrencilerin verileri kullanılmamıştır.

Eldeki veriler, gerek demografik gerekse öğrencilerin üniversite eğitimlerinde aldıkları tüm derslerle ilgili bilgiler bakımından zengin olduğundan, öğrencilerin üniversitedeki başarılarına etki edebilecek etmenler de belirlenmeye çalışılmıştır.

Veri setleri oluşturulurken öğrenci bilgi sisteminden gelen veriler Excel dosyalarına aktarılmıştır. Öğrencilerin sağlık durumlarıyla alakalı bilgiler Çizelge 4.1’de açıklanmıştır.

Çizelge 4.1: Öğrencilerin sağlık bilgilerinin bulunduğu veri setinin içeriği

<i>Öznitelik</i>	<i>Açıklama</i>	<i>Değerler</i>
Kayıt_No	Öğrenci numaraları yerine kullanılmış rastgele sayılar	[2 , 255]
Sigara	Öğrencinin sigara kullanıp kullanmadığı bilgisi	{Evet,Hayır}
Alkol	Öğrencinin alkol kullanıp kullanmadığı bilgisi	{Evet,Hayır}
Saglik_Durum	Öğrencinin derslerini etkileyebilecek rahatsızlığı olup olmadığı bilgisi	{Evet,Hayır}
Sorun_Tur	Öğrencinin sahip olduğu rahatsızlığın adı	-
Tedavi_Tur	Öğrencinin rahatsızlığında yönelik dikkat ettiği hususlar,tedaviler	-
Tedavi	Öğrencinin sürekli tedavi görmesini gerektiren hastalığı olup olmadığı ve adı	-
Acil_Tel_No	Acil durumlarda ulaşılacak öğrenci yakınının telefon numarası	-

Üniversiteye yeni kayıt yaptıklarında doldurdıkları demografik bilgileri içeren form ile ilgili açıklamalar Çizelge 4.2’de bulunmaktadır.

Çizelge 4.2: Öğrencilerin demografik bilgilerinin bulunduğu veri setinin içeriği

Öznitelik	Açıklama	Değerler
Kayıt_No	Öğrenci numaraları yerine kullanılmış rastgele sayılar	[2 , 255]
Tel_Esk	Öğrencinin Eskişehir'deki telefon numarası	-
Tel_Aile	Öğrencinin ailesinin telefon numarası	-
Cep_Ogr	Öğrencinin cep telefonu numarası	-
GSM	Öğrenciye ait cep telefonu servis sağlayıcısı	-
Burs	Öğrencinin burs alıp almadığı bilgisi	{Evet,Hayır}
Burs_Kurum	Öğrencinin burs aldığı kurumun adı	-
Gelir_Aile	Öğrencinin ailesinin aylık ortalama gelir bilgisi	-
Is_Durum	Öğrencinin bir işte ücretli olarak çalışmış ya da çalışıyor olması bilgisi	{Evet,Hayır}
Is_Turu	Öğrencinin çalışmış ya da çalışıyor olduğu işin türü ve çalışma süresi	-
Harclik	Öğrencinin ailesinden aldığı harçlık miktarı	-TL
Para_durum	Öğrencinin eline geçen parayı yeterli bulup bulmadığı bilgisi	{Yeterli,Yeterli değil, İdare eder, Temel ihtiyaç}
Sigara	Öğrencinin sigara kullanıp kullanmadığı bilgisi	{Evet,Hayır}
Alkol	Öğrencinin alkol kullanıp kullanmadığı bilgisi	{Evet,Hayır}
Anne_Egt	Öğrencinin annesinin tahsil durumu	{Tahsil gormemis, Ilkogretim, Ortaogretim,Universite,MasterDoktora}
Anne_Mes	Öğrencinin annesinin meslek bilgisi	{Ev hanimi,Serbet,Emekli, Devlet memuru, Isci,Diger}
Baba_Durum	Öğrencinin babasının hayatta olup olmadığı bilgisi	{Hayatta,Vefat etti}
Baba_Egt	Öğrencinin babasının tahsil durumu	{Tahsil gormemis, Ilkogretim, Ortaogretim,Universite,MasterDoktora}
Baba_Mes	Öğrencinin babasının meslek bilgisi	{Yok,Serbet,Emekli, Devlet memuru,Isci,Diger}
Sos_Guv	Öğrencinin bağlı bulunduğu sosyal güvenlik bilgisi	{Bagkur,Emekli sandigi,SSK,yok}
Anne_Baba	Öğrencinin anne ve babasının evli olup olmadığı bilgisi	{Evet,Hayır}
Kardes	Öğrencinin kendisi de dahil olmak üzere kardeş sayısı	-
Oku_Kardes	Öğrencinin kendisi de dahil olmak üzere okuyan kardeş sayısı	-
Kalacak_Yer	Öğrencinin kalacağı yer bilgisi	-

Öğrencilerin üniversite eğitimlerinin tamamında aldıkları dersler ve derslerle ilgili bilgilerin açıklaması Çizelge 4.3'te bulunmaktadır

Çizelge 4.3: Öğrencilerin aldıkları derslerle ilgili olan veri setinin içeriği

Öznitelik	Açıklama	Değerler
Kayit_No	Öğrenci numaraları yerine kullanılmış rastgele sayılar	[1 , 258]
Ders_Kod	Öğrencini almış olduğu derslerin kod olarak karşılığı	{101011001,.....,401111009}
Ders_Ad	Öğrencinin aldığı derslerin adları	-
Saat_Teo	Öğrencinin almış olduğu dersin teorik saati	[0 , 26]
Saat_Uyg	Öğrencinin almış olduğu dersin uygulama saati	[0 , 4]
Ders_Tur	Öğrencinin aldığı dersin türü	{0 , 3}
Kredi	Öğrencinin aldığı dersin kredisi	[0 , 5.5]
AKTS	Öğrencinin aldığı dersin AKTS kredisinin bilgisi	[0 , 60]
Not	Dersten alınmış olan başarı notu değeri	[0 , 100]
Harf_Kod	Öğrencinin derslerden almış olduğu harf kodunun bilgisi	{AA,BA,BB,CB,CC,DC,DD,FF,YT,YZ,MU}
Not_Donem	Dersin alınması gereken dönem bilgisi	{Bahar,Yaz,Guz}
Ders_Yil	Öğrencinin dersi aldığı yıl	[2002 , 2014]
Ders_Donem	Dersin öğrenci tarafından alındığı dönem	{Bahar,Yaz,Guz}
Cift_Yan	Dersin çift anadal, yandal ya da normal olarak alındığı bilgisi	{Cift,Yan,Normal}
Ders_Yer	Dersin kendi veya başka bölümlerden ya da başka üniversitelerden alındığı bilgisi	{Kendi_bolum,Diger_bolum,Diger_Uni}
Dis_Uni	Ders başka üniversiteden alınmış ise o üniversiteye ESOGÜ tarafından atanmış kod	-

Öğrencilerin genel not ortalaması, üniversiteye başlama ve bitirme tarihleri gibi genel bilgilerin bulunduğu Excel dosyası da Çizelge 4.4'te görüldüğü şekildedir.

Çizelge 4.4: Öğrencilerin bilgi sistemindeki bilgilerinin bulunduğu veri setinin içeriği

Öznitelik	Açıklama	Değerler
Kayit_No	Öğrenci numaraları yerine kullanılmış rastgele sayılar	[1 , 258]
Dogum_Yil	Öğrencini doğum yılı	[1967 , 1994]
Ders_Bas_Yil	Öğrencinin 1. sınıf derslerine başlama yılı	{2008,2009,2010,2011}
Haz_Durum	Öğrencinin hazırlık sınıfını ne şekilde bitirdiği bilgisi	{Basarisiz,Gecti,Muaf,Tanimsiz}
Haz_Not	Öğrencinin hazırlık sınıfındaki not ortalaması	[0 , 93]
Gelis_Tur	Öğrencinin üniversiteye başlama şekli	{Yeni_kayit, Lisans_tamamlama, Yatay_gecis, Dikey_gecis, Yabanci_uni}
Gelis_Yil	Öğrencinin üniversiteye başlama yılı	[2005 , 2012]
Gidis_Tur	Öğrencinin üniversiteden ayrılma şekli	{Yok,Kayit_yaptirmadi, Kendi_istegi_ile, Mezun, Tanimsiz, Yatay_gecis}
Gidis_Yil	Öğrencinin üniversiteden ayrılma yılı	[2009 , 2014]
GNO	Öğrencinin üniversitede dörtlük sistemdeki genel not ortalaması	[0 , 3.77]
Nufus_Il	Öğrencinin nüfusuna bağlı olduğu il	-
Dogum_Il	Öğrencinin doğduğu ilin adı	-
Adres_Il	Öğrencinin ikamet ettiği ilin adı	-
Ulke	Öğrencinin vatandaşı olduğu ülke	-
Cinsiyet	Öğrencinin cinsiyet bilgisi	{Erkek,Kiz}
Burs	Öğrencinin bursunun olup olmadığı bilgisi	{Var,Yok}
Kredi	Öğrencinin kredi alıp almadığı bilgisi	{Var,Yok}
Harc_Kredi	Öğrencinin üniversite harcı için kredi alıp almadığı bilgisi	{Var,Yok}
Burs_Kredi	Öğrencinin burs veya kredisinin olup olmadığı bilgisi	{Var,Yok}
Burs_Kredi_Tur	Öğrencinin aldığı burs veya kredinin türü	{ESOGU,KYK_Burs,KYK_3_misli}
Dis_Prog	Öğrenci yurtdışı değişim programlarından yararlandıysa hangi program olduğu bilgisi	{Yok,Erasmus,Farabi}
Mezun_Durum	Öğrencinin mezuniyet durumu	{Ilisigi kesildi,Mezun,Normal,Uyari}

Bu dört Excel dosyasının içindeki bilgiler de 2008, 2009, 2010 ve 2011 yılları için bölünmüş şekildedir.

Öncelikle veriler tekrardan düzenlenip yeni Excel dosyaları şeklinde ayarlanmıştır. Bu düzenleme esnasında ön işleme sürecinin bir kısmı da uygulanarak, çalışmaya temel oluşturacak veri setleri hazırlanmıştır. Çizelge 4.5'te oluşturulan Excel dosyalarının adları boyutları ve içerikleri verilmektedir.

Çizelge 4.5: Temel olarak oluşturulan Excel dosyaları ve kısa açıklamaları

<i>Dosya Adı</i>	<i>Girdi ve Öznitelik</i>	<i>Açıklamalar</i>
veriseti_2008_anket_v1_0.xls	44 x 24	Öğrencilerin üniversiteye kayıt yaptırdıkları sırada doldurdıkları form verilerinden oluşmaktadır. Demografik bilgileri içermektedir ve eğitim süreleri zarfında öğrenci bilgi sisteminde eksik bilgilerini tamamlama imkânları bulunmaktadır.
veriseti_2009_anket_v1_0.xls	39 x 24	
veriseti_2010_anket_v1_0.xls	43 x 24	
veriseti_2011_anket_v1_0.xls	14 x 24	
veriseti_2008-2011_anket_v1_0.xls	137 x 24	
veriseti_2008_obs_v1_0.xls	52 x 22	Öğrencilerin üniversiteye başlama ve bitirme tarihleri, doğum yılları, hazırlık sınıfı notu, burs kredi alıp almadıkları, 1. sınıfa başlama yılları, memleket bilgileri, ÖSS puanları, genel not ortalamaları ve mezuniyet durumlarıyla ilgili veriler yer almaktadır.
veriseti_2009_obs_v1_0.xls	52 x 22	
veriseti_2010_obs_v1_0.xls	60 x 22	
veriseti_2011_obs_v1_0.xls	89 x 22	
veriseti_2008-2011_obs_v1_0.xls	240 x 22	
veriseti_2008_ders_v1_0.xls	2617 x 16	Öğrencilerin eğitim süreleri boyunca aldıkları tüm derslerle ilgili bilgileri içerir. Derslerin teorik, uygulama saatleri, kredileri, öğrencilerin o derslerden aldıkları başarı puanları, harf kodları, hangi dönemde aldıkları gibi bilgiler bulunmaktadır.
veriseti_2009_ders_v1_0.xls	2607 x 16	
veriseti_2010_ders_v1_0.xls	3401 x 16	
veriseti_2011_ders_v1_0.xls	4278 x 16	
veriseti_2008-2011_ders_v1_0.xls	12900 x 16	
veriseti_2008_saglik_v1_0.xls	42 x 8	Öğrencilerin sağlık durumlarının bulunduğu veri setidir. Eğitim açısından engel teşkil edebilecek rahatsızlıkları var mı? Varsa rahatsızlığın türünün ne olduğu bilgisini içerir.
veriseti_2009_saglik_v1_0.xls	39 x 8	
veriseti_2010_saglik_v1_0.xls	42 x 8	
veriseti_2011_saglik_v1_0.xls	13 x 8	
veriseti_2008-2011_saglik_v1_0.xls	133 x 8	

Öğrenci bilgi sisteminden çekilen veriler ham halinde kullanılamayacak bir durumda bulunmaktadır. Veriler üzerinde yapılan ön işlemlerin ilk aşaması aşağıda özetlenmiştir.

