PAMUKKALE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

VERİ AMBARI VE VERİ MADENCİLİĞİ TEKNİKLERİ KULLANILARAK ÖĞRENCİ KARAR DESTEK SİSTEMİ OLUŞTURMA

YÜKSEK LİSANS
Gürler GÜLÇE

Anabilim Dalı : Bilgisayar Mühendisliği

Programı : Tezli Yüksek Lisans

Tez Danışmanı: Yrd. Doç. Dr. Gürhan GÜNDÜZ

Ağustos 2010

YÜKSEK LİSANS TEZ ONAY FORMU

Pamukkale Üniversitesi Fen Bilimleri Enstitüsü 071281003 nolu öğrencisi Gürler Gülçe tarafından hazırlanan "**VERİ AMBARI VE VERİ MADENCİLİĞİ TEKNİKLERİ KULLANILARAK ÖĞRENCİ KARAR DESTEK SİSTEMİ OLUŞTURMA**" başlıklı tez tarafımızdan okunmuş, kapsamı ve niteliği açısından bir Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Danışmanı: (Jüri Başkanı)	**Yrd. Doç. Dr. Gürhan GÜNDÜZ (PAÜ)**	
Jüri Üyesi:	**Doç. Dr. Abdullah T.TOLA (PAÜ)**	
Jüri Üyesi:	**Yrd. Doç. Dr. Emre ÇOMAK(PAÜ)**	

Pamukkale Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 22.09.2010. tarih ve ..24.113..... sayılı kararıyla onaylanmıştır.

Fen Bilimleri Enstitüsü Müdürü
Prof. Dr. Halil KARAHAN

Bu tezin tasarımı, hazırlanması, yürütülmesi, araştırılmalarının yapılması ve bulgularının analizlerinde bilimsel etiğe ve akademik kurallara özenle riayet edildiğini; bu çalışmanın doğrudan birincil ürünü olmayan bulguların, verilerin ve materyallerin bilimsel etiğe uygun olarak kaynak gösterildiğini ve alıntı yapılan çalışmalara atfedildiğini beyan ederim.

İmza:

Öğrenci Adı Soyadı: Gürler GÜLÇE

ÖNSÖZ

Üniversitelere kabul edilen öğrenci sayısı her geçen yıl artmaktadır. Öğrenci sayısının artmasıyla üniversitelere daha az hazırlıklı öğrencilerin giriş sayısı da yükselmektedir. Üniversitelerin akademik desteğe ihtiyaç duyacak öğrencileri henüz başarısızlık durumları oluşmadan belirleyerek uygun eğitim ve rehberlik hizmetlerinin düzenlenmesi için gerekli adımların atılması, hem kişilerin meslek hayatlarında daha başarılı bireyler olmalarında yardımcı olacak, hem de üniversite kalitesinin yükseltilmede büyük yarar sağlayacaktır.

Bu nedenle eldeki çalışmada üniversite öğrencilerinin akademik başarı durumlarının önceden tahmin edilerek, elde edilen bilgilerin ilgili kişilerin onayına sunulması ve bu sayede gelecekte alınacak eğitim ve öğretim plan ve programlama sürecine yönelik kararlarda yönetime destek olunması hedeflenmiştir. Bu amaçla, ilk olarak Pamukkale Üniversitesi öğrenci verileri üzerinde bir Veri Ambarı oluşturulmuştur. Tahmin işlemlerini yapabilmek için Veri Madenciliği algoritmaları olan karar ağaçları, yapay sinir ağları, naive bayes ve birliktelik kurallarından faydalanılmıştır. Bu algoritmaların sonuçlarına göre bir raporlama arayüzünden elde edilen sonuçlar karar vericilerin kullanımına sunulmuştur.

Bu tezin hazırlanması sürecinde göstermiş olduğu yardım ve anlayıştan dolayı saygı değer danışmanım Yrd. Doç. Dr. Gürhan GÜNDÜZ'E teşekkürlerimi sunarım. Aynı zamanda sevgili jüri üyelerim Doç. Dr. Abdullah TOLA ve Yrd..Doç. Dr. Emre ÇOMAK'A tezin iyileştirilmesi konusundaki değerli geri bildirimleri için teşekkür ederim. Son olarak gerek bir dost olarak gerek akademik olarak ihtiyacım olduğumda hep yanımda olan sevgili arkadaşlarım Özcan Dülger, Devrim İşli ve Murat Demir'e teşekkür ederim.

Ağustos 2010

Gürler Gülçe

(Bilgisayar Mühendisi)

İÇİNDEKİLER

	Sayfa No
ÖZET	**x**
SUMMARY	**xi**
1. GİRİŞ	**1**
2. VERİ AMBARI	**4**
2.1 Giriş	4
2.2 Veri Ambarı ile OLTP Sistemleri Arasındaki Farklar	6
2.3 Veri Ambarı Mimarisi	8
2.3.1 Kurumsal veri ambarı	8
2.3.2 Veri pazarı (Data Mart-Küçük Veri Ambarı)	8
2.3.3 Sanal veri ambarı	9
2.4 Karar Destek Sistemleri	10
3. VERİ MADENCİLİĞİ	**13**
3.1 Giriş	13
3.2 Veri Madenciliği Süreci	14
3.2.1 Veri temizleme	15
3.2.2 Veri bütünleştirme	16
3.2.3 Veri indirgeme	17
3.2.4 Veri dönüştürme	17
3.2.4.1 Min-Max normalleştirilmesi	17
3.2.4.2 Z-score standartlaştırma	18
3.2.5 Veri madenciliği algoritmasını uygulama	18
3.2.6 Sonuçları sunum ve değerlendirme	18
3.3 Veri Madenciliği Modelleri	19
3.3.1 Sınıflama ve regresyon	20
3. 3. 1. 1 Yapay sinir ağları	21
3. 3. 1. 2 Genetik algoritmalar	22
3. 3. 1. 3 K – En yakın komşu algoritması	23
3. 3. 1. 4 Karar ağaçları	23
3. 3. 1. 5 Bayes sınıflandırıcıları	24
3.3.2. Kümeleme modelleri	25
3.3.3 Birliktelik kuralları ve ardışık zamanlı örüntüler	26
3.4 Çalışmada Kullanılan Veri Madenciliği Modelleri	27
3.4.1 Microsoft Naive Bayes	28
3.4.2 Microsoft Decision Trees (Microsoft Karar Ağaçları)	29
3.4.3 Microsoft Association Rules (Microsoft Birliktelik Kuralları)	30
3.4.4 Microsoft Neural Network (Microsoft Sinir Ağları)	32
3.5 DMX (Data Mining Extensions)	34
3.6 Veri Madenciliğinin Eğitim Alanında Kullanımı	35

4. TEZ KAPSAMINDA KULLANILAN ÜRÜNLER **38**
4.1 Giriş 38
4.2 Microsoft Visual Studio 2008 39
4.3 Microsoft SQL Server 2008 39
4.4 Microsoft Analysis Services 2008 39
4.5 Microsoft Reporting Services 2008 40
5. ÖĞRENCİ KARAR DESTEK SİSTEMİ **41**
5.1 Giriş 41
5.2 Öğrenci İşleri Otomasyon Sistemi Yapısı 42
5.3 Veri Ambarı Oluşturma 44
5.4 Veri Madenciliği Modeli Oluşturma ve Algoritmasını Uygulama 46
5.4.1 Program ve cinsiyet'e göre akademik başarı tahmini 50
5.4.2 Öğrenci kimlik bilgilerine göre akademik başarı tahmini 66
5.4.3 Öğrenci kimlik bilgileri ve anket sonuçlarına göre akademik başarı tahmini 71
5.4.4 Aile eğitim ve gelir durumuna göre akademik başarı tahmini 78
5.4.5 Öğrencilerin ders başarılarının tahmin edilmesi 84
6. SONUÇLAR VE DEĞERLENDİRME **90**
KAYNAKLAR **93**

KISALTMALAR DİZİNİ

AB **:**Akademik Başarı

BKS **:**Bilgi Keşfi Süreci

CART **:**Classification and Regression Trees

DMX **:**Data Mining Extensions

DSS **:**Decision Support System

DSV **:**Data Source View

ETL **:**Extract Transform Load

KDD **:**Knowledge Discovery in Database

KDS **:**Karar Destek Sistemi

MSA **:**Microsoft Sinir Ağları

OİOS **:**Öğrenci İşleri Otomasyon Sistemi

OLAP **:**Online Analytical Processing

OLTP **:**On-Line Transaction Processing

SQL **:**Structured Query Language

TABLO LİSTESİ

Tablolar

2.1 :	İşletimsel sistemlerle veri ambarının karşılaştırılması	7
3.1 :	Microsoft Naive Bayes eğitim verileri	28
3.2 :	Microsoft Naive Bayes test verisi	29
3.3 :	Microsoft Karar Ağaçları eğitim verileri	30
3.4 :	Birliktelik kurallarında kullanılabilecek veri kümesi	31
5.1 :	Program ve cinsiyete göre AB tahmin modelinin giriş parametreleri	50
5.2 :	Program cinsiyet modelinin tahmin oranları	62
5.3 :	Cinsiyete göre başarı	65
5.4 :	Öğrenci kimlik bilgilerine göre AB tahmin modelinde giriş ve çıkış parametreleri	66
5.5 :	Akademik başarı modelinin tahmin oranları	70
5.6 :	Anket verileri destekli AB belirlemede kullanılan parametreler	72
5.7 :	Anket verilerine göre AB modelinin tahmin oranları	77
5.8 :	Aile gelir ve eğitim durumuna göre AB belirlemede kullanılan parametreler	79
5.9 :	Aile gelir ve eğitim modeli algoritmalarının uygulanan test verileri karşısında başarısı	83
5.10 :	Öğrenci ders başarıları modelinde kullanılan parametreler	85
5.11 :	Notlar ve anlamları	85
5.12 :	Öğrencilerin ders notları arasında bulunan bağıntılar	86

ŞEKİL LİSTESİ

Şekiller

2.1 :	Veri Ambarı Yapısı ve Kullanım Alanları	6
2.2 :	Karar Destek Sisteminin Bileşenleri	12
3.1 :	Veri Madenciliği Süreci	15
3.2 :	Farklı veri kaynaklarından alınan türlerin veri bütünleştirme işlemi	16
3.3 :	Veri Madenciliği Modelleri	20
3.4 :	Eğitim verilerine uygun karar ağacı	24
3.5 :	Örnek Nöron Gösterimi	33
5.1 :	Yeni kayıt olan öğrencilerle ilgili diyagram	44
5.2 :	Öğrencilerin aldığı tüm dersleri gösteren diyagram	45
5.3 :	Analiz hizmetleri servisinden veri ambarına bağlantı kurma	46
5.4 :	"Data Source View" kurma	47
5.5 :	İsimli sorgu tanımlama	48
5.6 :	En iyi algoritmayı bulma şeması	49
5.7 :	Veri Madenciliği algoritmasını seçme	51
5.8 :	Data Source View seçme	51
5.9 :	Modelde kullanılacak tabloyu belirleme işlemi	52
5.10 :	Eğitilecek verinin belirlenmesi	53
5.11 :	Kolon dönüştürme işlemi	54
5.12 :	Eğitim veri sayısı belirleme	55
5.13 :	Veri madenciliği modeli ekleme	56
5.14 :	Algoritmaların eğitilme işlemi	56
5.15 :	Program ve cinsiyet verilerine göre karar ağaçları modelinde oluşan sonuçlar	57
5.16 :	Fizyoterapi ve Rehabilitasyon programı kız öğrencilerinin karar ağacıyla başarı durumları	57
5.17 :	Rehberlik ve Psikolojik Danışmanlık programı erkek öğrencilerinin karar ağacıyla başarı durumları	58
5.18 :	Program ve cinsiyete göre veri madenciliği algoritmalarının “Başarısız” durumu ile ilgili değerleri	59
5.19 :	Program ve cinsiyete göre veri madenciliği algoritmalarının “Başarılı” durumu ile ilgili değerleri	60
5.20 :	Program ve cinsiyete göre veri madenciliği algoritmalarının “Başarısız” durumu ile ilgili değerleri	61
5.21 :	Program ve cinsiyet verilerine göre AB Tahmin Raporları	64
5.22 :	Giriş değerlerine göre AB Tahmin işlemi	65
5.23 :	Öğrenci Bilgilerine Göre AB Tahmin yönteminin “Başarısız” durumda lift chart değerleri	67
5.24 :	Öğrenci Bilgilerine Göre AB Tahmin yönteminin “Başarılı” durumda lift chart değerleri	68
5.25 :	Öğrenci Bilgilerine Göre AB Tahmin yönteminin “Çok Başarılı” durumdaki Lift chart değerleri	69

5.26 : Öğrenci bilgilerine göre AB tahmin raporu 71
5.27 : Öğrenci ve Anket Bilgilerine Göre AB tahmin yönteminin "Başarısız" durumundaki lift chart değerleri 74
5.28 : Öğrenci ve anket bilgilerine göre AB tahmin yönteminin "Başarılı" durumundaki lift chart değerleri 75
5.29 : Öğrenci ve anket bilgilerine göre AB tahmin yönteminin "Çok Başarılı" durumundaki lift chart değerleri 76
5.30 : Öğrenci ve anket bilgilerine göre akademik öğrencilerin akademik başarı tahmin raporları 78
5.31 : Aile eğitim ve gelir durumlarına göre AB tahmin yönteminin "Başarısız" değerleri aramadaki durumu 80
5.32 : Aile eğitim ve gelir durumlarına göre AB tahmin yönteminin "Başarılı" değerleri aramadaki durumu 81
5.33 : Aile eğitim ve gelir durumlarına göre akademik başarı tahmin yönteminin "Çok Başarılı" değerleri aramadaki durumu 82
5.34 : Aile Eğitim ve Gelir Durumlarına Göre Öğrencilerin Akademik Başarı Tahmin Raporu 84
5.35 : Dersler ve notlar arasındaki bağıntılar 87
5.36 : Dersler ve ders notları tahmin raporu 89

ÖZET

VERİ AMBARI VE VERİ MADENCİLİĞİ TEKNİKLERİ KULLANILARAK ÖĞRENCİ KARAR DESTEK SİSTEMİ OLUŞTURMA

Ülkemizde gerek artan genç nüfusa bağlı olarak gerekse üniversite eğitiminin çağdaş bir yaşam için gerekli iş olanağı ve bilişsel ve kişisel gelişimin önemli bir aracı olarak algılanmaya başlamasının bir sonucu olarak her yıl yüz binlerce öğrenci üniversiteye girmek amacıyla zorlu bir yarışa tabi tutulmaktadır. Buna karşılık öğrencilerin ancak sınırlı bir bölümü üniversiteye devam etme hakkı kazanmaktadır. Gittikçe artan bu talebi karşılamak amacıyla son yıllarda ülkenin dört bir köşesinde yeni üniversiteler açılmakta ve daha fazla öğrenci yüksek öğretimde eğitim alma olanağı elde etmektedir. Üniversite sayısındaki artışa paralel olarak üniversitede okumaya hak kazanan öğrenci sayısındaki hızlı artışa karşılık üniversitelerde öğrencilere sunulan akademik olanakların kalitesinde ve kaliteli bir eğitimin en önemli aktörlerinden birisi olan öğretim elemanının sayısında aynı oranda bir artış olmamaktadır. Bu da, öğrencilerin daha kalabalık sınıflarda öğretim elemanlarından daha az akademik ve sosyal destek alabilecekleri bir eğitim süreci anlamına gelmektedir. Daha da önemlisi, üniversiteye daha fazla öğrencinin kabulü daha düşük puanlarla ve akademik olarak daha az donanımlı öğrencilerin üniversite yaşamına ayak atmaya başlamaları anlamına gelmektedir. Akademik olarak daha az donanımlı bu öğrencilerin daha fazla desteğe ihtiyaç duymalarına karşılık gittikçe kalabalıklaşan sınıflarda daha az kişisel ilgi görebilecekleri ve böylece eğitim ortamlarında başarısız olma ihtimallerinin arttığı söylenebilir. Bu yüzden üniversitede akademik başarı açısından dezavantajlı öğrencilerin daha eğitim yaşantılarının ilk yıllarında tespit edilerek uygun akademik ve sosyal rehberlik programları aracılığıyla desteklenmesi yüksek öğretimde kalitenin artırılması ve hem kişisel hem ülke ekonomisi açısından gereksiz kayıpların önlenmesi açısından büyük önem taşımaktadır.

Son yıllarda bireyin davranışlarını tahmin etmek amacıyla sıklıkla kullanılan bir araç veri madenciliğidir. Bu amaçla bu çalışmada öğrencilerin akademik başarılarını belirleyen faktörlerin tespit edilerek başarısızlık riski taşıyan öğrencilerin belirlenmesi amacıyla çeşitli veri madenciliği modelleri kullanılmıştır. Eldeki verilerden üniversiteye devam etmekte olan öğrencilere ait bilgilere yönelik özel analizler yapılarak ileride idari anlamda alınacak kararlarda bir yol gösterici olarak rol oynaması amaçlanmıştır.

SUMMARY

DEVELOPING DECISION SUPPORT SYSTEM FOR STUDENT INFORMATION SYSTEM BY USING DATA WAREHOUSE AND DATA MINING TECHNICS

In our country, every year literally thousands of high school graduates compete to get an opportunity to continue their education through higher education system partially because the number of younger cohort is increasing and partially because there is an increasing awareness of the importance of higher education system both for obtaining a prestigious job and for personal development. In spite of this demand, only a small number of high school students could get into the higher education system. In order to close the gap between this demand and existing situation, many new universities have been opened all over the country. In line with the increase at the number of universities, more and more students have begun to get opportunity to continue their education through higher education system. However, the degree of the increase of the resources provided to the students along with the increase of the number of instructors is much slower than the degree of increase of the number of the students to be served. This means more crowded classes with less personalized instruction where the instructors could allocate less time and energy for each of the students. More importantly, because more universities are welcoming more students, their selection criteria are getting less strict with more students with lower exam scores could be accepted to the higher education system. This means an increase at the number of students who might be at risk of failing since they are less likely to be academically prepared. Combination of less personalized instruction within more crowded classroom with more academically disadvantaged students means more risk for failing during the university education process. For this reason, it is crucial to determine the academically disadvantaged students from the first year of their education life to provide appropriate educational support and experience to prevent both personal and national wide economical loss.

Data mining seems to be one of the most widely used tool to predict individuals' behavior patterns given some predetermined variables. Therefore, to predict the students who might be at risk of failing by determining the factors predicting students' achievement levels, data mining techniques is used in this thesis. By applying special analyses to the data obtained from attending students, it was aimed to provide information to guide management decisions.

1. GİRİŞ

Yaşadığımız zamanda bilişim teknolojileri çok hızlı bir şekilde gelişmektedir. Bilgisayar yazılım ve donanım alanlarında devamlı gelişen yeniliklerle karşılaşılmaktadır. Bu hızlı gelişimin önemli noktalarından birisi söz konusu teknoloji tabanlı ürünlere erişimin kolaylaşması, fiyatların ucuzlamasıdır. Şirketler ve kullanıcılar daha hızlı ve kullanışlı bilgisayar teknolojilerine kolayca sahip olabilmektedir.

Kolayca elde edilen bilgi teknolojileri yardımıyla artık her veri sayısal ortamlara kaydedilebilir duruma gelmiştir. Örneğin bir üniversitede öğrenci ve ders kayıtları ile ilgili tüm bilgiler bilgisayar ortamına kaydedilmektedir. Bu işlem binlerce kullanıcı olan otomasyonlarda çok sayıda veri saklanması anlamına gelir. Gelişen donanım teknolojileri sayesinde bu bilgileri kaydetmek maliyetler açısında yeterli olabilir. Ama asıl cevaplanması gereken soru bu mevcut değerlerden işe yarayan bilgilerin nasıl çıkarılacağıdır. Firmaların otomasyon sistemlerini kurduktan sonra cevap aradıkları önemli bir soru olan bu problemin cevabı ise firmaların çeşitli çözümleme işlemleri yaparak bu devasa veri yığınlarını kullanmak amacıyla karar destek sistemleri oluşturulması yoluyla gelecekte alınacak kararlar hakkında destek sağlanabileceği şeklindedir.

Veriler üzerinde çözümleme yapmak için çeşitli istatistiksel yöntemler kullanılabilir. Ama çok fazla veri içeren durumlarda işlem yapmak çok zorlaşacaktır. Otomasyonlarda kullanılan veritabanları OLTP(Online Transaction Processing, Çevrim İçi Hareket İşleme) sistemi kullanır. Bu yapılarda sistem alt yapısı veri ekleme-güncelleme-silme işlemleri için tasarlandığından ilişkisel veritabanı yapısını kullanırlar. Verilerin parçalı olarak bulunması çözümleme yapma işlemlerini oldukça zorlaştıracaktır. Bu nedenlerden dolayı, veriyi işlenmeye hazır hale getirmek için “Veri Ambarı”, verileri

çözümleyerek “Bilgi”ye dönüştürmek için de “Veri Madenciliği” teknolojileri ortaya çıkmıştır.

Bu çalışmanın temel amacı öğrencilerin kişisel bilgilerini kullanarak akademik başarısını önceden tahmin edebilmektir. Üniversite sınavına her yıl yüz binlerce kişi katılmakta ve üniversite kontenjanları bu talebi karşılamak üzere devamlı artırılmakta, daha düşük puan ve hazırlık düzeyine sahip öğrencilerin üniversiteye kabulleri artmaktadır. Bu öğrencilerin akademik başarı durumları önceden tahmin edilmeye çalışılarak, yardıma ihtiyaç duyacak öğrencilerin önceden belirlenmesi ve buna göre önlem alınması hedeflenmektedir. Ayrıca hangi programlara devam eden öğrencilerin başarısızlıkla daha çok karşılaştığı belirlenerek ilgili birimlerde bu iyileştirme yönünde düzenlemeler yapılabilir.

Akademik başarı tahminine ilave olarak, akademik başarıyı oluşturan derslerin başarısı hakkında da bilgili olmak gereklidir. Üniversitede birçok ders birbiriyle ilişkili olarak okutulmaktadır. Bu derslerden geçen veya kalan öğrencilerin hangi notları aldığı, hangi dersler arasında önceden tahmin edilemeyen ilişkilerin olduğunu belirlemek bu tezde irdelenen başka bir konudur. Bu işlemin gerçekleştirilmesiyle, öğrencilerin aldıkları ders ve dönem sonu notlarına göre ileride alacakları dersler hakkında bir uyarı mekanizması gerçekleştirilebilir. Bu sayede öğrenciler sistemin kendilerine önerdikleri konulara ekstra dikkat göstererek başarılı olma durumlarını iyileştirebilirler.

Çalışmada Pamukkale Üniversitesi Öğrenci İşleriyle ilgili veriler kullanılmıştır. Bu sistemde öğrenci kayıtlarıyla ilgili birçok bilgi bulunmaktadır. Ama veriler daha önce bu yönde bir çalışma için hiç kullanılmamıştır. Projede Öğrenci İşleri Otomasyon Sistemi(ÖİOS) üzerinde analiz işlemlerinin yapılacağı Veri Ambarı oluşturulmuş, bunun üzerinde tahmin işlemlerinin gerçekleştirileceği Veri Madenciliği algoritmaları denenmiştir. Elde edilen sonuçlar bir raporlama ara yüzünden karar verici kişilerin bilgisine sunulmuştur. Bu sayede üniversite yönetimine hizmet verebilecek bir Karar Destek Sistemi (KDS) hazırlanmıştır.

Tez çalışması altı bölümden oluşmaktadır. Birinci bölümde tez çalışması hakkında genel bilgi verilmiş, çalışmanın amaçlarından ve yapılacaklardan bahsedilmiştir. İkinci bölümde Veri Ambarları hakkında genel bilgiler verilmiş ve KDS’lerin tanımından bahsedilmiştir. Üçüncü bölümde veri madenciliğini gerçekleştirmek için hangi süreçlerden geçilmesi gerektiği anlatılmış ve Literatürde bulunan Veri Madenciliği

türlerinden bahsedilmiştir. Ayrıca bu bölümde çalışmada kullanılan algoritmalarla ilgili teknik bilgiler verilmiş ve eğitim konusunda yapılan Veri Madenciliği çalışmalarından bahsedilmiştir. Dördüncü bölümde tezde kullanılan bilgisayar programları hakkında bilgi verilmiştir. Beşinci bölümde Karar Destek Sisteminin gerçekleştirilme aşaması anlatılmış, tahmin işleminin gerçekleştirildiği beş farklı model üzerinde, Veri Madenciliği algoritmaları uygulanmış ve raporlama çalışmaları yapılmıştır. Son olarak altıncı bölümde ise uygulamanın sonuçları ve yorumları üzerinde durulmuştur.

2. VERİ AMBARI

2.1 Giriş

Bilgi teknolojilerinde yaşanan hızlı gelişmeler sayesinde bilgi sistemlerinde depolanan veriler çok büyük boyutlara ulaşmıştır. Sistemlerdeki otomasyonlar farklı organizasyon yapıları ve işlevleri nedeniyle verilerini birden çok veri kaynağına kaydetmek zorunda kalabilirler. Bu durumdaki veritabanlarında yüzlerce kullanıcının aynı anda işlem yapması sebebiyle OLTP yapısı kullanılır. Bu sistemlerde dağıtılmış tablolardan çok sayıda verinin Karar Destek Sistemlerine aktarılmasının güçlükleri nedeniyle veri ambarları ortaya çıkmıştır.

Veri Ambarları, birbiriyle ilişkisi olmayan veri kaynaklarından aldığı verileri birleştirip, bunları karar destek uygulamalarında kullanılmak üzerine oluşturulan çok boyutlu gösterim işlemidir[1].

Geleneksel veritabanı sistemleri, kullanıcı hareketlerine bağlı günlük işlemleri desteklemek için tasarlanmıştır ve bu sistemler işletimsel ya da hareketsel (operational / transactional) sistemler olarak adlandırılır. İşletimsel sistemler hareket ya da işlem yönlendirmeli, veri ambarları ise konu yönlendirmelidir. Örneğin, yüksek seviyeli ticari verinin esnek analitik işlenmesini sağlar. Veri ambarı, veritabanı hareketinden çok sorgulama ve analiz için kullanılmak üzere tasarlanmış ilişkisel bir veritabanıdır. Genelde hareket verisinden elde edilmiş tarihsel bilgileri içerdiği gibi başka dış kaynaklardan gelen verileri de içerebilir. Veri tabanı hareketlerinden kaynaklanan iş yüküyle analiz yükünü birbirinden ayırır. Bu sayede değişik kaynaklardan toplanan verilerin daha kolay bir şekilde organize edilmesini sağlar[2].

Veri ambarının en önemli özelliği yüklenen verinin salt okunur ve değişken olmayan yapıda olmasıdır. İşletimsel veri genellikle gerçek zamanlıdır fakat veri ambarında tarihsel veriler mevcuttur. Veri ambarında yüksek seviyeli tarihsel veri genellikle gelecekte ne olabileceğini tahmin etmek için raporlama ve analiz amacıyla kullanılır[2]. İşin amaç ve hedefine uygun bir yapıda tasarlanır.

Veri ambarının temel kavramlarından birisi veri ile bilgi arasındaki farkı anlamaktır. Veri daha çok çalışan ve hareket ifa eden sistemde gözlenebilir ve kaydedilebilir durum toplamıdır ve veriler son kullanıcıya ancak düzenlenip bilgi durumunda sunulduğunda yarar sağlayabilir. Bilgi de, birçok konunun birleştirilmesi ve bunun da karar destek işleminde temel olarak kullanılmasıdır[3].

Veri ambarı, özneye dayalı, bütünleşmiş, zaman dilimli ve yöneticinin karar verme işleminde yardımcı olacak biçimde toplanmış olan değişmeyen veriler topluluğudur[4]. Veri ambarında yer alan veri özelliklerini inceleyecek olursak:

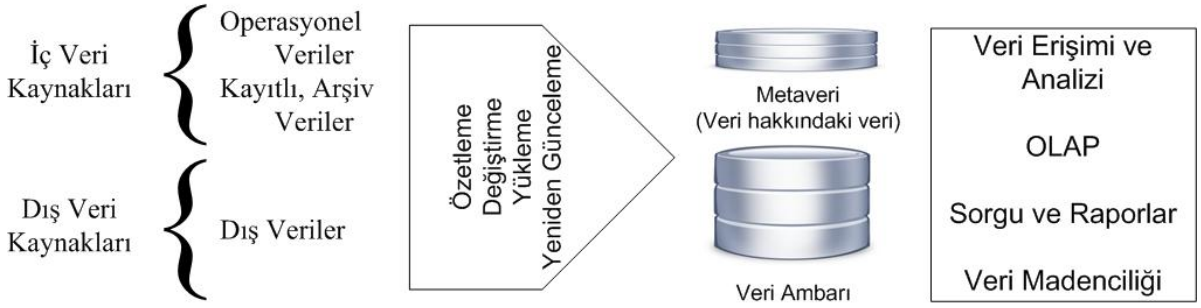
Konu Odaklı: Veri Ambarları, kullanıcının kendi verilerini analiz etmesine yardımcı olmak üzere tasarlanmıştır[5]. Yani, kullanılmayacak veriyi hariç tutar ve verinin basitleştirilmesini sağlar[6].

Bütünleşik: Veri Ambarı farklı birçok veri kaynağından verileri toplayıp, düzgün ve anlaşılabilir bir formatta birleştirmek zorundadır. Farklı veri kaynaklarından(farklı veri tabanı sistemleri, Excel ve text dosyaları) alınan veri tipleri ve ölçüm birimleri cinsinden aynı olmak zorundadır[7]. Bu işlemler ETL(Extract Transform Load) araçları kullanılarak yapılır[6].

Değişken Olmayan: Veri Ambarının bir diğer özelliği onun değişken olmayışıdır. Bir değer Veri Ambarına girdiği zaman bir daha değişemez. Çünkü Veri Ambarları sadece iki tür işlem gerçekleştirebilir; verinin yüklenmesi ve veri erişimi[5]. Veri ambarı bu nedenle işletimsel (operational) sistemler farklı olarak hareket işleme (transaction), geri alma (roolback) ve aynı zamanda işleme (concurrent processing) gibi işlemleri içermez[6].

Zaman Değişkenli: Veri Ambarının içindekiler, zamana bağlı olarak veri akışını temsil etmektedir. Veri Ambarının içindeki veri, erişimin herhangi bir zamanında doğrudur ve geniş bir zaman aralığında bulunmaktadır. Zaman değişkeninin asıl anlamı ise Veri Ambarının veri yapıları içinde anahtar bir bileşen gibi zamanın içsel olarak hazır bulunmasıdır[5].

Şekil 2.1'de Veri Ambarının yapısı ve hangi amaçlar için kullanılabileceği gösterilmiştir.



Şekil 2.1: Veri ambarı yapısı ve kullanım alanları[8]

Sonuç olarak veri ambarları veriye erişim türüne alternatif olmaktadır. Klasik veritabanı sistemlerinden farklı olarak SQL(Structured Query Language) sorgularının veritabanlarına gönderilmesi ve gelen sonucun düzenlenmesi yerine, istenen verinin daha önceden ilgili kaynaklardan çekilip, düzenlenmesi ve artık hazır durumda bulunan veri üzerinden gelen sorgulara cevap verilmesi amaçlanır. Bu işlemlerin yapılarak önceden sistemin hazır tutulması hem veri çekim işlemlerinde büyük performans artışlarını sağlamakta, hem de normal veritabanlarındaki iş yükünü de azaltarak diğer kullanıcılara rahat bir kullanım ortamı sunmaktadır[6].

2.2 Veri Ambarı ile OLTP Sistemleri Arasındaki Farklar

Veritabanı en genel tanımıyla, kullanım amacına uygun olarak düzenlenmiş veriler topluluğudur. Birbirleriyle ilişkileri olan verilerin tutulduğu, mantıksal ve fiziksel olarak tanımlarının olduğu bilgi depolarıdır. Veritabanları gerçekte var olan ve birbirleriyle ilişkisi olan nesneleri ve ilişkileri modeller. Kullanılan veritabanı modelleri ikiye ayrılmaktadır; İşletimsel veritabanları(OLTP) ve Veri Ambarlarıdır[9].

Veri ambarı ortamında, OLTP sistemi, organizasyonun işlemsel bir sistemdeki günlük işlerin uygulandığı bir veri tabanını belirtir. Verinin statik halinden ziyade, çoğu zaman eldeki güncel veriyi belirtir. Bu veritabanları ekleme, silme, güncelleme gibi hareket(transaction) içeren işlemlerin kullanımı için optimize edilmişlerdir. Örneğin öğretim elemanının öğrenci notunu sisteme işlediği anda, öğrencilerin notlarını aynı sistemde

nerdeyse eş zamanlı olarak görmesi OLTP sistemlerine bir örnektir. Genel anlamda alım-satım, hesaplama, üretim, bankacılık, eğitim gibi birçok işletmenin yüzlerce kullanıcısının bağlı olduğu ve yoğun hareket içeren otomasyon sistemleri OLTP yapısını kullanır. Veri ambarları ise, büyük miktarda ve zamana bağlı verileri depolamak için kullanılırlar. Kullanılan veri türleri statiktir.

OLTP ve Veri Ambarı arasında bulunan farklar Tablo 2.1'de özetlenmiştir.

Tablo 2.1: İşletimsel sistemlerle veri ambarının karşılaştırılması[2]

İşletimsel Sistemler	**Veri Ambarı Sistemleri**
Az seviyede raporlama içeren ve yüksek seviyeli hareket isleme amaçlı tasarlanmış sistemlerdir.(OLTP)	Analiz ve raporlama amaçlı tasarlanmış ve analitik işlemleri (OLAP) destekleyen sistemlerdir.
Genellikle işlem yönlendirmeli ya da işlem tabanlı (process-oriented / process-driven) sistemlerdir. Belirli ticari işlemleri ya da görevleri yerine getirmek için tasarlanırlar.	Veri ambarı sistemleri konu yönlendirmelidir. (subject-oriented) Konu alanları çoğunlukla bir ya da birden fazla işletimsel sistem verisinden oluşur.
Genellikle anlık veri ile ilişkilidir.	Genellikle tarihsel ve özet veri ile ilişkilidir.
İhtiyaç durumuna göre veri genellikle güncellenir.	Genellikle değişkenlik göstermeyen veriler içerir. Yeni veriler eklenir fakat bir kez yüklenir, çok nadiren veri güncellenir. Dolayısıyla veri çoğunlukla salt okunurdur.
Hızlı ekleme ve güncelleme amacıyla eniyileme yapılır.	Yüksek yoğunluktaki veri üzerinde hızlı veri erişimi amacıyla eniyileme yapılır.
Genellikle uygulamaya özel sistemlerdir. Birbirinden bağımsız uygulamalar tekrarlı veri oluşumunu doğurur.	Uygulama katmanında sistemler entegre edilir ve veri tekrarı engellenir.
Her kayıt analiz açısından pek çok gereksiz bilgi içerir.	Analiz açısından her kayıt büyük önem taşır.
Karmaşık sorguları yaratmak ve işleme koymak zahmetlidir.	Karmaşık sorgulara gerek kalmadan ihtiyaç duyulan verilere kolayca erişilebilir.
Karmaşık sorguların sonuçları saatlerle ölçülebilen zaman dilimlerini kaplayabilir.	Sorgular kısa sürede cevaplanır.
Sisteme girilmiş her türlü veri mevcuttur. Kullanıcı hataları ve eksik bilgiler bulunabilir.	Sadece ayıklanmış ve güvenilir veri kullanılır.
Nihai kullanıcının çok fazla bilgisayar becerisinin bulunması gerekmez.	Deneyimli kullanıcılara yöneliktir, yüksek seviyeli bilgisayar becerisi gerekir.