Öğrencilerin doğum yılları, üniversiteye girdikleri ve bitirdikleri tarihlerdeki yaş değerlerine çevrilmiştir. Öğrencilerin üniversiteye başladıkları ve üniversiteyi bitirdikleri tarihler arasındaki fark alınıp eğitim yılları oluşturulmuştur. Buna ek olarak hazırlık sınıfında geçirdikleri süre bu toplam eğitim yılından çıkarılarak bir öznitelik daha elde edilmiştir. Özniteliklerin bazıları çift haldedir. Özniteliklerden biri durumun açıklamasıyken diğeri sistemde bulunan kod olarak o durumun karşılığıdır. Bire bir örtüşükleri ve kodları anlaması güç olduğu için kod bilgileri kaldırılmıştır. Üç farklı adres türünden yalnız ikamet ettikleri adres bilgisi kullanılmış olup, adres bilgileri il adı olarak bulunduğu iller bulundukları coğrafi bölgenin adıyla değiştirilmiştir. Mezun oldukları lisedeki not ortalamaları, çok fazla eksik bilgiye sahip olduğu için kaldırılmıştır. Öğrencilerin Öss puanları birbirlerine yakın olması sebebiyle bu bilgi de veri setlerinden çıkarılmıştır. Dört farklı türde bulunan burs ve kredi bilgileri *BursKredi* özniteliği altında birleştirilmiştir; burs veya kredi alınıyorsa 1, alınmıyorsa 0 değer olarak atanmıştır.

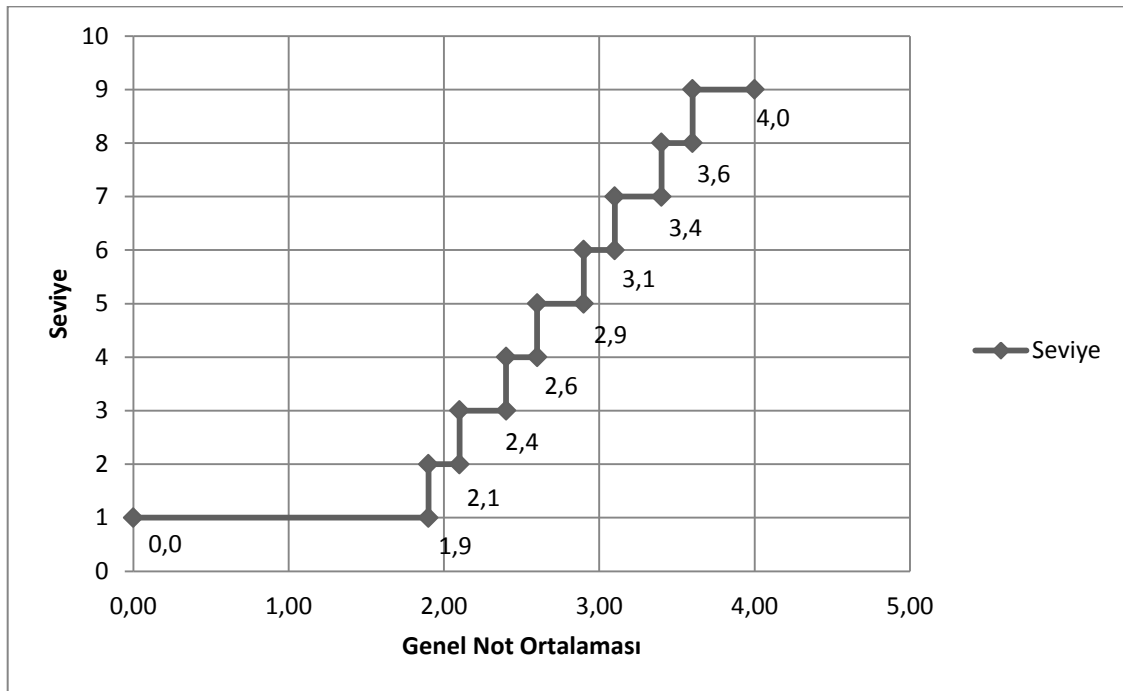
Öğrencilerin üniversite eğitimleri süresince aldıkları notların düzenlenmesinde yapılan işlemler yazının devamında verilmiştir. Eskişehir Osmangazi Üniversitesi'nde de, çoğu üniversitede olduğu gibi dörtlük notlandırma sistemi kullanılmaktadır. Çizelge 4.6'da dörtlük not ve harf kodu karşılıkları verilmektedir. Bu dönüşüm tablosuna göre öğrencilerin derslerden aldıkları harf notları dörtlük sistemdeki karşılık gelen notlara çevrilmektedir.

Öğrencilerin veri setindeki genel not ortalamaları 2.39, 3.21 gibi birbirinden farklı birçok değere sahiptir. Hazırlık sınıfı geçme sınavından aldıkları notlarsa 60, 72 gibi yüzlük not sistemindeki farklı değerlerden oluşmaktadır. Çok fazla farklı değerle çalışılması yerine, öğrencilerin notları özel olarak belirlenmiş aralıklara göre kategorileştirilmiştir.

Çizelge 4.6: Dörtlük not sistemi ile harf kodu karşılıkları

<i>Harf Kodu</i>	<i>Dörtlük Not</i>
AA	4,0
BA	3,5
BB	3,0
CB	2,5
CC	2,0
DC	1,5
DD	1,0
FF	0,0

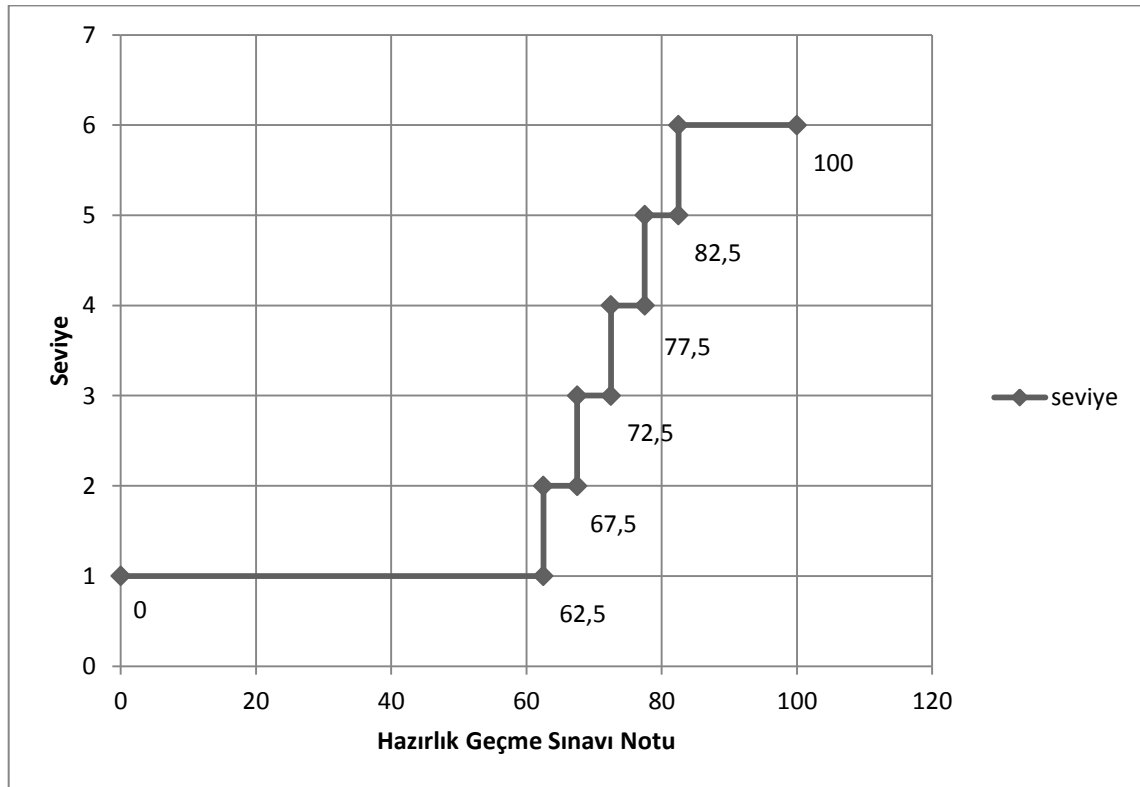
Genel not ortalamaları, birden dokuzca kadar olan tam sayılarla seviyelendirilmiş ve aralıklar için Şekil 4.5'teki dönüşüm grafiği kullanılmıştır. Şekilde de görüldüğü üzere dörtlük sistemin bölündüğü aralıklar eşit değildir. Önemli görülen not değerlerinin daha iyi gözlemlenebilmesi için böyle bir yaklaşımda bulunulmuştur.



Şekil 4.5: Genel not ortalamalarının dokuz kategoriye çevirim grafiği

Örneğin 2,0 notu CC harf notuna denk gelmektedir ve öğrenciler üniversiteden mezun olabilmeleri için genel not ortalamaları 2,0 ve üstü olması gerekmektedir. Genel not ortalaması 2,0 üstünde ve altında olan öğrenciler risk grubunda kabul edilmiştir ve $\pm 0,1$ değeriyle $[1,9 ; 2,1]$ aralığı oluşturulmuştur ve bu aralığı 2 tam sayısı ifade etmektedir. Not ortalaması 2,0 altında olan öğrenciler mezun olma şartına sahip olmadıklarından $[0 ; 1,9]$ aralığına 1 tam sayısı atanmıştır.

Hazırlık notları da genel not ortalamasına benzer bir biçimde dönüştürülmüştür ve Şekil 4.6'da gösterilmiştir. Hazırlık sınıfının başarı ile geçilebilmesi için final sınavı yapılmaktadır ve 70'in altında kalan öğrencilerin hazırlık sınıfını tekrar etmeleri gerekmektedir. Bu sebeple 70 değerinin $\pm 2,5$ aralığına 3 tam sayısı atanmıştır ve geri kalan aralıklar şekildeki gibi düzenlenmiştir.



Şekil 4.6: Hazırlık sınıfını geçme notlarının altı kategoriye çevirim grafiği

4.4. Veri Madenciliği ve Çıktıları

Veri madenciliği aşamasında iki yöntemsel yaklaşımda bulunulmuştur. Bunlardan birincisi sınıflama algoritmaları, ikincisi ise birliktelik kuralları çıkarımında kullanılan algoritmalar. Sınıflama algoritmalarından J48 ve kNN algoritmaları seçilmiştir. J48 algoritması oluşturabildiği için, kNN algoritması ise basit ve etkili yapıya sahip olduğu için seçilmiştir. Birliktelik kurallarından sıkça kullanılan Apriori ve onun bir türevi olan Predictive Apriori algoritmaları seçilmiştir.

Sınıflama algoritmalarının Weka ortamında seçilen parametreleri denemeler sonucunda belirlenmiş ve algoritma çıktılarının verildiği çizelgelerde en iyi sonuç verdikleri parametre değerleri gösterilmiştir. Kullanılan parametrelerin gösterimi ve açıklamaları aşağıda verilmektedir.

- $k = \{\text{pozitif tam sayı}\}$:kNN algoritmasında belirlenen komşuluk sayısı
- $1/d$:kNN algoritmasında uzaklığa ters orantılı ağırlıklandırma
- $c = \{\text{pozitif ondalıklı sayı}\}$:J48 algoritmasında güven faktörü, bu değer küçüldükçe oluşturulan ağaç daha çok budanmış olur.

Sınıflama algoritmalarının sonuçlarını içeren çizelgelerde, *TP* oranı (true positive), *FP* oranı (false positive), *kesinlik* (precision), *duyarlılık* (recall), *F-Ölçütü* ve *doğruluk* (accuracy) değerleri verilmektedir. J48 algoritmasında çizelgelere ek olarak ağaç yapıları da şekilsel olarak sunulmaktadır. Apriori algoritmalarından elde edilen kurallar ise maddeler halinde verilmektedir.

4.4.1. Öğrenci işleri bilgi sistemi

Öğrenci işleri bilgi sistemindeki verilerin bulunduğu "veriseti_2008_obs" gibi isminde "obs" içeren Excel dosyalarındaki (Bkz. Çizelge 4.5) verilerden anlamlı bilgiler elde edilmek istenmiştir. Özniteliklerin birbirleriyle olan ilişkileri ve öğrencilerin

üniversite başarılarına etki eden faktörler bulunmaya çalışılmıştır. Başarı ölçütü olarak öğrencilerin genel not ortalamaları ve toplam eğitim süreleri ele alınmaktadır.

Veri setlerinde 2008, 2009, 2010, 2011 yıllarında birinci sınıf derslerine başlayan öğrencilerin verileri bulunduğu için bu yıllara ait veriler kendi içlerinde ve bir araya getirilerek değerlendirilmiştir. 2011 yılında başlayan öğrencilerin büyük çoğunluğunun eğitimi devam etmekte olduğu için başarı ölçütlerinden birisi olan toplam eğitim yılında yanlış sonuçlara yönlendirebileceği kaygısıyla, 2011 yılında derse başlayan öğrencilerin verileri kendi içinde değerlendirilememektedir. Eğitimi devam etmekte olan öğrencilerin girdileri toplam eğitim yılının öznitelik olarak bulunduğu veri setlerinden çıkarılmıştır.

Özniteliklerin azaltılması, başarı ölçütlerinin sınıflanmasında doğruluk oranlarını nasıl etkilediğinin de gözlemlenebilmesi için her bir deneme iki veri setiyle yapılmıştır. Veri setlerinden birisi diğerinden daha çok öznitelik içermektedir. Bu yaklaşımla sonuçların hangi özniteliklere daha bağlı olduğu gözlemlenmek istenmiştir. Veri setlerinin adları ve içerikleri aşağıda verilmektedir.

V1 : {*TopEgtYılı, HazDuru, HazNot, AdresIl, Cinsiyet, BursKredi, GNO*}

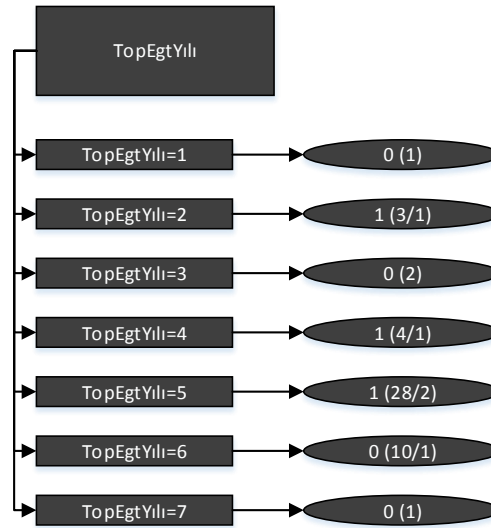
V2 : { *TopEgtYılı, BursKredi, GNO* }

2008 yılında birinci sınıfa başlayan öğrencilerin verilerinin bulunduğu veri setleri ile çalıştırılan sınıflama algoritmalarının sonuçları Çizelge 4.7’de verilmektedir. Çizelgede de görüldüğü üzere sınıf olarak GNO özniteliği seçildiğinde V1 ve V2 setleri arasında doğru sınıflandırma farkı fazla çıkmaktadır. Bu sonuç veri azaltma yoluna gidildiğinde GNO için önemli olabilecek bazı bilgilerin kaybolduğu anlamında gelmektedir. J48 algoritmasının 2008 verileri için daha iyi sonuç verdiği görülmektedir.

Burs veya kredi alan öğrencilerin sınıflaması en doğru sonuçlara sahip olup, J48 algoritmasının oluşturduğu Şekil 4.7’deki ağaç yapısı, V1 ve V2 veri setleri için birebir aynıdır. Şekildeki ağaç yapısına göre toplam eğitim yılı ile öğrencilerin burs veya kredi alıp almamasının arasında bağlantı olduğu gözükmemektedir. Eğitim yılı altı olan öğrencilerin çoğunun burs veya kredi almadığı fakat eğitim yılı dört ve beş olan öğrencilerin çoğunun burs veya kredi aldığı anlaşılmaktadır.

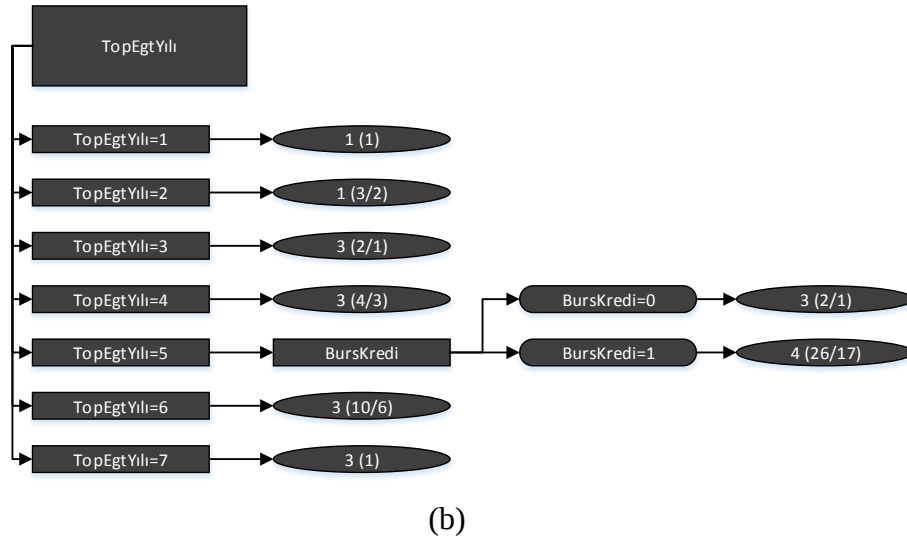
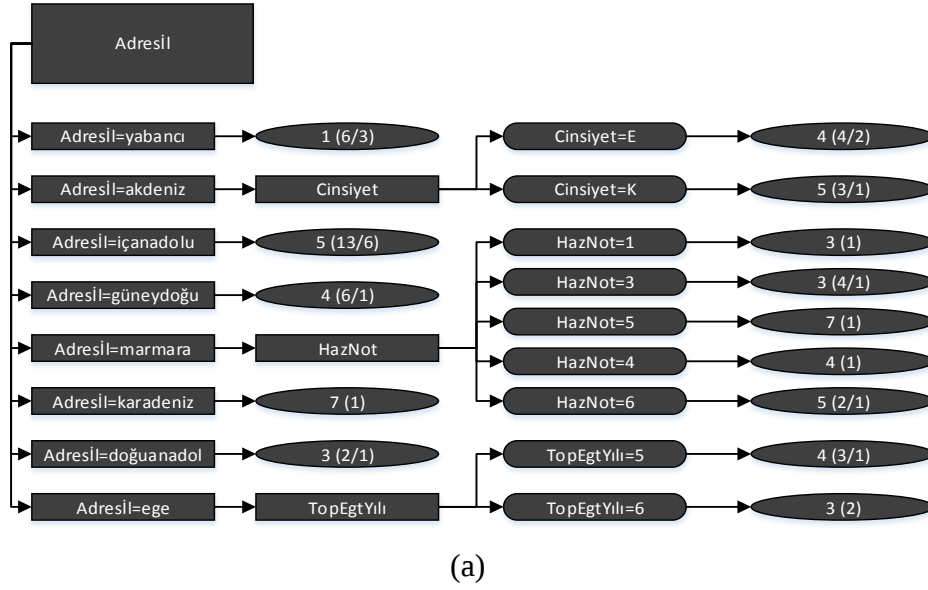
Çizelge 4.7: 2008 yılı, V1 ve V2 veri setlerine uygulanan sınıflama algoritmalarının çıktıları

<i>Sınıf</i>	<i>Sınıflandırıcı</i>	<i>Parametre</i>	<i>TP Oranı</i>	<i>FP Oranı</i>	<i>Kesinlik</i>	<i>Duyarlılık</i>	<i>F-Ölçütü</i>	<i>Doğruluk</i>	
GNO	V1	kNN	k=9,1/d	0.286	0.204	0.218	0.286	0.244	28.57
	V2	kNN	k=1,1/d	0.224	0.236	0.1	0.224	0.136	22.49
	V1	J48	c=0.25	0.388	0.147	0.331	0.388	0.357	38.57
	V2	J48	c=0.7	0.265	0.226	0.128	0.265	0.167	26.53
TopEgtYıl	V1	kNN	k=3,1/d	0.694	0.293	0.547	0.694	0.611	69.38
	V2	kNN	k=4,1/d	0.653	0.201	0.574	0.653	0.609	65.30
	V1	J48	c=0.25	0.714	0.205	0.572	0.714	0.631	71.42
	V2	J48	c=0.25	0.714	0.205	0.572	0.714	0.631	71.42
BursKredi	V1	kNN	k=3	0.796	0.301	0.793	0.796	0.788	79.59
	V2	kNN	k=3,1/d	0.816	0.263	0.814	0.816	0.812	81.63
	V1	J48	c=0.25	0.837	0.252	0.838	0.837	0.831	83.67
	V2	J48	c=0.25	0.837	0.252	0.838	0.837	0.831	83.67



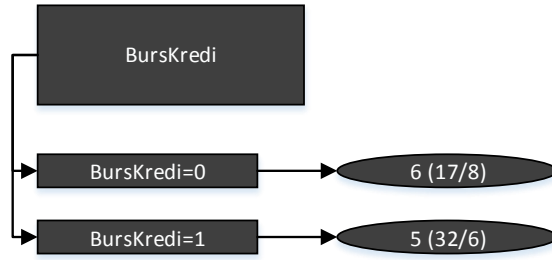
Şekil 4.7: 2008 yılı, J48 algoritmasının oluşturduğu BursKredi sınıfının ağaç yapısı

Genel not ortalamaları sınıflandırılmak istendiğinde J48 tarafından oluşturulan Şekil 4.8'deki ağaç yapısı incelenecek olursa, yararlı olabilecek kurallar çıkarılamadığı görülmektedir. Doğru sınıflandırma oranlarının düşük olmasından da anlaşılacağı üzere ağaç yapısı ayırt edici fazla bir bilgi içermemektedir.



Şekil 4.8: 2008 yılı, J48 algoritmasının oluşturduğu GNO sınıfının ağaç yapısı a) V1 veri seti için b) V2 veri seti için

Toplam eğitim yılı sınıf olarak seçildiğinde burs veya kredi alınmasının etkileri Şekil 4.9'daki ağaç yapısında görülmektedir. 2008 yılı verileri için burs veya kredi almayanların çoğu altı yılda mezun olurken burs veya kredi alanlar beş yılda yani zamanında mezun olmuşlardır. Ağaç yapısı V1 ve V2 veri setleri için birebir aynı çıkmıştır.



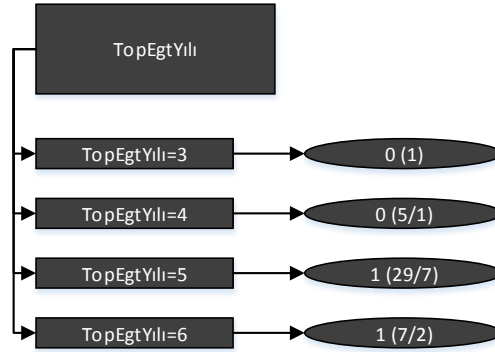
Şekil 4.9: 2008 yılı, J48 algoritmasının oluşturduğu TopEgtYılı sınıfının ağaç yapısı

2009 yılında birinci sınıfa başlayan öğrencilerin verilerinin bulunduğu veri setleri ile çalıştırılan sınıflama algoritmalarının sonuçları Çizelge 4.8’de verilmektedir. V1 ve V2 veri setleri arasındaki benzerlik 2008 verilerine göre daha fazla gözükmemektedir. Genel not ortalaması sınıf olarak seçildiğinde doğru tanıma oranı düşük seviyede olduğu görülmektedir. J48 algoritmasının bu veri setinde çoğunlukla daha iyi sonuçlar verdiği kanısına varılmaktadır.

Çizelge 4.8: 2009 yılı, V1 ve V2 veri setlerine uygulanan sınıflama algoritmalarının çıktıları

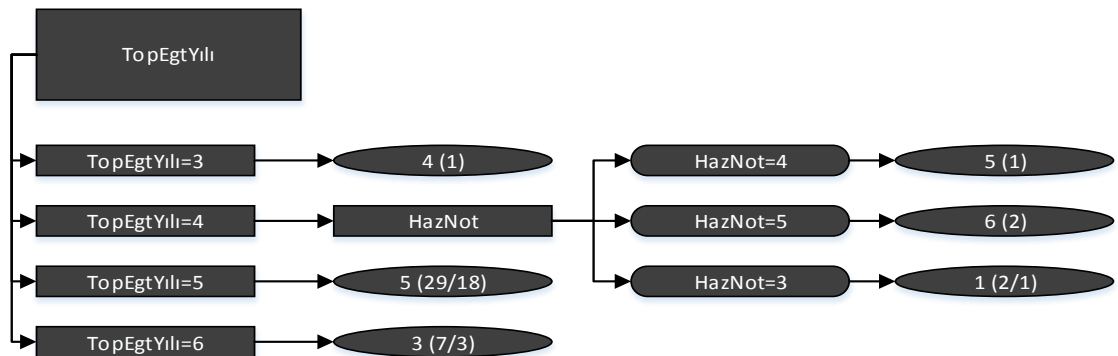
<i>Sınıf</i>	<i>Sınıflandırıcı</i>	<i>Parametre</i>	<i>TP Oranı</i>	<i>FP Oranı</i>	<i>Kesinlik</i>	<i>Duyarlılık</i>	<i>F-Ölçütü</i>	<i>Doğruluk</i>	
GNO	V1	kNN	k=4,1/d	0.19	0.33	0.145	0.19	0.16	19.05
	V2	kNN	k=4	0.333	0.269	0.279	0.333	0.3	33.33
	V1	J48	c=0.1	0.262	0.331	0.099	0.262	0.144	26.19
	V2	J48	c=0.1	0.262	0.317	0.165	0.262	0.192	26.19
TopEgtYıl	V1	kNN	k=4,1/d	0.667	0.591	0.545	0.667	0.597	66.67
	V2	kNN	k=3,1/d	0.714	0.539	0.652	0.714	0.648	71.42
	V1	J48	c=0.8	0.714	0.587	0.575	0.714	0.628	74.43
	V2	J48	c=0.3	0.69	0.491	0.558	0.69	0.617	69.05
BursKredi	V1	kNN	k=5	0.69	0.583	0.684	0.69	0.616	69.05
	V2	kNN	k=3,1/d	0.667	0.524	0.637	0.667	0.637	66.67
	V1	J48	c=0.35	0.714	0.536	0.724	0.714	0.657	71.43
	V2	J48	c=0.35	0.69	0.491	0.558	0.69	0.617	69.05

Burs veya kredi alan öğrencilerin çoğu beş veya altı yılda mezun olurken almayan öğrencilerin üç veya dört yılda mezun olduğu Şekil 4.10'da görülmektedir. V1 ve V2 veri setleri için aynı ağaç yapıları oluşmuştur.