Müşteri merkezlidir, bilgi teknolojisi profesyonelleri ve müşteriler tarafından bilgi ve sorgulama isleme için kullanılır.	Pazar merkezlidir ve analistler, uzmanlar ve yöneticiler tarafından veri analizi için kullanılır.
Varlık-bağıntı veri modelini kullanır ve uygulama merkezli veritabanı tasarımı vardır.	Yıldız ya da kar akısı modelini kullanır ve konu yönlendirmeli bir veri tabanı tasarımı vardır.
Veri ekleme, güncelleme ve silme sıklığı çok fazladır. Güncelleme sürekli yapılır.	Belirli zaman aralıklarında veri ekleme gerçekleştirilir.
Sisteme erisen kullanıcı sayısı fazladır.	Karar alıcı kademede bulunan belirli sayıdaki kullanıcılara yöneliktir.
Fazla detaylı güncel veriyi yönetir.	Karar destek amaçlı veri analizi için kullanılır. Büyük miktarlarda tarihsel veriyi yönetir.
Fiziksel kapasitesi MB ya da GB seviyelerindedir.	Fiziksel kapasitesi GB ya da TB seviyelerindedir.

2.3 Veri Ambarı Mimarisi

Veri Ambarı mimarisinin üç farklı kullanımı vardır. Bunlar Kurumsal Veri Ambarı, Veri Pazarı ve Sanal Veri Ambarından oluşmaktadır.

2.3.1 Kurumsal veri ambarı

Kapsamlı bir model gerektiren kurumsal veri ambarları organizasyonun tamamına ilişkin özetlenmiş veya detaylı verinin toplandığı ambarlardır. Tüm organizasyon için kullanılacak olan böyle bir sistem modeli geliştirmek, bunu besleyecek çeşitli kaynaklardan çekilecek verilerin ortaya konması uzun bir süre gerektirir ve esnekliği azaltır[6].

2.3.2 Veri pazarı (Data Mart-Küçük Veri Ambarı)

Bir Data Mart bir Veri Ambarının özelleştirilmiş sürümüdür. Veri Ambarları gibi Veri Pazarları da iş tarafına tarihsel eğilimler ve deneyimlerin analizinden yardımcı olacak

operasyonel verinin bir kopyasını saklamaktadır. Veri Pazarı oluşturmaktaki anahtar farklılık önceden tanımlanmış belli bir ihtiyaca yönelik bir grup bilgi ve seçilmiş veri yapılandırılmasıdır. Data Mart yapılandırılmasıyla ilgili veriye kolay erişim vurgulamaktadır[5].

Veri pazarı belirli bir konuya odaklı ya da bölüm düzeyindeki bir veri deposudur. Kurumsal veri ambarına göre tabandan tepeye yaklaşım daha kolay uygulanmaktadır. Bazı şirketler genel ve bütünsel bir veri ambarı kullanmak yerine bölüm düzeyindeki veriyi kullanmak isterler. Veri pazarı belirli bir konu veya bölüm bilgilerine odaklanırken, veri ambarı bütün şirketin bilgilerine odaklanır[6].

Veri Pazarları, Veri Ambarlarından daha az veri içerirler ve özel bir amaç için oluşturulmuşlardır. Veri Ambarı oluşturmak yerine Veri Pazarı oluşturmak aşağıdaki sebepler dolayısıyla çok tercih edilmektedir[7].

- Daha düşük maliyetli oluşturulabilmektedir.
- Veri Pazarları daha az veri içerirler, bu sayede sorgu çalıştırma zamanları daha kısadır.
- Veri Pazarları, Veri Ambarının basitleştirilmiş versiyonudur. Bu sebeple daha hızlı oluşturulur.
- Veri entegrasyonu ve bütünleştirme işlemleri daha hızlı gerçekleştirilir.
- Her Veri Pazarı özel bir veri kümesi için oluşturulduğu için, sistemi sadece veriyi ilgilendiren kişilerin kullanması gerekir. Bu sayede sistemi kullanacak kişilerin belirlenmesi daha kolay olur.

2.3.3 Sanal veri ambarı

Sanal veri ambarları, güncel olarak kullanılan operasyonel veritabanları üzerinde etkili sorgulama yapılabilmesi için oluşturulan özet veri kümeleridir. Oluşturması en kolay veri ambarı türüdür. Ancak veriler yine operasyonel veritabanı üzerinde tutulacağından çok fazla kapasite gereksinimi ortaya çıkar[6].

Veri ambarı sistemleri geliştirmede öncelikle en üst seviyede ortak bir veri modeli ortaya konması ve bölümsel veri pazarları tarafından da kullanılması önerilmektedir. Bu sayede veri ambarı ve veri pazarları aynı temeli paylaşırlar ve karşılaşılması muhtemel birçok tümleme probleminin önüne geçilmiş olur.

2.4 Karar Destek Sistemleri

Karar verme sadece bilgisayar alanında yapılan bir durum değil gündelik hayatımızda da devamlı yapmamız gereken bir durumdur. Devamlı karşılaştığımız durumlara göre bazı tespitler yapılır ve ona göre hareket edilir. Kararlar üç gruba ayrılır[10]. Yapısal kararlar devamlı yapılan ve rutin kararların olduğu gruptur. Yapısal olmayan kararlar, karar verecek kişinin hiçbir bilgiye sahip olmadan verdiğin kararlardır. Yarı Yapısal kararlar karşılaşılan sorun hakkında karar verinin sadece problemin bilinmeyen yönleriyle ilgili yargı vermesi durumudur.

Günümüz şartlarındaki bilişim teknolojisiyle yapılması günler alan işlemler saniyeler içinde yapılmaktadır. Bu karar verme sürecinin bilişim teknolojileri vasıtasıyla kullanılmaya başlanmasıyla karar destek sistemleri (KDS- Decision Support Systems-DSS) ortaya çıkmıştır.

KDS hakkında yapılan birkaç tanım aşağıda verilmiştir[11].

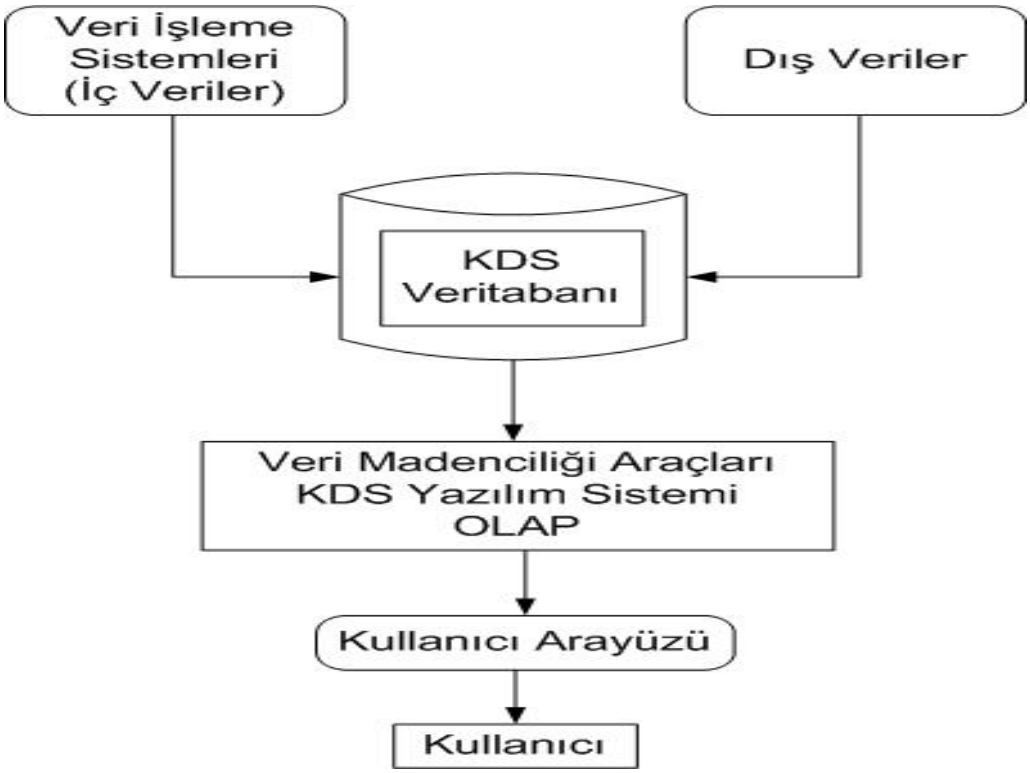
- Bir Karar Destek Sistemi, kullanıcıya yarı-yapısal ve yapısal olmayan karar verme işlemlerinde destek sağlamak amacıyla, karar modellerine ve verilere kolay erişimi sağlayan etkileşimli bir sistemdir.
- Karar Destek Sistemleri, kararın yapısal olmadığı durumlarda karar alma işlemine yardımcı olmak için tasarlanmış, esnek ve etkileşimli bilişim teknolojisi sistemleridir.
- Karar vericinin yerine geçmesinden ziyade onun kararlarını destekleyen, yarı yapısal ve yapısal olmayan problemlerin çözümü için karar vericiye karar vermesinde yardımcı olan etkileşimli sistemlerdir.

Diğer bir ifadeyle Karar Destek Sistemleri; karar vericinin, karar verebilmek amacıyla bilgisayar ile interaktif olarak faydalı bilgi üretmesine imkân veren, bilgisayarla bütünleştirilmiş yönetim bilgi sistemi araçlarıdır. Karar destek sistemleri, veri tabanı ve model tabanını bir yazılım sistemi yardımıyla işleyerek, ilgili karar verme probleminin çözümü amacıyla kullanıcı isteklerine göre alternatif çözümlerin oluşturulmasını sağlar. Oluşturulan bu alternatif çözümler doğrultusunda son karar kullanıcıya aittir[12].

Karar Destek Sistemlerinin özellikleri aşağıda listelenmiştir.

- Yarı-yapısal ve yapısal olmayan kararlarlarda kullanılır.
- Karar vericinin yerine geçmekten ziyade, ona karar vermesinde yardımcı olur.
- Karar verme prosesinin tüm aşamalarında destek sağlarlar
- Model kullanır.
- Sistem kullanıcının kontrolü altındadır.
- Kullanıcı etkileşimlidir

Karar destek sistemlerinin ana bileşenlerini veritabanı, yazılım sistemi ve kullanıcı ara yüzü oluşturur (Şekil 2.2). Veritabanı, bir kişisel bilgisayara yerleştirilecek kadar küçük olabildiği gibi, çok büyük bir veri deposu şeklinde de olabilir. Veritabanı, birçok uygulamadan elde edilen geçmişteki ve mevcutta bulunan verilerin bir araya gelmesinden oluşmaktadır. Yazılım sistemi, veri analizi için kullanılan yazılım araçlarını kapsar (örneğin C#, Delphi, Visual Basic, C++). Bu sistem, KDS kullanıcısının kolayca erişebileceği çeşitli veri madenciliği araçlarından veya matematiksel ve analitik modellerin bir araya gelmesinden oluşmaktadır. Veri madenciliği yazılım araçları, büyük veri havuzlarında gizlenmiş desenleri ve ilişkileri bulur, onlardan gelecekteki davranışların tahmin edilebilmesi için kurallar oluşturur ve kararın verilmesinde yol gösterir[12].



Şekil 2.2: Karar destek sisteminin bileşenleri[11]

Kullanıcı ara yüzü, karar vericilerin, KDS'lerine erişimini sağlar. Kullanıcı ara yüzü, kullanıcılar ile yazılım ve donanım arasındaki iletişime yardımcı olur. Kullanıcı, karar destek sistemini yöneten kişidir. Kullanıcı, ara yüzü yardımıyla karar destek sistemini yönlendirmektedir. Kullanıcı, karar problemi hakkında karar verici pozisyondadır. Ele aldığı problemin gerekleri doğrultusunda karar destek sistemini kullanarak sonuç raporlarından veya tablo analizlerinden hareketle, alternatif çözümler içerisinden en iyiyi bulmaya çalışır[11].

3. VERİ MADENCİLİĞİ

3.1 Giriş

Bilgisayar bilimcileri Moore kanununun belirttiği, bilgisayar işlem gücünün her 18 ayda bir ikiye katlanacağı kuralını sık sık dile getirirler. Daha az bilinen durum ise bilgisayarların depo kapasiteleri her dokuz ayda ikiye katlanmaktadır. Sanki bir gaz elementi gibi bilgisayar veritabanları da bulabildikleri tüm depo alanlarını doldururlar. Veritabanlarında bulunan bu çok miktardaki veri, ortaya çıkmamış bir kaynağı belirtir. Sanki bir altın madeni gibi bu veriler bilgiye çevrilebilir. Bu bilgilerde veri madenciliği teknikleri kullanarak "Değerli Bilgi" ye çevrilebilir.[13]

Şirketlerde, üniversitelerde, devlet kurumlarında ve diğer kuruluşlardaki çok büyük miktardaki kullanılmayan veriyi ifade etmek zorlaşmaktadır. İş yerleri her yıl bu verilere terabaytlarca yenileri eklemektedir. Bu verilerden bilgiyi çıkarmanın getirisi paha biçilemezdir. Veritabanlarındaki Bilgi Keşfi Süreci(BKS)(Knowledge Discovery in Database (KDD)) uzun zamandır bu amaç için geliştirilmektedir. Veri Madenciliği de bu BKS işlemi içerisindedir.[13]

Veri madenciliği, veriler arasında bilgi desenleri arayan karar destek sistemidir. Bir başka değişle; Veriler arasından yeni, geçerli, anlaşılır ve potansiyel olarak yararlı desen çıkarma işlemidir. Burada desen veriler arasındaki ilişkiyi belirtir[14]. Veri Madenciliği, daha önce sistemde bulunan ama bilinirliği olmayan, ilgi çekici, potansiyel olarak yararlı olan verinin bilgiye dönüştürülme sürecidir[15].

Veri Madenciliğini tanımlayan diğer yaklaşımlara bakacak olursak; Veri madenciliği, çok büyük miktardaki gözlenebilir verinin analiz edilmesiyle, beklenmedik veri ilişkilerinin ve sıra dışı sonuçların veri sahibine anlaşılır bir şekilde iletilmesidir.

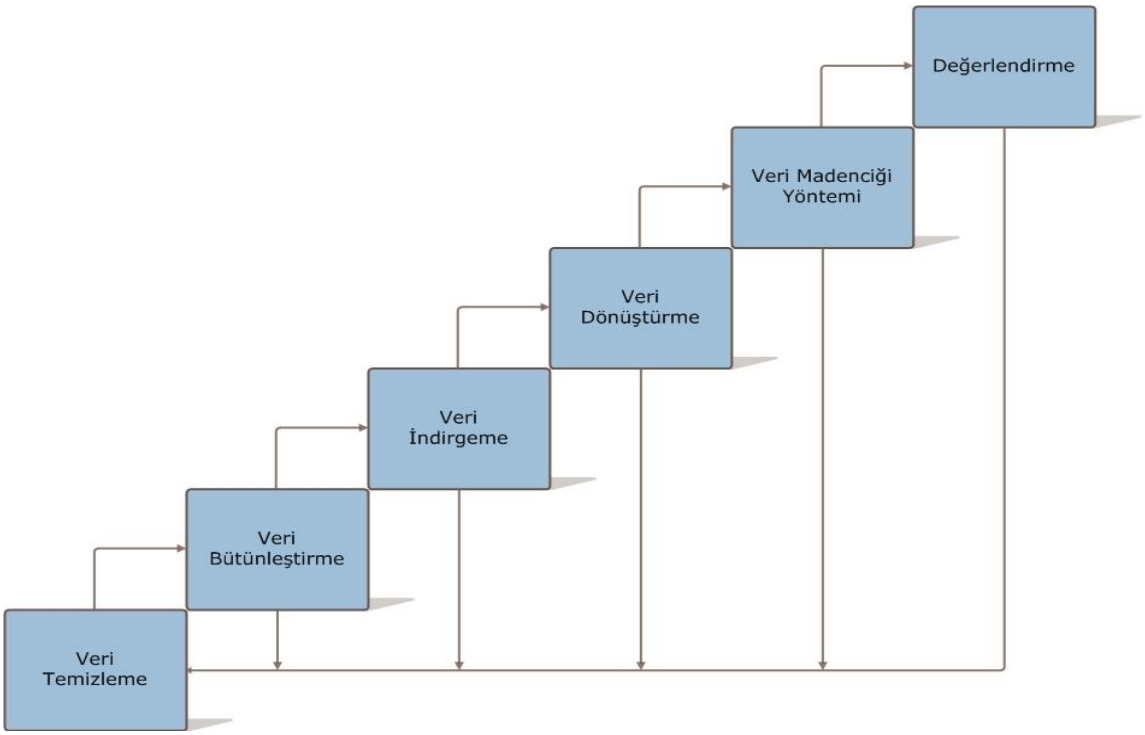
Diğer bir tanım ise şu şekildedir; Daha önceden bilinmeyen, geçerli ve uygulanabilir bilgilerin geniş veritabanlarından elde edilmesi ve bu bilgilerin kuruluş kararları verirken kullanılabilmesidir. Daha basit olarak veri madenciliği büyük ölçekli veriler arasından değerli olan bilgiyi bulup ortaya çıkarılmasına denir[16].

3.2 Veri Madenciliği Süreci

Veri Madenciliği büyük ölçekli verilerden anlamlı bilginin elde edilebilmesinde gerekli olan işlemlerin modellenmesi için yöntemler ve algoritmalar sunar. "Veri tabanlarında BKS" olarak da adlandırılan bu modelleme yöntemleri büyük veri kümeleri içerisinde gelecekle ilgili tahmin yapabilecek programların yazılmasında yardımcı olur[17]. Bu süreçlerin veri hazırlanmasını oluşturan ilk dört basamağı veri ambarı oluşturma sürecinde de belirtilebilir. Bu süreç birçok adımdan oluşan bir süreçtir. Süreç, aşağıdaki adımları içermektedir[18].

a. Veri temizleme
b. Veri bütünleştirme
c. Veri indirgeme
d. Veri dönüştürme
e. Veri madenciliği algoritmasını uygulama
f. Sonuçları sunum ve değerlendirme

Sitemin kurulması aşamasında en önemli kısım ilk dört basamaktır. Algoritmayı uygularken ortaya çıkacak sorunlarda, bu aşamaya geri dönülmesi ve verilerin yeniden düzenlenmesine neden olacaktır. Bu durum verilerin hazırlanması ve modelin kurulması aşamaları için, bir analistin veri keşfi sürecinin toplamı içerisinde enerji ve zamanının %50 - %85'ini harcamasına neden olmaktadır[19].



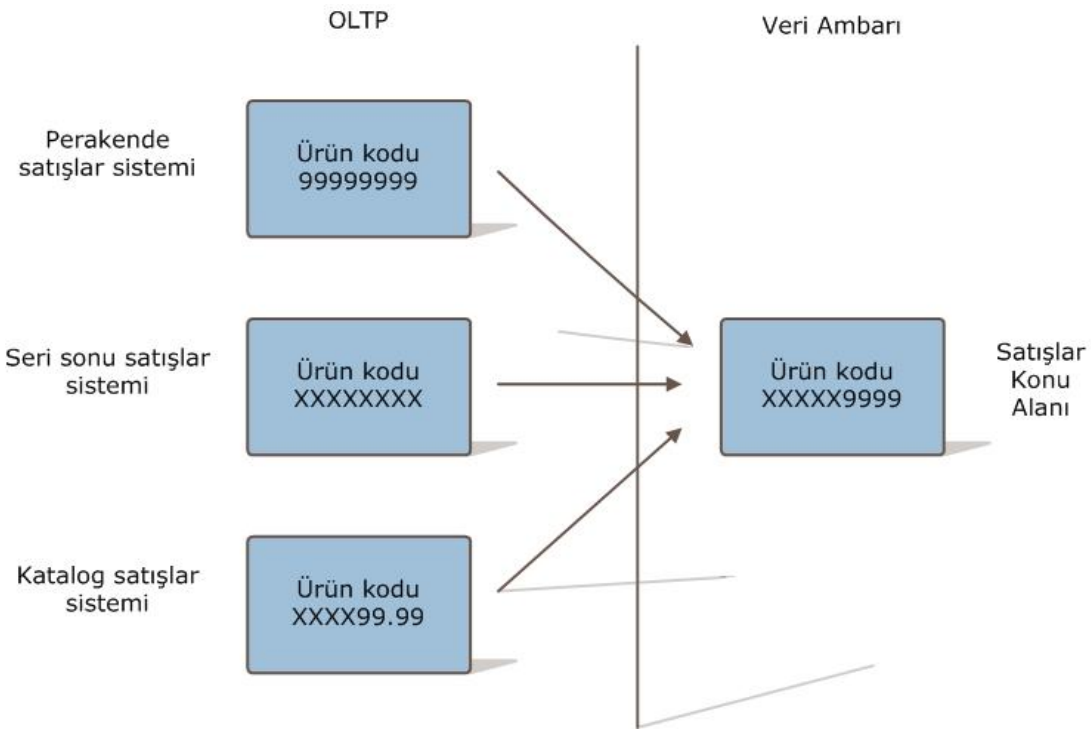
Şekil 3.1: Veri madenciliği süreci

3.2.1 Veri temizleme

Bazı problemlerde üzerinde işlem yapılacak verilerin istenilen özelliklere sahip olmadığı durumlar olabilir. Doğru verilerin yanında eksik veya hatalı verilerle karşılaşılabilir. Veri temizleme, diğer aşamalarda kullanılacak veri madenciliği modelinde veri kalitesini artırmak üzere yapılır[20]. Tutarsız ve hatalı veriler, veritabanı üzerinde yapılacak analizlerde doğru sonuç almamızı etkileyeceği için bu "gürültü" olarak tabir edilen verilerden sistemin temizlenmesi gerekir. Bu durumdan kurtulmak için birçok yol kullanılabilir. Gürültülü verinin sisteme etkisini önlemek için bu tip veriler analiz kümesinden çıkarılabilir veya bu tip verilerin yerine sabit bir değer koyulabilir ya da çok büyük veri kümeleriyle uğraşılmıyorsa manüel olarak da değişiklik yapılabilir. Örnek olarak "-1" değeri verilerek tüm hatalı verilerin bu değeri kullanması sağlanabilir (Tüm değerlerin aynı sabit değerle değiştirilmesi sistemi yanlış desenlere de yöneltebileceği göz önüne alınmalıdır). Başka bir uygulamada hatalı değerler yerine, doğru verilerin ortalaması yazılabilir. Bunu daha da özelleştirirsek tüm değerlerin ortalaması yerine hatalı kaydın yapısına benzer örneklerin ortalaması alınarak işlem yapılabilir.

3.2.2 Veri bütünleştirme

Birçok veri kaynağından çekilen farklı türdeki verilerin birlikte değerlendirmeye alınabilmesi için verilerin tek bir ortak türe dönüştürülmesi gerekmektedir. Örnek vermek gerekirse, cinsiyet ile ilgili bir alan uygulamada sadece simgeler "E" ve "K" kodlarıyla belirtilmiş olabilir. "E" kodu erkek, "K" kodu ise kadınları simgelemektedir. Bir başka uygulamada ise söz konusu cinsiyet ile ilgili alan 1 veya 0 değerleri ile ifade edilmiş olabilir. Farklı bir uygulamada direkt olarak "Erkek" ve "Kadın" ifadeleri cinsiyetleri belirtmek için kullanılmış olabilir. Başka bir örnekte uzunluk ölçüsü olarak cm, bazılarında inç, bazılarında ise metre olarak kullanılmış olabilir. Şekil 3. 2'de bu konu ile ilgili başka bir örnek anlatılmaktadır. Bu tip farklı bakış açıları veriler üzerinde çalışmayı imkânsız hale getirir. Bu nedenle bu tip verilerin analiz aşamasından önce ortak bir türe dönüştürülmesi yani veri bütünleştirmesi yapılması gerekir.



Şekil 3. 2: Farklı veri kaynaklarından alınan türlerin veri bütünleştirme işlemi

3.2.3 Veri indirgeme

Veri madenciliğinde çözümleme işlemleri bazen çok uzun süre alabilir. Veri kümesinde aynı tipte çok kayıt olduğu biliniyor ve bu kayıtlarının bazılarının çıkarılması sonucu değiştirmeyeceği düşünülüyorsa, kaynak verilerin sayısı azaltılabilir. Böyle bir ihtiyaç söz konusu olursa veri sıkıştırma kullanarak verilerin daha az yer kaplamaları sağlanabilir. Veya bazı özellikleri belirtmede büyük veri kümeleri yerine daha küçük veri grupları kullanılabilir. Başka bir durumda benzer karakteristiğe sahip birçok özellik yerine daha az özellik kullanılarak işlemlerin daha hızlı yapılması sağlanabilir.

3.2.4 Veri dönüştürme

Veri Madenciliğinde bazı zamanlar verileri aynen işleme katmak kurulan sistem için uygun olmayabilir. Bazı değişkenlerin ortalaması ve varyansları, diğer değişkenlerden çok büyük veya çok küçük olması durumunda, bu büyük fark yaratan değişkenlerin diğerleri üzerinde analiz aşamasında etkisi daha çok olur ve onların rollerini önemli ölçüde azaltır. Ayrıca değişkenlerin sahip olduğu çok büyük ve çok küçük değerler de çözümlemenin sağlıklı bir şekilde yapılmasını engeller. Bu durumda verinin standartlaşması için Min-Max normalleştirme veya Z-score standartlaştırma yöntemleri kullanılabilir. Bunların yanında seçilen veri madenciliği modeline göre değişkenlerin aralık değerlerini değiştirip gruplama yapılabilir. Örnek vermek gerekirse, 0'dan 100'e kadar olan sayılarla çalışmak yerine grubu beş farklı aralığa bölerek sistem analiz için daha uygun hale getirilebilir.

3.2.4.1. Min-Max normalleştirilmesi

Verileri 0 ile 1 arasındaki sayısal değerlere dönüştürmek için min-max normalleştirme yöntemi uygulanır. Bu yöntem, veri içindeki en büyük ve en küçük sayısal değerin belirlenerek diğerleri buna uygun biçimde dönüştürme esasına dayanmaktadır. Söz konusu dönüştürme yapısı denklem 3.1'de ifade edilmektedir:

$$X* = \frac{X - \text{Min}(X)}{\text{Max}(X) - \text{Min}(X)} \quad (3.1)$$

Burada X* dönüştürmüş değerleri, X gözlem değerlerini, Min(X) en küçük gözlem değerini ve Max(X) en büyük gözlem değerini ifade etmektedir.

3.2.4.2. Z-score standartlaştırma

Bu yöntem, verilerin ortalaması ve standart hatası göz önüne alınarak yeni değerlere dönüştürülmesi esasına dayanmaktadır. Söz konusu dönüşümlerde, 3.2'de gösterildiği şekilde bir bağıntıya yer verilir:

$$X* = \frac{X - \text{Mean}(X)}{\text{Std}(X)} \quad (3.2)$$

Burada X* dönüştürme değerleri, X gözlem değerini, Mean(X) verilerin aritmetik ortalamasını ve Std(X) gözlem değerinin standart sapmasını ifade etmektedir(3.3).

$$\text{Mean}(X) = 1/n \sum_{i=0}^{n} (X_i) \qquad \text{Std}(X) = \sqrt{\sum_{i=1}^{n} \left(X_i - \text{Mean}(X)\right)^2} \quad (3.3)$$

3.2.5 Veri madenciliği algoritmasını uygulama

Veri madenciliği yöntemlerini uygulayabilmek için yukarıda sıralanan işlemlerin uygun görünenleri yapılır. Veri hazır hale getirildikten sonra konuyla ilgili veri madenciliği algoritmaları uygulanır. Söz konusu algoritmalar sınıflandırma, kümeleme ve birliktelik kuralları konusunda olacaktır. 3.3'de bu algoritmalara değinilecektir.

3.2.6 Sonuçları sunum ve değerlendirme

Veri madenciliği algoritması veriler üzerinde uygulandıktan sonra, sonuçlar düzenlenerek karar vericilere sunulur. Bu sunum işlemi yazılı bir raporlama olabileceği gibi, web sayfasından kişilere bilgilendirme de olabilir. Bulunan sonuçlara göre

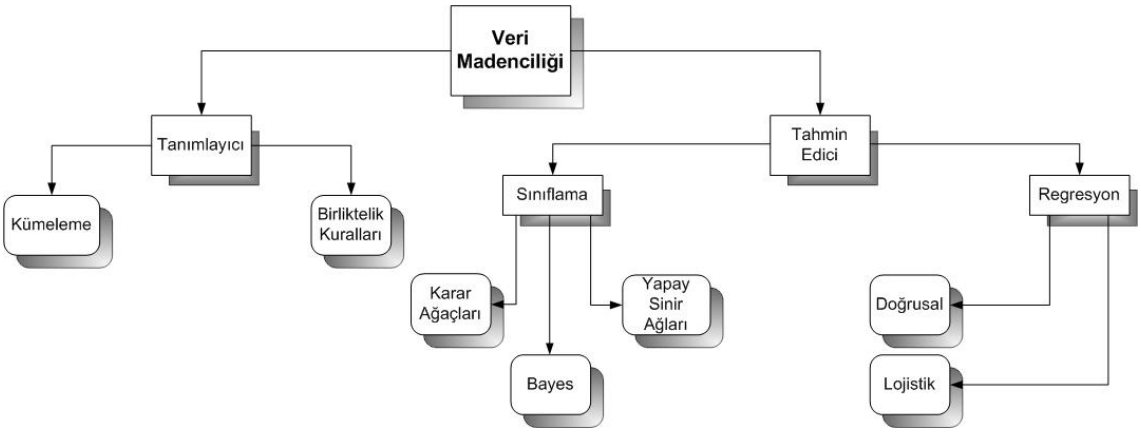
incelenen sistem üzerinde iyileştirme çalışmaları yapılır veya belirlenen durumlar karşısında önlemler alınır.

3.3 Veri Madenciliği Modelleri

Veri madenciliğinde kullanılan modeller, tahmin edici (predictive) ve tanımlayıcı (descriptive) olmak üzere iki ana başlık altında incelenmektedir. Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır. Örneğin bir banka önceki dönemlerde vermiş olduğu kredilere ilişkin gerekli tüm verilere sahip olabilir. Bu verilerde bağımsız değişkenler kredi alan müşterinin özellikleri, bağımlı değişken ise kredinin geri ödenip ödenmediğidir. Bu verilere uygun olarak kurulan model, daha sonraki kredi taleplerinde müşteri özelliklerine göre verilecek olan kredinin geri ödenip ödenmeyeceğinin tahmininde kullanılmaktadır. Tanımlayıcı modellerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır. X/Y aralığında geliri, evi ve arabası olan, ayrıca çocukları okul çağında olan aileler ile çocuğu olmayan ve geliri X/Y aralığından düşük olan ailelerin satın alma örüntülerinin birbirlerine benzerlik gösterdiğinin belirlenmesi tanımlayıcı modellere bir örnektir. Veri madenciliği modelleri gördükleri işlevlere göre üçe ayrılmaktır[21].

a. Sınıflama (Classification) ve Regresyon
b. Kümeleme (Clustering)
c. Birliktelik Kuralları (Association Rules) ve Ardışık Zamanlı Örüntüler (Sequential Patterns)

Sınıflama ve regresyon modelleri tahmin edici, kümeleme, birliktelik kuralları ve ardışık zamanlı örüntü modelleri tanımlayıcı modellerdir. Şekil 3.3.'de bu ilişkiler özetlenmiştir.



Şekil 3.3: Veri madenciliği modelleri

3.3.1 Sınıflama ve regresyon

İstenilen bir değişken bağımlı değişken ve diğerleri tahmin edici (bağımsız) değişkenler olarak adlandırılır. Amaç, girdi olarak tahmin edici değişkenlerin yer aldığı modelde, çıktının bağımlı değişkenin değerinin bulunduğu anlamlı bir model kurmaktır. Bağımlı değişken sayısal değil ise problem sınıflama problemidir. Eğer bağımlı değişken sayısal ise problem regresyon problemi olarak adlandırılır[21].

Sınıflama veri madenciliğinde sıkça kullanılan bir yöntem olup, veri tabanlarındaki gizli örüntüleri ortaya çıkarmakta kullanılır. Verilerin sınıflandırması için belirli bir süreç izlenir. Öncelikle var olan veritabanının bir kısmı eğitim amacıyla kullanılarak sınıflandırma kurallarının oluşturulması sağlanır. Daha sonra bu kurallar yardımıyla yeni bir durum ortaya çıktığında nasıl karar verileceği belirlenir.

Sınıflama ve regresyon modelinde kullanılan başlıca yöntemler aşağıda listelenmiştir.

- Yapay Sinir Ağları (Artificial Neural Networks)
- Genetik Algoritmalar (Genetic Algorithms)
- K-En Yakın Komşu (K-Nearest Neighbor)
- Karar Ağaçları (Decision Trees)
- Bayes

3. 3. 1. 1 Yapay sinir ağları

Yapay sinir ağları, temelde tamamen insan beyni örneklenerek geliştirilmiş bir teknolojidir. Bilindiği gibi; öğrenme, hatırlama, düşünme gibi tüm insan davranışlarının temelinde sinir hücreleri bulunmaktadır. İnsan beyninde tahminen $(10)^{11}$ adet sinir hücresi olduğu düşünülmektedir ve bu sinir hücreleri arasında sonsuz diyebileceğimiz sayıda sinaptik birleşme denilen sinirler arası bağ vardır. Bu sayıdaki bir birleşimi gerçekleştirebilecek bir bilgisayar sisteminin dünya büyüklüğünde olması gerektiği söylenmektedir[19].

Biyolojik sistemlerde öğrenme, nöronlar arasındaki sinaptik bağlantıların ayarlanması ile oluşturulur. İnsan yasamı süresince tecrübeler edinir, bu tecrübelerin sinaptik bağlantıları etkilediği ve öğrenmenin bu şekilde geliştiği düşünülmektedir. Yapay sinir ağlarında bu ayarlamayı yapmak ve öğrenmeyi sağlamak için ağırlık fonksiyonları kullanılmaktadır, insanın deneme yanılma yoluyla öğrenmesi yapay sinir ağlarında yinelemeli eğitim sayesinde gerçekleştirilmektedir[22].

Normal bir sinir ağı üç farklı tip katmandan meydana gelir, giriş katmanı, saklı katman ve çıkış katmanı. Buradaki önemli nokta, ağda üç farklı tipte katman olduğudur, sadece üç katman olduğu değil. Araştırmacının isteğine göre saklı katman sayısı artırılabilir. Bu, ne kadar kompleks bir yapı istendiğine bağlıdır. Giriş katmanı, giriş verisini içerir; çıkış katmanında ise tüm saklı katmanlarda işlem yapıldıktan sonra oluşturulan sonucu içerir. Giriş ve çıkış katmanları, saklı katman tarafından aracılı olduğundan, yapay sinir ağları çoğu zaman bir kara kutu(black box) olarak bilinir[23].

Yapay sinir ağları için farklı yapılar vardır ve bunların her biri verilen işleri yapmak için farklı yol ve öğrenme yöntemleri kullanırlar. Yapay sinir ağları, veri madenciliği için çok kullanışlı bir yöntem olmasının yanında anlaşılmaz modeller de ortaya çıkardığı için uygulaması çok uzun zaman gerektirebilir. Bu duruma rağmen yapay sinir ağları; sınıflama, kümeleme ve tahmin amaçları ile kolaylıkla kullanılabilecek genel amaçlı ve güçlü araçlardır.

3. 3. 1. 2 Genetik algoritmalar

Genetik algoritmalar, biyoloji biliminden yararlanılarak geliştirilmiş önemli makine öğrenimi yöntemlerinden birisidir. Doğada gözlemlenen evrimsel sürece benzer bir şekilde çalışan arama ve en iyileme yöntemidir. Karmaşık çok boyutlu arama uzayında en iyinin hayatta kalması ilkesine göre bütünsel en iyi çözümü arar. Genetik algoritmalar problemlere tek bir çözüm üretmek yerine farklı çözümlerden oluşan bir çözüm kümesi üretir. Böylelikle, arama uzayında aynı anda birçok nokta değerlendirilmekte ve sonuçta bütünsel çözüme ulaşma olasılığı yükselmektedir. Çözüm kümesindeki çözümler birbirinden tamamen bağımsızdır. Her biri çok boyutlu uzay üzerinde bir vektördür[19].

Genetik algoritmalar problemlerin çözümü için evrimsel süreci bilgisayar ortamında taklit ederler. Diğer en iyileme yöntemlerinde olduğu gibi çözüm için tek bir yapının geliştirilmesi yerine, böyle yapılardan meydana gelen bir küme oluştururlar. Problem için olası pek çok çözümü temsil eden bu küme genetik algoritma terminolojisinde popülasyon adını alır. Nüfuslar vektör veya birey adı verilen sayı dizilerinden oluşur. Birey içindeki her bir elemana gen adı verilir. Nüfustaki bireyler evrimsel süreç içinde genetik algoritma işlemcileri tarafından belirlenir[19]. Genlerin oluşturduğu yapılar olan kromozomlar üzerinden de hesaplamalar gerçekleştirilir.

Genetik algoritmalar açıklanabilir sonuçlar üretirler. Değişik tiplerdeki verileri işleme özelliğine sahip olan genetik algoritmalar en iyileme (optimization) amacı ile kullanılabilirler. Ayrıca genetik algoritmalar yapay sinir ağları ile ortaklasa çalışarak başarılı sonuçlar üretmektedirler. Genetik algoritmalar yapay sinir ağlarının eğitilmesi, bellek tabanlı yöntemlerde birleşim fonksiyonunun oluşturulması gibi işlerde de kullanılmışlardır[24].

Tüm bu olumlu yönlerine rağmen genetik algoritmaların kullanımlarında bazı sıkıntılar da yaşanmaktadır. Bunlardan en belirgin olanı karmaşık sorunların genetik kodlanmasının çok zor olmasıdır. Ayrıca en iyi (optimal) sonucun üretildiğine dair bir garanti de bulunmamaktadır[24].

3. 3. 1. 3 K – En yakın komşu algoritması

En yakın komşu metodu 1950'lerin başında ilk defa adını duyurmuştur. Büyük veri kümelerinin eğitilmesinde çok büyük işlem gücü gerektirdiği için sistem, işlem gücü fazlalaşmış bilgisayarların 1960 'da ulaşılabilir olmasına kadar yaygınlık kazanmamıştır. En yakın komşu algoritması örneklerin benzerliklerine göre sınıflandırma üzerine kurulmuştur. Test verilerini, eğitim verilerinin benzerliğine göre karşılaştırarak sınıflama işlemini yapar. Örnek vermek gerekirse, gelen veri kümesi n kadar özelliği olduğunu düşünelim. Bu durumda tüm eğitim verileri n boyutlu desen uzayına gösterili olacaktır. Yeni bir veri geldiği zaman k-en yakın komşu algoritması, verinin k. merkeze yakın olduğunu bulmak için bir yakınlık hesaplaması yapar. Bu işlemde Öklid uzaklık formülü kullanılabilir(3.4)[25].

$$\text{Uzaklik}(X_1, X_2) = \sum_{i=1}^{n} (X_{1i} - X_{2i})^2 \qquad (3.4)$$

$X_{11+}X_{12+} \ldots + X_{1n}$ $X_{2=}X_{21+}X_{22+} \ldots + X_{2n}$

Buradaki k tane komşudaki sonuca göre test verisi kendine en yakın komşuya yerleştirilir. Kümeler daha sonra yeni merkezi bulmak için tekrar analiz edilir. Eldeki tüm veriler için bu işlem tekrarlanır[26].

3. 3. 1. 4 Karar ağaçları

Veri madenciliğinde karar ağaçları, kurulmasının ucuz olması, yorumlanmalarının kolay olması, güvenilir olmaları ve veri tabanı sistemlerine kolayca entegre edilebilmeleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahip yöntemdir[21].

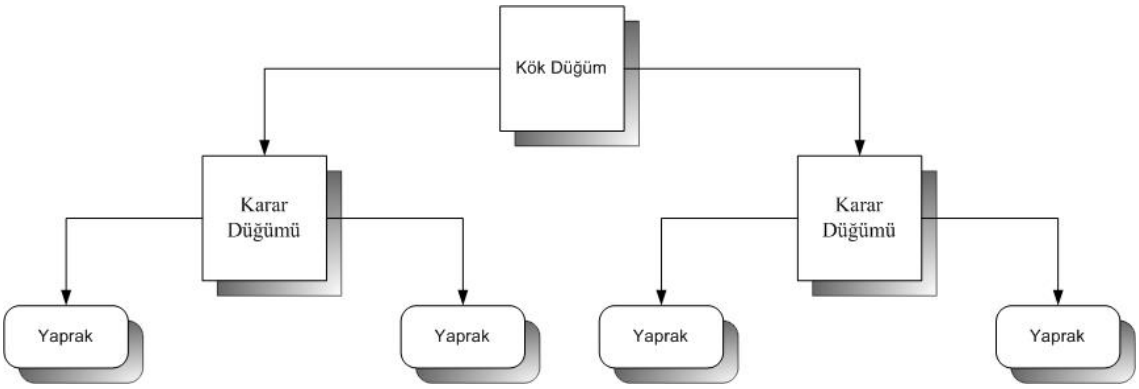
Karar ağacı karar düğümleri, dallar ve yapraklardan oluşur. Karar düğümü, gerçekleştirilecek testi belirtir. Bu testin sonucu ağacın veri kaybetmeden dallara ayrılmasına neden olur. Her düğümde test ve dallara ayrılma işlemleri ardışık olarak gerçekleşir ve bu ayrılma işlemi üst seviyedeki ayrımlara bağımlıdır. Ağacın her bir dalı sınıflama işlemini tamamlamaya adaydır. Eğer bir dalın ucunda sınıflama işlemi gerçekleşemiyorsa, o dalın sonucunda bir karar düğümü oluşur. Ancak dalın sonunda

belirli bir sınıf oluşuyorsa, o dalın sonunda yaprak vardır. Bu yaprak, veri üzerinde belirlenmek istenen sınıflardan biridir. Karar ağacı işlemi kök düğümünden başlar ve yukarıdan aşağıya doğru yaprağa ulaşana dek ardışık düğümleri takip ederek gerçekleşir[19].

Karar ağacı algoritmaları içinde ID3, C4.5, C5.0, M5P, J48 algoritmaları sıklıkla kullanılan algoritmalardır. İyi bir karar ağacı algoritması hem ayrık hem de sürekli veriler üzerinde çalışabilmelidir. Sürekli veri üzerinde oluşturulan modellere regresyon ağacı adı verilmektedir[27].

Karar ağacı algoritmalarında her zaman sorgulanan kategorilerde en uygun dala ayrılmaya çalışılır. Bu ayrılma işleminin hesaplanmasında da farklı hesaplama türleri kullanılır. Sınıflama ve regresyon(Classification and Regression) ağaçları çeşitlik indeksini(index of diversity) kullanırken, ID3 algoritması entropi değerine göre düğümlerin ayrılmasını hesaplar[28].

Örnek bir karar ağacı yapısı şekil 3.4'de gösterilmiştir.



Şekil 3. 4: Eğitim verilerine uygun karar ağacı

3. 3. 1. 5 Bayes sınıflandırıcıları

Sınıflandırma işleminde istatistiksel teknikler de kullanılmaktadır. Bunlardan birisi de Bayes teoremine dayanmaktadır. Değişkenlere ait alt kümeler arasındaki koşullu bağımsızları tanımlanarak bayes sınıflandırıcıları kullanılabilir.

Naive bayes algoritmasında her kriterin sonuca olan etkilerinin olasılık olarak hesaplanması temeline dayanmaktadır. Veri Madenciliği işlemini en çok verilen örneklerden biri ile açıklayacak olursak elimizde tenis maçının oynanıp oynanmamasına

dair bir bilgi olduğunu düşünelim. Ancak bu bilgiye göre tenis maçının oynanması veya oynanmaması durumu kaydedilirken o anki hava durumu, sıcaklık, nem ve rüzgar durumu bilgileri de alınmış olsun. Biz bu bilgileri değerlendirdiğimizde varsayılan tahmin yöntemleri ile "hava bugün rüzgârlı, tenis maçı bugün oynanmaz" şeklinde kararları farkında olmasak da veririz. Ancak Veri Madenciliği bu kararların tüm kriterlerin etkisi ile verildiği bir yaklaşımdır. Dolayısıyla biz ileride öğrettiğimiz sisteme bugün hava güneşli, sıcak, nemli ve rüzgar yok şeklinde bir bilgiyi verdiğimizde sistem eğitildiği daha önce gerçekleşmiş istatistiklerden faydalanarak tenis maçının oynanma ve oynanmama ihtimalini hesaplar ve bize tahminini bildirir[19].

3.3.2 Kümeleme modelleri

Nesnelerin kendilerini ya da diğer nesnelerle olan ilişkilerini tarif eden bilgileri kullanarak nesneleri gruplara ayırma işlemine kümeleme denir. İnsanoğlu için beş ya da altı niteliğe sahip bir nesneler topluluğunu, niteliklerdeki benzerliklere bakarak gruplamak mümkünken bu sayının çok daha fazla olması durumunda bu işlem olanaksız hale gelmektedir. Kümelemede amaç; grup içindeki nesneleri, diğer gruplardaki nesnelerden olabildiğince ayrı/bağımsız, kendi aralarında ise birbirine benzer/bağımlı olacak şekilde oluşturmadır[29]. Bu noktada tanımlayıcı bir veri madenciliği yöntemi olan kümeleme devreye girmekte ve veriyi çeşitli tekniklerle önceden sayısı bilinmeyen kümelere bölmektedir[27].

Kümeleme modellerinde amaç üyelerinin birbirlerine çok benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veritabanındaki kayıtların bu farklı kümelere bölünmesidir. Kümeleme analizinde; veritabanındaki kayıtların hangi kümelere ayrılacağı veya kümelemenin hangi değişken özelliklerine göre yapılacağı konunun uzmanı olan bir kişi tarafından belirtilebileceği gibi veritabanındaki kayıtların hangi kümelere ayrılacağını geliştirilen bilgisayar programları da yapabilmektedir[19].

3.3.3 Birliktelik kuralları ve ardışık zamanlı örüntüler

Veritabanı içinde yer alan kayıtların birbirleriyle ilişkilerini inceleyerek, hangi olayların eş zamanlı olarak birlikte gerçekleşebileceğini ortaya koymaya çalışan veri madenciliği yöntemleri bulunmaktadır. Bu ilişkilerin belirlenmesi ile "Birliktelik Kuralları" (Association Rules) elde edilir.

Bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi, müşteriye daha fazla ürünün satılmasını sağlama yollarından biridir. Satın alma eğilimlerinin tanımlanmasını sağlayan birliktelik kuralları ve ardışık zamanlı örüntüler, pazarlama amaçlı olarak pazar sepeti analizi (Market Basket Analysis) adı altında veri madenciliğinde yaygın olarak kullanılmaktadır. Bununla birlikte bu teknikler, tıp, finans ve farklı olayların birbirleri ile ilişkili olduğunun belirlenmesi sonucunda değerli bilgi kazanımının söz konusu olduğu ortamlarda da önem taşımaktadır[30].

Pazar sepet analizleri yardımıyla bir müşteri herhangi bir ürünü aldığında, sepetine başka hangi ürünleri de koyduğu bir olasılığa göre tahmin edilir. Birlikte satın alınan ürünler belirlendiğinde, mağazalarda raflar ona göre düzenlenerek müşterilerin bu tür ürünlere daha kolayca erişmeleri sağlanabilir. Bu modelin en bilinen temsilcisi "Apriori" algoritmasıdır.

Birliktelik kuralları aşağıda sunulan örneklerde görüldüğü gibi eş zamanlı olarak gerçekleşen ilişkilerin tanımlanmasında kullanılır[21].

- Müşteriler bira satın aldığında, % 75 ihtimalle patates cipsi de alırlar,
- Düşük yağlı peynir ve yağsız yoğurt alan müşteriler, %85 ihtimalle diet süt de satın alırlar.

Ardışık zamanlı örüntüler ise aşağıda sunulan örneklerde görüldüğü gibi birbirleri ile ilişkisi olan ancak birbirini izleyen dönemlerde gerçekleşen ilişkilerin tanımlanmasında kullanılır.

- X ameliyatı yapıldığında, 15 gün içinde % 45 ihtimalle Y enfeksiyonu oluşacaktır,

- İMKB endeksi düşerken A hisse senedinin değeri % 15'den daha fazla artacak olursa, üç iş günü içerisinde B hisse senedinin değeri % 60 ihtimalle artacaktır,
- Çekiç satın alan bir müşteri, ilk üç ay içerisinde % 15, bu dönemi izleyen üç ay içerisinde % 10 ihtimalle çivi satın alacaktır.

3.4 Çalışmada Kullanılan Veri Madenciliği Modelleri

Veri madenciliği çalışmalarında önceki bölümlerde bahsedildiği üzere çok fazla algoritma ve yöntem vardır. Yapılan çalışmaların her model için uygun olacağı ya da kullandığımız modelin incelediğimiz durumda nasıl tepki verebileceği algoritmadan algoritmaya farklılık gösterecektir. Bu durumda, incelenen özellik için birden çok model geliştirilerek sonuçlar arasında farkların incelenmesi, bizi en doğru sonuca ulaştırmada önemlidir.

Veri madenciliğinde kullanılabilecek birçok program vardır. Bu programların sayısı onlarla ifade edilmektedir[31]. Her program mevcut ana algoritma kalıplarından yararlanarak kendi sistemlerini geliştirmişlerdir. Örnek vermek gerekirse karar ağaçları algoritması tek bir ana başlık olarak incelenmesine rağmen detayında bir çok farklı algoritma tanımlanmıştır; CART, ID3, C4.5. Ayrıca bu temel karar ağacı yapılarını kullanan başka algoritmalarda vardır. Bu programlar arasından, bu çalışmada kullanılmak üzere verilerin saklanması ve yönetimi için Microsoft SQL Server 2008, analizi içinde Microsoft Analysis Services 2008 seçilmiştir. KDS de analiz işlemleri içinde, Microsoft Naive Bayes, Microsoft Decision Trees, Microsoft Association Rules, Microsoft Neural Network algoritmaları kullanılmıştır.

Tüm bu algoritmalar daha önce anlatılan özelliklerin temel yapılarını karşılamaktadır. Algoritmaların içerikleri ve teknik işleyişleriyle ilgili bilgi aşağıda verilmiştir.

3.4.1 Microsoft Naive Bayes

Matematiksel modeli Thomas Bayes tarafından geliştirilen bu metot, koşullu ve koşulsuz olasılıkların birlikteliğini kullanır. Yapı kullanılan diğer algoritmalara göre daha az işlem karmaşıklığı içerir. Bu sebeple hızın önemli olduğu durumlarda kullanılması tavsiye edilir[35]. Aşağıda bu algoritmanın nasıl çalıştığıyla ilgili bir örnek verilmiştir.

Tablo 3.1'de örnek veriler gösterilmektedir. Çeşitli aktivitelere göre farklı cinsiyetteki kişilerin, ilgili işi yapmak isteyip istemediği belirtilmiştir. Kadın ve erkek olarak, "Dizi İzleme", "Sinemaya Gitme", "Maç Seyretme" ve "Bilgisayarla Vakit Geçirme" aktivitelerinden hangilerini yapmak isteyip istemedikleri belirtilmiştir. Tablo 3.2'de ise dört aktivitede neler yapmak istediğini belirtmiş bir kişinin cinsiyeti sorgulanmak istenmektedir. Bu işlem Tablo 3.1 deki eğitim verilerine göre bulunur.

Tablo 3.1: Microsoft Naive Bayes eğitim verileri

	Dizi İzleme		Sinemaya Gitme		Maç Seyretme		Bilgisayarla Vakit Geçirme		Cinsiyet	
	E	K	E	K	E	K	E	K	E	K
Evet	29	90	45	79	113	18	109	89	130	100
Hayır	101	10	85	21	17	82	21	11		
Evet Yüzdesi	%22	%90	%36	%79	%90	%18	%84	%89	%57	%43
Hayır Yüzdesi	%78	%10	%64	%21	%10	%82	%16	%11		

Elimizdeki kişi sayısı incelenecek olursa eldeki 230 kişinin %57'si erkek, % 43'ü kadındır. Sadece bu veriye göre bakacak olursak erkek P(E)= %57, kadın P(K) = %43 olasılığa sahip olduğunu söyleyebiliriz. Diğer özellikleri de tahmin etme işlemine dâhil edersek Tablo 3.2 deki tercihleri yapan bir kişinin cinsiyetinin olasılığı

Tablo 3.2: Microsoft Naive Bayes test verisi

Dizi İzleme	Sinemaya Gitme	Maç Seyretme	Bilgisayarla Vakit Geçirme
Evet	Hayır	Hayır	Evet

P(E) olasılığı = 0,22 * 0,64 * 0,1 * 0,84 * 0,57 = 0,0638

P(K) olasılığı = 0,9 * 0,21* 0,82 * 0,89 * 0,43 = 0, 1379

olarak bulunur. Burada anlaşılacağı üzere aranan özelliklerdeki kişinin cinsiyetinin kadın olma olasılığı daha fazladır.

P(E) = 0,0638 /(0,0638 + 0, 1379) ≅ %31

P(K) = 0, 1379/(0,0638 + 0, 1379) ≅ %69

Kısaca eğer elimizde bir H hipotezi ve E olayı olma durumu varsa, H'nin gerçekleşme olasılığı denklem 3.5'de gösterildiği gibidir.

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)} \quad (3.5)$$

3.4.2 Microsoft Decision Trees (Microsoft Karar Ağaçları)

Karar ağaçları veri madenciliğinde kullanılan en popüler tekniklerden biridir. Hızlı eğitim süresinin olması, iyi derecedeki tahmin yeteneği ve kolay anlaşılır olması bu durumun en önemli sebeplerinden bazılarıdır. Microsoft Decision Trees algoritması sınıflama ve regresyon işlemlerini yapabilen hibrit güçlü bir karar ağaçları algoritmasıdır. Ayrıca birliktelik kurallarının izlenmesinde de kullanılabilmektedir. Aşağıdaki örnekte algoritmanın temel mantığının nasıl çalıştığı ve eğitim sırasında verdiğimiz durumlara göre ağaç yapısının nasıl oluştuğu anlatılmıştır.

Bir bankanın kredi verme durumlarıyla ilgili hayali bir veri Tablo 3.3'de gösterilmiştir.

Tablo 3.3: Microsoft Karar Ağaçları eğitim verileri

		Borç Miktarı		Gelir Düzeyi	
		Çok	Az	Yüksek	Düşük
Kredi Verme Durumu	Hayır	700	300	500	500
	Evet	400	1600	1100	900

Bu verilerle eğitilen bir karar ağacı entropi (Shannon's Entropy) ve bayes değerlerine göre hangi özellik için bir karar düğümü oluşturacağını veya dala ayrılacak uygun bir değer olup olmadığını ortaya çıkarır[32]. Entropi değeri 3.6'da gösterildiği şekilde bulunur;

$$\text{Entropi}(p_1, p_2, .., p_n) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 \ldots - p_n \log_2 p_n \quad (3.6)$$

Buradaki p değerleri tahminde kullanılan her özelliğin olasılığını belirtir. Aralarında $p_1+p_2\ldots+p_n = 1$ bağıntısı vardır. Bu durumda "Borç Miktarı" ve "Gelir Düzeyi" durumlarının değerleri bulunur.

Borç Mikrarı = Entropi(700,400) + Entropi(300,1600) = 0,946 + 0,629 = 1,571

Gelir Düzeyi = Entropi(500,1100) + Entropi(500,900) = 0,896 + 0,941 = 1,837

Ağaç en düşük entropi değeri hangi değerdeyse ondan başlayarak dallara ayrılır. Burada Borç miktarı daha düşük olduğu için ilk başta bu değer üzerinde ilk ayrılma gerçekleşecektir. Bundan sonra sonraki değerler için ayrılan dal sayısı kadar işlem tekrarlanarak, ağaç oluşum şekli tamamlanır.

3.4.3 Microsoft Association Rules (Microsoft Birliktelik Kuralları)

Microsoft Birliktelik Kuralları, etkili bir birliktelik algoritması olan Apriori Birliktelik algoritmasını kullanır. Algoritmanın temeli eldeki veri kümesinde bulunan

değerlerin sayılarak birbirleri arasında ilişkilerinin belirtilmesidir. Bunun için öncelikle veri kümesinde bulunan tüm hareketin(örmek vermek gerekirse alışverişte satın alınan malların) listesini çıkarmaktır. Buna veri kümesi denir[33]. Alışverişte kullanılabilecek basit bir satış raporu Tablo 3.4'de gösterilmiştir. Bu değerler örnek veri kümesi olarak kullanılabilir.

Tablo 3.4: Birliktelik kurallarında kullanılabilecek veri kümesi

1. Alışveriş	2. Alışveriş	3. Alışveriş	4. Alışveriş	…
Gazoz, Mendil, Kek	Gazoz, Ekmek, Süt	Kek, Kola, Süt	Peynir, Et	…

İkinci aşamada birlikte satın alınan malların, tüm satın alınan mallar içerisindeki sayıları bulunmalıdır. Bu algoritmanın destek(support) değeridir. Bize tüm rapor içinden ilgili desenden kaç tane olduğunu söyler. Tablo 3.4 deki değerlere göre gazozun destek değeri

Destek({Gazoz}) = UrunleIlgiliKacTaneHareketOldugu(Gazoz) = 2

Burada incelenen tek bir tane üründür. Birden çok değere bakacak olunsaydı, aynı ürünlerin birlikteliğine göre sayım işlemi yapılacaktı.

Bu sayılar arasında bazı ürünler arasındaki ilişkilerin çıkarılması(rule belirlenmesi) olasılık hesaplarına göre olur. A ürününü alan birinin B ürününü de alma olasılığı(probability- confidence) denklem 3.7'de gösterildiği gibi hesaplanır.

Olasılık(A => B) = Olasılık(B|A) = Destek(A,B) / Destek(A) (3.7)

Birliktelik kurallarında veri kümesi, destek, olasılık hesapları algoritmanın çalışmasında bir durum daha kullanılır. Bu da önem(importance) değeridir. Bu değer veri kümelerini ve bunlar arasında bulunan ilişkileri ölçmeye yarar. Veri kümeleri için 3.8'de gösterilen formülüne göre önem değeri hesaplanır.

Önem({A,B}) = olasılık(A,B) / (olasılık(A) * olasılık(B)) (3.8)

İlişkiler(Kurallar-rules) için ise denklem 3.9'da verilmiştir[24].

Önem(A=>B) = log(p(B|A)/p(B|not A)) (3.9)

Kurallarda önemin 0 olması, A ve B arasında bir ilişki olmadığını belirtir. Eğer pozitif bir değer olursa A alındığı zaman, B'nin alınma olasılığı artar demektir. Negatif bir önem değerinde de ise bunun tersi geçerlidir.

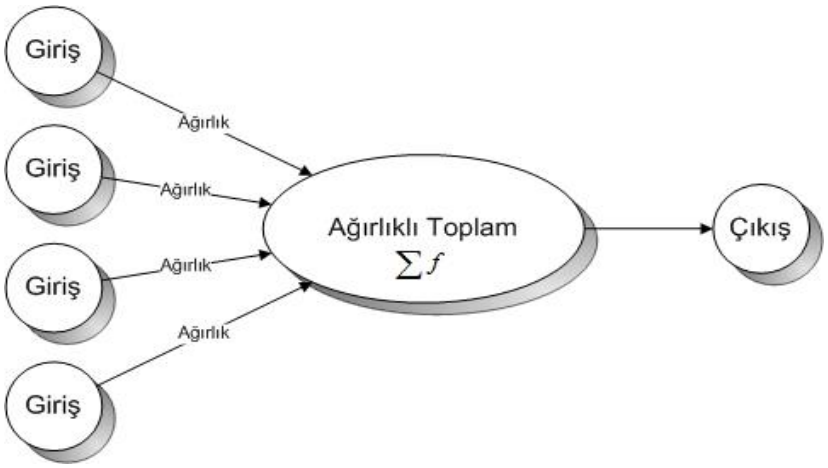
3.4.4 Microsoft Neural Network (Microsoft Sinir Ağları)

Microsoft Sinir Ağları(MSA) sınıflandırma ve regresyon ve birliktelik kurallarının incelenmesinde kullanılabilen bir algoritmadır[34]. Karar ağaçlarında olduğu gibi, MSA'da veriler arasında bulunan lineer olmayan ilişkileri bulabilir. Negatif özelliği ise eğitim süresinin diğer algoritmalara göre daha uzun sürmesidir.

MSA nöronlardan ve bu nöronlar arasındaki bağlardan(ağırlıklardan) oluşmuştur. Üç farklı tip nöron vardır; giriş, saklı ve çıkış nöronları. Bu nöronların birçok girişi ve bir çıkışı olur. Giriş nöronları sisteme giriş yaptığımız özellikleri belirtir(akademik ortalama, cinsiyet gibi). Saklı katman, giriş ve çıkış katmanları arasındaki nöronların bulunduğu gizli katmandır. Giriş katmanında; gelen değerlere, ilgili ağırlıklar ve katsayılara göre hesaplama işlemleri olur. Sisteme giren veriler çoğu zaman (0-1) aralığında kalan ondalıklı sayılardır. Çıkış katmanında ise saklı katmandan gelen değerlere göre tahmin işlerini yürütür. Çıkış katmanında birden çok nöron olabilir.

MSA ileri beslemeli(feed-forward) bir ağdır, geri beslemeli bir yapı kullanmaz. Bu sebeple giriş değerleri her zaman bir sonraki düğüme(nörona) aktarılır.

Yapay sinir ağları, canlılardaki sinir ağlarındaki sinyal iletimde olduğu gibi bir giriş ve çıkış değerine sahiptir. MSA'da nörona gelen tüm değerleri birleştirerek gerekli hesaplama işlemlerini yapar ve çıkış değeri üretir ve sonra aktivasyonu gerçekleştirir. Birleştirme işlemi için en popüler metot olan "ağırlıklı toplam" (weighted sum) hesaplaması kullanılır. Bu işlemde nörona gelen her değerin, geliş yönündeki ağırlığa göre çarpma işlemi gerçekleştirerek toplaması gerekir(Şekil 3.5).



Şekil 3. 5: Örnek nöron gösterimi

Aktivasyon fonksiyonu olarak iki yapı kullanılmaktadır; tanh ve sigmoid. Bu yapılar ağın öğrenme yeteneğini geliştirerek sisteme lineer olmayan özellikler kazandırır. Bu denklemler 3.10 ve 3.11'de belirtilmiştir.

$$\text{sigmoid} : {}^{1}/_{(1 + e)^{a}} \tag{3.10}$$

$$\tanh : (e^{a}\text{-}e^{-a}) / (e^{a}\text{+}e^{-a}) \tag{3.11}$$

Temel olarak MSA'da algoritmanın işleyişi şu şekildedir; Tüm nöronlara (0-1) aralığında belli bir başlangıç değeri verilir. Gelen değerler o anda bulunan nöronlar arası ağırlık değerlerine göre hesaplama işlemlerini gerçekleştirir. En son olarak çıkış katmanı sigmoid ve tanh fonksiyonlarına göre bir sonuç üretir. Algoritmadaki kilit nokta ise o anki ağırlıklara göre bir çıkış üreten sistemin aslında ne kadar başarılı olduğunun ölçülmesidir. Eğitim verisinden yararlanılarak sistemin o andaki çıkış ile gerçek değer arasında karşılaştırma yapılır ve hata fonksiyonu[error function-loss function] yardımıyla sistemin olması gereken değerden ne kadar farklı olduğu hesaplanır. Hesaplanan bu veriye göre de sistemdeki tüm diğer düğümler ve ağırlıklar tekrar düzenlenir. MSA'daki bu kilit duruma geri yayılım(backpropagation) denir. Hata denklem 3.12'de gösterildiği gibi hesaplanır[35];

$$\text{Hata}_{i} = O_{i}(1 - O_{i})(T_{i} - O_{i}) \tag{3.12}$$

Burada T_i sonucun olması gereken asıl değeri, O_i ise nöronun o andaki çıkışını belirtmektedir. Ara katmandaki nöronların hata değeri hesabı 3.13'de gibi olur.

$$\text{Hata}_i = O_i(1 - O_i) \sum j \ \text{Hata}_j w_{ij} \quad (3.13)$$

O_i , saklı katmandaki i nöronunun çıkışı, j ise bir sonraki katmana giden çıkışları belirtir. Hata_j, ilgili nöronun hata değerini ve w_{ij} de, i ve j de nöronlar arası ağırlık hesabını belirtir.

Bu hata değerlerine göre tüm düğümlerdeki ağırlıklar, aşağıdaki metodu kullanarak tekrar düzenlenir(3.14).

$$w_{ij} = w_{ij} + 1 * O_i * \text{Hata}_j \quad (3.14)$$

Buradaki "1" değeri öğrenme hızı belirtir. Eğer değer "1"den küçükse değişiklik daha küçük olur ve öğrenme daha yavaş gerçekleşir, "1"den büyük olursa istenen hedefe doğru daha hızlı sıçrama işlemini yapabilir. Bu hata değerlerini geriye yansıtırken, her durumdan sonra ağ ağırlıklarını değiştirmek yerine, tüm eğitim verileri bittikten sonra toplu olarak güncelleme işlemi yapar.

MSA algoritmasında tek bir saklı katman kullanarak sonuca gitmeyi çalışılır. Saklı katmanda kullanılacak nöron sayısı da 3.15'deki formülden hesaplanır;

$$\text{Nöron Sayısı} = c * \sqrt{m * n} \quad (3.15)$$

Buradaki m değeri giriş yapan, n ise çıkışta kullanılan nöron sayısını belirtir. "c" değerini ise Microsoft standart bir değer olarak 4 almıştır.

3.5 DMX (Data Mining Extensions)

Veri Madenciliği ile ilgili birçok farklı program olması ve tüm bu programların çalışabilmeleri için kendilerine özel farklı türde kod yazım standartı ve algoritmalara sahip olması, tekniği uygulayacak sistemlerin veri madenciliği konusunda entegrasyon konusunda zorluklar yaşamasına neden olmaktaydı. Bu nedenden Veri Madenciliği dünyasına ortak bir yapı ve söz diziminin gerekliliği ortaya çıkmıştır. DMX(Data Mining Extensions) bu sebeple Veri Madenciliği işlemlerini ortak bir dil altında yapabilmek için 1999 yılında ortaya çıkarılan bir standarttır[35]. Tez kapsamında

Microsoft SQL Server 2008 Analysis Services kullanıldığı için, çalışmada işlem yapılırken DMX sorguları kullanılmıştır.

Normal bir OLTP sisteminde nasıl veri sorgulamak için SQL(Structured Query Language) veya OLAP işlemleri için MDX(Multi-Dimensional Expressions) kullanılıyorsa Veri Madenciliği işlemleri için de DMX kullanılır. Yazımı kolay anlaşılabilmesi için SQL'e benzetilmiştir ama yinede kendisine özel yazım kuralları ve yapıları vardır. DMX yardımıyla program ara yüzleri üzerinden yapabildiğimiz Veri Madenciliği yapısı(Structure) ve modeli oluşturma, güncelleme ve silme, ayrıca model gerçekleştirildikten sonraki yapılan tahmin(predict) işlemlerini kod bazında yapabilmeyi sağlar.

3.6 Veri Madenciliğinin Eğitim Alanında Kullanımı

Daha önce belirtildiği gibi bilgi teknolojilerinin ve bilgisayar sayısının artması, bununla beraber elde edilen verilerin saklanabilirliğin çok yüksek düzeylere çıkması büyük veri yığınlarının ortaya çıkmasına sebebiyet vermiştir. Eldeki ham verinin bize yarar sağlayacak bilgiye dönüştürülmesi veri madenciliği süreciyle gerçekleşir. Bunun için veri madenciliği birçok bilim dalında, onlarca farklı konuda kullanılmış ve bu konular üzerinde incelemeler yapılmıştır. Bu sebeple burada tezin amacı olan öğrenci, eğitim ve akademik başarı olduğu için, bu alanlarda yapılan çalışmalardan örnekler verilmiştir.

Dimitris ve Chritos[36] 2006 yılında, Hellenic Open Universitesi'ne kayıtlı olan uzaktan eğitim öğrencilerinin final sınavlarındaki başarılarını tahmin etmeye çalışmışlar, bunun içinde ev ödevlerini ölçüt olarak kullanmışlardır. Analiz işleminde karar ağaçları ve genetik algortitma kullanmışlardır.

Bozkır ve diğerleri[27] 2008 yılında, ÖSYM'nin ÖSS'ye giren öğrenciler üzerine yaptığı sosyal, eğitim hayatları ve ebeveyn eğitimleri ile ilgili anket çalışması üzerinde sınav başarılarını etkileyen faktörleri tespit etmişlerdir. Karar ağaçları algoritması kullanarak analiz işlemleri yapmışlardır. ÖSS puan türlerine göre başarıyı etkileyen faktörleri belirlemişlerdir.

Erdoğan ve Timor [26] 2005 yılında, Maltepe Üniversitesi öğrencilerinin, üniversiteye giriş sınavında yer aldığı yüzdelik dilim ile üniversite akademik başarıları arasında ilişki olup olmadığı araştırılmıştır. Kümeleme algoritmalarıyla analiz işlemi gerçekleştirilmiştir. Burslu öğrencilerin daha çok bulunduğu Fen Edebiyat Fakültesi ortalamalarının daha yüksek, İktisadi ve İdari Bilimler ve İletişim Fakültesi öğrencilerinin daha düşük ortalamalara sahip olduğu belirlenmiştir.

Ayık ve diğerleri [19] 2007 yılında, Atatürk Üniversitesinde okuyan öğrencilerin lise türü ve lise mezuniyet derecelerine göre üniversitede de kazandıkları bölümler arasında ilişki incelenmiştir. Sınıflama algoritmaları yardımıyla üniversite giriş sınavında yüksek puanlı öğrenci alan fakülte öğrencilerinin lisede daha başarılı oldukları gözlenmiştir.

Al-Radaideh ve diğerleri [37] 2006 yılında, karar ağaçları yardımıyla öğrencilerin üniversitede aldıkları derslerin başarısını etkilen faktörleri bulmaya çalışmıştır. Bu amaçla örmek olarak seçilen bir dersi alan öğrencilerin dönem sonu final notları tahmin edilmeye çalışılmıştır.

Karabatak ve İnce'nin [30] 2004 yılında yaptıkları çalışmalarında, Fırat Üniversitesi Teknik Eğitim Fakültesi Elektronik ve Bilgisayar Eğitim öğrencilerinin aldıkları matematik, fizik, kimya, Türk dili ve Atatürk İlkeleri Ve İnkılâp Tarihi gibi derslerin notları arasında apriori algoritmasıyla ilişkiler aranmış ve bulunan birliktelik kuralları listelemiştir.

Güneri ve Apaydın'ın [38] (2004) çalışmalarında, Gazi Üniversitesi Ticaret ve Turizm Eğitim Fakültesi öğrencileri üzerinde program, cinsiyet, lise ortalaması, ÖSS puanı gibi parametrelere göre akademik başarı tahmini yapılmaya çalışılmıştır. Bunun için yapay sinir ağları ve lojistik regresyon algoritmaları kullanılmış ve performans karşılaştırması yapılmıştır.

Ho Yu ve diğerleri [23] 2010 yılında, Arizona State University'deki öğrencilerin ne tür nedenlerle okula devam etmekten vazgeçtiklerini sınıflama ağaçları(Classification Trees), MARS(Multivariate adaptive regression splines) ve yapay sinir ağları yardımıyla bulunmaya çalışmışlardır. Etnik köken ve ikametgâh gibi kavramların, üniversiteye devam kararında lise başarısından daha önemli olduğu bulunmuştur.

Vandamme ve diğerleri [28] 2007 yılında, üniversiteye yeni kayıt yaptıran öğrenciler üzerinde akademik başarıyı en çok hangi parametrelerin etkilediğini bulmaya çalışmışlardır. Öğrenciler düşük-orta-yüksek risk grubu şeklinde sınıflandırarak akademik yaşantılarında uygun rehberlik sürecinin sağlanması ve öğrenci gelişiminin desteklenmesinin amaç edinildiği çalışmada aynı zamanda diskriminant analizi, sinir ağları ve karar ağaçları kullanarak akademik başarıları tahmin edilmeye çalışılmıştır.

Özçınar'ın [22] 2006 yılındaki çalışmasında, Pamukkale Üniversitesi Eğitim Fakültesi Sınıf Öğretmenliği programındaki öğrencilerin derslerde aldıkları notlar, akademik ortalamaları ve öğretim türleri yardımıyla KPSS sınav başarıları tahmin edilmeye çalışılmıştır. Tahmin işleminde, yapay sinir ağları ve regresyon analizi kullanılmıştır.