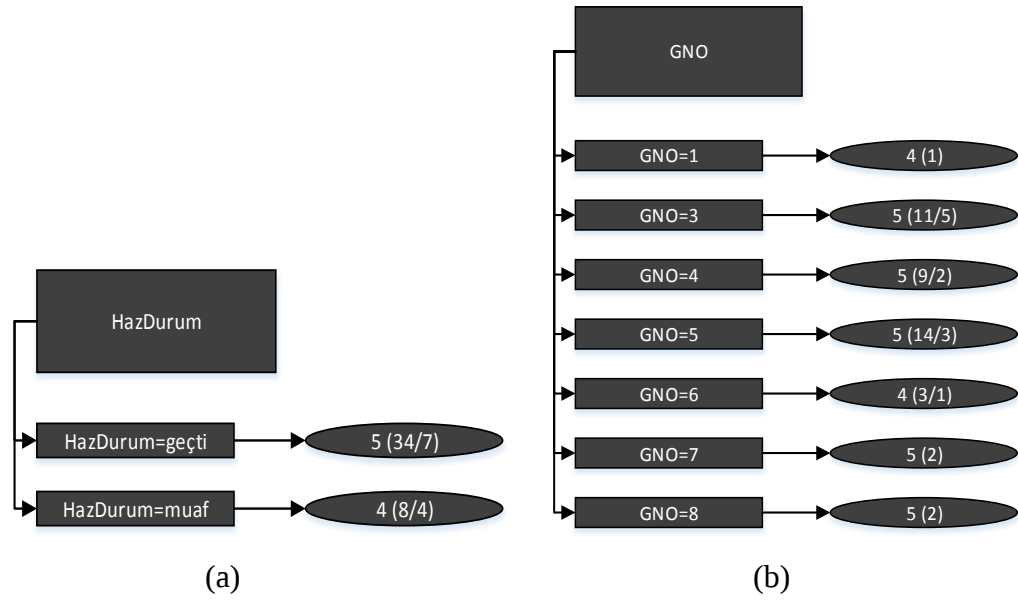
Şekil 4.10: 2009 yılı, J48 algoritmasının oluşturduğu BursKredi sınıfının ağaç yapısı

Genel not ortalamasının sınıf olarak seçildiği durumda V1 veri seti için oluşan ağaç yapısı Şekil 4.11'de verilmektedir. Ağaç yapısı toplam eğitim yılıyla ve hazırlık sınıfı notuyla ilişkilendirilmiştir. Not ortalaması iyi olan öğrencilerin zamanında mezun oldukları tahmin edilebilir bir bilgi olduğundan, katkısı azdır.



Şekil 4.11: 2009 yılı, J48 algoritmasının oluşturduğu GNO sınıfının ağaç yapısı

Toplam eğitim yılı sınıf olarak seçildiğinde Şekil 4.12’de görüldüğü üzere V1 ve V2 veri setleri için farklı ağaç yapıları oluşmuştur. Çizelgede toplam eğitim yılı için doğru sınıflama oranının düşüş göstermesi bilgi kaybı olduğu yönünde yorumlanabilir (Bkz. Çizelge 4.8).



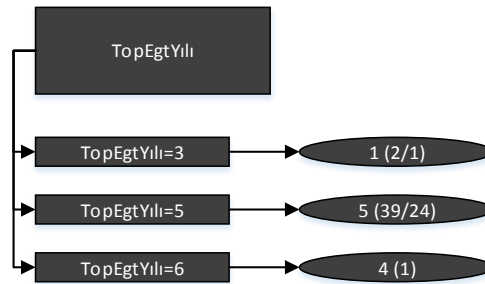
Şekil 4.12: 2009 yılı, J48 algoritmasının oluşturduğu TopEgtYılı sınıfının ağaç yapısı a) V1 veri seti için b) V2 veri seti için

2010 yılında birinci sınıfa başlayan öğrencilerin verilerinin bulunduğu veri setleri ile çalıştırılan sınıflama algoritmalarının sonuçları Çizelge 4.9’da verilmektedir. Toplam eğitim yılının sınıf olarak kullanıldığı durumdaki sonuçlar dikkat çekmektedir. Doğru sınıflama oranının yüzde 93'lere varmış olduğu gözükmemektedir. İyi sonuç verdiği düşünülse de FP oranına bakıldığında yanlışlıkla bir sınıfa dahil edilen girdilerin sayısının da çok olduğu gözükmemektedir. Fazla girdi sayısına sahip sınıflar baskın oldukları için sonuçları bu şekilde etkileyebilmektedir.

Çizelge 4.9: 2010 yılı V1 ve V2 veri setlerine uygulanan sınıflama algoritmalarının çıktıları

<i>Sınıf</i>	<i>Sınıflandırıcı</i>	<i>Parametre</i>	<i>TP Oranı</i>	<i>FP Oranı</i>	<i>Kesinlik</i>	<i>Duyarlılık</i>	<i>F-Ölçütü</i>	<i>Doğruluk</i>	
GNO	V1	kNN	k=12	0.31	0.366	0.116	0.31	0.169	30.95
	V2	kNN	k=7	0.333	0.332	0.161	0.333	0.209	33.33
	V1	J48	c=0.25	0.357	0.357	0.128	0.357	0.188	35.71
	V2	J48	c=0.25	0.33	0.363	0.122	0.333	0.179	33.33
TopEgtYıl	V1	kNN	k=3,1/d	0.929	0.929	0.862	0.929	0.894	92.86
	V2	kNN	k=1,1/d	0.929	0.929	0.862	0.929	0.894	92.86
	V1	J48	c=0.25	0.929	0.929	0.862	0.929	0.894	92.86
	V2	J48	c=0.25	0.929	0.929	0.862	0.929	0.894	92.86
BursKredi	V1	kNN	k=7,1/d	0.762	0.792	0.613	0.762	0.68	76.19
	V2	kNN	k=3,1/d	0.786	0.786	0.617	0.786	0.691	78.57
	V1	J48	c=0.25	0.786	0.786	0.617	0.786	0.691	78.57
	V2	J48	c=0.25	0.786	0.786	0.617	0.786	0.691	78.57

Sadece V2 veri setinde GNO için Şekil 4.13'teki ağaç yapısı oluşmuştur diğer sınıflar için J48 algoritması ağaç yapısı oluşturamamıştır. Toplam eğitim yılı beş olan öğrencilerin 39'unun GNO değeri de 5 gözükmektedir. Bu 39 öğrenciden 24'ü yanlış olarak bu sınıfa atanmıştır.



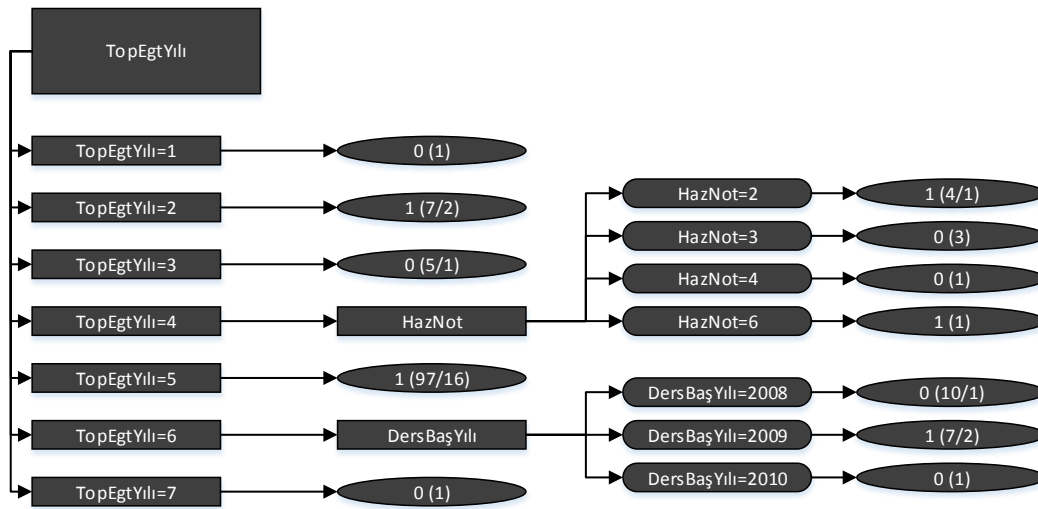
Şekil 4.13: 2010 yılı, J48 algoritmasının oluşturduğu GNO sınıfının ağaç yapısı

2008 ve 2011 aralarındaki yıllara ait veriler de olmak üzere birleştirilip aynı işlemler tekrar edilmiştir. V1 veri setine, öğrencilerin derse başlama yılları da öznel olarak eklenmiştir. Çizelge 4.10'da sınıflama algoritmalarının sonuçları verilmiştir.

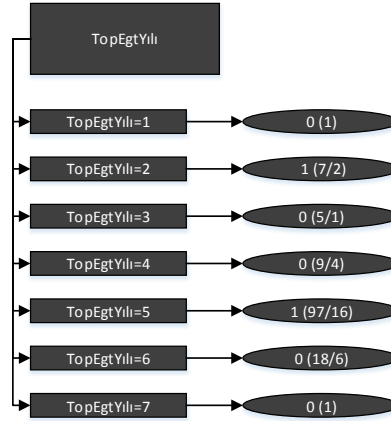
Çizelge 4.10: 2008-2011 yılları V1 ve V2 veri setlerine uygulanan sınıflama algoritmalarının çıktıları

Sınıf		Sınıflandırıcı	Parametre	TP Oranı	FP Oranı	Kesinlik	Duyarlılık	F-Ölçütü	Doğruluk
GNO	V1	kNN	k=13	0.297	0.259	0.215	0.297	0.233	29.71
	V2	kNN	k=1,1/d	0.29	0.261	0.158	0.29	0.195	28.98
	V1	J48	c=0.15	0.268	0.264	0.14	0.268	0.171	26.81
	V2	J48	c=0.25	0.312	0.253	0.181	0.312	0.212	31.15
TopEgtYıl	V1	kNN	k=3	0.725	0.554	0.605	0.725	0.641	72.46
	V2	kNN	k=12,1/d	0.652	0.676	0.49	0.652	0.56	65.22
	V1	J48	c=0.15	0.696	0.473	0.583	0.696	0.629	69.56
	V2	J48	c=0.25	0.703	0.703	0.494	0.703	0.58	70.29
BursKredi	V1	kNN	k=3	0.703	0.576	0.664	0.703	0.661	70.29
	V2	kNN	k=11,1/d	0.696	0.565	0.658	0.696	0.661	69.56
	V1	J48	c=0.3	0.79	0.427	0.787	0.79	0.768	78.98
	V2	J48	c=0.25	0.775	0.377	0.765	0.775	0.766	77.53

J48 algoritması *BursKredi* özniteliği için V1 ve V2 veri setlerinde farklı ağaç yapıları oluşturmuştur. Şekil 4.14'te görüldüğü üzere V1 veri setinde toplam eğitim yılı altı olan öğrencilerin burs veya kredi alması yıllara göre değişiklik göstermektedir. V2 veri seti için oluşan ağaç yapısında ise sadece toplam eğitim yılının etkili olduğu gözükmemektedir (Şekil 4.15)

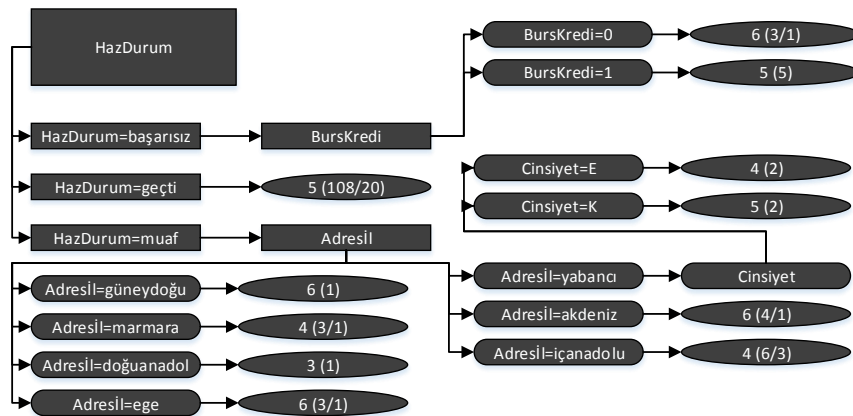


Şekil 4.14: 2008-2011 yılları, J48 algoritmasının V1 veri seti için oluşturduğu BursKredi sınıfının ağaç yapısı



Şekil 4.15: 2008-2011 yılları, J48 algoritmasının V2 veri seti için oluşturduğu BursKredi sınıfının ağaç yapısı

Genel not ortalaması için oluşturulan ağaç yapısı, doğruluk oranı düşük olduğundan yanıltıcı sayılabilecek bilgiler çıkarmıştır. Toplam eğitim yılı sınıfı olarak seçildiğinde sadece V1 veri seti için ağaç yapısı oluşmuştur. Şekil 4.16'daki ağaç yapısından elde edilen, hazırlığı geçen öğrenciler beş yılda mezun olurlar kuralı, çoğunluk için doğru olmakla beraber önemli bilgi olarak görülmemiştir. Normal mezun olma süresi hazırlık sınıfı da dahil beş yıl olduğu için bu durumun dışındaki öğrencilerin girdilerinin düzgün sınıflanmadığı ortaya çıkmaktadır.



Şekil 4.16: 2008-2011 yılları J48 algoritmasının oluşturduğu TopEgtYılı sınıfının ağaç yapısı

Birliktelek kuralları algoritmalarının ürettiği kurallar çok fazla sayıda olabildiği için her bir algoritma için en iyi derecelere sahip olan on kural sunulmaktadır. Apriori algoritması tarafından çıkarılan kurallar Şekil 4.17’de verilmiştir. Predictive Apriori tarafından üretilen birliktelek kurallarıysa Şekil 4.18’de verilmiştir. Kurallarda parantez içinde yazılan sayılar girdi sayısını vermektedir. Örneğin “BursKredi=1 (97) ==> TopEgtYıl=5 (81)” burs veya kredi alan 97 kişiden 81’i beş yılda mezun olmuştur anlamına gelmektedir.