Kovačić'ın [39] 2010 tarihli çalışmasında Yeni Zelanda Open Polytechnic'de bulunan öğrencilerin yaş, cinsiyet, etnik kimlik, ders ve program gibi birçok parametreye göre başarıyı tahmin etmeye yönelik çalışmasında CART algoritması kullanılarak sınıflandırma ve tahmin işlemleri gerçekleştirilmiştir.

Bozkır ve diğerleri [40] 2008 yılında, Hacettepe Üniversitesindeki lisans öğrencilerinin eğitim amacıyla internet kullanımıyla ilgili sonuçları karar ağaçları, kümeleme ve birliktelik kuralları ile incelemişlerdir. Bilgisayar kullanma konusunda teknik altyapının, üniversitede bilgisayarla ilgili ders vermekten daha etkili olduğu görülmüştür.

Yang'ın [41] 2006 yılındaki, Texas Woman's University'deki yıllara göre kayıtları inceleyerek, veriler arasından anlamlı desenler ve ilişkiler çıkarılmaya çalışılan çalışmasında kümeleme analizi ile öğrenci sürekliliği incelenmiştir.

4. TEZ KAPSAMINDA KULLANILAN ÜRÜNLER

4.1 Giriş

Binlerce öğrencinin kayıtlı olduğu zengin verilere sahip üniversiteler sistemlerinde kayıtlı bulunan öğrencilerin durumlarını incelemek üzere veri analizlerine ihtiyaç duyarlar. Bu nedenle bu kuruluşların en büyük ihtiyaçlarından bir tanesi raporlamadır. Bu raporlar sayesinde öğrenci ve programların akademik başarı durumları gözlenebilir.

Tüm verilerin veritabanlarında tutulduğu çağımızda verilerin analizi ve üniversite içerisinde alınacak önemli stratejik kararların alınması bu veritabanları üzerinde gerçekleştirilecek yapılara bağlıdır. Veri madenciliği, bu stratejik kararlar almaya yardımcı olacak iyi tahminler çıkarabilir ve karar destek sistemi oluşturmasına yardımcı olabilir.

Bütün bu bilgiler göz önünde bulundurularak eldeki tez kapsamında, veri ambarı ve veri madenciliği teknolojileri kullanılarak üniversitede bulunan öğrenciler üzerinde akademik başarı ve ders başarıları hakkında birçok kritere göre tahmin işlemi yapılması sağlanmaktadır.

Bu veri toplama, veri analizi ve raporlama işlemler için ilk başta verilerin uygun bir ortamda saklaması ve sorgulanması gerekmektedir. Bu kapsamda verilerin ilk olarak ham olarak bulunduğu OLTP yapısı, verilerin dönüştürülüp saklandığı veri ambarı, veri madenciliği algoritmasının çalıştığı analiz bölümü ve işlemlerin sonuçlarının gösterildiği raporlama ara yüzüyle ilgili kullanılan yazılımlara ait bilgiler aşağıda özet olarak verilmiştir.

4.2 Microsoft Visual Studio 2008

Microsoft firması tarafından sunulan, Windows, Web ve Veri Madenciliği uygulamaları geliştirilmesini sağlayan yazılım platformudur. .NET çatısı, yazılım geliştiriciler için zengin Windows Forms uygulamalarının geliştirilebilmesi için temiz, nesneye dayalı ve genişletilebilir sınıf kümelerini sunar. Uygulamalar, çok katmanlı dağıtık çözümlerde yerel kullanıcı arayüzü gibi davranır[6].

4.3 Microsoft SQL Server 2008

Microsoft SQL Server 2008 Veritabanı Yönetim Sistemi, ilişkisel veri tabanı sistemidir. Express, Express Advanced, Web, Workgroup, Standard, Enterprise ve Development sürümleri mevcuttur. Verilerin tutulduğu tablolar, belirlenmiş olan sınırlamalara bağlı kalarak diğer tablolar ile kurulmuş ilişkiler üzerinden veri bütünlüğünü ve doğruluğunu sağlamaktadır. Veri tabanı, sistem, kullanıcı ve kontrol bilgilerinin tutulduğu fiziksel ve mantıksal yapıları içerir[6]. Tez kapsamında oluşturulan veri ambarı, SQL Server 2008 Development sürümü üzerinde yer almaktadır.

4.4 Microsoft Analysis Services 2008

SQL Server 2008'in en önemli servislerinden biri olan SQL Server Analysis Services, karar destek motorunun ve araçlarının yer aldığı ortamdır. Karar destek mekanizmasına ait iki içerik olan Veri Madenciliği ve OLAP bu ürün kapsamında desteklenmektedir.

Analysis Services mimarisi istemci ve sunucu bölümleri olmak üzere ikiye bölünebilir. İstemci bölümü, uç kullanıcılar için arayüz desteğini sağlarken sunucu

bölümü, istemci servislere işlevsellik ve güç veren motorların çalışmasını sağlar. OLTP ile doğrudan bağlantı kurarak çalışabildiği gibi birçok veri tabanı sistemine ODBC üzerinden bağlanabilmektedir[6].

4.5 Microsoft Reporting Services 2008

SQL Server Reporting Services (Raporlama Servisleri), SQL 2000 ile birlikte bir ek olarak sunulmuş olup SQL Server kaynakları üzerinde raporlama yapmak için kullanılır. SQL 2008 ile birlikte daha da olgunlaşan bu ürün piyasadaki diğer araçlara göre yapılandırılması ve kullanımı kolay olan bir araçtır. Reporting Services, SQL Server 2008′in bir parçası haline getirilmiş olup SQL Server içerisinde bir servis olarak sunulur. Bu aracı kullanarak veri kaynaklarındaki sorguların sonuçları XML, CSV, TIFF, EXCEL, PDF, Tek dosyalı Web sayfası, Word formatında dışarı verilebilir. Bu aracın en güzel yanı Microsoft tabanlı birçok ürünle ilişkili çalışıyor olması ve .NET uygulamalarında gömülü olarak kullanılıyor olmasıdır[6].

5. ÖĞRENCİ KARAR DESTEK SİSTEMİ

5.1. Giriş

Veri ambarı bir ya da daha fazla veri kaynağından (MS SQL Server, Oracle, Access gibi) gelen verilerin temizlenerek özetlendiği depo alanıdır. Bu nedenle, veri ambarına hangi kaynaklardan veri taşınacağını belirlememiz gerekir. Tez kapsamında öğrenci, program ve üniversite ile ilgili gerekli bilgilerin Pamukkale Üniversitesi Bilgi İşlem Daire Başkanlığı sunucuları üzerinden çalışan Öğrenci İşleri Otomasyon Sistemi(ÖİOS) veri tabanı, veri kaynağı olarak alınmıştır. Bu sistem 2008 yılında geliştirilen bir sistem olup farklı veri kaynaklarına sahip birçok otomasyonu tek bir veritabanı altında toplamıştır. Bu sayede öğrencinin üniversiteye ilk kaydından üniversiteden ayrılışına kadar geçen sürede elde edilen her türlü nüfus bilgileri, anket girdileri, ve ders kayıt bilgilerine ait tüm veriler tek bir veritabanı altında toplanmıştır. Aktif olarak işlem yapılan bu OLTP sistemi, Microsoft SQL Server 2008 veritabanı yönetim sistemini kullanmaktadır.

ÖİOS için öğrenci, dersler ve akademik başarı ile ilgili birçok rapor hazırlanmıştır. Ama bu raporlar belli bir kişi veya topluluk(program veya akademik birim) temelli raporlardır. Sadece sorgulanan değer ile ilgili sonuç sağlayabilirler. Örneğin bu raporlar; sorumlu olduğu derslerin notlarını göstermesi ve akademik ortalamayı hesaplaması, programdaki tüm öğrencilerin akademik ortalamaya göre başarı sırasının oluşturulması veya belli bir dersteki öğrencilerin not durumlarının incelenmesi gibi raporlardır. Ama bu tezle amaçlanan daha geniş veri grupları içerisinde aynı özellikleri paylaşan kişilerin durumlarının belirlenmesi ve gelecekte benzer durumlarla işlem yapmaya çalışıldığında bize eldeki analiz verilerinden çeşitli raporlar sunarak sorgulanan konuda karar almamızı kolaylaştırıcı bilgiler sunmasıdır. Aynı zamanda ileri ki bölümlerde

görüleceği gibi kullanılan çeşitli algoritmalar aracılığıyla tahminlerde bulunarak eldeki verilerden bir sınıflandırma işlemi de yaparak mevcut verilerin durumu hakkında da bilgi vermektedir.

Bu çalışmada tahmin işlemlerinde en çok sorgulanan konu öğrenci başarısıdır. Üniversitelerde kaliteyi belirleyen en önemli parametrelerden biri an akademik başarı(AB) gerçekleştirilen beş modelin dördünde ana amaç olarak belirlenmiştir. Bunun yanında öğrencilerin ders başarıları da gerçekleştirilen modelle izlenmiştir.

Eldeki tez çalışması kapsamında, üniversitenin elinde lisans hakkı olması ve ayrıca mevcut veri kaynağının da SQL Server 2008 olması sebebiyle veri tabanı yönetim sistemi ve veri madenciliği algoritmalarını kullanmaya yarayan uygulamalardan Microsoft SQL Server 2008 ve Microsoft Reporting Service kullanılması kararlaştırılmıştır. Veri Madenciliği modelini gerçekleştirmek için de Microsoft Visual Studio 2008 kullanılmıştır.

Uygulamanın ana hedef kitlesi, stratejik kararları almakta yardımcı olması için tasarlandığından üniversite yönetimi olmuştur. Ama raporlar özelleştirilerek öğrencilerin kendilerini ilgilendiren tarafların görüntülenmesi sağlanabilir. SQL Server 2008 üzerinden gelen verileri Analiz Hizmetleri(Analysis Services) kullanılarak veri madenciliği algoritması gerçekleştirilmiş ve kullanıcılara rapor olarak sonuçların gösterilmesi amaçlanmıştır. Bu bölümde ÖİBS'nin yapısı ve uygulamanın nasıl geliştiği anlatılacaktır.

5.2. Öğrenci İşleri Otomasyon Sistemi Yapısı

Üniversite bünyesinde kullanılmakta olan Öğrenci işleri Otomasyon sistemi birçok modülden oluşmaktadır. Bu modüller

- Öğrenci Ön Kayıt Modülü
- Anket İşlemleri Modülü
- Ders Şubesi Açma ve Görevlendirme Modülü
- Ders Kayıt Modülü
- Yönetim Kurul Kararları İşlemleri Modülü
- Not Giriş Modülü

- Geçici Ders Kayıt Modülü
- Yaz Okulu Ders Kayıt Modülü
- Ders Eğitim Planı Hazırlama Modülü

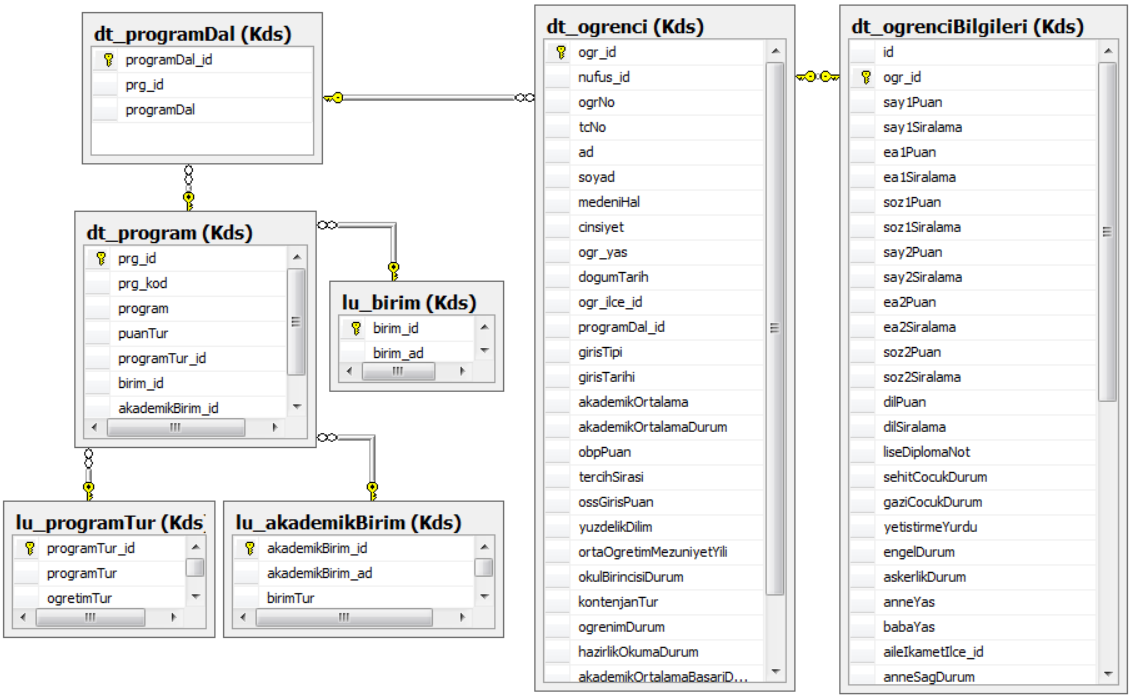
Şu an kullanılmakta olan otomasyon sistemi 2008 yılında ilk defa birinci sınıf öğrencilerine hizmet vermeye başlayarak kullanıma açılmıştır. 2009-2010 güz dönemiyle beraber eski öğrenci işleri otomasyonundaki verilerde yeni sisteme aktarılmış ve tek bir sistem üzerinden işlemlerin yapılması sağlanmıştır. Sistemin öğrenciler, öğrenci işleri personeli ve akademisyenler olmak üzere çok geniş bir kullanım kitlesi vardır. Sistem üzerinde öğrenciler not, transkript izleme, ders kayıt ve anket doldurma işlemlerini, öğretim elemanları not giriş ve öğrenci bilgileri izleme işlemlerini, öğrenci işleri de raporlamalar, ders şubeleri açma-görevlendirme, yönetim kurul kararı işlemlerini ve burada sayılamayacak daha birçok işlevi yürütmektedirler.

ÖİOS'ine her dönem milyonlarca yeni kayıt eklenmektedir. Bu kayıtlar çok çeşitlilik arz etmektedir. Örnek vermek gerekirse üniversiteye yeni kayıt olma aşamasında öğrenciden elde edilen kişisel bilgilere ait veri (lise başarı puanları, aile yapısı, gelir durumları…), on binlerce öğrencinin ders kaydı ve dönem sonlarında bu derslere ilişkin öğrenci notları, başarı durumları ve benzeri gibi pek çok bilgi sisteme dahil edilmektedir. Bu veriler sadece belli sabit raporları almak için kullanılmaktadır. Örnek olarak transkriptte, öğrenci ders listesi ve akademik ortalaması hesaplaması yapılmaktadır. Bir başka raporda ise sabit bir şekilde öğrencilerin adları ve notları listelenerek ders listeleri oluşturulması amaçlanmıştır. Ama verinin bu şekilde sadece gösterim amaçlı kullanılması bize ekstra bir şey kazandırmamaktadır. Bu veri yığını içerisinden veri madenciliği algoritmaları yardımıyla anlamlı desenler çıkarılabilir. Bu sayede fazla kullanılmayan veri gruplarından anlamlı değerler elde ederek üniversitedeki öğrenci profili hakkında daha önce irdelenmemiş sonuçlara ulaşılabilir ve yararlı bilgiler çıkarılabilir.

Bilgiye erişmek için öncelikli olarak verilerin hazırlanması gerekir. Bu sebeple ilk olarak ÖİOS'nin kullandığı OLTP veritabanından gerekli verilerin çekilip, dönüştürülerek, analizin yapılacağı veri ambarına aktarılması gerekmektedir. Daha sonra analiz hizmetleri vasıtasıyla veri madenciliği algoritması kullanılarak desen arama, inceleme ve sağlama işlemleri yapılacaktır. Son olarak bulunan desenler karar destek sisteminde kullanılmak üzere raporlanacaktır.

5.3. Veri Ambarı Oluşturma

Veri madenciliğinde kullanılmak üzere gerekli olan veriler ÖİOS veritabanından gerekli görülen alanlar dönüştürüp alınarak hazırlanacaktır. Veri ambarında tablolar belli bazı özelliklerine göre gruplandırılmıştır. Bütün tablolar kds(karar destek sistemi) şeması altında bulunmaktadır. Tabloların başlangıcında bulunan ön ekler, tablonun lu(lookup table-tanım tablosu), dt(data table- veri tablosu) olduğu belirtmektedir. Şekil 5.1 de üniversiteye yeni kayıt yaptıran öğrencilerle ilgili diyagram gösterilmiştir.



Şekil 5.1**:** Yeni kayıt olan öğrencilerle ilgili diyagram

Burada bulunan tabloları açıklayacak olursak;

- dt_ogrenci: Öğrenci ile ilgili tüm genel bilgilerin tutulduğu tablodur.
- dt_ogrenciBilgileri: Yeni kayıt yaptıran öğrencilerin anket sonuçlarını tutan tablodur.
- dt_programDal: Öğrencilerin hangi program dala bağlı olduklarını belirten tablodur.

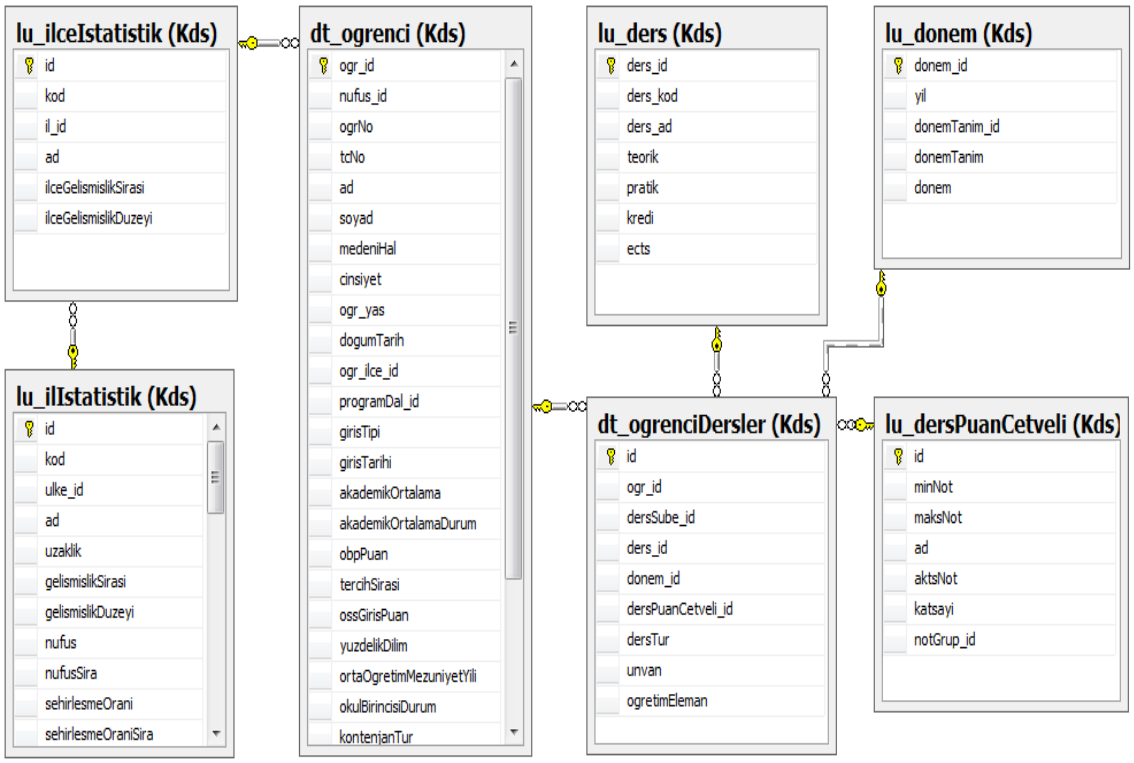
- dt_program: Öğrencinin bağlı olduğu programlar hakkında bilgileri içermektedir.

- lu_programTur: Öğrencinin lisans-önlisans ve normal öğretim-ikinci öğretim bilgilerini tutan tablodur.

- lu_birim: Programın hangi birime bağlı olduğunu belirten tablodur.

- lu_akademikBirim: Programın hangi akademik birime bağlı olduğunu getiren tablodur.

Şekil 5.2 de öğrencilerin kayıt yaptırdıkları dersleri belirten diyagram gösterilmiştir.



Şekil 5.2: Öğrencilerin aldığı tüm dersleri gösteren diyagram

Bazı genel tablolar(dt_ogrenci gibi) birden fazla diyagram içerisinde olabilir(Hem Şekil 5.1 ve hem de Şekil 5.2'de gösterimi bulunabilir). Bunları tekrar belirtmeden diğerlerinin genel açıklaması şu şekildedir;

- dt_ogrenciDersler: Öğrencilerin aldığı tüm dersleri listeleyen tablo.

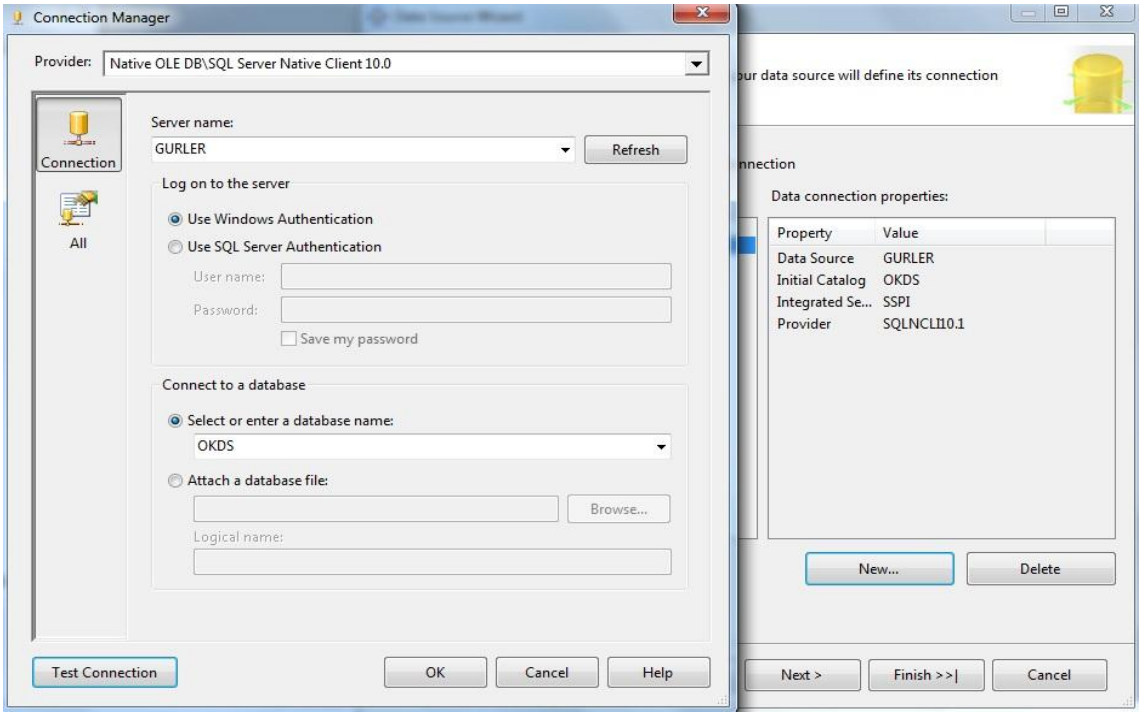
- lu_ders: Derslerin adları, kredileri ve diğer özelliklerinin tutulduğu tablo.

- lu_donem: Dersi hangi dönemde aldığını belirten tablo.
- lu_dersPuanCetveli: Derslerden hangi notları aldığını belirten tablo.
- lu_il: İllerin listesinin tutulduğu tablo.
- lu_ilce: İlçelerin listesinin tutulduğu tablo.

5.4. Veri Madenciliği Modeli Oluşturma ve Algoritmasını Uygulama

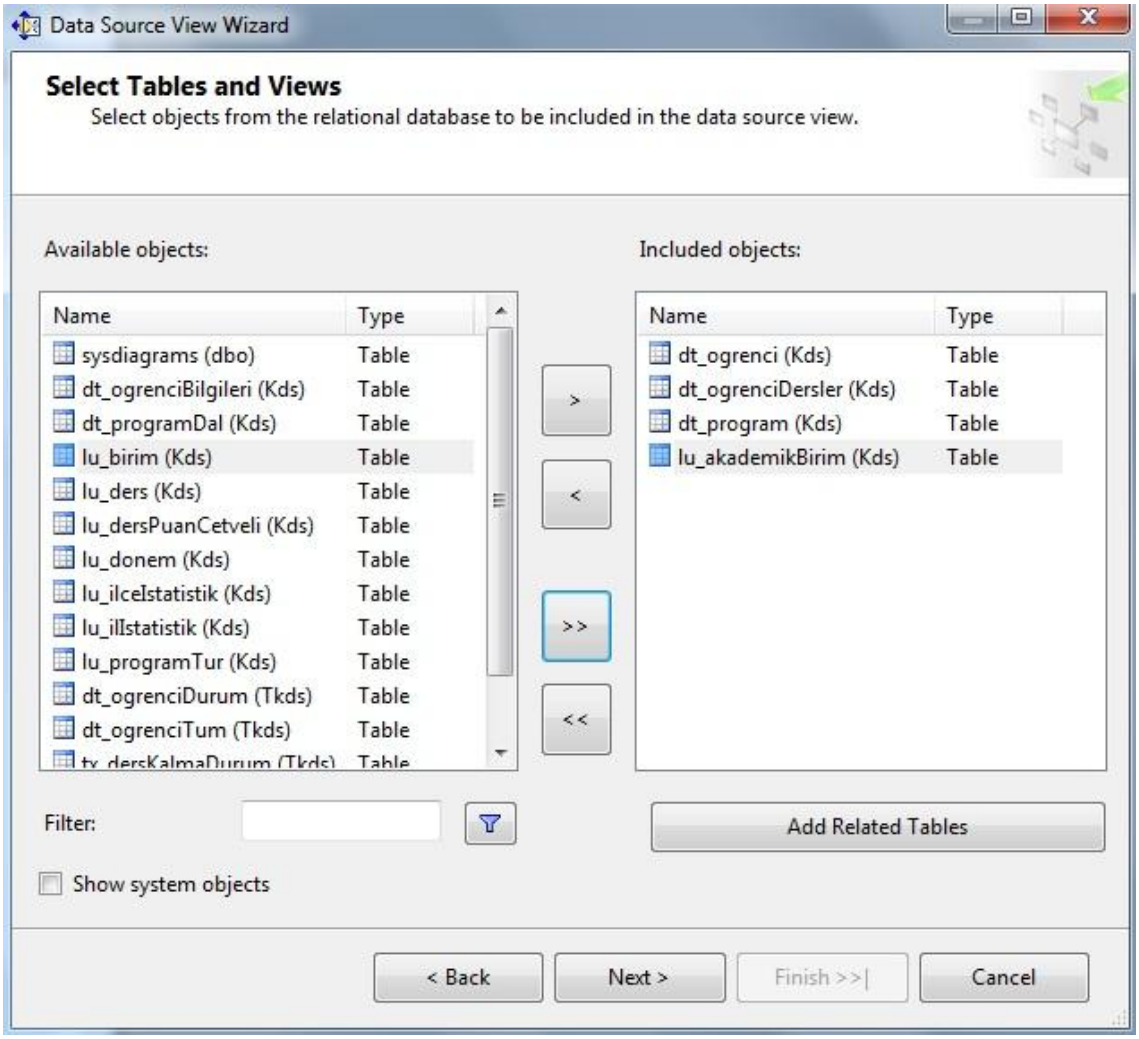
Veri Madenciliği ara yüzü olarak Visual Studio 2008 kullanılmıştır. Programa erişim ve gerekli tanımların yapılması algoritmaların gerçekleştirilmesi için gereklidir. Bu aşamalar, örnek olması açısında sadece tek bir defa açıklanacaktır.

Veri madenciliği modeli tanımlamak için Visual Studio 2008 üzerinden File kısmından "Analysis Services Database" i tıklayıp ilgili veri ambarını tıklayarak çalışması açılır. Burada ilk basamak, üzerinde işlem yapılacak veri ambarına bağlantı kurulmasıdır. Şekil 5.3 de işlem gösterilmiştir.



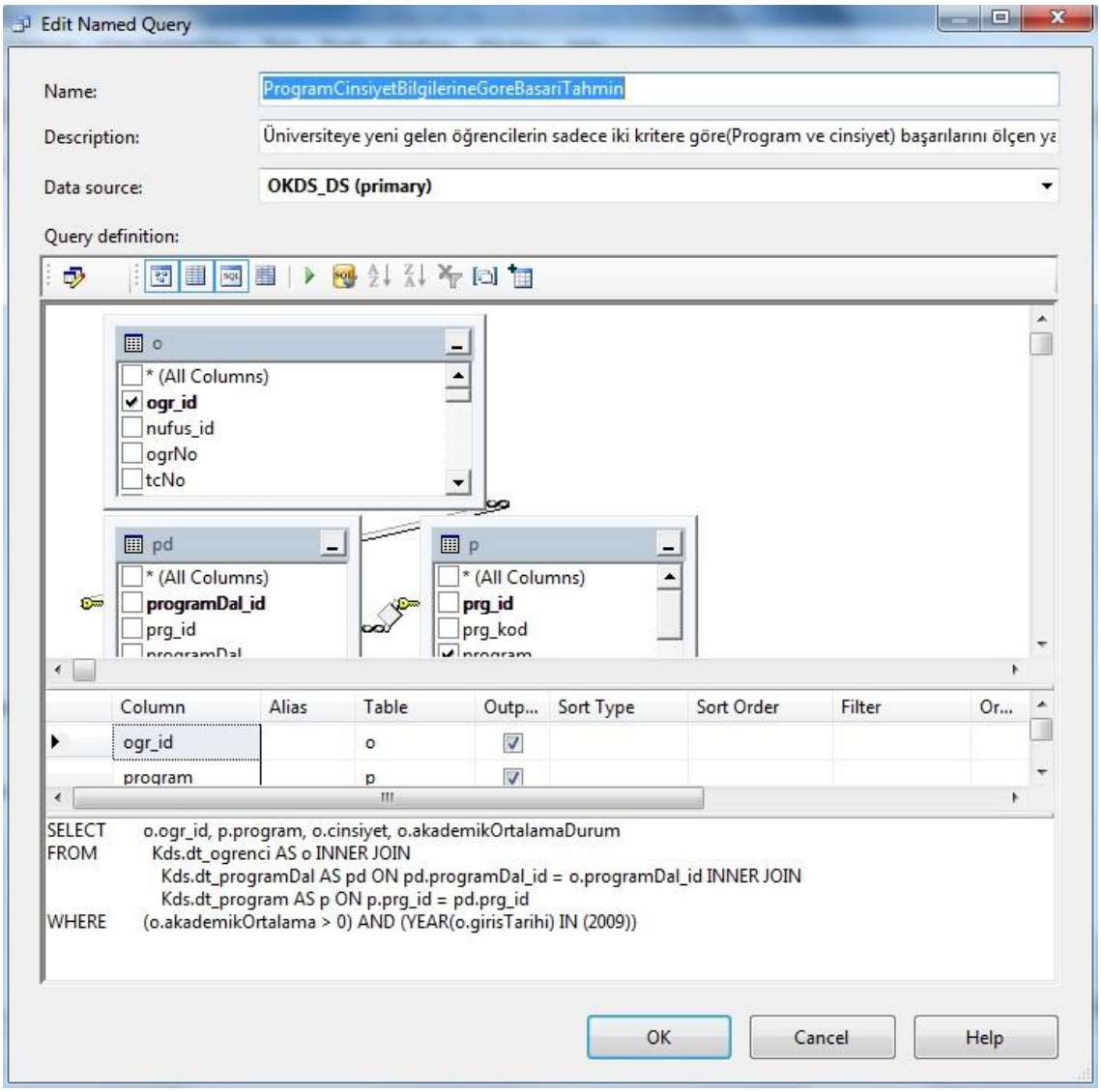
Şekil 5. 3: Analiz hizmetleri servisinden veri ambarına bağlantı kurma

Proje sunucunun bulunduğu bilgisayarda çalışıldığı için sunucu adı yerel bilgisayar adı ve bağlantı şekli "Windows Authentication" olarak belirtilmiştir. Veri kaynağı ile bağlantı sağladıktan sonra veri madenciliği algoritmalarını bağlantı sağlayacağı "Data Source View(DSV)" lar tanımlanması gerekir. Bu DSV'ler kurulan veri tabanı bağlantısından gelen verilerin gösterildiği bir ara yüz oluşturur. DSV tanımlaması Şekil 5.4'de gösterilmiştir.



Şekil 5.4: "Data Source View" kurma

DSV yapılacak tüm işlemlerin tutulduğu alandır. Yapılan çalışmada kullanılan tüm tablolar, ekstra tanımlanmış sorgu kümeleri ve hesaplanmış kolonlar burada tutulur. Veri üzerinde çalışabilmemiz için birçok isimli sorgu(Named Query) tanımlanmıştır. Şekil 5.5'de bu tanımlamanın nasıl yapıldığı gösterilmiştir. Kısaca isimli sorgular veritabanından veri çekmemize yarayan yapılardır.



Şekil 5.5: İsimli sorgu tanımlama

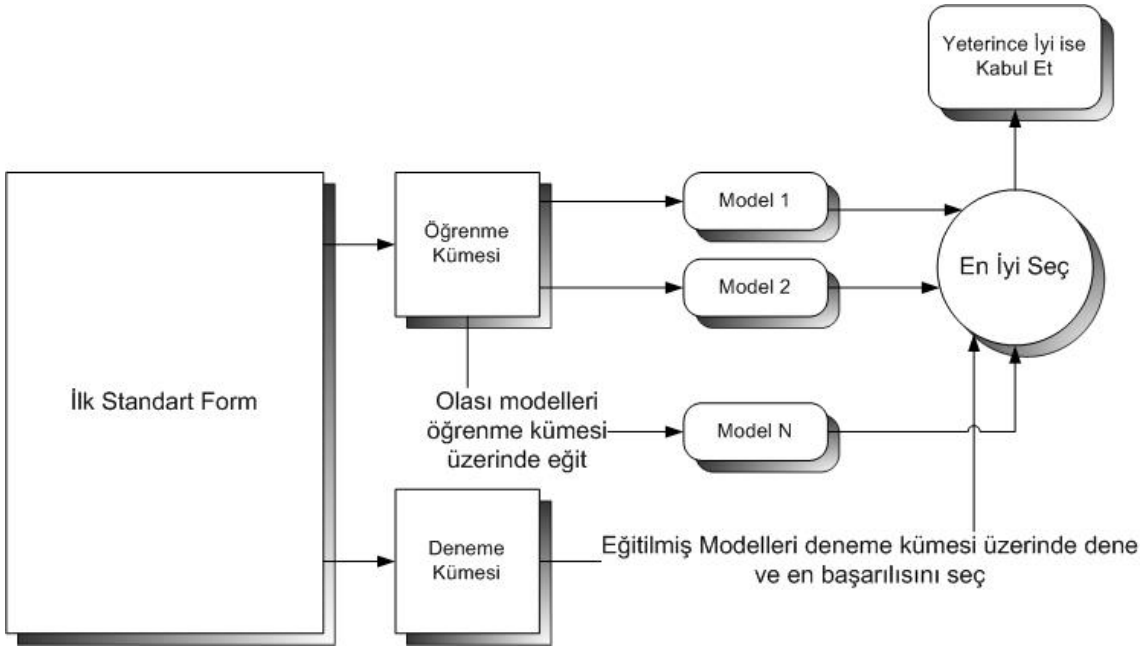
Veri madenciliği yapısı, veri kümesi içinden veri madenciliği algoritmasında kullanmak üzere seçilen kolonlardır. Veri madenciliği algoritmalarında desenler üzerinde tahmin sınıflandırma işlemleri yaparken veri ambarına aktardığımız tüm kayıtları kullanma zorunluluğu yoktur. Çok önemli olmadığı düşünülen bazı değişkenler yapıda yer alamazken, bazı önemsiz görünenler ön plana çıkabilir. Bu değişkenlerin kontrolünü direkt kullanıcıda yapabilir veya sistemin yol göstermesi sağlanabilir.

Çalışmada hedeflenen öğrenci başarı durumlarını önceden tahmin edebilmektir. Bunun için veri madenciliği algoritması olarak "Microsoft Karar Ağaçları", "Microsoft Naive Bayes", "Microsoft Birliktelik Kuralları", "Microsoft Sinir Ağları" kullanılmıştır. Akademik başarıların(AB) tahmin edilmeye çalışıldığı modellerde her zaman "Microsoft Karar Ağaçları", "Microsoft Naive Bayes", "Microsoft Sinir Ağları"

kullanılmıştır. Bu sebeple AB ölçümünün yapıldığı modellerde algoritmalar dendiğinde bu üç tip kastedilecektir. Ders başarılarının ölçüldüğü çalışmada ise "Microsoft Birliktelik Kuralları" kullanılmıştır.

Oluşturulan birden fazla veri madenciliği modelinde isimlendirmede bir standarda gidilmiştir. Tüm modeller model ana isim sonuna alt çizi ve algoritmanın baş harfleri gösterilecek şekilde düzenlenmiştir; modelAnaİsmi_AlgoritmaBaşHarfleri. Bu harfler DT (Decision Trees - Karar Ağaçları), NN(Neural Network – Sinir Ağları), NB(Naive Bayes) ve AR(Association Rules – Birliktelik Kuralları) algoritmalarını gösterir. Örnek verilirse "AkademikBasariTahmin_DT" modeli, karar ağaçları algoritmasıyla oluşturulan bir Veri Madenciliği modelidir. Model isimleriyle ilgili grafiklerde bu sebeple ayrıca bir açıklama yapılmayacaktır.

Birden fazla sayıda algoritma kullanılmasının sebebi, sabit bir modelin her problemde iyi performans vermemesidir. Veri Madenciliği modeli problemden probleme performans farkı göstermektedir. Bunun için yukarda bahsedilen algoritmalar her problem için sırasıyla uygulanmıştır. Test aşamasında bu algoritmaların başarısı test edilmiş ve raporlamada kullanılmak üzere en iyi performansa sahip model seçilmiştir. Şekil 5.6'de raporlamada kullanılacak modelin belirlenme yöntemi verilmiştir.



Şekil 5.6: En iyi algoritmayı bulma şeması[42]

5.4.1. Program ve cinsiyet'e göre akademik başarı tahmini

AB ölçümü bu tez çalışmasının en önemli amacını oluşturduğundan akademik başarı birçok farklı desen ve parametreye göre tahmin edilmeye çalışılmıştır. Birçok tahminsel (predictive) algoritmalar aynı zamanda sınıflandırma işlemi de yaptığı için üniversitenin mevcut durumu hakkında da bilgi sahibi olunmaktadır. Bu modelde amaç öğrencilerin genel bilgileri üzerinden akademik başarıları üzerinde bir sınıflandırma yapmak, hangi bölümlerde daha başarılı oldukları veya hangilerinde başarısız olduklarını gözlemleyebilmektir. Ayrıca minimum girdi sağlayarak, modelin tahmin başarısı ölçülecektir.

Bu modelin yapısını iki giriş ve bir tahmin parametresi oluşturmaktadır. Bu parametreler modelin eğitiminde kullanılmıştır. Tanımları ve alabileceği değerler Tablo 5.1'de verilmiştir.

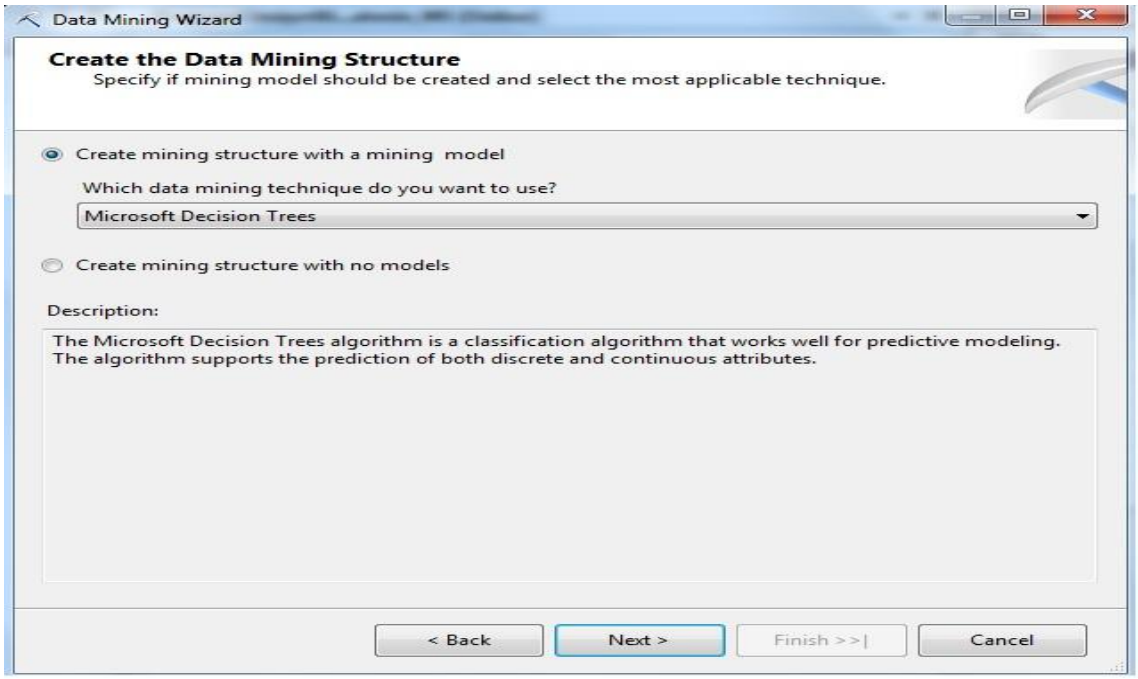
Tablo 5.1: Program ve cinsiyete göre AB tahmin modelinin giriş parametreleri

Cinsiyet	Cinsiyet alanı; Erkek, Kadın
Program	Üniversitede aktif olan programlar; Bilgisayar Mühendisliği, Fizik, Sosyoloji…
Akademik Başarı Durum	Tahmin Edilecek Değer. Akademik ortalamayı üç kısıma ayırarak tahmin işlemi yapılır. Başarısız (0 - 1,99), Başarılı (2 - 2,99), Çok Başarılı (3 ve üzeri) olarak ayırma gerçekleştirilmiştir.

Veri kümesi olarak 2009 yılında üniversiteye ilk defa kayıt yaptırmış öğrenciler kullanılmıştır. Test verisi olarak 3588 kişi modeli eğitmek amacıyla kullanılmıştır. Bu sistemde üç farklı Veri Madenciliği algoritması kullanılmıştır.

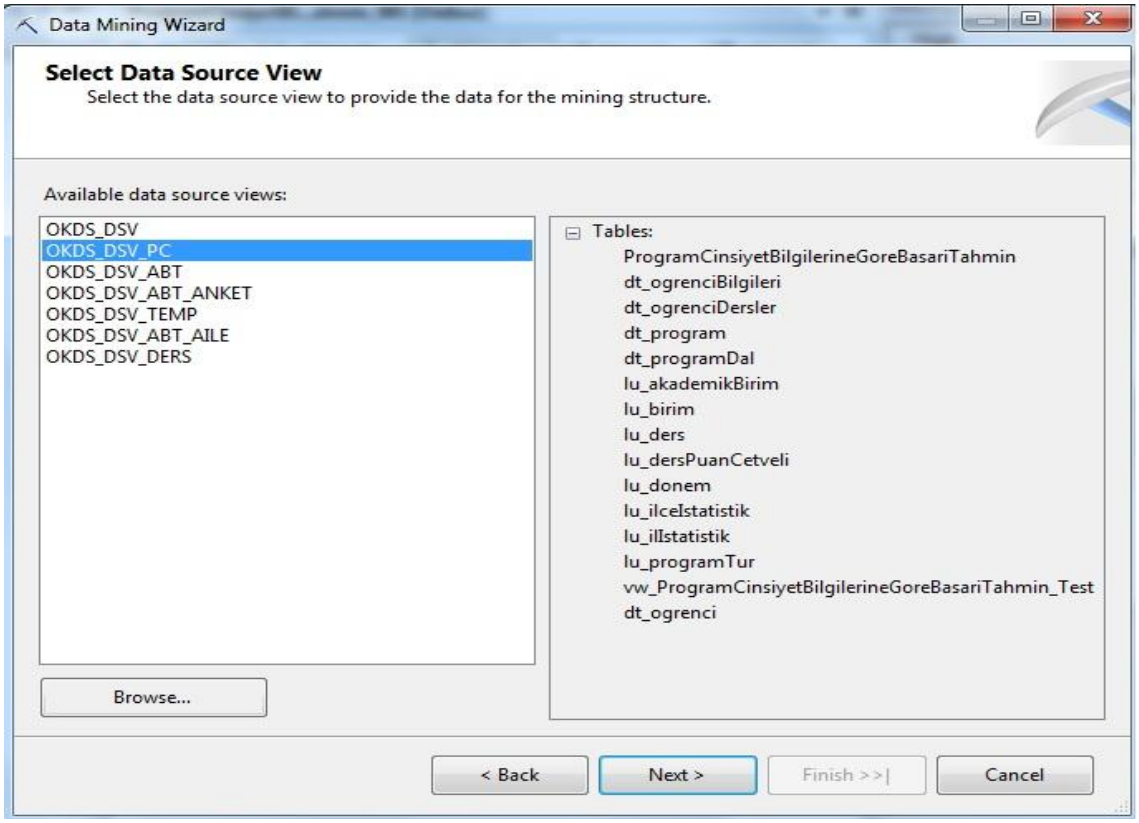
Bu modelde sistemin kurulma aşamaları adım adım anlatılmıştır. Diğer kullanılan modellerde tekrarı önlemek amacıyla bu ön bilgilere girilmeyerek direkt sonuç bilgileri gösterilecektir.

İlk olarak sol taraftaki "Solution Explorer"dan yeni bir veri madenciliği yapısı oluşturulması seçilir. Açılış mesajından sonra hangi algoritmanın kullanılacağı belirtilir(Şekil5.7).



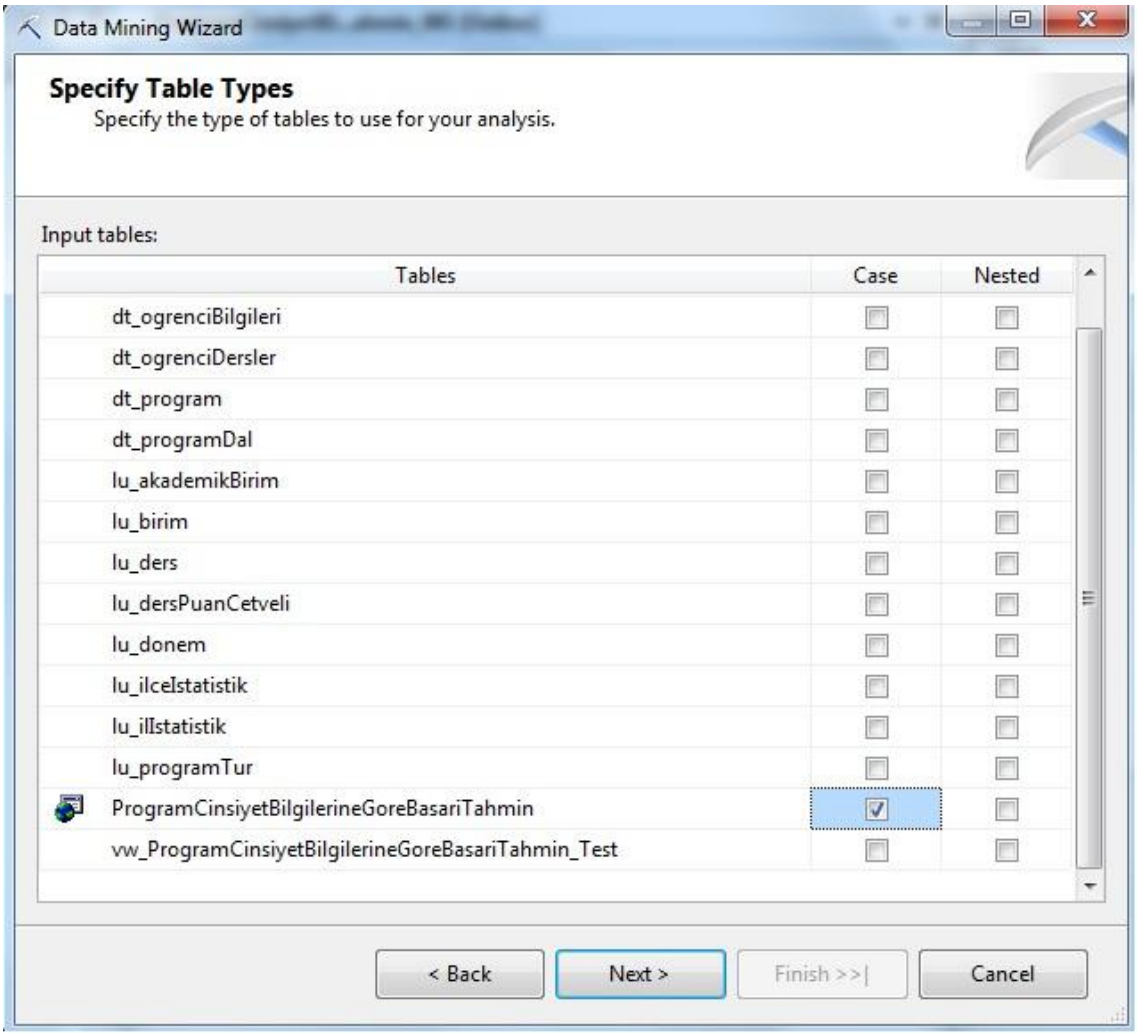
Şekil 5.7: Veri Madenciliği algoritmasını seçme

Üzerinde çalışılacak algoritma belirlendikten sonra hangi DSV üzerinde çalışacağı sorulur (Şekil5.8).



Şekil 5.8: Data Source View seçme

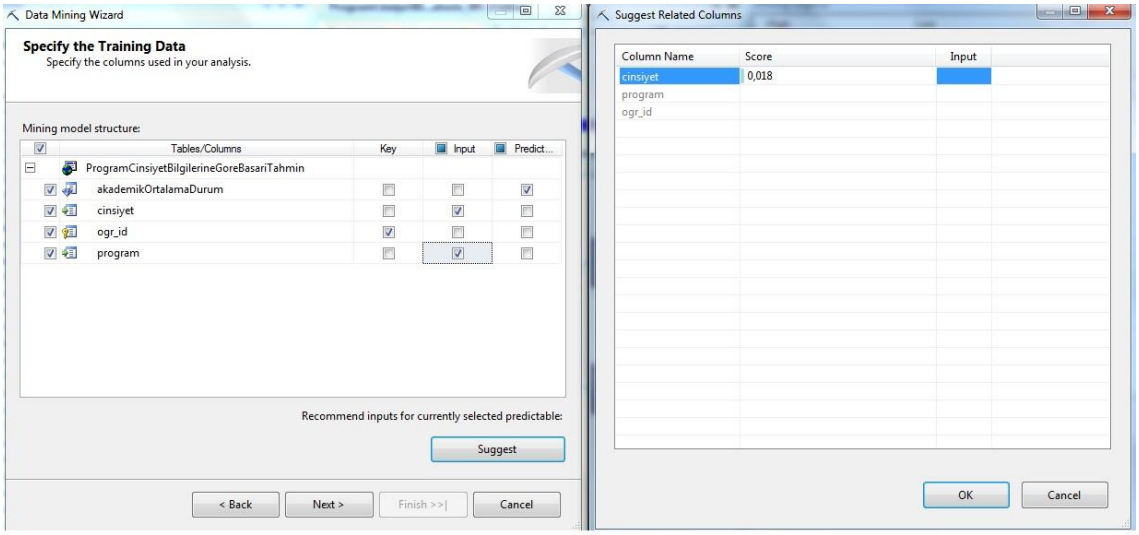
İşlem yapılacak tablolar veya hazırlanan isimli sorguların bu seçtiğimiz alanda tanımlı olması gerekir. Bu ekrandan sonra DSV üzerinde bulunan tablo dönüşlü sorgular ve tablolar ekrana gelir. Burada daha önce isimli sorgu olarak tanımlanan "ProgramCinsiyetBilgilerineGoreBasariTahmin" tablosu seçilir(Şekil 5.9).



Şekil 5.9: Modelde kullanılacak tabloyu belirleme işlemi

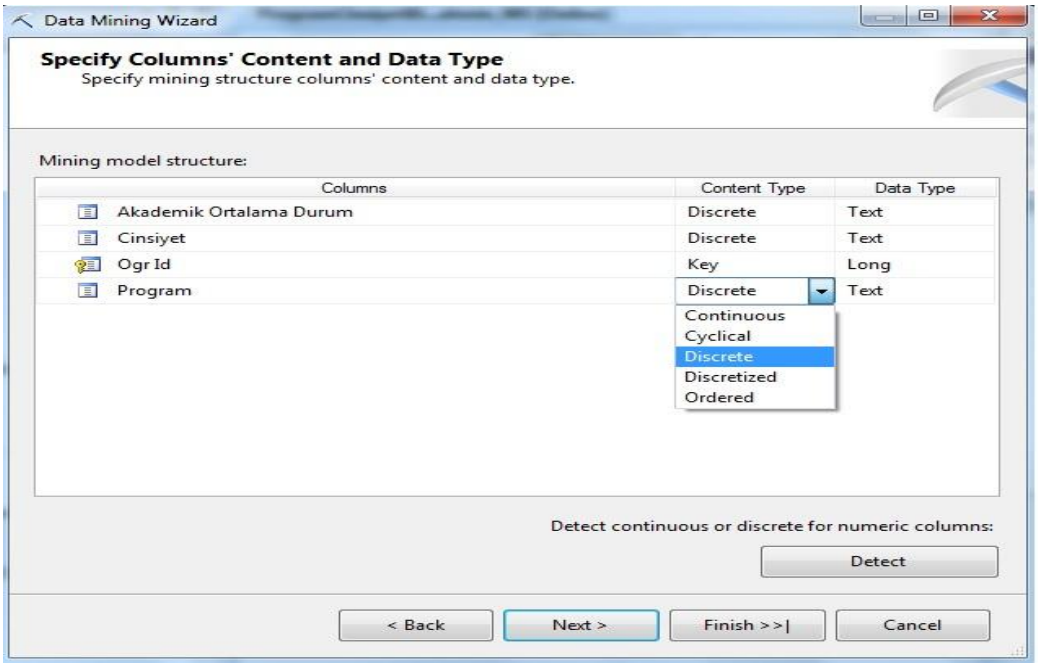
Bu ekrandaki "Case" ifadesi o tablonun ana tablo olduğunu belirtir. Sadece tek bir tablo "Case" olarak işaretlenebilir. İlk tablonun yanına ekstra tablo eklenmek istendiğinde bunlar "Nested" olarak işaretlenebilir. Bunun anlamı ana tablomuzdaki değerin diğer seçtiğimiz tablolarla ilişki olduğudur. Ama bu yapılan çalışmada sadece tek bir ana tablo işlem yapmaya yeterlidir. İleri denildiğinde seçilen tablonun içeriği gösterilecektir. Burada eğitilecek verinin hangilerinin olacağı, giriş değerlerinin neler olacağı ve eğer daha önceden tablolara birincil anahtar tanımlanmamışsa tablolara anahtar kolon tanımları yapılır. Veri kaynaklarından getirilen tüm sütunlar modeli

gerçekleştirmek için işe yarar olmayabilir. Bu durumda ya elle filtreleme işlemi uygulanır veya "Suggest" tuşuna tıklanarak sistemin mantıklı sonuçlar bulunması sağlanabilir(Şekil 5.10).



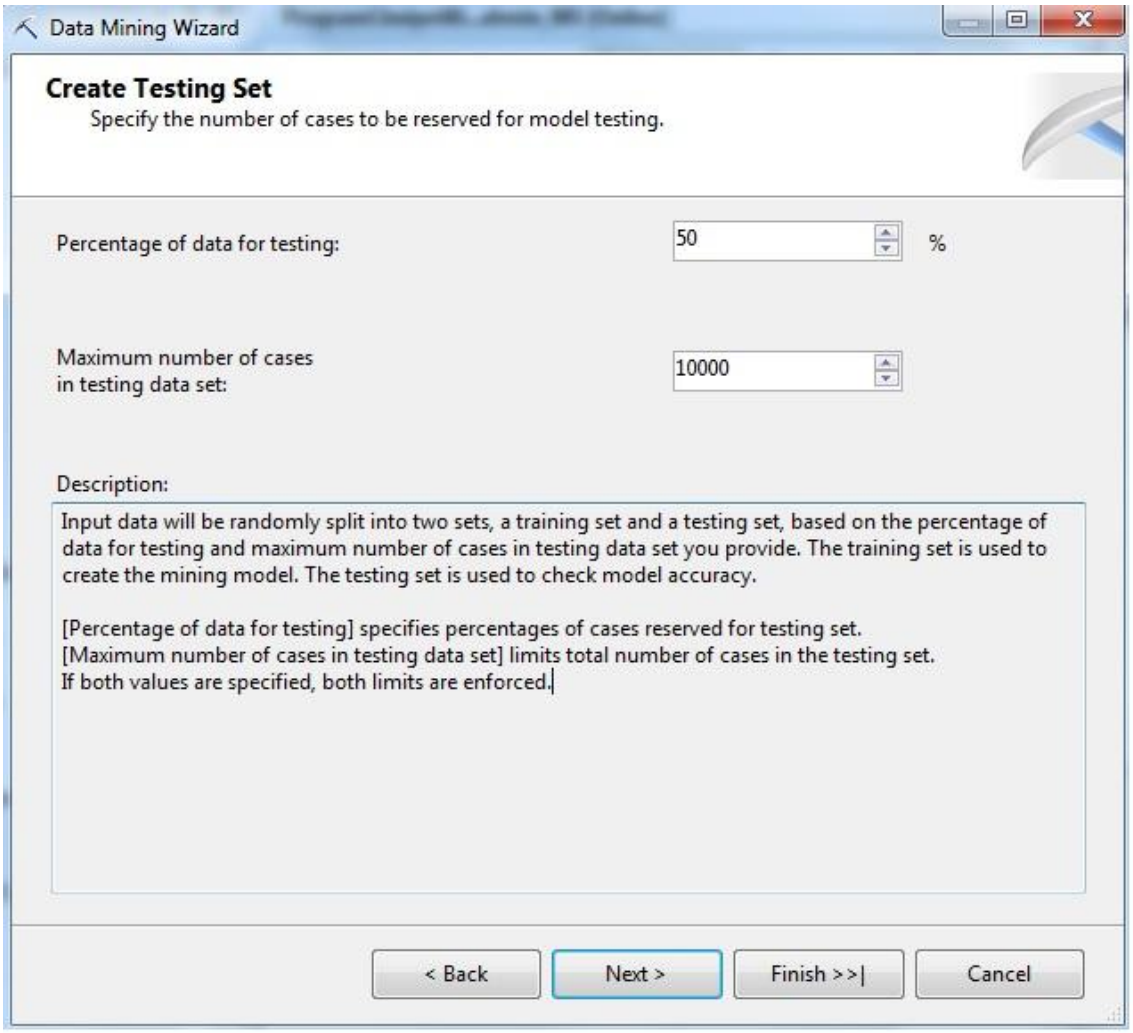
Şekil 5.10: Eğitilecek verinin belirlenmesi

Sonraki aşama Veri Madenciliği modeline gönderilecek kolanların tiplerini düzenlemeye yarar. Burada istenirse sayısal dönüştürme işleminden geçirerek algoritmada kullanılması sağlanabilir. Karar Ağaçları modeli sayısal verileri bu şekilde bir gruplama işlemi yaptıktan sonra kullandırır. Örneğin elimizde 1'den 100 e kadar olan değerler olsaydı sistem bunları 1-10 arası birinci grup, 11-20 arası ikinci grup… şeklinde gruplayarak algoritmaya öyle alacaktır. Şekil 5.11 de görüldüğü gibi sistemdeki tüm değerler karakter tabanlı olduğu için herhangi bir dönüştürme işlemi yapılmasına gerek yoktur.



Şekil 5.11: Kolon dönüştürme işlemi

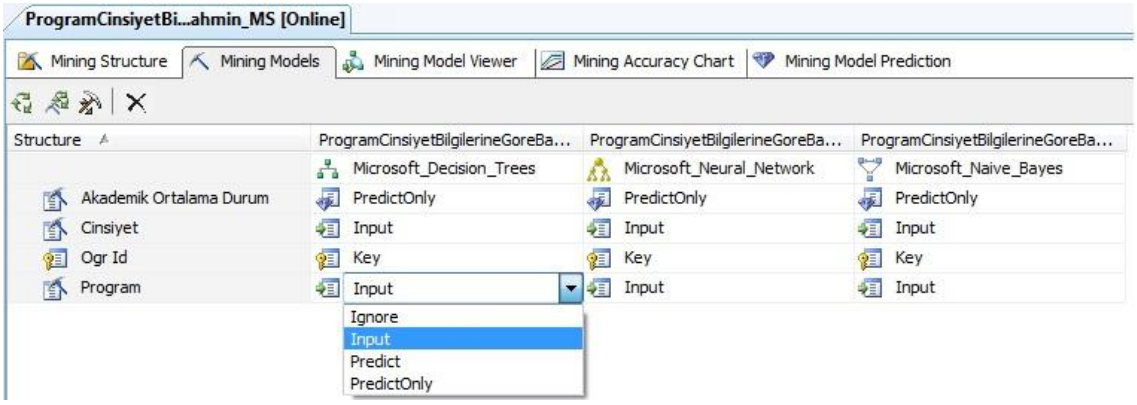
Bu ekrandan sonra ne kadar test verisi kullanılacağı belirtilir. Bu değer elimizdeki tüm kayıt sayının belli bir yüzdesi veya belli sayısal değer olabilir veya her ikisi de (hem yüzde sınırı koyma hem de kayıt sayısı sınırı koyma) işaretlenebilir(Şekil 5.12).



Şekil 5.12: Eğitim veri sayısı belirleme

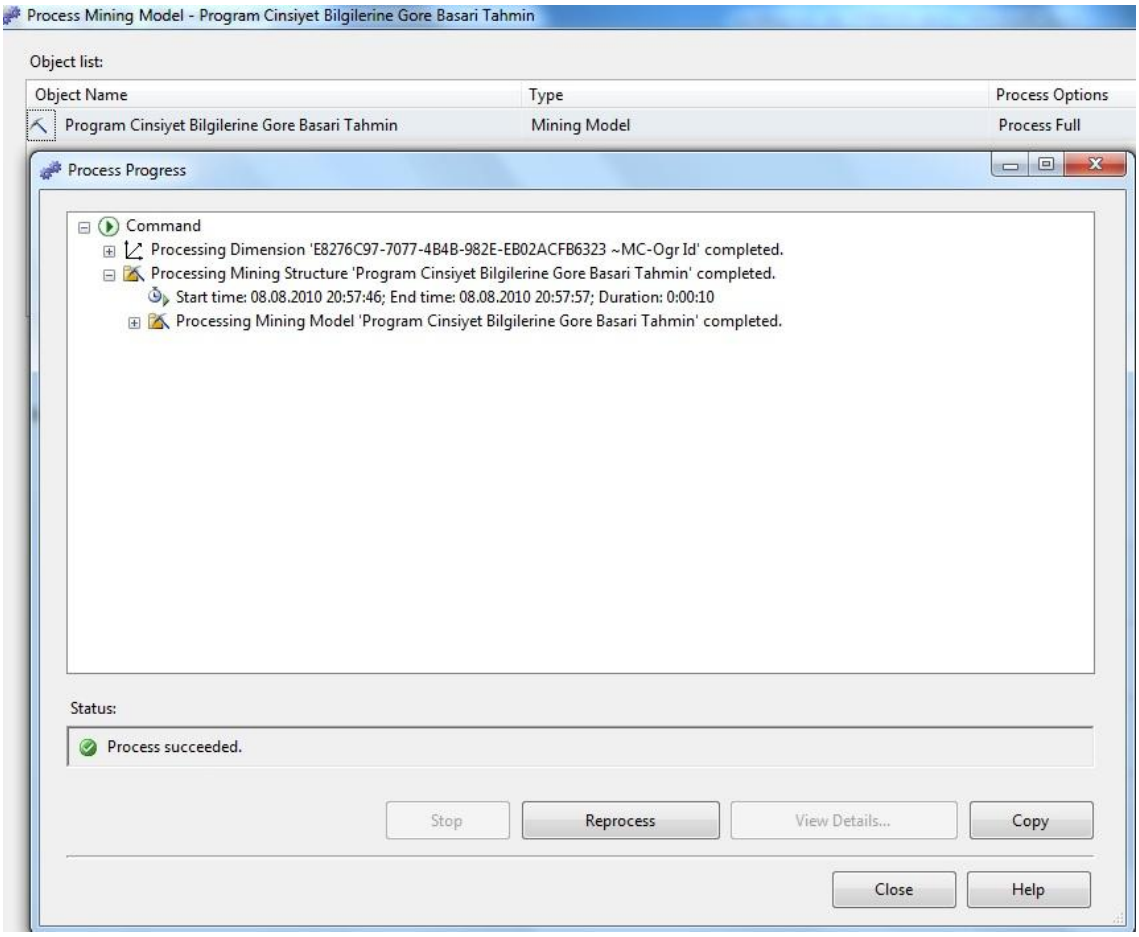
Bu işlemden sonra veri madenciliğinde kullanacağımız veri yapımız ve modelimiz hazır olacaktır.

Yapımız oluşturulduktan sonra karşımıza üzerinde çeşitli sekmeleri olan yeni bir ekran gelecektir. Bir açılan sayfada “mining structure” denilen yapıyı değiştirebiliriz. Yeni kolonlar eklenip eskileri çıkarılabilir. Yanındaki sekmede veri madenciliği modeli üzerinde değişiklik yapabilmemize imkân sağlar. Ayrıca ayrı yapı üzerinden birden fazla algoritma çalıştırmaya yardımcı olur. Bu sayede farklı algoritmalar üzerinden performans karşılaştırması yapabiliriz. Mevcut oluşturulan modelin yapısına başka bir algoritma eklenme işlemi Şekil 5.13 de görülebilir.



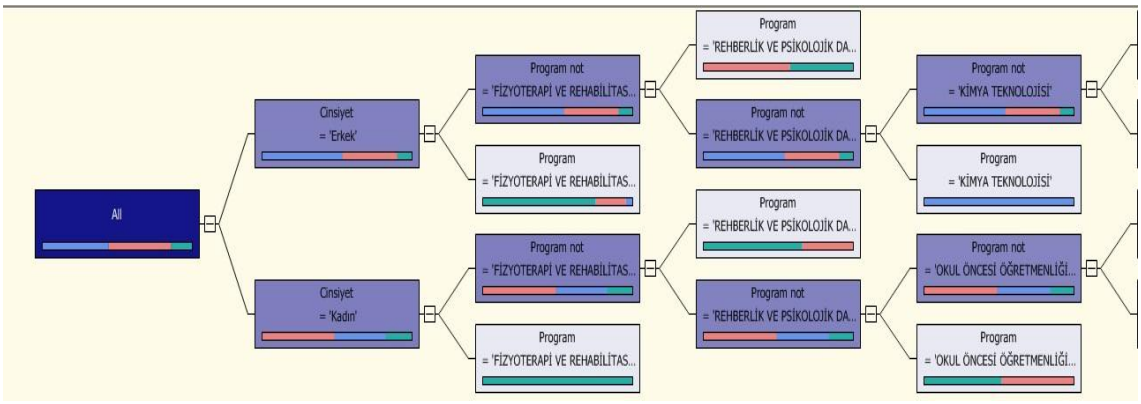
Şekil 5.13: Veri madenciliği modeli ekleme

Modelleri tanımladıktan sonra "Database" seçeneğinden "Process" i tıklayarak algoritma test verileri üzerinden işlem yapması sağlanır(Şekil 5. 14).



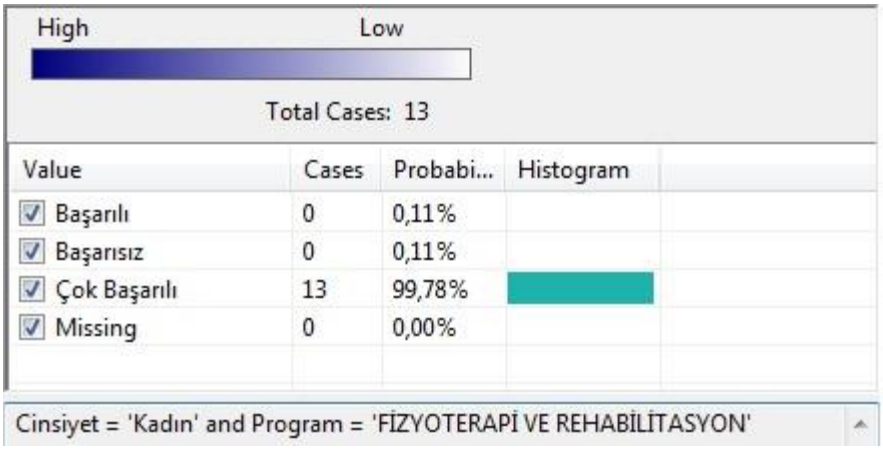
Şekil 5.14: Algoritmaların eğitilme işlemi

Veri madenciliği modeli sayfasının üçüncü sekmesinde algoritmaların eğitim verilerine göre buldukların sonuçların listesi görüntülenir. Karar ağaçları kullanılarak yapılan modelle şu şekilde bir sonuç döndürmüştür(Şekil 5. 15).



Şekil 5.15: Program ve cinsiyet verilerine göre karar ağaçları modelinde oluşan sonuçlar

Her ağaç dalının başarı durumlarıyla ilgi ayrı yüzdeleri vardır. Kişi kendi cinsiyet ve program bilgisini girdiği zaman daha özel tahmin yüzdeleri ile karşılaşacaktır. Örnek olması açısından Fizyoterapi ve Rehabilitasyon programındaki kız öğrencilerin tahmin yüzdelerini şekil 5.16' de gösterilmektedir.