<i>Kural</i>	<i>Birliktelek Kuralları</i>	<i>Güv</i>
1	BursKredi=1 GNO=4 (21) ==> TopEgtYıl=5 (20)	0.95
2	BursKredi=1 GNO=5 (31) ==> TopEgtYıl=5 (28)	0.9
3	TopEgtYıl=5 GNO=4 (23) ==> BursKredi=1 (20)	0.87
4	BursKredi=1 (97) ==> TopEgtYıl=5 (81)	0.84
5	TopEgtYıl=5 (97) ==> BursKredi=1 (81)	0.84
6	TopEgtYıl=5 GNO=5 (34) ==> BursKredi=1 (28)	0.82
7	GNO=5 (42) ==> TopEgtYıl=5 (34)	0.81
8	GNO=4 (29) ==> TopEgtYıl=5 (23)	0.79
9	GNO=5 (42) ==> BursKredi=1 (31)	0.74
10	GNO=4 (29) ==> BursKredi=1 (21)	0.72

Şekil 4.17: 2008-2011 yılları için Apriori algoritmasının çıkardığı kurallar

<i>Kural</i>	<i>Birliktelek Kuralları</i>	<i>Doğ.</i>
1	BursKredi=1 GNO=7 (7) ==> TopEgtYıl=5 (7)	0.96
2	BursKredi=1 GNO=4 (21) ==> TopEgtYıl=5 (20)	0.94
3	TopEgtYıl=5 GNO=8 (5) ==> BursKredi=1 (5)	0.94
4	BursKredi=1 GNO=8 (5) ==> TopEgtYıl=5 (5)	0.94
5	GNO=9 (2) ==> BursKredi=1 (2)	0.86
6	TopEgtYıl=3 GNO=4 (2) ==> BursKredi=0 (2)	0.86
7	TopEgtYıl=5 GNO=1 (2) ==> BursKredi=1 (2)	0.86
8	BursKredi=1 GNO=5 (31) ==> TopEgtYıl=5	0.85
9	TopEgtYıl=5 (97) ==> BursKredi=1 (81)	0.82
10	BursKredi=1 (97) ==> TopEgtYıl=5 (81)	0.82

Şekil 4.18: 2008-2011 yılları için Predictive Apriori algoritmasının çıkardığı kurallar

4.4.2. Demografik özelliklerin öğrenci başarısı üzerine etkisi

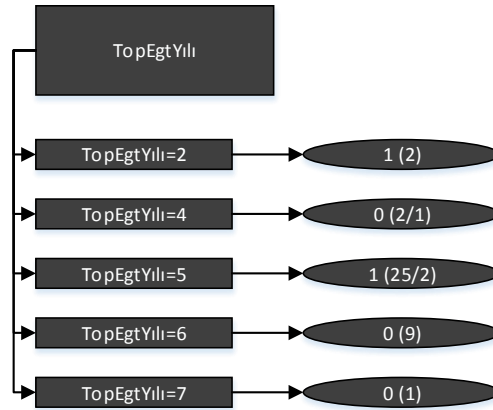
Demografik özelliklerin öğrenci başarısını ne yönde etkilediği ve özelliklerin birbirleri arasındaki ilişkinin belirlenmesi amaçlanmıştır. Yıllara göre ayrı ve bütün yılların bir arada olduğu veri setleri oluşturulmuştur. Veri setlerinde kullanılan öznitelikler şunlardır: *TopEgtYılı*, *AdresIl*, *Cinsiyet*, *BursKredi*, *Parasal*, *AnneEgt*, *AnneMes*, *BabaEgt*, *BabaMes*, *KalacakYer* ve *GNO*.

2008 yılında birinci sınıf derslerine başlamış öğrencilerin demografik verilerinde çalıştırılan sınıflama algoritmalarının sonuçları Çizelge 4.11’de verilmektedir. Çizelge, *GNO*, *TopEgtYılı*, *BursKredi* sınıflarının doğru sınıflara atanma oranlarının önceki veri setinden elde edilen sonuçlardan yüksek olduğunu göstermektedir (Bkz. Çizelge 4.7). J48 algoritmasının doğru sınıflandırma değerlerinin kNN algoritmasından daha büyük olduğu gözükmemektedir. Öğrencilerin babalarının eğitim durumları ve mesleklerinin doğru sınıflandırılmadığı görülmektedir. Öğrencilerin annelerinin eğitim bilgilerinin yarıya yakını doğru sınıflandırılırken, meslek bilgilerinin ise yüzde 75’e varan oranlarda doğru sınıflandırıldığı görülmektedir. Bu yüksek sonuçların çıkmasındaki sebeplerden birisi ev hanımı olan annelerin sayısının diğer meslek gruplarında olan annelerin sayısından çok yüksek olmasıdır. Sınıflayıcılar baskın olan sınıfa atama yaptıklarında oranlar yüksek çıkmaktadır. FP oranlarının yüksek çıkması da bu savı destekler niteliktedir.

Çizelge 4.11: 2008 yılı demografik verilere uygulanan sınıflama algoritmalarının çıktıları

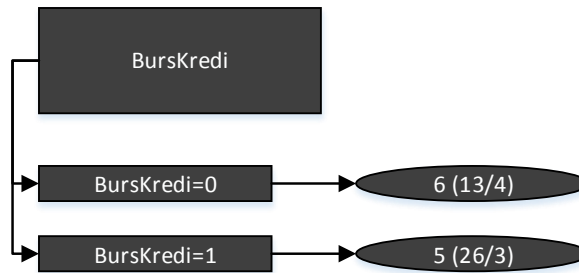
Sınıf	Sınıflandırıcı	Parametre	TP Oranı	FP Oranı	Kesinlik	Duyarlılık	F-Ölçütü	Doğruluk
GNO	kNN	k=10	0.359	0.203	0.323	0.359	0.332	35.89
	J48	c=0.1	0.41	0.164	0.349	0.41	0.377	41.02
TopEgtYılı	kNN	k=13,1/d	0.615	0.649	0.405	0.615	0.488	61.53
	J48	c=0.25	0.821	0.168	0.727	0.821	0.767	82.05
BursKredi	kNN	k=1,1/d	0.846	0.192	0.846	0.846	0.846	84.61
	J48	c=0.25	0.872	0.218	0.875	0.872	0.867	87.17
AnneEgt	kNN	k=10,1/d	0.436	0.429	0.393	0.436	0.387	43.58
	J48	c=0.7	0.487	0.271	0.488	0.487	0.486	48.71
AnneMes	kNN	k=5,1/d	0.744	0.771	0.587	0.744	0.656	74.36
	J48	c=0.25	0.769	0.769	0.592	0.769	0.669	76.92
BabaEgt	kNN	k=1,1/d	0.231	0.353	0.223	0.231	0.225	23.07
	J48	c=0.25	0.359	0.307	0.346	0.359	0.351	35.89
BabaMes	kNN	k=3,1/d	0.359	0.24	0.325	0.359	0.327	35.89
	J48	c=0.7	0.282	0.256	0.3	0.282	0.269	28.20

BursKredi sınıfının J48 tarafından oluşturulmuş Şekil 4.19'daki ağaç yapısına bakıldığında toplam eğitim süreleri altı yıl olan öğrencilerin burs veya kredi almadığı, eğitim süreleri beş yıl olan öğrencilerinse burs veya kredi aldığı görülmektedir.



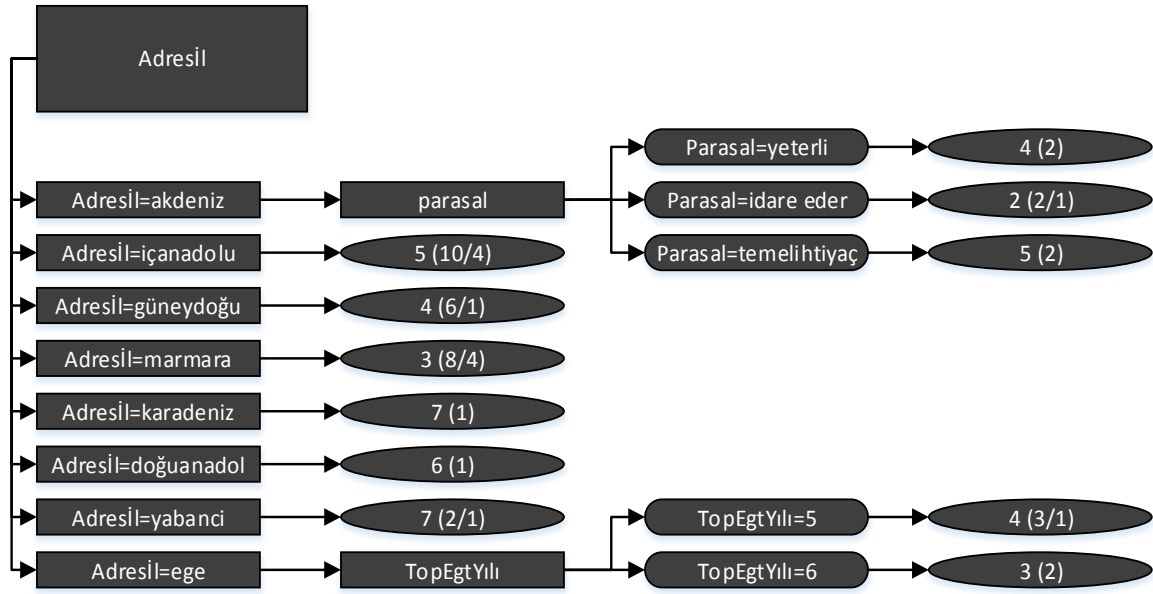
Şekil 4.19: 2008 yılı, demografik verilere uygulanan J48 algoritmasının oluşturduğu *BursKredi* sınıfının ağaç yapısı

TopEgtYılı sınıfı *BursKredi* özneliği kullanılarak sınıflandırılmıştır. Şekil 4.20'de eğitim süresinin öğrencilerin burs veya kredi almalarıyla bağlantılı olduğu sonucu çıkmaktadır.



Şekil 4.20: 2008 yılı, demografik verilere uygulanan J48 algoritmasının oluşturduğu *TopEgtYılı* sınıfının ağaç yapısı

GNO sınıfı için oluşturulmuş ağaç yapısı Şekil 4.21’de verilmiştir. Öğrencilerin ikamet ettikleri coğrafi bölgelerle not ortalamaları arasında bağlantı kurulduğu gözükmemektedir. Yalnız doğru sınıflama oranı düşük olduğu için ağaç yapısından elde edilecek kuralların güvenilirliği çok az olmaktadır.

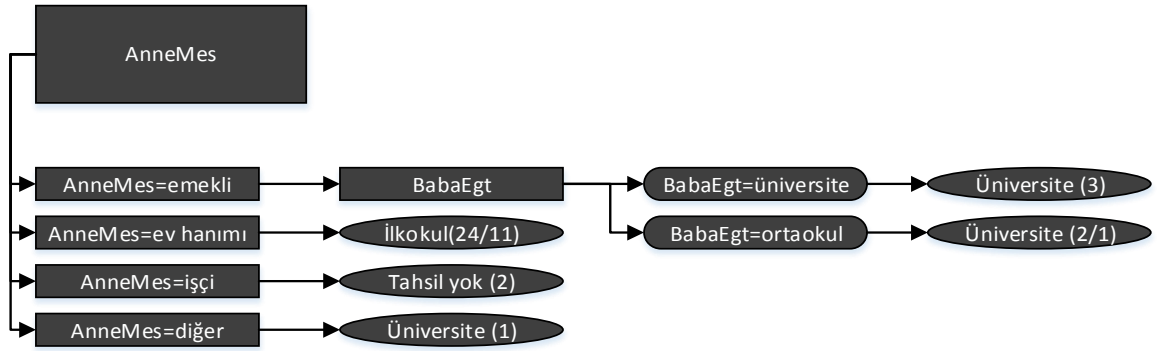


Şekil 4.21: 2008 yılı, demografik verilere uygulanan J48 algoritmasının oluşturduğu GNO sınıfının ağaç yapısı

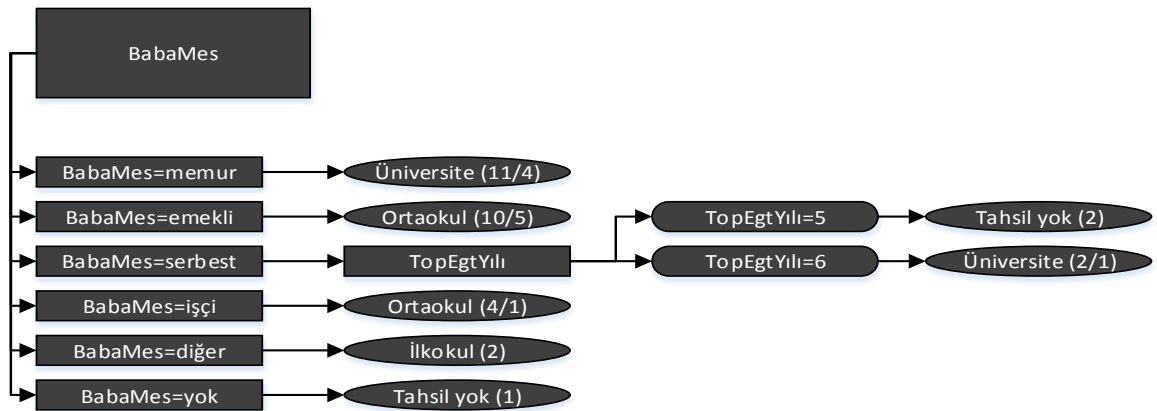
2009 yılında birinci sınıf derslerine başlamış olan öğrencilerin sınıflama algoritmaları tarafından elde edilen çıktıları Çizelge 4.12’de verilmektedir. *BursKredi* sınıfı hariç diğer iki sınıf için sonuçlar daha yüksek doğruluğa sahiptir (Bkz. Çizelge 4.8). J48 algoritması çoğu durumda, bu veri seti için de kNN algoritmasından yüksek doğruluk değerlerine sahiptir. Şekil 4.22’de öğrencilerin ebeveynlerinin eğitim durumları yine ebeveynlerinin meslekleri ile ilişkilendirilmiştir. Öğrencilerin annelerinin eğitim durumları çoğunluğu için ilköğretim olduğundan diğer eğitim seviyeleri de yanlışlıkla ilköğretim olarak sınıflandırılmıştır.

Çizelge 4.12: 2009 yılı Demografik verilere uygulanan sınıflama algoritmalarının çıktıları

Sınıf	Sınıflandırıcı	Parametre	TP Oranı	FP Oranı	Kesinlik	Duyarlılık	F-Ölçütü	Doğruluk
GNO	kNN	k=8,1/d	0.313	0.312	0.243	0.313	0.254	31.25
	J48	c=0.25	0.469	0.149	0.504	0.469	0.48	46.87
TopEgtYılı	kNN	k=1,1/d	0.813	0.47	0.772	0.813	0.781	81.25
	J48	c=0.25	0.75	0.75	0.563	0.75	0.643	75.0
Burs Kredi	kNN	k=4,1/d	0.688	0.731	0.51	0.688	0.586	68.75
	J48	c=0.7	0.688	0.528	0.667	0.688	0.674	68.75
AnneEgt	kNN	k=9	0.344	0.389	0.261	0.344	0.286	34.37
	J48	c=0.25	0.5	0.318	0.498	0.5	0.428	50.0
AnneMes	kNN	k=7,1/d	0.719	0.756	0.556	0.719	0.627	71.87
	J48	c=0.7	0.719	0.397	0.703	0.719	0.71	71.87
BabaEgt	kNN	k=1	0.313	0.33	0.356	0.313	0.317	31.25
	J48	c=0.25	0.5	0.236	0.564	0.5	0.483	50.0
BabaMes	kNN	k=3,1/d	0.313	0.313	0.245	0.313	0.245	31.25
	J48	c=0.1	0.344	0.293	0.33	0.344	0.313	34.37



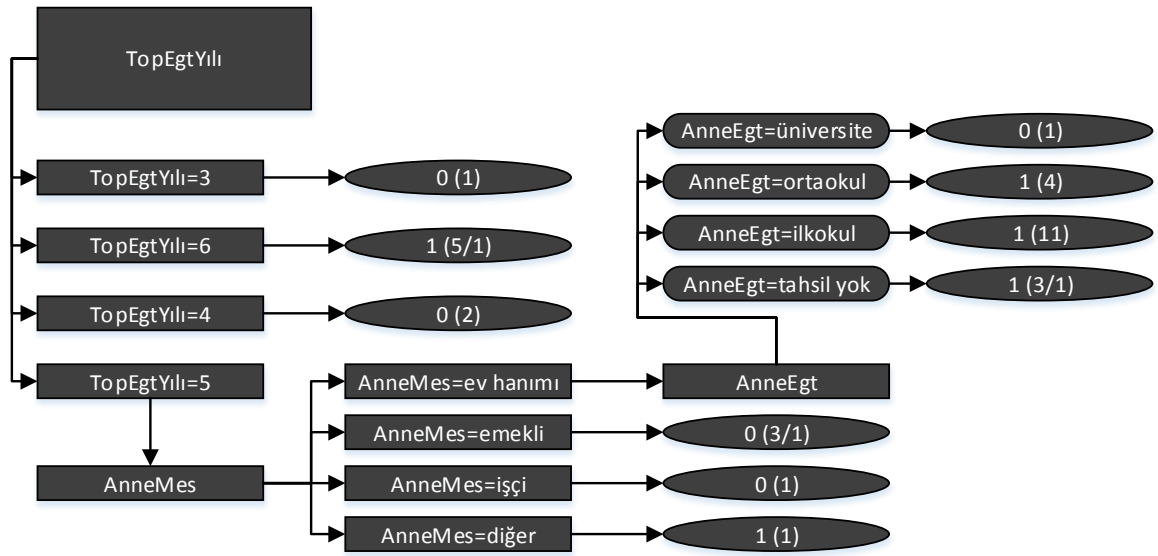
(a)



(b)

Şekil 4.22: 2009 yılı, demografik verilere uygulanan J48 algoritmasının oluşturduğu ağaç yapıları a) AnneEgt sınıfı için b) BabaEgt sınıfı için

Öğrencilerin burs veya kredi almaları, öğrencilerin annelerinin eğitim seviyeleri, meslekleri ve öğrencilerin eğitim süreleri ile ilişkilendirilmiştir (Şekil 4.23). Anneleri tahsil görememiş, ilköğretim ve ortaöğretime gitmiş öğrencilerin burs veya kredi aldıkları saptanmaktadır.



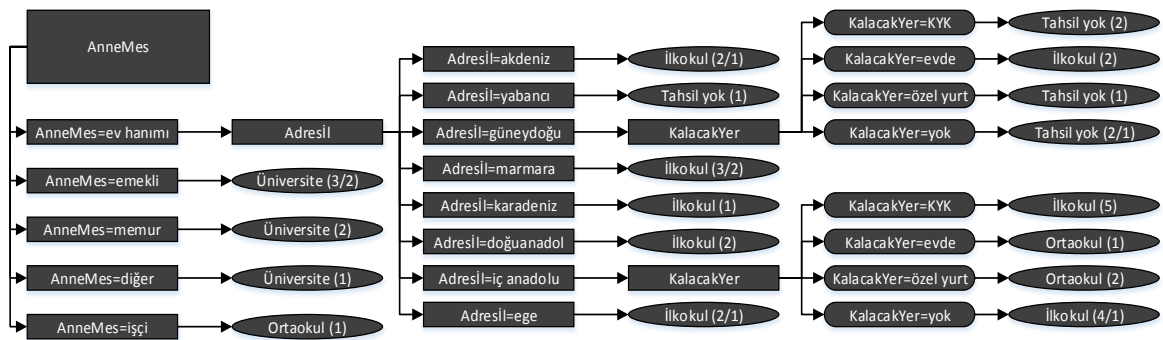
Şekil 4.23: 2009 yılı, demografik verilere uygulanan J48 algoritmasının oluşturduğu BursKredi sınıfının ağaç yapısı

Birinci sınıf derslerine 2010 yılında başlayan öğrencilerin verilerinde çalıştırılan sınıflama algoritmalarının çıktıları Çizelge 4.13'te sunulmaktadır. Toplam eğitim yılı çok yüksek oranda bir doğrulukla sınıflandırılmasına karşın FP oranının yüksek olmasından yanlış sınıflama oranının da büyük olduğu anlaşılmaktadır. Algoritma tarafından az olan sınıfların düzgün sınıflandırılmadığı sonucu çıkmaktadır. Bu veri setinde kNN algoritması J48'e göre daha iyi sonuçlar elde etmiştir.

Çizelge 4.13: 2010 yılı Demografik verilere uygulanan sınıflama algoritmalarının çıktıları

Sınıf	Sınıflandırıcı	Parametre	TP Oranı	FP Oranı	Kesinlik	Duyarlılık	F-Ölçütü	Doğruluk
GNO	kNN	k=3	0.243	0.282	0.204	0.243	0.176	24.32
	J48	c=0.25	0.351	0.226	0.238	0.351	0.28	35.13
TopEgtYili	kNN	k=2,1/d	0.919	0.947	0.893	0.919	0.906	91.89
	J48	c=0.25	0.946	0.946	0.895	0.946	0.92	94.59
BursKredi	kNN	k=3	0.811	0.595	0.79	0.811	0.775	81.08
	J48	c=0.7	0.784	0.603	0.749	0.784	0.755	78.37
AnneEgt	kNN	k=4,1/d	0.459	0.474	0.351	0.459	0.344	45.94
	J48	c=0.7	0.405	0.353	0.383	0.405	0.392	40.54
AnneMes	kNN	k=1	0.811	0.697	0.712	0.811	0.756	81.08
	J48	c=0.7	0.757	0.36	0.732	0.757	0.744	75.67
BabaEgt	kNN	k=2	0.378	0.222	0.373	0.378	0.362	37.83
	J48	c=0.25	0.351	0.268	0.32	0.351	0.329	35.13
BabaMes	kNN	k=2	0.297	0.27	0.258	0.297	0.269	29.72
	J48	c=0.7	0.297	0.25	0.27	0.297	0.285	29.72

Öğrencilerin annelerinin eğitim durumları annelerinin meslekleri ile bağlantılı olduğu görülmektedir. Ek olarak öğrencilerin ikametgâh ettikleri coğrafi bölge ile de bağlantı kurulmaktadır (Şekil 4.24). Mesleği ev hanımı, ikamet edilen bölge Güneydoğu Anadolu olan annelerin çoğunluğu hiç okula gitmemiş olduğu çıkarılmaktadır.



Şekil 4.24: 2010 yılı, demografik verilere uygulanan J48 algoritmasının oluşturduğu AnneEgt sınıfının ağaç yapısı

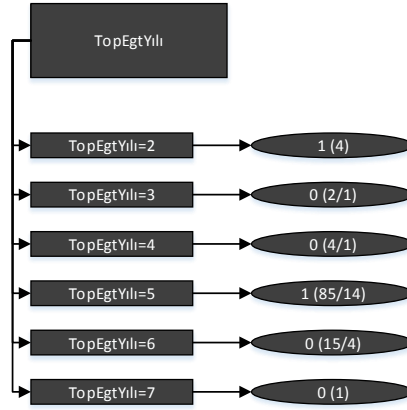
2008 ve 2011 yılları ile aralarında kalan yıllarda birinci sınıf derslerine başlayan öğrencilerin verileri birleştirilerek, veri madenciliği işlemleri tekrar edilmiştir. Çıktılar Çizelge 4.14’te verilmektedir.

Çizelge 4.14: 2008-2011 yılları Demografik verilere uygulanan sınıflama algoritmalarının çıktıları

Sınıf	Sınıflandırıcı	Parametre	TP Oranı	FP Oranı	Kesinlik	Duyarlılık	F-Ölçütü	Doğruluk
GNO	kNN	k=13,1/d	0.378	0.242	0.263	0.378	0.297	37.83
	J48	c=0.25	0.243	0.258	0.168	0.243	0.189	24.32
TopEgtYılı	kNN	k=7,1/d	0.757	0.767	0.585	0.757	0.66	75.67
	J48	c=0.7	0.685	0.547	0.647	0.685	0.665	68.46
BursKredi	kNN	k=5,1/d	0.712	0.673	0.64	0.712	0.645	71.17
	J48	c=0.25	0.784	0.458	0.769	0.784	0.764	78.37
AnneEgt	kNN	k=10,1/d	0.441	0.388	0.369	0.441	0.379	44.14
	J48	c=0.7	0.514	0.269	0.524	0.514	0.517	51.35
AnneMes	kNN	k=3	0.784	0.72	0.674	0.784	0.707	78.37
	J48	c=0.7	0.748	0.403	0.711	0.748	0.729	74.77
BabaEgt	kNN	k=9,1/d	0.351	0.3	0.336	0.351	0.341	35.13
	J48	c=0.7	0.495	0.216	0.47	0.495	0.481	49.54
BabaMes	kNN	k=10,1/d	0.36	0.257	0.297	0.36	0.321	36.03
	J48	c=0.1	0.405	0.255	0.389	0.405	0.371	40.54

J48 algoritmasının *BursKredi* özniteliği sınıf olarak seçildiğinde oluşturduğu ağaç yapısı Şekil 4.25’te verilmiştir. Geri kalan öznitelikler için ağaç yapıları oluşmadığı ya da çok fazla dallandığı için sunulmamıştır. Tüm yıllara ait veriler kullanıldığında oluşan ağaç yapısı 2008 için oluşan ağaç yapısıyla benzerlikler göstermektedir (Bkz. Şekil 4.18).

Apriori ve Predictive Apriori algoritmaları tarafından çıkarılan en iyi onar kural Şekil 4.26’da verilmiştir. Kurallar görüldüğü gibi öğrencilerin annelerinin eğitim durumları, meslekleri ile ilgili bilgiler etrafında yoğunlaşmaktadır. Öğrencilerin burs veya kredi alması ve parasal durumları da annelerinin bu bilgileriyle alakalı olduğu gözükmemektedir.