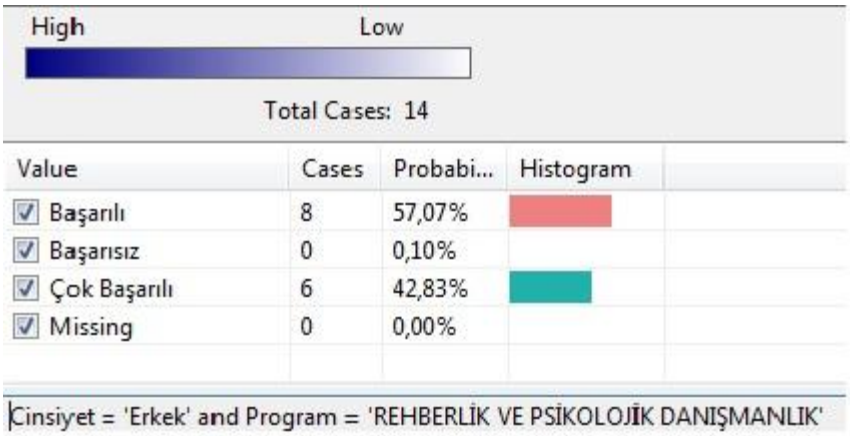


Şekil 5.16: Fizyoterapi ve Rehabilitasyon programı, kız öğrencilerin karar ağacıyla başarısı durumu

Burada akademikOrtalamaDurum adlı alan akademik ortalamanın hesaplanmasından meydana gelmiştir. Genel olarak gerçekleştirilen tüm modellerde akademik ortalama değerleri şu şekilde kabul edilmiştir;

- 0.001 ile 1.999 arası "Başarısız"
- 2.0 ile 2.999 arası "Başarılı"
- 3.0 ve üstü "Çok Başarılı"

Başka bir örnekte, Rehberlik ve Psikolojik Danışmanlık programındaki erkek öğrencilerin başarı tahmin durumları şu şekilde gösterilmektedir(Şekil 5.17)

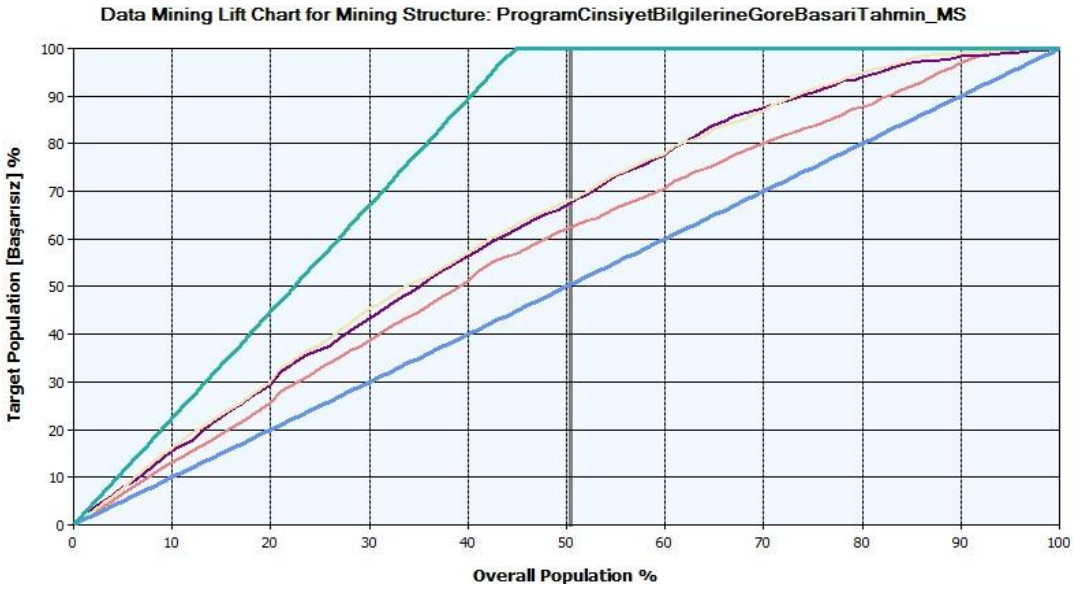


Şekil 5.17: Rehberlik ve Psikolojik Danışmanlık programı erkek öğrencilerinin karar ağacıyla başarı durumları

Veri madenciliği modeli ekranındaki dördüncü sekme, yapılan çalışmanın ne kadar doğru olduğunu belirlemede kullanılır. Eğitilmiş verilerin başka veriler karşısında döndürdüğü sonuçlar kontrol edilerek modelin ne kadar başarılı olduğu belirlenir. "Lift Chart" kısmında eğitilen tüm algoritmalar belli değerlere göre sınanmasıyla modellerin başarıları ölçülür. Bu yapıdaki "belli değerler" akademik ortalama durum alanındaki üç farklı (başarısız, başarılı, çok başarılı) değeri ifade etmektedir.

Buradaki amaç minimum kayıt gezerek hedeflenen tüm kayıtlara ulaşmaktır. Ne kadar kısa sürede %100 e yakın veya eşit değerlere çıkılırsa algoritma o kadar iyi eğitilmiş demektir. Buradaki mavi çizgi standart bir arama işlemindeki ilerleyişi temsil etmekte, pembe çizgi ise en ideal durumda sonuca ulaşma yolunu gösterir. Diğerleri üç algoritmanın sonuçlarını gösterir.

Şekil 5.18'de algoritmaların "Başarısız" durumundaki sonuçları gösterilmektedir.

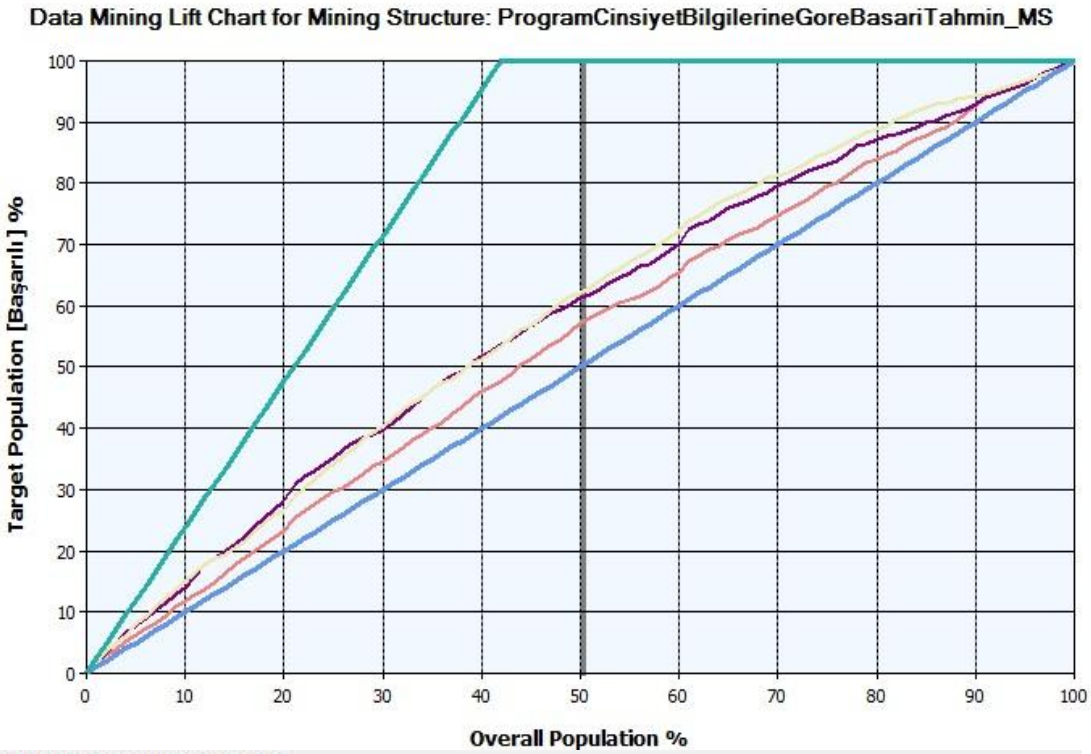


Population percentage: 50,49%

Series, Model	Score	Target population	Predict probability
ProgramCinsiyetBilgilerineGoreBasariTahmin_DT	0,75	62,93%	37,19%
ProgramCinsiyetBilgilerineGoreBasariTahmin_NN	0,80	68,41%	45,37%
ProgramCinsiyetBilgilerineGoreBasariTahmin_NB	0,81	68,72%	44,09%
Random Guess Model		51,00%	
Ideal Model for: ProgramCinsiyetBilgilerineGoreB...		100,00%	

Şekil 5.18: Program ve cinsiyete göre veri madenciliği algoritmalarının "Başarısız" durumu ile ilgili değerleri

Ayın durumda "Başarılı" durumunda elde edilen değerler Şekil 5.19'de ifade edilmiştir.

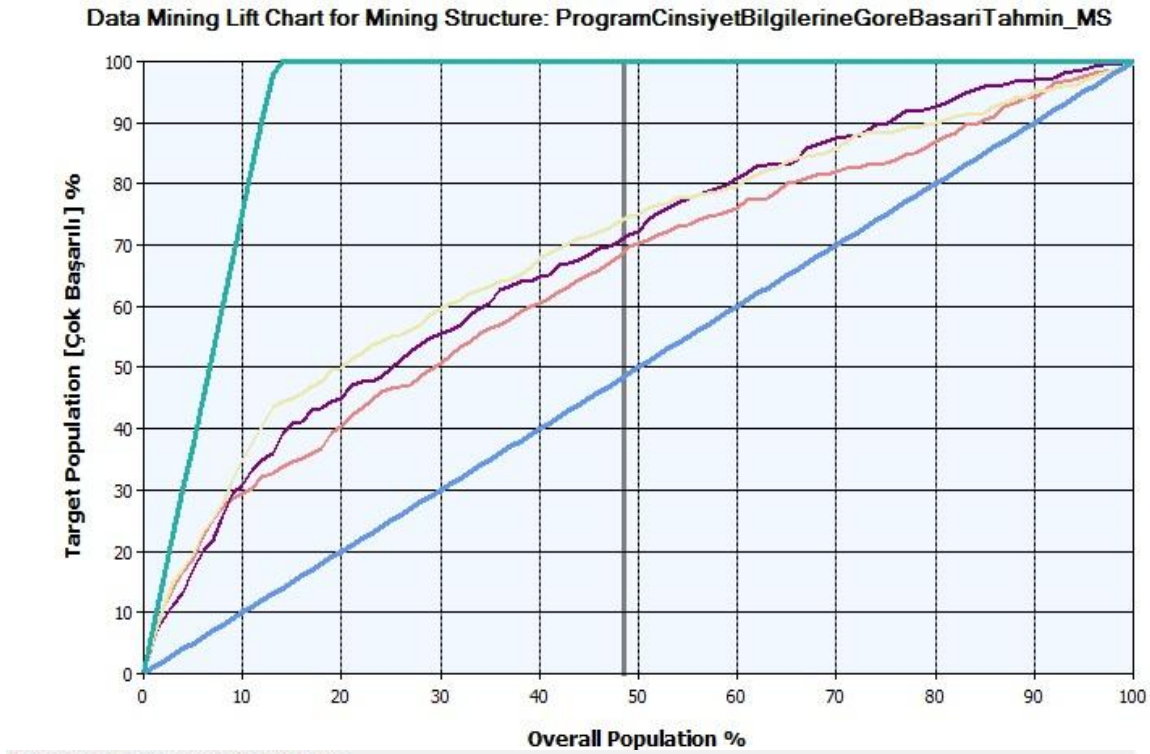


Population percentage: 50,49%

Series, Model	Score	Target population	Predict probability
ProgramCinsiyetBilgilerineGoreBasariTahmin_DT	0,69	58,01%	36,41%
ProgramCinsiyetBilgilerineGoreBasariTahmin_NN	0,73	61,93%	40,46%
ProgramCinsiyetBilgilerineGoreBasariTahmin_NB	0,74	62,99%	40,57%
Random Guess Model		51,00%	
Ideal Model for: ProgramCinsiyetBilgilerineGore...		100,00%	

Şekil 5.19: Program ve cinsiyete göre veri madenciliği algoritmalarının "Başarılı" durumu ile ilgili değerleri

Son olarak "Çok Başarılı" değeri için başarı grafikleri Şekil 5.20'de gösterilmiştir.



Population percentage: 48,54%

Series, Model	Score	Target population	Predict probability
ProgramCinsiyetBilgilerineGoreBasariTahmin_DT	0,69	69,81%	13,85%
ProgramCinsiyetBilgilerineGoreBasariTahmin_NN	0,73	71,70%	10,59%
ProgramCinsiyetBilgilerineGoreBasariTahmin_NB	0,74	74,63%	9,70%
Random Guess Model		49,00%	
Ideal Model for: ProgramCinsiyetBilgilerineGore...		100,00%	

Şekil 5.20: Program ve cinsiyete göre veri madenciliği algoritmalarının "Başarısız" durumu ile ilgili değerleri

Daha önce bahsedildiği üzere model adları, ana isim sonuna alt çizi ve algoritmanın baş harfleri gösterilecek şekilde düzenlenmiştir. Bu harfler DT (Decision Trees - Karar Ağaçları), NN(Neural Network – Sinir Ağları) ve NB(Naive Bayes) algoritmalarını gösterir.

Eğitilen Veri Madenciliği modellerinin test verileri karşında verdiği doğru tahminler onun başarısını belirtir. Üç farklı algoritmanın test verileri karşısındaki sonuçları Tablo 5.2'de gösterildiği gibidir.

Tablo 5.2: Program cinsiyet modelinin tahmin oranları

Microsoft Karar Ağaçları

Tahmin Edilen	Çok Başarılı (Gerçek)	Başarılı (Gerçek)	Başarısız (Gerçek)	Tahmin
Çok Başarılı	89	79	12	18,70%
Başarılı	255	811	624	53,90%
Başarısız	133	615	969	60,40%
	Tahmin Oranı	52,10%		

Microsoft Sinir Ağları

Tahmin Edilen	Çok Başarılı (Gerçek)	Başarılı (Gerçek)	Başarısız (Gerçek)	Tahmin
Çok Başarılı	95	87	31	19,90%
Başarılı	247	827	464	55,00%
Başarısız	135	591	1110	69,20%
	Tahmin Oranı	56,60%		

Microsoft Naive Bayes

Tahmin Edilen	Çok Başarılı (Gerçek)	Başarılı (Gerçek)	Başarısız (Gerçek)	Tahmin
Çok Başarılı	86	60	16	18,00%
Başarılı	268	834	460	55,40%
Başarısız	123	611	1129	70,30%
	Tahmin Oranı	57,10%		

Tablonun sol tarafı model tarafından tahmin edilen değerleri göstermektedir. Kolonlar ise verinin gerçekteki durumunu gösterir. Sinir ağları algoritmasına bakılacak olursa 95 tane "Çok Başarılı" tahmini yapılmış ve bunlar doğru çıkmıştır. Ama bunun yanında 87 öğrenciye de normalde "Başarılı" durumundayken "Çok Başarılı" tahmininde bulunmuş ve hataya sebebiyet vermiştir. Kısacası diyagonalde bulunan sayılar doğru yapılan tahminleri dışındakilerde yanlış yapılan tahminleri gösterir. Bu değerler ışığında Naive Bayes modeli en iyi performansı verdiği için raporlamada kullanılmak üzere seçilmiştir.

Rapor gösterimi için Microsoft Reporting Services, rapor verilerini getirmek içinse DMX sorguları kullanılmıştır. Kullanılan bu modelin örnek test verileri için oluşturulan DMX yapısı aşağıda verilmiştir.

```
SELECT
  (t.[ogrNo]) as [OgrNo],
  (t.[isim]) as [İsim],
  (t.[cinsiyet]) as [Cinsiyet],
  (t.[program]) as [Program],
  ([ProgramCinsiyetBilgilerineGoreBasariTahmin_NB].[Akademik Ortalama
Durum]) as [Tahmin],

(PredictProbability([ProgramCinsiyetBilgilerineGoreBasariTahmin_NB].[A
kademik Ortalama Durum],'Başarısızz')) as [Başarısız Tahmin Yüzdesi],

(PredictProbability([ProgramCinsiyetBilgilerineGoreBasariTahmin_NB].[A
kademik Ortalama Durum],'Başarılı')) as [Başarılı Tahmin Yüzdesi],

(PredictProbability([ProgramCinsiyetBilgilerineGoreBasariTahmin_NB].[A
kademik Ortalama Durum],'Çok Başarılı')) as [Çok Başarılı Tahmin
Yüzdesi]
From
  [ProgramCinsiyetBilgilerineGoreBasariTahmin_NB]
PREDICTION JOIN
  OPENQUERY([OKDS_DS],
    'SELECT
      [ogrNo],
      [isim],
      [cinsiyet],
      [program]
    FROM
      [Kds].[vw_ProgramCinsiyetBilgilerineGoreBasariTahmin_Test]
    ') AS t
ON
  [ProgramCinsiyetBilgilerineGoreBasariTahmin_NB].[Program] =
t.[program] AND
     [ProgramCinsiyetBilgilerineGoreBasariTahmin_NB].[Cinsiyet]       =
t.[cinsiyet]
```

DMX sorguları genel olarak standart SQL sorgularına benzerdir. Sorgunun en önemli kısmı olan “PREDICTION” ifadesinde eldeki model ile test verisi birleştirilerek tahmin işlemi yapılmaktadır. Birliktelik Kuralının kullanıldığı model hariç, diğer modellerde de ana sorgu mantığı çalışmaktadır.

Hazırlanan test verisine göre oluşturulan raporlar Şekil 5.21’de gösterildiği gibidir.

Program ve Cinsiyet Bilgilerine Göre Akdemik Başarı Belirleme

Sıra No	Cinsiyet	Program	Başarı Tahmin	Başarısız Tahmin Yüzdesi	Başarılı Tahmin Yüzdesi	Çok Başarılı Tahmin Yüzdesi
1	Kadın	ANESTEZİ	Çok Başarılı	0.03%	0.03%	99.92%
2	Erkek	ANESTEZİ	Çok Başarılı	0.03%	0.03%	99.92%
3	Kadın	ANTRENÖRLÜK EĞİTİMİ	Başarılı	24.07%	75.87%	0.03%
4	Erkek	ANTRENÖRLÜK EĞİTİMİ	Başarılı	38.85%	61.09%	0.03%
5	Kadın	ARKEOLOJİ	Başarılı	18.60%	68.51%	12.86%
6	Erkek	ARKEOLOJİ	Başarılı	32.31%	59.37%	8.30%
7	Kadın	ARKEOLOJİ (İ.Ö.)	Başarılı	11.27%	88.68%	0.03%
8	Erkek	ARKEOLOJİ (İ.Ö.)	Başarılı	20.28%	79.66%	0.03%
9	Kadın	BANKACILIK VE SİGORTACILIK	Başarısız	58.52%	36.59%	4.86%

Program ve Cinsiyet Bilgilerine Göre Akdemik Başarı Belirleme

	Erkek				Kadın			
	Akademik Ortalama Durum	Başarısız Tahmin Yüzdesi	Başarılı Tahmin Yüzdesi	Çok Başarılı Tahmin Yüzdesi	Akademik Ortalama Durum	Başarısız Tahmin Yüzdesi	Başarılı Tahmin Yüzdesi	Çok Başarılı Tahmin Yüzdesi
ANESTEZİ	Çok Başarılı	0.03%	0.03%	99.92%	Çok Başarılı	0.03%	0.03%	99.92%
ANTRENÖRLÜK EĞİTİMİ	Başarılı	38.85%	61.09%	0.03%	Başarılı	24.07%	75.87%	0.03%
ARKEOLOJİ	Başarılı	32.31%	59.37%	8.30%	Başarılı	18.60%	68.51%	12.86%
ARKEOLOJİ (İ.Ö.)	Başarılı	20.28%	79.66%	0.03%	Başarılı	11.27%	88.68%	0.03%
BANKACILIK VE SİGORTACILIK	Başarısız	74.45%	23.22%	2.31%	Başarısız	58.52%	36.59%	4.86%
BANKACILIK VE SİGORTACILIK (İ.Ö.)	Başarısız	56.05%	43.89%	0.03%	Başarılı	38.89%	61.06%	0.03%
BEDEN EĞİTİMİ VE SPOR ÖĞRETMENLİĞİ	Başarılı	33.83%	58.03%	8.11%	Başarılı	19.67%	67.61%	12.69%

Şekil 5.21: Program ve cinsiyet verilerine göre AB Tahmin Raporları

Rapordaki gösterilen "Akademik Ortalama Durum" alanı gelen verilere göre sistemin yaptığı tahmini göstermektedir. Diğer "Başarısız, Başarılı, Çok Başarılı Tahmin Yüzdesi" alanları kişinin belirtilen durumla ilgili tahminden ne kadar yakın veya uzak olduğunu gösterir. Bir örnek vermek gerekirse, ilk rapordaki "3" sıra numaralı kişi bilgileri şu şekildedir; Tahmin yüzdeleri sırayla(Başarısız, Başarılı, Çok Başarılı) 24.07, 75.87 ve 0.03 değerlerini almıştır. Sistem tahmin sansının maksimum

yapmak istediğinden öğrenci için en uygun sonucun "Başarılı" olarak nitelendirildiği görülmektedir.

Bu raporlarda bir test kümesi kullanılarak raporlar hazırlanmıştır. Yani birden çok kişinin tahmin işlemini bir anda yapılmaktadır. Ama kullanıcılar giriş parametrelerini kendileri belirleyerek, özel durumlar hakkında sonuç almak isteyebilirler. Bu durumlar için hazırlanan rapor aşağıdaki şekilde gösterilmiştir. Diğer tahmin işlemlerinde de bu şekilde dışarıdan parametre alarak sonuç döndüren rapor yapılmıştır. Örnek olması açısından sadece program-cinsiyet verilerine göre sonuç döndüren rapor şekil 5.22'de gösterilmiştir. Diğer modellerde de farklı parametreler alarak tahmin işlemi gerçekleştiren raporlar hazırlanabilir.

Şekil 5.22: Giriş değerlerine göre AB Tahmin işlemi

Program ve cinsiyet bilgilerine göre yapılan bu işlemde 2009 yılında üniversiteye yeni kayıt yaptıran öğrenciler kullanılmıştır. Akademik başarı tahmini ve raporlamanın yanında aşağıdaki sonuçlarda belirlenmiştir.

- Naive Bayes yöntemi kadınların, erkeklere göre daha başarılı olduğunu göstermektedir.

Tablo 5.3: Cinsiyete göre başarı

	Başarısız	Başarılı	Çok Başarılı
Erkek	60,345	43,141	36,062
Kadın	39,655	56,859	63,938

- Karar Ağaçları yöntemi Fizyoterapi ve Rehabilitasyon, Rehberlik ve Psikolojik Danışmanlık, Arkeoloji, Kimya Teknolojisi ve Türkçe

Öğretmenliği programları dışında bulunan erkek öğrencilerin üniversitede başarı yüzdesini %46 olarak bulmuştur. Yeni gelen erkek öğrencilerin yarıdan fazlası üniversitenin ilk senesinde başarısız olmaktadır.

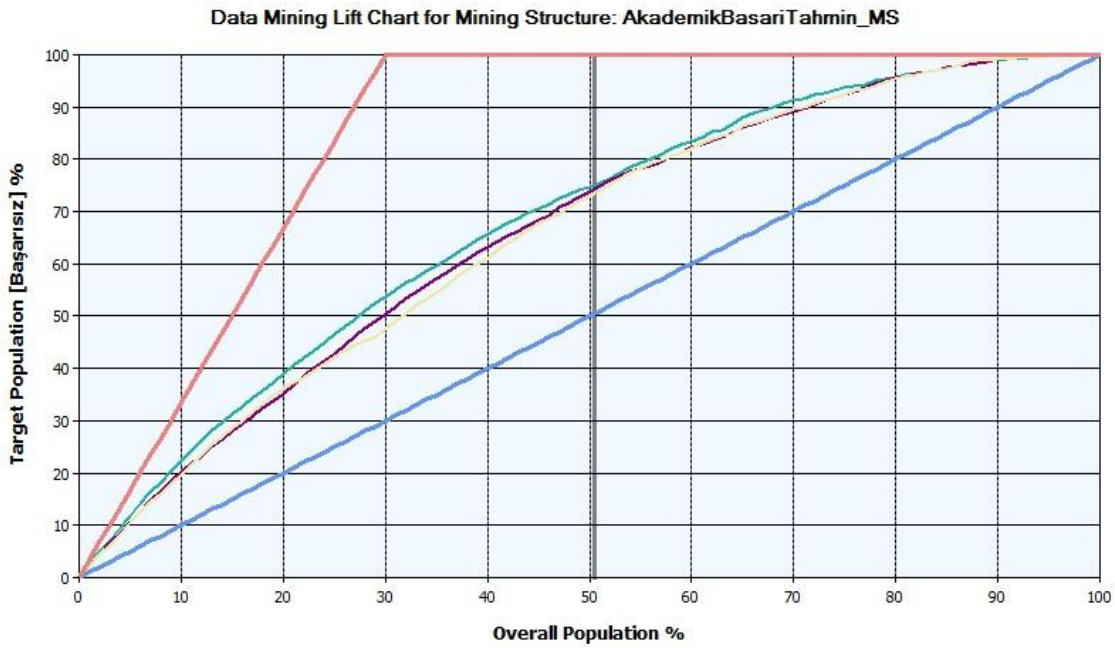
5.4.2. Öğrenci kimlik bilgilerine göre akademik başarı tahmini

Bu yapıda temel öğrencilerin bilgilerinden faydalanarak öğrencilerin başarısı tahmin edilmeye çalışılmıştır. Bilgilerin alındığı 2010 yaz ayında 26000'in üzerinde öğrenci üniversiteye kayıtlı durumdadır. Model için kullanılmaya elverişli öğrencilerden 12296 tanesi eğitim veri olarak seçilmiştir. Bu öğrenciler arasından 2009 yılında hazırlık okumuş olan öğrenciler çıkarılmıştır. Tahmin işleminde kullanılacak giriş parametreleri Tablo 5.4'de belirtilmiştir.

Tablo 5.4: Öğrenci kimlik bilgilerine göre AB tahmin modelinde giriş ve çıkış parametreleri

Akademik Birim Ad	Öğrencinin bağlı olduğu akdemik birim(Mühendislik Fakültesi, Sağlık Yüksek Okulu...)
Akademik Ortalama Durum	Tahmin Edilecek Değer. Akademik ortalamayı üç kısıma ayırarak tahmin işlemi yapılır. Başarısız (0 - 1,99), Başarılı (2 - 2,99), Çok Başarılı (3 - 4) olarak ayrılma gerçekleştirilmiştir.
Birim Tur	Bağlı bulunana akademik birimin türü(fakülte, meslek yüksek okulu,...)
Cinsiyet	Erkek, Kadın
Giris Tipi Ozet	Özel Yetenek,ÖSS,Af ile Dönenler,Dikey Geçiş,Yatay Geçiş,Ek Kontenjan ve Diğer durumlar
Medeni Hal	Boşanmış,Evli,Dul,Bekar
Ogr Yas	Öğrencinin üniversiteye kayıt yaptırdığı sıradaki yaşı
Okul Birincisi Durum	Lisede okul birincisi olup olmadığı
Orta Ogretim Mezuniyet Yili	Orta öğretim kurumundan hangi yılda mezun olduğunu gösterir.
Oss Giris Puan	Üniversiteyi kaç puanla kazandığı
Program	Üniversitede aktif olan programlar; Bilgisayar Mühendisliği, Fizik, Sosyoloji…
Tercih Sirasi	Bölüm tercihlerini yaparken, öğrencinin kazandığı programı kaçıncı tercihi olarak yazdığı
Yuzdelik Dilim	ÖSS puanına göre öğrencinin sınavda bulunduğu yüzdelik dilim

Model önceki yapıda olduğu gibi üç farklı algoritma için çalıştırılmıştır. Bunlar Microsoft Karar Ağaçları, Microsoft Sinir Ağları ve Microsoft Naive Bayes algoritmalarıdır. Elde edilen "Lift Chart" değerleri aşağıda her durum için ayrı ayrı gösterilmiştir.



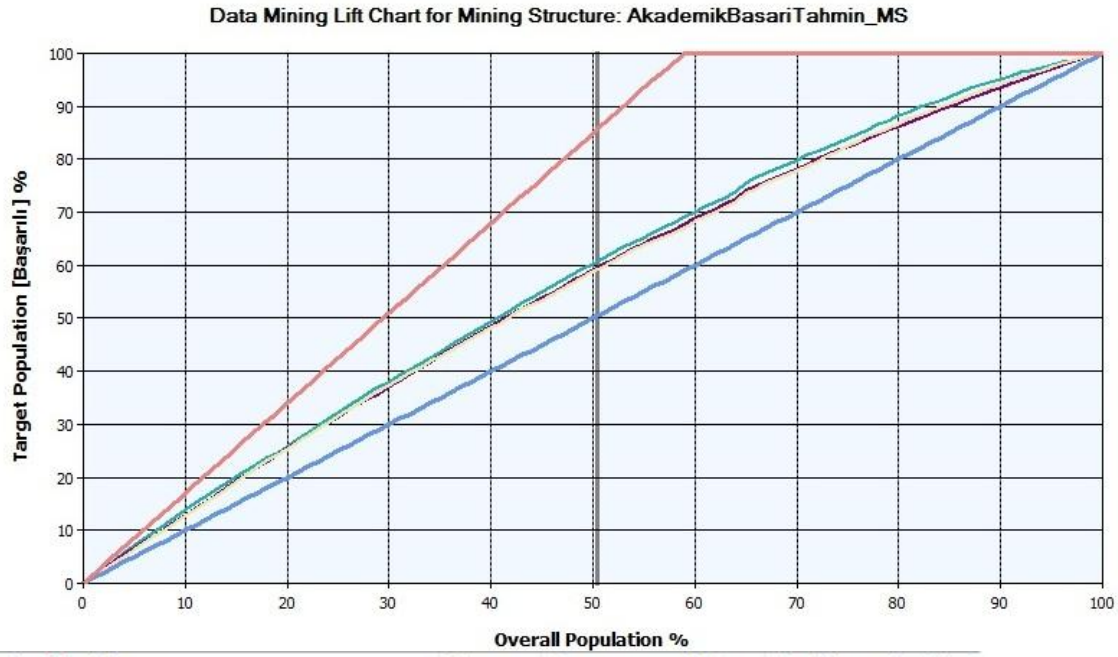
Population percentage: 50,50%

Series, Model	Score	Target population	Predict probability
AkademikBasariTahmin_DT	0,80	75,59%	23,71%
AkademikBasariTahmin_NN	0,78	74,88%	24,70%
AkademikBasariTahmin_NB	0,77	74,12%	22,51%
Random Guess Model		51,00%	
Ideal Model for: AkademikBasariTahmin...		100,00%	

Şekil 5.23: Öğrenci bilgilerine göre AB tahmin yönteminin "Başarısız" durumda lift chart değerleri

"Başarısız" durumunda algoritmaların başarı durumları Şekil 5.23'de gösterilmiştir. Buradaki amaç diğer benzer grafiklerde olduğu gibi minimum kayıttan geçerek, aranan değere ulaşabilmektir. Kırmızı çizgi sistemin en ideal durumunu belirtir. Bu kısım %30 civarı bir rakama denk gelmektedir. Bu demektir ki tüm başarısızların toplam küme içerisindeki oranı %30 dur. En ideal durumda tek seferde tüm hedef kitleyi bulmak için en iyi ihtimalle kayıtların sadece %30'luk kısmına bakmak yeterlidir. Mavi çizgi ise rastgele incelendiğinde istenilen kayıtlara ulaşılma durumunu belirtir. Diyagonal bir

çizgiden meydana gelmektedir. Diğer şekiller algoritmaların başarısını göstermektedir. Grafik ideal şekle ne kadar yakında o kadar başarılıdır.

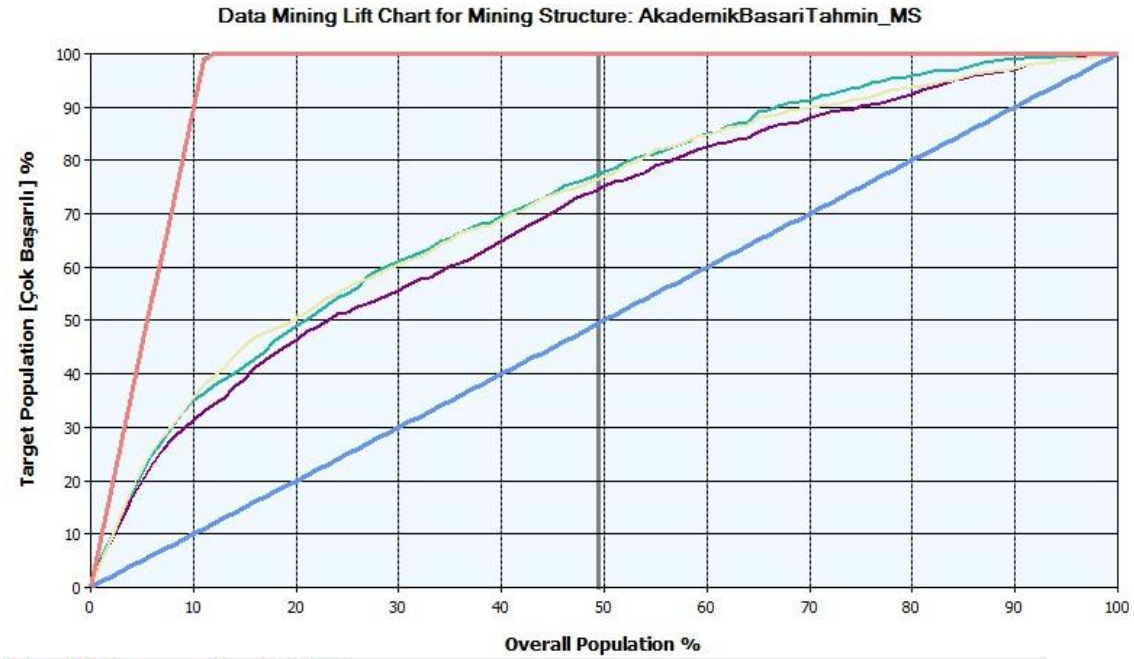


Series, Model	Score	Target population	Predict probability
AkademikBasariTahmin_DT	0,81	61,36%	58,49%
AkademikBasariTahmin_NN	0,79	60,15%	57,29%
AkademikBasariTahmin_NB	0,79	59,77%	63,08%
Random Guess Model		51,00%	
Ideal Model for: AkademikBasariTahmin...		86,57%	

Şekil 5.24: Öğrenci bilgilerine göre AB tahmin yönteminin "Başarılı" durumda lift chart değerleri

Şekil 5.24'de "Başarılı" durumunda algoritmaların sonuçları listelenmiştir.

"Çok Başarılı" olan öğrencileri arama işleminde modelin başarı durumu Şekil 5.25'da gösterilmiştir.



Population percentage: 49,50%

Series, Model	Score	Target population	Predict probability
AkademikBasariTahmin_DT	0,76	77,87%	9,52%
AkademikBasariTahmin_NN	0,73	75,09%	10,82%
AkademikBasariTahmin_NB	0,76	76,92%	4,54%
Random Guess Model		50,00%	
Ideal Model for: AkademikBasariTahmin...		100,00%	

Şekil 5.25: Öğrenci bilgilerine göre AB tahmin yönteminin "Çok Başarılı" durumdaki lift chart değerleri

Spesifik bir durumu bulmada Karar Ağaçlarının azda olsa diğerlerinden daha başarılı olduğu görülmektedir. Test verileri karşısında sistemin başarı durumu Tablo 5.5'de gösterilmiştir.

Tablo 5.5: Akademik başarı modelinin tahmin oranları

Microsoft Karar Ağaçları

Tahmin Edilen	Çok Başarılı (Gerçek)	Başarılı (Gerçek)	Başarısız (Gerçek)	Tahmin
Çok Başarılı	162	147	34	11,80%
Başarılı	1053	5971	1953	82,40%
Başarısız	154	1126	1696	46,00%
	Tahmin Oranı	63,70%		

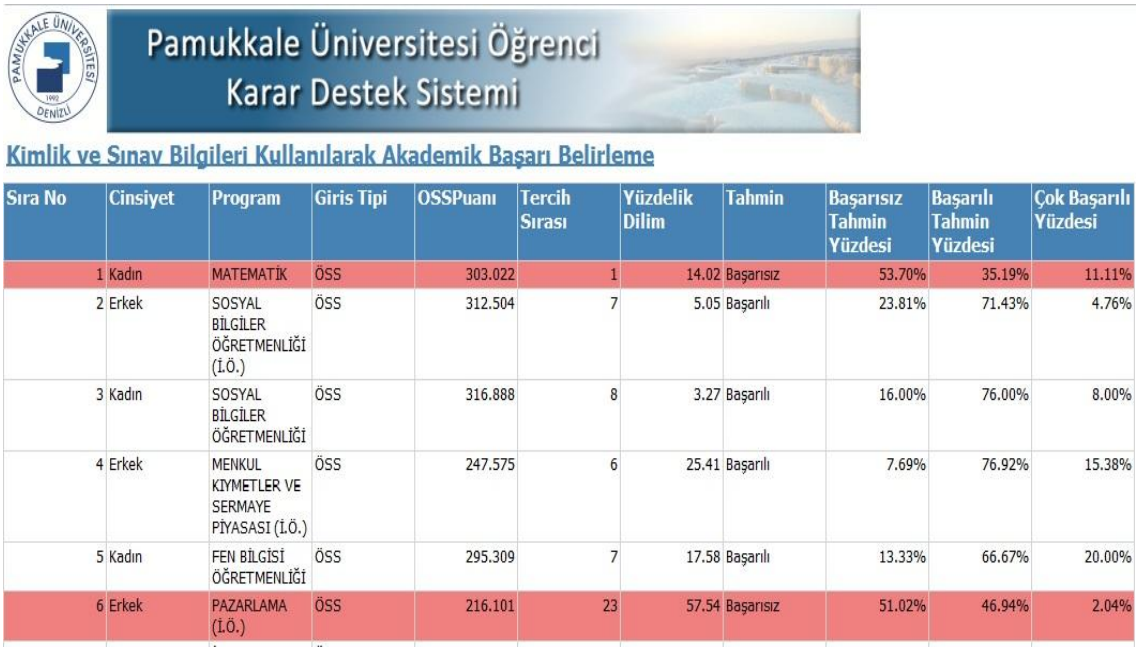
Microsoft Sinir Ağları

Tahmin Edilen	Çok Başarılı (Gerçek)	Başarılı (Gerçek)	Başarısız (Gerçek)	Tahmin
Çok Başarılı	257	290	81	18,80%
Başarılı	962	5803	2184	80,10%
Başarısız	150	1151	1418	38,50%
	Tahmin Oranı	60,80%		

Microsoft Naive Bayes

Tahmin Edilen	Çok Başarılı (Gerçek)	Başarılı (Gerçek)	Başarısız (Gerçek)	Tahmin
Çok Başarılı	342	352	27	25,00%
Başarılı	721	5157	1827	71,20%
Başarısız	306	1735	1829	49,70%
	Tahmin Oranı	59,60%		

"Başarısız" ve "Çok Başarılı" durumundaki öğrencilerde Naive Bayes algoritması daha iyi tahmin yüzdelerine sahiptir. "Başarılı" öğrencilerin tahmininde de Karar Ağaçları modeli daha iyi performans sergilemiştir. Genel tahmin değerlerinde de Karar Ağaçları diğer algoritmalara göre biraz daha başarılıdır. Bu sebeple raporda Microsoft Karar Ağaçları modeli kullanılacaktır. Bu model kullanılarak karar destek sistemi için hazırlanan raporlar şekil 5.26'de gösterilmiştir.