Şekil 4.25: 2008-2011 yılları, demografik verilere uygulanan J48 algoritmasının oluşturduğu BursKredi sınıfının ağaç yapısı

Kural	Birliktelik Kuralları	Güv.
1	BursKredi=1 AnneEgt=Ilkogretim (37) ==> AnneMes=EvHanımı (36)	0.97
2	TopEgtYıl=5 BursKredi=1 AnneEgt=Ilkogretim (35) ==> AnneMes=EvHanımı (34)	0.97
3	BursKredi=1 AnneEgt=Ilkogretim (37) ==> TopEgtYıl=5 (35)	0.95
4	BursKredi=1 AnneEgt=Ilkogretim AnneMes=EvHanımı (36) ==> TopEgtYıl=5 (34)	0.94
5	TopEgtYıl=5 AnneEgt=Ilkogretim (41) ==> AnneMes=EvHanımı (38)	0.93
6	AnneEgt=Ilkogretim (52) ==> AnneMes=EvHanımı (48)	0.92
7	BursKredi=1 AnneEgt=Ilkogretim (37) ==> TopEgtYıl=5 AnneMes=EvHanımı (34)	0.92
8	BursKredi=1 AnneMes=EvHanımı (68) ==> TopEgtYıl=5 (62)	0.91
9	Parasal=YeterliDegil (42) ==> AnneMes=EvHanımı (38)	0.9
10	TopEgtYıl=5 AnneMes=EvHanımı (69) ==> BursKredi=1 (62)	0.9

(a)

Kural	Birliktelik Kuralları	Doğ.
1	TopEgtYıl=5 BursKredi=1 GNO=5 (25) ==> AnneMes=EvHanımı (25)	0.99
2	Parasal=YeterliDegil AnneEgt=Ilkogretim (21) ==> AnneMes=EvHanımı (21)	0.99
3	TopEgtYıl=5 AdresIl=ICANADOLU BursKredi=1 (20) ==> AnneMes=EvHanımı (20)	0.99
4	TopEgtYıl=5 AnneEgt=Ortaogretim AnneMes=EvHanımı (19) ==> BursKredi=1 (19)	0.99
5	AnneEgt=Ilkogretim KalacakYer=KYK (18) ==> AnneMes=EvHanımı (18)	0.99
6	BursKredi=1 AnneMes=EvHanımı BabaMes=Memur (18) ==> TopEgtYıl=5 (18)	0.99
7	AdresIl=GUNEYDOGU (17) ==> AnneMes=EvHanımı (17)	0.99
8	BursKredi=1 KalacakYer=Evde Ark (17) ==> AnneMes=EvHanımı (17)	0.99
9	BursKredi=1 GNO=4 (17) ==> TopEgtYıl=5 (17)	0.99
10	Parasal=YeterliDegil GNO=5 (16) ==> AnneMes=EvHanımı (16)	0.98

(b)

Şekil 4.26: 2008-2011 yılları demografik verilerde elde edilen birliktelik kuralları a)

Apriori ile b) Predictive Apriori ile

4.4.3. Atılma politikasının öğrenci başarısı üzerine etkisi

Üniversitelerden atılmanın 2011 yılında kalkmasının öğrencilerin başarılarını ne yönde etkilediğinin anlaşılabilmesi için farklı bir veri seti oluşturulmuştur. Her bir öğrencinin, 2011 öncesi ve sonrasındaki not ortalamaları hesaplanmıştır. Girdiler öğrencinin 2011 öncesi ve 2011 sonrası not ortalamaları şeklinde belirlenmiştir. Tahmin edileceği üzere 2011 sonrası notları olmayan öğrenciler olacağı için, her öğrenci veri setine iki girdi sağlayamayacaktır. Çizelge 4.15'teki örnekte veri setinin ilk oluşturulma aşaması verilmiştir. Veri madenciliği aşamasına geçildiğinde *KayıtNo* özniteliği veri setinden çıkarılmıştır. Toplam öğrenim sürelerinde halen devam etmekte olan öğrenciler “-1” ifadesi ile gösterilmiş olup bu değere sahip olan girdiler veri setinden çıkarılmıştır.

Çizelge 4.15: 2011 yılı öncesi ve sonrası sınıflaması için oluşturulmuş veri seti örneği

<i>KayıtNo</i>	<i>GNO</i>	<i>NotOrt</i>	<i>TopEgtYıl</i>	<i>Zaman</i>
62	4	3	6	once
62	4	5	5	sonra
63	5	5	5	sonra
63	5	5	5	once
65	4	5	5	once
66	2	3	-1	sonra

Veri seti veri madenciliğine uygun bir biçime dönüştürüldüğünde 220 girdi ve 4 öznitelikten oluşmaktadır. Öznitelikler şunlardır:

- *GNO* : Öğrencilerin 9 seviyeye dönüştürülmüş genel not ortalamaları
- *NotOrt* : Öğrencilerin 2011 öncesindeki ve sonrasındaki not ortalamaları
- *TopEgtYıl* : Öğrencilerin yıl bazında toplam üniversite öğrenim süreleri

- *Zaman* : Girdinin sahip olduğu *önce*(2011 öncesi) ya da *sonra*(2011 sonrası) değerleri

Öncelikle probleme sınıflama problemi gibi yaklaşılmış ve *Zaman* özniteliği sınıf olarak belirlenmiştir ve kNN ve J48 algoritmaları kullanılmıştır. Sınıflamanın başarısına göre *önce* ya da *sonra* ayrımının sonuçları ortaya konulmak istenmiştir.

Sınıflama algoritmaları Çizelge 4.16'daki sonuçları elde etmiştir. İki algoritma da %60'a varan doğru sınıflama oranına sahip olsa da 2011 sonrası olan girdilerin hepsini 2011 öncesi şeklinde sınıflamıştır. 2011 öncesi girdi sayısı sonrasında daha yüksek olduğu için tanıma oranları yüksek çıkmış ama istenilen başarıyı sağlayamamıştır.

Çizelge 4.16: Algoritmaların zaman sınıfı için ürettiği sonuçlar

<i>Sınıf</i>	<i>Sınıflandırıcı</i>	<i>Parametre</i>	<i>TP Oranı</i>	<i>FP Oranı</i>	<i>Kesinlik</i>	<i>Duyarlılık</i>	<i>F-Ölçütü</i>	<i>Doğruluk</i>
Zaman	kNN	k=9,1/d	0.614	0.515	0.59	0.614	0.571	61.36
	J48	c=0.25	0.605	0.605	0.365	0.605	0.456	60.45

Apriori ve Predictive Apriori algoritmalarıyla özniteliklerin birbirleriyle olan ilişkileri değerlendirilmeye çalışılmıştır. Algoritmaların çıkardığı kurallar kayda değer bilgiler içermemektedir. Değerli sayılabilecek bilgilerin elde edilememesinin altında yatan neden, toplam öğrenim süresinin eklenmesiyle çıkarılan girdilerle yararlı olabilecek bir takım bilgilerin de çıkarılmış olmasıdır. Eğitimleri devam etmekte olan öğrencilerin girdileri tekrar veri setine eklenip, *TopEgtYıl* özniteliği veri setinden çıkarılmıştır.

Veri seti tekrar düzenlenerek sadece, *GNO*, *Not* ve *Zaman* özniteliklerini içerecek şekilde ayarlanmıştır. Veri seti bu haliyle 334 girdi ve 3 öznitelikten oluşmaktadır. Bu veri setinde çalıştırılan sınıflama algoritmalarının sonuçları Çizelge 4.17'de verilmiştir.

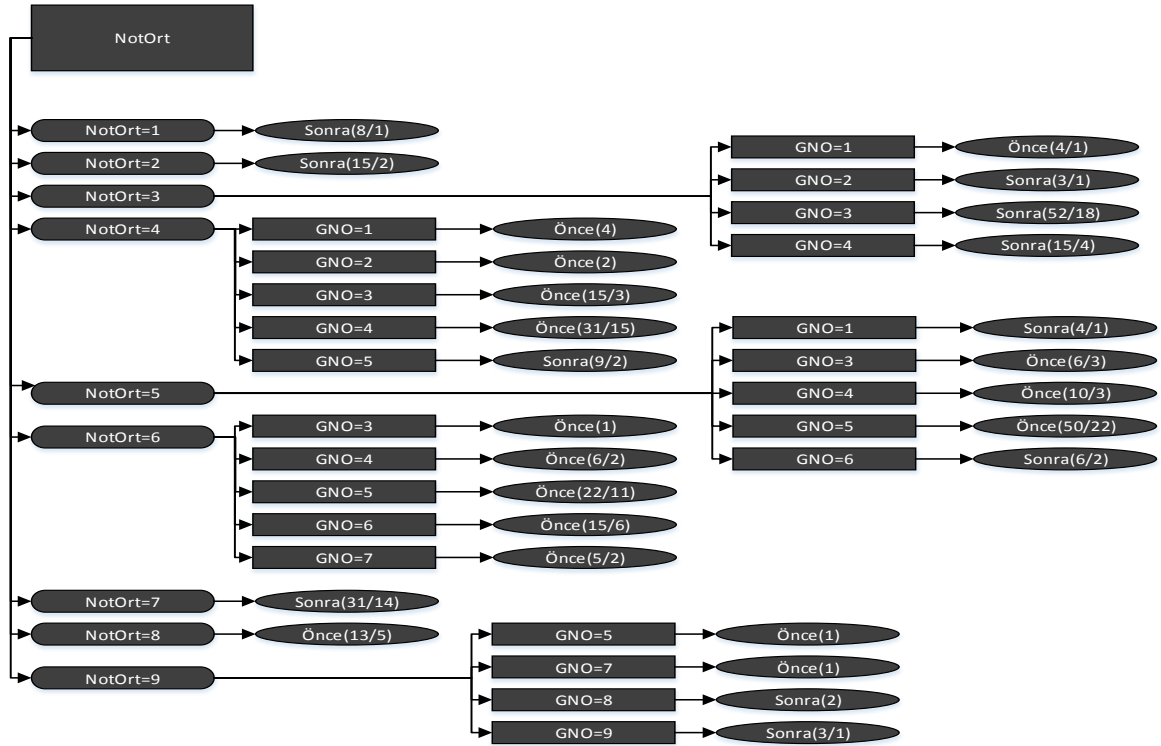
Çizelge 4.17: Algoritmaların üç öznitelikle zaman sınıfı için ürettiği sonuçlar

<i>Sınıf</i>	<i>Sınıflandırıcı</i>	<i>Parametre</i>	<i>TP Oranı</i>	<i>FP Oranı</i>	<i>Kesinlik</i>	<i>Duyarlılık</i>	<i>F-Ölçütü</i>	<i>Doğruluk</i>
Zaman	kNN	k=3,1/d	0.554	0.446	0.556	0.554	0.554	55.39
	J48	c=0.25	0.533	0.462	0.538	0.533	0.533	53.29

Algoritmaların ulaştığı sonuçlar dikkate alındığında, doğru sınıflamada yüzdeleri düşmüş olsa da *FP*(false positive) oranları da düştüğü için yanlış sınıflandırmada azalma gözükmemektedir. *FP* oranı girdinin yanlışlıkla bulunmaması gereken sınıfta bulunduğunu gösterdiği için oranın düşük olması, önceki veri setine göre gelişme olduğu anlamına gelmektedir. J48 algoritmasının başarısı kNN algoritmasından düşük olmasına rağmen Şekil 4.27’deki ağaç yapısını oluşturduğundan kullanışlı olabilecek bilgiler çıkarmıştır. Şekildeki ağaç yapısına bakılacak olursa *NotOrt* özniteliği 1, 2 ve 3 grubunda olanların çoğunluğunun 2011 sonrası sınıfına atandığı görülmektedir. Örneğin ağaç yapısındaki “*NotOrt*=2 => sonra(15/2)” gösterimi, notu 2 seviyesine giren 15 girdi 2011 sonrası şeklinde sınıflandırılmıştır ve 2 girdinin bu sınıfa yanlışlıkla atanmış olduğu anlamına gelmektedir.

Düşük not kategorilerine giren öğrencilerin 2011 sonrası olarak sınıflandırılması üniversitelerden atılmanın kalkmasıyla öğrenci başarısında düşüş olduğu şeklinde yorumlanabilir.

Apriori ve Predictive Apriori algoritmalarından elde edilen kurallar Şekil 4.28’de verilmektedir. Predictive Apriori algoritmasının ürettiği kurallara bakıldığında J48 ile elde edilen sonuçların desteklendiği gözükmemektedir. Beşinci kuraldaki notu 2 kategorisine giren 15 öğrenciden 13’ünün ve altıncı kuraldaki notu 1 kategorisine giren 8 öğrenciden 7’sinin 2011 sonrasına ait olduğu sonucu çıkmaktadır.



Şekil 4.27: J48 algoritmasının 2011 öncesi ve sonrası için çıkardığı ağaç yapısı

Kural	BirlikteK Kuralları	Güv.
1	GNO=3 NotOrt=3 (52) ==> Zaman=sonra (34)	0.65
2	NotOrt=3 (74) ==> Zaman=sonra (48)	0.65
3	GNO=3 (87) ==> Zaman=sonra (53)	0.61
4	NotOrt=4 (61) ==> Zaman=once (36)	0.59
5	NotOrt=5 (76) ==> Zaman=once (41)	0.54
6	GNO=5 (88) ==> Zaman=once (45)	0.51

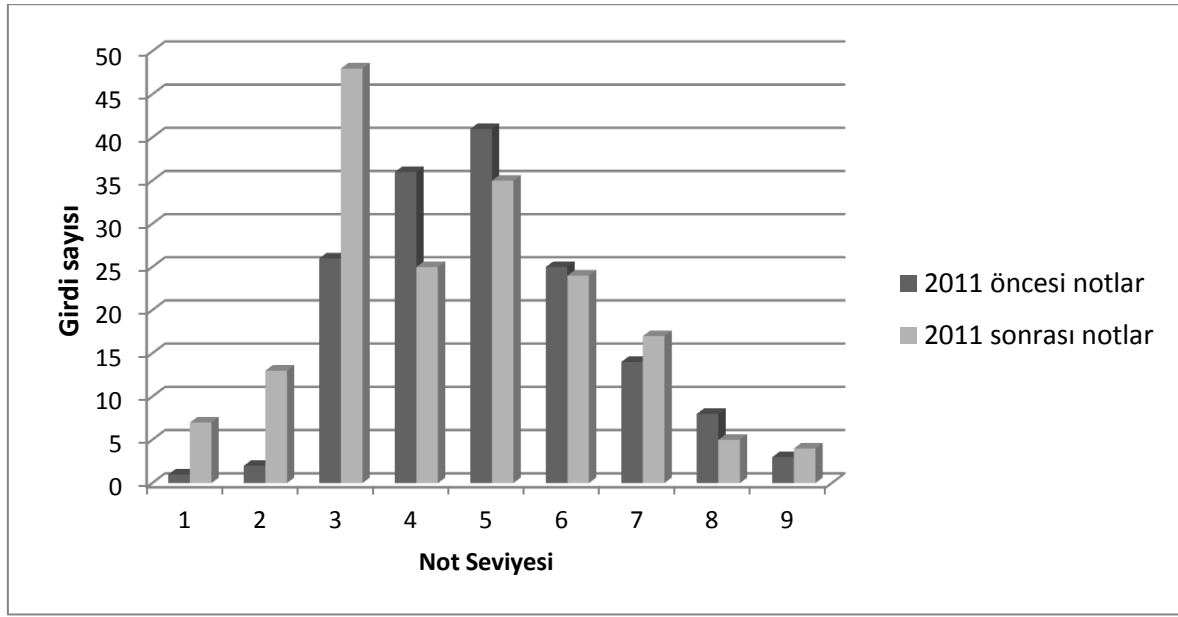
(a)

Kural	BirlikteK Kuralları	Doğ.
1	GNO=3 NotOrt=2 (8) ==> Zaman=sonra (8)	0.95
2	GNO=3 NotOrt=1 (5) ==> Zaman=sonra (5)	0.89
3	GNO=1 NotOrt=4 (4) ==> Zaman=once (4)	0.85
4	GNO=2 NotOrt=2 (3) ==> Zaman=sonra (3)	0.8
5	NotOrt=2 (15) ==> Zaman=sonra (13)	0.76
6	NotOrt=1 (8) ==> Zaman=sonra (7)	0.74

(b)

Şekil 4.28: 2011 yılı öncesi ve sonrası için elde edilen birliktelik kuralları a) Apriori ile
b) Predictive Apriori ile

2011 yılı öncesi ve sonrası not durumları Şekil 4.29'daki gibi görselleştirilmiştir. Şekilde de görüldüğü üzere orta ve üst seviye sayılabilecek not seviyeleri 2011 öncesi ve sonrası için birbirlerine yakındır. Üç ve aşağısındaki not seviyelerinde ise 2011 yılı sonrası notlar çoğunluktadır.



Şekil 4.29: 2011 öncesi, 2011 sonrası sınıflarının girdi sayısına göre dağılımı

5. SONUÇLAR

Bu tez kapsamında, Eskişehir Osmangazi Üniversitesi Bilgisayar Mühendisliği Bölümü öğrencilerinin verileri üzerinde veri madenciliği yöntemleri uygulanmakta ve öğrencilerin başarılarına etki eden unsurlar incelenmektedir. Tezin ana amacı 2011 yılında değişen atılma politikasının öğrenci başarısını ne yönde etkilediğini belirlemektir.

Veriler 2008,2009,2010 ve 2011 yıllarında Bilgisayar Mühendisliği Bölümü birinci sınıf derslerine başlayan öğrencilere aittir. Veriler üzerinde üç farklı yaklaşımda bulunmaktadır. Birinci yaklaşımda, öğrenci işleri bilgi sisteminde bulunan genel not ortalaması, eğitim süreleri, gibi eksi bilginin az olduğu veri setleri ile çalışılmaktadır. İkinci yaklaşımda, öğrencilerin doldurduğu anket ve formlardan edinilmiş bilgilerin çoğunlukta olduğu veri setleri kullanılmaktadır. Üçüncü yaklaşımda ise öğrencilerin üniversite eğitim süreleri boyunca aldıkları derslerin 2011 yılı öncesi ortalamaları ve 2011 yılı sonrası ortalamaları hesaplanmakta ve not ortalamaları eklenmesi ile oluşturulan veri seti veri madenciliği yöntemleri ile incelenmektedir.

Birinci yaklaşımda, burs veya kredi alan öğrencilerin, hazırlık sınıfı da dahil eğitim sürelerinin beş yıl, almayan öğrencilerin ise altı yıl olduğu bulunmuştur. Bu kısımda yapılan denemelerde burs veya kredi alınması ile eğitim süreleri arasında bağlantı olduğu gözükmemektedir. Apriori algoritması bu bağlantıya yönelik kurallar çıkarmıştır. Kurallardan birisinin “Burs veya kredi alanlar beş yılda eğitimlerini tamamlamaktadırlar.” şeklinde olduğu gözükmemektedir. Ayrıca, veri azaltmanın sonuçları nasıl etkilediği de incelenmektedir. Özniteliklerin çıkarılmasıyla oluşturulan veri seti, eksiksiz veri setiyle çoğu denemede benzer sonuçlar vermektedir.

İkinci yaklaşımda, anket verilerinin de dahil edilmesi sebebiyle bilgileri eksiksiz olan öğrenciler veri setine dahil edilmiş ve girdi sayısı yarıya düşmüştür. Oluşturulan veri setinde öğrencilerin annelerinin eğitim durumları, babalarının eğitim durumlarına göre daha iyi sınıflandırılmıştır. Öğrencilerin annelerinin eğitim durumlarının daha iyi sınıflandırılmasının altında yatan sebebin, çoğu öğrencinin annesinin eğitim seviyesinin ilk öğretim olmasından kaynaklandığı düşünülmektedir. Sınıflama algoritmalarının, azınlıkta kalan eğitim seviyelerini de ilk öğretim olarak tanımlamaya yatkın olduğu görülmektedir.

Annelerin çoğu ev hanımı olduğu için sınıflama algoritmaları, girdilerin çoğunu ev hanımı sınıfına atmaktadır. Yüksek çıkan doğru sınıflama oranı da bu yüzden yanıltıcı olmaktadır. Karşılaşılan bu gibi sorunlara rağmen 2010 yılında birinci sınıf derslerine başlayanların veri setinde elde edilen bazı birliktelik kuralları yararlı olabilecek bilgiler içermektedir. Parasal durumu sıkıntılı olan öğrencilerin çoğunun annelerinin ev hanımı olduğu ve burs veya kredi aldıkları gözükmemektedir.

Son yaklaşımda, üniversitelerden atılmanın kalkmasının öğrencilerin başarılarına etkileri belirlenmektedir. Sınıflama algoritmalarının doğru sınıflama oranı düşük çıkmasına rağmen J48 algoritması ürettiği ağaç yapısında, 2.4 ve aşağısında not ortalamasına (1,2 ve 3 kategorisine giren not gruplarının) sahip öğrencilerin çoğunluğunun 2011 yılı sonrasına ait olduğunu göstermektedir. Benzer sonuçlar Predictive Apriori algoritması ile de elde edilmektedir. Apriori algoritmasının oluşturduğu kurallarsa biraz daha yüzeysel kalmaktadır.

Sonuç olarak atılmanın kalkmasıyla alt seviyelerde sayılabilecek not ortalamasına sahip girdi sayısı artmıştır. Atılma politikasındaki bu değişikliğin, öğrencilerin bir kısmının başarısını kötü yönde etkilediği kanısına varılmaktadır. Risk grubunda sayılabilecek öğrencilerin, üniversitelerden atılma ihtimalleri ortadan kalkınca notlarının düşmesine kayıtsız kaldıkları düşünülebilir. Eğitimlerini uzatmaları hayata daha geç atılmalarına, kendilerine ve ailelerine maddi ve manevi açıdan da daha fazla yük oluşturmalarına sebep olabilir.

İleriye dönük çalışmalarda öncelikli olarak veri toplama kısmında daha az ön işleme yapmayı gerektirecek bir sistemde verilerin biriktirilmesi düşünülebilir. Aşama aşama veri setlerinin boyutları arttırılabilir. Örneğin bir sonraki çalışmada tüm mühendislik bölümlerinin öğrencilerinin verileri dahil edilebilir. Sonraki aşamada ülke geneli gibi genişleyen bir düzende ilerlenebilir. Geniş kapsamlı bir veri seti daha sağlıklı sonuçlar alınmasını kolaylaştıracaktır. Bu sayede algoritmaların sınıfları eşit şekilde öğrenmesi sağlanmış olacaktır. Veri setinde mümkünse her sınıf için eşit sayıda girdi kullanılabilir. Her sınıf için eşit sayıda girdi olması, algoritmaların sınıfları eşit şekilde öğrenmesini ve bulguların daha sağlıklı olmasını sağlayabilir.

KAYNAKLAR DİZİNİ

- Agrawal, R., Srikant, R., 1994, Fast algorithms for mining association rules, 20th Int. Conf. Very Large Data Bases, 1215, p. 487-499
- Altınışık, U., 2006, Öğrenci bilgi sisteminde veri madenciliğinin uygulanması, Kocaeli Üniversitesi
- Aydın, S., 2007, Veri madenciliği ve Anadolu Üniversitesi uzaktan eğitim sisteminde bir uygulama, Anadolu Üniversitesi
- Borne, K., 2009, Scientific data mining in astronomy, arXiv:0911.0505.
- Brossette, S. E., Sprague, A. P., Hardin, J. M., Waites, K. B., Jones, W. T., vd., 1998, Association rules and data mining in hospital infection control and public health surveillance, Journal of the American medical informatics association, 5, 4,373-381.
- Brown, D. E., Oxford, R. B., 2001, Data mining time series with applications to crime analysis, Systems Man and Cybernetics, 2001 IEEE International Conference, 3, p. 1453-1458
- Cooley, R., Mobasher, B., Srivastava, J. 1999, Data preparation for mining world wide web browsing patterns, Knowledge and information systems, 1,1, 5-32
- Ekim, U., 2011, Veri madenciliği algoritmalarını kullanarak öğrenci verilerinden birliktelik kurallarının çıkarılması, Selçuk Üniversitesi
- Erdoğan, Ş. Z., 2004, Veri madenciliği ve veri madenciliğinde kullanılan k-means algoritmasının öğrenci veri tabanında uygulanması, İstanbul Üniversitesi
- Fayyad, U, Piatotsky-Shapiro, G., Smyth, P., 1996, From data mining to knowledge discovery in databases, AI Magazine Volume 17 Number 3, 37-54

KAYNAKLAR DİZİNİ (devam ediyor)

- Guifang, G., Youshi, H., 2010, July, Research on the application of data mining to customer relationship management in the mobile communication industry, Computer Science and Information Technology, 2010 3rd IEEE International Conference ,2, p. 597-599
- Gülçe, G., 2010, Veri ambarı ve veri madenciliği teknikleri kullanılarak öğrenci karar destek sistemi oluşturma, Pamukkale Üniversitesi
- Güvenç, E., 2001, Yüksek öğrenimde öğrenci performansının veri madenciliği teknikleri ile belirlenmesi, Boğaziçi Üniversitesi,
- Han, J., Kamber, M., 2006, Data mining: concepts and techniques, Morgan Kaufmann San Francisco USA, p.
- Hormozi, A. M., Giles, S., 2004, Data mining: a competitive weapon for banking and retail industries, Information systems management, 21, 2 , 62-71
- İnan, O., 2003, Öğrenci işleri veri tabanı üzerinde veri madenciliği uygulamaları, Konya Selçuk Üniversitesi
- Kesavaraj, G., Sukumaran, S., 2013, A study on classification techniques in data mining, Computing Communications and Networking Technologies, 2013 Fourth International Conference, p. 1-7
- Khan, F. S., Anwer, R. M., Torgersson, O., Falkman, G., 2008, Data mining in oral medicine using decision trees, World Academy of Science, Engineering and Technology, 37, 225-230.
- Kumar, V., Chadha, A., 2012, Mining association rules in student's assessment data, International Journal of Computer Science Issues, 9, 5, 211-216.
- Ling, C. X., Li, C. ,1998, Data mining for direct marketing: problems and solutions, KDD, 98, 73-79

KAYNAKLAR DİZİNİ (devam ediyor)

- Liu, Y., 2010, Study on application of apriori algorithm in data mining, Computer Modeling and Simulation, Second International Conference, 3, p. 111-114
- Marquez-Vera, C., Morales, C. R., Soto, S. V., 2013, Predicting school failure and dropout by using data mining techniques, Tecnologias del Aprendizaje, IEEE Revista Iberoamericana de, 8, 1, 7-14.
- Oğuzlar, A., 2004, Veri madenciliğinde birliktelik kuralları, Marmara Üniversitesi Sosyal Bilimler Enstitüsü Hakemli Dergisi Öneri, 6, 22, s. 315-321
- Özarслан, S., Barışçı, N., 2014, Öğrenci performansının veri madenciliği ile belirlenmesi, ISITES'2014, Karabük Üniversitesi
- Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Hsu, M. C., 2004, Mining sequential patterns by pattern-growth: the prefixspan approach, Knowledge and Data Engineering, 16, 11, 1424-1440.
- Romero, C., Ventura, S., 2007, Educational data mining: a survey from 1995 to 2005, Expert Systems With Applications 33, 135-146.
- Saygılı, A., 2013, Veri madenciliği ile mühendislik fakültesi öğrencilerinin okul başarılarının analizi, Yıldız Teknik Üniversitesi
- Scheffer, T., 2005, Finding association rules that trade support optimally against confidence, Intelligent Data Analysis, 9, 4, p. 381-395.
- Shah, C., Jivani, A. G., 2013, Comparison of data mining classification algorithms for breast cancer prediction, Computing Communications and Networking Technologies, 2013 Fourth International Conference, p. 1-4
- Singh, S., Lal, S., P., 2013, Using feature selection and association rule mining to evaluate digital courseware, Eleventh International Conference on ICT and Knowledge Engineering, p. 7

KAYNAKLAR DİZİNİ (devam ediyor)

Talavera, L., Gaudioso, E., 2004, Mining student data to characterize similar behavior groups in unstructured collaboration spaces, 16th European conference on artificial intelligence, p. 17-23

Taşdemir, M., 2012, Veri madenciliği: öğrenci başarısına etki eden faktörlerin regresyon analizi ile tespiti, Dicle Üniversitesi

Tüzüntürk, S., 2010, Veri madenciliği ve istatistik, Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, XXIX, 1, s. 65-90

Witten, I. H., Frank, E., Hall, M. A., 2011, Data mining practical machine learning tools and techniques, Morgan Kaufmann Publishers is an imprint of Elsevier, 665 p.