Kimlik ve Sınav Bilgileri Kullanılarak Akademik Başarı Belirleme

Sıra No	Cinsiyet	Program	Giris Tipi	OSSPuanı	Tercih Sırası	Yüzdelik Dilim	Tahmin	Başarısız Tahmin Yüzdesi	Başarılı Tahmin Yüzdesi	Çok Başarılı Yüzdesi
1	Kadın	MATEMATİK	ÖSS	303.022	1	14.02	Başarısız	53.70%	35.19%	11.11%
2	Erkek	SOSYAL BİLGİLER ÖĞRETMENLİĞİ (İ.Ö.)	ÖSS	312.504	7	5.05	Başarılı	23.81%	71.43%	4.76%
3	Kadın	SOSYAL BİLGİLER ÖĞRETMENLİĞİ	ÖSS	316.888	8	3.27	Başarılı	16.00%	76.00%	8.00%
4	Erkek	MENKUL KIYMETLER VE SERMAYE PİYASASI (İ.Ö.)	ÖSS	247.575	6	25.41	Başarılı	7.69%	76.92%	15.38%
5	Kadın	FEN BİLGİSİ ÖĞRETMENLİĞİ	ÖSS	295.309	7	17.58	Başarılı	13.33%	66.67%	20.00%
6	Erkek	PAZARLAMA (İ.Ö.)	ÖSS	216.101	23	57.54	Başarısız	51.02%	46.94%	2.04%

Kimlik ve Sınav Bilgileri Kullanılarak Akademik Başarı Belirleme

		Erkek					Kadın				
		Lise Mezuniyet Yılı	Tercih Sırası	Yüzdelik Dilim	ÖSS Giriş Puanı	Tahmin	Lise Mezuniyet Yılı	Tercih Sırası	Yüzdelik Dilim	ÖSS Giriş Puanı	Tahmin
⊞ BEKİLLİ MESLEK YÜKSEKOKULU		2006	23	57.54	216.101	Başarısız	2003	1	20.81	253.135	Başarılı
⊞ BULDAN MESLEK YÜKSEKOKULU		2005	8	24.86	248.21	Başarılı	2007	18	45.76	226.786	Başarılı
⊟ ÇİVRİL ATASAY KAMER MESLEK YÜKSEKOKULU	DIŞ TİCARET	2007	6	17.72	257.216	Çok Başarılı	2007	5	13.36	263.673	Çok Başarılı
	DIŞ TİCARET (İ.Ö.)	2006	24	19.51	254.826	Başarılı	2003	12	26.57	246.251	Başarılı
	İŞLETME YÖNETİMİ	2007	1	0	0	Başarılı	2004	1	0	0	Başarılı
	İŞLETME YÖNETİMİ (İ.Ö.)	2006	21	30.12	242.333	Başarısız	2007	12	31.62	240.744	Başarılı
	MUHASEBE VE VERGİ UYGULAMALARI	2005	5	0	0	Başarısız	2005	1	0	0	Başarısız

5. 26: Öğrenci bilgilerine göre AB tahmin raporu

Raporda eğitim kümesinde kullanılan alanlara göre bir test kümesi oluşturulmuş ve giriş parametrelerine göre öğrencilerin akademik başarısı tahmin edilmeye çalışılmıştır.

5.4.3. Öğrenci kimlik bilgileri ve anket sonuçlarına göre akademik başarı tahmini

Bir önceki modelde üniversitede aktif olarak bulunan öğrenciler arasında sadece öğrenci temel bilgileri kullanılarak tahmin işlemi yapılmaya çalışılmıştır. 2008 ve 2009 yıllarında üniversiteye kayıt yaptıran öğrencilere birçok farklı alan üzerine sorular

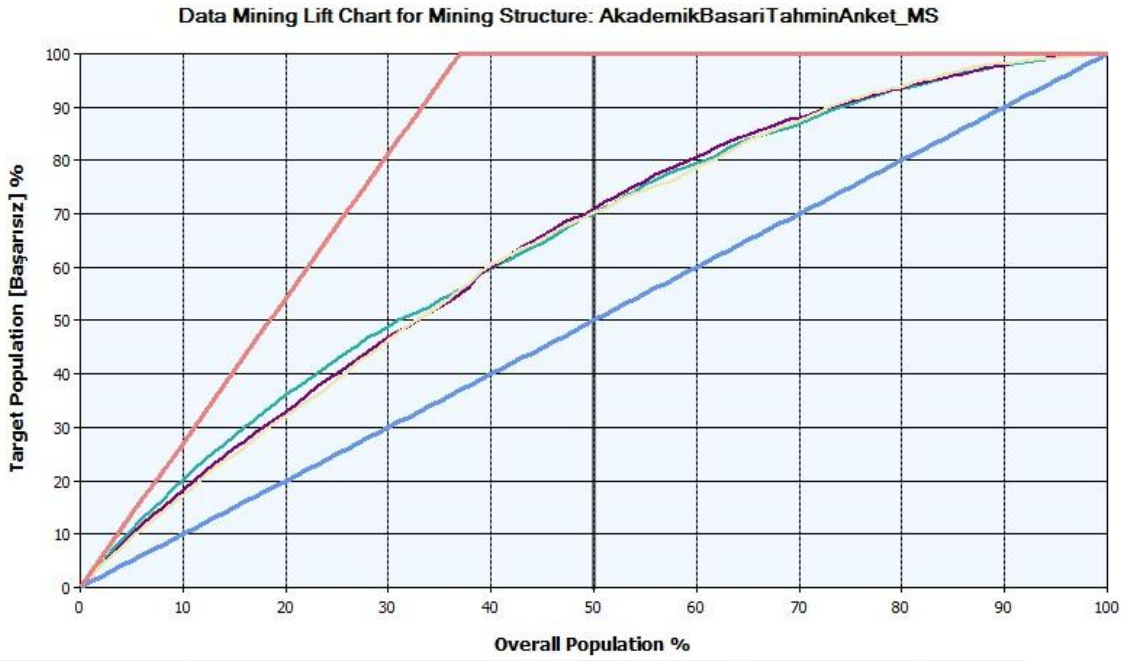
içeren bir anket uygulanmıştır. Bu nedenle sadece 2008 ve 2009 yılı girişli öğrencileri ilgilendirecek, temel öğrenci bilgileri dışında kullanılabilecek anket verilerini de içeren ayrı bir model oluşturulması düşünülmüştür. Öğrencilerin cevaplandırdığı ankette otuzun üstünde soru bulunmaktadır ama bunların çoğu analiz işleminde farkındalık oluşturmadığı için işleme alınmamıştır. Bu yöntemde sisteme giriş parametresi olarak kullanılan alanlar aşağıda verilmiştir.

Tablo 5.6: Anket verileri destekli AB belirlemede kullanılan parametreler

Aile Yasam Standartdı	Alabilecek değerler ; "Açlık Sınırı Altında", "Yoksulluk Sınırı Altında","Normal Aralıkta". Hangi aralığın belirleneceği anne ve babanın gelirleri ve çocuk sayılarına göre hesaplanır. Verilerin eski kaldığı göz önüne alınarak 2008 yılında girilen veriler üzerine %8.7 ve %2.5 oranlarında memur zamları uygulanır[43-44]. 2009 yılında veriler üzerine de %2.5 zam uygulanır. Bu gelirler ve çocuk sayıları üzerinden, [45]'de anlatıldığı üzere açlık ve yoksulluk sınırları belirlenir. Tüm bu kaynaklara rağmen bu sınıflandırmanın tam doğru yapılmamasını sağlayan sistemin kendi içindeki bazı durumlar mevcuttur. Öncelikle 2008 ve 2009 yılındaki ücretlerle ilgili cevap şıklarında değişliğe gidilmiş ve iki anketin birleştirilmesi zorlaşmıştır. Ayrıca ücretlerin sayısal veri olarak değil de, belli bir aralık belirterek yazı olarak sisteme girmiş olması(örnek ücretin 1000TL ile 1500TL arasında olduğunun girilmesi) ücretlerin tam doğru aralıklarda hesaplanamamasına neden olmuştur. Bu alan oluşturulurken eldeki veriler mümkün olduğunda temizlenip, standartlaştırdıktan sonra kullanılmıştır.
Akademik Birim Ad	Öğrencinin bağlı olduğu akademik birim(Mühendislik Fakültesi, Sağlık Yüksek Okulu...)
Akademik Ortalama Durum	Tahmin Edilecek Değer. Akademik ortalamayı üç kısıma ayırarak tahmin işlemi yapılır. Başarısız (0 - 1,99), Başarılı (2 - 2,99), Çok Başarılı (3 - 4) olarak ayrıl gerçekleştirilmiştir.
Anne Baba Ayrilik	Anne babanın birlikte veya ayrı oldukları belirtir.
Anne Yas	Annenin Yaşı
Askerlik Durum	Öğrencinin askerlikte ilişkisinin olup olmadığı
Birim Tur	Bağlı bulunana akademik birimin türü(fakülte, meslek yüksek okulu, ...)
Cinsiyet	Erkek, Kadın
Dershane Devam	Üniversiteye hazırlık aşamasında dershaneye gidilip, gidilmediği
Giris Tipi Ozet	Özel Yetenek, ÖSS, Af ile Dönenler, Dikey Geçiş, Yatay Geçiş, Ek Kontenjan ve Diğer durumlar
Kardes Sayisi	Öğrencinin kendisi hariç kardeş sayısı

Lise Diploma Not	Orta Öğretim kurumundan 100 üzerinden kaç puanla mezun olduğu
Medeni Hal	Boşanmış, Evli, Dul, Bekâr
Ogr Yas	Öğrencinin üniversiteye kayıt yaptırdığı sıradaki yaşı
Okul Birincisi Durum	Lisede okul birincisi olup olmadığı
Orta Ogretim Mezuniyet Yili	Orta öğretim kurumundan hangi yılda mezun olduğunu gösterir.
Oss Giris Puan	Üniversiteyi kaç puanla kazandığı
Program	Üniversitede aktif olan programlar; Bilgisayar Mühendisliği, Fizik, Sosyoloji…
Sehit Cocuk Durum	Öğrencinin Şehit çocuğu olup olmadığı
Tercih Sirasi	Bölüm tercihlerini yaparken, öğrencinin kazandığı programı kaçıncı tercihi olarak yazdığı
Yuzdelik Dilim	ÖSS puanına göre öğrencinin sınavda bulunduğu yüzde

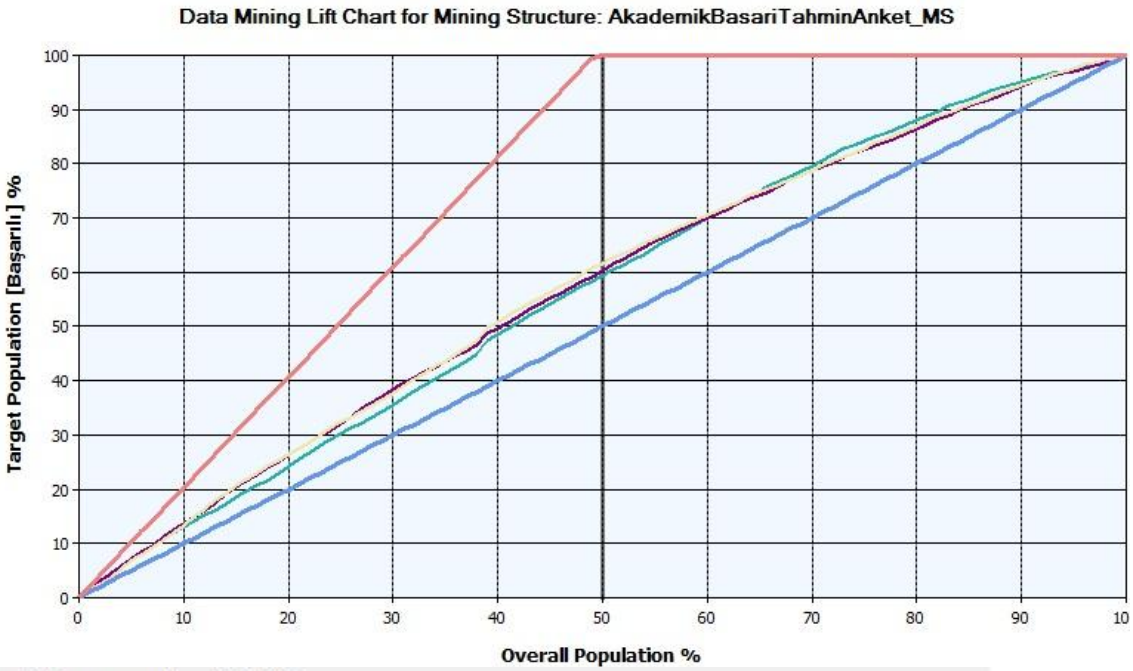
Eğitim verisi olarak 7092 kayıt kullanılmıştır. Bu öğrenciler arasından 2009 yılında hazırlık okumuş olan öğrenciler çıkarılmıştır. Önceki yöntemlerde kullanılan üç algoritma bu modelde de kullanılmıştır. Modelin "Başarısız" öğrencileri bulma başarısı Şekil 5.27 belirtilmiştir.



Series, Model	Score	Target population	Predict probability
AkademikBasariTahminAnket_DT	0,79	70,16%	35,71%
AkademikBasariTahminAnket_NN	0,79	70,92%	28,63%
AkademikBasariTahminAnket_NB	0,78	70,27%	36,66%
Random Guess Model		50,00%	
Ideal Model for: AkademikBasariTahmin...		100,00%	

Şekil 5.27: Öğrenci ve anket bilgilerine göre AB tahmin yönteminin "Başarısız" durumundaki lift chart değerleri

Aynı modelin "Başarılı" durum için çalışması Şekil 5.28'da gösterildiği gibi olmuştur;

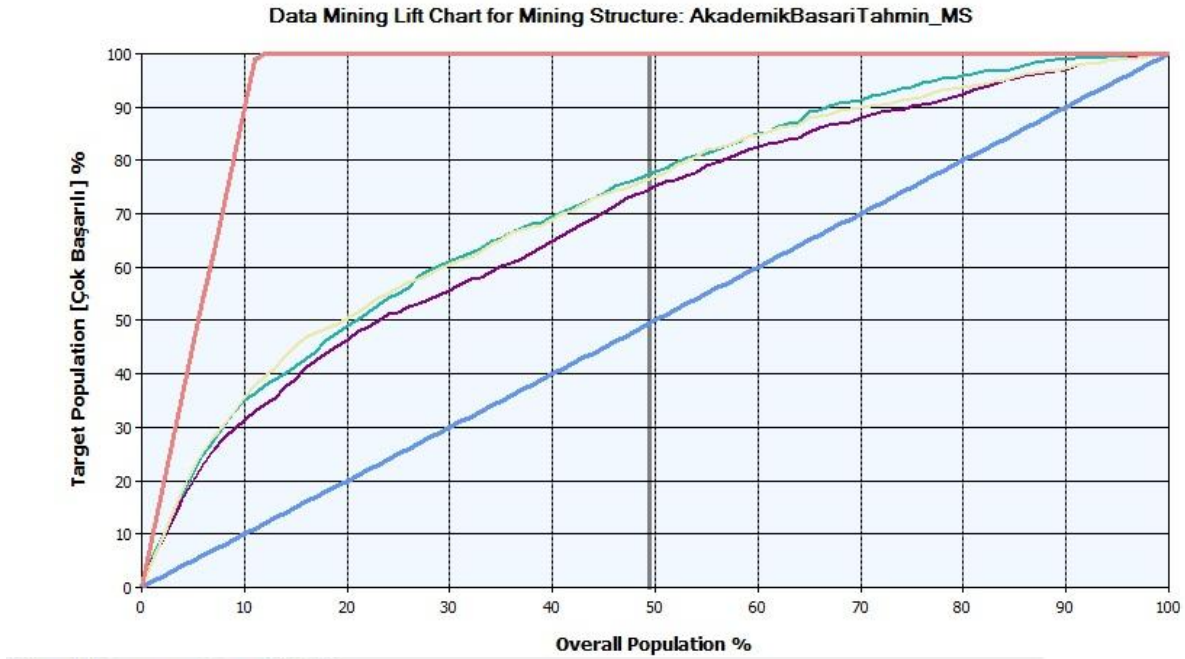


Population percentage: 50,00%

Series, Model	Score	Target population	Predict probability
AkademikBasariTahminAnket_DT	0,75	59,39%	51,35%
AkademikBasariTahminAnket_NN	0,76	60,30%	54,15%
AkademikBasariTahminAnket_NB	0,76	61,62%	44,42%
Random Guess Model		50,00%	
Ideal Model for: AkademikBasariTahmin...		100,00%	

Şekil 5.28: Öğrenci ve anket bilgilerine göre AB tahmin yönteminin "Başarılı" durumundaki lift chart değerleri

"Çok Başarılı" öğrencileri tahmin etme başarısı Şekil 5.29'daki grafikte gösterilmiştir.



Population percentage: 50,00%

Series, Model	Score	Target population	Predict probability
AkademikBasariTahminAnket_DT	0,75	78,78%	11,11%
AkademikBasariTahminAnket_NN	0,73	73,57%	10,41%
AkademikBasariTahminAnket_NB	0,73	72,65%	6,04%
Random Guess Model		50,00%	
Ideal Model for: AkademikBasariTahmin...		100,00%	

Şekil 5.29: Öğrenci ve anket bilgilerine göre AB tahmin yönteminin "Çok Başarılı" durumundaki lift chart değerleri

Diğer modellerde olduğu gibi en iyi ve en kötü arasındaki üç algoritma, ne kadar yukarıdaki kırmızı(ideal tahmin) çizgiye yaklaşırsa o kadar iyi ve ne kadar mavi(rastgele tahmin) çizgisine yaklaşırsa da o kadar kötü durumdadır. Algoritmaların test veriler karşısında tahmin başarısı Tablo 5.7'de verilmiştir.

Tablo 5.7: Anket verilerine göre AB modelinin tahmin oranları

Microsoft Karar Ağaçları

Tahmin Edilen	Çok Başarılı (Gerçek)	Başarılı (Gerçek)	Başarısız (Gerçek)	Tahmin
Çok Başarılı	190	175	40	19,40%
Başarılı	680	2630	1332	75,30%
Başarısız	110	689	1245	47,60%
	Tahmin Oranı	57,30%		

Microsoft Sinir Ağları

Tahmin Edilen	Çok Başarılı (Gerçek)	Başarılı (Gerçek)	Başarısız (Gerçek)	Tahmin
Çok Başarılı	125	112	27	12,80%
Başarılı	692	2643	1361	75,60%
Başarısız	163	739	1229	47,00%
	Tahmin Oranı	56,40%		

Microsoft Naive Bayes

Tahmin Edilen	Çok Başarılı (Gerçek)	Başarılı (Gerçek)	Başarısız (Gerçek)	Tahmin
Çok Başarılı	278	266	28	28,40%
Başarılı	433	2113	952	60,50%
Başarısız	269	1115	1637	62,60%
	Tahmin Oranı	56,80%		

Algoritmaların sonuçları birbirine oldukça yakın çıkmıştır. "Çok Başarılı" öğrencilerin tahmininde Karar Ağaçları, "Başarılı" öğrencilerin tahmininde Sinir Ağları, "Başarısız" öğrencilerin tahmininde de Naive Bayes algoritması diğer algoritmalara göre daha iyi performans göstermiştir. Ama genel duruma bakılacak olursa karar ağaçları diğer modellere göre biraz daha başarılı tahminlerde bulunmuştur. Bu nedenle raporda karar ağaçları algoritması kullanılacaktır. Rapor çıktıları Şekil 5.30'de gösterilmiştir;

Öğrenci Kimlik ve Anket Bilgileri Kullanılarak Akademik Başarı Belirleme

Sıra No	Cinsiyet	Program	Lise Diploma Not	Okul Birincisi Durum	Dershaneye Devam	Kardeş Sayısı	Aile Yaşam Standardı	Anne Baba Ayrılık	Tahmin
1	Kadın	MATEMATİK	89.39	Normal	Evet	1	Yoksulluk Sınırı Alt	Hayır.	Başarılı
2	Erkek	SOSYAL BİLGİLER ÖĞRETMENLİĞİ (İ.Ö.)	76.72	Normal	Evet	1	Açlık Sınırı Altında	Hayır.	Başarılı
3	Kadın	SOSYAL BİLGİLER ÖĞRETMENLİĞİ	83.61	Normal	Evet	1	Açlık Sınırı Altında	Hayır.	Başarılı
4	Erkek	MENKUL KIYMETLER VE SERMAYE PİYASASI (İ.Ö.)	55.8	Normal	Evet	1	Yoksulluk Sınırı Alt	Hayır.	Başarısız
5	Kadın	FEN BİLGİSİ ÖĞRETMENLİĞİ	74.2	Normal	Evet	1	Açlık Sınırı Altında	Hayır.	Başarılı

Öğrenci Kimlik ve Anket Bilgileri Kullanılarak Akademik Başarı Belirleme

		Erkek						Kadın	
		Lise Diploma not	Aile Ayrılık	Gelir Durumu	Kardeş Sayısı	Yüzdelik Dilim	Tahmin	Lise Diploma not	Aile Ayr
⊞ BEKİLLİ MESLEK YÜKSEKOKULU		76.6	Hayır.	Yoksulluk Sınırı Alt	1	57.54	Başarısız	67.2	Hayır.
⊞ BULDAN MESLEK YÜKSEKOKULU		58.4	Hayır.	Açlık Sınırı Altında	3 den fazla	24.86	Başarılı	88.17	Hayır.
⊟ ÇİVRİL ATASAY KAMER MESLEK YÜKSEKOKULU	DIŞ TİCARET	61.49	Hayır.	Açlık Sınırı Altında	3	17.72	Başarılı	76.63	Hayır.
	DIŞ TİCARET (İ.Ö.)	56	Hayır.	Yoksulluk Sınırı Alt	Yok	19.51	Başarısız	82.8	Hayır.
	İŞLETME YÖNETİMİ	74.74	Hayır.	Yoksulluk Sınırı Alt	Yok	0	Başarısız	66.8	Evet.
	İŞLETME YÖNETİMİ (İ.Ö.)	62.8	Hayır.	Yoksulluk Sınırı Alt	2	30.12	Başarılı	67.22	Hayır.
	MUHASEBE VE VERGİ UYGULAMALA	72.6	Hayır.	Açlık Sınırı Altında	3	0	Başarısız	53.6	Hayır.

Şekil 5.30: Öğrenci ve anket bilgilerine göre akademik öğrencilerin akademik başarı tahmin raporları

5.4.4. Aile Eğitim ve Gelir Durumuna Göre Akademik Başarı tahmini

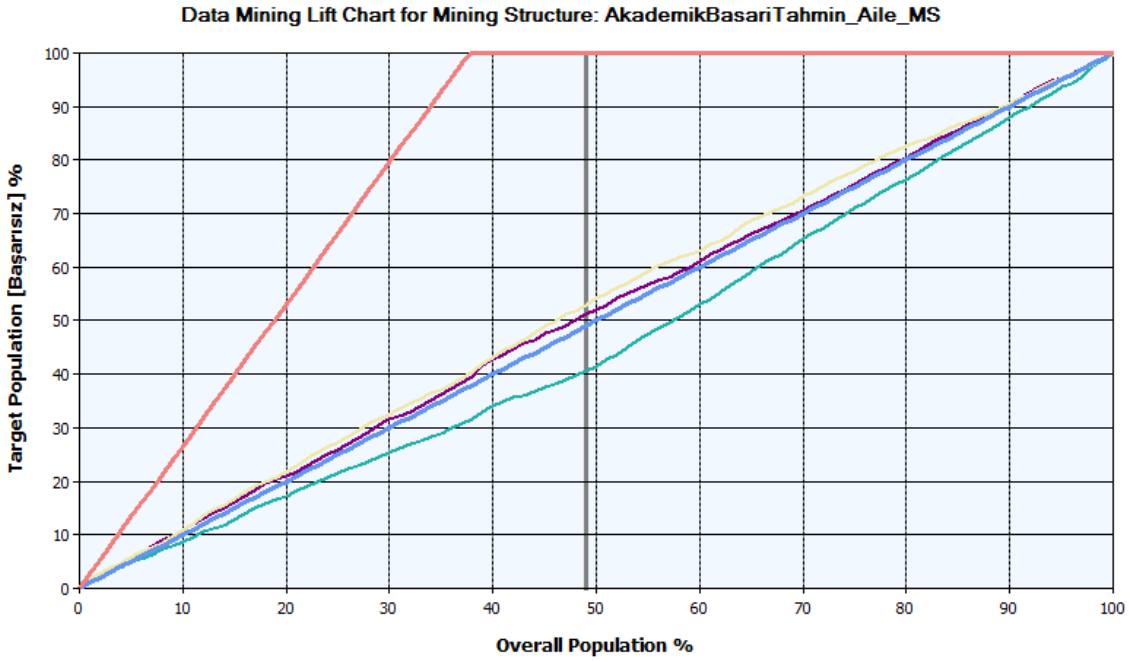
2008- 2009 yılı girişli öğrencilere yapılan anketle anne - baba gelir düzeyleri ve eğitim durumları hakkında öğrencilerden bilgiler toplanmıştır. Ailenin ekonomik durumu ve eğitim düzeyinin öğrenci üzerinde başarıyı artırıcı veya azaltıcı bir rolü olup

olmadığını belirlemek amacıyla Pamukkale Üniversitesindeki veriler analiz edilmiştir. Modeli eğitmede Tablo 5.8'de belirtilen parametreler kullanılmıştır.

Tablo 5.8: Aile gelir ve eğitim durumuna göre AB belirlemede kullanılan parametreler

Aile Disardan Geliri	Ailenin normal kazandığının yanında ekstra gelen gelirler; 250 TL den çok,101-250 TL arası,100 TL'den az, Yok değerlerini alabilir.
Aile Yasam Standardi	Ailenin gelir ve çocuk sayısına göre hesaplanan özet gelir düzeyi. Alabilecek değerler ; "Açlık Sınırı Altında", "Yoksulluk Sınırı Alt","Normal Aralıkta". Ayrıntılı açıklama Tablo 5.6'da anlatılmıştır.
Akademik Ortalama Durum	Tahmin Edilecek Değer. Akademik ortalamayı üç kısıma ayırarak tahmin işlemi yapılır. Başarısız (0 - 1,99), Başarılı (2 - 2,99), Çok Başarılı (3 - 4) olarak ayrılma gerçekleştirilmiştir.
Anne Aylik Gelir	Annenin aylık kazandığı maaşı: 2000 TL'den çok,1001-2000 TL arası,600-1.000 TL arası,600 TL'den az, Yok
Anne Egitim	Annenin mezuniyet durumu: Yükseköğretim, Ortaöğretim, İlköğretim
Baba Aylik Gelir	Babanın aylık kazandığı maaşı : 2000 TL'den çok,1001-2000 TL arası,600-1.000 TL arası,600 TL'den az,Yok
Baba Egitim	Babanın mezuniyet durumu: Yükseköğretim, Ortaöğretim, İlköğretim
Kardes Sayisi	Öğrencinin kendisi hariç kardeş sayısı: Yok, 1, 2, 3, 3'den fazla

Modeli eğitmek için 7092 kayıt kullanılmıştır. Diğer örneklerde olduğu gibi akademik ortalaması olmayan hazırlık öğrencileri eğitim verilerinin dışında tutulmuştur. Algoritmaların lift chart grafiğinde üç farklı durum içinde sonuçlar listelenmiştir. Şekil 5.31'de sistemin "Başarısız" değerini bulma grafiği gösterilmiştir.

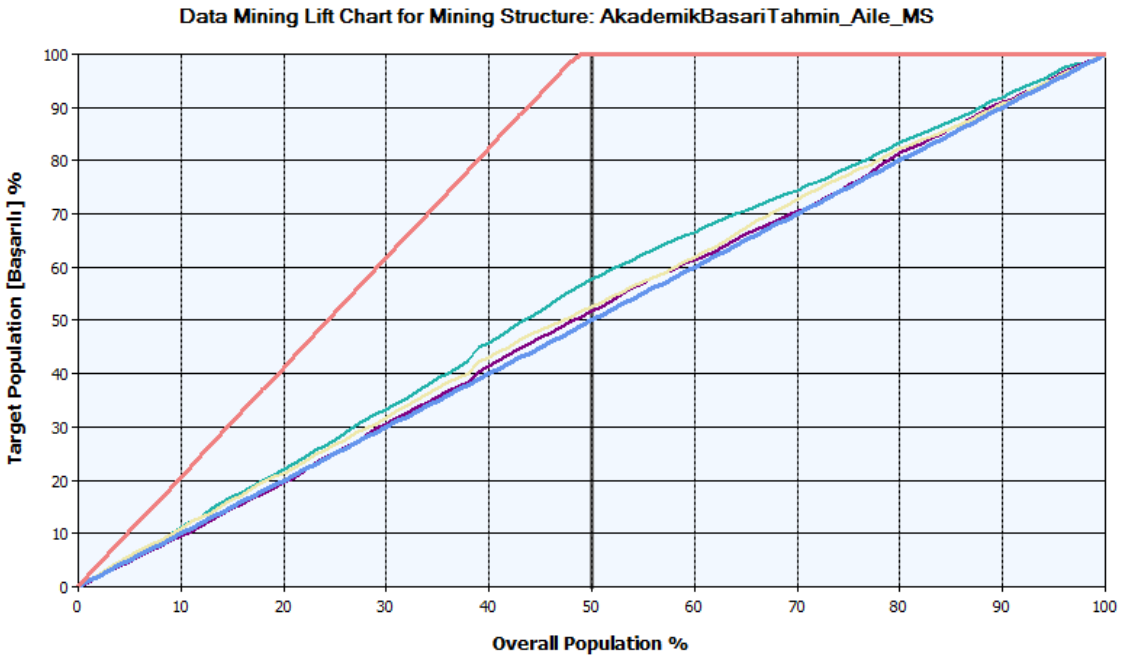


Population percentage: 49,02%

Series, Model	Score	Target population	Predict probability
AkademikBasariTahmin_Aile_DT	0,57	40,59%	36,18%
AkademikBasariTahmin_Aile_NN	0,63	51,25%	28,43%
AkademikBasariTahmin_Aile_NB	0,65	52,86%	35,57%
Random Guess Model		49,00%	
Ideal Model for: AkademikBasariTahmin...		100,00%	

Şekil 5.31: Aile eğitim ve gelir durumlarına göre AB tahmin yönteminin "Başarısız" değerleri aramadaki durumu

"Başarılı" durumundaki öğrencileri algoritmanın bulma başarısı Şekil 5. 32'deki grafikte gösterilmiştir.

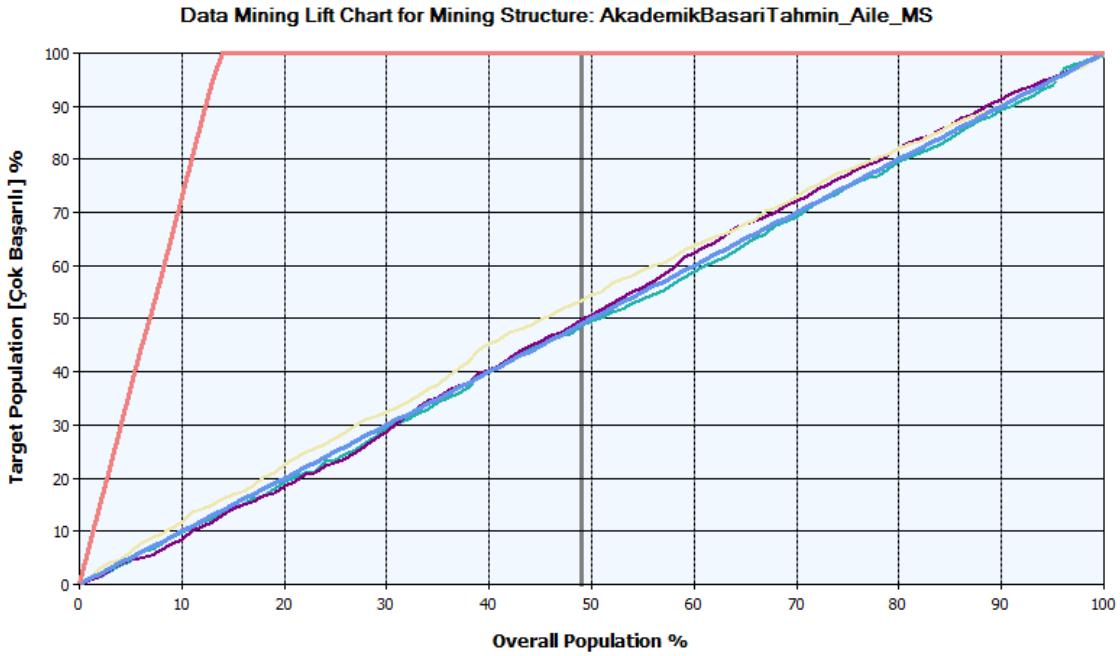


Population percentage: 50,00%

Series, Model	Score	Target population	Predict probability
AkademikBasariTahmin_Aile_DT	0,71	57,85%	50,44%
AkademikBasariTahmin_Aile_NN	0,67	51,67%	58,00%
AkademikBasariTahmin_Aile_NB	0,69	52,68%	51,40%
Random Guess Model		50,00%	
Ideal Model for: AkademikBasariTahmin...		100,00%	

Şekil 5.32: Aile eğitim ve gelir durumlarına göre AB tahmin yönteminin "Başarılı" değerleri aramadaki durumu

"Çok Başarılı" durumundaki öğrencileri algoritmanın bulma başarısı Şekil 5.33'de gösterilmiştir.



Population percentage: 50,00%

Series, Model	Score	Target population	Predict probability
AkademikBasariTahmin_Aile_DT	0,53	49,64%	13,38%
AkademikBasariTahmin_Aile_NN	0,55	50,77%	13,30%
AkademikBasariTahmin_Aile_NB	0,57	54,58%	12,97%
Random Guess Model		50,00%	
Ideal Model for: AkademikBasariTahmin...		100,00%	

Şekil 5.33: Aile eğitim ve gelir durumlarına göre akademik başarı tahmin yönteminin "Çok Başarılı" değerleri aramadaki durumu

Bu grafiklerden görüldüğü kadarıyla eldeki verilerin tahmin işleminin sağlıklı işlemesi için yeterli olmadığı görünüyor. Algoritmaların test veriler karşısında gösterdiği başarı Tablo 5.9'da belirtilmiştir.

Tablo 5.9: Aile gelir ve eğitim modeli algoritmalarının uygulanan test verileri karşısında başarısı

Microsoft Karar Ağaçları

Tahmin Edilen	Çok Başarılı (Gerçek)	Başarılı (Gerçek)	Başarısız (Gerçek)	Tahmin
Çok Başarılı	0	0	0	0,00%
Başarılı	961	3417	2623	99,10%
Başarısız	10	30	50	1,90%
	Tahmin Oranı	48,90%		

Microsoft Sinir Ağları

Tahmin Edilen	Çok Başarılı (Gerçek)	Başarılı (Gerçek)	Başarısız (Gerçek)	Tahmin
Çok Başarılı	5	13	28	0,50%
Başarılı	949	3365	2578	97,60%
Başarısız	17	69	67	2,50%
	Tahmin Oranı	48,50%		

Microsoft Naive Bayes

Tahmin Edilen	Çok Başarılı (Gerçek)	Başarılı (Gerçek)	Başarısız (Gerçek)	Tahmin
Çok Başarılı	0	0	0	0,00%
Başarılı	873	3123	2373	90,60%
Başarısız	98	324	300	11,20%
	Tahmin Oranı	48,30%		

Tablo 5.9'de tüm algoritmaların özellikle "Çok Başarılı" ve "Başarısız" öğrencilerin tahminin neredeyse tamamen etkisiz olduğu görünmektedir. Bu bize ailenin gelir ve mezuniyet durumlarının öğrenci üzerinde bir etkisini olmadığını gösterir. Her ne kadar ailenin sosyoekonomik yapısının çocuk gelişimi üzerindeki etkileri inkar edilemez olsa da bu etkilerin, ilköğretim ve ortaöğretim yıllarını içeren gelişiminin erken dönemlerinde daha belirgin olmasına karşın üniversiteye devam eden bireyler üzerinde etkisiz olduğu görülmektedir.[28]'de yaptığı çalışmasında aile eğitim düzeyinin üniversite eğitimi sırasında akademik başarıya etkisi olmadığını belirtir. Tahmin

ihtimalleri kötü olmasına rağmen bu modelle ilgili karar ağaçları yöntemiyle oluşturulan rapor Şekil 5.34'de verilmiştir.

Aile Eğitim ve Gelir Duruma Göre Akademik Başarı Belirleme

Sıra No	Cinsiyet	Program	Anne Eğitim	Anne Aylık Gelir	Baba Eğitim	Baba Aylik Gelir	Dışardan Gelen Gelir	Tahmin	Başarısız Tahmin Yüzdesi	Başarılı Tahmin Yüzdesi	Çok Başarılı Tahmin Yüzdesi
1684	Kadın	BANKACILIK VE SİGORTACILIK (İ.Ö.)	İlköğretim	600 TL'den az	İlköğretim	600 TL'den az	Yok	Başarılı	36.18%	50.44%	13.38%
1685	Kadın	SOSYAL BİLGİLER ÖĞRETMENLİĞİ (İ.Ö.)	İlköğretim	Yok	İlköğretim	600 TL'den az	Yok	Başarılı	36.18%	50.44%	13.38%
1686	Erkek	PAZARLAMA	İlköğretim	Yok	Ortaöğretim	600 TL'den az	Yok	Başarılı	36.18%	50.44%	13.38%
1687	Kadın	MATEMATİK	İlköğretim	Yok	İlköğretim	600 TL'den az	Yok	Başarılı	36.18%	50.44%	13.38%
1688	Kadın	SİYASET BİLİMİ VE KAMU YÖNETİMİ	İlköğretim	Yok	Ortaöğretim	Yok	Yok	Başarılı	36.18%	50.44%	13.38%
1689	Kadın	FELSEFE (İ.Ö.)	Ortaöğretim	Yok	Ortaöğretim	600-1.000 TL arası	Yok	Başarılı	36.18%	50.44%	13.38%
1690	Erkek	JEOLOJİ MÜHENDİSİ İÇİ	Yükseköğretim	Yok	İlköğretim	1001-2000 TL arası	Yok	Başarılı	36.18%	50.44%	13.38%

Şekil 5.34: Aile eğitim ve gelir durumlarına göre öğrencilerin akademik başarı tahmin raporu

5.4.5. Öğrencilerin ders başarılarının tahmin edilmesi

Şu ana kadar yapılan çalışmalarda çeşitli parametrelere göre öğrencilerin akademik başarıları üç farklı algoritmayla belirlenmeye çalışılmıştır. Her problem için üç farklı algoritma kullanılmış, bunlar arasından en iyi performansı veren algoritma tahmin işleminde kullanılmak üzere seçilmiştir. Daha sonra karar vericilerin oluşturulan bu alt yapıyı kullanabilmesi içinde raporlar hazırlanmıştır.

Bu modelde ise öğrencilerin ders bazında gösterecekleri başarıların tahmin edilmesi amaçlanmıştır. Öğrencilerin belirli derslerden aldıkları notları, başka derslerin başarısını etkileyip etkilemediği sorgulanmıştır. Eğitim kümesi olarak 2009 yılında üniversiteye kayıt olan ve hazırlık okumayan 52 Bilgisayar Mühendisi öğrencisi ve onların 2009 yılına ait güz ve bahar dönenimde aldıkları dersler kullanılmıştır. Modelde, diğer algoritmalardan farklı olarak Microsoft Birliktelik Kuralları algoritmasıyla işlem yapılmıştır.

Diğer modellerde sadece ana tablo bazlı (case table) modeller oluşturulmuştur. Bu modelde ise ana tabloların yanında bağlı tablolar (nested table) da modellin içinde yer almaktadır. Her öğrenci için 2009 güz ve bahar döneminde aldığı dersler ayrı ayrı

nested tablo olarak tanımlanarak işleme sokulmuştur. Modelde kullanılan parametrelerin açıklaması Tablo 5.10'da belirtilmiştir.

Tablo 5.10: Öğrenci ders başarıları modelinde kullanılan parametreler

Güz döneminde alınan dersler	2009 güz döneminde alınan dersler belirtilmektedir. Oluşturulan ders kavramı iki bilginin birleştirilmesi ile oluşturulmuştur; dersin adı ve öğrencinin dersle ilgili dönem sonu başarı notu. Bu iki verinin birleştirilmesi ile modelde kullanılacak ders alanı oluşturulmuştur. Bu parametre sisteme giriş değerlerini oluşturmaktadır. Bu sayede güz döneminde aldıkları dersler ve notlarına göre bahar dönemi derslerin tahmini mümkün kılınmıştır.
Bahar döneminde alınan dersler	2009 bahar döneminde alınan dersler belirtilmektedir. Oluşturulan ders kavramı iki bilginin birleştirilmesi ile oluşturulmuştur; dersin adı ve öğrencinin dersle ilgili dönem sonu başarı notu. Bu iki verinin birleştirilmesi ile modelde kullanılacak ders alanı oluşturulmuştur. Tahmin alanı bu kolondur.

Modelde kullanıldığı için Pamukkale Üniversitesi not sistemi açıklanması önem arz etmektedir. Tablo 5.11'de not sistemi özet olarak gösterilmiştir.

Tablo 5.11: Notlar ve anlamları

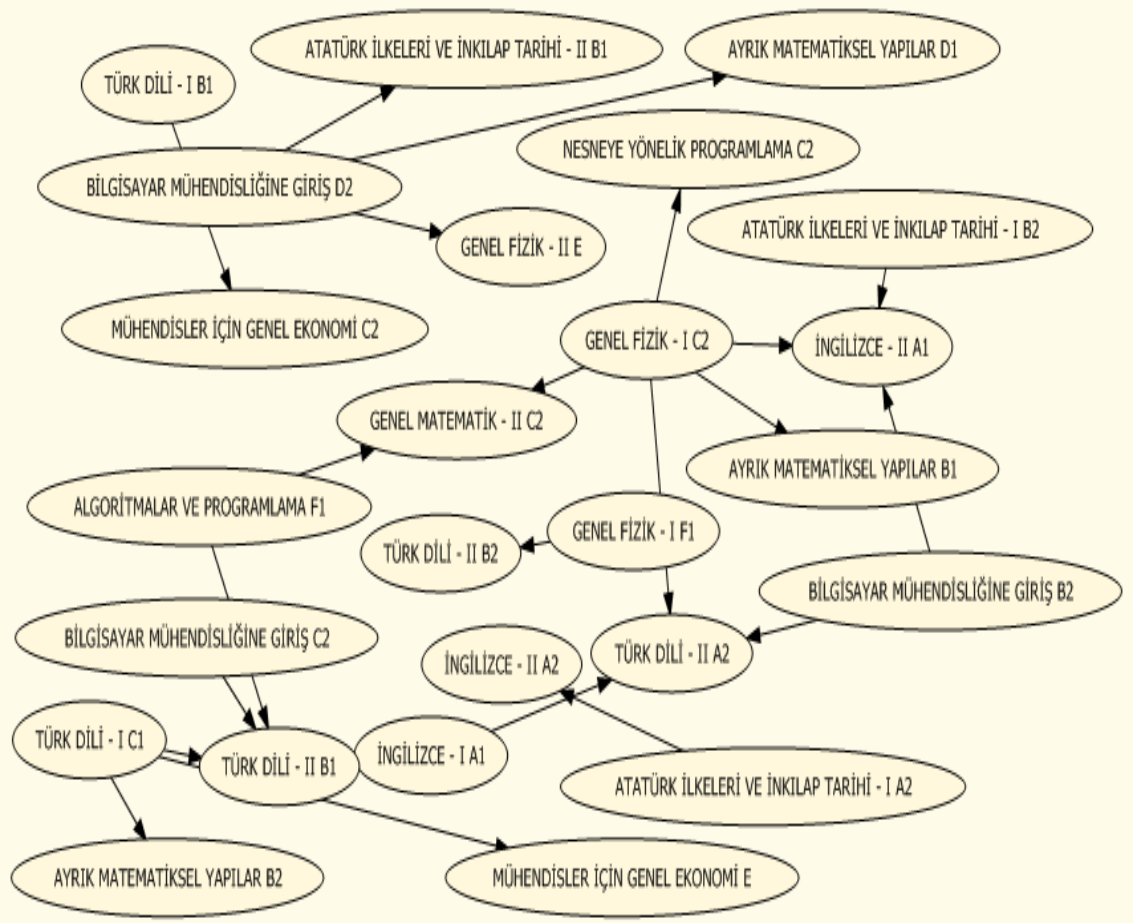
Not Aralığı	**Harf Karşılık**	**Sonuç**
90-100	A1	Başarılı
80-89	A2	Başarılı
75-79	B1	Başarılı
70-74	B2	Başarılı
65-69	C1	Başarılı
60-64	C2	Başarılı
55-59	D1	Koşullu Geçer
50-54	D2	Koşullu Geçer
40-49	E	Başarısız
0-39	F1	Başarısız
Finale Girmedi	F2	Başarısız

Birliktelik kurallarını kullanarak öğrencilerin dersleri arasında bulunan 46 adet kural belirlenmiş, bu kurallardan yüksek olasılığa sahip olanlar Tablo 5.12'de belirtilmiştir.

Tablo 5.12: Öğrencilerin ders notları arasında bulunan bağıntılar

Olasılık	Güven	Kural
1	0,82391	GENEL FİZİK - I F1 -> TÜRK DİLİ - II B2
1	0,82391	GENEL FİZİK - I F1, BİLGİSAYAR MÜHENDİSLİĞİNE GİRİŞ D2 -> TÜRK DİLİ - II B2
1	0,69897	ATATÜRK İLKELERİ VE İNKILAP TARİHİ - I A2, İNGİLİZCE - I A1 -> İNGİLİZCE - II A2
1	0,69897	TÜRK DİLİ - I C1, ALGORİTMALAR VE PROGRAMLAMA F1 -> TÜRK DİLİ - II B1
1	0,69897	BİLGİSAYAR MÜHENDİSLİĞİNE GİRİŞ B2, GENEL FİZİK - I C2 -> İNGİLİZCE - II A1
1	0,60206	BİLGİSAYAR MÜHENDİSLİĞİNE GİRİŞ C2 -> TÜRK DİLİ - II B1
1	0,60206	BİLGİSAYAR MÜHENDİSLİĞİNE GİRİŞ B2, İNGİLİZCE - I A1 -> TÜRK DİLİ - II A2
1	0,60206	TÜRK DİLİ - I C1, GENEL MATEMATİK - I E -> TÜRK DİLİ - II B1
1	0,60206	GENEL MATEMATİK - I E, ALGORİTMALAR VE PROGRAMLAMA F1 -> TÜRK DİLİ - II B1
1	0,52288	BİLGİSAYAR MÜHENDİSLİĞİNE GİRİŞ B2, GENEL FİZİK - I C2 -> TÜRK DİLİ - II A2
1	0,52288	BİLGİSAYAR MÜHENDİSLİĞİNE GİRİŞ B2, GENEL MATEMATİK - I C2 -> TÜRK DİLİ - II A2
1	0,45593	TÜRK DİLİ - I B1 -> MÜHENDİSLER İÇİN GENEL EKONOMİ C2
0,8	0,73848	GENEL FİZİK - I C2, İNGİLİZCE - I A1 -> İNGİLİZCE - II A1
0,8	0,61354	TÜRK DİLİ - I C1 -> TÜRK DİLİ - II B1
0,8	0,51663	BİLGİSAYAR MÜHENDİSLİĞİNE GİRİŞ B2 -> TÜRK DİLİ - II A2
0,8	0,51663	GENEL FİZİK - I C2, İNGİLİZCE - I A1 -> TÜRK DİLİ - II A2
0,75	1,20412	GENEL FİZİK - I C2, GENEL MATEMATİK - I C2 -> NESNEYE YÖNELİK PROGRAMLAMA C2
0,75	0,727	TÜRK DİLİ - I C1, ALGORİTMALAR VE PROGRAMLAMA F1 -> MÜHENDİSLER İÇİN GENEL EKONOMİ E
0,75	0,60206	BİLGİSAYAR MÜHENDİSLİĞİNE GİRİŞ B2, İNGİLİZCE - I A1 -> İNGİLİZCE - II A1
0,75	0,60206	GENEL MATEMATİK - I C2, ALGORİTMALAR VE PROGRAMLAMA F1 -> GENEL MATEMATİK - II C2
0,667	0,83727	ALGORİTMALAR VE PROGRAMLAMA F1, İNGİLİZCE - I A1 -> GENEL FİZİK - II D1

Derslerin şekil üzerinde birbiriyle olan ilişkisi ise Şekil 5.35'da gösterilmiştir.



Şekil 5.35: Dersler ve notlar arasındaki bağıntılar

Bu modellerin raporlanması için de diğer raporlar gibi DMX sorguları kullanılmıştır. Yalnız burada önemli fark, önceki raporlarda sadece ana tablo(case table) ifadesi alan sorgular, burada bir ana tablo, iki tanede bağlı tablo almaktadır. Yazılan sorgu şu şekildedir;

```
SELECT FLATTENED
  (t.[ogrNo]) as [OgrNo],
  (t.[isim]) as [İsim],
  (t.[cinsiyet]) as [Cinsiyet],
  (t.[ogr_yas]) as [Yaş],
  (t.[akademikOrtalamaDurum]) as [Akademik Ortalama],
  (PredictAssociation([BilgisayarMuhDersTahmin_AR].[Bilgisayar Muh Bahar Dersler],30))
as [Tahmin]
From
  [BilgisayarMuhDersTahmin_AR]
PREDICTION JOIN
  SHAPE {
  OPENQUERY([OKDS_DS],
    'SELECT
      [ogrNo],
      [isim],
      [cinsiyet],
      [ogr_yas],
      [akademikOrtalamaDurum],
```

```
      [ogr_id]
     FROM
      (
       SELECT o.ogr_id, o.ogrNo, o.ad + '' '' + o.soyad AS isim, o.cinsiyet, o.ogr_yas,
o.akademikOrtalamaDurum
       FROM Kds.dt_ogrenci AS o INNER JOIN
            Kds.dt_programDal AS pd ON pd.programDal_id = o.programDal_id INNER JOIN
            Kds.dt_program AS p ON p.prg_id = pd.prg_id
              WHERE (o.akademikOrtalama > 0)
                     AND (p.prg_kod IN (253))
                     AND EXISTS
                     (SELECT id
                      FROM Kds.dt_ogrenciDersler AS od2
                      WHERE (ogr_id = o.ogr_id) AND (donem_id IN (151))) AND
(YEAR(o.girisTarihi) IN (2009)
                     )
         ) [BilgisayarMuhDersTahmin_Test]
    ORDER BY
      [ogr_id]')}
  APPEND
  ({OPENQUERY([OKDS_DS],
    'SELECT
      [alinanDers],
      [ogr_id]
     FROM
     (
               SELECT od.id, od.ogr_id, d.ders_ad + '' '' + dpc.ad AS alinanDers
               FROM Kds.dt_ogrenciDersler AS od INNER JOIN
                       Kds.lu_ders AS d ON d.ders_id = od.ders_id INNER JOIN
                       Kds.lu_dersPuanCetveli AS dpc ON dpc.id = od.dersPuanCetveli_id
INNER JOIN
                       Kds.lu_donem AS do ON do.donem_id = od.donem_id INNER JOIN
                       Kds.dt_ogrenci AS o ON o.ogr_id = od.ogr_id INNER JOIN
                       Kds.dt_programDal AS pd ON pd.programDal_id = o.programDal_id
INNER JOIN
                       Kds.dt_program AS p ON p.prg_id = pd.prg_id
               WHERE (od.donem_id IN (151))
                      AND (p.prg_kod IN (253))
                      AND (dpc.ad NOT LIKE ''M%'')
                      AND (YEAR(o.girisTarihi) IN (2009))
       ) [BilgisayarMuhGuzDersler_Test]
    ORDER BY
      [ogr_id]')}
    RELATE
      [ogr_id] TO [ogr_id])
    AS
      [BilgisayarMuhGuzDersler_Test] AS t
ON
  [BilgisayarMuhDersTahmin_AR].[Bilgisayar Muh Guz Dersler].[Alinan Ders] =
t.[BilgisayarMuhGuzDersler_Test].[alinanDers]
```

Yapıda “FROM” hangi model üzerinden işlemde yapıldığını gösterir. “PREDICTION” ise sorgunun ana yapısının oluşturur ve bir tahmin işleminin yapılacağını belirtir. “OPENQUERY” ifadesi modeli dolduracak veri kaynaklarının alınması için bir bağlantı oluşturur. “SHAPE” ifadesi ana tabloya(case table) , “APPEND” ise bağlı tabloya (nested table) modelden sonuç alabilmek için gerekli olan verileri sağlar. “OPENQUERY” den sonra gelen ifadeler standart T-SQL(Transact-SQL) ifadesidir. “RELATE” anahtar sözcüğü ile case-nested tablolar arasındaki ilişkiyi sisteme açıklar. Son olarak da “ON” ifadesi ile bağlı tablodan gelen değerler ile modelin verileri eşleştirilerek tahmin gerçekleştirilir.

"SELECT" ifadesi bize DMX yapısında hangi sonuçların getirileceğini belirtir. Kimlik bilgileri sayılan "ogrNo", "İsim", "Cinsiyet" gibi alanlar herhangi bir işlemden geçmeden direkt ana tablo kolonlarından yazdırılır. "PredictAssociation" ifadesi, tahmin değerleri arasından öğrenci için en iyi tahmin olan 30 değeri getirir. Son olarak "FLATTENED" ifadesi, verilerin düz liste biçiminde bir sonuç kümesi olarak dönmesini sağlar.

Modelin rapor olarak sunumu Şekil 5.36'da verilmiştir.

Sıra No	Cinsiyet	Yaş	Akademik Ortalama	Tahmin Alinan Ders
50	Erkek	18	Çok Başarılı	GENEL FİZİK - II D2
51	Erkek	18	Çok Başarılı	AYRIK MATEMATİKSEL YAPILAR C2
52	Erkek	18	Çok Başarılı	GENEL MATEMATİK - II F1
53	Erkek	18	Çok Başarılı	GENEL MATEMATİK - II E
54	Erkek	18	Çok Başarılı	AYRIK MATEMATİKSEL YAPILAR B2
55	Erkek	18	Çok Başarılı	AYRIK MATEMATİKSEL YAPILAR B1
56	Erkek	18	Çok Başarılı	TÜRK DİLİ - II C1
57	Erkek	18	Çok Başarılı	NESNEYE YÖNELİK PROGRAMLAMA C2
58	Erkek	18	Çok Başarılı	NESNEYE YÖNELİK PROGRAMLAMA B1
59	Erkek	18	Çok Başarılı	GENEL MATEMATİK - II C1
60	Erkek	18	Çok Başarılı	BİLGİSAYAR MÜHENDİSLİĞİ SEMİNERİ B1
61	Kadın	20	Başarılı	TÜRK DİLİ - II A2
62	Kadın	20	Başarılı	İNGİLİZCE - II A1
63	Kadın	20	Başarılı	BİLGİSAYAR MÜHENDİSLİĞİ SEMİNERİ A1
64	Kadın	20	Başarılı	NESNEYE YÖNELİK PROGRAMLAMA F1
65	Kadın	20	Başarılı	GENEL FİZİK - II D1
66	Kadın	20	Başarılı	ATATÜRK İLKELERİ VE İNKILAP TARİHİ - II A2

Şekil 5.36: Dersler ve ders notları tahmin raporu

6. SONUÇLAR VE DEĞERLENDİRME

Tez kapsamında Öğrenci İşleri Otomasyon Sistemi üzerinde bulunan verilerin incelenmesi yapılmaktadır. Oluşturulan veri ambarı ve kullanılan veri madenciliği modelleri vasıtasıyla öğrencilerin akademik ortalamaları ve ders başarıları tahmin edilmeye çalışılmıştır. Bu amaçla veriler analiz için bir Veri Ambarı yapısında toplanmış ve çeşitli giriş parametrelerine göre Veri Madenciliği algoritmaları kullanılarak anlamlı sonuçlar elde edilmiştir. Analiz işlemi için birden çok model kullanılmıştır. Bu sayede probleme göre modellerin başarısı belirlenmiş ve en iyi performansa sahip modelin sistemde kullanılabilmesi mümkün olmuştur. Üniversite yönetiminin bu sonuçlardan yaralanması için, hazırlanan raporlar Karar Destek Sistemi(KDS) altında kullanıma sunulmuştur.

Üniversitede sadece statik olarak yer kaplayan verilerden üniversitenin gelişmesine yönelik çalışmalarda kullanılmak üzere sonuçlar çıkarılmıştır. Bu sonuçlarla cinsiyetin akademik başarıya etkisi araştırılmış ve değerlerin programdan programa değişkenlik göstermesine rağmen kız öğrencilerin üniversite bünyesinde daha başarılı olduğu saptanmıştır. Öğrencilerin hangi programlarda akademik performans olarak daha fazla zorlandığı ve kötü notlar aldığı, hangi programlarda başarıların yüksek olabileceği tahmin edilebilir olmuştur.

Öğrenci kimlik bilgileri ve anket sonuçlarına göre tahmin işlemleri yapılmış, öğrenci akademik başarısını yordamak amacıyla birçok girdi değerine göre doğru çıktı alınmaya çalışılmıştır. Mevcut veya üniversiteye yeni kayıt olmuş öğrencilerin akademik başarıları hakkında, üniversite yönetimi karar destek sistemiyle bir tahmin şansı yakalamıştır. Bu sayede risk grubundaki öğrenciler önceden belirlenerek, sorun yaşayacak öğrenciler üzerine önlem alma çalışmaları başlatılabilir.

Ders başarılarının birbiriyle ilişkileri bulunmasıyla, öğrencilerin hangi derslere dikkat etmeleri, hangi derslere daha fazla çaba sarf etmeleri gerektiği veya hangi derslerin kişiye uygun olup olmayacağı KDS'nin dersler notlarını tahmin işlemiyle gerçekleşebilir. Bu sayede öğrencinin de genel akademik başarısının artacağı düşünülmüştür.

Tez bünyesinde araştırılan her özellik beklendik sonuç vermemiştir. Ailenin eğitim ve gelir durumlarıyla ilgili yapılan analizde ailenin eğitim düzeyi ve gelir düzeyi gibi sosyo-kültürel özelliklerinin üniversite öğrencilerinin akademik başarıları üzerindeki etkilerinin önemli olmadığı tespit edilmiştir. Yani ebeveynlerin eğitim seviyeleri ve gelirlerinin çok yüksek olması öğrenci başarısı üzerinde belirleyici bir etkiye sahip değildir.

Sistemi oluştururken kullanılan algoritmaların tahmin yüzdelerinin akademik başarı (AB) tahmininde çok yüksek seviyelere çıkmadığı görülmektedir. Bunun sebebi olarak, sistemi analiz ederken daha özelleşmiş veri grupları(örnek olarak belli sayıda programlar) yerine tüm üniversite verileriyle çalışılmış olması gösterilebilir. Ayrıca öğrencilerin anket cevaplarına göre AB tahmin işleminde, anketlerin istenilen performansı göstermediği düşünülmektedir. Bunun nedeni olarak anket sorularının bu şekilde bir çalışmada kullanılmak için hazırlanmamış olmasının etkisi büyüktür. İleride anket sorularının da özelleşerek karar destek sistemine daha yarar sağlayacak seviyeye getirileceği düşünülmektedir. Ayrıca çalışmanın bir sonraki basamağında her program için özelleşmiş bir veri madenciliği modeli kullanıp öğrencilere uygulanacak yeni anketler yaparak modellerin başarı durumunu maksimize edilmesi amaçlanmıştır.

Özet olarak anlatılacak olursa, bilgi çağımız olan günümüzde üniversitenin gerek kişilik gerek mesleki gelişimdeki önemine ilişkin yaygın farkındalık her geçen gün daha fazla bireyin eğitim öğretim yaşantısına üniversiteyle devam etme kararı almasına neden olmaktadır. Bireylerin üniversiteye talebinin bir sonucu olarak da sayısı gittikçe artan üniversitelere daha fazla öğrenci kabul edilmektedir. Bu durum ister istemez akademik olarak daha az hazırlıklı daha fazla sayıda öğrencinin üniversiteye giriş yapmasına neden olmaktadır. Bu kalabalık öğrenci grubu içinde akademik olarak yardıma ihtiyaç duyan öğrencileri ilk yıllarında tespit etme ve ihtiyaç duyulan rehberliğin zamanında sağlanması açısından geliştiren Karar Destek Sisteminin bu işlemlerde yardımcı olacağı açıktır. Böylelikle bu tez çalışmasının gerek Öğrenci İşleri,

gerekse Üniversite Yönetimi tarafından kullanımının üniversite içersinde kaliteyi artıracağı düşünülmektedir. Tezin temel amacı da karar almayı kolaylaştırıcı sistem tasarlamak olduğu için hedeflenen değerlere ulaşıldığı söylenebilir.

KAYNAKLAR

[1] **Seetharaman, M.**, 2008: Data Warehousing: Case study in data quality improvement, School of Information Systems and Engineering Technology, Institute of Technology, State University of New York.

[2] **Düzgünoğlu, S.**, 2006: Veri Ambarı ve OLAP Teknolojilerinden Yararlanılarak Karar Destek Amaçlı Raporlama Aracı Gerçekleştirimi, Yüksek Lisans Tezi, Bilgisayar Mühendisliği Anabilim Dalı, Fen Bilimleri Enstitüsü, Hacettepe Üniversitesi

[3] **Whiting, R.**, 2004: Bigger&Better. Information week <http://www.informationweek.com/news/storage/showArticle.jhtml?articleID=18400975>, alındığı tarih 04.05.2010

[4] **Inmon, W.H.**, 2002: Building the Data Warehouse. Wiley Computer Publishing

[5] **Aköz, E.**, 2007: Otomotiv Sektöründe Veri Ambarları ve Bir Uygulama, Matematik Bilgisayar Ana Bilim Dalı, Fen Bilimleri Enstitüsü, Beykent Üniversitesi

[6] **İşli, D.**, 2009: Veri Ambarı ve OLAP Teknolojilerinden Yararlanılarak Raporlama Aracı Gerçekleştirimi, Bilgisayar Mühendisliği Anabilim Dalı, Fen Bilimleri Enstitüsü, Pamukkale Üniversitesi

[7] **Türkmen, G.**, 2007: Developing A Data Warehouse For A University Decision Support System, Department of Computer Engineering, Graduate School of Natural and Applied Sciences, Atılım University

[8] **Döşlü, A.**, 2008: Veri Madenciliğinde Market Sepet Analizi ve Birliktelik Kurallarının Belirlenmesi, Bilgisayar Mühendisliği Anabilim Dalı, Fen Bilimleri Enstitüsü, Yıldız Teknik Üniversitesi

[9] **Giran, Ö.**, 2008: İnşaat İşletme Yönetiminde Bilgi Ambarlama, İnşaat Mühendisliği Anabilim Dalı, Fen Bilimleri Enstitüsü, İstanbul Üniversitesi

[10] **Turban, E.**, 1995: Decision support and expert systems: management support systems, Englewood Cliffs, N.J., Prentice Hall

[11] **Çetinyokuş, T.**, Gökçen H., 2002: Borsada Göstergelerle Teknik Analiz İçin Bir Karar Destek Sistemi, Gazi Üniv. Müh. Mim. Fak. Der. Cilt 17, No 1, 43-58

[12] **İç, Y., T.**, 2006: İşleme Merkezlerinin Seçiminde Kullanılacak Bir Karar Destek

Sisteminin Geliştirilmesi, Makine Mühendisliği Anabilim Dalı, Fen Bilimleri Enstitüsü, Gazi Üniversitesi

[13] **Cloyd, J., D.**, 2002: Data Mining with Newton's Method, Department of Computer and Information Sciences, East Tennessee State University

[14] **Al-Hudairy, H., H.,M.,A., A.**, 2004: Data Mining and Decision Making Support in The Governmental Sector, Department of Computer Engineering an Computer Science, University of Louisville

[15] **Shen, Y.**, 2007: A Formal Ontology for Data Mining: Principles, Design, and Evolution ,Département de mathématiques et d'informatique ,Universite du Quebec a Trois-Rivieres

[16] **Silahtaroğlu, G.**, 2008: Kavram ve Algoritmalarıyla Veri Madenciliği, Papatya Yayıncılık Eğitim

[17] **Sezer, Ü.**, 2008: Karar Ağaçlarının Birliktelik Kuralları ile İyileştirilmesi, Bilgisayar Mühendisliği Anabilim Dalı, Fen Bilimleri Enstitüsü, Kocaeli Üniversitesi

[18] **Özkan, Y.**, 2008: Veri Madenciliği Yöntemleri, Papatya Yayıncılık Eğitim

[19] **Ayık, Y.,Z., Özdemir, A., Yavuz, U.**, 2007: Lise Türü ve Lise Mezuniyet Başarısının, Kazanılan Fakülte ile İlişkisinin Veri Madenciliği Tekniği ile Analizi, Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, Cilt 10, Sayı 2

[20] **Zhu, J**, 2003: Sampling- An Efficient Solution For Data Mining Of Association Rules, Department of Computer Science, Dalhousie University

[21] **Akpınar, H.**, 2000: Veritabanlarında Bilgi Keşfi ve Veri Madenciliği, İstanbul Üniversitesi, İşletme Fakültesi Dergisi, c:29, s:1-22

[22] **Özçınar, H.**, 2006: KPSS Sonuçlarının Veri Madenciliği Yöntemleriyle Tahmin Edilmesi, Bilgisayar Mühendisliği, Fen Bilimleri Enstitüsü, Pamukkale Üniversitesi

[23] **Ho Yu, C., DiGangi, S., Jannasch-Pennell, A., Kaprolet, C.**, 2010: A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year, Journal of Data Science 8(2010), 307-325

[24] **Tantuğ, A., C.**, 2002: Veri Madenciliği ve Demetleme, Bilgisayar Mühendisliği Anabilim Dalı, Elektrik-Elektronik Fakültesi, İstanbul Teknik Üniversitesi

[25] **Han, J., Kamber, Kamber, M.**, 2006: Data Mining Concepts and Techniques,

Morgan Kaufmann

[26] **Erdoğan, Ş., Z., Timor, M.**, 2005: A Data Mining Application In a Student Database. Journal Of Aeronautics and Space Technologies, July 2005 Volume 2 Number 2 (53-57)

[27] **Bozkır, A., S., Sezer, E., Gök, B.,** 2009: Öğrenci Seçme Sınavında(ÖSS) Öğrencilerin Başarımını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti, Uluslar arası İleri Teknolojiler Sempozyumu (İATS'09)

[28] **Vandamme, J., P., Meskens, N., Superby, J., F.**, 2007: Predicting Academic Performance by Data Mining Methods , Education Economics , Vol. 15 , 405-419.

[29] **Özdamar, E. Ö.**, 2002: Veri Madenciliğinde Kullanılan Teknikler ve Bir Uygulama, İstatistik Anabilim Dalı, Fen Bilimleri Enstitüsü, Mimar Sinan Üniversitesi

[30] **Karabatak, M., İnce, M., C.**, 2004: Apriori Algoritması ile Öğrenci Başarısı Analizi, Elektrik Elektronik Bilgisayar Mühendisliği Sempozyumu, ELECO 2004, Elektronik-Bilgisayar sayfa 348-352

[31]**Url_1**:<http://www.business.com/directory/computers_and_software/software_applications/data_management/data_mining/>, alındığı tarih 18.07.2010

[32]**Url_2**:<http://msdn.microsoft.com/en-us/library/ms175382.aspx>, alındığı tarih 19.07.2010

[33]**Url_3**:<http://msdn.microsoft.com/en-us/library/ms174916.aspx>, alındığı tarih 19.07.2010

[34]**Url_4**:<http://msdn.microsoft.com/en-us/library/ms174941.aspx>, alındığı tarih 19.07.2010

[35] **Maclennan, J., Tang, Z., Crivat, B.**, 2009: Data Mining with Microsoft SQL Server 2008, Wiley Publishing

[36] **Kalles, D., Pierrakeas, C.**, 2006: Analyzing Student Performance in Distance Learning With Genetic Algorithms and Decision Trees, Applied Artificial Intelligence, Volume 20, Issue, pages 655 - 674

[37] **Al-Radaideh, Q., A., Al-Shawakfa, E., M., Al-Nijjar, M., I.**, 2006: Mining Student Data Using Decision Trees, International Arab Conference on Information Technology

[38] **Güneri, N., Apaydın, A.**, 2004: Öğrenci Başarılarının Sınıflandırılmasında

Lojistik Regresyon Analizi ve Sinir Ağları Yaklaşımı.Ticaret ve Turizm Eğitim Fakültesi Dergisi,Yıl: 2004, Sayı: 1, ss. 170 – 188.

[39] **Kovačić, Z., J.**, 2010: Early Prediciton of Student Success: Mining Students Enrolment Data, Proceedings of Informing Science & IT Education Conference (InSITE)

[40] **Bozkır, A., S.**, Gök B., Sezer E., 2008: Üniversite Öğrencilerinin İnterneti Eğitimsel Amaçlar İçin Kullanmalarını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti, Bilimde Modern Yöntemler Sempozyumu

[41] **Yang, M.**, 2006: Data Mining Techniques Applied to Texas Woman's University's Enrollment Data – What Can The Data Tell Us?, College of Arts and Sciences, Texas Woman's University

[42] **Alpaydın, E.**, 2000: Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri, Bilişim 2000 Eğitim Semineri

[43]**Url_5**: <http://www.cnnturk.com/2008/ekonomi/genel/08/29/hukumet.ile.memurlar.zam.konusunda.anlasti/491913.0/index.html>, alındığı tarih 05.05.2010

[44]**Url_6**: <http://www.memurlar.net/haber/147674/>, alındığı tarih 05.05.2010

[45]**Url_7**: : <http://www.memursen.org.tr/file/A%C3%A7l%C4%B1k%20ve%20Yoksulluk%20S%C4%B1n%C4%B1r%C4%B1%20Raporu%20-%20Mart%202010%20-%20T%C3%BCrkiye%20Verileri.doc>, alındığı tarih 06.05.2010

ÖZGEÇMİŞ

Ad Soyad: Gürler Gülçe

Doğum Yeri ve Tarihi: Uşak - 08.11.1983

Adres: Fatih Mah. Huzurkent B-1 Blok Daire: 5 Uşak

Lisans Üniversite: Pamukkale Üniversitesi

Bilgisayar Mühendisliği(2007)

Yayın Listesi: -